Redbooks
ibm.com/redbooks

IBM Cloud Pak for Data Version 4.5

A practical, hands-on guide with best practices, examples, use cases, and walk-throughs

Simon Cambridge
Lakshmana Ekambaram
Stephen D. Gawtry
Vasfi Gucer
Audrey Holloman
Frank Ketelaars
Darren King
Karen Medhat
Mark Moloney
Payal Patel
Neil Patterson
Deepak Rangarao
Mark Simmonds
Malcolm Singh
Tamara Tatian
Henry L.Quach

Cloud

Data and AI

IBM

Redbooks

IBM Redbooks

IBM Cloud Pak for Data Version 4.5: A practical, hands-on guide

November 2022

SG24-8522-00

Note: Before using this information and the product it supports, read the information in “Notices” on page ix.

First Edition (November 2022)

This edition applies to IBM Cloud Pak for Data Version 4.5.

Contents

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

Foreword

Thank you for reading this IBM Redbooks publication. *IBM Cloud Pak for Data Version 4.5: A practical, hands-on guide with best practices, examples, use cases, and walk-throughs* was created at the request of customers whom I worked with over the years.

Organizations long recognized the value that IBM Redbooks provide in guiding them with best practices, frameworks, and hands-on examples as part of their solution implementation.

This book is a collaboration involving many skilled and talented authors that were selected from our IBM global technical sales, development, Expert Labs, Client Success Management, and consulting services organizations, using their diverse skills, experiences, and technical knowledge of IBM Cloud Pak for Data.

I want like to thank the authors, contributors, reviewers, and the IBM Redbooks team for their dedication, time, and effort in making this publication a valuable asset that organizations can use as part or their journey to AI.

I also want to thank Mark Simmonds and Deepak Rangarao for taking the lead in shaping this request into yet another successful IBM Redbooks project.

It is my sincere hope that you enjoy *Hands on with IBM Cloud Pak for Data* as much as the team who wrote and contributed to it.

Steve Astorino, IBM VP Development Data and AI, and Canada Lab Director

Preface

IBM® Cloud Pak for Data platform is a unified, integrated and collaborative data and artificial intelligence (AI) platform that helps organizations collect, organize, and analyze data and infuse AI within and across business process and applications. Its ultimate goal is to deliver deeper business insights that enable smarter business outcomes.

IBM Cloud Pak for Data delivers a set of capabilities core to a data fabric. A data fabric can help organizations improve productivity and reduce complexities when accessing, managing, and understanding disparate, siloed data that is distributed across a hybrid cloud landscape.

The platform offers a wide selection of IBM and third-party services that span the entire data lifecycle. Deployment options include an on-premises software version that is built on the Red Hat® OpenShift® Container Platform, and a fully or partially managed version that can run on the IBM Cloud, and other hyper-scalers, such as Amazon Web Services (AWS) and Microsoft Azure.

This IBM Redbooks® publication provides a broad understanding of the IBM Cloud Pak for Data concepts and architecture, and the services that are available in the product.

In addition, several common use cases and hands-on scenarios are included that help you better understand the capabilities of this product.

Code samples for these scenarios are available at this GitHub web page.

This publication is for IBM Cloud Pak for Data customers who seek best practices and real-world examples of how to best implement their solutions while optimizing the value of their existing and future technology, data, and skills investments.

Note: This book is based on IBM Cloud Pak for Data Version 4.5.

Authors

This book was produced by a team of specialists from around the world.

Simon Cambridge is a Principal Customer Success Manager with IBM in North America. He has over 25 years of experience in the computing industry, working on data and AI solutions. He holds multiple product and industry certifications from IBM, Amazon Web Services (AWS), Microsoft Azure, Nvidia, and Snowflake. Simon's current role focuses on customer success initiatives with strategic partners, including cloud service providers and global systems integrators. Simon specializes in data fabric, management, analytics, and MLOps solutions. He holds a Bachelor of Business degree in Business Information Systems from Massey University in New Zealand.

Lakshmana Ekambaram is an IBM Senior Technical Leader with over 30 years of experience in database development, advanced analytics, and building Hybrid cloud solutions. He is currently part of the IBM Expert Labs SWAT organization, leading the data fabric and trusted AI journey for customers worldwide. He has developed many IBM certification courses and co-authored books about data science, AI, and data fabric.

Stephen D. Gawtry is a senior managing consultant with the IBM Expert Labs Data Science and AI team. He manages solution delivery for many customers, frequently advises on new solutions and technologies, provides mentorship, and is a frequent speaker and author. With nearly 30 years developing cutting-edge analytics solutions in both the public and private sectors, he has delivered with many different tools and is glad to be working at IBM with the best of them. Steve has been delivering solutions with Cloud Pak for Data since its inception and holds multiple Cloud Pak for Data certifications.

Vasfi Gucer is a project leader with the IBM Systems WW Client Experience Center. He has more than 20 years of experience in the areas of systems management, networking hardware, and software. He writes extensively and teaches IBM classes worldwide about IBM products. His focus has been primarily on storage and cloud computing for the last 8 years. Vasfi also is an IBM Certified Senior IT Specialist, Project Management Professional (PMP), IT Infrastructure Library (ITIL) V2 Manager, and ITIL V3 Expert.

Audrey Holloman works on the Cloud Pak Platform team with IBM Expert Labs as a Certified Cloud Pak for Data Solutions Architect specializing in Data science implementations and solution architecture of Cloud Pak for Data. She follows a Data & AI Consultant methodology and works with customers to put use cases into production on Cloud Pak for Data.

Frank Ketelaars is an Information Architecture and Integration technical lead working in the IBM Europe, Middle-East, and Africa Data and AI technical sales team with over 30 years of IT experience. He has been working on many Cloud Pak for Data engagements with customers, IBM business partners, and system integrators since the inception of the product. He has recently focused on operationalizing the IBM Cloud Paks. In addition to being certified on Cloud Pak for Data, Frank holds multiple Red Hat (OpenShift) certifications.

Darren King is a trusted advisory partner to the C-Suite and successfully builds and leads high-performance, globally diverse teams. He emphasizes Business Value Engineering best practices and strategy to lower Total Cost of Ownership, while maximizing Total Economic Impact, ROI, and IRR. Darren holds certifications in Enterprise Architect (OEA), Enterprise Cloud Architect (ECA), Project Management Professional (PMP), Lean Six Sigma Green Belt (LSSGB), ITIL Foundations (ITIL), and Microsoft Networking Fundamentals/Essentials.

Karen Medhat is a Customer Success Manager Architect in the UK and the youngest IBM Certified Thought Leader Level 3 Technical Specialist. She is the Chair of the IBM Technical Consultancy Group and an IBM Academy of technology member. She holds an MSc degree with honors in Engineering in AI and Wireless Sensor Networks from the Faculty of Engineering, Cairo University, and a BSc degree with honors in Engineering from the same faculty. She co-creates curriculum and exams for different IBM professional certificates. She also created and co-created courses for IBM Skills Academy in various areas of IBM technologies. She serves on the review board of international conferences and journals in AI and wireless communication. She also is an IBM Inventor and experienced in creating applications architecture and leading teams of different scales to deliver customers' projects successfully. She frequently mentors IT professionals to help them define their career goals, learn new technical skills, or acquire professional certifications. She has authored publications on Cloud, IoT, AI, wireless networks, microservices architecture, and Blockchain.

Mark Moloney is a Theoretical Physics graduate (BA Hons) from Trinity College, University of Dublin. Throughout his degree, he worked with C++, Python, Mathematica and Data Analysis. Mark joined IBM in 2019 as a Data Scientist as part of the Cloud Pak Acceleration Team. From there, Mark became an expert in Cloud Pak for Data and Red Hat OpenShift. Mark now works as Customer Success Manager - Architect, helping IBM®'s large banking clients adopt and operationalize IBM's containerized software running on Red Hat OpenShift.

Payal Patel is a Solutions Architect in IBM Expert Labs, with a focus on IBM Cloud® Pak for Data and Business Analytics solutions. She has worked in various technical roles across the financial services, insurance, and technology industries. She holds a Bachelor of Science in Information Science from UNC Chapel Hill, and a Masters in Analytics from North Carolina State University.

Neil Patterson is an executive architect in the World Wide SWAT organization for Cloud Paks. Over 30 years experience within IT. Currently providing thought leadership to clients around the globe that are implementing the IBM Cloud Paks.

Deepak Rangarao is an IBM Distinguished Engineer and CTO responsible for Technical Sales-Cloud Paks. Currently, he leads the technical sales team to help organizations modernize their technology landscape with IBM Cloud Paks. He has broad cross-industry experience in the data warehousing and analytics space, building analytic applications at large organizations and technical pre-sales with start-ups and large enterprise software vendors. Deepak has co-authored several books on topics, such as OLAP analytics, change data capture, data warehousing, and object storage and is a regular speaker at technical conferences. He is a certified technical specialist in Red Hat OpenShift, Apache Spark, Microsoft SQL Server, and web development technologies.

Mark Simmonds is a Program Director in IBM Data and AI. He writes extensively on AI, data science, and data fabric, and holds multiple author recognition awards. He previously worked as an IT architect leading complex infrastructure design and corporate technical architecture projects. He is a member of the British Computer Society, holds a Bachelor's Degree in Computer Science, is a published author, and a prolific public speaker.

Malcolm Singh is a Product Manager for IBM Cloud Pak® for Data, where he focuses on the Technical Strategy and Connectivity. Previously, he was a Solution Architect for IBM Expert Labs in the Data and AI Platforms Team. As a Solution Architect in the Expert Labs, he worked with many top IBM clients worldwide, including Fortune 500 and Global 500 companies, and provided guidance and technical assistance for their Data and AI Platform enterprise environments. Now as a Product Manager, he works on establishing the roadmap for future features and enhancements for Cloud Pak for Data. This effort includes collecting the requirements and defining the scope of new features and enhancements and then, working with the engineering teams to devise the technical specifications. Malcolm is based at the IBM Canada Lab in Toronto, working in the Data and AI division within IBM Software. He holds a Bachelor of Science degree in Computer Science from McMaster University.

Tamara Tatian is a technical sales lead in IBM Data and AI's Europe, Middle East, and Africa technical sales Geo team. Tamara specializes in IBM Cloud Pak for Data and Data Fabric solutions, supporting key customer engagements across EMEA cross-industry, providing product SME expertise, and acting as a trusted advisor to IBMers, IBM Business Partners and customers. She holds an MEng/MTech degree in IT and Computer Science, an MSc degree in Project Management, and is a certified IBM Cloud Pak for Data Solution Architect, an IBM Recognized Speaker and Presenter, and an IBM Recognized Teacher and Educator.

Henry L.Quach is the manager of the Learning Architecture and Design team and the Technical Solution Architect. He designs and builds product environments that customers use for training. Before this, Henry was the Lead Curriculum Architect designing and building technical training courses across Cloud Pak for Data. Henry has over 15 years of experience working in the learning and training team with a strong technical background across our Data and AI products and solutions. Henry is a certified Cloud Pak for Data Solution Architect and Administrator.

Thanks to the following people for their contributions to this project:

- Steve Astorino
- Adrian Houselander
- Campbell Robertson
- Jeroen van der Schot
- Ramy Amer
- David Aspegren
- Daniel Zyska
- Erika Agostinelli
- Maxime Allard
- Joe Kozhaya
- Graziella Caputo
- Kazuaki Ishizaki
- Sandhya Nayak
- Vijayan T
- Eric Martens
- Jacques Roy
- Mark Hickok
- Frank Lee
- Anthony O'Dowd
- Jerome Tarte
- Mike Chang
- Sandro Corsi
- Patrik Hysky
- Erica Wazewski
- Angie Giacomazza

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks® residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our books to be as helpful as possible. Send us your comments about this book or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Cloud Pak for Data concepts and architecture

This chapter provides an overview of the Cloud Pak for Data concepts, architecture, and business value. It also includes an overview of the data fabric concepts, architecture, and use cases that are built on data fabric.

This chapter includes the following topics:

1.1 Cloud Pak for Data concepts

While many organizations are struggling with the challenges of data sprawl and complexity, some organizations are finding success as they embrace a modern data strategy. Data-savvy organizations are more likely to use data in a manner that informs decision making to strategically address unmet needs with new data-driven business models.

If an organization can provide enterprise-wide access to previously siloed data, configure governance policies, and address data-quality concerns, it is ready to make large strategic AI investments that can ultimately lead to outperforming revenue targets, their competition, and thereby increase profitability and market share.

The IBM Institute of Business Value (IBV) conducts regular surveys of organizations to identify market outperformers and looks for patterns that set them apart. The 20th edition of the C-Suite study was published in 2020 and draws input from over 13,000 respondents across multiple C-suite roles, industries, and countries.

In this most recent edition of the study, companies are categorized based on their ability to create value from data and the degree to which they have integrated their data and business strategy. Identified as *torchbearers*, 9% of companies that were surveyed showed the most leadership in this area. The following striking numbers in this study are about these *torchbearer* companies:

- They are 88% more likely to make data-driven decisions to advance their corporate strategies.
- They are 112% more likely to find gaps and fill them with data-driven business models.
- They are 300% more likely to enable the free sharing of data across silos and different business functions.
- They are 149% more likely to make large strategic investments in AI technologies.
- Most importantly, they are 178% more likely to outperform others in their industry in the areas of revenue and profitability.[1]

The bottom line: You must outperform your competitors or risk being outperformed by them.

1.1.1 Application and data sprawl

Many organizations have hundreds of disparate applications, data marts, databases, data warehouses, data lakes, and data lake houses sprawled across and beyond the enterprise. Somehow, all of these components must be seamlessly accessible to collect, organize, analyze, inform, and infuse AI into existing and new business processes and applications.

Cloud, *containerization*, and *Kubernetes* are words that are synonymous with modern day development practices and information architectures. Cloud technologies can help enable and provision assets as a set of location-independent and platform-independent services effectively and efficiently by delivering infrastructure, platform, data, software, security, and more “as a Service”. Each can be composed as set of containerized micro services and managed through a platform, such as the Red Hat OpenShift Container Platform.

Managing all these silos of data and applications, many of which were never designed to be integrated, represents a major challenge for organizations when trying to meet the demands of the business to deliver the right insights to the right people at the right time.

[1] Source: IBM Institute of Business Value Study of 13,000 c-suite leaders: https://www.ibm.com/thought-leadership/institute-business-value/c-suite-study

Some vendors or companies publish APIs to a range of data, ML, and AI services, but that alone still infers a level of technical ability that might be out of reach for many. APIs are just one small aspect of the overall data science experience. Although some people might like to build a vehicle from a kit or individual components, most of the public prefers to buy a ready-to-drive vehicle that meets their long-term needs.

1.1.2 AI journey

What IBM learned from countless AI projects is that every step of the journey is critical. AI is not magic; it requires a thoughtful and well-architected approach. For example, most AI failures are due to problems in data preparation and data organization, not the AI models. Success with AI models depends on achieving success first with how you collect and organize data.

The IBM AI Ladder, shown in Figure 1-1, represents a prescriptive approach to help customers overcome data challenges and accelerate their journey to AI, no matter where they are on their journey. It enables them to simplify and automate how an organization turns data into insights by unifying the collection, organization, and analysis of data, regardless of where it lives. By climbing the ladder to AI, enterprises can build a governed, efficient, agile, and everlasting approach to AI.

The AI Ladder™ features the following four steps (often referred to as *rungs*), as shown in Figure 1-1:

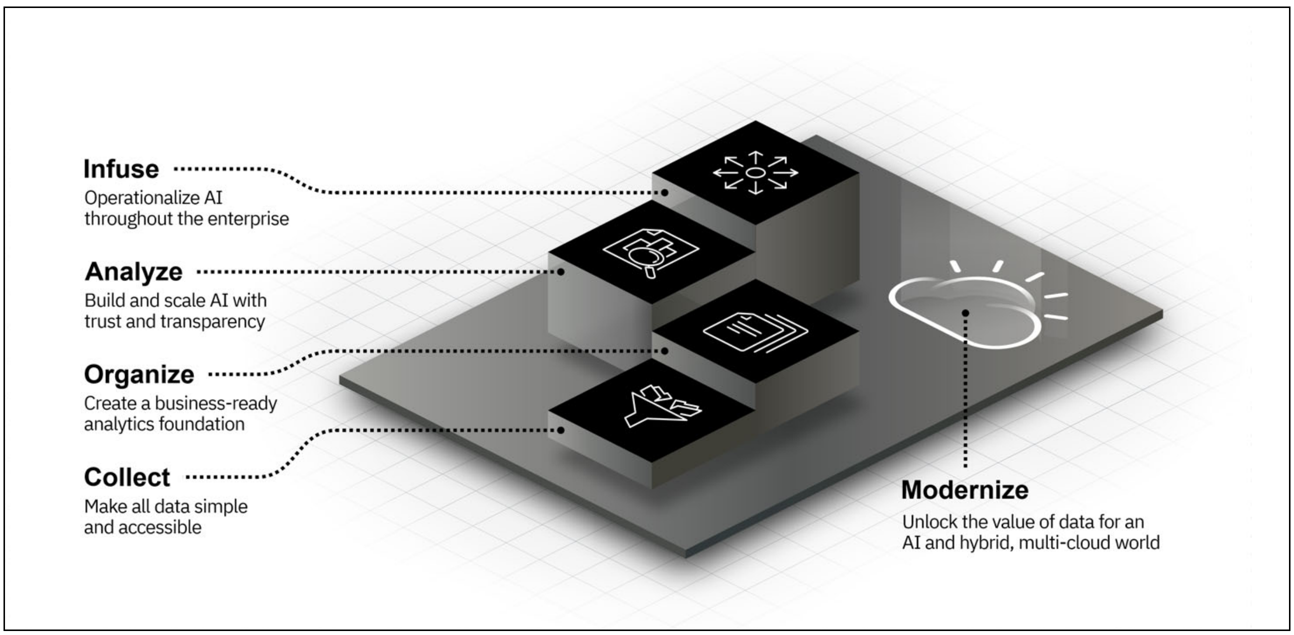

Figure 1-1 Four steps of the AI Ladder

- Collect: Make data simple and accessible

  Collect data of every type, regardless of where it is stored, which enables flexibility in the face of ever-changing data sources. The term “collect” does not refer to putting data all in one place. In fact, it is the opposite. It means to virtualize the data to allow access to wherever it is stored as though it were consolidated.

- Organize: Create a business-ready analytics foundation

  Organize collected data into a trusted, business-ready foundation with built-in governance, protection, and compliance.

- Analyze: Build and scale AI with trust and transparency

  Analyze data in automated ways and benefit from AI models that empower teams to gain new insights and make better, smarter decisions.

- Infuse: Operationalize AI throughout the business

  Infuse AI throughout the business (across multiple departments and within various processes), drawing on predictions, automation, and optimization.

These steps can be further broken down into a set of key capabilities, as shown in Figure 1-2.

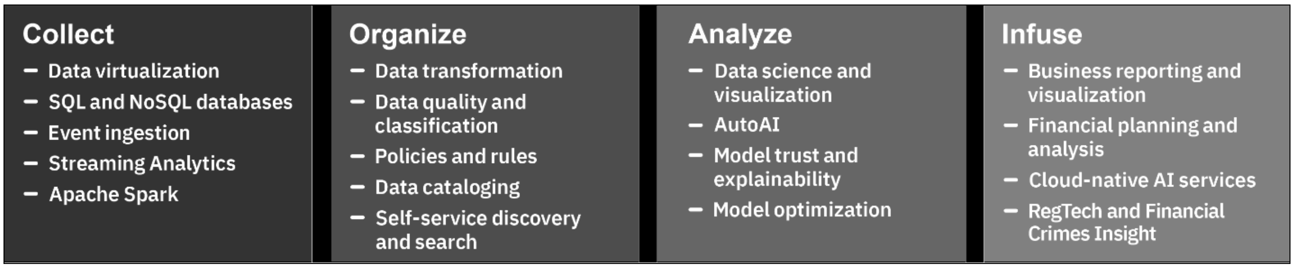

Figure 1-2 AI Ladder capabilities

Supporting the AI Ladder is the concept of modernization, which is how customers can simplify and automate how they turn data into insights by unifying the collection, organization, and analysis of data (regardless of where it is stored) within a secure hybrid cloud platform.

The following priorities are built into the IBM technologies that support this AI ladder:

- Simplicity: Different kinds of users can use tools that support their skill levels and goals, from “no code” to “low code” to programmatic.
- Integration: As users go from one rung of the ladder to the next, the transitions are seamless.
- Automation: The most common and important tasks have intelligence included so that users focus on innovation rather than repetitive tasks.

1.1.3 Delivering a hybrid cloud data and AI platform

IBM Cloud Pak for Data embodies a unified IBM Enterprise Insight® Platform (EIP) that runs on multiple vendors’ clouds and infrastructures. *EIP* is a term that is used by industry analysts and consultants as a category for describing integrated sets of data management, analytics, and development tools.

The first core tenet of Cloud Pak for Data is that you can run it anywhere. You can colocate it where you are making your infrastructure investments, which means that you can deploy Cloud Pak for Data on the major cloud vendor’s platforms and the IBM Cloud.

You also can deploy it on-premises for the case in which you are developing a hybrid cloud approach. Finally, on IBM Cloud, you can subscribe to Cloud Pak for Data-as-a-Service if you need a fully managed option where you pay only for what you use.

Cloud Pak for Data helps organizations to have deployment flexibility to run anywhere (see Figure 1-3 on page 11).

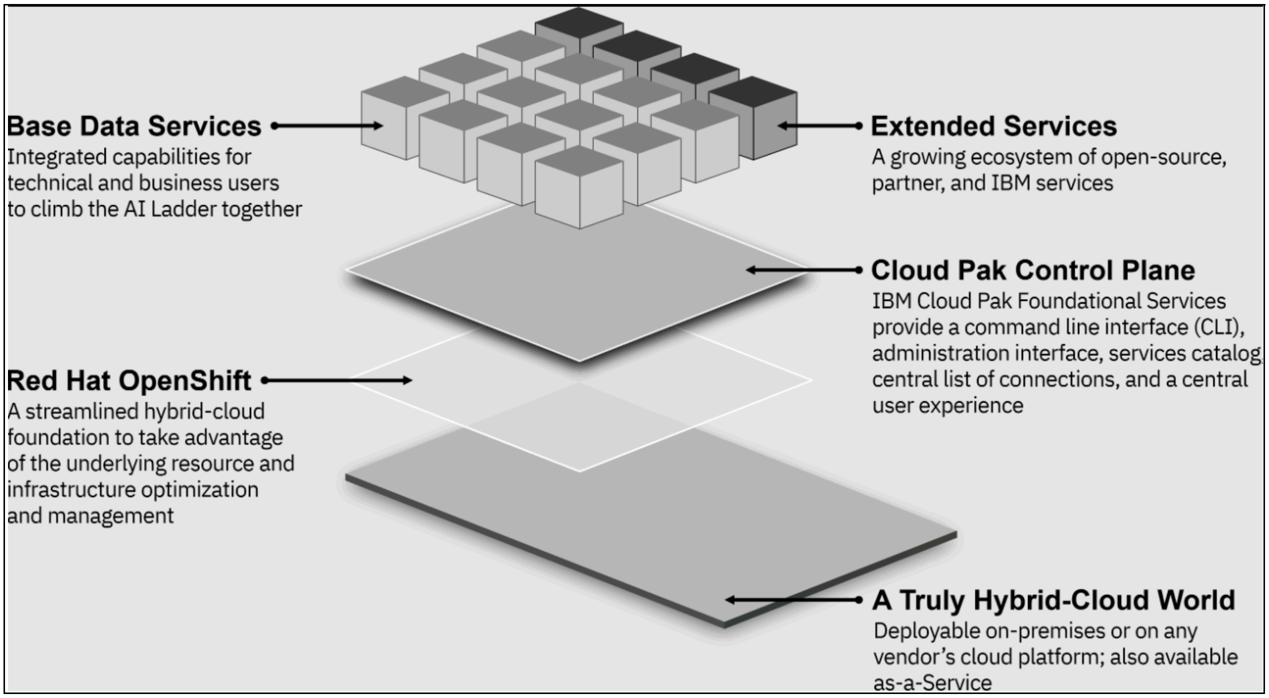

Figure 1-3 Cloud Pak for Data

1.1.4 Unified and modular

Cloud Pak for Data is built on the foundation of Red Hat OpenShift. This foundation provides the flexibility for customers to scale across any infrastructure by using the leading open source steward: Red Hat.

Red Hat OpenShift is a Kubernetes-based platform with which IBM deploys software through a container-based model that delivers greater agility, control, and portability. Based on business value alone, Cloud Pak for Data is an ideal entry-point into the containerization space and a foundational building-block for more Cloud Paks and Cloud Pak Services.

IBM's Cloud Pak offerings, including Cloud Pak for Data, all share a common control plane that simplifies and standardizes administration and integration of diverse services.

Cloud Pak for Data provides a set of pre-integrated data services with which an organization can collect information from any repository, such as databases, data lakes, and data warehouses. The design point here is for customers to leave the data in all the places where it exists, but to its users it seems as though the enterprise data is in one spot by using the platform's data virtualization technologies.

After an enterprise's data is connected, industry-leading data organization services can be deployed that allow for the development of an enterprise data catalog. This capability enables a "shop for data" type of experience that helps enforce governance across all data sources, which enables data consumers to have a single place to go for all their data needs.

With your enterprise data connected and cataloged, Cloud Pak for Data presents various data analysis tools that are available for immediate use. For example, a wealth of data science capabilities is available that cater to all skill levels (no-code, low-code, and all code). Users can quickly grab data from the catalog and instantly start working toward generating insights in a common workflow that is built around the "project" concept.

For more capabilities, a large set of extended services is available for Cloud Pak for Data that presents more-specialized data management and analytics capabilities. These services range from powerful IBM solutions, such as business analytics, data management, multicloud integration, trustworthy AI, data governance and privacy, and customer care, to solutions from IBM Business Partners that offer business ontology creation, open source databases, and more (see Figure 1-4).

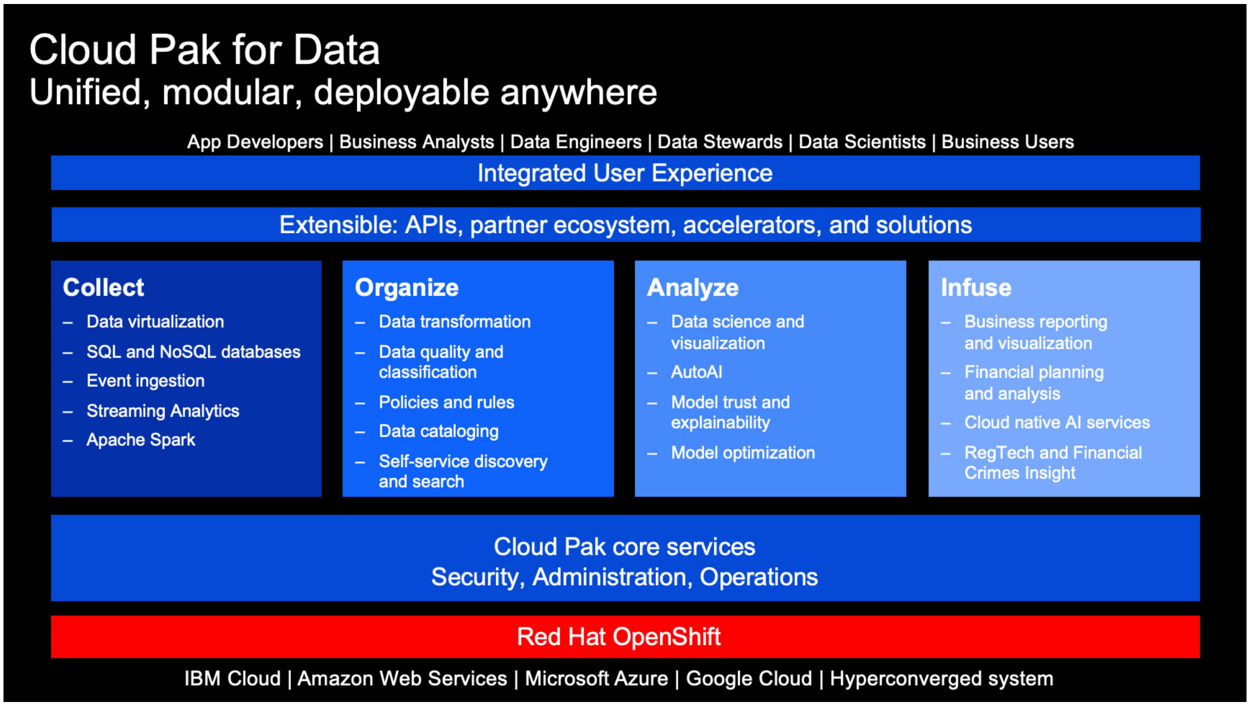

Figure 1-4 Cloud Pak for Data – Unified, Modular, Deploy Anywhere

IBM Cloud Pak for Data delivers a unified, collaborative user experience across the AI ladder. A key design point was to enable users (regardless of skill level) to extract insights from data, whether an experienced data scientist or a line-of-business user, or developers to build, test, deploy, and manage machine learning models, or security officers and administrators to manage all of the platform’s services.

As shown in Figure 1-4, Cloud Pak for Data became the platform to deliver all IBM Data and AI services in a consistent and integrated way. Essentially, it is a development platform to enable cloud-native, modernization, and digital transformation.

1.1.5 Your data and AI: How and where you want it

IBM’s open information architecture for AI is built upon Cloud Pak for Data on Red Hat OpenShift, which is built for a hybrid cloud world. But what does this mean? In one word: flexibility. Consider the following points:

- If your organization is in a place where you must manage as little IT as possible, you can use Cloud Pak for Data entirely through an as-a-service model by subscribing to the integrated family of data services on the IBM Cloud.
- If your organization needs the flexibility and control of running the data infrastructure in your own data center or on Infrastructure as a Service (IaaS) from your preferred cloud vendor, you can deploy Red Hat OpenShift and then, Cloud Pak for Data on your local or cloud estate.

- If high performance and total control are needed, you can choose the Cloud Pak for Data System, which is a hyper-converged infrastructure (an optimized appliance) that combines compute, storage, and network services that are optimized for Red Hat OpenShift and data and AI workloads.

Regardless of the form factor and the degree of management control that is needed, Cloud Pak for Data provides cloud-native data management services that modernize how businesses collect, organize, and analyze data and then, infuse AI throughout their organizations.

If you have an Red Hat OpenShift deployment on IBM Cloud, Amazon Web Services (AWS), Microsoft Azure, or Google Cloud, you can deploy Cloud Pak for Data on your cluster.

If you prefer to keep your deployment behind a firewall, you can run Cloud Pak for Data on your private, on-premises cluster.

The different deployment categories are summarized in Table 1-1.

Table 1-1 Deployment categories

	Self-managed Red Hat OpenShift	Managed Red Hat OpenShift	IBM Satellite™
Description	Base Red Hat OpenShift that is managed by the customer.	Vendor-hosted Red Hat OpenShift. Versioning, maintenance, and daily operations are handled by the vendor.	Hybrid cloud solution for developing and monitoring IBM Cloud services on-premises, edge, and public cloud environments.
Value	Full control over the Red Hat OpenShift cluster.	Eliminates the complexity of and cost of managing and maintaining container environments.	Provides a consistent and easily managed experience for IBM Cloud capabilities on a ROKS cluster. Allows for simple configuration of access and security controls.
Customers	For customers with resources and skills to manage an Red Hat OpenShift cluster.	For customers who benefit from reduced operational complexity or do not have the technical skills to manage a Red Hat OpenShift footprint.	For customers who benefit from IBM Satellite monitoring and consistent UX across clouds
Supported forms	► AWS ► Azure ► IBM Cloud	► AWS ROSA ► Azure ARO ► IBM ROKS	AWS

1.1.6 Multitenancy

Cloud Pak for Data supports different installation and deployment mechanisms for achieving multitenancy, and environment in which multiple independent instances of one or multiple applications operate in a shared environment. The instances (tenants) are logically isolated, but physically integrated.

At the platform level, Cloud Pak for Data (the Cloud Pak for Data control plane) can be installed many times on the same cluster by installing each instance of Cloud Pak for Data in a separate project (Kubernetes namespace).

The following components are installed after they are on the cluster and shared by any instances of Cloud Pak for Data on the cluster:

- IBM Cloud Pak foundational services
- Scheduling service

The Cloud Pak for Data platform also supports many mechanisms for achieving service multitenancy. However, not all services support the same mechanisms. For example, the platform offers the following mechanisms:

- Installing a service one time in *each* project where the control plane is installed. (This method is the most common for achieving multitenancy.)
- Installing a service one time in the same project control plane and provisioning multiple instances of the service in that project.
- Installing a service one time in a project that is tethered to the project where the control plane is installed.
- Installing a service one time in the same project as the control plane and deploying instances of the service to projects that are tethered to the project where the control plane is installed.

In summary, Cloud Pak for Data is designed to provide a unified, integrated user experience to collect, organize, and analyze data and infuse AI throughout the enterprise by using a data fabric approach and architecture.

Many of the complexities of managing and orchestrating data and other artifacts can be abstracted through the data fabric architectural approach. IBM Think® of the data fabric as the “magic” that can help make more of an organization’s data, applications, and services ready for AI by automating and augmenting many of the steps that otherwise must be undertaken by large groups of architects, administrators, and data scientists.

1.2 Data fabric

Enterprises face all sorts of complexities in implementing their uses cases by using current approaches, such as providing a 360-degree view of the data, and Master Data Management use cases, data governance, data privacy, regulatory compliance, operational analytics, business intelligence, trustworthy AI, and data science, for example.

As many infrastructures grow, enterprises can often face higher compliance, security, and governance risks. These risks can result in complexity and a high level of effort to enforce policies and perform stewardship.

Complex infrastructures also can lead to higher costs of integrating data and stitching data pipelines across multiple platforms and tools. In turn, these platforms and tools can bring more reliance on IT, which makes collaboration more challenging and possibly slow time to value. However, business-led self-service analytics, insights, and democratization of data can help deliver greater business agility. The many attempts to pull disparate data silos together all fell short of business and user expectations.

What is needed is a new design or approach that provides an abstraction layer to share and use data (with data and AI governance) across a hybrid cloud landscape ***w***ithout a massive pendulum swing to having everything de-centralized. It is a balance between what must be logically or physically decentralized and what must be centralized. For example, an enterprise can have multiple catalogs, but only one source of truth can exist for the global catalog.

A data fabric is a data management architecture that helps optimize access to distributed data and intelligently curate and orchestrate it for self-service delivery to data consumers. Some of a data fabric's key capabilities and characteristics include the following examples:

- Designed to help elevate the value of enterprise data by providing users with access to the right data just in time, regardless of where or how it is stored.
- Architecture independent of data environments, data processes, data use, and geography, while integrating core data management capabilities.
- Automates data discovery, governance, and use and delivers business-ready data for analytics and AI.
- Helps business users and data scientists access trusted data faster for their applications, analytics, AI, and machine learning models, and business process automation, which helps to improve decision making and drive digital transformation.
- Helps technical teams use simplify data management and governance in complex hybrid and multicloud data landscapes while significantly reducing costs and risk.

The data fabric approach enables organizations to better manage, govern, and use data to balance agility, speed, SLAs, and trust. Trust covers deep enforcement of governance, security, and compliance and the total cost of ownership and performance (TCO/P). Trust covers the deep enforcement of governance, security, and compliance. The total cost of ownership and performance (TCO/P) factor also must be considered. TCO/P covers costs that are associated with integration, egress, bandwidth, and processing.

versus performance, and so on.

A data fabric approach is beneficial to the business and IT users:

- For business teams and chief data officers (CDO):
  - Gain faster and more accurate insights because of easy access to high-quality data.
  - Ability to focus time on analyzing rather than finding and preparing data.
  - Frustration-free full self-service data shopping experience.
  - Avoidance of biased analysis because of data restrictions.
  - Increased compliance and security despite full analytics utilization.
- For technical teams and chief technology officers (CTO):
  - Decreased effort to maintain data quality standards because of fewer data versions.
  - Reduced infrastructure and storage cost (consolidated data management tools and reduction in data copies).
  - Faster and simplified data delivery processes because of fewer targets and advance optimization of data flows.
  - Reduction in efforts for data access management as it becomes automated by global data policy enforcement.

1.2.1 Data fabric use cases

IBM focuses on five key use cases for a data fabric, as shown in Figure 1-5.

Figure 1-5 Data fabric use cases

Next, we review these use cases in terms of their capabilities and core differentiators.

Data governance and privacy: Automate data governance and privacy to enable data trust, protection, security, and compliance

This use case features the following capabilities:

- Know your data: Access a 360° view of their business-ready data and relationships.
- Trust your data: Onboard and automatically enrich metadata to make trusted data available to data consumers.
- Protect your data: Enable global automatic policy enforcement for increased data protection and privacy.
- Govern your data: Set up a robust framework for data governance.

This use case features the following benefits:

- Automatically apply industry-specific regulatory policies and rules to your data assets.
- Quickly establish an environment for highly automated and consistent governance.
- Automatically secure data across the enterprise.

This use case features the following core differentiators:

- AI-augmented data catalog with which business users can easily understand, collaborate, enrich, and access the right data.
- Metadata and governance layer for all data, analytics, and AI initiatives that increases visibility and collaboration on any cloud.
- Dynamically and consistently mask data at a user-defined, granular level.
- Facilitate anonymized training data and the creation of test sets while maintaining integrity of the data.

> **Note:** For more information, see Chapter 3, “Data governance and privacy” on page 115. You also can download the ebook *Data governance and privacy for data leaders* (log-in required).

Multicloud integration: Integrate data across hybrid cloud to accelerate time to value by democratizing data for AI, business intelligence, and applications

This use case features the following capabilities:

- Connect to, refine, and deliver data across a hybrid and multicloud landscape.
- Democratize data access: Deliver data where you need it, whether it be on-premises or on any cloud, in near real-time, batch, or powered by a universal SQL engine.
- Continuous availability for mission-critical data: Real-time synchronization of operational and analytics data stores across hybrid cloud environments with high throughput and low-latency.
- Empower new data consumers: Allow data consumers to access trusted, governed data and comprehensive orchestration to manage data pipelines. Enable flexibility with open APIs and SDKs.

This use case features the following benefits:

- Create a unified view of all enterprise data, which enables consistency between operational applications.
- Consolidate and simplify IT infrastructures, run anywhere (on-premises or any cloud), and automate data operations to deliver trusted data to business users.
- Prevent delays and disruptions of mission-critical data through data resilience and easy data access.

This use case features the following core differentiators:

- Augmented data integration flows making the best use of ETL, data virtualization, and real-time capture to optimize access to many diverse data sources.
- Capability to build quality analysis and remediation natively into data pipelines without costly downstream processing.
- Support for full DataOps lifecycle through governance, quality, master data, integration, and collaboration.

Note: For more information about this use case, see Chapter 4, “Multicloud data integration” on page 163. You can also download the ebook *Multicloud data integration for data leaders* (log-in required).

Data observability: Use data observability for the reliability of your enterprise data

This use case features the following capabilities:

- Detect earlier: Pinpoint unknown data incidents and reduce mean time to detection (MTTD) from days to minutes.
- Resolve faster: Improve mean time to resolution (MTTR) with incident alerts and routing from weeks to hours.
- Deliver trustworthy data: Enhance reliability and data delivery SLAs, provide visibility into pipeline quality issues that otherwise go undetected.

This use case features the following benefits:

- Automatically and proactively manage dynamic and static data pipelines, which provide quicker time to value.
- Simultaneously address data quality and reliability continuously.
- Continuously manage AI reliability across data and models.

This use case features the following core differentiators:

- Composable and integrated capability that eliminates need for point solutions.
- Modular architecture allows for integration with other data use cases.
- Best in class end-to-end observability solution for static and in motion data.

> **Note:** For more information about this use case, see Chapter 5, "Trustworthy artificial intelligence concepts" on page 267.

Customer 360: Provide comprehensive view of customer by integrating data across domains

Customer 360 provides a modern Master Data Management (MDM) solution with ML-augmented self-serve access to governed, quality, and compliant data.

Customer 360 enables customer data that is:

- Agile and user-accessible: Quickly match new sources and ensure data quality to find and act on customer data by using knowledge graph with a low-code approach.
- Protected and compliant: Achieve agility in customer data access while quickly meeting and delivering compliance requirements.
- Ready for the enterprise: Provides hyper-personalization and next-best-offers, while infusing compliance, privacy, and fraud identification capabilities for better overall customer care.

This use case features the following benefits:

- Spend more time on applying AI and analytics to business challenges versus wasting time on hunting quality data that is understood.
- Business outcomes at pace with competitive needs.
- Break down data silos with an integrated view of data.

This use case features the following core differentiators:

- Resolve entities across different sources with self-service capabilities and much faster time to value to enable analytics and AI.
- Simplified data modeling for data-first approach.
- ML-powered probabilistic matching for self-service.

> **Note:** For more information about this use case, download the ebook *Customer 360 for data leaders* (log-in required).

Machine Learning Operations (MLOps) and trustworthy AI: Operationalize AI with governed data integrated throughout AI lifecycle for trusted outcomes

This use case features the following capabilities:

- Trust in data: A complete view of quality data that is private, self-served, and ready for analysis by multiple personas.
- Trust in model: MLOps that is infused with fairness, explainability, and robustness.
- Trust in process: Automation to drive consistency, efficiency, and transparency for AI at scale.

This use case features the following benefits:

- Unlock trustworthy AI starting with governed data access for data scientists.
- Automated MLOps that is infused with trust throughout the entire AI lifecycle.
- AI governance by introducing transparency and monitoring for each stage of the AI lifecycle.

This use case features the following core differentiators:

- Architecture that is built with compliance and scalability into first design principles.
- Model monitoring with explainability, draft, bias, fairness, and quality.
- Integrate data governance with data science to provide end-to-end trust.

> **Note:** For more information about this use case, see Chapter 5, “Trustworthy artificial intelligence concepts” on page 267 for more information. You can also download the ebook *MLOps and trustworthy AI for data leaders* (log-in available).

1.3 Architecture

The simpler the architecture, the more effective the solution is, and it is this premise upon which IBM Cloud Pak for Data architecture is built.

The Cloud Pak for Data architecture is based on microservices architecture and consists of different pre-configured microservices that run on a multi-node Red Hat OpenShift cluster. These microservices enable the connection to data sources to make the necessary governance, profiling, transformation, and analysis on data from a single view, which is the dashboard of the Cloud Pak for Data.

1.3.1 Reference architecture

Figure 1-6 shows the reference architecture for Cloud Pak for Data version 4.5.

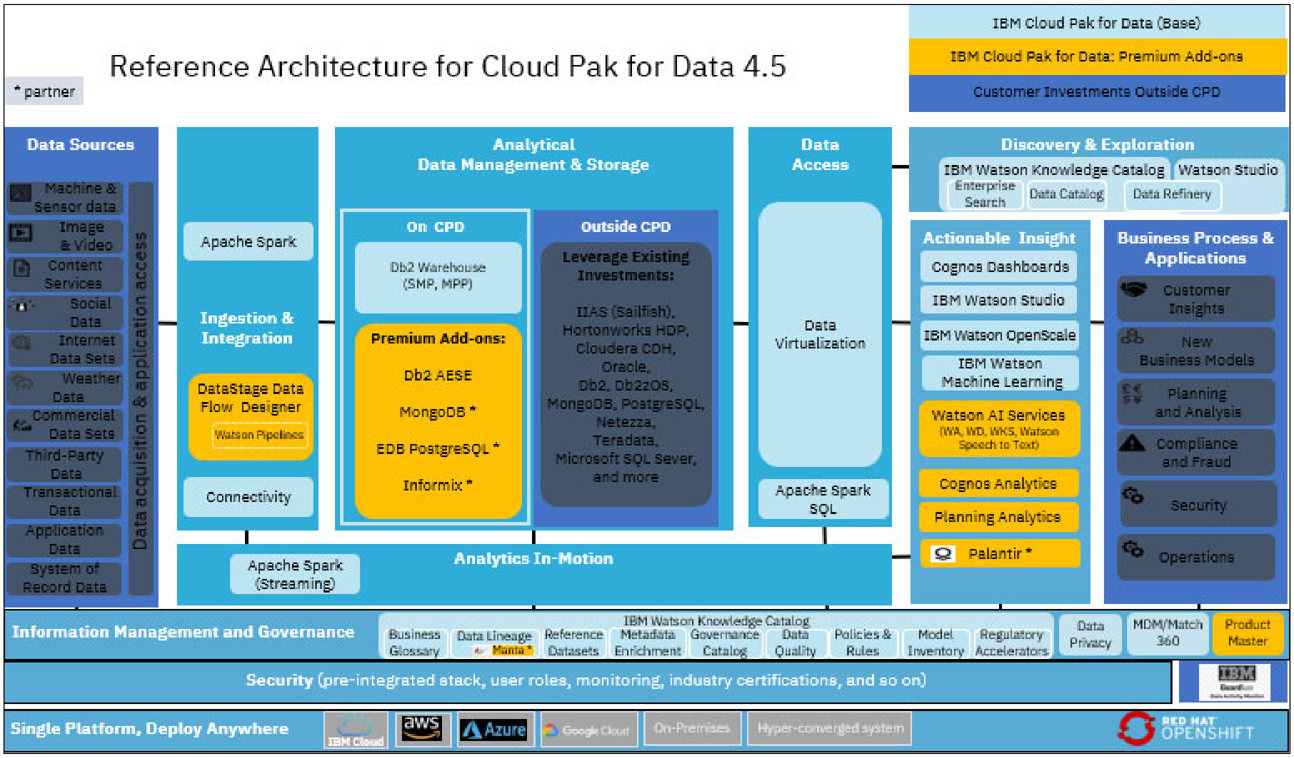

Figure 1-6 Reference architecture for the Cloud Pak for Data

The following base components or tools are included:

- Apache Spark
- Connectivity
- Data virtualization
- Db2
- IBM Watson Knowledge Catalog
- IBM Watson Studio
- IBM Watson OpenScale
- IBM Watson Machine Learning
- Master Data Management
- IBM Cognos®

Other add-ons also are available for Cloud Pak for Data, such as IBM Watson AI services, DataStage®, Cognos Analytics, and Planning Analytics.

The Cloud Pak for Data can be integrated with other external data sources, business process and applications for security, operations, analytics, and business models.

1.3.2 Cloud-native design

The design of the Cloud Pak for Data supports managing multiple applications on different environments, which support organizations in building cloud-native applications efficiently and accelerating their modernization journeys.

Cloud Pak for Data architecture supports the journey to AI because it is considered an extensible cloud-native architecture that is based on data fabric. It brings together cloud, data and AI capabilities for collecting, organizing, and analyzing data as containerized microservices to deliver the AI ladder in a multi-cloud environment.

IBM Ladder to AI

The AI ladder consists of the following steps to process data to make it meaningful and ready to be processed by the AI business models:

1. Collect your data: Where the connections for all the sources of data are made without migrating the data.
2. Organize your data: A creation for a business-ready foundation occurs, which simplifies the preparation of data, secures it, and insures its compliance.
3. Analyze your data: The data is ready to be analyzed and the focus is to build, deploy, and manage the AI capabilities that can be easily scalable.
4. Infuse AI: The business operates based on AI with trust, transparency, and agility.

Figure 1-7 shows the Cloud Pak for Data components that are used in each one of the steps of the AI ladder.

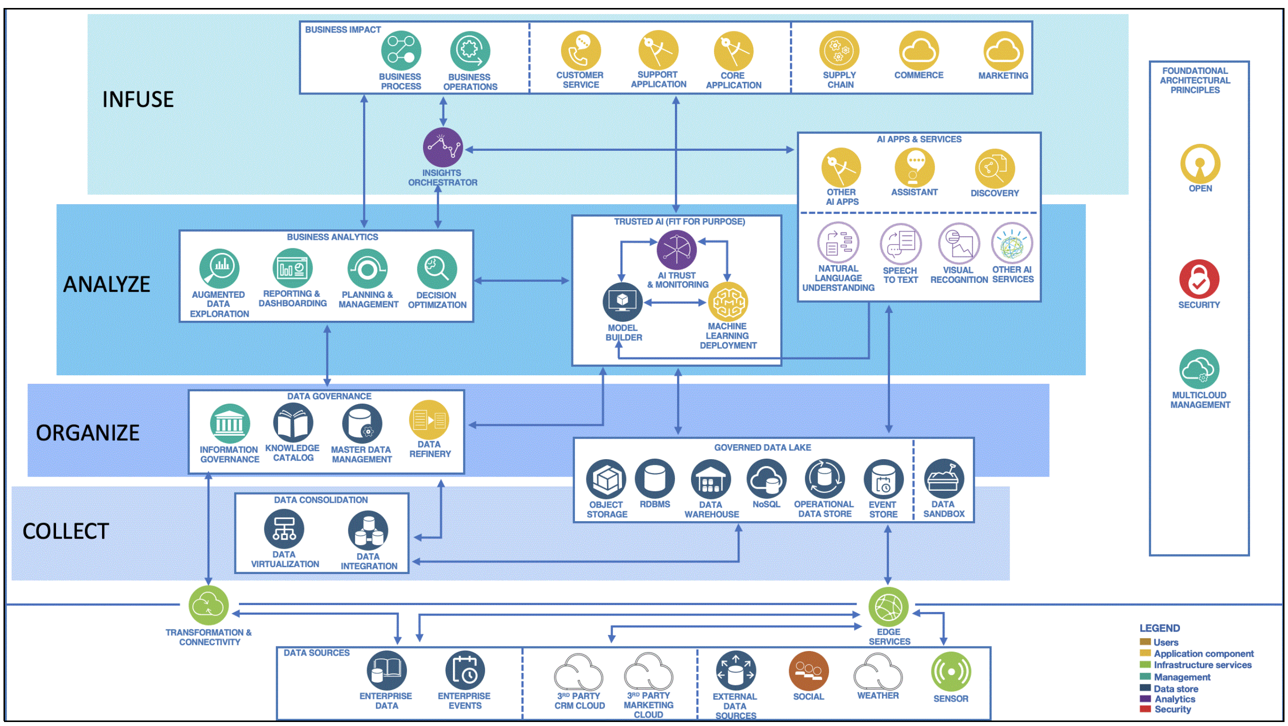

Figure 1-7 Steps of AI ladder and the Cloud Pak for Data components in each step

Build phase

The following flow of activities occurs during the build phase to develop the model and continuously improve it over time while the intelligent application is running in production:

- In the first step, which is part of the infuse phase, the business impact is shown; the business identifies the need that can be addressed by building a predictive model.

- In the second step, which is in the analyze phase, the role of trusted AI starts to play through the model builder where the business analyst collaborates with data steward, data scientist, and cognitive architect collaborate to build the predictive model by using IBM Watson Studio.
- In the third step, which is in the organize phase, the team look at enterprise data lake or data warehouse to identify the optimal informing features toward building an effective predictive model. They decide on the data structures, environments, the design goals (whether the model is to be a data lake or warehouse), and the deployment environment.
- In the fourth step, which is in the organize phase, the data governance is considered where the business analyst, data steward, data scientist, and cognitive architect decide on the signals that can be extracted from the data and which of them to virtualize and what to federate.
- In the fifth step, which is in the collect phase, by using a cloud-native data service, the signals are pulled into the data collections. According to the decisions that are made in the third step, new data sources might be included in the environment.
- In the sixth step, which is part of the organize phase, various portions of the informing data collections are refined to engineer features for the predictive model by using Data Refinery. Often, data must be curated, aggregate totals might be created, or some other modification or enhancement is needed.
- In the seventh step, which is part of the analyze phase, Watson Natural Language Understanding (NLU) APIs are used on the unstructured data to extract entities and taxonomies to enrich the unstructured information with relevant metadata.
- In the eighth step, which is part of the organize phase, the methods that are used to enrich the unstructured data are used as an ongoing data enrichment given the iterative nature of building models to reach quickly to changes.
- In the ninth step, which is part of the analyze phase, with multiple batch experiments and tuned hyper-parameters, a performant model is available for use.
- In the 10th step as part of the analyze phase, the model is deployed by using Watson Machine Learning, which provides a scalable model with continuous learning.

Run phase

The run time involves the following analyzing and infusing activities, where the components included in these two phases work together to create business agility:

- In the first step (analyze phase), the AI model is deployed as a service or runs within a near real-time analytics processing application or an embedded device.
- In the second step (analyze phase), an AI orchestrator, such as IBM Watson OpenScale, monitors the runtime performance and fairness across the customer population.
- In the third step (infuse phase), the predictions are interpreted by an AI insights orchestrator, such as one that is deployed on IBM Cloud.
- In the fourth step (analyze phase), the orchestrator consults the Decision Management or Optimization components for the next best action or offer to be provided to the customer.
- In the fifth step, (infuse phase), this offer is communicated within the customer's workflow by infusing it into the active customer care and support interactions, such as a personal assistant implementation, such as IBM Watson Assistant.
- In the sixth step, (analyze phase), during the runtime, the AI orchestrator detects model skew that indicates that the model must be retrained.

- In the seventh step (analyze phase), the detection of a skew triggers deeper analysis of the divergence from the original model by using a machine learning workbench. Similar steps to the original model design are taken to retrain the model with optimal features and new training data.
- In the eighth step (analyze phase), an updated, better-performing model is deployed to the model-serving component as another turn in the iterative model lifecycle governance.
- In the ninth step (infuse phase), customers react favorably to the offers, and the business starts to see lower customer churn and higher satisfaction ratings.

Component description

Table 1-2 lists the components that are shown in Figure 1-7 on page 21 and the corresponding products in Cloud Pak for Data.

Table 1-2 Main components of the Cloud Pak for Data on the AI Ladder

Component	Description	Products
Business process	Business processes lay the foundation for back-office and front-office business functions, from managing invoices and records to quickly opening customer accounts and offering real-time promotional offers to prospects. Business processes allow all the different parts of an organization to efficiently and effectively work together toward their common goal of serving customers better.	Business process management
Business operations	Business operations include a set of processes, applications, and tools that aim at optimizing the harvested value from all the assets that are owned by a business. Those assets include physical assets and assets, such as intellectual property.	Enterprise-specific
Customer service	Customer service supports customer care across the entire transaction lifecycle and all commerce channels, where customer care personnel can see the behaviors of a customer in more than one channel. Customer service includes CRM and loyalty management. In most instances, these core systems store transaction data.	Enterprise-specific
Support application	Core applications run business transactions. Organizations also use supporting applications, such as human resources, accounting, and finance to support the core business of an enterprise.	Enterprise-specific

Component	Description	Products
Core application	Core applications drive the business transactions of an enterprise. Examples of core applications are core banking, claims management, order management, ERP systems, transportation systems, and logistics management. Core applications differ for different industries. IBM Order Management software tracks orders from inception to fulfillment and manages processes and data that are connected to the order as it moves through its lifecycle. It tracks all information and processes, including order entry, inventory management, supply-chain order fulfillment, and after-sales services. It also provides real-time visibility into orders from multiple channels.	Order management software
Supply chain	Supply-chain management is the handling of the entire production flow of a good or service, starting from the raw components all the way to delivering the final product to the consumer. To accomplish this task, a company creates a network of suppliers (the “links” in the chain) that move the product from the suppliers of raw materials to the organizations that deal directly with users.	► IBM Planning Analytics ► IBM Watson Supply Chain
Commerce	Digital, cross-channel, social, and sentiment analytics collect data collect data about customer behavior that occurs across channels to personalize shopping interactions.	► IBM Watson Customer Experience Analytics ► IBM Watson Marketing Insights ► IBM Predictive Customer Intelligence ► IBM SPSS
Marketing	Marketing provides capabilities to get deeper insights that help in making smarter marketing decisions. Marketing drives results with personalized customer experiences across all channels, drives lasting loyalty with exceptional end-to-end customer experiences, and boosts engagement and conversions wherever the customers are with SMS, push notifications, and more.	IBM Watson Marketing
Insights orchestrator	This component orchestrates insight creation by accessing data to be scored and calling a model deployment service and any other service that is required, such as a decision optimization service.	► IBM Watson Studio ► IBM Watson Machine Learning

Component	Description	Products
Other AI applications	These applications become more powerful with prediction and optimization from AI. They typically span various domains, such as marketing, finance, healthcare, automotive, media, utilities, or agriculture. These applications build upon trusted AI models and often take advantage of packaged automated tools to simplify and democratize AI deployments.	Enterprise-specific
Assistant	Assistant is an application for conversational AI agents that integrates across channels and devices to help improve self-service and experience for customers and employees.	IBM Watson Assistant
Discovery	Discovery conducts search and data exploration by using indexing and machine learning technology to show the user matches of various documents and records from internal and external data sources.	IBM Watson Discovery
Natural language understanding	Natural language understanding is a set of AI techniques for natural language processing to read, understand, and extract useful and valuable insights from natural language, such as keywords, entities, and concepts to improve the interaction between humans and AI machines.	► IBM Watson Natural Language Understanding ► IBM Watson Natural Language Classifier ► IBM Watson Knowledge Studio
Speech to text	Speech to text converts voice to text.	IBM Watson Speech to text
Visual recognition	Visual recognition identifies objects such as faces and individual items.	► IBM PowerAI Vision ► IBM Watson Visual Recognition
Other AI services	These services are extra AI services that the application needs.	Enterprise-specific
Augmented data exploration	Augmented data exploration uncovers insights in your data by using plain language, visual exploration, and machine learning. These applications can be used to ask questions and get answers in plain language. They also can be used to discover hidden patterns in your data with visual exploration tools that help to avoid bias.	► IBM Business Analytics ► IBM Cognos Analytics

Component	Description	Products
Reporting and dashboarding	These tools and offerings make it easy to visualize, analyze, and share insights about your business. They help you prepare and share data, uncover what drives performance, visualize that performance, and share those insights with your team by using dashboards and pixel perfect reports.	► IBM Business Analytics ► IBM Cognos Analytics ► IBM Planning Analytics
Planning and management	Planning and management tools automate decisions by capturing and running business rules or complex event processing.	► IBM Operational Decision Manager on Cloud: Cloud-managed comprehensive decision automation platform to capture, analyze, automate, deploy, and govern rules-based business decisions ► IBM Operational Decision Manager: Comprehensive decision automation platform to capture, analyze, automate, deploy, and govern rules-based business decisions ► IBM Blueworks Live: Intuitive, cloud-based business process discovery and modeling tool that generates industry-standard BPMN 2.0 layouts, documentation, and output.
Decision optimization	Decision optimization uses powerful analytics to solve planning and scheduling challenges by reducing the effort, time, and risk that are associated with creating tailored solutions that improve business outcomes.	IBM Decision Optimization Center
AI trust and monitoring	These applications track the performance of production AI and its impact on business goals with actionable metrics, which creates a continuous feedback loop that improves and sustains AI outcomes. They also maintain regulatory compliance by tracing and explaining AI decisions across workflows. They intelligently detect and correct bias to improve outcomes.	IBM Watson OpenScale
Model builder	Data scientists rely on model builders to iterate on the development and training of AI models by using optimal machine learning and deep learning techniques. They use open source frameworks and tools to manage the models and provide users and lines of business with multitenancy and role-based access controls.	► IBM Watson Studio ► IBM Watson Machine Learning ► SPSS® Modeler ► IBM Watson Machine Learning Accelerator ► AutoAI with IBM Watson Studio ► Deep Learning Power AI

Component	Description	Products
Machine learning deployment	Machine learning deployment implements a machine learning model into a production environment to make practical business decisions that are based on data.	► IBM Watson Studio ► IBM Watson Machine Learning
Open	Open refers to software and related data that are freely licensed for anyone to use, copy, study, and change in any way.	Enterprise-specific
Security	Security enables identity and access management, and data and application protection. It provides actionable security intelligence across cloud and enterprise environments.	Enterprise-specific
Multicloud management	Multicloud management is an integrated framework to monitor, govern, manage, and optimize multiple work-loads across multiple cloud providers.	IBM Cloud Pak for Multicloud Management
Information governance	Information governance provides the policies and capabilities that enable the analytics environment to move, man-age, and govern data.	► InfoSphere Information Analyzer (for data classification) ► IBM InfoSphere® Information Server for Data Quality
Knowledge catalog	The knowledge catalog helps to you find, understand, and use needed data. It also helps users to discover, curate, categorize, and share data assets, data sets, analytical models, and their relationships with other members of an organization. The catalog serves as a single source of truth.	IBM Watson Knowledge catalog
Master Data Management	Master Data Management is a method that is used to define and manage the critical data of an organization to provide, with data integration, a single point of reference. The data that is mastered can include reference data, which is the set of permissible values and the analytical data that supports decision-making.	Master Data Management
Data Refinery	Data Refinery is a data-preparation capability in support of self-service analytics. It can be used for the quick transformation of large amounts of raw data into consumable, quality information that is ready for analytics.	► Data Refinery, available by using IBM Watson Studio and Watson Knowledge Catalog (Pro) ► IBM InfoSphere Advanced Data Preparation
Object storage	Object storage typically supports exponential data growth for cloud-native workloads. It supports built-in, high-speed file transfer capabilities, cross-region offerings, integrated services, and security.	IBM Cloud Object Storage

Component	Description	Products
Data RDBMS	A relational database management system (RDBMS) is a database that stores and processes data in a structured, tabular format as a collection of tables that consist of columns and rows, with relational operators to query data through Structured Query Language (SQL).	► IBM Db2® Family ► IBM Db2 on Cloud ► IBM Db2 Hosted ► IBM Db2 for z/OS®
Data warehouse	A data warehouse is a consolidated repository of integrated, conformed, and aggregated data from multiple and disparate data sources in support of business analytics and reporting. Data warehouses typically process structured data in tabular or relational form, often with history, on scalable relational database technology platforms that support large numbers of concur-rent users and complex queries across large data sets.	► IBM Integrated Analytics System ► IBM Db2 Warehouse on Cloud ► IBM Db2 Warehouse ► IBM Cloud Pak for Data System with IBM Performance Server for PostgresSQL
Event store	An event store is a database management system that implements the concept of event sourcing. Event stores persist all state-changing events for an object with a timestamp, creating time series for individual objects. The cur-rent state of an object can be inferred by replaying all events for that object from time 0 until the current time.	IBM Db2 Event Store
Data virtualization	Data virtualization is technology that connects all these data sources into a single self-balancing constellation. No longer are analytics queries that are performed on data that is copied and stored to a centralized location. The analytics application submits a query that is processed on the device where the data source is persisted. The results of the query are consolidated within the constellation and returned to the origin application. No data is copied. It remains persisted only at the source.	Data virtualization
Edge services	Edge services provide network capability to deliver content through the Internet (DNS, CDN, firewall, and load balancer). They handle the request and get it to the right destination. When the web application server completes its tasks, it delivers the resulting content back through the firewall, which passes the content to the user's browser.	Enterprise-specific

1.3.3 Cluster architecture

The Cloud Pak for Data runs on Red Hat OpenShift, which supports running it on any public cloud that supports Red Hat OpenShift and an on-premises private cloud cluster. The requirements of the cluster depend on the number of Cloud Pak for Data instances to be installed, which services to be installed on the top of the Cloud Pak for Data, and the types of workloads.

Typically, the Cloud Pak for Data can be deployed on a 3-node cluster for any nonproduction environments and the nodes to be increased on production environments. The number of increased nodes for the production environment ensures the high availability, which is a main requirement for production environments.

Figure 1-8 shows the topology of a production-level cluster.

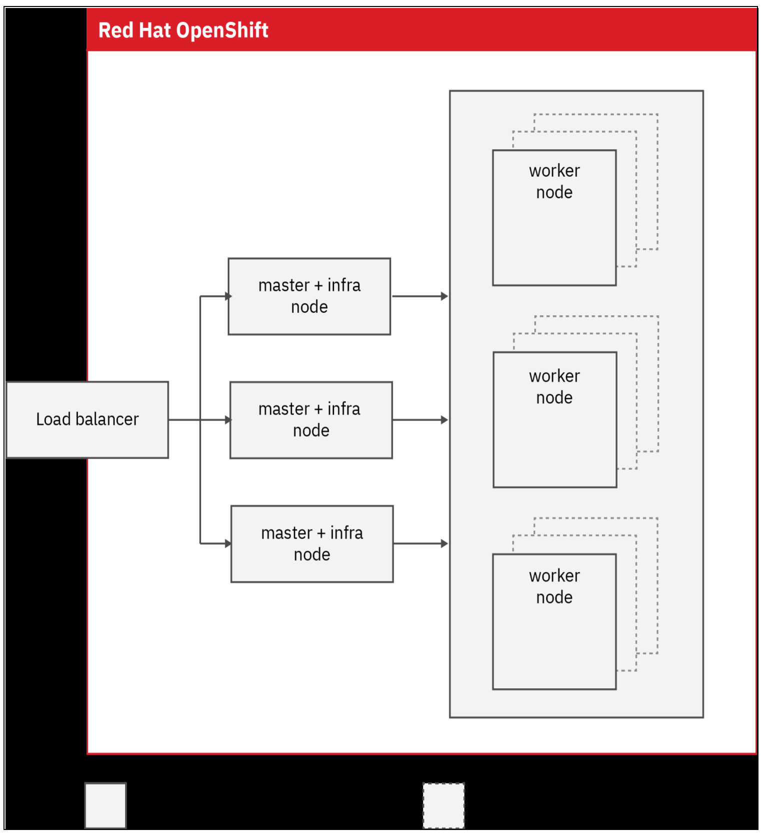

Figure 1-8 Production-like cluster for the Cloud Pak for Data

The load balancer can be in the cluster or external to the cluster. However, in a production-level cluster, an enterprise-grade external load balancer is recommended.

The load balancer distributes requests between the three master and infra nodes. The master nodes schedule workloads on the worker nodes that are available in the cluster. A production-level cluster must have at least three worker nodes, but you might need to deploy extra worker nodes to support your workload.

Modular platform

The platform consists of a light-weight installation that is called the Cloud Pak for Data control plane. The control plane provides an administration interface, a services catalog, and the central user experience. The control plane must be installed with each instance of the Cloud Pak for Data. It enables the needed tools, whether through CLI, user interface, administrator interface or the catalog to coordinate and interact with the services that are deployed with each instance.

Figure 1-9 shows the main components of the control plane.

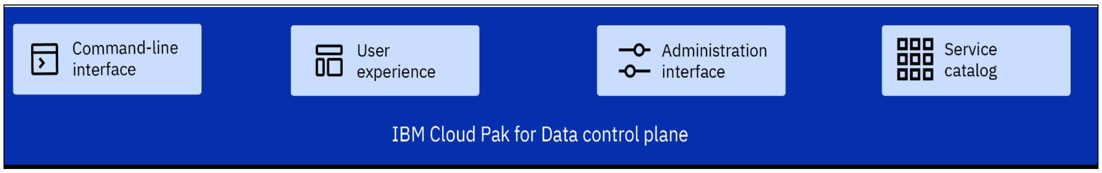

Figure 1-9 Cloud Pak for Data components of the control plane

Note: Figure 1-9 shows the command-line interface as part of the control pane, but it is not installed by default. It can be downloaded separately and used to connect to the control plane to perform various activities.

1.3.4 Multitenancy support

Gartner, uses the following definition for multitenancy:

Multitenancy is a reference to the mode of operation of software where multiple independent instances of one or multiple applications operate in a shared environment. The instances (tenants) are logically isolated, but physically integrated. The degree of logical isolation must be complete, but the degree of physical integration will vary.[2]

Achieving multitenancy with multiple instances of Cloud Pak for Data

In this pattern, the installation of multiple instances of Cloud Pak for Data is done on a single Red Hat OpenShift cluster. Each instance of Cloud Pak for Data is installed in a separate Red Hat OpenShift project (namespace).

This configuration offers complete logical isolation of each instance of Cloud Pak for Data with limited physical integration between the instances.

Red Hat OpenShift cluster administrator can create multiple projects (Kubernetes namespaces) to partition your cluster. Within each project, resource quotas can be assigned. Each project acts as a virtual cluster with its own security and network policies. In addition to being logically separated, different authentication mechanisms can be used for each Cloud Pak for Data deployment.

This tenancy model addresses the following use cases:

- Partitioning nonproduction environment from production environment in a continuous integration, continuous delivery (CICD) pipeline. In this model, tenants work in discrete, isolated units with a clear separation of duties.

[2] https://www.gartner.com/it-glossary/multitenancy

- Creating instances for different departments or business units that have distinct roles and responsibilities within your enterprise. In this model, each tenant has their own authentication mechanism, resource quotas, and assets.

This tenancy model also offers the following advantages:

- Minimizes overhead costs by deploying multiple instances on the same cluster.
- The cluster administrator can:
  - Establish tenant-specific quality of service characteristics in each instance
  - Assign project administrators to manage an instance of Cloud Pak for Data
- The project administrator can control which services are deployed in the project and can manage the resources that are associated with the project. However, the project administrator cannot access cluster-level settings or change the resource quotas for their project.

Achieving multitenancy within a single instance of Cloud Pak for Data

In this pattern, a single instance of Cloud Pak for Data is installed on Red Hat OpenShift cluster. The instance uses a single authentication mechanism for all users, and each user is assigned to the suitable role within the instance.

In this configuration, tenancy occurs at the resource level and users can see only resources to which they are granted access.

The following types of resources support logical isolation:

- Analytics projects

  Users must be added as collaborators to access the contents of a project. In this way, you can enforce logical isolation between projects. For example, you can create analytics projects to support specific teams or departments within your organization.

- Analytics deployment spaces

  Users must be added as collaborators to access the contents of an analytics deployment space. In this way, you can enforce logical isolation between deployment spaces.

- Services that support service instances

  Some services, such as integrated databases, can be deployed multiple times within a single deployment of Cloud Pak for Data. These deployments are called *service instances*. Users must be granted specific access to a service instance to interact with it. In this way, you can enforce logical isolation between service instances.

For information about services that support service instances, see this IBM Documentation web page.

For an extra layer of isolation, service instances can be deployed to separate projects, called tethered projects. For more information, see this IBM Documentation web page.

However, some services do not support service instances. The resources that are associated with those services are available to any users who can access the service. In some cases, all of the users who can access the instance of Cloud Pak for Data also can access the service.

Although this configuration is physically integrated, it does not support complete logical isolation. Also, you cannot partition the system to isolate tenant workloads or establish tenet-level resource quotas.

Tethered projects

A tethered project for a service instance is created during installation to isolate the service instance by deploying it to separate project. The service instance in the tethered project can be managed by Cloud Pak for Data, but is otherwise isolated from Cloud Pak for Data and the other services that run in the Cloud Pak for Data project.

Deploying a service instance to a tethered project is needed in the following cases:

- Running a custom application that needs to access a specific service instance, but for security reasons, the application must not access other services that are running in Cloud Pak for Data.
- Running a custom application or service instance that requires specific compute resources or a particular quality of service.

Because the tethered project is logically isolated from the main Cloud Pak for Data project, the tethered project can have its own network policies, security contexts, and quotas.

1.3.5 Operator installation architecture

The way that Cloud Pak for Data software operators are installed on the Red Hat OpenShift cluster depends on whether a need exists to enable the IBM Cloud Pak for Data platform operator to complete specific tasks (express installation) or whether more control is needed over how components are deployed (specialized installation).

Express installations

An express installation requires elevated permissions and does not enforce strict division between Red Hat OpenShift Container Platform projects (Kubernetes namespaces).

In an express installation, the IBM Cloud Pak foundational services operators and the Cloud Pak for Data operators are in the same project. The operators are included in the same operator group and use the same NamespaceScope Operator. Therefore, the settings that you use for IBM Cloud Pak foundational services also are used by the Cloud Pak for Data operators.

Specialized installations

A specialized installation allows a user with project administrator permissions to install the software after a cluster administrator completes the initial cluster setup.

A specialized installation also facilitates strict division between Red Hat OpenShift Container Platform projects (Kubernetes namespaces).

In a specialized installation, the IBM Cloud Pak foundational services operators are installed in the `ibm-common-services` project and the Cloud Pak for Data operators are installed in a separate project (typically `cpd-operators`).

Each project includes the following dedicated components:

- Operator group, which specifies the OwnNamespace installation mode.
- NamespaceScope Operator, which allows the operators in the project to manage operators and service workloads in specific projects.

In this way, different settings for the IBM Cloud Pak foundational services and the Cloud Pak for Data operators can be specified.

1.3.6 Storage architecture

Cloud Pak for Data supports NFS, Portworx, Red Hat OpenShift Container Storage, and IBM Cloud File Storage, as described next.

NFS storage

In this configuration, where an external NFS server can be used, a sufficiently fast network connection is required to reduce latency and ensure performance. NFS is installed on a dedicated node in the same VLAN as the cluster.

Red Hat OpenShift Container Storage

In this configuration, dedicated storage nodes can be used. Storage nodes can coexist with the worker nodes.

Because Red Hat OpenShift Container Storage uses three replicas, it is recommended to deploy Red Hat OpenShift Container Storage in multiples of three. Doing so makes it easier to scale up the storage capacity.

IBM Spectrum Scale Container Native

IBM Spectrum® Scale Container Native connects to a Spectrum Scale Storage Cluster through a remote network mount to provide access to the high-performance General Parallel File System (GPFS). IBM Spectrum Scale Container Native provides persistent data storage through the IBM Spectrum Scale Container Storage Interface Driver.

IBM Spectrum Scale Container Native and IBM Spectrum Scale Container Storage Interface Driver are deployed on the worker nodes of your Red Hat OpenShift cluster.

Portworx storage

Raw disks on the Red Hat OpenShift worker nodes must be added to use for storage (they can be the same nodes as the worker nodes where the services run). When Portworx is installed on the cluster, the Portworx service takes over those disks automatically and uses them for dynamic storage provisioning.

IBM Cloud File Storage

The `ibmc-file-gold-gid` and `ibm-file-custom-gold-gidstorage` classes are supported. The relative location of the storage is managed by the Red Hat OpenShift deployment on IBM Cloud.

For more information about specific requirements and considerations, see this IBM Documentation web page.

1.3.7 Base services

In this section, we discuss the base services that are available.

Common core services

Several Cloud Pak for Data services require similar features and interfaces. To streamline the platform, these features are provided by the Cloud Pak for Data common core services. These services are installed once in a project (namespace) and can be used by any service that requires one or more of the features.

The common core services provide data source connections, deployment management, job management, notifications, projects, and search (see Figure 1-10 on page 34).

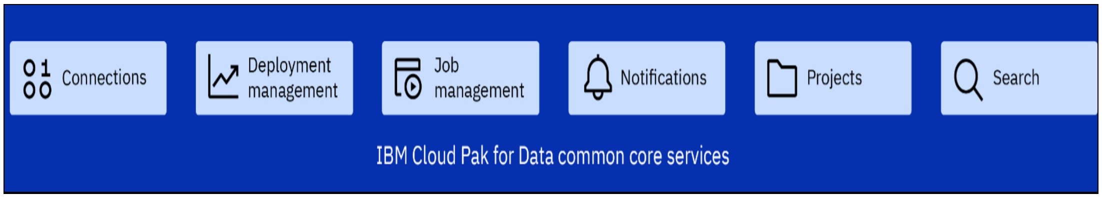

Figure 1-10 Core services of the Cloud Pak for Data

The common core services are automatically installed when you install a service that relies on them. If the common core services are already installed in the project (namespace), the service uses the existing installation.

Integrated data and AI services

The services catalog includes a broad range of offerings from IBM and from third-party vendors. The catalog contains the following types of services:

- AI
- Analytics
- Dashboards
- Data governance
- Data sources
- Developer tools
- Industry solutions
- Storage

Figure 1-11 shows the Integrated data and AI services that can be installed and configured by the control plane.

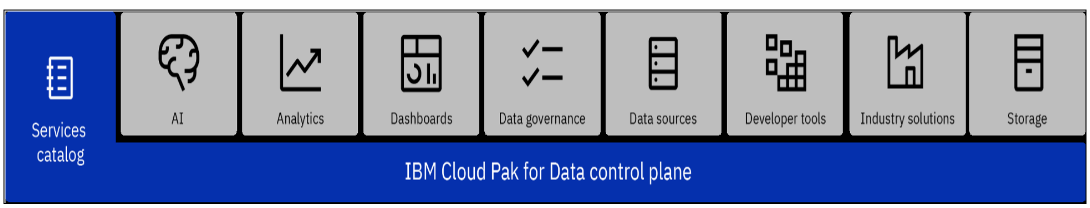

Figure 1-11 Data and AI services in the Cloud Pak for Data

For the example that is shown in Figure 1-12, if data governance and data science is a concern, installation is done for several AI services and analytics services, data governance services, and developer tools that support the developers and data scientists who use Cloud Pak for Data. In addition, an integrated database can be deployed to store the data science assets, which are generated as a result of the use of Cloud Pak for Data.

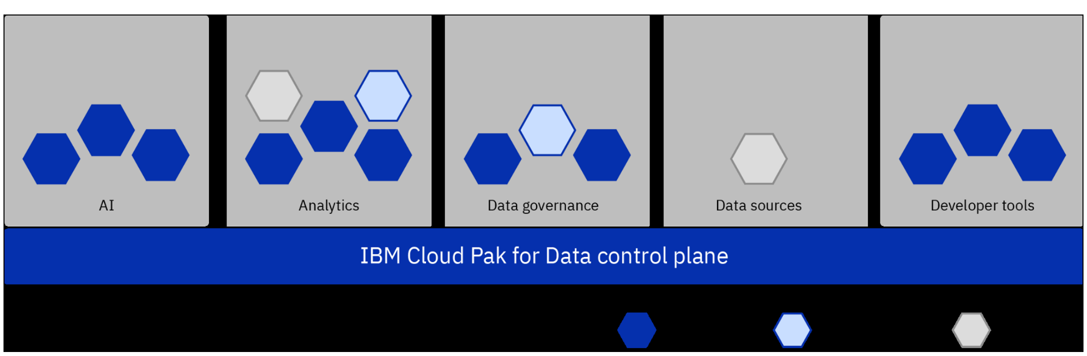

Figure 1-12 Instances of Data and AI services that are installed on top of the Cloud Pak for Data

The number of services that are installed on the Cloud Pak for Data control plane and the workloads that run for each service determine the needed resources.

1.4 Business value

The year 2022 saw more implementations in the data fabric space, but many such implementations fail to meet business expectations. The reason is simple: data issues.

1.4.1 IBM Cloud Pak for Data version 4.5

IBM Cloud Pak for Data provides an insight platform that combines data management with data science and AI development to help firms realize value. IBM commissioned Forrester Consulting to conduct a Business Value Engineering Assessment based on their Total Economic Impact Methodology.

The original study[3] was conducted in December 2020 for version 3.0 of Cloud Pak for Data and the assumptions and data were applied against version 4.5, which was released in July 2022[4]. Improvements to function, performance, and business value are impressive from the earlier version to the latest instance.

IBM commissioned one of their certified Enterprise Architects to review the original Forrester study and extrapolate to version 4.5, taking into account all of the improvements, while considering the latest technological environment in 2022.

Forrester interviewed four customers in late 2020 who had experience in the use of Cloud Pak for Data version 3.0.

The interviewed customers provided their assessment of the data fabric landscape before undertaking Cloud Pak for Data version 3.0. The following challenges were reported:

- Cloud Migration preparation
- No cohesive governance strategy
- Difficulty in managing multiple point solutions

Migrating to Cloud Pak for Data provided significant potential in the following areas:

- Containers and container management efficiencies
- Data governance and data virtualization
- Data science, ML, and AI integration

A composite organization was developed from four companies who implemented Cloud Pak for Data. The composite company is a global organization with $2 billion in annual revenue, 8,000 employees, and deployed on-premises solutions in all four functional areas of Cloud Pak for Data (Collect, Organize, Analyze, and Infuse).

The Business Value Engineering framework identified the following potential investment factors:

- Cost
- Benefit
- Flexibility
- Risk

[3] New Technology: The Projected Total Economic Impact Of IBM Cloud Pak For Data https://www.ibm.com/downloads/cas/V5GNQKGE

[4] New Technology: The Projected Total Economic Impact Of Explainable AI And Model Monitoring In IBM Cloud Pak For Data https://www.ibm.com/downloads/cas/DZ8N68GD

Due diligence

In this study, IBM stakeholders and Forrester analysts were interviewed to gather data that was relative to Cloud Pak for Data.

Early-implementation customer interviews

Four organizations were interviewed during early stages of adoption, such as pilots or betas.

Composite organization

A composite organization was developed based on the conducted interviews.

Projected financial model framework

A Business Value Engineering model was created by using Forrester's Total Economic Impact methodology.

Case study

The following elements of the Total Economic Impact methodology were used:

- Cost
- Benefit
- Flexibility
- Risk

Interviewed organizations: Cloud Pak for Data

For this study, Forrester conducted four interviews with IBM Cloud Pak for Data customers. Interviewed customers include the following types:

- Industry
  - Mobile and telecom services
  - Financial services
  - Financial services (North America)
  - Consulting services
- Region

  The participants were from the following regions:
  - APAC
  - EMEA
  - North America Global
  - Global
- Interviewee

  Interviewee job roles included the following examples:
  - General manager, telecommunications, and emerging technologies
  - Director of data management
  - Information architect
  - Managing director of artificial intelligence innovation
- Number of employees:
  - 80
  - 5,000
  - 50,000
  - 100,000+

- Annual revenue:
  - $15 million
  - $2 billion
  - $10 billion+
  - $10 billion+

Key challenges before IBM Cloud Pak for Data

The four interviewed organizations felt Cloud Pak for Data was the optimal solution for them for the following reasons:

- Difficulty migrating fully to the cloud in the past.
- Ever-increasing volumes of data without a governance strategy with previous solutions.
- Managing multiple point solutions was cumbersome and inefficient before Cloud Pak for Data.
- Desire to integrate Data Science, ML, and AI into a single platform was seamless with Cloud Pak for Data.

Key projected results with IBM Cloud Pak for Data

The interviews indicated that customers invested in Cloud Pak for Data to address the challenges previously identified, and were seeing early benefits indications in the following areas:

- Containers and container management efficiencies

  With Cloud Pak for Data, companies can improve their readiness for cloud migration, improve licensing flexibility with IBM, and reduce hardware purchases and infrastructure management efforts.

- Data virtualization and governance benefits

  Data virtualization “democratizes” data visibility across the organization, improves data governance and security, and allows companies to avoid costly data migration projects.

- Data science, ML, and AI benefits

  Data scientists are more productive with Cloud Pak for Data and can deploy models to market faster. Also, because of Cloud Pak for Data’s integrated platform, companies avoided costs that are associated with older analytics tools or otherwise build a comparable solution internally.

Composite organization

The composite organization features the following characteristics:

- It is a global enterprise with $2 billion in annual revenue and 8,000 employees.
- Has five separate, large-scale data management infrastructure (for example, data stores) that are in different countries.
- Already employs five data scientists and uses various data analytical tools.
- Has organization-wide decision to pursue container management.
- It deployed on-premises solutions in all four functional areas of Cloud Pak for Data (Collect, Organize, Analyze, and Infuse).

Risk treatment for benefits and costs projections

A risk factor was applied to cost and benefit calculations.

Total Economic Impact methodology

The following are the defined terms for the Total Economic Impact methodology:

- Projected Benefits represent the projected value to be delivered to the business by the product.
- Projected Costs consider all expenses necessary to deliver the proposed value, or benefits, of the product.
- Flexibility represents the strategic value that can be obtained for some future extra investment that builds on top of the initial investment that was made.
- Risks measure the uncertainty of benefit and cost estimates given the likelihood that estimates meet original projections and that estimates are tracked over time.

1.4.2 IBM OpenPages with Watson

IBM provides an integrated governance, risk, and compliance (GRC) platform that helps its customers identify risk, anticipate uncertainty, and proactively respond to changes in market conditions and inside their organization.

Five customers were interviewed[5]. IBM OpenPages was selected to:

- Provide a scalable, integrated GRC platform.
- Deliver a holistic view of risk and compliance.
- Make GRC accessible to business users across the enterprise.

A composite organization was modeled. The composite organization is a global financial services company, and it deployed IBM OpenPages with 750 user accounts.

All values are reported in risk-adjusted, three-year present value (PV) unless otherwise indicated.

Interviewed organizations - IBM OpenPages with Watson

Five interviews, all from financial services, were conducted with IBM OpenPages customers and the following data points Identified:

- Region

  Interviewed one customers from North America, and other global customers.

- Interviewee:
  - Director
  - Operational risk coordinator
  - Vice president (VP)
  - Head of risk
  - Senior manager
- Number of employees:
  - 5,000
  - 20,000+
  - 20,000+
  - 15,000
  - 5,000

[5] The Total Economic Impact Of IBM OpenPages® with Watson https://www.ibm.com/downloads/cas/5WOQRKQP

Key challenges before IBM OpenPages with Watson

Before the investment in OpenPages, interviewees described the following challenges with their previous solution:

- Risk management tools and silos
- Reactive decision making
- Difficulty satisfying auditors and regulators

Why IBM OpenPages with Watson?

IBM OpenPages was chosen for the following reasons:

- Single, integrated platform
- Improved user experience
- Reduced risk management effort
- Regulatory penalty and fines avoidance

Composite organization

The composite organization included the following details:

- Global financial services organization with $10 billion in annual revenue and 10,000 employees.
- Eight departments responsible for GRC, which previously used disparate GRC solutions and spreadsheets to track organizational risk.
- Deployed IBM OpenPages with:
  - Watson SaaS including Operational Risk Management
  - IT Governance
  - Policy Management
  - Internal Audit Management
  - Regulatory Compliance Management modules

The composite organization plans to implement Financial Controls Management and Third-Party Management modules in the future.

- A total of 750 OpenPages user accounts:
  - 600 first-line users: 500 managers and 100 business unit risk professionals
  - 100 second-line users (enterprise risk professionals)
  - 50 third-line users (internal audit professionals)

IBM OpenPages overview

IBM OpenPages provides the following solutions that span risk and compliance domains and can be used independently or together:

- Operational Risk Management enables businesses to integrate risk data within a single environment and automate identifying, measuring, monitoring, analyzing, and managing operational risk processes.
- Business Continuity Management helps organizations in their effort to develop and manage business continuity and IT resiliency by strengthening their preparedness to respond to future unforeseen events with visibility across risks.
- Regulatory Compliance Management reduces the time and costs of understanding regulatory requirements. It also reduces risks, such as sanctions and fines, because of noncompliance. It catalogs regulations and creates actionable tasks.
- IT Governance reduces complexity of IT risk management by aligning IT operations management with business initiatives, strategy, and regulatory requirements.

- Policy Management simplifies compliance with numerous industries, ethics, privacy, and government regulations. Identifies similarities between regulations to reduce duplicated efforts.
- Internal Audit Management provides internal auditors with a uniquely configured, cross-departmental view into organizational GRC. Automates and standardizes auditing procedures.
- Model Risk Governance combines a flexible data model for key stakeholders with document risk management, workflow, and business intelligence.
- Third-Party Risk Management enables customers to understand their third-party risk exposures. Customers can access the Santa Fe Group's shared assessments SIG questionnaires to support the process.
- Financial Controls Management reduces costs and simplifies compliance with Sarbanes-Oxley Act and similar global reporting regulations.

1.4.3 IBM Watson Assistant

IBM Watson® Assistant is IBM's chatbot with which users interact with business systems by using natural human language.

Several customers were interviewed across the following use cases[6]:

- Agent assist
- Customer self-service
- Employee self-service

Composite organization

The composite organization featured the following characteristics:

- Revenue: $10 billion
- Geography: Headquartered in Europe with worldwide operations
- Employees: 40,000
- Monthly conversations: 1 million

Key findings

The following are the key findings:

- Quantified benefits

  The organization achieves the following benefits:

  - The organization realizes cost savings by using IBM Watson Assistant.
  - Employee self-service drives containment and reassignment of HR and IT help desk agents.
  - Chatbot-augmented agents reduce handle time.
  - Correctly routed conversations per correctly routed call.
  - Creates a self-serve, digital-first experience provides a competitive advantage. Agent experience also improves.
  - IBM Watson Assistant can be integrated into the channels most are used by customers.

[6] The Total Economic Impact Of IBM Watson Assistant
https://www.ibm.com/watson/assets/duo/pdf/watson_assistant/The_Total_Economic_Impact_of_IBM_Watson_Assistant-March_2020_v3.pdf

- IBM Watson adds capacity Constant, 24x7x365 automated coverage, reduces time-to-resolution, and provides help to customers when they need it.
- Brand perception improves when combined with AI.

- Costs
  - IBM licenses
  - Internal labor costs for implementing workflows
  - Conversation analysts
  - Professional services fees

Interviewed organizations: IBM Watson Assistant

The interviewed organizations featured the following characteristics:

- Industries:
  - Software
  - Financial services
  - Financial services
  - Automotive
- Headquarters size:
  - $2 billion, 10,000 employees
  - $12 billion, 71,000 employees
  - $74 billion, 109,000 employees
  - $185 billion, 300,000 employees
- Regions:
  - North America
  - Europe
  - South America
  - Europe
- Interviewee job roles:
  - Senior product manager
  - Head of digital supply chain innovation
  - AI manager, research, and innovation
  - Product owner

Key challenges

The following key challenges were identified:

- Limited service hours created a poor customer experience.
- Multi-step routing journeys and long wait times created a frustrating customer experience.
- Traditional call centers were costly and difficult to scale to new channels.
- Agents did not have the right knowledge and data.

Why IBM Watson Assistant?

IBM Watson Assistant was chosen for this study for the following reasons:

- Freedom to tailor Watson Assistant to multiple use cases
- Digital experience and cost savings
- Features and technical product capabilities
- Ability to deploy agilely

Use cases

The following use cases were considered in this study:

- Agent assist
- Customer self-service
- Employee self-service

Composite organization

The global enterprise is headquartered in Europe, generates $10 billion in revenue, and has 40,000 employees. The organization is in a highly regulated industry with nuanced products.

1.4.4 IBM Watson Natural Language Processing

Natural language processing (NLP) solutions can help organizations turn unstructured data into insights. NLP also makes information easy for customers and knowledge workers to access and understand, which improves knowledge worker efficiency, boosts business growth, and increases knowledge worker and customer satisfaction.

Key findings

The following key findings were made[7]:

- The organization achieved the following benefits:
  - Knowledge workers that previously spent 20% of their time on text analysis or search tasks reduced that time by 50%.
  - Established tools can be replaced by the IBM NLP solutions.
  - The efficiency and accuracy of IBM NLP tools delivered another 5% in business growth per year.
- The following costs were involved:
  - IBM license
  - Developer and subject matter expert (SME) time to build and train the organizations' NLP applications
  - Training for the NLP application users

Knowledge worker definition

The following knowledge worker definitions were used in this study:

- Analyzing data to establish relationships.
- Assessing input to evaluate complex or conflicting priorities.
- Identifying and understanding trends.
- Making connections.
- Understanding cause and effect.

Knowledge workers depend on information to do their jobs well, and their expertise is the backbone of most enterprises. These workers span industries and specialties, and include many positions, such as the following examples:

- Physicians and pharmacists
- Programmers
- Lawyers
- Maintenance professionals

[7] The Total Economic Impact Of IBM Watson Natural Language Processing (NLP) Solutions
https://www.ibm.com/downloads/cas/XMRMP7XK

- Researchers and analysts
- Design thinkers
- Accountants
- Media specialists

Key challenges

The following key challenges were identified:

- The organizations' business operations relied on understanding voluminous amounts of unstructured data.
- Deficiencies in traditional search and text analytics tools reduced the efficiency of knowledge workers.
- Business operations did not scale with established tools and processes that were in place.

IBM Natural Language Processing solutions

The interviewed organizations used a combination of the following IBM solutions:

- IBM Watson Discovery

  IBM Watson Discovery finds precise answers and extracts high-quality insights from enterprise documents and web page data including, PDFs, HTML, tables, and images. It is an intelligent, AI-powered search and text analytics platform. IBM Watson Discovery uses NLP to understand the unique language of the business, regardless of domain or industry.

- IBM Natural Language Understanding

  IBM Watson Natural Language Understanding (NLU) uses deep learning to extract metadata, such as keywords, concepts, sentiment, named entity recognition (NER), and categories from natural language text. Watson Natural Language Understanding can be used in Watson Discovery or as a stand-alone service.

Composite organization

The composite organization is a large organization in a services industry that employs highly skilled knowledge workers to deliver revenue-generating services to customers. These knowledge workers review complex documents and collect information from various sources to perform their jobs.

They can access rudimentary text analytics and search tools to help their data collection efforts, but, given the limitations of these tools, they still spend a significant amount of time on these tasks.

Deployment characteristics

The composite organization invests in the portfolio of IBM Watson NLP solutions (IBM Watson Discovery and IBM Watson NLU) to build a single application that can extract key information from documents and surface that information for user queries.

Developer staff build the application, and SMEs work alongside those developers to train the application to understand the structure of the documents that information is pulled from and to ensure that insights and results are accurate.

Initially, the application is rolled out to 200 users. However, this number grows to 400 total users in years 2 and 3, as shown in Figure 1-13.

Benefits - Risk Adjusted	Year 1	Year 2	Year 3	Total	PV @ 10%
Container Management	$ 2,940,000	$ 2,940,000	$ 2,940,000	$ 8,820,000	7,272,97
Data Virtualization	$ 197,400	$ 197,400	$ 197,400	$ 592,200	488,32
Data Science, ML, AI	$ 482,900	$ 837,100	$ 1,185,800	$ 2,505,800	2,066,28
GRC/OpenPages	$ 2,005,200	$ 2,935,950	$ 1,626,000	$ 6,567,150	5,415,27
Customer Care/Watson Assistant	$ 4,439,736	$ 11,138,532	$ 18,992,526	$ 34,570,794	28,507,07
Customer Care/Watson NLP	$ 931,788	$ 2,933,581	$ 5,029,525	$ 8,894,894	7,334,72
TOTALS	$ 10,997,024	$ 20,982,563	$ 31,979,587	$ 61,950,838	51,084,66

Costs - Risk Adjusted	Initial	Year 1	Year 2	Year 3	Total	PV @ 10%
Licensing, Professional Services, Support	$ 347,550	$ 60,900	$ 60,900	$ 60,900	$ 530,250	437,24
Implementation and Operations	$ 80,600	$ 822,900	$ 822,900	$ 822,900	$ 2,468,700	2,035,69
GRC - Licensing, Support, Implementation	$ 328,900	$ 710,850	$ 483,150	$ 483,150	$ 2,006,050	1,654,18
Customer Care/Watson Assistant	$1,211,910	$1,025,955	$1,927,236	$2,676,975	$ 6,842,076	5,641,97
Customer Care/Watson NLP	$ 617,689	$ 303,479	$ 278,783	$ 278,783	$ 1,478,735	1,219,36
TOTALS	$2,586,649	$2,924,084	$3,572,970	$4,322,708	$ 13,325,811	10,988,46

Benefits - Risk Adjusted	Costs - Risk Adjusted	Risk	TEI
$ 51,084,661	$ 10,988,464	15%	365%

Figure 1-13 Key findings

2

Cloud Pak for Data services overview

In this chapter, we introduce the services that are available in Cloud Pak for Data. A high-level overview is provided for each of the services, describing the purpose and the high-value features.

In this overview, a brief explanation is included about working with the service. Set up and configuration advice also is provided that you can consider when you are deploying the service.

Then, we include references for you to get more in-depth information about the service.

The chapter includes the following topics:

2.1 Introduction

IBM Cloud Pak for Data is a single interface-integrated platform that is designed to help unify and simplify collection, organization, and analysis of data. It offers a set of unified, pre-integrated data, and AI services that are delivered within a modular cloud native platform and architecture.

The platform offers a wide range of capabilities across the entire data and AI lifecycle, including data management, ETL, data engineering and ingestion, DataOps, data governance, data analysis, AI/ML, Data Science and MLOps, and business intelligence and visualization.

Collectively, Cloud Pak for Data services implement the Ladder to AI, as described in Chapter 1, "Cloud Pak for Data concepts and architecture" on page 7 and help accelerate your AI journey. The platform's architectural design also makes it suited for data fabric scenarios. Cloud Pak for Data is a data fabric solution that enables various data fabric use cases, such as Multi-Cloud Data Integration, MLOps and Trustworthy AI, Customer 360, Governance & Privacy, and more.

> **Note:** The various services that are available and the platform's composable architecture also means that you do not have to use everything that is available on the platform. Instead, you deploy, scale, and use exactly what you need for your specific use cases and business challenges by choosing any relevant combination of Cloud Pak for Data services and service sizes.

The following concepts are common across all services:

- Connections are data assets that contain the information that is necessary to establish a connection to a data source. After a connection is set up, it can be used to access and ingest metadata and data from the source system.

  Credentials that are used to define the connection determine the baseline level of access to source data (for example: read only, read and write, all tables, and files and schemas or only specific files). Extra security measures can then be layered on within Cloud Pak for Data.

  Connections can be designated as Shared (credentials for the source are entered once during the connection setup and then are reused by all users of the connection) or Personal (users must unlock access to the data by entering their own credentials to the data source).

  No data is moved or copied from the connection by default; instead, the data stays at source until and unless relevant analytical and data integration and movement activities are started through the platform.
- Connected data assets are pointers to data that is accessed through a connection to an external data source and represent relevant tables, views, or files within the platform. Connected assets rely on connections and are created by specifying the relevant connection, any intermediate structures or paths, and a relational table or view, a set of partitioned data files, or a file. Similar to connections, connected data assets are metadata representations of remote tables, views, or files and do not involve ingestion, movement, or copying of the entire data set from the source system to the platform by default.
- Catalogs help organize and group data assets (connections, connected assets, and local files) and control and restrict access to those by user and user group. Each user and user group can have a different role assignment for different catalogs (Administrator, Editor, or Viewer). If a user is not assigned to a catalog, they do not see it in the system.

- Projects are shared collaborative workspaces that provide various tools for achieving a specific goal; for example, building a data model or integrating data. The tools that available depend on which services are deployed on your Cloud Pak for Data cluster.

  Each project can have a different set of data (connections, connected assets, files, and so on) added to it directly or by using catalogs, and a different collaborator list (users, user groups) and user role assignment (Administrator, Editor, or Viewer). If a user is not assigned to a project, they do not see it in the system.
- Service instances are individual deployments (copies) of a service on your Cloud Pak for Data cluster. Some services can exist only as a single shared instance (one per cluster) and is deployed during service installation, while some services support multiple service instance deployments. For the latter, following service installation, one or more instances of the service need must be provisioned.

  Each of the instances can then have their own set up and users (user access and service role assignment). The user who deployed the instance typically is assigned the service-specific administrator role for that instance automatically. No other user can access any of the instances by default, and must be assigned as a user of the instance and allocated relevant roles that are available with the service. Examples of instance-based services include IBM Data Virtualization and IBM Db2 Warehouse.
- Deployment spaces are workspaces that help you organize your model deployments. They contain deployable assets, such as model deployments, jobs, associated input and output data, and the associated environments.

  Because deployment spaces are not associated with a project, you can deploy assets from multiple projects into a single space. You might have a space for test, pre-production, and production. For more information, see 2.5.3, "Watson Machine Learning" on page 90.

The remainder of this chapter focuses on each service that is available in Cloud Pak for Data. A high-level overview for each of the services is provided so that you can get a general understanding of the main purpose and features of the service.

2.2 Data sources

This section highlights the data source services that are available with Cloud Pak for Data.

2.2.1 Data Virtualization

The Data Virtualization service is a virtual data platform with which you can access your data from multiple locations across your enterprise as a single data view. This single data view is a virtualized table that is read-only and is used to access your data from a central location, no matter where your data is physically located or hosted. You query your virtual tables by using standard SQL through common interfaces from basic command line, to development tools, and embedded applications.

This virtual data platform improves control over your enterprise information by centralizing access control. It provides a robust security infrastructure, and reduces physical copies of your data. This feature enables your virtualized tables to be used by multiple users in different projects in a controlled and trusted environment.

These virtual tables provide your users with a simplified representation of your enterprise data without having to know the complexities of the physical data layer or where the data is stored. This feature delivers a complete view of your data for insightful analytics in real time without moving data, duplication, ETL, or other storage requirements.

Creating virtual tables

You use the Data Virtualization service to create virtual objects from a single table, multiple tables, or files. It also is possible to create a join view from multiple virtualized tables. Figure 2-1 shows the three main phases when creating virtual objects:

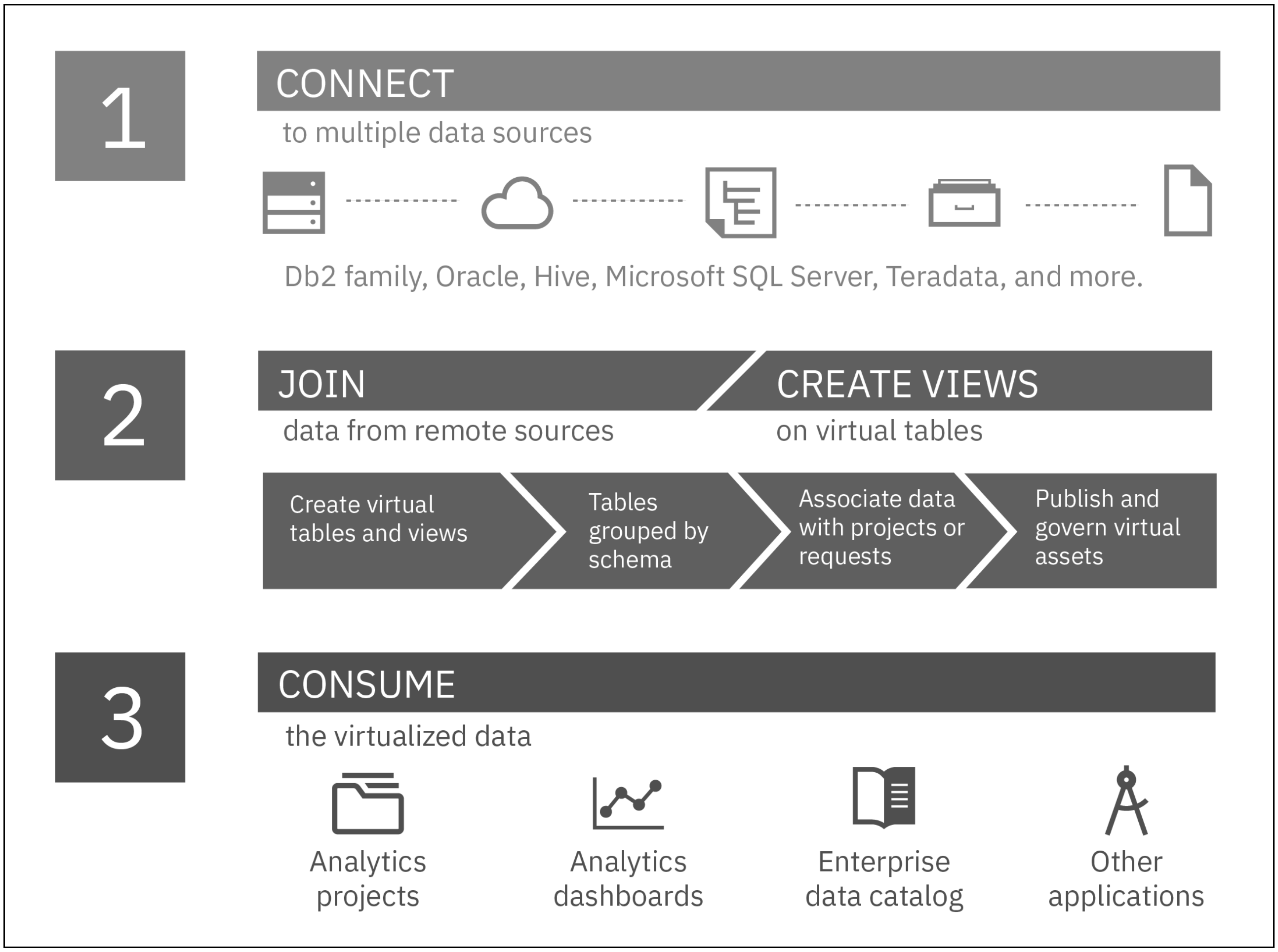

Figure 2-1 Virtualizing data with the Data Virtualization service instance

1. Connect: You start by connecting to your single data source or multiple data sources
2. Join, create, and then govern: You can create your virtual tables, which includes grouping your tables by schema. After this process is complete, you can govern your virtual tables as assets when you associate the data with projects.
3. Consume: Now you are ready to use your virtual table in an analytics project, dashboard, data catalogs, or other applications.

Initial setup and configuration considerations

Installing the Data Virtualization service also provisions an IBM Db2 Data Management Console, if one does not exist. After you install the Data Virtualization service, you must first provision an instance of the service and configure the network (optionally, you can install the remote connector).

When you are provisioning an instance, you must specify the number of worker nodes and the number of cores and memory to allocate. Then, you specify the node storage and cache storage.

After the instance is provisioned, you must find the exposed ports for your applications to connect to the Data Virtualization service instance in Cloud Pak for Data. This process might include configuring your local firewall rules or load balancers. Now that the Data Virtualization service instance is provisioned and the network is configured, you can manage users, connect to multiple data sources, create and govern virtual assets, and use the virtualized data.

For more information about the Data Virtualization service, see the following resources:

- IBM Documentation:
  - Data Virtualization on Cloud Pak for Data
  - Preparing to install the Data Virtualization service
  - Installing Data Virtualization
  - Postinstallation setup for Data Virtualization
- Tutorials:
  - Data virtualization on IBM Cloud Pak for Data
  - Create a single customer view of your data with Data Virtualization
- APIs
- Chapter 4, “Multicloud data integration” on page 163

2.2.2 Db2 Data Management Console

The Db2 Data Management Console service is a database management tool platform that you can use to administer and optimize the performance of your integrated IBM Db2 databases on Cloud Pak for Data. These integrated databases include Db2, Db2 Warehouse, Db2 Big SQL, and Data Virtualization, which you can manage and monitor from a single user interface console.

By using this console, you can perform the following tasks for your integrated databases:

- Administer databases
- Work with database objects and utilities
- Develop and run SQL scripts
- Move and load large amounts of data into databases for in-depth analysis
- Monitor the performance of your Cloud Pak for Data integrated Db2 database

Using the Db2 Data Management Console

The console home page provides an overview of all of the Cloud Pak for Data integrated databases that you are monitoring. This home page includes the status of database connections and monitoring metrics that you can use to analyze and improve the performance of your databases.

From the console, you can perform the following tasks:

- Explore integrated databases through its schemas, tables, views, and columns.
- Monitor integrated databases through key metrics such as Availability, Responsiveness, Throughput, Resource usage, Contention, and Time Spent.
- Run SQL and maintain scripts for reuse.
- Load data from flat files that are stored on various storage types.
- Tune single SQL statements and query workloads.
- Create and schedule jobs.
- Manage alerts.

- Create monitoring reports to compare and analyze different data sets.
- Set up and manage monitor profiles and event monitor profiles.

Figure 2-2 shows the Summary page.

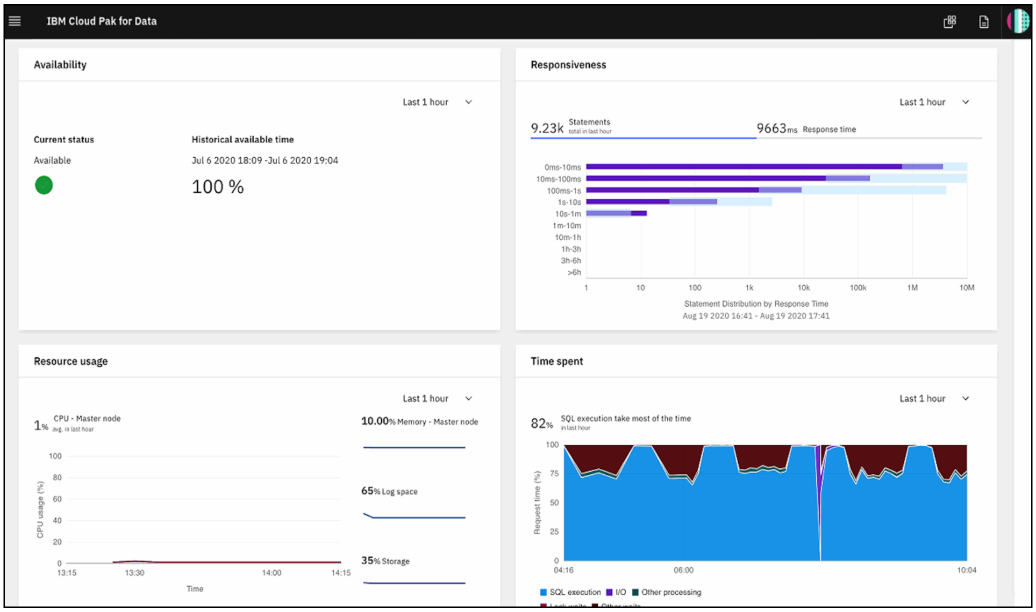

Figure 2-2 Db2 Data Management Console Summary Page

Initial setup and configuration considerations

After installing the Db2 Management Console, you must provision an instance of the service. Only one instance of the console can exist in a Cloud Pak for Data deployment.

To provision an instance, you first select the plan size for the compute resources: small, medium, or large. Then, you configure the storage resources by providing the storage class and the amount of storage for your persistent storage.

When the console instance is provisioned, you can start to use the console to manage and maintain your integrated databases.

Learn more

For more information about the Db2 Data Management Console service, see the following resources:

- IBM Documentation:
  - IBM Db2 Data Management Console on Cloud Pak for Data
  - Installing Db2 Data Management Console
  - Provisioning the service (Db2 Data Management Console)
- Db2 Data Management Console for Cloud Pak for Data demonstration
- APIs
- Chapter 4, “Multicloud data integration” on page 163.

2.2.3 IBM Db2

IBM Db2 database is a world class, enterprise relational database management system (RDBMS). Db2 provides advanced data management and analytics capabilities for your online transactional workloads (OLTP).

The scalability of Db2, which includes the number of cores, memory size, and storage capacity, provides an RDBMS that can handle any type of workload. These capabilities are available in the Db2 service that is deployed as a set of microservices that is running in a container environment. This containerized version of Db2 for Cloud Pak for Data makes it highly secure, available, and scalable without any performance compromises.

Db2 databases are fully integrated in Cloud Pak for Data, which enables them to work seamlessly with the data governance and AI services to provide secure in-depth analysis of your data.

By using the Db2 operator and containers in Cloud Pak for Data, you can deploy Db2 by using a cloud-native model, which provides the following benefits:

- Lifecycle management: Similar to a cloud service, it is easy to install, upgrade, and manage Db2.
- Ability to deploy your Db2 database in minutes.
- A rich ecosystem that includes Data Management Console, REST, and Graph.
- Extended availability of Db2 with a multitier resiliency strategy.
- Support for software-defined storage, such as Red Hat OpenShift Data Foundation, IBM Spectrum Scale CSI, and other world leading storage providers.

Figure 2-3 shows the layers of the Db2 architecture.

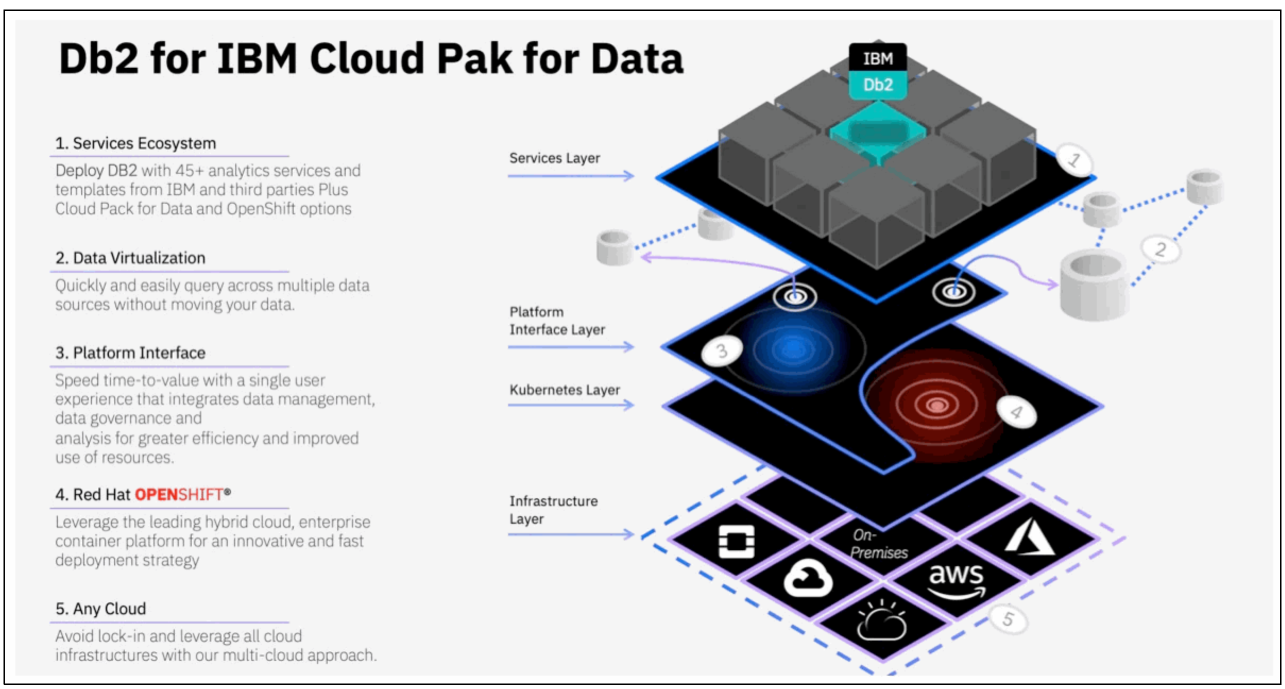

Figure 2-3 Db2 for Cloud Pak for Data

Working with a Db2 database

Integrating a Db2 database with Cloud Pak for Data can be useful in the following situations:

- You need your transactional data to be governed, such as data from a website, bank, or retail store.
- You want to create a replica of your transactional database so that you can run analytics without affecting regular business operations.
- You need to ensure the integrity of your data by using an ACID-compliant database.
- You need a low-latency database.
- You need real-time insight into your business operations.

After you create a Db2 database, you can use the integrated database console to perform common activities to manage and work with the database. From the console, you can perform the following tasks:

- Explore the database through its schemas, tables, views, and columns, which include viewing the privileges for these database objects.
- Monitor databases through key metrics, such as Availability, Responsiveness, Throughput, Resource usage, Contention, and Time Spent.
- Manage access to the objects in the database.
- Load data from flat files that are stored on various storage types.
- Run SQL and maintain scripts for reuse.

Initial setup and configuration considerations

Setting up the Db2 service and databases in Cloud Pak for Data requires some extra steps and considerations compared to some of the other Cloud Pak for Data services.

Before installing the Db2 service, consider the use of dedicated compute nodes for the Db2 database. In a Red Hat OpenShift cluster, compute nodes or worker nodes runs the application.

Installing Db2 on a dedicated compute node is recommended for production and is important for databases that are performing heavy workloads. Setting up dedicated nodes for your Db2 database involves Red Hat OpenShift taint and toleration to provide node exclusivity. You also must create a custom security context constraint (SCC) that is used during the installation.

After installing the Db2 service and before creating your database, consider disabling the default automatic setting of interprocess communication (IPC) kernel parameters so that you can set the kernel parameters manually. Also, consider enabling the `hostIPC` option for the cluster so that kernel parameters can be tuned for the worker nodes in the cluster. Doing so allows you to use the Red Hat OpenShift Machine Config Operator to tune the worker IPC kernel parameters from the control or the master nodes.

Now you can create your database in your Cloud Pak for Data cluster. You can specify the number of nodes that can be used by the database, including the cores per node and memory per node. However, you also can specify to use dedicated nodes by specifying the label for the dedicated nodes.

You also can set the page size for the database to 16 K or 32 K. One of the last steps is to set the storage locations for your system data, user data, backup data, transactional logs, and temporary table space data. This data can be stored together in a single storage location, but it is advised to consider the use of separate locations, especially among the user data, transactional logs, and backup data.

After the database is created, you can now start using the database by creating your first set of tables and loading data into the tables.

Learn more

For more information about the Db2 service, see the following resources:

- IBM Documentation
  - Db2 on Cloud Pak for Data
  - Preparing to install the Db2 service
  - Installing the Db2 service
  - Postinstallation setup for the Db2 service
- Video: Db2 on IBM Cloud Pak for Data platform
- Blog: The Hidden Historyof Db2

2.2.4 Db2 Warehouse

IBM Db2 Warehouse is an enterprise ready data warehouse that is used globally. Db2 Warehouse provides in-memory data processing, columnar data store, and in-database analytics for online analytical processing workloads (OLAP).

The scalability and performance of Db2 Warehouse through its massively parallel processing (MPP) architecture provides a data warehouse that can handle any type of analytical workloads. These workloads include complex queries and predictive model building, testing, and deployment.

IBM Cloud Pak for Data automatically creates the suitable data warehouse environment. For a single node, the warehouse uses symmetric multiprocessing (SMP) architecture for cost-efficiency. For two or more nodes, the warehouse is deployed by using an MPP architecture for high availability and improved performance.

By using the Db2 Warehouse operator and containers in Cloud Pak for Data, you can deploy Db2 Warehouse that uses a cloud-native model and provides the following values:

- Lifecycle management: Similar to a cloud service, it is easy to install, upgrade, and manage Db2 Warehouse.
- Ability to deploy your Db2 Warehouse database in minutes.
- A rich ecosystem: Data Management Console, REST, and Graph.
- Extended availability of Db2 Warehouse with a multitier resiliency strategy.
- Support for software-defined storage, such as Red Hat OpenShift Data Foundation, IBM Spectrum Scale CSI, and other world leading storage providers.

Using the Db2 Warehouse database

Integrating a Db2 Warehouse database with Cloud Pak for Data can be useful in the following situations:

- You have developers who must create small-scale database management systems for development and test work. For example, if you need to test new applications and data sources in a development environment before you move them to a production environment.
- You want to accelerate line-of-business analytics projects by creating a data mart service that combines a governed data source with analytic techniques.

- You musty deliver self-service analytics solutions and applications that use data that is generated from new sources and is imported directly into the private cloud warehouse.
- You want to migrate a subset of applications or data from an on-premises data warehouse to a private cloud.
- You want to save money or improve performance by migrating on-premises data marts or an on-premises data warehouse to a cloud-native data warehouse.
- You want to support data scientists who are coding, must store data locally, and need to use a logical representation.
- You want to reduce network traffic and improve analytic performance by storing your data near your Analytics Engine.
- You have multiple departments, and each department requires their own database management system.

After you create a Db2 Warehouse database, you can use the integrated database console to perform the following common tasks to manage and work with the database:

- Explore the database through its schemas, tables, views, and columns, which include viewing the privileges for these database objects.
- Monitor databases through key metrics, such as Availability, Responsiveness, Throughput, Resource usage, Contention, and Time Spent.
- Manage access to the objects in the database.
- Load data from flat files that are stored on various storage types.
- Run SQL and maintain scripts for reuse.

Initial setup and configuration considerations

Setting up the Db2 Warehouse service and data warehouse databases in Cloud Pak for Data requires some extra steps and considerations compared to some of the other Cloud Pak for Data services.

Before installing the Db2 Warehouse service, consider the use of dedicated worker nodes for the Db2 Warehouse database, which is important for data warehouse databases. Setting up dedicated nodes for your Db2 Warehouse database involves taint and toleration to provide node exclusivity.

If you plan to use an MPP configuration, you must designate specific network communication ports on the worker nodes, and ensure that these ports are not blocked. You also can improve performance in an MPP configuration be establishing an inter-pod communication network. Also, create a custom security context constraint (SCC) that is used during the installation.

After installing the Db2 Warehouse service and before creating your data warehouse database, consider disabling the default automatic setting of interprocess communication (IPC) kernel parameters so that you can set them kernel parameters manually. Also, consider enabling the `hostIPC` option for the cluster so that you can tune kernel parameters for the worker nodes in the cluster. Doing so allows you to use the Red Hat OpenShift Machine Config Operator to tune the worker IPC kernel parameters from the master nodes.

Now, you can create your data warehouse database in your Cloud Pak for Data cluster. You can choose to use the SMP or MNP architectures with the following configurations:

- Single physical node with one logical partition (default).
- Single physical node with multiple logical partitions.
- Multiple physical nodes with multiple logical partitions.

These configurations can be deployed on dedicated nodes by specifying the label for the dedicated nodes.

One of the last steps is to set the storage locations for your system data, user data, backup data, transactional logs, and temporary table space data. This data can be stored together in a single storage location, but it is advised to consider the use of separate locations, especially among the user data, transactional logs, and backup data.

After the data warehouse database is created, you can now start using the database by creating your first set of tables and loading data into the tables.

Learn more

For more information about the Db2 Warehouse service, see the following resources:

- IBM Documentation:
  - Db2 Warehouse on Cloud Pak for Data
  - Preparing to install the Db2 Warehouse service
  - Installing the Db2 Warehouse service
  - Postinstallation setup for the Db2 Warehouse service
- Chapter 4, “Multicloud data integration” on page 163.

2.2.5 Db2 Data Gate service

The IBM Db2 Data Gate service provides a gateway to synchronize data from Db2 for z/OS that is hosted on IBM Z® to any IBM Cloud Pak for Data environment. This gateway can extract, load, synchronize, and propagate your mission-critical data to a target database on Cloud Pak for Data for quick access to your high-volume, read-only transactional and analytic applications.

You can choose your target database based on your business needs. For example, you might set up Db2 as your target database for your new high-intensity transactional workloads. Or, you might set up Db2 Warehouse as your target database for your analytic or AI workloads. This service also provides one-click integration with IBM Watson Knowledge Catalog that simplifies incorporating Db2 Data Gate metadata within Cloud Pak for Data.

The Db2 Data Gate service uses an integrated data synchronization protocol to ensure that your data is current, consistent, and secure. The fully zIP-enabled synchronization protocol is lightweight, high throughput, and low latency. It enables near real-time access to your data without degrading the performance of your core transaction engine.

Accessing data in the Db2 Data Gate target database

After the Db2 Data Gate instance is provisioned in your Cloud Pak for Data cluster, you can now work with the associated Db2 or Db2 Warehouse database that is populated by Db2 Data Gate. When the Db2 Data Gate instance was defined the workload type (transactional or analytical) dictates the target database:

- Transactional workloads use IBM Db2 as the target database
- Analytical workloads use IBM Db2 Warehouse as the target database

Now, your transactional and analytical applications can use the data that is stored in the Cloud Pak for Data target database.

Figure 2-4 shows you the Db2 Data Gate architecture and components.

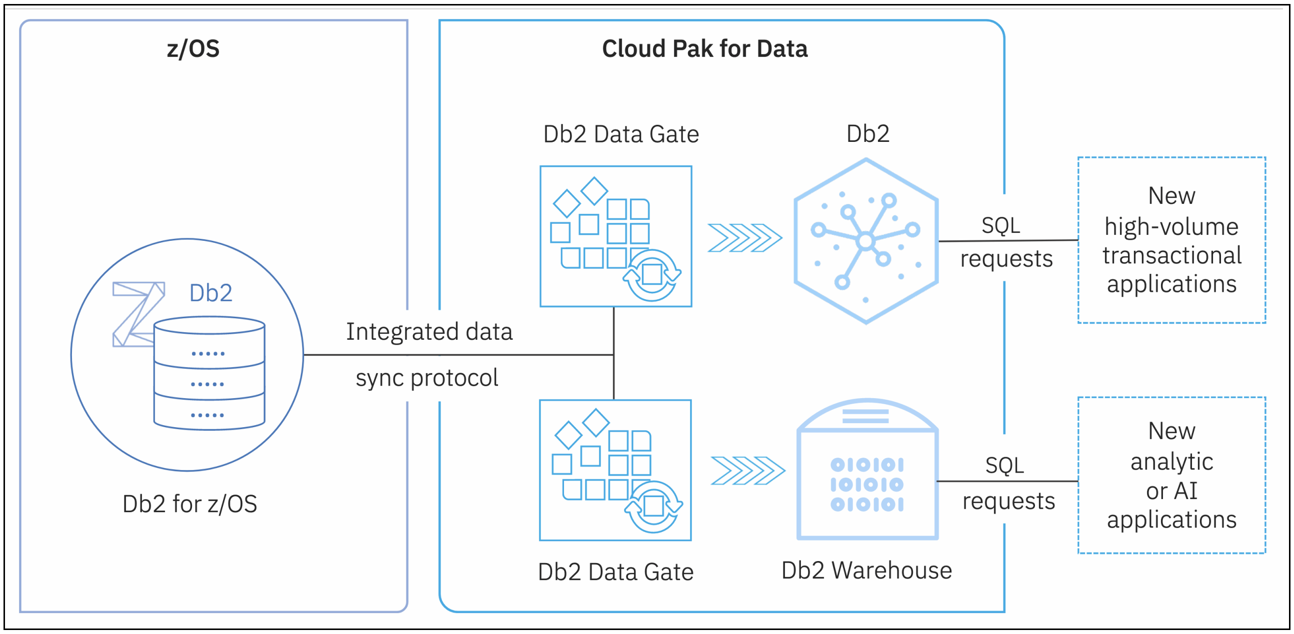

Figure 2-4 Db2 Data Gate architecture

Initial setup and configuration considerations

Installing and configuring the Db2 Data Gate service and instance requires extra steps compared to some of the other Cloud Pak for Data services and instances. These steps include configuring the Db2 for z/OS database on the z System and provisioning a Db2 database or Db2 Warehouse database.

The following high-level tasks are used to configure the z System (for more information, see this IBM Dpocumentation web page):

1. Configure inbound access for Db2 for z/OS.
2. Encrypt outbound network access from Db2 for z/OS to the Db2 Data Gate service.
3. Install the Db2 Data Gate back end on the IBM z Systems® System.
4. Configure Db2 for z/OS to support Db2 Data Gate.
5. Create and set Db2 Data Gate stored procedures.
6. Create Db2 Data Gate users and grant privileges on the z System.

After this process is complete, install, provision, and configure a Db2 instance of Db2 or Db2 Warehouse for the target database on Cloud Pak for Data. The instance that is used depends on your workload type:

- Transactional workloads use IBM Db2 as the target database
- Analytical workloads use IBM Db2 Warehouse as the target database

Learn more

For more information about the Db2 Data Gate service, see this IBM Documentation web page.

2.2.6 IBM Business Partner databases

IBM has a vast Business Partner Program, which includes numerous world leading database platform vendors for enterprise environments. The Cloud Pak for Data team partnered with several of these database vendors to expand the types of databases that you can deploy on Cloud Pak for Data.

The following database platforms support your enterprise deployments:

- EDB Postgres for a PostgreSQL database
- MongoDB for a NoSQL database
- IBM Informix® for a database that integrates TimeSeries, Spatial, NoSQL, and SQL data

The use of these databases extends the type of data that you can store and access on Cloud Pak for Data from structured to unstructured data, including semi-structured data.

EDB Postgres

EDB Postgres offers a secure, enterprise-class database that is based on open source PostgreSQL. This database platform enables you to use your data type of choice from structured data to semi-structured and unstructured data, such as JSON, geospatial, XML, and key-value.

The EDB Postgres service provides two versions with which you can provision and manage on Cloud Pak for Data: EDB Postgres Standard and EDB Postgres Enterprise. Both services provide an enhanced version of the open source PostgreSQL.

Use the EDB Postgres Standard service to provision and manage a PostgreSQL database. For your production environments, use EDB Postgres Enterprise to access all of the capabilities of EDB Postgres Standard, but with security and performance features for enterprises.

EDB Postgres Enterprise enables users to deploy EDB Postgres Advanced Server databases with the following enhancements:

- Performance diagnostics
- Enterprise security
- Oracle database compatibility
- Enhanced productivity capabilities for DBAs and developers

MongoDB

MongoDB offers a NoSQL database management platform that specializes in documented oriented data. This database platform provides an engine for storing and processing JSON object data without having to use SQL.

The MongoDB service provides the MongoDB Enterprise Advanced edition, which is a highly performant, highly available database with automatic scaling in your Cloud Pak for Data cluster so that you can govern the data and use it for in-depth analysis.

Integrating a MongoDB database into Cloud Pak for Data can be useful in the following situations:

- You need an operational database that supports a rapidly changing data model.
- You want lightweight, low-latency analytics integrated into your operational database.
- You need real-time views of your business, even if your data is in silos.

- You develop applications and need a database that can:
  - Store large amounts of data with different data types, such as structure, unstructured, and polymorphic data
  - Support millions of users
  - Personalize the content that you deliver to customers
- You need to store large amounts of data from Internet of Things devices or sensors.
- You need to maintain a catalog.
- You need to store and serve many different types of content.

Informix

Informix offers high-performance database engine for integrating SQL, NoSQL, JSON, time-series, and spatial data, with easy access by way of MQTT, REST, and MongoDB APIs. The Informix service provides an Informix database on Cloud Pak for Data so you can use the rich features of an on-premises Informix deployment without the cost, complexity, and risk of managing your own infrastructure.

Integrating an Informix database into Cloud Pak for Data can be useful in the following situations:

- You need an operational database that supports a rapidly changing data model.
- You want lightweight, low-latency analytics integrated into your operational database.
- You need to store large amounts of data from Internet of Things devices or sensors.
- You need to store and serve many different types of content.

Learn more

For more information about the Business Partner Database services, see the following resources:

- EDB Postgres:
  - EDB Postgres on Cloud Pak for Data
  - EnterpriseDB.com
- MongoDB:
  - MongoDB on Cloud Pak for Data
  - Mongodb.com
- Informix

2.2.7 Virtual Data Pipeline platform

IBM Virtual Data Pipeline (VDP) is a highly scalable, test data management platform that virtualizes databases to improve the resiliency, agility, and cloud mobility of your business. IBM VDP enables you to capture data from production systems, manage it in the most efficient way possible, and use virtual or physical copies of the data whenever and wherever they are needed.

The IBM VDP solution is deployed outside of a Cloud Pak for Data cluster to capture point in time copies of production databases and provide virtual clones for Cloud Pak for Data and Data Virtualization.

For more information about this external offering, see *Getting Started with Virtual Data Pipeline Copy Data Management*.

Initial setup and configuration considerations

The IBM VDP solution includes two components that are delivered as software appliances:

- VDP appliance, which provides virtual clones of databases.
- InfoSphere Virtual Data Pipeline - Global Manager appliance, which is a management console for one or more VDP appliances.

For more information and steps to deploy an IBM VDP application, see the following resources:

- Installation: This IBM Support web page
- Deployment: *Support Matrix: IBM InfoSphere Virtual Data Pipeline 8.1.1.3*

2.3 Governance services

This section describes the Governance services that are available in IBM Cloud Pak for Data:

- IBM Watson Knowledge Catalog
- IBM Data Privacy
- IBM Product Master

2.3.1 IBM Watson Knowledge Catalog

IBM Watson Knowledge Catalog service provides the following capabilities:

- Data governance
- Metadata management
- Data cataloging
- Data profiling and classification
- Data quality
- Data lineage
- Data masking and access controls
- Self-service data consumption

The service allows you to catalog, categorize, classify, and curate data. You also can set up and manage a trusted and governed data foundation and enable data democratization within your enterprise.

The governance and privacy capabilities of IBM Watson Catalog are a key part of a modern data fabric and are central to IBM's data fabric approach.

Figure 2-5 on page 60 shows a profiled data asset in a catalog.

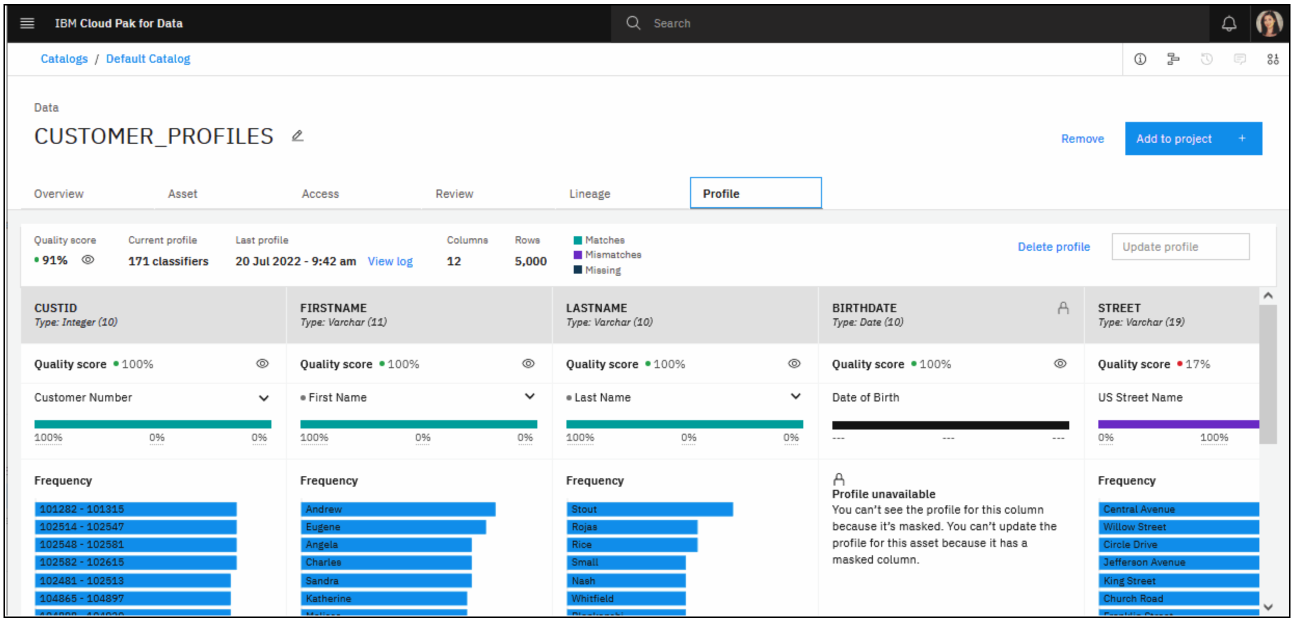

Figure 2-5 Cataloged data asset: Profile tab view

By using IBM Watson Knowledge Catalog, you can perform the following tasks:

- Create a common centralized repository of governance artifacts that are relevant to your organization; that is, business glossary terms, data classes (types), data classifications, reference data sets, policies, governance rules, and data protection (masking) rules. You also can organize and control access to those artifacts by using Categories.
- Take advantage of accelerator content to speed up governance framework setup and adoption. The service includes:
  - Over 165 pre-built Data Classes (data type classifiers that can be used for automated data profiling and classification, and automated data quality analysis).
  - Four pre-built Classifications for designating levels of sensitivity of data (Sensitive Personal Information, Personally Identifiable Information, Personal Information and Confidential).
  - An extensive set of industry-specific content packs that contain key industry terms and reference data sets (Knowledge Accelerator content). These assets can be reused and adapted to your needs, and bespoke governance artifacts can be set up as and where needed.
- Enforce change control by using predefined workflows to manage the process of creating, updating, and deleting governance artifacts, or create your own custom workflows.
- Discover, categorize, classify, tag, catalog, and curate data, while leaving the data where it is. IBM Watson Knowledge Catalog includes automated ML-powered Profiling and Metadata Import and Enrichment capabilities and allows you to import metadata from your source systems and organize your connections and data assets into one or more catalogs, all without physically moving or copying your actual data. That metadata can be enriched further by using your own governance foundation setup, which allows you to put business context to your data sets.
- Establish relevant masking and access controls and rules to satisfy your organization's internal privacy guidelines and mask data dynamically and consistently at the required level of granularity.

- Analyze and understand data quality by using pre-built data quality dimensions, intelligent automated data quality analysis engine, a repository of reusable quality rules, and the ability to build your own bespoke data quality rules.
- Use model inventory and AI FactSheets capabilities to track the model lifecycle of machine learning models that are developed by your organization, from training to production, and facilitate efficient ModelOps governance.
- Browse, preview, and self-serve data, governance artifacts, and other metadata by using the in-built semantic search, asset preview, and governed data asset access control capabilities.
- Automatically synchronize catalog assets and governance artifacts with select external repositories by using the Open Data Platform initiative (ODPi) Egeria connector.
- Map, understand, and automatically generate lineage information (an add-on [MANTA Automated Data Lineage] is required to generate lineage information).
- Set up an external reporting data mart and generate reports to get insights about metadata that is held in IBM Watson Knowledge Catalog.

Working with the service

The service has two deployment options: core (which installs a subset of service capabilities) and base (full set of current capabilities is installed). It uses the optional installation concept for relevant extra features, which at the time of this writing include:

- Knowledge Graph (enables Semantic Search and NextGen lineage features)
- Data Quality (enables NextGen Data Quality features in Analytics Projects)
- Factsheets (enables Model Inventory features)
- Advanced Metadata Import (enables NextGen Advanced Metadata Import features)

> **Note:** An IBM Db2 or a PostgreSQL instance (external to Cloud Pak for Data, or provisioned as a Cloud Pak for Data service) is required for the Reporting capability setup.

The Data Refinery service, as described Chapter 7, “Business analytics” on page 483, is an included feature that is installed automatically when IBM Watson Knowledge Catalog is deployed.

Knowledge Accelerator content packs are optionally imported by way of API calls postinstallation, if required.

The following other roles are created or modified upon service installation:

- Business Analyst
- Data Engineer
- Data Quality Analyst
- Data Scientist
- Data Steward
- Developer
- Reporting Administrator

To set up the service, the following key tasks often are performed:

- Users and user groups setup, relevant Platform roles assignment.
- Categories and categories hierarchy setup. Allocation of Admin, Owner, Editor, Reviewer or Viewer roles to relevant users and user groups for each category.
- Governance artifact setup, including relevant category assignment. Governance workflows set up.

- Catalogs set-up. Allocation of Viewer, Editor, or Admin roles to relevant users and user groups for each catalog.
- Connections set up: Platform connections, and connections local to individual catalogs or projects, where relevant.
- Metadata import, metadata enrichment, quality and lineage analysis, followed by catalog publish.
- More Data Quality setup and analyses.
- More setup where relevant (reporting, Model Inventory, and Knowledge Accelerator content import).

Learn more

For more information about the IBM Watson Knowledge Catalog service, see the following resources:

- IBM Documentation:
  - Watson Knowledge Catalog on Cloud Pak for Data
  - Preparing to Install Watson Knowledge Catalog
  - Installing Watson Knowledge Catalog
  - Postinstallation setup for Watson Knowledge Catalog
  - Governance and Catalogs (Watson Knowledge Catalog)
  - Managing workflows
  - IBM Knowledge Accelerators
  - Governance
- Tutorial: Trust Your Data, Protect Your Data and Know Your Data Tutorials are available for IBM Watson Knowledge Catalog on Cloud (SaaS version of the service)
- APIs
- Chapter 3, “Data governance and privacy” on page 115.

2.3.2 IBM Data Privacy

IBM Data Privacy service allows you to create physically masked copies or subsets of your structured (relational) data and semi-structured data (CSV, Avro, partitioned data, or Parquet files).

The service adds masking flow capabilities for creating format-preserving anonymized data sets for training, test, and other purposes.

Jobs that are running the masking flows can be scheduled, and their outputs can be saved as a file within Cloud Pak for Data, or saved into a data repository of your choosing.

Figure 2-6 shows the job setup window.

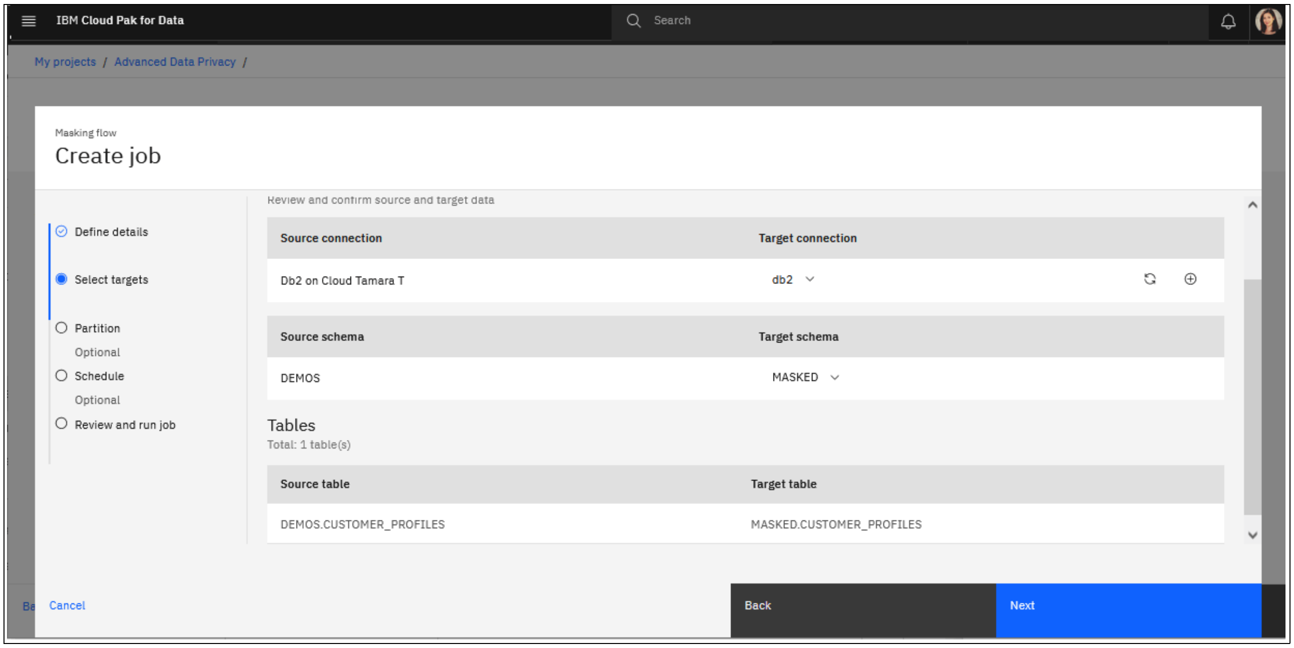

Figure 2-6 Data Privacy service - setting up a job running a masking flow

With the release of Cloud Pak for Data version 4.5, advanced data masking capabilities, which were part of the Data Privacy service, were rolled into Watson Knowledge Catalog core capability set instead. Therefore, as of version 4.5, installation and use of the IBM Data Privacy service is needed only if the ability to produce physically masked copies of data is required in addition to dynamic at-view level advanced masking of data assets.

Working with the service

To use IBM Data Privacy, the prerequisite IBM Watson Knowledge Catalog service and IBM Analytics Engine Powered by Apache Spark service must be installed on the cluster. Consider the following points:

- IBM Data Privacy is an add-on to IBM Cloud Pak for Data's Watson Knowledge Catalog Service. The add-on relies on and uses data masking rules (Data Protection Rules) that are defined as part of Watson Knowledge Catalog setup.
- IBM Data Privacy's ability to run masking jobs is underpinned by the IBM Analytics Engine Powered by Apache Spark service capabilities.

No service instance provisioning is required after installation. The IBM Data Privacy service is deployed as a single shared instance per cluster.

Data administrators, data engineers, and data stewards are collectively responsible for the prerequisite Cloud Pak for Data platform and Watson Knowledge Catalog setup that must be in place before masking flows can be created, including the following examples:

- Data protection rules design and setup;
- Cataloging, enrichment and preparation of data for masking
- Configuration of user permissions and masked data access

After these prerequisites are in place, any relevant users of data (for example, data scientists, business analysts, testers, and developers) and any of the personas that were involved in setting and administering the prerequisites can participate in masking flow creation.

However, this access is subject to the following conditions being met for each user:

- Access catalogs permission added as part of the user's security scope.
- Ability to create projects or assignments of an Administrator or Editor role within a project.
- The user can access the source data asset in a catalog.
- If the resulting masked copy must be copied to a data source (rather than as a file within the project), the user must be allowed to write to the target data source (for example, have write access to it).

Learn more

For more information about the Data Privacy service, see the following resources:

- IBM Documentation:
  - Data Privacy on Cloud Pak for Data
  - Installing Data Privacy
  - Masking Data with Data Privacy
  - QuickStart: Project Data
  - Data protection rules (Watson Knowledge Catalog)
- Data Privacy and Security smart paper: *Data Leaders: Turn compliance into competitive advantage*
- Data governance and privacy tutorial: Protect you data
- Chapter 3, "Data governance and privacy" on page 115.

2.3.3 IBM Product Master

The IBM Product Master service provides Master Data Management (MDM) capabilities for your product and service-related information.

This Product Information Management (PIM) solution acts as middleware that helps enterprises establish a scalable, flexible, integrated, and centralized repository of up-to-date, 360° views of product and services information.

The provided PIM capabilities enable aggregation of data from upstream systems, enforcement of business processes to ensure data accuracy and consistency, and synchronization of the resultant trusted information with your downstream systems.

Figure 2-7 shows a sample product quality dashboard that is provided with the Product Master service.

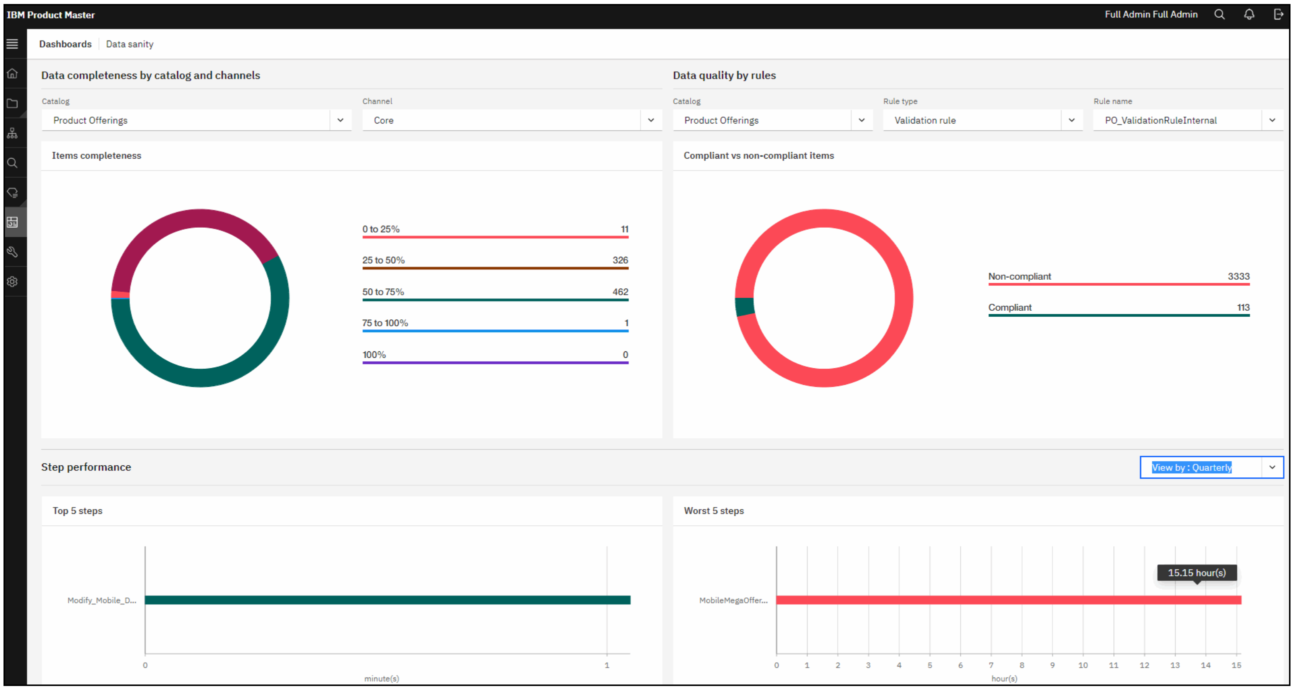

Figure 2-7 Product Master service - the data completeness by catalog and channel dashboard

Working with the service

IBM Product Master uses Db2 (IBM Db2 Version 11.5 is supported and recommended at the time of this writing), or Oracle Database (Version 19c at the time of this writing). A separate database of either type serves as the back-end data repository and must be created for the Product Master.

> **Note:** The prerequisite database services and instances are not installed with Product Master and must be installed separately.

The service is instance-based; that is, following service installation, one or more instances of Product Master must be provisioned, and relevant users assigned service-specific roles within each instance. IBM Product Master provides a persona-based UI and ships with 11 standard role templates:

- Administrator, IT, and system personas: Admin, Full Admin, Service Account, and Solution Developer.
- Business personas: Basic, Catalog Manager, Category Manager, Content Editor, Digital Asset Management, Merchandise Manager, and Vendor.

With IBM Product Master, you can perform the following tasks:

- Capture, create, link, and manage product, location, trading partner, organization, and terms-of-trade information by using the provided tools.
- Use pre-built data models to begin working quickly. Adapt your data model as needed.
- Use Digital Asset Management capabilities to manage product images, videos, brochures, and other unstructured data.

- Set up business user workflows to support multi-departmental and multi-enterprise processes.
- Customize user interfaces.
- Collaborate internally and with external parties (for example, suppliers and partners).

Figure 2-8 shows the product editing experience in the Product Master service.

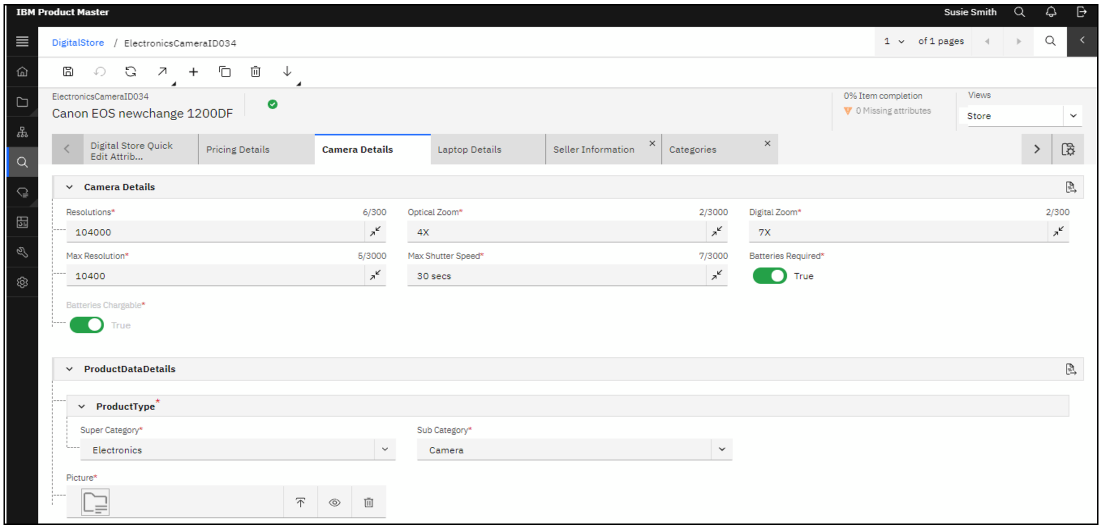

Figure 2-8 Product Master service: Editing an entry in a product catalog

Learn more

For more information about the Product Master service, see the following resources:

- IBM Documentation:
  - Product Master on Cloud Pak for Data
  - Preparing to install the Product Master service
  - Installing Product Master
  - Postinstallation setup for Product Master service
  - Managing master data by using Product Master
- Product site
- White Paper: *IBM Product Master for Product Information Management: Achieve better operational efficiency, manage compliance and drive data-based Digital Transformation*

2.4 Analytics Services

This section highlights the Analytics Services that you use on Cloud Pak for Data.

2.4.1 IBM DataStage

IBM DataStage is an Extract-Transform-Load (ETL) service that enables design, scheduling, and execution of data flows that transform and move data.

Although data engineers, data integration specialists, and IT teams often are the main users of this service, technical and nontechnical users can participate in ETL pipelines design and execution in practice because no coding is required for flow build, job scheduling, scaling, or execution.

The service provides a graphical drag-and-drop flow designer interface with which users can easily compose transformation flows from various pre-built standard data source connector and processing stage objects.

Flow design and flow execution are logically separated out.

After the flows are built, they are compiled and a job instance that runs the steps of the flow is then created and run on the in-built parallel processing and execution engine that enables almost unlimited scalability, performant workload execution, built-in automatic workload balancing and elastic scaling. The jobs can be run on-demand and scheduled.

This design paradigm accelerates development, increases productivity, promotes reusability, and enables a consistent, standardized way of creating integration workloads to satisfy data integration tasks of any complexity.

Figure 2-9 shows the DataStage tile that you use to add the flow asset to a Project.

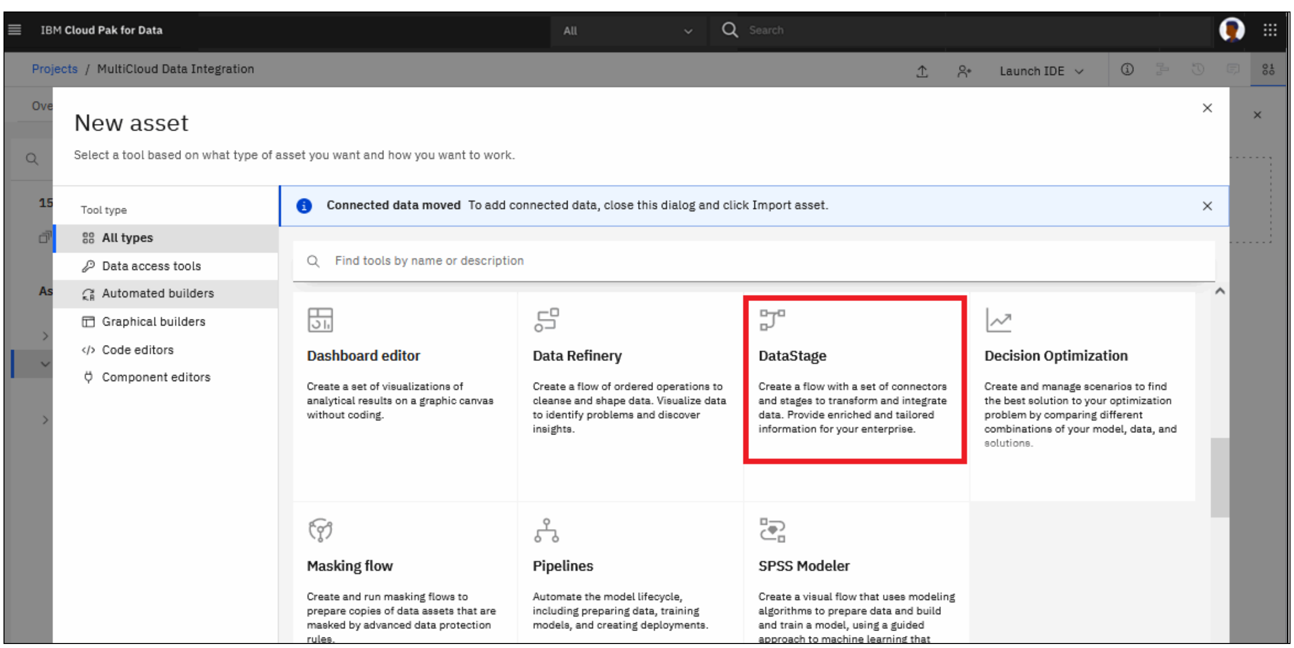

Figure 2-9 Creating an IBM DataStage flow

The service supports batch, real-time, and streaming scenarios. It also provides built-in search, automatic metadata propagation, and simultaneous highlighting of all compilation errors. Different job runtime sizes and specifications can be chosen to suit your job complexity and data processing volumes.

Working with the service

You can choose from two versions of the service: IBM DataStage Enterprise and IBM DataStage Enterprise Plus.

IBM DataStage Enterprise provides all the described baseline capabilities.

The IBM DataStage Enterprise Plus version includes more data quality stages and capabilities. It adds transformation stages for data cleansing (by identifying potential anomalies and metadata discrepancies) and duplicates identification and handling (by using data matching and probabilistic matching of data entities between two data sets).

IBM DataStage Enterprise Plus also is the prerequisite service version for address verification interface (AVI) use cases and scenarios. It includes baseline functions that enable AVI use (the Address Verification stage). To use AVI, a separate purchase of relevant AVI Reference Data packs is required.

Jobs that are built by using the DataStage service are parallel jobs. The Watson Studio Pipelines service is required to build and run sequence jobs that allow you to link and daisy-chain multiple parallel job executions, and incorporate branching, looping, and other programming controls.

IBM DataStage flow and job design and scheduling are performed within projects. After the relevant service version is installed on your Cloud Pak for Data cluster, DataStage functions and tools are available for use within Analytical Project workspaces.

Different Projects can be used to organize, group, and control access to different transformation and ETL activities and initiatives.

Transformation flow design typically starts with capturing and defining of the wanted source and target data sources, files, and applications. Those components must be defined first within the relevant Project.

After the required connections and data assets are defined, DataStage Flow Designer canvas can be started by adding a New Asset of type DataStage.

Figure 2-10 shows the DataStage Flow Designer canvas.

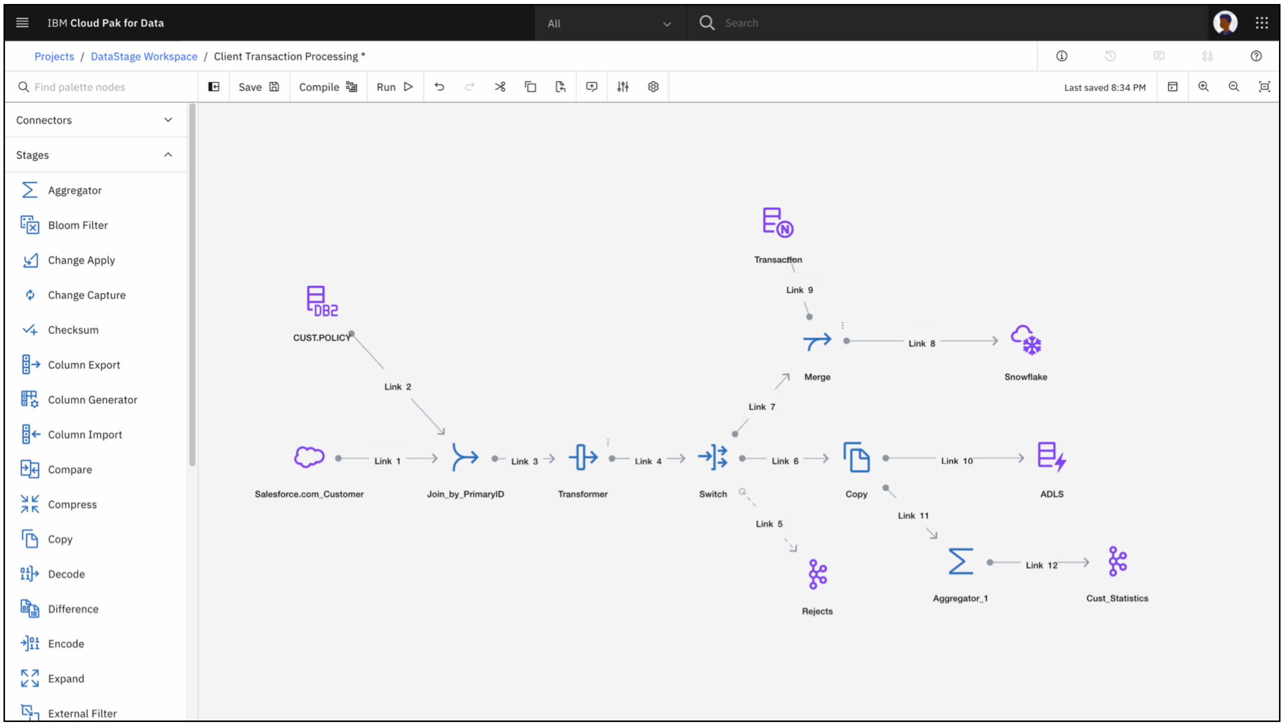

Figure 2-10 DataStage Flow Designer canvas

The flow is then designed by selecting, dragging and dropping, and connecting and arranging relevant connectors and stages on the canvas. Then, relevant properties can be added and edited for the connectors and stages of the flow.

After the wanted flow design is completed, the flow can be saved and compiled. The resulting logs and status messages flag errors if any are present at this stage. The errors must be rectified before a job that runs the flow can be run.

The jobs can be created and run on-demand and scheduled, and different run times can be chosen for different jobs. Each job run generates a log that contains execution status, warnings, and errors.

In addition to DataStage flows and jobs, the service allows you to create DataStage components that can further be reused across different flows (subflows, data definitions, standardization rules, or schema library components), and Parameter Sets that capture multiple job parameters with specified values for reuse in jobs.

Learn more

For more information about the DataStage and Watson Studio Pipeline services, see the following resources:

- IBM Documentation:
  - DataStage on Cloud Pak for Data
  - Installing DataStage
  - Transforming data (DataStage)
  - Watson Studio Pipelines
  - Orchestrating flows with Watson Pipelines

- Data fabric tutorals
- Chapter 3, “Data governance and privacy” on page 115.

2.4.2 IBM Data Refinery

IBM Data Refinery is a graphical, interactive data exploration, data preparation, and data wrangling service that can be used to easily and quickly shape and cleanse your data.

The most common and frequently used cleansing and shaping operations are included in the tool, with which you can easily fix or remove incorrect, incomplete, or improperly formatted data. You also can join multiple data sets, remove duplicates, filter data, sort data, combine or remove columns, and create derived columns, all with no coding required.

The service also helps teams gain more insight into their data, which can serve as a precursory data exploration steps for data science model development, dashboarding, and other analytical tasks.

Figure 2-11 shows the Data Refinery service.

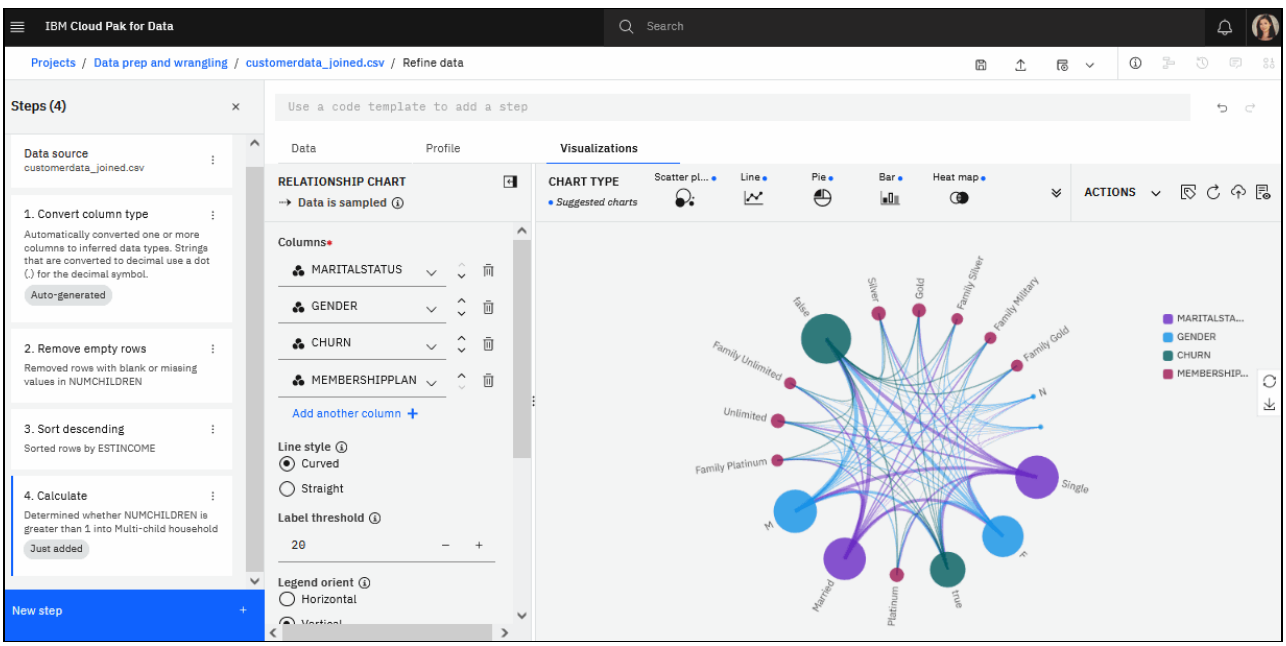

Figure 2-11 Data Refinery service

Data Refinery brings the following data transformation and discovery features and capabilities Into the hands of line-of-business teams personas that are involved in business intelligence, data analytics, and data science work:

- Dedicated sandbox environments for data set exploration, discovery, visualization, and modeling of your wanted data shaping and cleansing activities. You can safely experiment with transformations and dynamically explore data without affecting data at the source.
- A rich set of features to help understand your original data set, dynamically validate your data preparation steps in real time with built-in data set previews, data set profiles with data classification and statistical, frequency, and distribution analyses, and a data visualization interface with over 20 customizable charts.

- Build your own data transformation recipes by using over 100 pre-built operations. The design transformation steps sequence can be saved as a data shaping and cleansing flow for reuse.
- Another option is available to use R functions and code to complement the pre-built UI-based operations, if needed.
- Create, run on-demand or schedule processing jobs that are based on your saved transformation flows and choose where and how to land the results. The jobs process the entire source data set and can land the transformed outcome as a local file within the project where the flow is, or write it as a file or a table to your target data source of choice.
- Choose and use different runtime sizes for different jobs to best suit the size and complexity of the source data set.

Working with the service

To use the service, IBM Watson Studio service or IBM Watson Knowledge Catalog service must be deployed on your Cloud Pak for Data cluster. The Data Refinery service is a feature that is bundled with these services and does not require a separate installation.

Data Refinery jobs can use various runtime configurations. Default Data Refinery XS is the default in-built run time. To use more Spark and R environments and templates, or to run Data Refinery jobs directly on a Hadoop cluster, the Analytics Engine Powered by Apache Spark service of Cloud Pak for Data also must be deployed on your cluster.

Data Refinery works with tabular data formats only (tables and files). To work with a data asset in Data Refinery, the asset must be added to a Project first. The Refinery capabilities are then started by clicking **Refine** at the upper right of the asset's Preview tab.

Then, Cloud Pak for Data creates and starts a Data Refinery sandbox instance that can be used for data exploration and visualization and data transformation flow build.

Figure 2-12 shows the Data Refinery flow design canvas and a selection of the pre-built data preparation operations that the service provides.

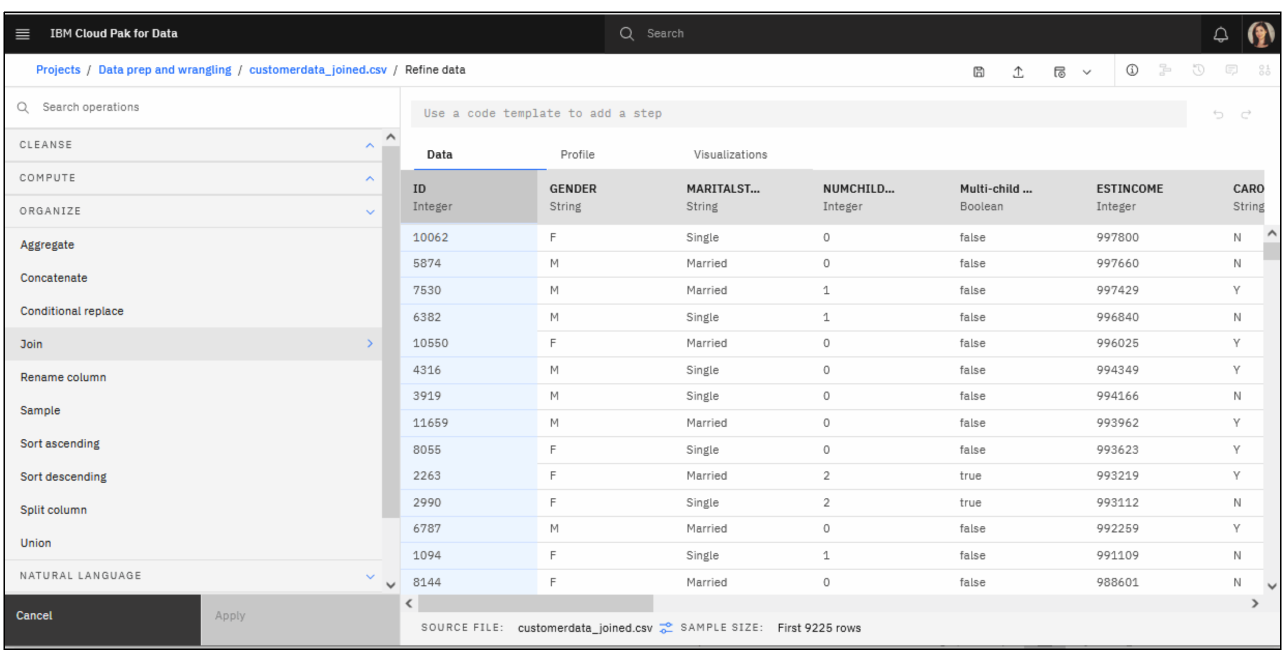

Figure 2-12 Data transformation design with Data Refinery service

After the wanted sequence of transformations steps is created, the flow is saved and can be used and reused to generate flows. It also can be used to schedule and run Data Refinery jobs that run the transformation steps on the source data set and create the wanted resulting target data set.

Learn more

For more information about Data Refinery, see the following resources:

- IBM Documentation:
  - Data Refinery on Cloud Pak for Data
  - Data Refinery environments (Watson Studio and Watson Knowledge Catalog)
  - Refining data (Data Refinery)
- Tutorial: Quick start: Refine data
- Chapter 3, "Data governance and privacy" on page 115.

2.4.3 SPSS Modeler

SPSS Modeler allows data scientists to create flows that can:

- Discover, prepare, and blend data
- Develop and manage models
- Visualize the results

They can do all of this without any coding required. SPSS Modeler flows build machine learning pipelines that you can use to iterate rapidly during the model building process.

With SPSS Modeler, you can build predictive models to improve decision making based on business expertise. SPSS Modeler offers various modeling methods that are taken from machine learning, artificial intelligence, and statistics.

The methods that are available on the node palette allow you to create predictive models. Each method has its strengths and is suited for particular types of problems.

SPSS Modeler integrates with Watson Studio. In an analytics project, you create a SPSS Modeler flow, which is a visual flow that uses modeling algorithms to prepare data and build and train a model.

Working with the service

The following overall process is used to develop a flow:

1. Add data to your project to use as a source or target node in a flow.
2. Open the node palette and drag nodes onto the canvas. Different nodes have different functions. For example, to import data to be used in a flow, you use the Import node. If you wanted to filter out some unwanted data, use the Filter node.
3. Create a flow of different nodes to represent how you want to develop your model.
4. You can preview the output at the end of each node.
5. Run the flow to generate the output.
6. Save the flow.

Figure 2-13 shows the canvas with the nodes and the flows.

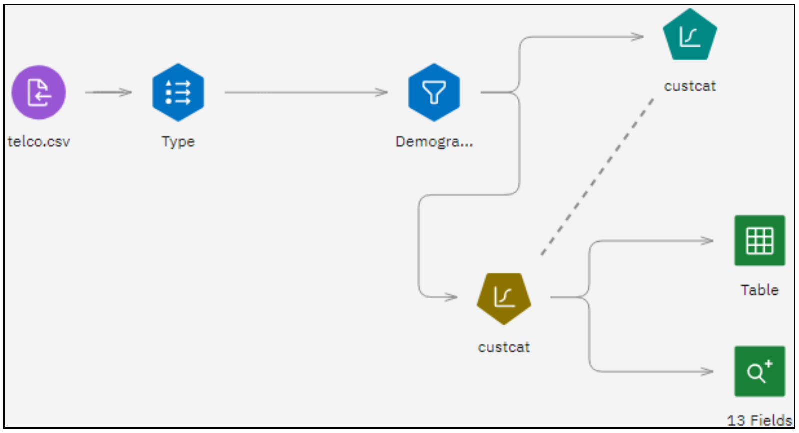

Figure 2-13 Canvas with the nodes and the flows

Learn more

For more information about the SPSS Modeler service, see the following resources:

- IBM Documentation:
  - SPSS Modeler on Cloud Pak for Data
  - Installing SPSS Modeler
- Tutorial: Download and import the example projects

2.4.4 Decision Optimization

Prescriptive analytics uses data to determine an optimal course of action. The IBM Decision Optimization service provides CPLEX® optimization engines that enable you to make optimal business decisions by evaluating millions of possibilities to find the most suitable prescriptive solutions.

Decision Optimization integrates with Watson Studio and Watson Machine Learning. Before you install Decision Optimization, be sure both of those services are installed and running.

With Decision Optimization, you can:

- Prepare data.
- Import or create Decision Optimization models in Python, OPL, or Natural Language (by using Modeling Assistant).
- Solve models and compare multiple scenarios.
- Visualize data, solutions, and produce reports.
- Save models to deploy with Watson Machine Learning.

You can build models bu using notebooks or the Decision Optimization experiment UI. You can add the Decision Optimization UI within your Watson Studio analytics project.

In the UI, you can create or edit models in various languages. Then, you can deploy the models to Watson Machine Learning deployment spaces.

Figure 2-14 shows the overview process of the use of Decision Optimization.

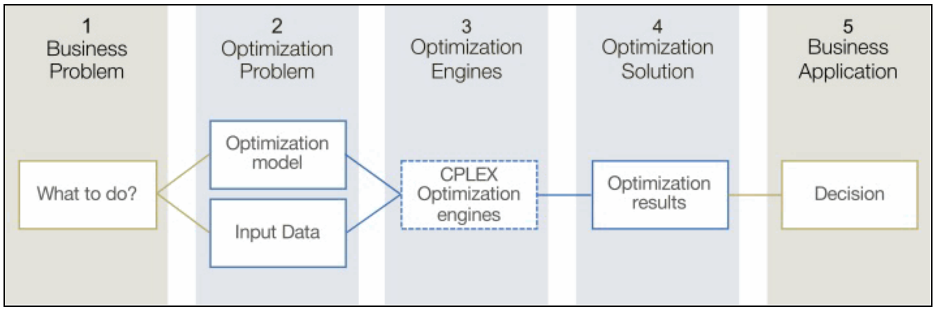

Figure 2-14 Decision Optimization

Optimization refers to finding the most suitable solution to a precisely defined situation. Typically, this process includes the following steps:

1. Define the business problem, such as planning, scheduling, pricing, inventory, or resource management.
2. Create the optimization model with the suitable input data. The model specifies the relationship among the objectives, constraints, limitations, and choices that are involved in the decisions. Combined with the input data, the optimization model represents an instance of the optimization problem.
3. Optimization engines, or the solvers, apply mathematical algorithms to find a solution within the constraints and limitations of the problem.

   The solutions are the values for all the decisions that are represented in the model.
4. The objective and solution values are summarized in a tabular or graphical view.

Learn more

For more information about the Decision Optimization service, see the following resources:

- IBM Documentation:
  - Decision Optimization on Cloud Pak for Data
  - Installing Decision Optimization
- Tutorials:
  - Solving a model using the Modeling Assistant
  - Solving a Python DOcplex model
  - Decision Optimization notebooks

2.4.5 Analytics Engine powered by Apache Spark

The Analytics Engine powered by Apache Spark is a compute engine for running analytical and machine learning jobs. If you have Watson Studio installed, you can specify Spark as the environment run time for your analytics project. You also can create custom Spark templates in a project. An administrator must first install and provision an instance of this service before it can be used.

Whenever you submit a job, a Spark cluster is created for the job. You can specify the size of the driver and executor, and number of executors for the job. These specifications enable you to achieve predictable and consistent performance.

When a job completes, the cluster is automatically cleaned up so that the resources are available for other jobs. An interface also is available for you to analyze the job's performance and to debug problems. You can run jobs through an analytics project or directly by using the Spark job API.

Spark environments

If your environment has Watson Studio installed, you can select the Spark environment when creating your notebook. After you select the environment, you can run your notebook by using the Spark run time.

Every notebook that is associated with the environment has its own dedicated Spark cluster and no resources are shared. If you have two notebooks by using the same Spark environment template, two Spark clusters are started with its own Spark driver and set of Spark executors.

Running Spark application job

You can time that you submit Spark jobs that run data transformation, data science, and machine learning model jobs by using the Spark job API.

The following process is used to submit a Spark job using the API:

1. Locate the Spark service instance.
2. Under Access information, copy the Spark jobs endpoint.
3. Generate an access token.
4. Submit the job by using the endpoint and the access token.

The minimal required parameters are shown in Example 2-1.

Example 2-1 Minimal required parameters

```
curl -k -X POST <V3_JOBS_API_ENDPOINT> -H "Authorization: Bearer <ACCESS_TOKEN>"
-d '{
    "application_details": {
        "application": "/opt/ibm/spark/examples/src/main/python/wordcount.py",
        "application_arguments":
["/opt/ibm/spark/examples/src/main/resources/people.txt"]
        }
}'
```

After the job is submitted, you can view its status by listing all the active jobs, as shown in Example 2-2.

Example 2-2 Viewing job status

```
curl -k -X GET <V3_JOBS_API_ENDPOINT> -H "Authorization: Bearer <ACCESS_TOKEN>"
```

Then, get the status of the job:

```
curl -k -X GET <V3_JOBS_API_ENDPOINT>/<job_id> -H "Authorization: Bearer
<ACCESS_TOKEN>"
Here is an example response:
{
      "application_id": "<application_id>",
      "state": "RUNNING",
      "start_time": "Monday' 07 June 2021 '14:46:23.237+0000",
      "spark_application_id": "app-20210607144623-0000"
  }
```

Depending on how you configured the service, advanced features are available that support application development and monitoring. These advanced features must be enabled before the Spark instance is created on the cluster.

To enable advanced features, as the Red Hat OpenShift administrator, you must run a patch command to update the `AnalyticsEngine` kind in Red Hat OpenShift project where the service was installed. Advanced features provide you with an interface in which you can manage your applications or monitor jobs.

Learn more

For more information about the Analytics Engine powered by Apache Spark, see the following resources:

- IBM Documentation:
  - Using advanced features in Analytics Engine Powered by Apache Spark
  - Analytics Engine Powered by Apache Spark on Cloud Pak for Data
  - Installing Analytics Engine Powered by Apache Spark
  - Postinstallation setup for Analytics Engine Powered by Apache Spark
- Chapter 7, “Business analytics” on page 483.

2.4.6 Execution engine for Apache Hadoop

This service integrates the Watson Studio service with the remote Apache Hadoop cluster. Data scientists can use this service for the following tasks:

- Browse remote Hadoop data through connections.
- Cleanse and shape remote Hadoop data with Data Refinery.
- Run Data Refinery jobs on the Hadoop Spark cluster.
- Run a notebook session on the remote Hadoop system.
- Access Hadoop systems with basic utilities from RStudio and Jupyter notebooks.

This service generates a secure URL for each Watson Studio cluster that is integrated with the remote Hadoop cluster. You can build and train models on the Hadoop cluster.

If you have data in a Hive or HDFS storage system on a Hadoop cluster, you can work with that data directly on the cluster.

Within a Watson Studio analytics project, you can find Hadoop environment templates on the Environments page. You can use the Hadoop environment in the following ways:

- Train a model on the Hadoop cluster by selecting a Hadoop environment in a Jupyter notebook
- Manage a model on the Hadoop cluster by running Hadoop integration utility methods within a Jupyter notebook
- Run Data Refinery jobs on the Hadoop cluster by selecting a Hadoop environment for the Data Refinery job

Figure 2-15 shows how data scientists work with an analytics project can train a notebook on a Hadoop cluster with data on the Hadoop cluster.

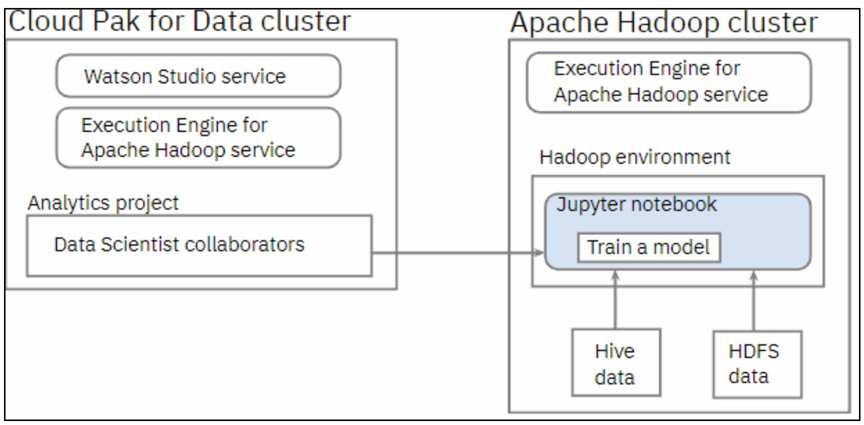

Figure 2-15 Watson Studio analytics project

Learn more

For more information about the Execution Engine for Apache Hadoop, see the following IBM Documentation web pages:

- Execution Engine for Apache Hadoop on Cloud Pak for Data
- Installing Execution Engine for Apache Hadoop
- Machine learning models on a remote Apache Hadoop cluster in Jupyter Python

2.4.7 Cognos Analytics

The Cognos Analytics service for Cloud Pak for Data provides an AI-infused Business Intelligence and Analytics platform. This platform integrates dashboards, stories, explorations, modeling, and reporting so that you can understand your organization’s data, and make effective business decisions.

You can use dashboards and stories to communicate your insights and analysis through a view that contains visualizations, such as a graph, chart, plot, table, and map. The Dashboard component lets you analyze powerful visualizations of your data and discover patterns and relationships that impact your business.

Figure 2-16 shows an example of a Dashboard.

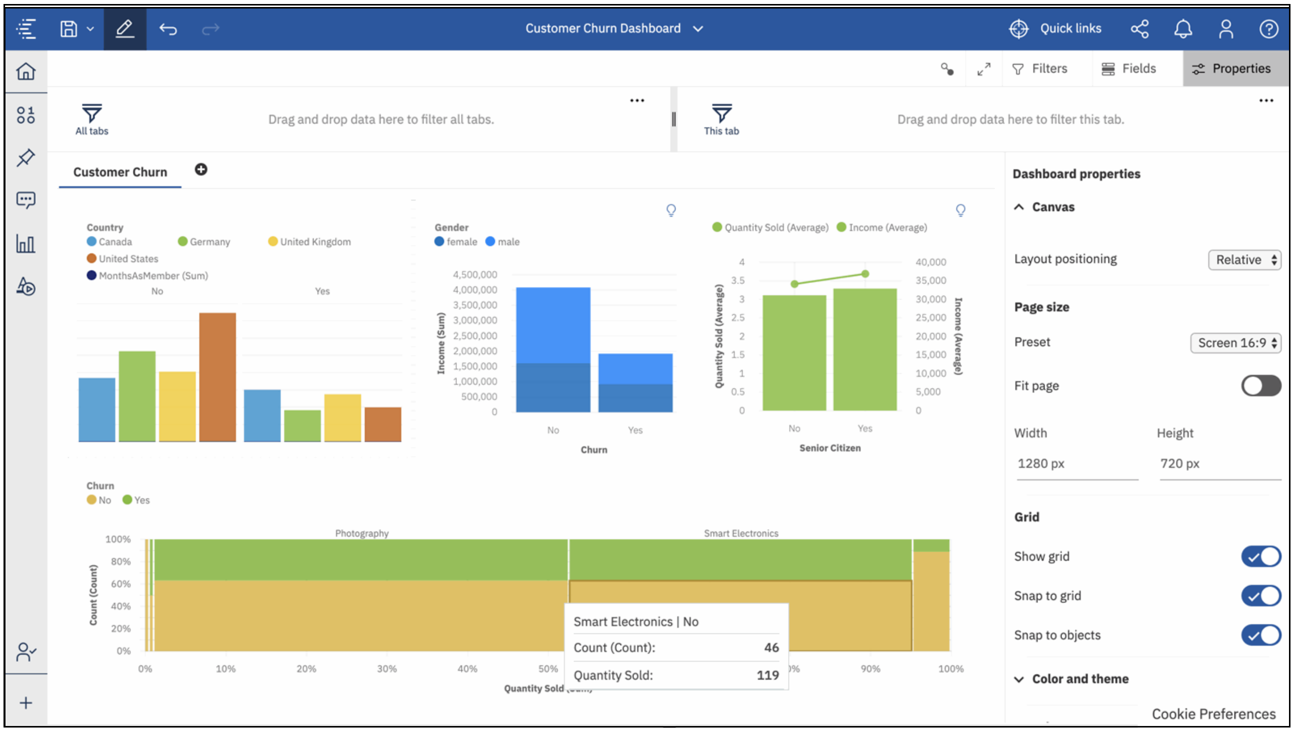

Figure 2-16 Cognos Analytics Dashboard

The Stories component can help you inform and engage your audience by creating scenes that visualize your data and to tell a narrative. To further enhance your analysis, you can use the Explorations component, which is a flexible workspace where you can discover and analyze data, including visualizations from a dashboard or story.

No Business Intelligence and Analytics platform is complete without report. With the Reporting component, you can create any reports that your organization requires, such as invoices, statements, and weekly inventory reports. The Reporting component is a web-based report authoring tool that professional report authors and developers use to build sophisticated, multi-page, multiple-query reports against multiple databases.

2.4.8 Using the Cognos Analytics tools

The following Cognos Analytics components can be used to complete various tasks:

- Dashboards and Stories:
  - Monitor events or activities at a glance by providing key insights and analysis about your data.
  - Explore your data through visualizations by using the interactive title, drilling up or down columns, and viewing details of a data point.
  - Create forecasts to discover and model trends, seasonality, and time dependence in data.
  - Communicate your insights through a story with scenes and animations in sequence over time.
  - Share your dashboards and stories.

- Exploration:
  - Start an exploration from visualizations in a dashboard or story, or from a data module.
  - Analyze relationships in your data by using the relationship diagram.
  - Explore the relationships in your data by changing the scope of your relationship diagrams, along with the strength of the related fields that are displayed.
- Reporting:
  - Create sophisticated, multi-page, multiple-query reports against multiple data modules.
  - Run variations of reports with report views.
  - Subscribe to reports where you control when it is delivered and the format.

Initial setup and configuration considerations

After you install the Cognos Analytics service, you must provision an instance. During the provisioning of the instance, you must specify the size of the deployment as Fixed Minimum, Small, Medium, or Large, which is determined during the sizing phase when reviewing your workload requirements.

You also must set up and configure a database as a content store after you install the Cognos Analytics service. The content store is used to store configuration data, global settings, data server, connections, and product-specific content.

Learn more

For more information about the Cognos Analytics service, see the following resources:

- IBM Documentation:
  - Cognos Analytics on Cloud Pak for Data
  - Installing Cognos Analytics
  - Postinstallation setup for Cognos Analytics
  - Get started with Dashboards and Stories
  - Explorations
  - Getting started in IBM Cognos Analytics - Reporting
  - Tutorial: Cognos Analytics dashboards
- Chapter 4, "Multicloud data integration" on page 163.

2.4.9 Planning Analytics service

The Planning Analytics service for Cloud Pak for Data provides an AI-powered Extended Planning and Analysis (xP&A) solution platform. This platform is an integrated planning solution that integrates, streamlines, and scales planning, budgeting, and forecasting across every part of your organization, including finance, sales, supply chain, marketing, and beyond.

It includes dashboards and scorecards to help you uncover deep insights to identify trends and drill down into the data with reporting and analysis capabilities. These capabilities are further enhanced with built-in predictive analytics, multidimensional analysis, and intelligent workflows to improve the accuracy and frequency of your forecasts.

You also can test and compare your assumptions by creating what-if scenarios in your own personal sandbox to see the effect before making a decision. By using these tools and features, you can quickly create more accurate plans and forecasts to accelerate decision making and pivot in-real time with a complete up-to-date view of your organization.

Within the Planning Analytics service, Planning Analytics Workspace is a web-based interface that provides full modeling, reporting, planning, and administrative capabilities.

In addition to Planning Analytics Workspace, the Planning Analytics service provides access to the following Planning Analytics components:

- IBM Planning Analytics Workspace: A web-based interface that provides full modeling, reporting, and administrative capabilities.
- Planning Analytics for Microsoft Excel: An Excel-based tool that you can use to build sophisticated reports in a familiar spreadsheet environment.
- Planning Analytics IBM TM1® Web: A web-based interface that you can use to interact and perform administrative tasks with Planning Analytics data.

You also can use the next generation Planning Analytics database. This database is in technical preview for Cloud Pak for Data 4.5.

Planning Analytics Engine is enterprise class, cloud-ready, and available exclusively on the Cloud Pak for Data platform that uses container and Kubernetes container infrastructure. Planning Analytics Engine includes the following key features:

- Database as a service: Planning Analytics Engine runs as a service with which you can manage all your Planning Analytics Engine databases through a single service endpoint.
- High Availability: Planning Analytics Engine can run individual databases in High Availability mode. When running in this mode, the service manages multiple replicas of the database in parallel, which ensures that all changes are propagated to all replicas while dispatching requests in such a way as to spread the load on the overall system.
- Horizontal scalability: Planning Analytics Engine allows the number of replicas of any database to be increased or decreased without any downtime. This feature allows customers to scale up during peak periods and scale during quiet times without any interruption to users.

Figure 2-17 shows you an example of Planning Analytics report.

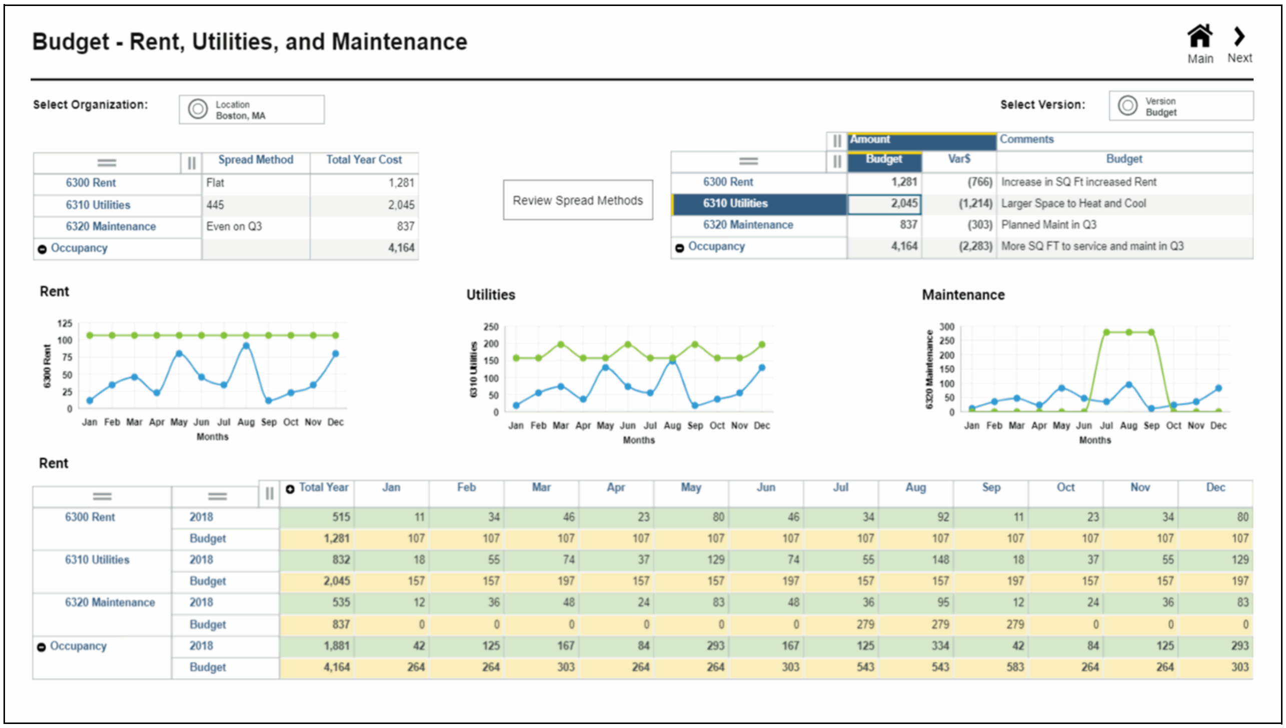

Figure 2-17 Planning Analytics Report

Using the Planning Analytics tools

The Planning Analytics service instance provides the Planning Analytics Workspace, along with access to the Planning Analytics for Microsoft Excel and Planning Analytics TM1 Web.

Use the Planning Analytics Workspace to:

- Build a model of your business.
- Import and maintain business data.
- Manage users and administer security.
- Monitor and administer the Planning Analytics database where your data is stored.
- Automate common processes and chores.
- Create and share complex reports, complete with visualizations.
- Build private sandboxes where you can explore unlimited "what-if" scenarios.
- Contribute to the planning process with advanced data entry capabilities.

Use the Planning Analytics for Microsoft Excel to:

- Build and share various report types, from simple to sophisticated.
- Enhance reports with standard Excel functions and a full range of custom Planning Analytics functions.
- Understand trends and anomalies.
- Share your findings with others.
- Perform multidimensional analysis and exploration of large data sources.

Use the TM1 Web to:

- View, analyze, edit, and chart your Planning Analytics data.
- Analyze cube data.
- View and edit data in formatted Excel reports.
- Drill, pivot, select, and filter data.
- Build charts from cube data.

Initial setup and configuration considerations

After you installed the Planning Analytics service, you must provision an instance. During the provisioning of the instance, you must specify the size of the deployment as Small, Medium, or Large, which is determined during the sizing phase when reviewing your workload requirements.

Also, during the provisioning stage, you must specify the location of your TM1 databases. If you enabled the tech preview, you can use the Planning Analytics Engine database.

When the provisioning is complete, the Planning Analytics service is available for you and your users to model, report, and plan with your business data.

Learn more

For more information about the Planning Analytics service, see the following resources:

- IBM Documentation:
  - Planning Analytics on Cloud Pak for Data
  - Installing Planning Analytics
  - Postinstallation setup for Planning Analytics
- IBM Blog: Experience faster planning, budgeting, and forecasting cycles on IBM CLoud Pak for Data
- Videos:
  - Extend planning across the enterprise
  - IBM Planning Analytics with Watson for Microsoft Excel
  - Planning Analytics Reports and Dashboards
  - Using what-if scenario analysis to make smarter decisions
- Chapter 7, “Business analytics” on page 483.

2.4.10 Db2 Big SQL

The Db2 Big SQL service is an advanced SQL engine that is optimized for workloads on data that is stored in your big data environments. This SQL engine combines open source big data technology with powerful Db2 SQL processing to drive interactive and batch advanced query analytics on your enterprise data.

The high-performance SQL engine can access your data that is stored in Hadoop environments or object stores by using a single database connection or a single query. As a result, you can efficiently and securely query your data across your enterprise. This data can be stored in Hadoop clusters by using open source components and object stores such as, HDFS, Hive metastore, and S3.

Using Db2 Big SQL

The use of Db2 Big SQL with IBM Cloud Pak for Data can be useful in the following situations:

- You need to query large amounts of data that is stored on a Hadoop secured (Kerberized) or unsecured clusters.
- You need to query large amounts of data that is stored on public or private cloud object storage.
- You need highly optimized queries for multiple open source data formats, including Parquet, ORC, Avro, and CSV.

After you provision a Db2 Big SQL instance, you can access your data that is stored on remote Hadoop clusters and cloud object stores. Now, you can start using SQL queries to explore your data and then, analyze your data further. This analysis can include the other Cloud Pak for Data analytic services, such as Watson Studio and Jupyter notebooks.

Initial setup and configuration considerations

A Db2 Big SQL instance that is deployed on Cloud Pak for Data can be configured in one of three ways:

- Hadoop side-car

  The Db2 Big SQL engine runs on your Cloud Pak for Data cluster, but the metadata and data are stored on the same remote cluster's Hive metastore and the HDFS. This configuration allows you to offload your resource intensive analytic workloads from the Hadoop cluster to Cloud Pak for Data while retaining your metadata, data, security policies. The SQL engine reads the data over the network from the HDFS in parallel.

- Pure object store

  The Db2 SQL engine runs on your Cloud Pak for Data cluster, but the data is stored on an object service; the metadata is stored in a Hive metadata service that is associated with the Db2 Big SQL instance. The object storage can be from a public cloud storage provider, or an on-premises storage cluster that is local to the Cloud Pak for Data cluster.

- Hybrid deployment that is configured for both

  This type of deployment is possible if the remote Hadoop cluster is configured to access an object store. You also can deploy Db2 Big SQL in a hybrid configuration to access tables on HDFS and object store.

Before you install the Db2 Big SQL service or provision the instance, you must first ensure that the HDFS NameNode, Hive metastore, and data nodes on the remote Hadoop cluster can be accessed from the Cloud Pak for Data cluster. You also must grant access to the Hive warehouse directory.

For the cloud object store, ensure that the credentials that are to be used by the Db2 Big SQL instance includes read and write access on the object storage buckets.

After you installed the Db2 Big SQL service, you must provision an instance. When you provision an instance, you can specify the number of worker nodes along with the number of cores and memory size. Then, you can specify the persistent storage to be used by the Db2 Big SQL instance.

The last step consists of setting up the connections to your remote data sources that you configured for access by the Db2 Big SQL instance when you provision an instance. When the instance is provisioned, you can access your data for the cloud object that is stored in these remote locations by using SQL.

Learn more

For more information about the Db2 Big SQL After you have installed the Db2 Big SQL service, see the following IBM Documentation web pages:

- Db2 Big SQL on Cloud Pak for Data
- Preparing to install Db2 Big SQL
- Installing Db2 Big SQL
- Postinstallation setup for Db2 Big SQL
- Working with Db2 Big SQL databases
- Analyzing data stored on S3 or S3 compatible object store services (Db2 Big SQL)
- Using a Jupyter notebook with Db2 Big SQL

2.5 Artificial Intelligence services and developer tools

This section highlights the AI services, along with the developer tools that you can use with Cloud Pak for Data.

2.5.1 Watson Studio

Watson Studio is one of the core components of Cloud Pak for Data. Designed for data scientists and business analysts, Watson Studio's architecture revolves around the Analytics Project.

Figure 2-18 shows how analytics projects are organized.

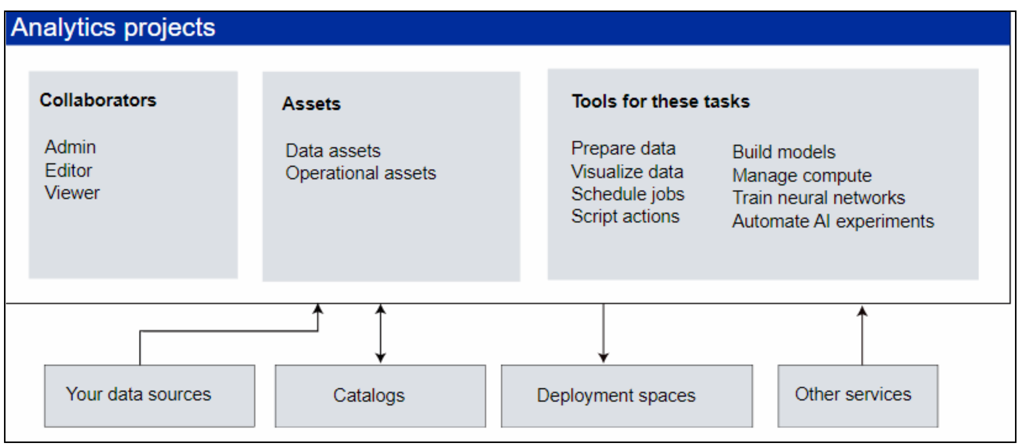

Figure 2-18 How analytics projects are organized

Different types of resources are available in a project, depending on the extra services that are installed, including the following examples:

- *Collaborators* are the people on the team that work with the data.
- *Data assets* point to your data that is in uploaded files or through connections to data sources.
- *Operational assets* are the objects that you create, such as scripts, models, and Jupyter notebooks that run code on the data.

- The *tools* are the software or capabilities that you can use to derive insights from the data, including the following examples:
  - Data Refinery to prepare and visualize data
  - Jupyter notebooks to explore data and build models.
  - AutoAI experiments to create models without using code or programming.

Watson Studio fully integrates with catalogs and deployment spaces. Catalogs are provided by the Watson Knowledge Catalog service with governance and data protection rules. For more information, see 2.3.1, “IBM Watson Knowledge Catalog” on page 59.

Deployment spaces are provided by Watson Machine Learning to easily move assets between projects that are in different deployment spaces, such as pre-production, test, and production. For more information, see 2.5.3, “Watson Machine Learning” on page 90.

Analytics Projects

An Analytics Project is a collaborative workspace where you work with data and other assets to solve a specific goal. You can use Watson Studio to prepare and analyze data and build models. With extra services, such as Watson Machine Learning, you can deploy models to different deployment spaces or create AutoAI experiments.

Figure 2-19 shows the assets that you can add to your Analytics Projects.

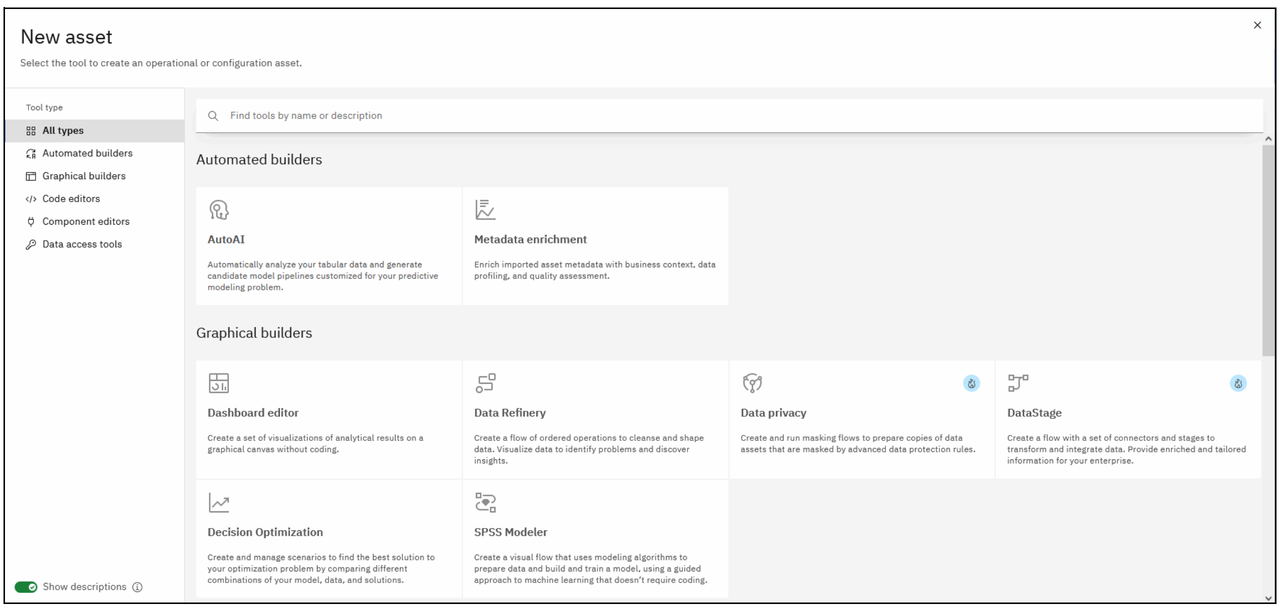

Figure 2-19 Assets that you can add to your Analytics projects 1/2

Figure 2-20 shows the assets that you can add to your Analytics Projects.

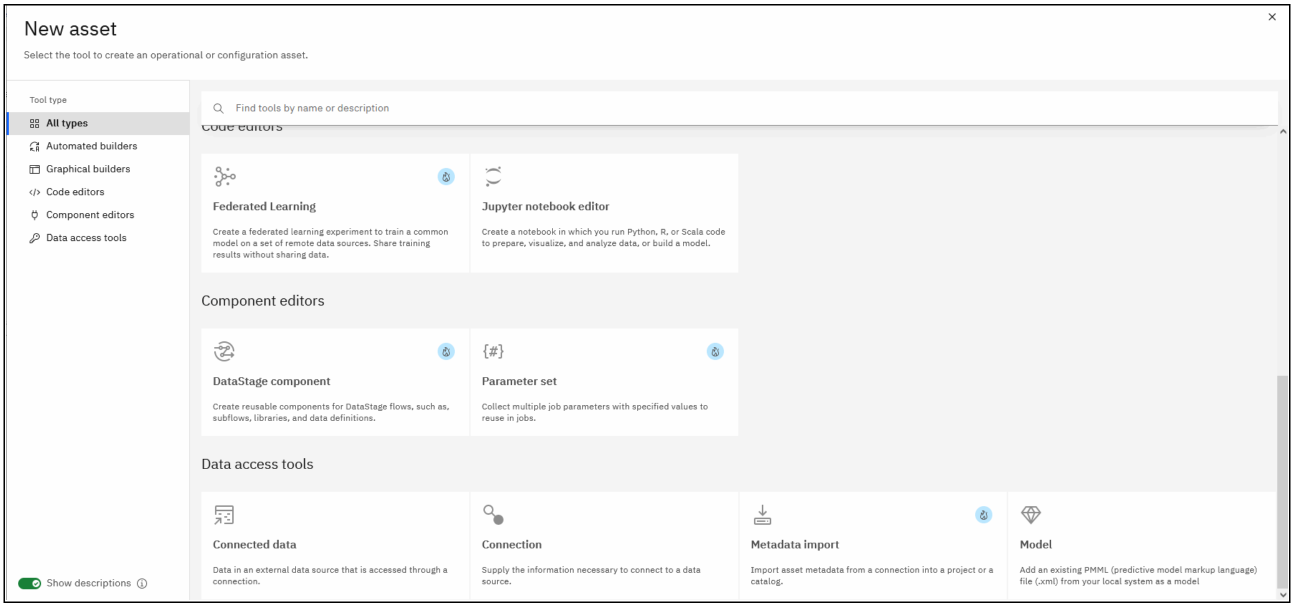

Figure 2-20 Assets that you can add to your Analytics Projects 2/2

Notebooks

The Jupyter notebook editor provides the platform on which you can develop Jupyter notebooks that are written in Python, Scala, or R. *Notebooks* are an interactive environment that can run small pieces of code within cells with its results returned right beneath the cell.

Notebooks include the following building blocks that you need to work with data:

- The data
- The code computation that processes the data
- Visualization of the results
- Text and rich media

When a notebook is in edit mode, only the editor can make changes. All other users see a lock icon and cannot edit that notebook at the same time. Only the project administrator or the editor can unlock the notebook.

Notebooks uses runtime environments to process the data. Different runtime environments are available that you can select. Some of these environments are available immediately, such as the runtime for Python 3.9. For others, you must install and make them available to the platform.

For more information, see 2.5.2, “Developer tools for Watson Studio” on page 88.

Figure 2-21 shows a sample notebook.

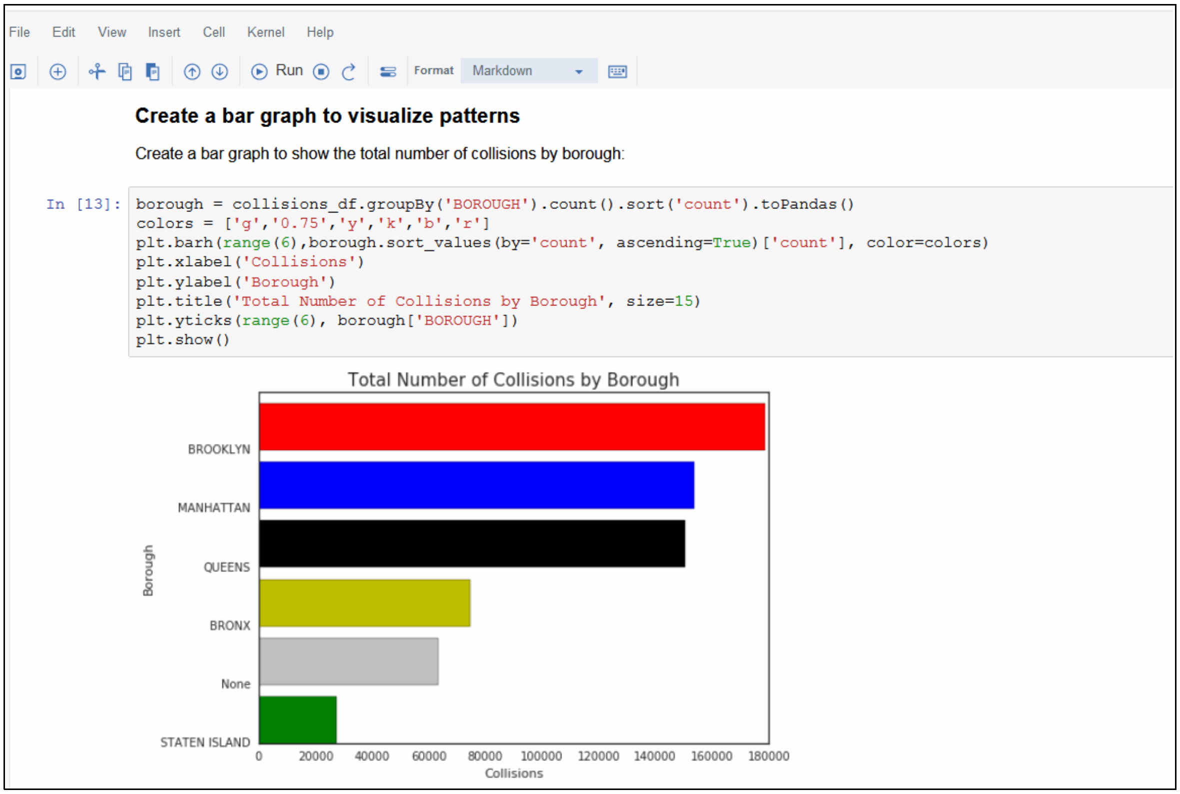

Figure 2-21 Sample notebook

Environment run times

When you install Watson Studio, several runtime environments are available. When you open a notebook in edit mode, one run time is associated with that notebook. Therefore, if two notebooks use with the same run time, two kernels are started, one for each of the notebooks. The resources for those run times are shared.

You can manually restart a kernel, if necessary. However, all execution results are lost if you restart the kernel. If the kernel loses connection and you reconnect to the kernel and the notebook is connected to the same kernel session, the saved results are available.

The following options are available for environment run times for notebooks:

- Execution Engine for Apache Hadoop
- Analytics Engine Powered by Apache Spark
- Jupyter notebooks with Python with GPU and Jupyter notebooks with R 3.6 (as part of the Watson Studio Runtimes package)

For more information see “IBM Watson Studio Runtimes package” on page 88, Chapter 7, “Business analytics” on page 483, and 2.5.3, “Watson Machine Learning” on page 90.

JupyterLab

In Watson Studio, you can work with notebooks by using two different tools: Jupyter notebooks and JupyterLab.

JupyterLab offers an IDE-like environment that includes notebooks. The modular structure of the interface is extensible and open to developers, which enables working with several open notebooks in the same window. The integration with GIT supports collaboration and file sharing.

Learn more

For more information about the Watson Studio service, see the following resources:

- IBM Documentation:
  - Watson Studio on Cloud Pak for Data
  - Installing Watson Studio
  - Postinstallation tasks for the Watson Studio service
  - Creating notebooks (Watson Studio)
  - Coding and running a notebook (Watson Studio)
- APIs
- Chapter 8, “IBM Cloud Pak for Data Operations” on page 569.

2.5.2 Developer tools for Watson Studio

The tools that are described in this section integrate with Watson Studio.

Anaconda Repository

By using Anaconda Repository, you can control the open source packages that data scientists can use in Jupyter notebooks and JupyterLab in Watson Studio analytics projects. You can receive Conda package updates in real time with access to:

- Anaconda packages in Python and R
- Open Source packages in Conda-Force, CRAN and PyPI
- Your own proprietary packages

With the Anaconda Repository, you can control the environment, which packages are allowed, who can access them, and any dependencies that are required by the enterprise.

Note: Installing Anaconda Repository must be done on a Linux system. It cannot be installed on a Red Hat OpenShift node. For more information, see the following IBM Documentation web pages:

- Installing Anaconda Repository for Cloud Pak for Data
- Customizing an environment for Anaconda Repository for Cloud Pak for Data

IBM Watson Studio Runtimes package

The IBM Watson Studio Runtimes package includes runtime environments that you can use with your notebooks. The default run time that is installed when you install Watson Studio is the Jupyter notebooks with Python 3.9.

If you have GPU nodes in your cluster, data scientists can use the Jupyter notebooks with Python 3.9 with GPU to run experiments and train compute intensive machine learning models in Watson Studio. If you require R, install the Jupyter notebooks with the R.3.6 run time environment.

Data scientists that choose to use Jupyter notebooks with Python 3.9 for GPU must create a custom environment definition in each analytics project. They also must specify GPU as the compute engine type and Python 3.9 as the software configuration.

The experiment build tool that is provided by the Watson Machine Learning Service requires GPU environments. Before you can use the GPU runtime, an administrator must install the service.

For more information, see this IBM Documentation web page.

Data scientists who use R can use the Jupyter notebooks with R3.6. This run time provides compute environments to run Jupyter notebooks in the R 3.6 coding language in Watson Studio. To use this run time when you create a notebook, an administrator must first install the Watson Machine Learning Service.

For more information, see this IBM Documentation web page.

RStudio Server with R 3.6

RStudio provides an IDE for working with R in Watson Studio to create R Shiny applications. This service provides RStudio Server 1.4 and R 3.6.

R is a popular statistical analysis and machine-learning package that enables data management that includes tests, models, analysis, and graphics. RStudio provides the IDE for working with those in Watson Studio. To use RStudio, you create an Analytics project and then, start the IDE from within your project.

Figure 2-22 shows the RStudio IDE.

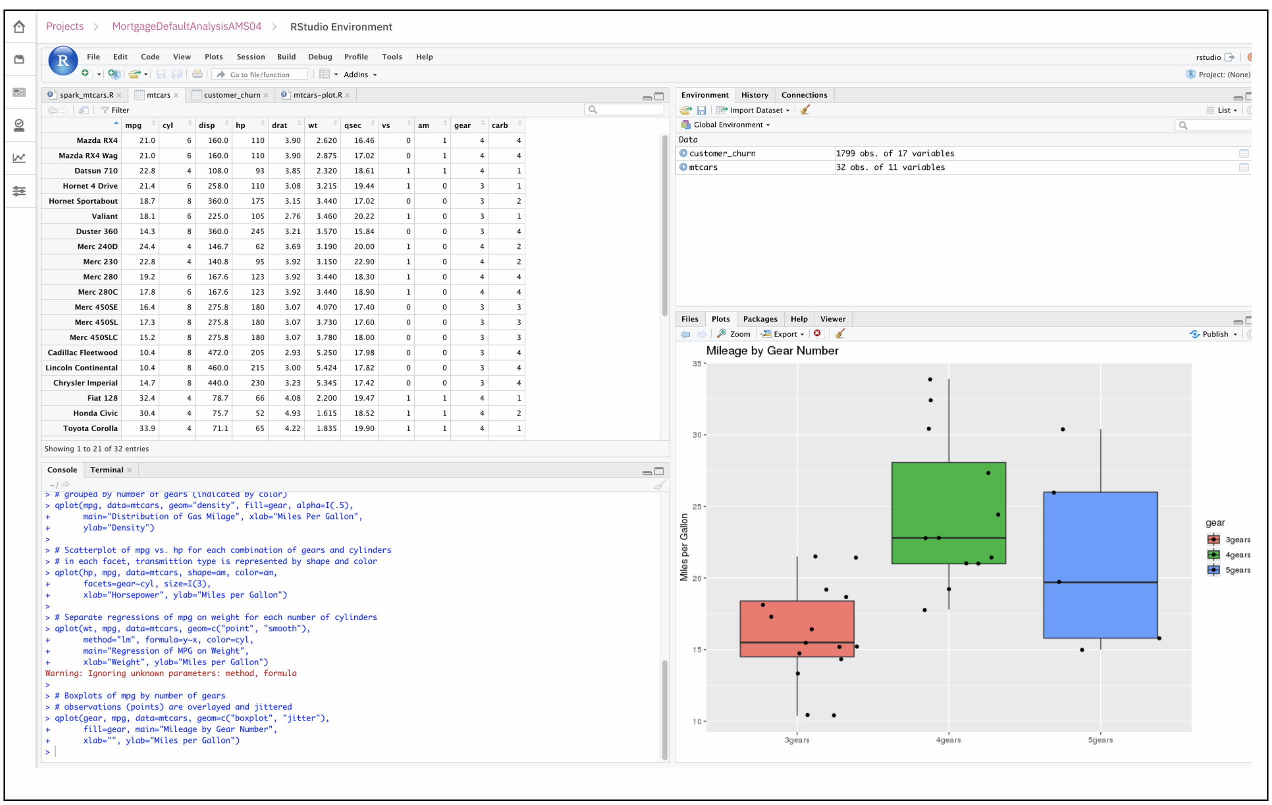

Figure 2-22 RStudio IDE

Before you can use RStudio, an administrator must first install this service. For more information, see this IBM Documentation web page.

Learn more

For more information about the Developer tools, see the following IBM Documentation web pages:

- Anaconda Repository for IBM Cloud Pak for Data on Cloud Pak for Data
- Jupyter Notebooks with Python 3.9 for GPU on Cloud Pak for Data
- Jupyter Notebooks with R 3.6 on Cloud Pak for Data
- RStudio Server with R 3.6 on Cloud Pak for Data

2.5.3 Watson Machine Learning

You can use Watson Machine Learning to build, manage, and deploy machine learning models. Watson Machine Learning is often installed with Watson Studio to provide a full suite of tools that data scientists and developers can use to build sophisticated analytical models.

> **Note:** Unless specified otherwise, all of the capabilities that are discussed in this section assume that you installed Watson Studio *and* Watson Machine Learning.

Watson Studio provides a holistic view and approach to the use of analytics projects with added integrated services. This section focuses on the Watson Machine Learning capabilities that work with Watson Studio.

The following capabilities are available with Watson Machine Learning:

- AutoAI experiment builder is a simple-to-use tool to automatically process structured data to generate model-candidate pipelines. You can select the best performing pipelines to save as a machine learning model and then, deploy for scoring. No coding experience is needed to create your models with AutoAI.
- Jupyter notebooks provide an interactive programming environment for uses who are working with data, testing models, and rapid prototyping. You can use notebooks to develop and code your own models by using programming languages, such as, Python, Scala, or R.
- Deep Learning experiments automate running hundreds of training runs while tracking and scoring results. You can use this capability to build neural networks for use cases, such as image processing.
- Analytic deployment spaces provide the tool to manage model deployments. You use this tool to manage your models as they proceed through the end-to-end model lifecycle.

Building models by using AutoAI

AutoAI is a graphical tool in Watson Studio with Watson Machine Learning that analyzes your data and builds predictive models without any required coding.

The overall process that is used to build an AutoAI experiment includes the following steps:

1. Provide the data through a data connection or a data file.
2. Create and run the AutoAI experiment, which automatically goes through the following tasks:
   - Data pre-processing
   - Automated model selection
   - Automated feature engineering
   - Hyperparameter optimization

For more information about the AutoAI implementation, see this IBM Documentation web page.

3. View the results of the pipeline generation process. You can select the leading model candidate and evaluate them before saving a pipeline as a model. This model then can be deployed to a deployment space for further testing before going into production.

For more information about building an AutoAI model, see this IBM Documentation web page.

Deep learning experiments

With IBM Watson Machine learning, you can train a neural network by using the experiment building feature in Watson Studio. To build deep learning experiments, you also need the Watson Machine Learning Accelerator. For more information, see 2.5.4, “Watson Machine Learning Accelerator” on page 92.

The idea behind deep learning is to train thousands of models to identify the correct combination of data with hyperparameters to optimize the performance of your neural networks. Watson Machine Learning accelerates this process by simplifying the process to train models in parallel with auto-allocated GPU compute containers.

For more information about deep learning experiments, see this IBM Documentation web page.

Analytic deployment spaces

An AI lifecycle involves the following major phases:

- Build: In this phase, you build, train, and test your models.
- Deploy: In this phase, you deploy, score, and monitor your models in a dashboard.
- Trust: In this phase, you evaluate your deployments for bias or drift. retrain your models, and update your deployments when specific thresholds are met to maintain accuracy.

The use of deployment spaces are how you organize your model deployments. It contains deployable assets, such as model deployments, jobs, associated input and output data, and the associated environments.

Figure 2-23 shows the Deployments dashboard.

Figure 2-23 Deployments dashboard

The Deployments dashboard is an aggregate view of the deployment activities across deployment spaces. You can create, delete, and manage access to deployment spaces. You also can export and import spaces.

Learn more

For more information about the Watson Machine Learning Service, see the following resources:

- IBM Documentation:
  - Watson Machine Learning on Cloud Pak for Data
  - Installing Watson Machine Learning
- APIs
- Chapter 5, “Trustworthy artificial intelligence concepts” on page 267.

2.5.4 Watson Machine Learning Accelerator

Deep learning is the subset of machine learning. Deep learning can interpret text, images, audio, and video at scale, and generate patterns for recommendation engines, sentiment analysis, financial risk modeling, and anomaly detection.

These deep learning models process neural networks with which it learns from large amounts of data through each of its layers. Although a neural network with a single layer can still make accurate predictions, multiple layers (including hidden layers) can help to optimize and refine for accuracy.

Watson Machine Learning Accelerator is a deep learning platform that data scientists can use to build, train, and deploy deep learning models. Watson Machine Learning Accelerator can be connected to Watson Machine Learning to take advantage of the resources that are available across Watson Machine Learning projects.

Watson Machine Learning Accelerator provides the following benefits:

- Distributed deep learning architecture and hyper-parameter search and optimization that simplifies the process of training deep learning models across a cluster.
- Large model support that helps increase the amount of memory that is available for deep learning models per network layer, which enables more complex models with larger, more high-resolution data inputs.
- Advanced Kubernetes scheduling, including consumer resource plans, the ability to run parallel jobs and dynamically allocate GPU resources.
- Included deep learning frameworks, such as TensorFlow and PyTorch.
- An administrative console for GPU cluster management and monitoring.

Working with the service

As a data scientist, you can use Watson Machine Learning Accelerator Service to build your deep learning models. The overall process to build the models is as follows:

1. Connect to the data.
2. Train the model.
3. Monitor training progress.
4. Deploy the model.
5. Test the deployed model.

These steps might seem familiar because this process also is used to develop a machine learning model. Watson Machine Learning Accelerator provides specific tools to accelerate this effort.

You can connect to the Watson Studio Experiment Builder to train your neural networks. You can use the API to do your training. You can use a Jupyter notebook to build and train your models. You can view the console to monitor the training progress.

Learn more

For more information about the Watson Machine Learning Accelerator Service, refer to the following links:

- IBM Documentation:
  - Watson Machine Learning Accelerator on Cloud Pak for Data
  - Installing Watson Machine Learning Accelerator
- Tutorials:
  - Use AI to assess construction quality issues that impact home safety
  - Expedite retail price prediction with Watson Machine Learning Accelerator hyperparameter optimization
  - Drive higher GPU utilization and throughput
- Chapter 5, “Trustworthy artificial intelligence concepts” on page 267.

2.5.5 Watson OpenScale

Trusted AI became critical because of the adoption of AI across industries. As enterprises adopt AI, you must analyze AI with trust and transparency to understand how those AI models make decisions.

You also must detect and mitigate bias (implicit and explicit) and drift in the accuracy of your models over time. You must increase the quality and accuracy of your model's prediction.

Watson OpenScale provides an environment for AI applications with the visibility into how the AI is built and used. By using Watson OpenScale, you can scale adoption of trusted AI across enterprise applications.

For an example, we use an insurance company. The system suggests rejecting an applicant that seems to meet all of the necessary criteria for a claim.

However, how can the claims processor understand why? T

The explain feature enables them to get details about the decision. Traceability allows them to trace the process back to the source documents the AI drew its decision from. And if there is bias, the bias feature pinpoints how it occurs and auto mitigates it. Over time, if the model loses accuracy, the drift feature monitors and alerts the user through the OpenScale dashboard when it reaches a threshold.

Working with the service

Watson OpenScale consists of the following main areas.

- Insights provide the dashboard that displays the models that you are monitoring and the status of the results of the model evaluations.
- Explain a transaction describes how the model determined a prediction. It lists some of the most important factors that led to the prediction so that you can be confident in the results.
- By using Configuration Traceability, you can set up the database to be used as the data mart for the models. You also can specify your own machine learning provider, or use the Watson Machine Learning Service within Cloud Pak for Data. You can also integrate with other services, such as Watson OpenPages, for model lifecycle management. Watson OpenPages, see 2.5.6, "Watson OpenPages" on page 97.
- Support is available if you need help with Watson OpenScale. You can access product documentations, or even connect to the IBM Support team by creating tickets.

When you configure a model to be monitored with Watson OpenScale, the following default monitors can be used:

- Quality monitor describes the model's ability to provide correct outcomes that are based on labeled test data, which is known as *Feedback data*.
- Fairness monitor describes how evenly the model delivers favorable outcomes between groups. The fairness monitor looks for biased outcomes in your model.
- Drift monitor warns you of a drop in accuracy or data consistency.
- You also can create custom monitors for use with your model deployments.

Figure 2-24 shows the Explain transaction for the sample model.

Figure 2-24 Explain transaction for the sample model

For each model deployment that you want to track and monitor by using Watson OpenScale, you must set up and enable the evaluation by using the following process:

1. Select the model.
2. Set the data type for the payload logging.
3. Review the configuration for each of the monitors that you use.

Fairness metrics overview

The fairness metric evaluates your model to determine whether it produces biased outcomes. When you set up your fairness metrics, you can specify features from your model that you want to monitor for bias.

You also must specify values for each feature. Among those features, you must select a monitored group and a reference group. For example, you can set the Female value as the monitored group and the Male value as the reference group for the Sex feature.

You also must specify the output schema for a model or function in Watson Machine Learning to enable fairness monitoring in Watson OpenScale.

Quality metrics overview

The quality metric evaluates how well your model predicts outcomes. Quality metrics are calculated with manually labeled data and monitored deployment responses. For correct monitoring, you must regularly provide feedback data to Watson OpenScale. Some of the quality metrics include, but are not limited to:

- Area under ROC
- Area under PR
- Accuracy
- True Positive rate and False Positive rate
- Recall
- Precision
- F1-Measure
- Logarithmic loss

Drift detection overview

Drift detection evaluates the degradation of predictive performance over time. Watson OpenScale detects and highlights drift so that you can prevent errors in your model. It detects the drop in accuracy and data consistency over time, which can lead to a negative effect on the business if it is not resolved.

Initial setup and configuration of Watson OpenScale

After the OpenScale service is installed, a default instance is created. You can create more instances on the same platform. Each instance is independent, but they share a pool of services and hardware resources. If more separation is required, each instance can be set up with its own Db2 data mart.

It is recommended that you perform the automated setup, which runs the demonstration scenario that is provided when you first start Watson OpenScale. You are prompted to specify your local instance of Watson Machine Learning.

Then, you also must provide an instance of a Db2 database to be used as the data mart. The process takes some time to complete as it loads the sample model deployment and the pre-configured monitors.

From there, you can take the self-guided tour or explore the model and monitors on your own. This process confirms that your setup and configuration is successful, and you can build and load your own models.

Configuring model risk management and model governance

Watson OpenScale combined with Watson OpenPages provide an end-to-end model governance solution. For more information about Watson OpenPages, see 2.5.6, “Watson OpenPages” on page 97.

This section explains how you can integrate the Watson OpenPages directly with Watson OpenScale.

For the two services to be fully integrated, you must add your Watson OpenPages URL and authentication credentials to Watson OpenScale. To add these credentials, in the Configure section of the Watson OpenScale sidebar, select **Integrations** and enter the URL, username, and API key for the Watson OpenPages instance.

After the integration is set up, you can perform the model analysis in Watson OpenScale. As a result, you now can send all the metrics to the Watson OpenPages where the model was originally developed.

For more information about this end-to-end process with Watson OpenPages and Watson OpenScale, see 2.5.5, “Watson OpenScale” on page 94.

Learn more

For more information about the Watson OpenScale service, see the following resources:

- IBM Documentation:
  - Watson OpenScale on Cloud Pak for Data
  - Installing Watson OpenScale
  - Configuring model monitors
  - Configure Watson OpenScale with advanced setup
- APIs
- Chapter 5, “Trustworthy artificial intelligence concepts” on page 267.

2.5.6 Watson OpenPages

Watson OpenPages is an integrated governance, risk, and compliance (GRC) suite that empowers enterprises to manage, monitor, and report on risk and compliance. This service provides several tool sets that span across various use cases, including the following examples:

- Management of:
  - Business continuity
  - Financial controls
  - Data privacy
  - Internal audit
  - Operation risk
  - Policy
  - Regulatory compliance
  - Third-party risk
- Governance of:
  - IT
  - Model risk

Watson OpenPages integrates with Watson OpenScale. The models that are monitored by using Watson OpenScale take advantage of Watson OpenPages’ model risk governance capabilities.

Figure 2-25 shows the Watson OpenPages dashboard.

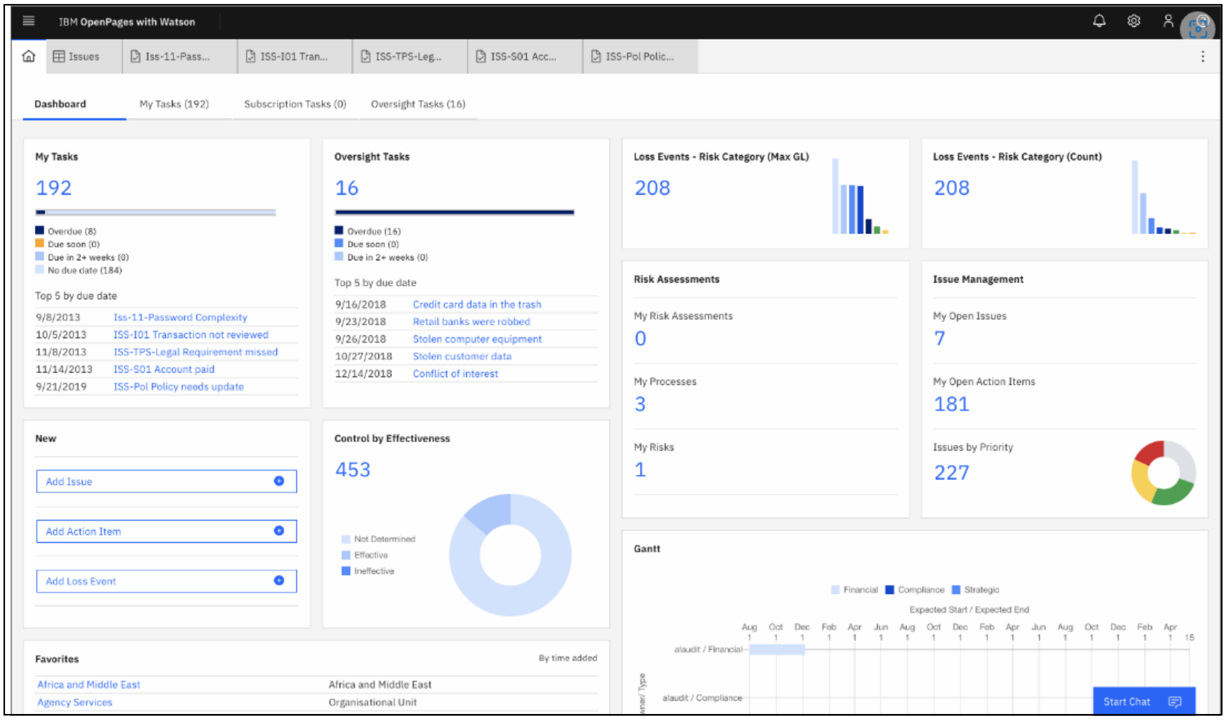

Figure 2-25 Watson OpenPages dashboard

Working with the service

Watson OpenPages combined with Watson OpenScale offers an end-to-end model risk management solution. Watson OpenPages manages a comprehensive model inventory with model risk governance. Watson OpenScale offers model evaluation, metrics, and measure outcomes of the models in deployment.

You start by setting up the model in Watson OpenPages and set the Monitored with OpenScale option. From there, you take the model through the candidate and development workflow for model evaluation.

Then, you link the model to a sample Watson OpenScale model.

Finally, you export the Watson OpenScale metrics to Watson OpenPages as part of the pre-implementation validation process and explore ways to view and interpret these metrics by using Watson OpenScale.

After you return to Watson OpenPages with the metrics that you received from Watson OpenScale, you change the model status to be Approved for Deployment. After it is approved, you can move the new model to production in Watson OpenScale to continue gathering metrics for the production model. Use the analysis of the model to refine the model as needed.

Learn more

For more information about the Watson OpenPages service, see the following resources:

- IBM Documentation:
  - OpenPages on Cloud Pak for Data
  - Preparing to install OpenPages
  - Installing OpenPages
  - Postinstallation setup for OpenPages
- GitHub Tutorial: End-to-end model risk management with Watson OpenPages
- Chapter 5, "Trustworthy artificial intelligence concepts" on page 267.

2.5.7 Watson Assistant

The Watson Assistant service enables you to build conversational interfaces into any application, device, or channel. Much more than a chatbot, which often is only a Question and Answer (Q&A) tool, Watson Assistant analyzes the messages from the customer and provides specific answers or tasks to further clarify the questions.

Watson Assistant uses training data to generate a customized machine model, so it knows when to search for an answer in a knowledge base, to ask for clarification, or to direct users to a person. The machine learning model provides the logic behind the assistant to understand the customer request and to answer them correctly.

Watson Assistant routes the input from the customer to the suitable skill that you, as the developer of the assistant, creates. Two skills can be developed with each assistant service:

- Dialog skill: Interprets the customer's input and gathers any information that it needs to respond, or uses that information to perform a task on the customer's behalf.

  For example, if the customer is looking for information that is stored in another database or system, Watson Assistant can query that database to gather the necessary information to return back to the customer.

- Search skill: Integrates with another service, called *Watson Discovery*, to perform complex searches across disparate sources. Watson Discovery treats the customer's input as a search query and it finds and returns the results back to Watson Assistant so that the response is sent back to the customer. This entire process occurs seamlessly from the customer's perspective.

With the skills that you develop, your Watson Assistant service can answer simple or complex questions, perform tasks, such as opening tickets, updating account information, or placing orders.

Working with the service

One of the questions that people often ask when they hear about the power behind Watson Assistant is how can they build their own assistant.

This section highlights the workflow that is needed to define the scope of the assistant project to create the details behind the logic of the business opportunity you are trying to solve.

The typical workflow for an assistant project includes the following steps:

1. Define a narrow set of key requirements that you want to solve with your project. Start small and iterate.
2. Create intents that represent the customer needs that you identified in the previous step; for example, `#store_hours` or `#place_order`.

3. Build a dialog that detects the defined intents and addresses them with a simple response or a dialog flow that collects more information.
4. Define any entitles that are needed to clarify the customer's meaning.
5. Test each function that you add to the assistant by using the Try it window.
6. After you create a few key tasks, add an integration that deploys the assistant to a development environment. Test the development assistant and make refinements.
7. After you have an effective assistant, take a snapshot of the dialog skill and save it as a version. The version gives you the ability to go back if subsequent changes you make to the skill decreases its effectiveness.
8. Deploy the version of the assistant to the test environment and then, test it.
9. Monitor the assistant transcript logs of the test assistant to determine whether you must make improvements to your training data or dialog.
10. When you are satisfied with the version of the assistant, deploy that version into a production environment.
11. Monitor the logs from the conversations that users have with the deployed assistant.

The process of monitoring the logs is critical to continuing to improve the quality and accuracy of your assistant. For example, if you see that an intent is used for too many questions incorrectly, you want to adjust that intent. Or, if many customers are asking something that the assistant does not know anything about, you want to create another intent. Or, if two or more intents are not distinct enough and you are consistently getting incorrect responses, you want to update the intents.

Figure 2-26 shows the development workflow.

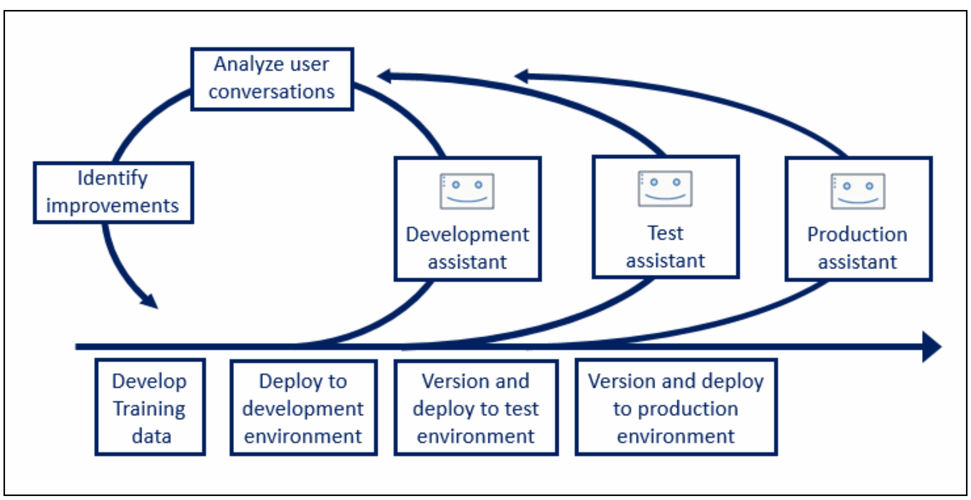

Figure 2-26 Development workflow

Initial setup and configuration

After an administrator installs the service, it must be provisioned. You can provision up to 30 instances per deployment. Each instance contains their own access level, so you can create different instances for different teams, as an example.

Learn more

For more information about the Watson Assistant service, see the following resources:

- IBM Documentation:
  - Watson Assistant on Cloud Pak for Data
  - Installing Watson Assistant
  - Postinstallation setup for Watson Assistant
- Tutorials:
  - Building a complex dialog
  - Adding a node with slots to a dialog
  - Improving a node with slots
  - Understanding digressions
- APIs
- Chapter 6, “Customer care” on page 439

2.5.8 Watson Discovery

The Watson Discovery service provides an AI-powered, intelligent search and text analytics platform that helps you uncover valuable insights from complex business documents. Watson Discovery can search through your enterprise documents by using advanced search capabilities, such as reading comprehension, curations, and table retrievals.

A built-in contract understanding function searches and interprets legal contracts. It can also conduct in-depth analysis of unstructured text such, as images.

Working with the service

You start by connecting your data to Watson Discovery. Then, you teach Watson Discovery to understand the language of that document, including the concepts that are unique to your business and your industry.

Use Watson Discovery to enrich your data with custom Natural Language Understanding (NLU) technology so that you can identify key patterns and information.

Finally, you build search solutions to find answers to queries, explore data to uncover patterns, and use the search results in an automated workflow, such as with Watson Assistant.

Watson Discovery is made up of four main concepts:

- Projects: A project is a way to organize and manage resources in your Watson Discovery application.
- Collections: A collection is a set of documents that you upload or crawl from where it is stored on a connected data source.
- Fields: As documents are crawled, unstructured text is organized into fields, such as author, file type, and text.
- Enrichment: Enrichments are the AI capabilities that you can apply to fields to identify and extract relevant information from your document.

The following project types are available and each includes the correct set of enrichments applied to the documents automatically:

- Document Retrieval: Search and find the most relevant answers from your data.
- Document Retrieval for Contracts: Apply contract enrichments to English-language contracts that can recognize and tag contract-related concepts in your data.
- Conversational Search: Returns information from a connected data collection as answers to questions that customers ask an assistant.
- Content Mining: Discover hidden insights, trends, and relationships in your data.
- Custom: Because no automatic enrichments are applied, you can add the ones that you need for your use case.

Figure 2-27 shows the general development steps for the use of Watson Discovery.

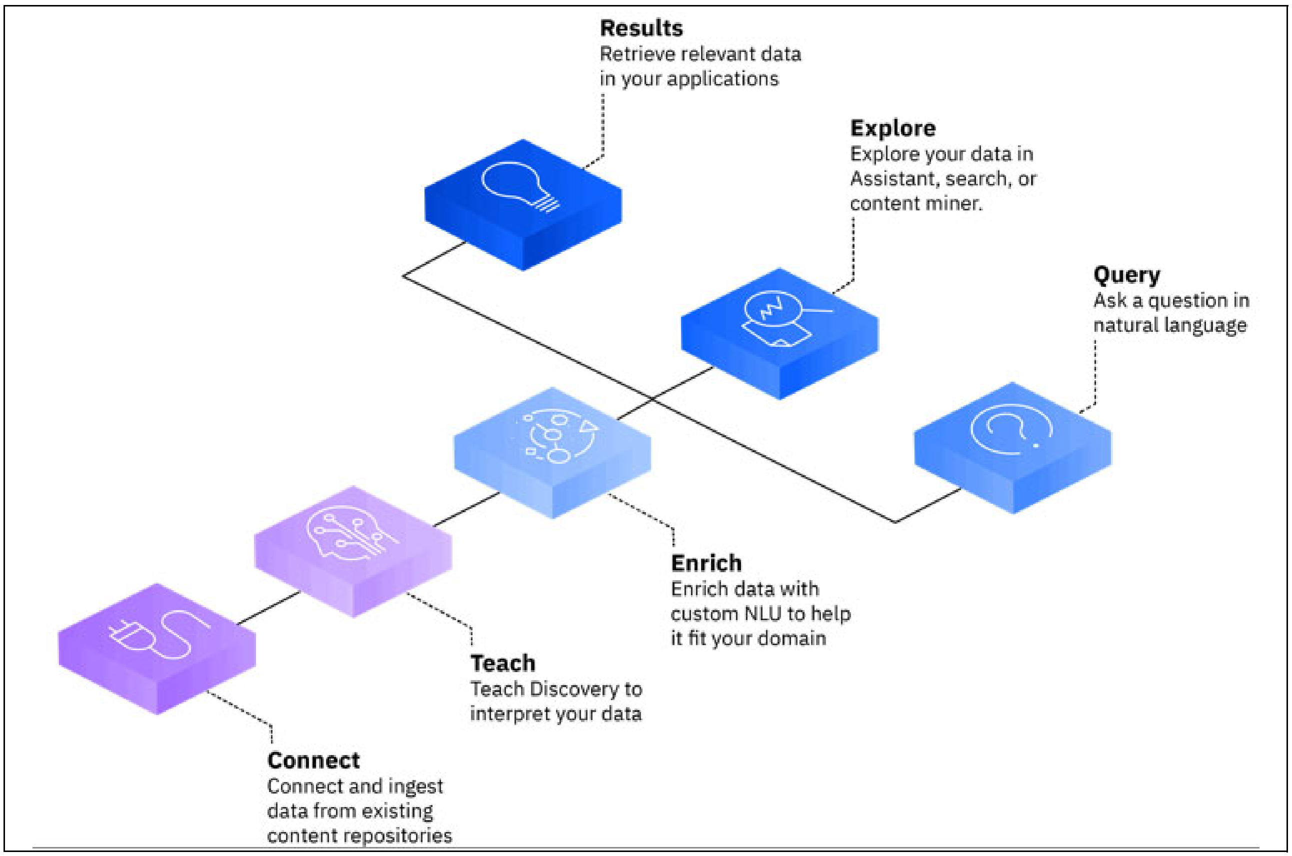

Figure 2-27 Development steps for using Watson Discovery

You can decide how you want to organize your source content into collections. One example is if you receive content from different sources, you can create a collection for each, such as a website and Salesforce. Each collection adds data from a single source. Then, when they are built together in a single project, a user can search across both sources at the same time.

You can also create Smart Document Understanding (SDU) models. These models help you identify content that is based on the structure of the document. For example, if you have 20 PDF files from one department and another 20 PDF files from a different department, you can use SDU to build a model for each source department with their own document structures.

The model can define custom fields that are unique to the source documents.

Enrichments are a method with which you can tag your document in your collection to fit your domain. With IBM Watson Natural Language Processing (NLP), you can add pre-built enrichments to your documents. The following enrichments are available:

- Entities: Recognizes proper nouns, such as people, cities, and organizations that are mentioned in the content.
- Keywords: Recognizes significant terms in your content.
- Parts of Speech: Identifies the parts of speech in the content.
- Sentiment: Understands the overall sentiment of the content.

Initial setup and configuration

After the service is installed by an administrator, a service instance must be provisioned. You can provision up to 10 instances, each with their own level of access control.

Learn more

For more information about the Watson Discovery service, see the following resources:

- IBM Documentation:
  - Watson Discovery on Cloud Pak for Data
  - Installing Watson Discovery
  - Postinstallation setup for Watson Discovery
- Tutorials:

  > **Note:** Although the tutorials that are listed here were designed for the service on IBM Cloud, they also apply to the software version on Cloud Pak for Data.

  - Help your chatbot answer frequently asked questions
  - Get quick answers from existing help content
  - Use Smart Document Understanding to improve search results
- APIs

2.5.9 Watson Speech to Text

Watson Speech to Text provides speech recognition capabilities for your applications. The service is ideal for customers who must extract high-quality speech transcripts from call center audio.

The service uses machine learning to combine knowledge of grammar, language structure, and the composition of audio and voice signals to accurately transcribe the human voice. As more speech audio is received, the machine learning model updates and refines its transcription.

You also can customize the service to suit your language and application needs. You can send a continuous stream of data or pass prerecorded files. The service always returns a complete transcript of the audio that you send.

For speech recognition, the service supports synchronous and asynchronous HTTP REST interfaces. It also supports a web socket interface that provides a full-duplex, low latency communication channel. Clients send requests and audio to the service and receive results over a single connection asynchronously.

Working with the service

The quickest way to get started is to use the curl-based service. Make sure you have the curl command installed on your system. Then, from the Cloud Pak for Data service instance page, open the speech-to-text instance to obtain the token and the service URL.

The next step is to transcribe audio with no options. Call the `POST /v1/recognize` method to request a basic transcript of a FLAC audio file with other request parameters.

Complete the following steps:

1. Download the sample audio-file.flac at this web page.
2. Issue the command that is shown in step 5 to call the service's `/v1/recognize` method for basic transcription with no parameters.
3. Replace {*token*} with the access token for the service instance.
4. Replace {*url*} with the URL for the service instance.
5. Modify {*path_to_file*} to specify the location of the audio file:

```
curl -X POST \
--header "Authorization: Bearer {token}" \
--header "Content-Type: audio/flac" \
--data-binary @{path_to_file}audio-file.flac \
"{url}/v1/recognize"
```

   The service returns the following transcription results:

```
{
  "result_index": 0,
  "results": [
    {
      "alternatives": [
        {
          "confidence": 0.96
          "transcript": "several tornadoes touch down as a line of severe
thunderstorms swept through Colorado on Sunday "
        }
      ],
      "final": true
    }
  ]
}
```

To transcribe audio with options, modify the command to call the service's `/v1/recognize` method with two extra parameters.

For example, `{url}/v1/recognize?timestamps=true&max_alternatives=3` sets the timestamp parameter to true to indicate the beginning and end of each word in the audio stream and the maximum alternative to `3` to receive the three most likely alternatives for the transcription.

The service returns the following transcription results, which include timestamps and three alternative transcriptions:

```
{
  "result_index": 0,
  "results": [
    {
      "alternatives": [
```

```
        {
          "timestamps": [
            ["several", 1.0, 1.51],
            ["tornadoes":, 1.51, 2.15],
            ["touch":, 2.15, 2.5],
            . . .
          ]
        },
        {
          "confidence": 0.96
          "transcript": "several tornadoes touch down as a line of severe
thunderstorms swept through Colorado on Sunday "
        },
        {
          "transcript": "several tornadoes touched down as a line of severe
thunderstorms swept through Colorado on Sunday "
        },
        {
          "transcript": "several tornadoes touch down as a line of severe
thunderstorms swept through Colorado and Sunday "
        }
      ],
      "final": true
    }
  ]
}
```

Next-generation languages and models

Watson Speech to Text supports a growing collection of next-generation models that improve upon the speech recognition capabilities of the previous-generation models. Next-generation models also provide noticeably better transcription accuracy.

When you use next-generation models, the service analyzes audio bidirectionally. That is, the model evaluates the information forwards and backwards to predict the transcription, effectively listening to the audio twice.

The following types of next-generation models are available:

- Telephony models are intended specifically for audio that is communicated over a telephone. As with previous-generation narrow band models, telephony models are intended for audio that has a minimum sampling rate of 8 kHz.
- Multimedia models are intended for audio that is extracted from sources with higher sampling rate, such as video. Use a multimedia model for any audio, other than telephonic audio. As with previous-generation broadband models, multimedia models are intended for audio that has a minimum sampling rate of 16 kHz.

Select the model that best suits your needs.

Next-generation language models include, but are not limited to, the following examples:

- Czech: cs-CZ_Telephony
- English (United Kingdom): en-GB-Telephony
- French (France): fr-FR_Multimedia
- German: de-DE_Multimedia
- English (all supported dialects): en-WW-Medical_Telephony

Note: The en-WW-Medical_Telephony model is in beta as of this writing. Common use cases of this model include conversations between a patient and a medical provider where the model can understand the terminology from the medical profession.

Initial setup and configuration

After the installation process is complete, you provision a service instance and provide access to the users of the platform. From the service instance page, you obtain the instance URL and token that is used in the commands to make the API call.

Learn more

For more information about the Watson Speech services, see the following resources:

- IBM Documentation:
  - Watson Speech services on Cloud Pak for Data
  - Preparing to install Watson Speech services
  - Installing Watson Speech services
  - Postinstallation setup for Watson Speech services
- Demonstrations:
  - Watson Speech to Text (US English)
  - Watson Speech to Text (All languages)

2.5.10 Watson Text to Speech

Watson Text To Speech provides speech synthesis capability for your application to convert written text to natural sounding speech. The service streams the audio back to the customer.

The Watson Text to Speech service is suitable for voice-driven and screenless applications, where audio is the preferred method of output. The service offers HTTP REST and WebSocket interfaces. With the WebSocket interface, the service can return word timing information to synchronize the input text and the resulting audio.

The Watson Text to Speech Service can synthesize text to audio in many formats. It also can produce speech in male and female voices for many languages and dialects. The service accepts plain text and that text is annotated with XML-based Speech Synthesis Markup Language. It also provides a customization interface that you can use to specify how to pronounce certain words that occur in your input.

Working with the service

You can use the curl-based commands to quickly get started with the service. Make sure you have the curl command installed on your system. Then, from the Cloud Pak for Data service instance page, open the text-to-speech instance to obtain the token and the service URL.

Complete the following code to synthesis the string “hello world” to audio: `POST /v1/synthesize`. The result is a .WAV file that is named `hello_world.wav`.

Consider the following points:

- Replace the *{token}* with the access token for the service instance.
- Replace *{url}* with the URL for the service instance:

```
curl -X POST \
--header "Authorization: Bearer {token}" \
--header "Content-Type: application/json" \
--header "Accept: audio/wav" \
--data "{\"text\":\"hello world\"}" \
--output hello_world.wav \
"{url}/v1/synthesize?voice=en-US_MichaelV3Voice"
```

The result is in an English-speaking voice that is named `en-US_MichaelV3Voice`. You can use a different voice, such as a female voice, by using the following parameter in the command:

```
{url}/v1/synthesize?voice=en-US_AllisonV3Voice
```

You also can synthesize text in Spanish:

```
curl -X POST \
--header "Authorization: Bearer {token}" \
--output hola_mundo.wav \
"{url}/v1/synthesize?accept=audio%2Fwav&text=hola%20mundo&voice=es-ES_EnriqueV3Voi
ce"
```

Initial setup and configuration

After the installation process completes, you provision a service instance and provide access to the users of the platform. From the service instance page, you obtain the instance URL and token that is used in the commands to make the API call.

Learn more

For more information about the Watson Speech services, see the following resources:

- IBM Documentation: See the web pages that are listed in the Learn More section in the previous section.
- Demonstration: Watson Text to Speech
- Tutorial: Getting started with Watson Speech to Text

> **Note:** Although this tutorial is designed for the service on IBM Cloud, it also applies to the software version on Cloud Pak for Data.

2.5.11 Watson Knowledge Studio

Watson Knowledge Studio allows you to create machine learning (ML) models that identify entities and relationships from your domain in unstructured text. You can build these models in a collaborative environment with developers and domain experts, without needing to write code.

Then, you can use those models in Watson Discovery. The goal of Watson Knowledge Studio allows for an end-to-end cycle of domain adaptation of unstructured texts.

You use Watson Knowledge Studio to identify custom entitles and relations to train the custom model that then recognizes those entitle in text, as shown in the following example:

```
ABC Motors has received great reviews for its new 2020 Lightning.
```

The custom entity and relations model can be trained to recognize “2020 Lightning” as a `Vehicle` entity, and “ABC Motors” as a `Manufacturer` entity. The model also can be trained to recognize that the two entities are connected by an `isManufacturedBy` relation.

Working with the service

The Watson Knowledge Studio service provides tools for annotating unstructured domain documents and uses those annotations to create a custom ML model that understands the language of the domain. The accuracy of the model improves through iterative testing and can ultimately recognize patterns in a large collection of documents.

The overall process is used:

1. Import source document.

   Based on a set of domain-specific source documents, the development team creates a system that defines the entity types and relation types.

2. Annotate and adjudicate

   A group of two or more human annotators annotates a small set of source documents to label words that represent entity types, identify relation types, and define co-references, which identify different mentions of the same thing or same entity. Any inconsistencies are resolved, and one set of optimally annotated documents is built, which forms the ground truth.

3. Train the model by using the ground truth.
4. Apply the new model to documents that are new to the system.

Figure 2-28 shows the overview process to build and apply a model.

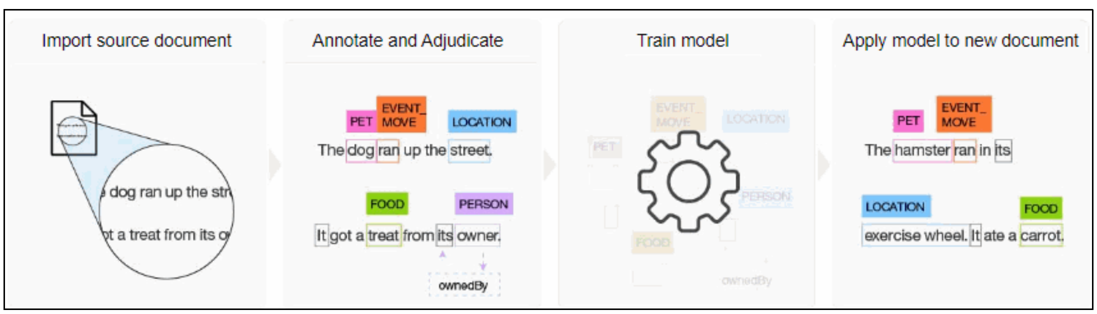

Figure 2-28 Build and apply a model

Watson Knowledge Studio provides a rule editor that simplifies the process of building rules to capture common patterns. You can then create a model by using those rule patterns.

To deliver a successful project, you must form a team of subject matter experts (SMEs), project managers (PMs), and users who can understand and interpret statistical models. Then, you create a workspace that contains the artifacts and resources that are needed to build the model. You then train the model to produce a custom model that you can apply to new documents.

The machine learning model creation workflow (see Figure 2-29) shows the steps that are performed by the PM and the human annotators, often known as the SMEs or domain experts.

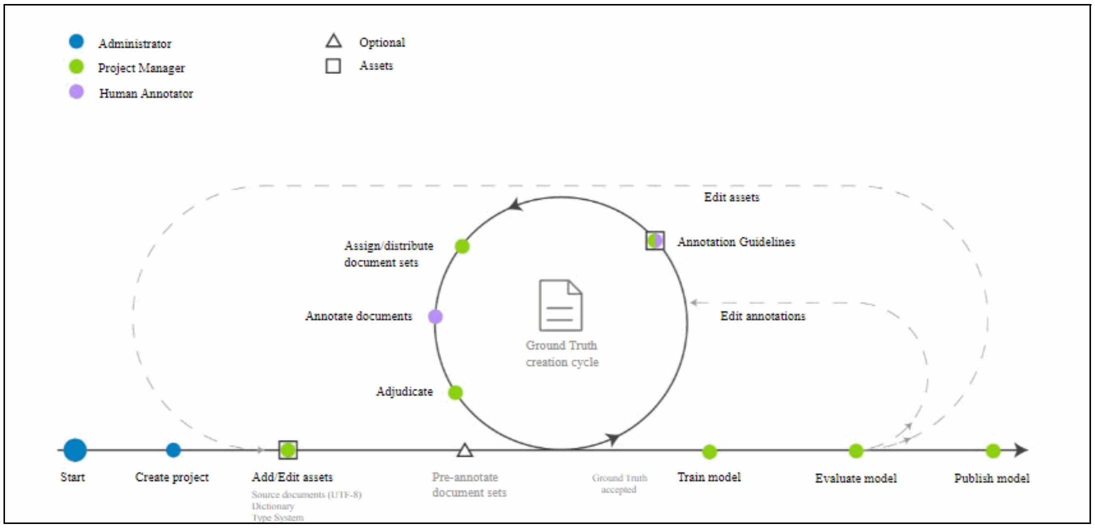

Figure 2-29 ML model creation workflow

At the start of the project, an administrator creates the project and then adds the suitable users into the system. Then, the PM adds the assets, including documents and other resources into the workspace.

Then, it goes through a cycle of annotation by the SMEs to establish what is known as the ground truth. This ground truth plays an important part in the accuracy of the model creation.

After the ground truth Motors is established, the PM trains the model and evaluates its accuracy. If necessary, it returns to the ground truth creation cycle until the model evaluation produces an optimal model. That model is then published and used on never-before-seen documents, often with the Watson Discovery service.

Initial setup and configuration

After an administrator installed the service, a service instance also must be provisioned. Access must be granted to the users that want to access and use the service. In addition, the administrator also must grant the users permission from within the Watson Knowledge Studio service. This permission is another layer of access that must be granted to use this service.

Learn more

For more information about the Watson Knowledge Studio service, see the following resources:

- IBM Documentation:
  - Watson Knowledge Studio on Cloud Pak for Data
  - Installing Watson Knowledge Studio
- Tutorials:

  > **Note:** Although these tutorials are designed for the service on IBM Cloud, they also apply to the software version on Cloud Pak for Data.

  - Creating a machine learning model
  - Creating a rule-based model
  - Pre-annotating documents
  - Creating an advanced rules model

2.5.12 IBM Match 360 with Watson

IBM Match 360 with Watson is a Modern Master Data Management (MDM) solution that combines Operational MDM and Analytical MDM capabilities.

The service helps consolidate data from disparate data sources and resolve duplicate person, organization, and other customer data to fit-for-purpose 360-degree views of your customers.

IBM Match 360 with Watson helps generate data models automatically that are based on your source data. IT also further enriches, extends, and customizes them with more attributes and custom entity types.

The built-in machine learning-assisted matching technology helps minimize data matching, mapping, and profiling effort. It also trains and tunes your matching algorithm in line with the business needs of your enterprise.

The service is designed with user needs in mind. It provides rich self-service capabilities for consumers of customer entity data and allows them to tailor the presentation layer according to their own preferences.

Figure 2-30 shows the data mapping setup page.

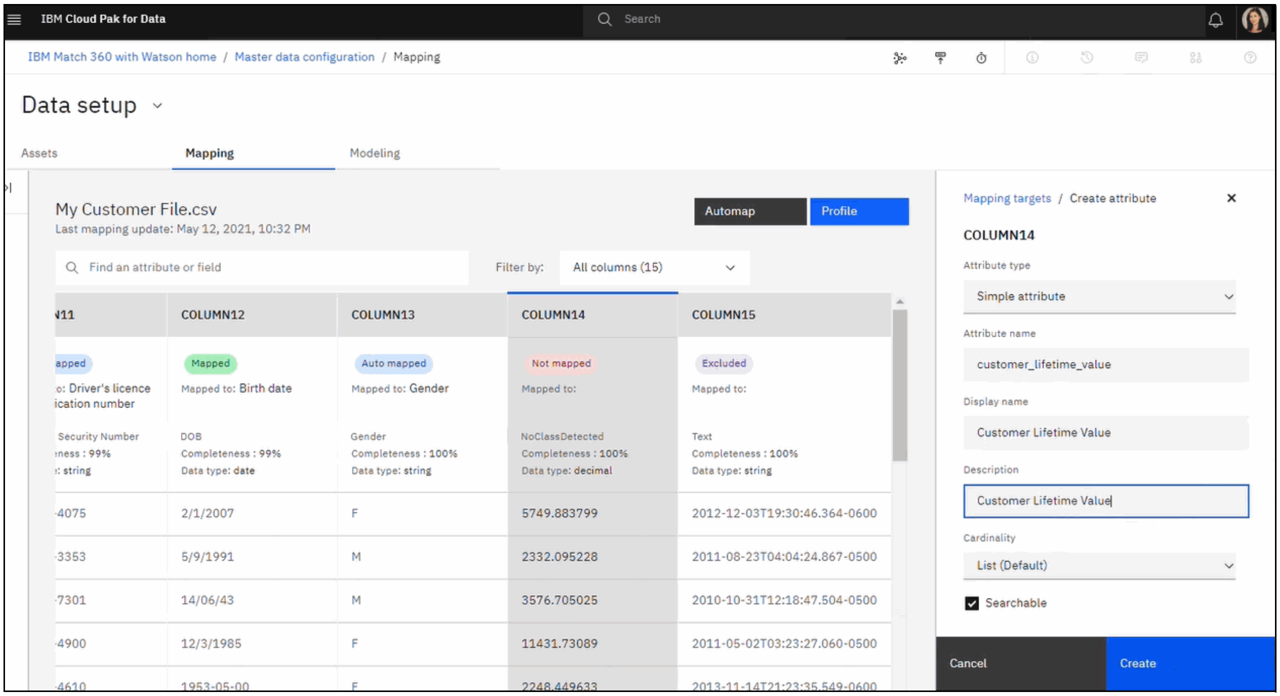

Figure 2-30 Data-mapping setup page

Work with the service

IBM Match 360 with Watson is instance-based. After the service is installed on your IBM Cloud Pak for Data cluster, a single instance of the service is provisioned. The user who provisions the instance (often the `admin` user), then is automatically assigned as the administrator of that instance. No other users can access the service instance by default.

The service includes four dedicated roles (personas), each with a specific permissions profile:

- Data Engineer
- Data Steward
- Publisher User
- Entity Viewer

To start working with the service instance, relevant users must be added to it and assigned one of these dedicated roles.

Also, a project must be created and associated with the instance as part of the prerequisite service set-up.

IBM Match 360 with Watson provides the following integrated complementary user experiences:

- Master Data Configuration for preparing and configuring master data

  Users-assigned Data Engineer user profiles for the service entity often use the set of capabilities features that is bundled into this experience to upload, connect, and configure data sources; run profiling and mapping; tweak and enhance the generated data and data models; and tune and configure the matching algorithm.

- The Master Data for self-service exploration and analytics

  All users and roles can use this experience to search, view, and analyze master data entities and records, and configure entity and record search results presentation preferences. Users with Data Steward user profile also can add, edit, and export records and entities.

Figure 2-31 shows the Master data explorer experience.

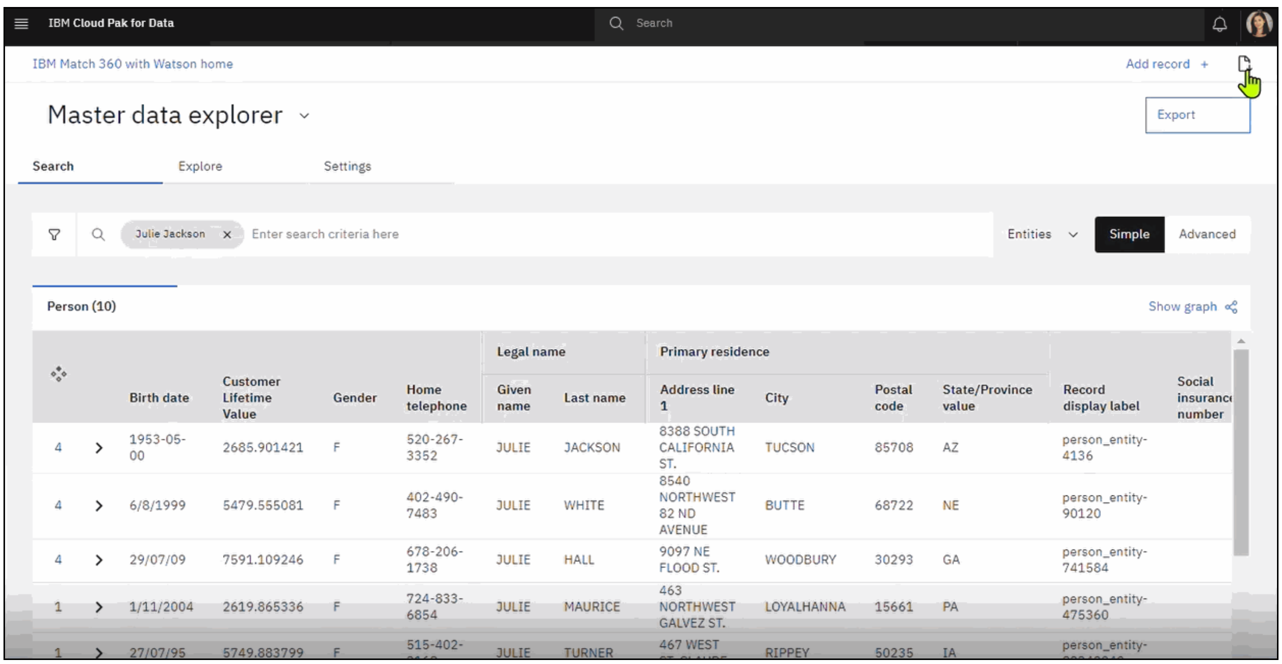

Figure 2-31 Master data explorer experience

IBM Match 360 with Watson is tightly integrated with the IBM Watson Knowledge Catalog service, and stand-alone IBM Master Data Management Advanced Edition and Standard Edition solutions (IBM MDM AE/SE). Consider the following points:

- The auto-mapping and profiling features of the service use the capabilities and set up from Watson Knowledge Catalog. If those features are required, IBM Watson Knowledge Catalog service must be installed on your IBM Cloud Pak for Data cluster, and a catalog must be associated with the Match 360 service instance. The auto-mapping and auto-profiling capabilities are supported only for person and organization record types (in version 4.5.2 at the time of this writing).
- Master data from stand-alone IBM MDM AE/SE can be exported and published into IBM Match 360 with Watson by using the standard MDM Publisher tool that is available for and bundled with IBM MDM AE/SE. Then, it further enriched and combined with data from other sources within IBM Match 360 with Watson.

Note: The Publisher tool requires separate sizing and installation, with the installation physically close to the source IBM MDM AE/SE instance, which is recommended as a best practice. Licensing guidelines are out of scope of this paper. Contact your IBM Sales representative to ensure suitable entitlement coverage.

Learn more

For more information about the Match 360 service, see the following resources:

- IBM Documentation:
  - IBM Match 360 on Cloud Pak for Data
  - Preparing to install IBM Match 360 with Watson
  - Installing IBM Match 360 with Watson
  - Postinstallation setup for IBM Match 360 with Watson
  - Managing master data by using IBM Match 360 with Watson
- Tutorial: Onboarding and matching data in IBM Match 360 with Watson

2.6 Dashboards

This section highlights the Dashboard service that is available on Cloud Pak for Data: Cognos Dashboards.

2.6.1 Cognos Dashboard service

The Cognos Dashboard service provides a graphical canvas to help you accelerate visualizing your data to uncover accurate insights and see patterns within complex data. You can share these insights through the visualizations that you created by using this service to encourage data-driven decision making quickly and confidently.

By using the Db2 Cognos Dashboard editor, you can drag data onto the canvas and use various visualizations to check correlations and connections or understand relationships or trends in your data. Then, you can quickly build sophisticated visualizations to help you answer your important questions or provide a foundation for more in-depth analysis.

Working with dashboards to visualize your data

The dashboards editor provides a graphical canvas to begin investigating data for insights and patterns. You do not need to understand coding or SQL to explore the data and gain insights.

You can use the dashboard editor in projects to build visualizations of your analytics results, and communicate the insights that you discovered in your data on a dashboard. Alternatively, you can transfer the dashboard to a data scientist for deeper analysis and predictive modeling.

The data that you can use with the dashboard editor are data in flat files (CSV) and from various database tables, such as IBM Db2, IBM Data Virtualization, and PostgreSQL.

You also can create a dashboard from many different templates that contain predefined designs and grid lines for easy arrangement and alignments of the visualizations. Then, you can select your source data from one of your data source connections.

Now, you can start to visualize your data from the selected data source connection by adding different types of graphs and widgets to enhance your dashboard. When you are finished creating your dashboard, you can publish it so that it can be shared.

Learn more

For more information about the Cognos Dashboard service, see the following IBM Documentation web pages:

- Cognos Dashboards on Cloud Pak for Data
- Installing Cognos Dashboards
- Visualizing data with Cognos Dashboards

3

Data governance and privacy

Data governance and privacy are key challenges for many enterprises where a need exists to balance the benefit of data access with the need to protect sensitive data.

Cloud Pak for Data provides the capabilities that your enterprise needs to automate data governance and privacy so you can ensure data accessibility, trust, protection, security, and compliance.

This chapter includes the following topics:

3.1 Introduction

The data governance and privacy use case covers the key tasks that are used to manage metadata and policies to support the activities of self-service data access, data quality, and data privacy. It also provides the base governance capabilities to support regulatory compliance with regulations, such as General Data Protection Regulation (GDPR) and Basel Committee on Banking Supervision (BCBS) 239.

The Watson Knowledge Catalog service in Cloud Pak for Data provides the tools and processes that your organization needs to implement a data governance and privacy solution. The following key steps are described in this use case:

- Establish:
  - Users of the system.
  - Category, project, and catalog that is to be used for running tasks.
  - Governance foundation
- Import technical metadata into a project.
- Enrich the technical metadata with the glossary information.
- Create data quality rules to check the quality of the technical metadata.
- Publish the enriched technical metadata to a catalog.
- Create data protection rules to control access to the contents of the technical metadata.
- Report on the business and technical metadata that is contained in our instance of Watson Knowledge Catalog.
- Use advanced metadata import capabilities to bring in lineage information for our data assets.

To begin, we create user groups for the users of our system. The use case serves several personas that are involved in tasks that include the creation, enrichment, and consumption of assets. These personas are represented by using the user, group, and role constructs of the Cloud Pak for Data platform.

The standard installation of Cloud Pak for Data provides the Admin and User roles within the platform. With the addition of Watson Knowledge Catalog, the following new roles are introduced to the platform that are the producers of information that is related to the governance and privacy use case:

- Data Steward
- Business Analyst
- Reporting Administrator
- Data Engineer
- Data Quality Analyst

In addition, the following new personas are introduced to the platform that use the artifacts that are created by the personas:

- Data Scientist
- Developer

For this use case, we set up the following user groups:

- Test admins are the platform administrators. They are assigned platform administrator permissions and are responsible for setting up the projects and categories within the platform

- Test stewards are the platform users that create the governance foundation artifacts. They are created with the Data Steward and Data Engineer Cloud Pak for Data roles. They have editor roles in the categories and projects that are established by the test administrator role.

Complete the following steps to create these user groups:

1. From the hamburger (navigation) icon, select **Access control** and then, select the **User groups** tab. Click **Create new usergroup**. Initially, we create a user group for administrators, as shown in Figure 3-1. Because we also create a static group of users, ensure that the **Assigned** option is selected.

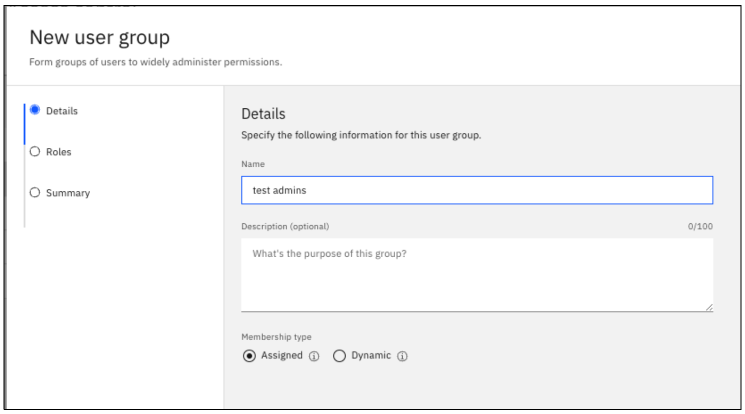

Figure 3-1 Creating the test admin group

2. Click **Next**. The available users on the cluster are shown. It is assumed here that Cloud Pak for Data is installed and the internal Cloud Pak for Data user registry was populated (not recommended for production use), or the Identity and Access Management integration with the Cloud Pak Foundational Services was enabled and connected to a user repository.

   Add a single user to the group as an administrator (see Figure 3-2).

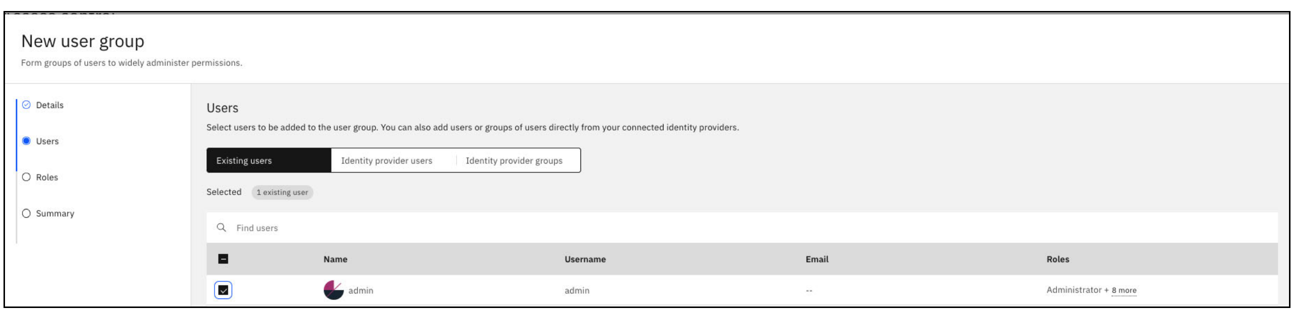

Figure 3-2 Adding users to the test admins group

3. The role for the new group must be established. Therefore, select the administrator role, as shown in Figure 3-3.

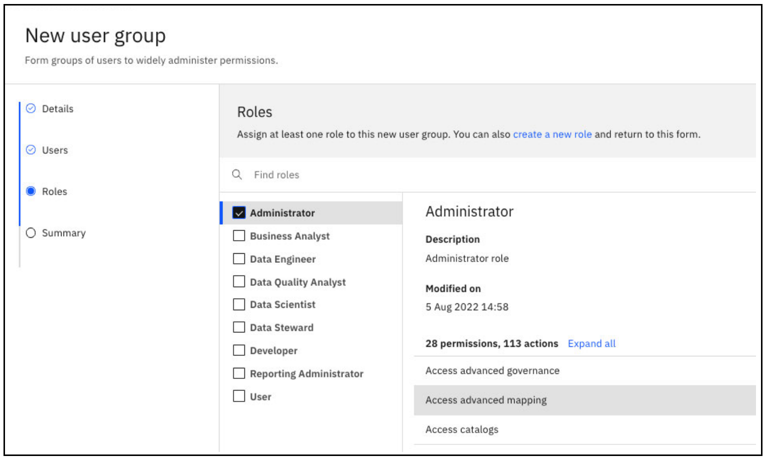

Figure 3-3 Adding the administrator role to the new user group

4. Create a user group to represent the data stewards that are responsible for curating data. Create a user group that is named `test stewards`, as shown in Figure 3-4.

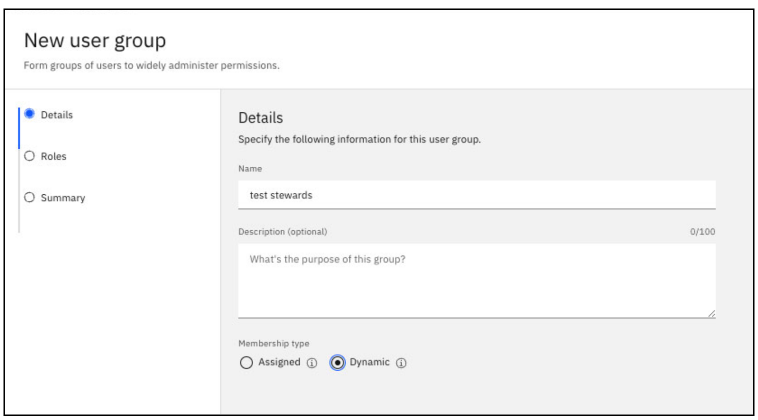

Figure 3-4 Adding the test stewards user group

5. Add at least one user to the user group, as shown in Figure 3-5.

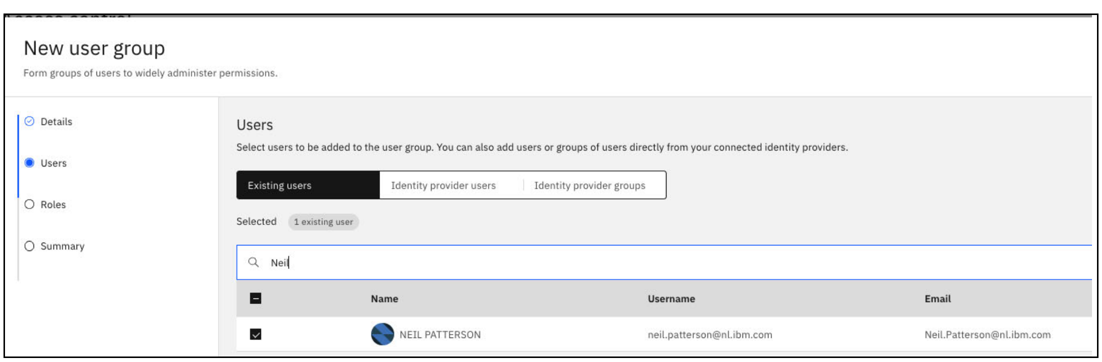

Figure 3-5 Adding users to the test stewards group

These users must have the data steward and data engineer role assigned, as shown in Figure 3-6.

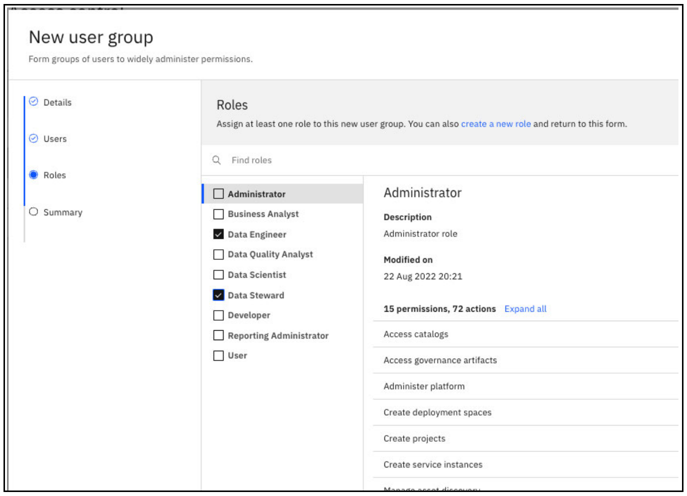

Figure 3-6 Assign roles to test stewards group

The rest of this chapter provides the details of the data governance and privacy use case that sets the foundation for the other Data Fabric use cases, such as MLOps.

3.2 Establishing the governance foundation

Establishing a governance foundation is a critical step for an organization to transition to a data-centric organization. The governance foundation provides the link to the business terminology, policies, and rules. Several steps that are described in this section must be completed to create the base of this governance foundation.

3.2.1 DataOps assessment and initial scope

The complexity of the organization that is required for implementing a data governance and privacy use case vary among enterprises. Heavily regulated industries (such as financial services) likely have a well-structured data governance organization with leadership coming from the executive board of an organization.

Other industries might not be as far along in the process, but with an ever increasing focus on regulatory regimes the need for this function within an organization is becoming more pressing.

Before embarking on a data governance initiative, the scope must be clearly defined. DataOps teams must focus on aligning the delivery of necessary data with the value it can bring to the business. The teams start small and it is important that they begin by focusing on a problem statement or business initiative that can deliver instant value back to the business.

Within CPD, governance is provided by the use of governance artifacts. The governance artifacts consist of categories, terms, policies, rules, reference data sets, classifications, and data classes. All of these artifacts are accessible from the main Cloud Pak for Data menu in the governance section, as shown in Figure 3-7.

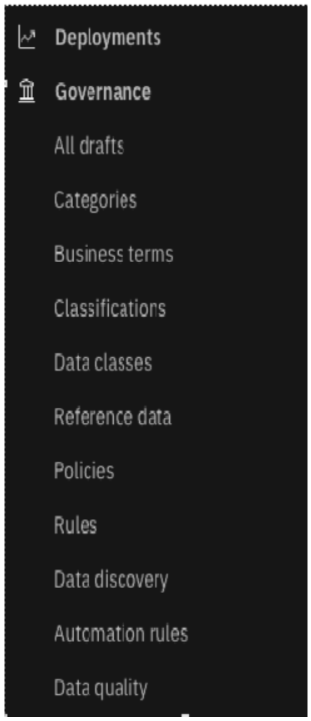

Figure 3-7 Glossary artifact menu

The following sections describe each of the governance artifacts.

Categories

Categories are the containers for the governance artifacts and are used to organize them. Categories can be nested.

Before users of the platform can start creating governance artifacts and curating data, you must create governance categories, add users to categories, and set up workflow configurations to control how governance artifacts are created and published.

Categories provide access control to the governance artifacts by using category collaborators. Cloud Pak for Data users are added to the categories through the Access Control tab of a category with a specific category role of administrator, editor, or viewer. The category role determines the permissions a Cloud Pak for Data user has on the contained governance artifacts.

By default, a single category `[uncategorized]` is created when Watson Knowledge Catalog is installed. This category is populated with a standard set of Classifications and Data Classes.

Complete the following steps to create a category that is to be used to contain your enterprises glossary artifacts:

1. To create the category from the governance artifact menu, select **Categories** and then, select **New Category**, as shown in Figure 3-8

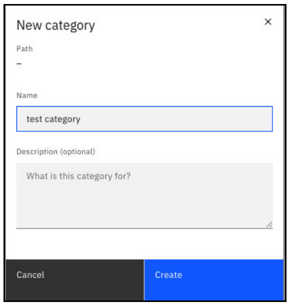

Figure 3-8 Creating the category

2. The collaborators for the newly created category now must be configured. To do so, switch to the **Access** control tab of the category. By default, the category creator is the owner and the special group `All users` is given view access to the category, as shown in Figure 3-9. If you want to prohibit all users from viewing the category, the `All users` group must be removed from the category.

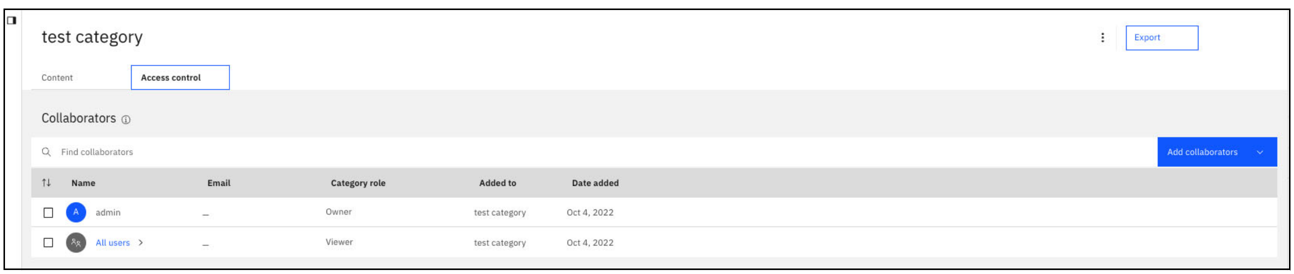

Figure 3-9 Add all users as viewers to test category

3. Select **Add collaborators** → **Add user groups** and then, select **User groups**. The available user groups are presented as shown Figure 3-10.

Figure 3-10 Add groups to test category

4. Select the test admin user group and assign the group the Admin role, select the user group **test stewards** and then, assign the group Editor, Reviewer, and Viewer roles as shown in Figure 3-11.

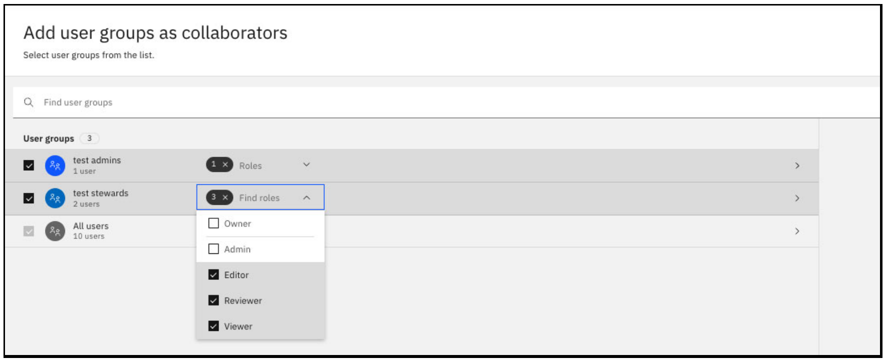

Figure 3-11 Adding user groups as collaborators to the test category

Business terms

Business terms represent the language of the business. They standardize the definitions of business concepts so that is a uniform understanding exists of the concepts across an enterprise. Business terms include a well-defined structure and can be related to each other. Through manual processes or the process of metadata enrichment, they can be associated with IT assets.

Policies

A *policy* is a natural language description of an area of governance. Policies define an organization's guidelines, regulations, processes, or standards. Policies are composed of one or more rules. Policies can be nested and associated with business terms to further enhance the meaning of a term.

Rules

Rules can be split into the following sets:

- *Governance rules* are a natural language description of the actions to be taken to implement a specific policy. These rules can be associated with terms to enhance the terms with natural language descriptions of the rules that apply to the term. These rules form the basis for rule definitions and data rules. For more information, see 3.3.3, "Data quality" on page 146.
- *Data protection rules* are defined to describe how data must be protected, obfuscated, or redacted based on the identity of the user that uses the data. For more information, see 3.3.6, "Data privacy and data protection" on page 158.

Reference data sets

Reference data sets provide logical groupings of code values (reference data values), such as product codes and country codes. These codes typically are sets of allowed values that are associated with data fields and can be assigned to business terms.

Classifications

Classifications are governance artifacts that you can use to classify assets that are based on the level of sensitivity or confidentiality to your organization. They can be used similar to tags to control groupings of assets in your company.

Unlike data classes, which include logic to match data values, classifications are more like labels. A small set of classifications is included with the base Watson Knowledge Catalog. The use of the Knowledge Accelerators introduces another set.

Data classes

Data classes describe the type of data that is contained in data assets. Data classes are used during the data enrichment process to determine the type of data within a data asset by running rules against the data that is contained within a data asset.

The platform includes 160 data classes. The data classes must be checked for validity because all might not apply for your specific use case. New data classes can be created.

Knowledge Accelerators

To provide an enterprise with a head start in their governance initiatives, IBM provides Knowledge Accelerators. These Knowledge Accelerators are prepackaged content and include categories, terms, data classes, and reference data. Knowledge Accelerators are available for the following industries:

- Financial services
- Energy and utilities
- Health care
- Insurance
- Cross industry

In addition to these enterprises, a separate set of business scopes are available. Whereas an enterprise vocabulary might contain tens of thousands of terms, these business scopes are much smaller (generally, 200 - 500 terms), and are organized around specific business problems; for example, personal data, credit card data, and clinical order imaging.

For more information, see this Cloud Pak for Data IBM Documentation web page.

In this use case, we use the Banking Customers business scope (from the Knowledge Accelerator for Financial Services model). This business scope includes the glossary artifacts that we use to form the scope of our data governance and privacy initiative.

With the 4.5 release of Cloud Pak for Data, a project administrator can import each Knowledge Accelerators Business Scope into Watson Knowledge Catalog by using an API call. For more information, see this IBM Documentation web page.

At this web page, follow the instructions to install the Banking Customers business scope into Cloud Pak for Data.

The Knowledge Accelerators provide a reference for the glossary artifacts that we include in our scope. As an enterprise, we might want to customize this scope by updating the glossary artifacts in the scope or by creating glossary artifacts. For the new glossary artifacts, we use the test category that was set up earlier.

Next, we describe how the workflow for governance artifacts can be configured for the test category.

3.2.2 Defining the governance workflows

Managing governance artifacts occurs in a production environment. A single instance of Watson Knowledge Catalog is used to process the artifacts from a draft state to a published state. The process of moving the governance artifacts through these states is controlled by the Watson Knowledge Catalog workflow capabilities.

A set of predefined governance artifact workflows is included with the platform. In addition, custom workflows can be created by using Flowable products. These custom workflows can be uploaded to the platform.

By managing the workflow configurations for governance artifacts, you can control which workflow configurations apply to which triggers, and which people work on each step of the workflow.

To manage the workflow permissions, the Cloud Pak for Data user must be granted the Manage workflows permission, which can be assigned by an authorized user by clicking **Administration** → **Access Control**.

Consider the following points:

When you create a workflow configuration, you choose a template that contains a set of steps. You assign one or more users to each step. Each step is associated with a task, unless a step is skipped.

One of the assignees must complete the task to continue the workflow.

The default workflow configuration for all governance artifacts that are subject to workflows (named Default) allows users to publish artifacts without generating any tasks.

Complete the following steps to define the governance workflow:

1. Access the workflow page from the navigation menu by selecting **Workflows** in the Administration section. The Workflows window opens, as shown in Figure 3-12 on page 124.

   The page has two tabs:

   – The Task status tab provides details about the active tasks that were generated by the governance artifact workflows.

   – The Workflow types tab provides a tile that is named Governance artifact management with which the workflows are configured for the governance artifacts.

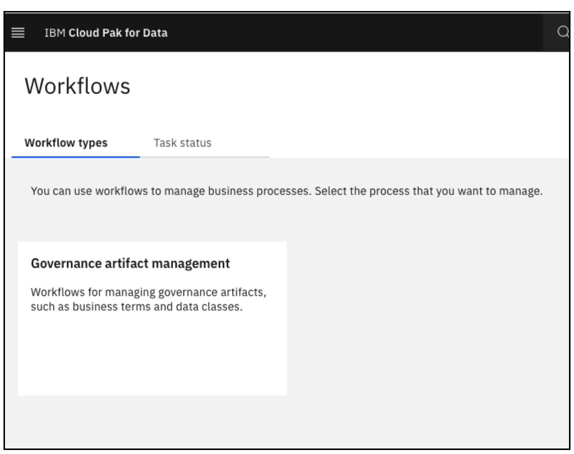

Figure 3-12 Workflows window

2. Click this tile to get to the Governance artifact management window (see Figure 3-13). This window includes three tabs:
   - Overview: Provides an overview of the artifacts that are used to trigger a workflow and the templates that are used for the workflows.
   - Workflow configurations: Allows administrators to configure the type of workflow and the categories that trigger a workflow for a specific glossary artifact.
   - Template files: Provides the flowable templates that are used for instances of the workflows. A standard set of templates are delivered with the platform. A user can create their own workflows by using these templates, but this process is beyond the scope of this publication.

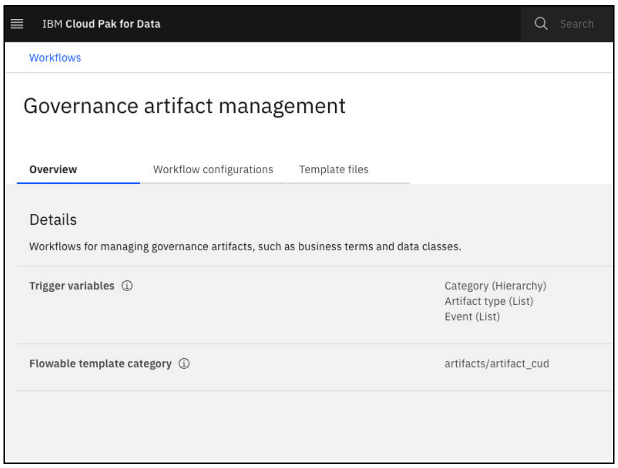

Figure 3-13 Governance artifact management window

The Workflow configurations tab (see Figure 3-14 on page 125) provides details about workflow configurations and allows users to add configurations.

A freshly installed instance of Watson Knowledge Catalog shows a single configuration, which is the Default workflow that is configured for all artifact types and all governance categories. This default workflow features a single step that automatically changes the state of an artifact to the published state with no tasks for the update that is being generated.

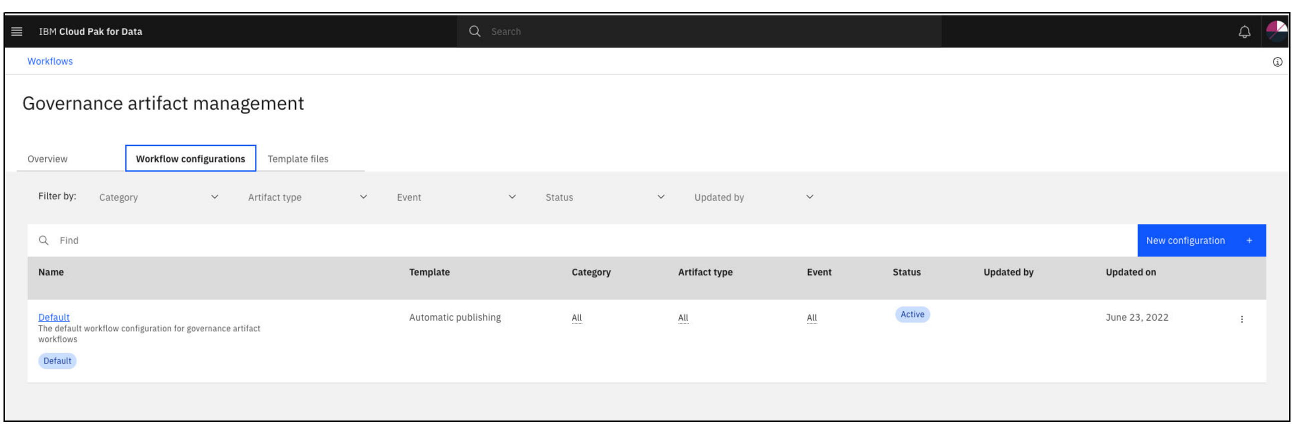

Figure 3-14 Workflow configuration

3. To create a workflow configuration from the available templates, click **New configuration** in the upper right of the window. The New workflow configuration window opens.

   A new workflow is configured that consists of two approval steps and a review step. Select the suitable process. Enter a name for the new configuration (for our example, we select **Two Step Approval**. Click **Create**. The configuration flow is shown for the two approval step workflow (see Figure 3-15).

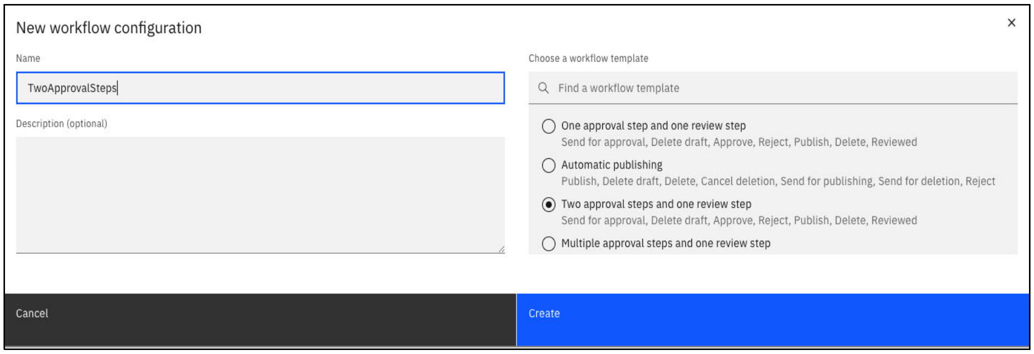

Figure 3-15 New TwoStepApproval workflow

4. The conditions that trigger the workflow must be created. Click **Add Condition** to add a condition to the workflow (see Figure 3-16).

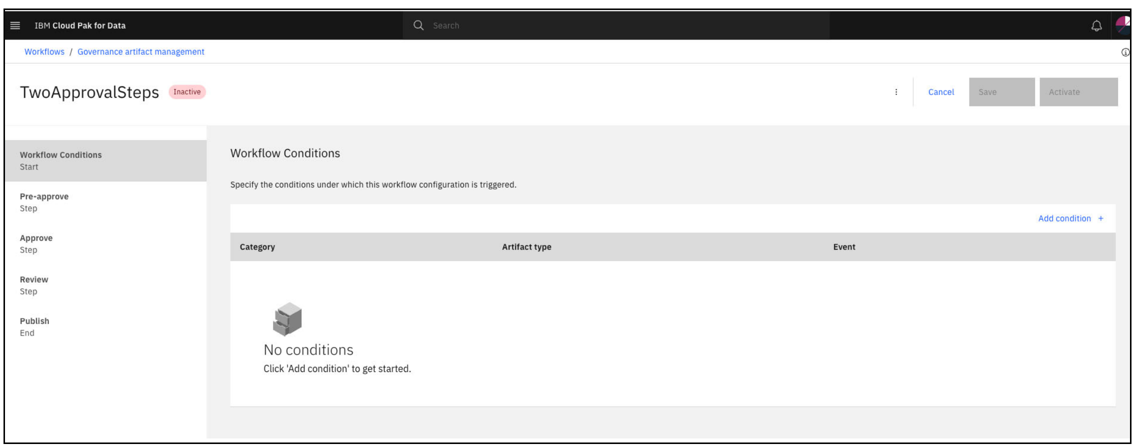

Figure 3-16 Adding trigger conditions to workflow

5. The conditions window is displayed. It is in this window that the workflow is configured. All categories or specific categories can be selected for the workflow. For our example, we select the Test category. The category can be selected as shown in Figure 3-17.

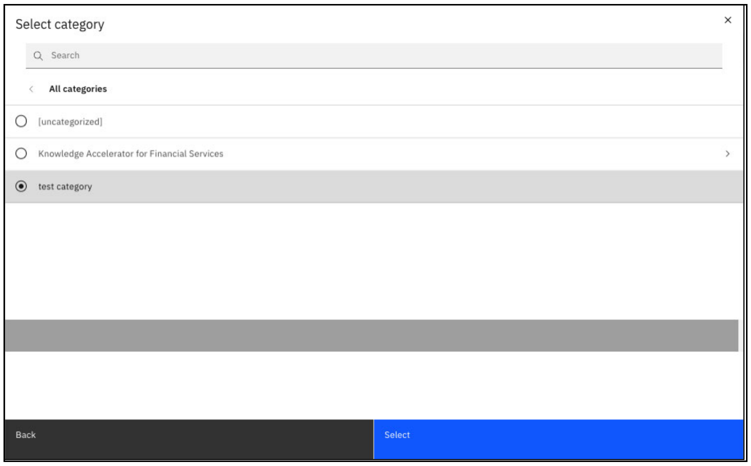

Figure 3-17 Select category for the workflow

6. The workflow can be configured to be triggered by creating, updating, deleting, or importing a specific governance artifact type. For our work flow example, we configure all governance artifact types and all triggers to start this new workflow.

   Clicking in each of the tiles replaces the default workflow by the workflow configuration that we are creating (see Figure 3-18). The fact that the new workflow is configured for a specific cell is indicated by the strike-through text.

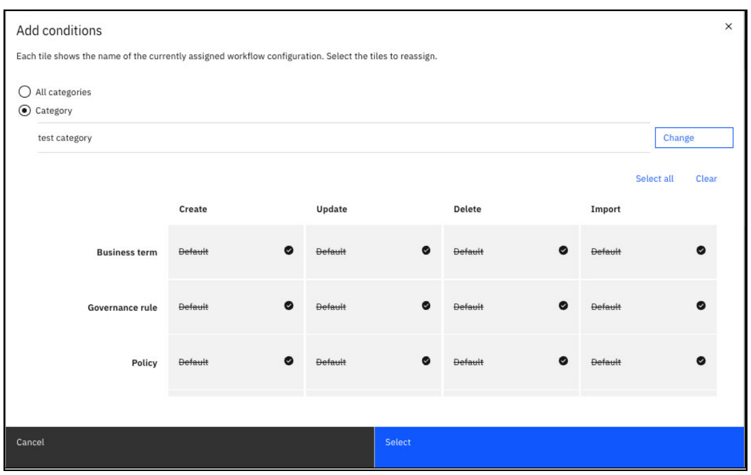

Figure 3-18 Triggers and glossary types for the two-step approval process

7. When all the changes are made for all the governance artifact types, click **Select**. A summary of the workflow conditions is shown (see Figure 3-19).

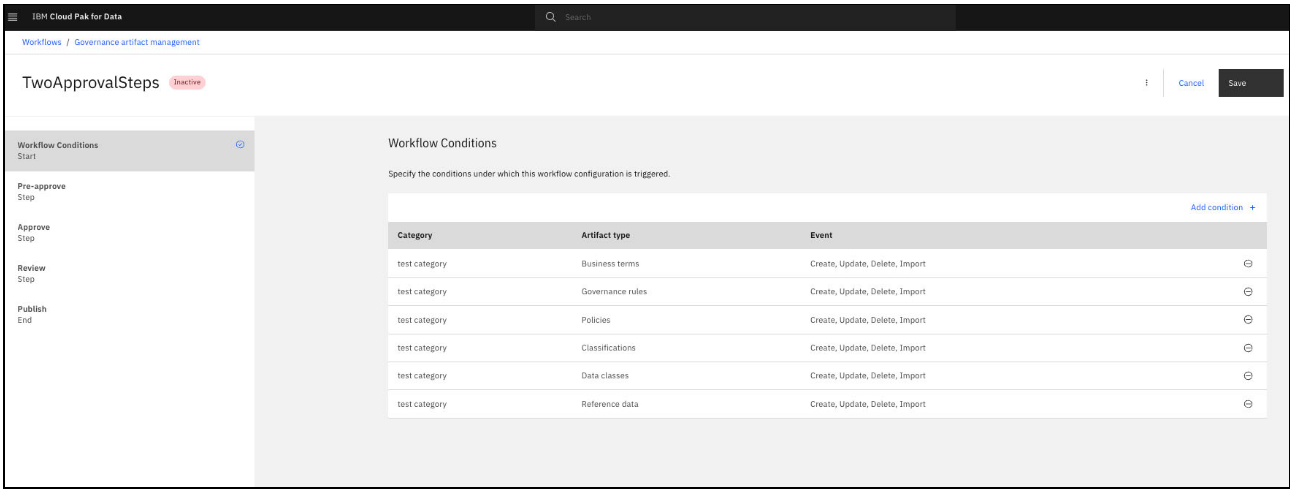

Figure 3-19 Workflow conditions summary

8. The user and groups that are used in the Pre-approve, Approve, Review, and Publish step of the workflow must be configured.

   For each step, we use the test stewards group that we created. We configure the assignee for the step to be that group and we configure that group to receive all notifications for each step. This process is an example and when implemented within a production environment, more thought must be taken for each step in the process.

   The pre-approval step is shown in Figure 3-20. When this step is completed, click **Save** and then, complete the remaining steps of the process by using the same information.

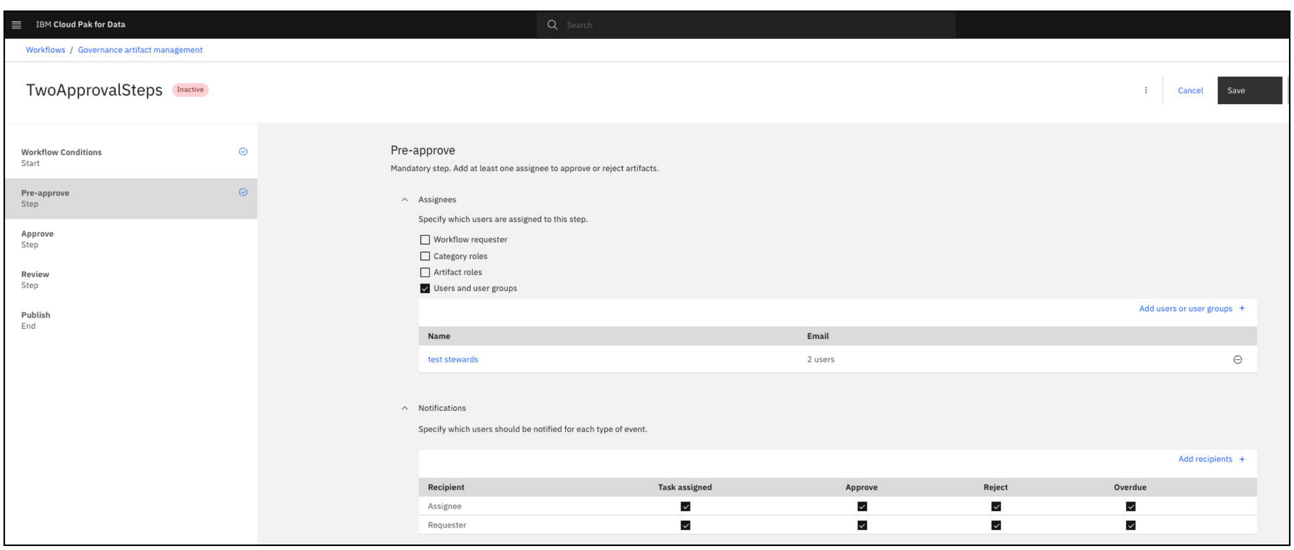

Figure 3-20 Configure pre-approval steps and notifications

9. After the final publish step is completed, click **Activate** to activate this workflow. A dialog box is displayed, as shown Figure 3-21.

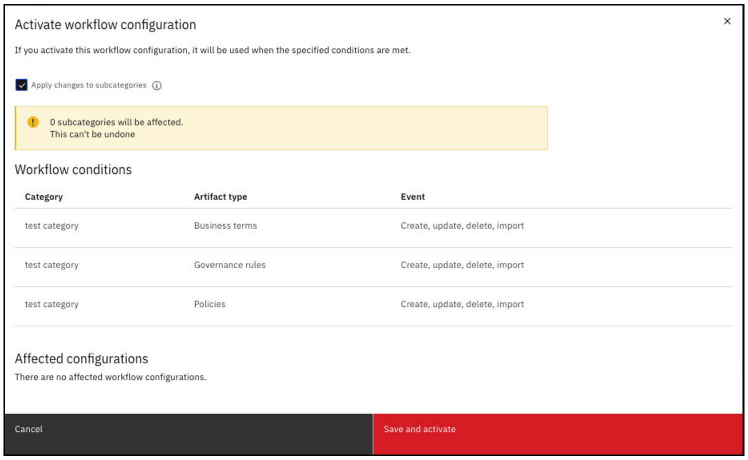

Figure 3-21 Activate workflow configuration

This dialog summarizes the changes that are to be made. Select **Apply changes** to see subcategories (Ignore the warning message. We only have a single top-level category.) Click **Save and Activate** to save the workflow and activate it.

Next, we describe how to create the initial set of governance artifacts that support our use case.

3.2.3 Creating initial governance artifacts

Creating any of the governance artifact types follows a similar process, which is governed by the workflow configuration that was created.

In this section, we show creating a single business term and data class to demonstrate the process. In a real-world situation, these tasks are repeated for each of the glossary artifacts that are in the scope of the initial governance initiative.

Now, create a business term in the test category. Because we configured a workflow for this category, we must complete several steps to get the business term into a published state that is visible to other platform users. (Although our single user can complete all of the steps, the approval, review, and publish steps likely are assigned to different users and groups within an organization in a real-world situation.)

Complete the following steps:

1. Log on to Cloud Pak for Data as a user that is a member of the data stewards group. From the navigation menu, select **Business terms** under the governance section. The list of published business terms is shown (see Figure 3-22).

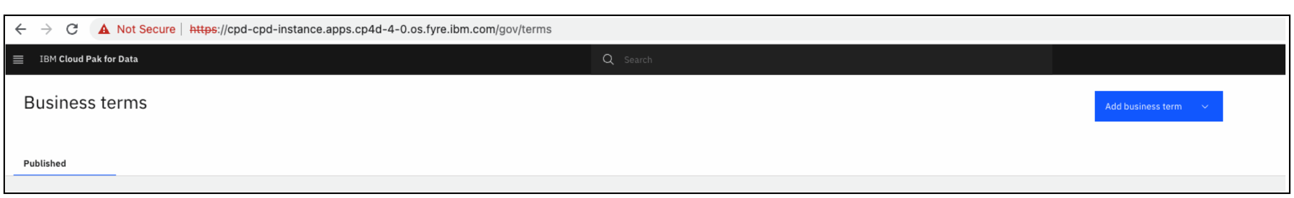

Figure 3-22 Business terms

2. Click **Add business term** to add a term, as shown in Figure 3-23. Ensure that the test category is selected as the primary category and then, click **Save as draft** to save the new business term.

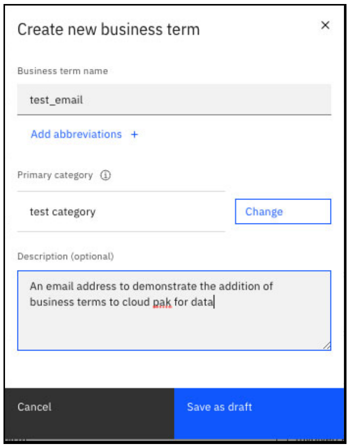

Figure 3-23 Creating draft of test email business term

3. Because the test category was configured with a two-step approval process, the draft term features a status of `Not started`, as shown in Figure 3-24. By clicking **Send for approval**, a dialog opens in which you can enter a comment for the approval request and a due date for the next step.

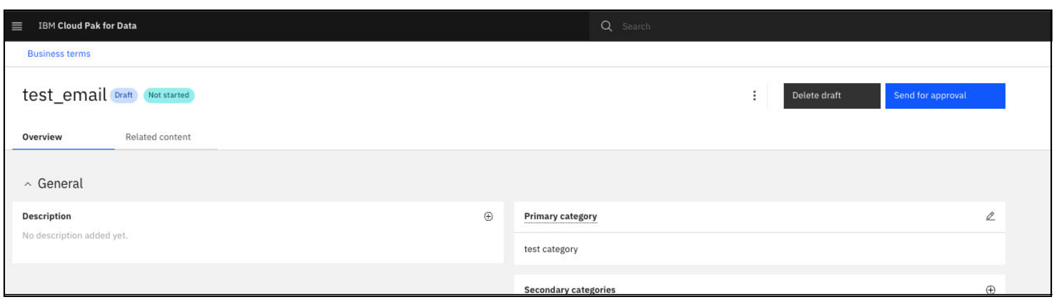

Figure 3-24 Draft test email term created

4. Earlier on, we set up the test stewards group so they were approvers, reviewers, and publishers, which means that the user you are logged as can process all the workflow tasks that are created. (You might have different groups with different users for each of these tasks.)

What follows shows that the workflow tasks are created. The term is now ready for approval. Click **Take action** → **Approve** (see Figure 3-25) to move to the next step of the workflow. Another dialog box is presented in which you can enter any comments and a deadline for the approval step.

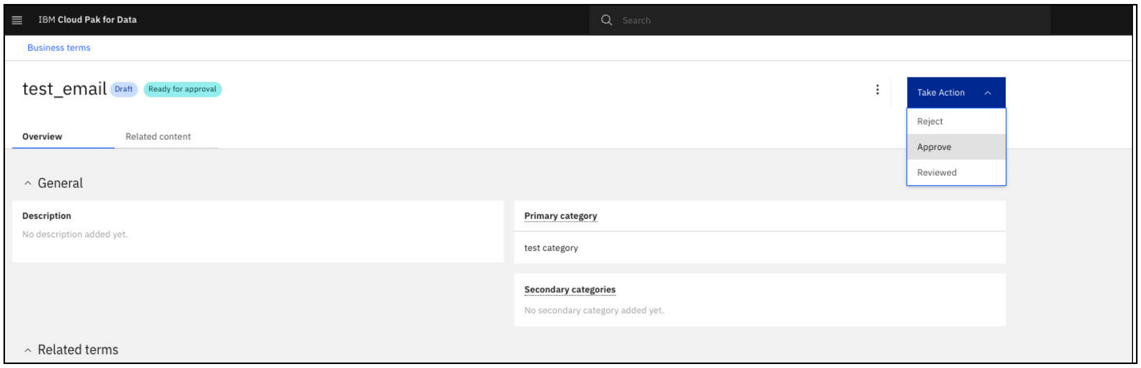

Figure 3-25 test email ready for approval

5. The term is now shown as ready for second approval. Click **Approve**, as shown in Figure 3-26. Enter any comments and a deadline for the second approval.

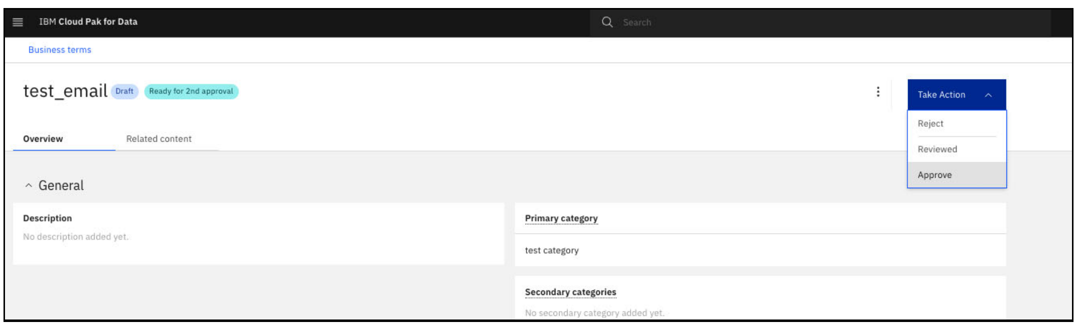

Figure 3-26 Test email ready for second approval

6. The term is now ready for publication, as shown in Figure 3-27. Click **Publish** to publish the term and enter any comments and a deadline for the task.

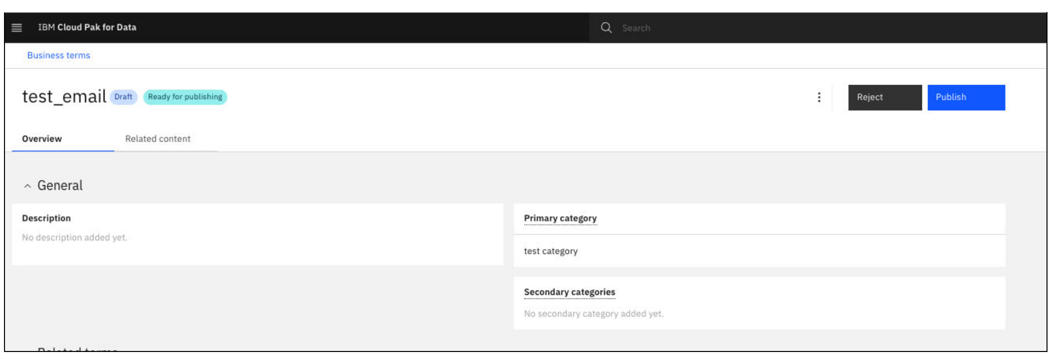

Figure 3-27 Test email ready for publish

Now, the term features a Published status, as shown in Figure 3-28. It is available for the metadata enrichment process that is described next.

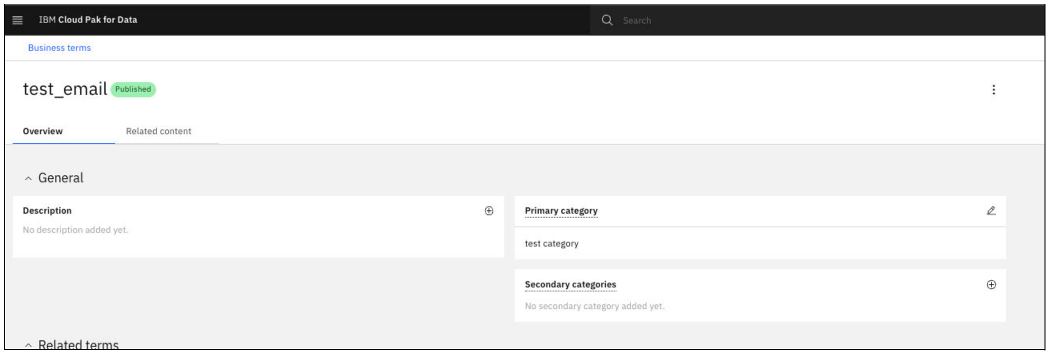

Figure 3-28 Test email published

7. At any point, the current set of tasks (see Figure 3-29) open and completed and assigned or requested by the logged on user can be seen in the task inbox (see Figure 3-29). The task inbox can be accessed by clicking the **navigation menu** → **Task** inbox. In addition, the home window provides a quick link to this inbox.

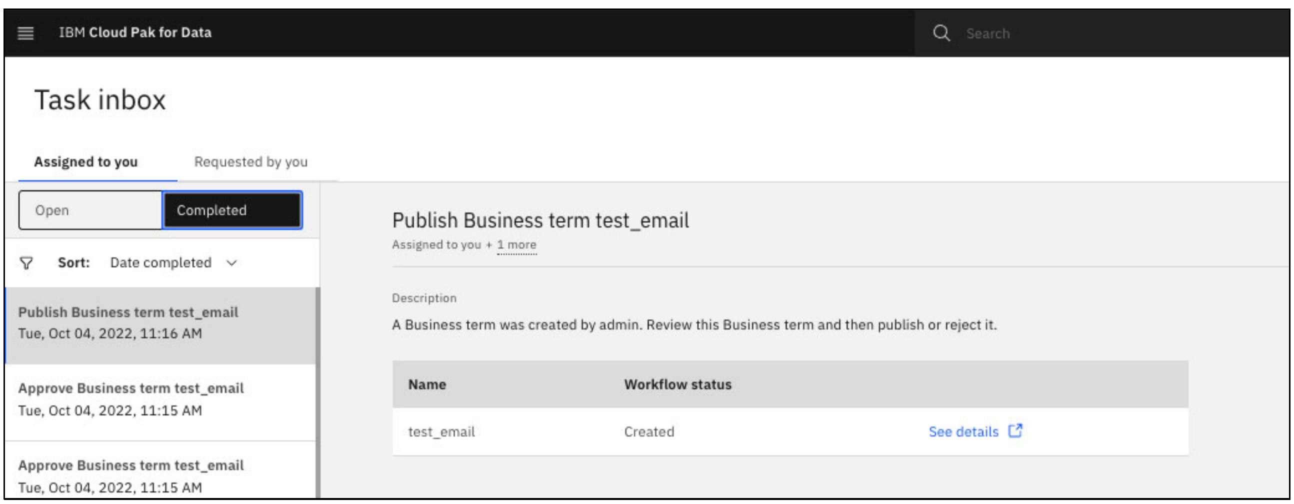

Figure 3-29 Task inbox

8. A key glossary artifact that is used when enriching the technical metadata is the data class. The data class is used by the metadata enrichment process to determine whether the contents of a data asset match specific criteria. If a match is found, the data class is assigned to the data asset.

   The data class can contain reference to business terms and classifications, but this linkage is not made immediately. If this linkage is made and if a data class is triggered by the enrichment process, the associated classifications and terms are added to the data asset.

   The data classes that are included with the platform are found in the `[uncategorized]` category. Log off from the Cloud Pak for Data platform and log in as your admin user because this user has permissions to edit the artifacts in this category.

From the navigation menu, select the **Data classes** menu item. Search for the Email Address data class, as shown in Figure 3-30.

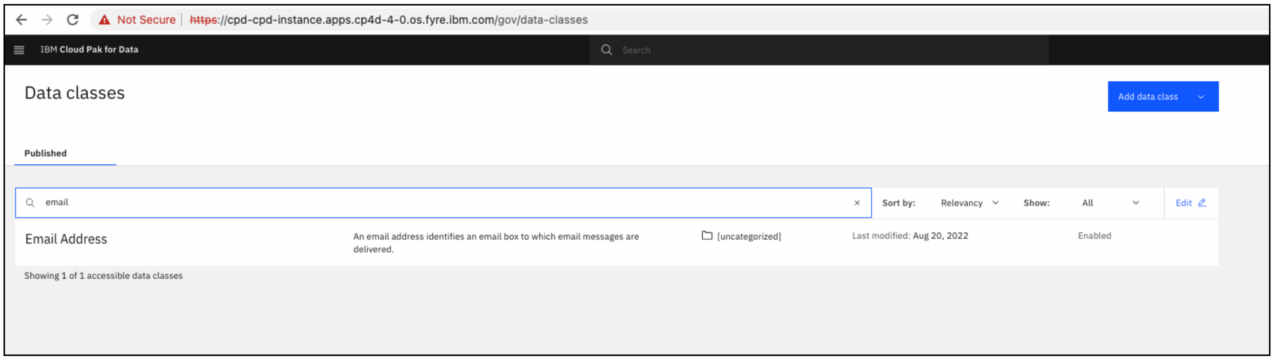

Figure 3-30 Email address for data class

9. Click the **+** sign that is next to the business terms section to add a business term to the data class, as shown in Figure 3-31.

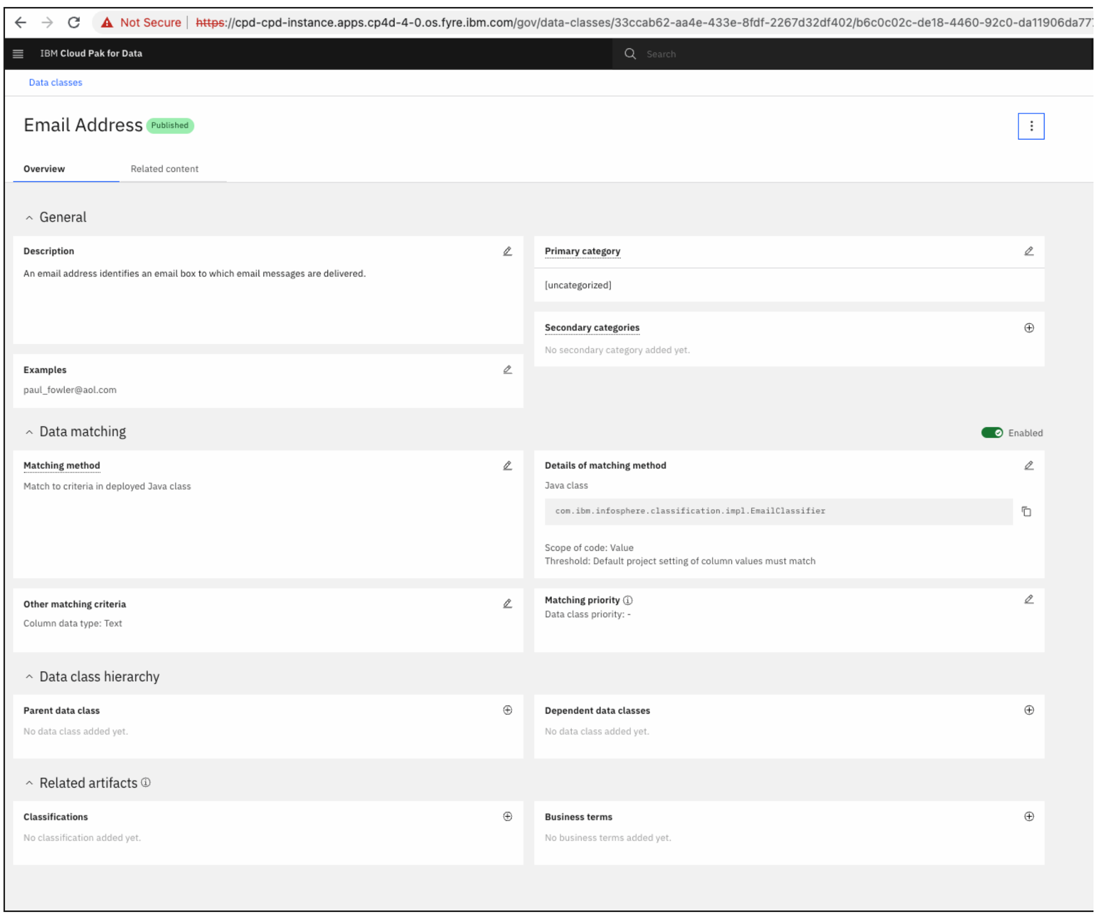

Figure 3-31 Email address data class details

10. Search for the test email business term. Select the term and add it to the data class, as shown in Figure 3-32. The data class is published directly because only the default workflow is configured for the `[uncategorized]` category.

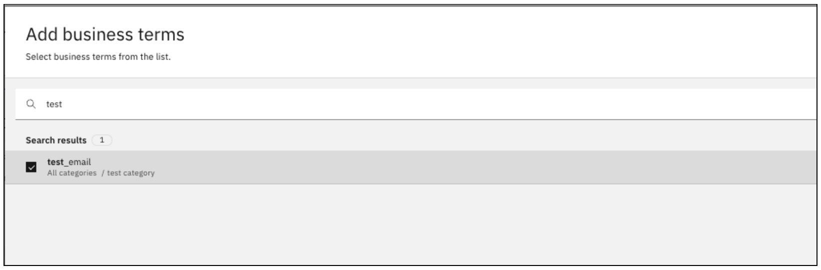

Figure 3-32 Adding the test email term to the data class

These two glossary artifacts are used in the next section to enrich the data assets that we import into Cloud Pak for Data.

3.3 Curating and managing data assets

The key components of Watson Knowledge Catalog that are used for governing IT assets are catalogs and projects. Work on the assets, such as importing the metadata and enriching the metadata, occurs in projects. This metadata is then published to one or more catalogs for consumption by downstream users of the platform.

For the rest of this use case, we use a project (called “test project”) for performing the curation work. We also publish the output to the “test catalog” for use by the platform users.

The project is created by clicking the **navigation menu** → **All projects** → **Click on new project** and the, clicking the **Create empty project** tile and entering a name. We use `test project` for the name and then, click **Create**.

The assets that are published are created from the **navigation menu** → **All catalogs -** → **Create catalog**. The parameters for the catalog are completed as shown in Figure 3-33. Enforce data protection rules were selected and we use the capability in the data privacy and data protection section later. We do *not* want duplicates in the catalog (assets that are published to the catalog update any assets).

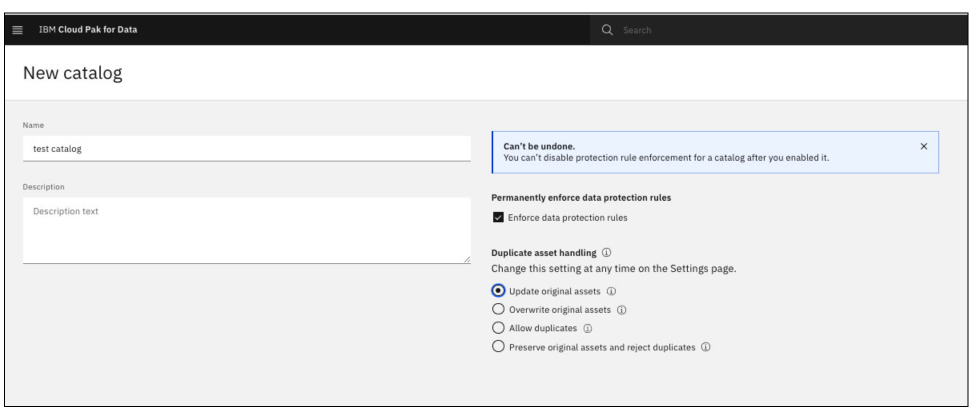

Figure 3-33 Creating the test catalog

Before we start, we need databases that contain data that we can use. A sample set of banking data is available in the GitHub repository.

A server also is needed that is accessible from your Cloud Pak for Data cluster. The server must have Docker or Podman installed. The IP address or DNS name of this server is needed when we create a connection to this data.

To set up the sample data, clone this GitHub repository to the machine that you use to host the samples database by using the following command:

```
git clone
https://github.com/IBMRedbooks/SG248522-Hands-on-with-IBM-Cloud-Pak-for-Data.git
```

To set up the samples database, change to the `usecase-data-governance-and-privacy` directory of the downloaded repository. Ensure that the `sample-data.sh` script is executable and then, run the script.

The script downloads the Db2 image from the Docker registry and starts it. A Db2 instance is created and populated with sample data. The new instance is set up to listen on port 50000 by default.

We then create a specific user for connecting to this database. The `configure-db-user.sh` script creates this user. Again, ensure that the script is executable and then, run the script. The script adds a test user to the Db2 instance with a default password of `password`. These user credentials are used in the rest of this chapter for connecting to the database.

To use this sample data, we add the connection to the Platform Connections catalog that is available from the navigation menu by clicking **Add connection** and then, selecting **IBM Db2** as the connection type (see Figure 3-34).

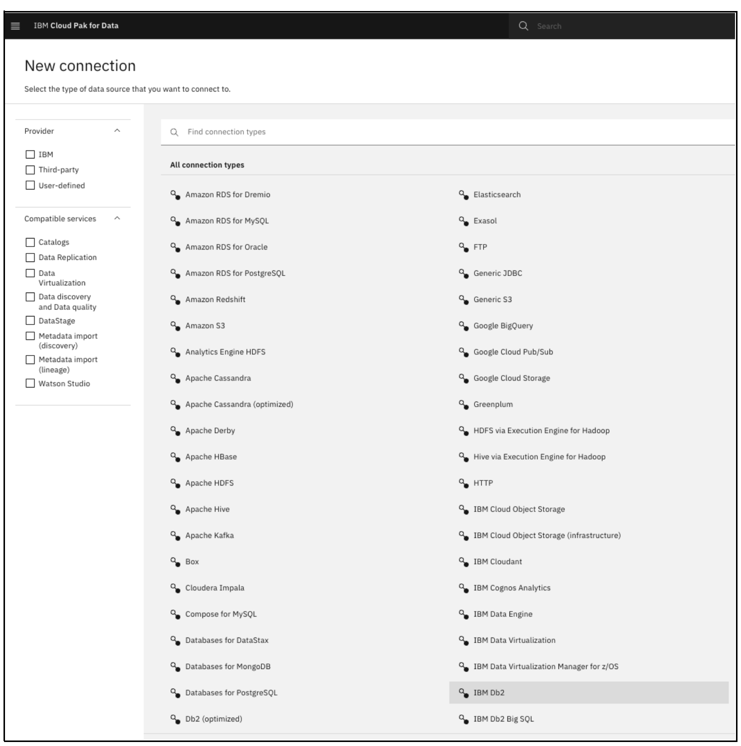

Figure 3-34 Selecting the connection type for the banking database

11.Enter the details of the banking database as shown in Figure 3-35. The hostname is set to the IP address of the host that is running the newly created database.

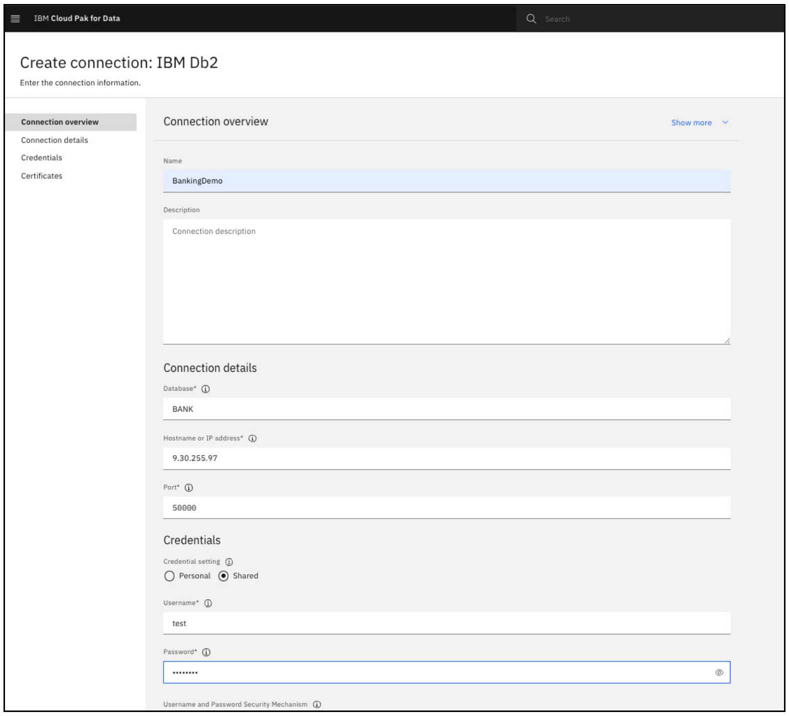

Figure 3-35 Enter the details for the banking demo database

12.When complete, click **Test connection** to see whether the connection is valid. When the validity is confirmed, click **Create** to create the connection. The connection is added to the platform assets catalog (we use it to demonstrate several capabilities).

Initially, we want to work with the connection to import metadata and work on enriching the metadata. This work occurs in the context of a project.

From the navigation menu, click **All projects** and then, select the test project that we created earlier. From the project, click the **Assets** tab and then, click **Add new asset**. Click the **Connection** tile. When the New Connection window is shown, click the **From platform** tab (see Figure 3-36). Select the **BankingDemo connection** and then, click **Create** to add the connection to the project.

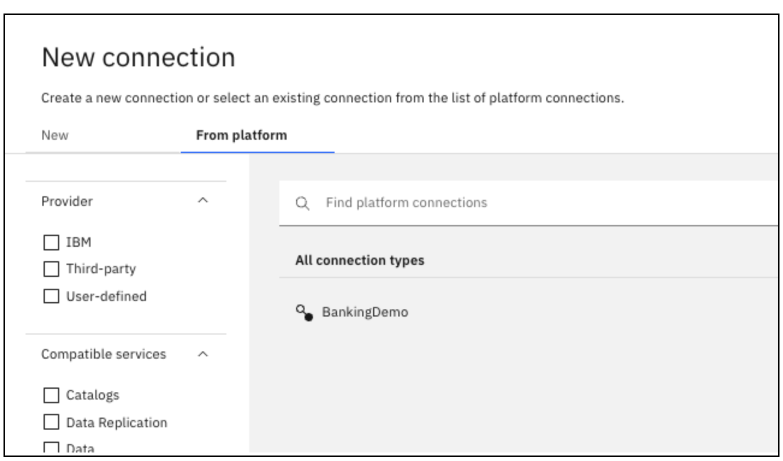

Figure 3-36 Add the BankingDemo connection to the project

In the next section, we describe how this connection can be used to bring metadata into the platform.

3.3.1 Importing metadata

Importing metadata is the recommended way to bring asset metadata into the Watson Knowledge Catalog. Complete the following steps:

1. To import new metadata into the project, click **New Asset** to create an asset within the project. The list of available asset types is shown.
2. Click the **Metadata import** tile to start creating metadata import for the test project (see Figure 3-37).

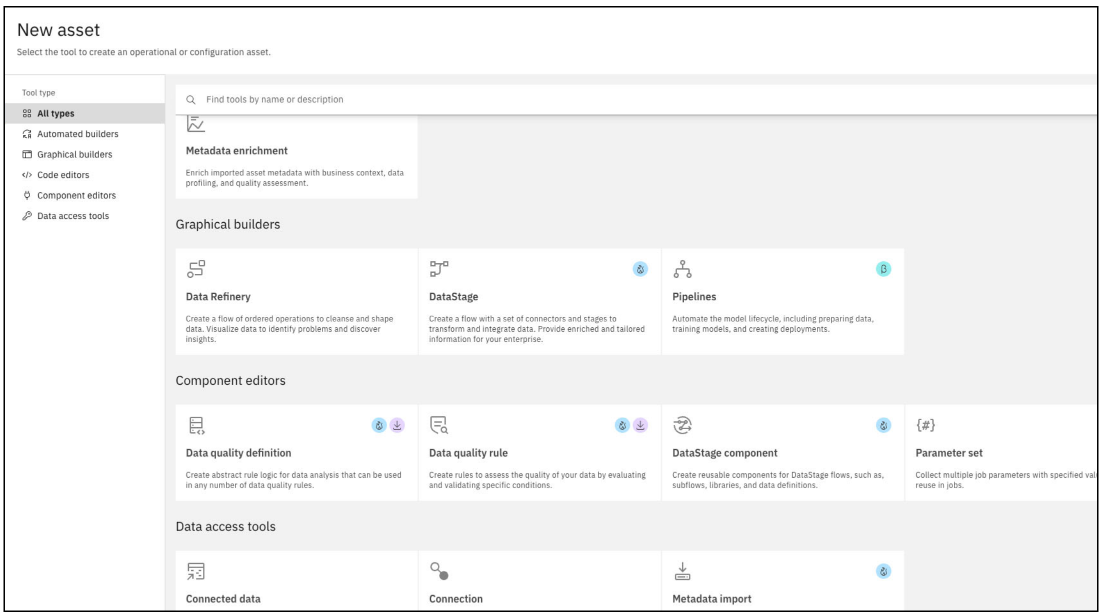

Figure 3-37 Starting a metadata import

3. A dialog window appears in which you are prompted about the goal of this import (see Figure 3-38). For now, we use the Discover option (the Lineage option is discussed in 3.3.4, “Data lineage” on page 152).

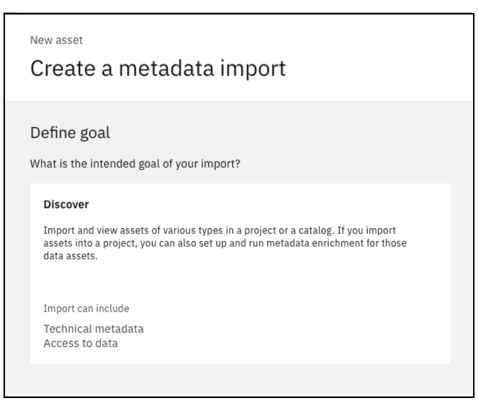

Figure 3-38 Set the goal of the import

4. Click the **Discover** tile to start the metadata import configuration. A set of dialog windows guide you through the configuration process. For the select target step, ensure that **test project** for this project is selected and then, click **Next**.

   For the scope, select the **BankingDemo** connection and the three BANK schemas, as shown in Figure 3-39.

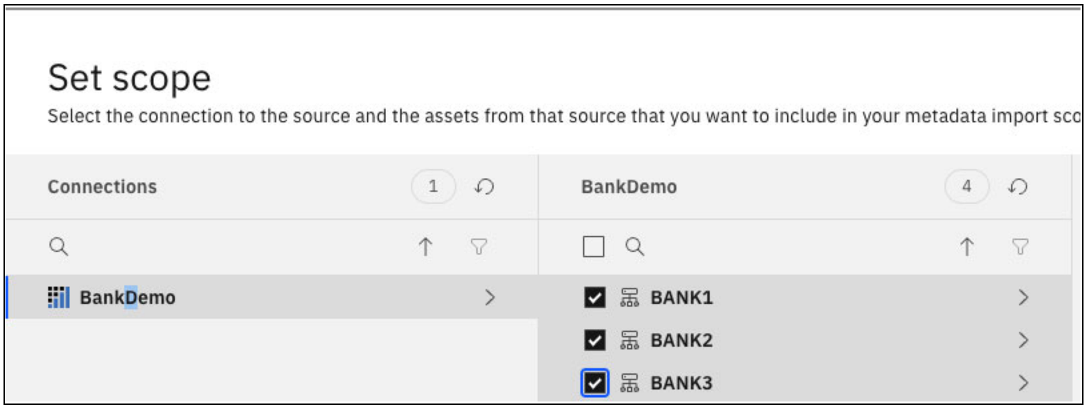

Figure 3-39 Setting the scope of the metadata import

5. Click **Next**. A summary of the import is shown. Click **Next**. Because we are conducting a one-off import, we do not need a schedule. Therefore, click **Next** to create the metadata import.

   The metadata import creates a job, which can be tracked from the Jobs tab of the project page. A job is shown with the same name as the metadata import. Clicking this job shows the job runs that are associated with the job.

   For our newly created metadata import, a single job run is created. Clicking the job run shows the logs for the run. Returning to the assets page and clicking M**etadata import** also shows the status of the import process, but without the logging information.

   The import shows as finished when the job completes successfully, as shown in Figure 3-40.

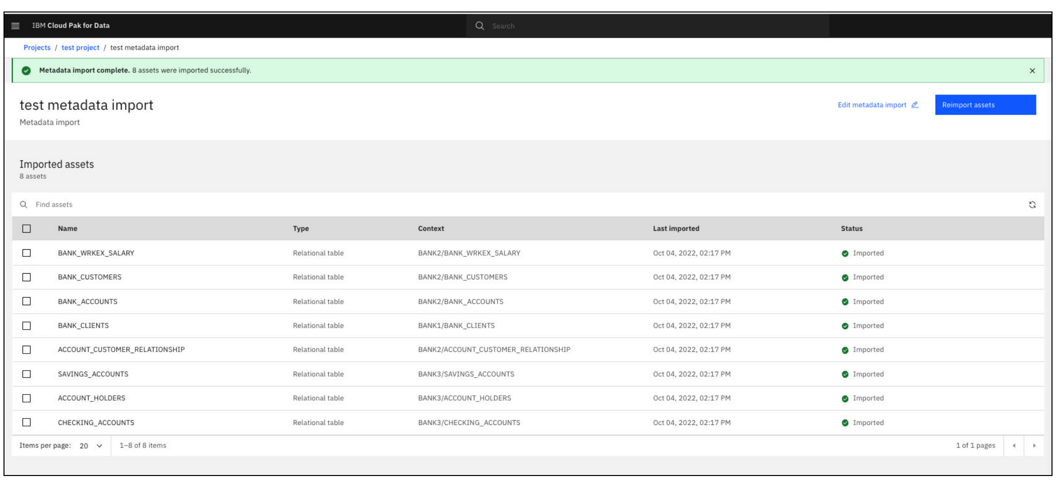

Figure 3-40 Completed metadata import

We can publish this metadata to our catalog for consumption downstream, but only the technical metadata is available now. Therefore, we must enrich this data with the glossary artifacts that we created. To do this, we use the metadata enrichment capability that is described next.

3.3.2 Metadata enrichment

Data is useful only if its context, content, and quality are trusted. To keep it that way, data must continuously be evaluated, and suitable remediation be taken, if required.

Metadata enrichment is a capability of Watson Knowledge Catalog that provides this extra context. Metadata Enrichment uses several techniques to automate the enrichment of information assets with the glossary artifacts that we discussed in 3.2, “Establishing the governance foundation” on page 119.

This automated enrichment is the first step in the curation process. The Metadata Enrichment can be seen as a work area that is created in the project for each distinct enrichment that is run. This work area is the place where data stewards make decisions as to the real nature of the enrichment based on the provided automated suggestions.

The information that is added to assets through metadata enrichment also helps to protect data because it can be used in data protection rules to mask data or to restrict access.

Complete the following steps enrich the metadata:

1. Select **New asset** from the Test Project assets tab. On the new assets page, select the **Metadata Enrichment** tile to start the process of configuring the metadata enrichment (see Figure 2 on page 140).

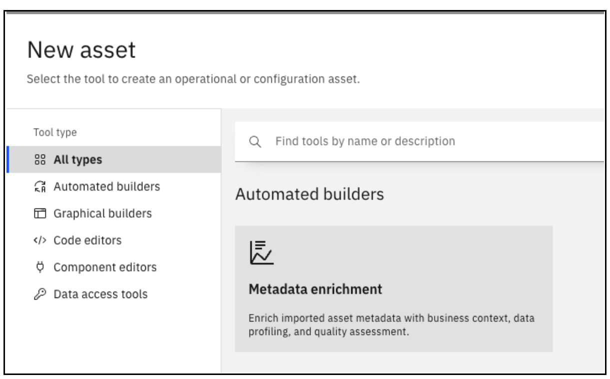

Figure 3-41 Selecting the metadata enrichment tile to start an enrichment process

2. Enter a name and description for the metadata enrichment, as shown in Figure 3-42.

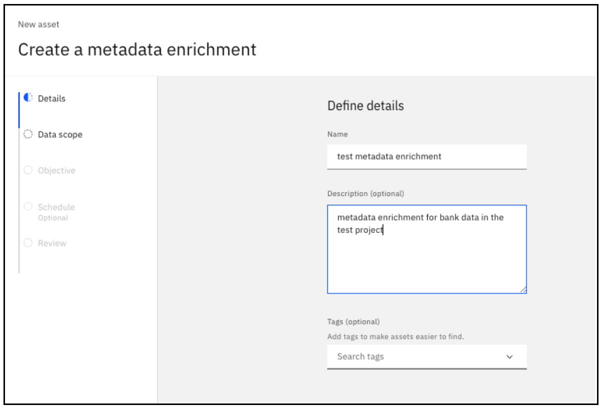

Figure 3-42 Providing details of the metadata enrichment

3. The scope for the enrichment must be specified. The scope can consist of the individual data assets that were imported by the metadata import or more simply by using the metadata import. For our example, we use the metadata import (See Figure 3-43).

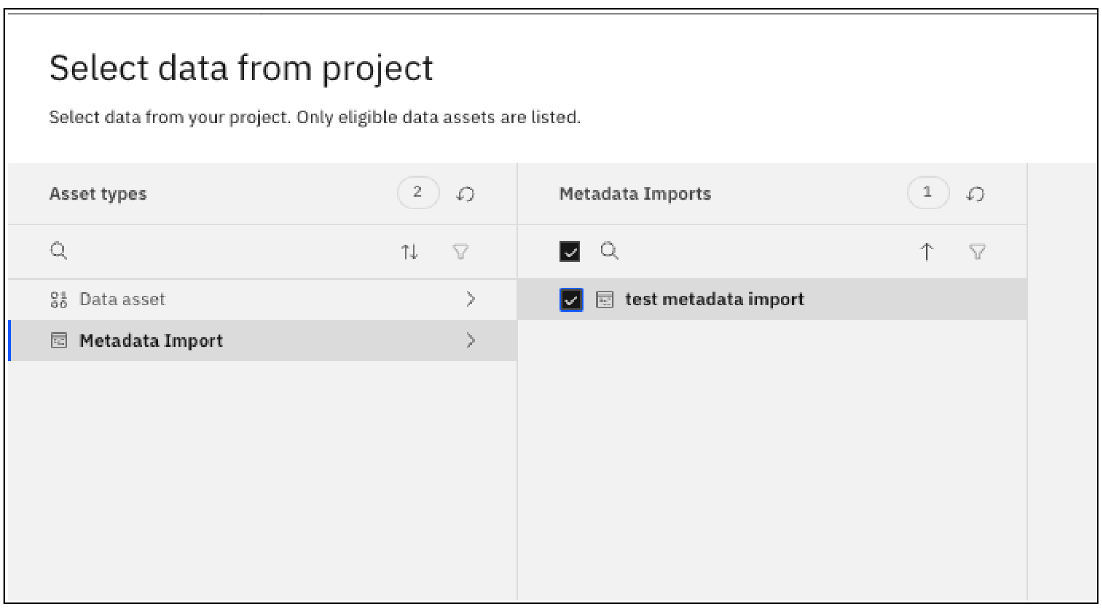

Figure 3-43 Selecting the scope for the metadata enrichment

4. Click **Select** to establish the scope. A window opens in which the scope is summarized. Click **Next** to open the metadata enrichment objective window, as shown in Figure 3-44.

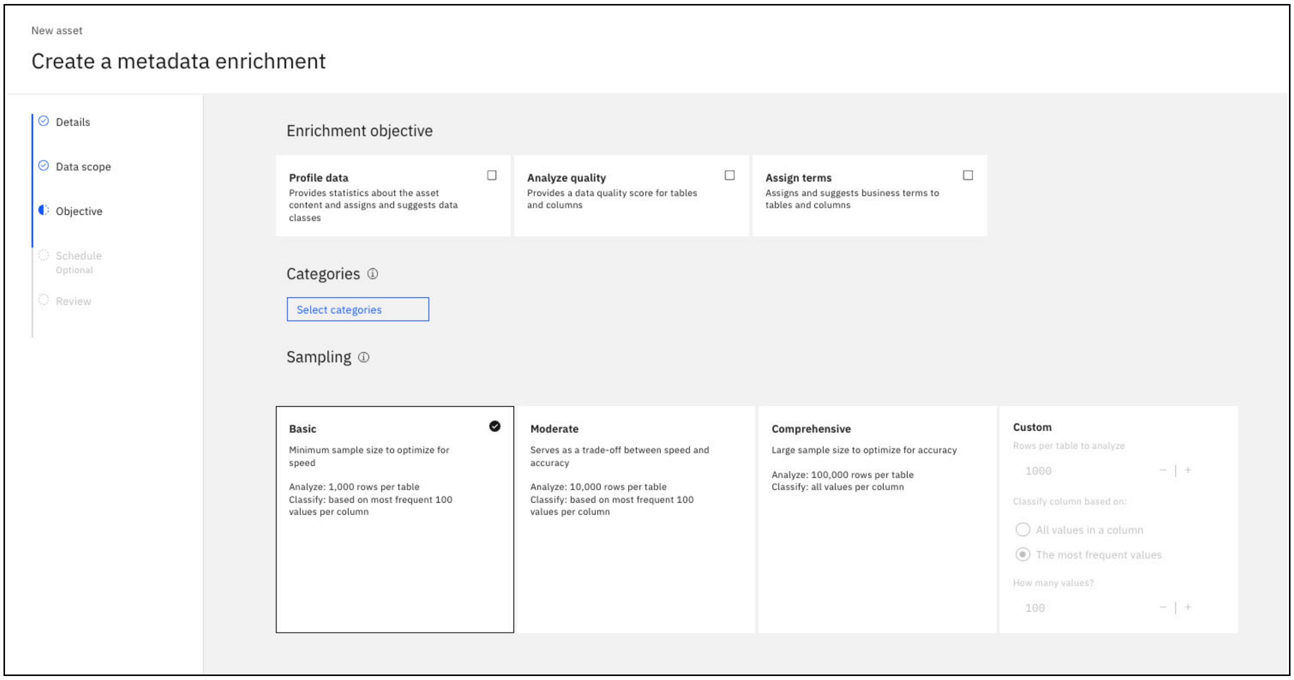

Figure 3-44 Configure metadata enrichment scope

In this window, the metadata enrichment is configured. The following key options are available, of which one or more can be selected for the enrichment process:

- Profile data: Provides statistics about the asset content and assigns and suggests data classes. This results in statistics for the data asset, such as the percentage of matching, mismatching, or missing data. The frequency distribution for all values is identified in a column and for each column, the minimum, maximum, and mean values and the number of unique values are displayed.

  In addition, a data classes area is assigned to describe the contents of the data in the column. For more information about data classes, see 3.2.3, “Creating initial governance artifacts” on page 129. For more information about data privacy, see 3.3.6, “Data privacy and data protection” on page 158.

- Analyze quality: Provides a data quality score for tables and columns. Data quality analysis can be done in combination with profiling only. Therefore, the Profile data option is automatically selected when you choose to analyze data quality. Data quality scores for individual columns in the data asset are computed based on quality dimensions. The overall quality score for the entire data asset is the average of the scores for all columns.

- Assign terms: Automatically assigns business terms to columns and tables, or suggests business terms for manual assignment. Those assignments or suggestions are generated by a set of services, which include linguistic matching, rule-based matching, and AI algorithms.

Depending on which term assignment services are active for your project, the term assignment might require profiling.

To scope the terms and data classes that are used in the enrichment processes, the categories that contain these governance artifacts must be selected. Only top-level categories can be selected, as shown in Figure 3-45.

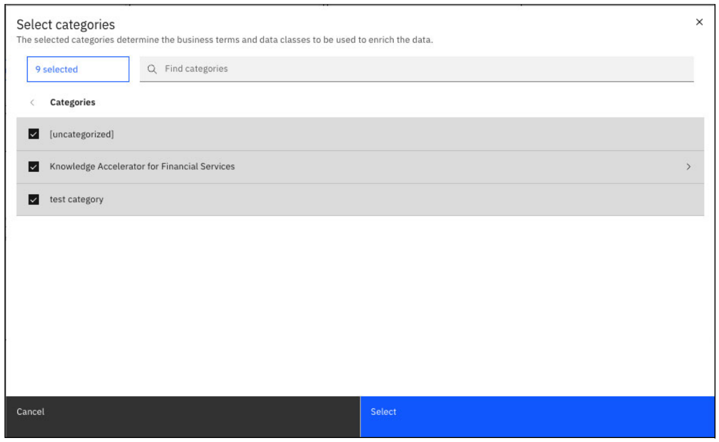

Figure 3-45 Selecting categories for enrichment process

5. Select the categories and click **Select**. The metadata enrichment page is updated with the categories that are in scope. Select the type of sampling that you want to use for the metadata enrichment job. The following options are available (see Figure 3-46 on page 143):
   - Basic: A total of 1,000 rows per table are analyzed. Classification is done based on the most frequent 100 values per column.
   - Moderate: A total of 10,000 rows per table are analyzed. Classification is done based on the most frequent 100 values per column.
   - Comprehensive: A total of 100,000 rows per table are analyzed. Classification takes all values per column into account.
   - Custom: Define your own approach.

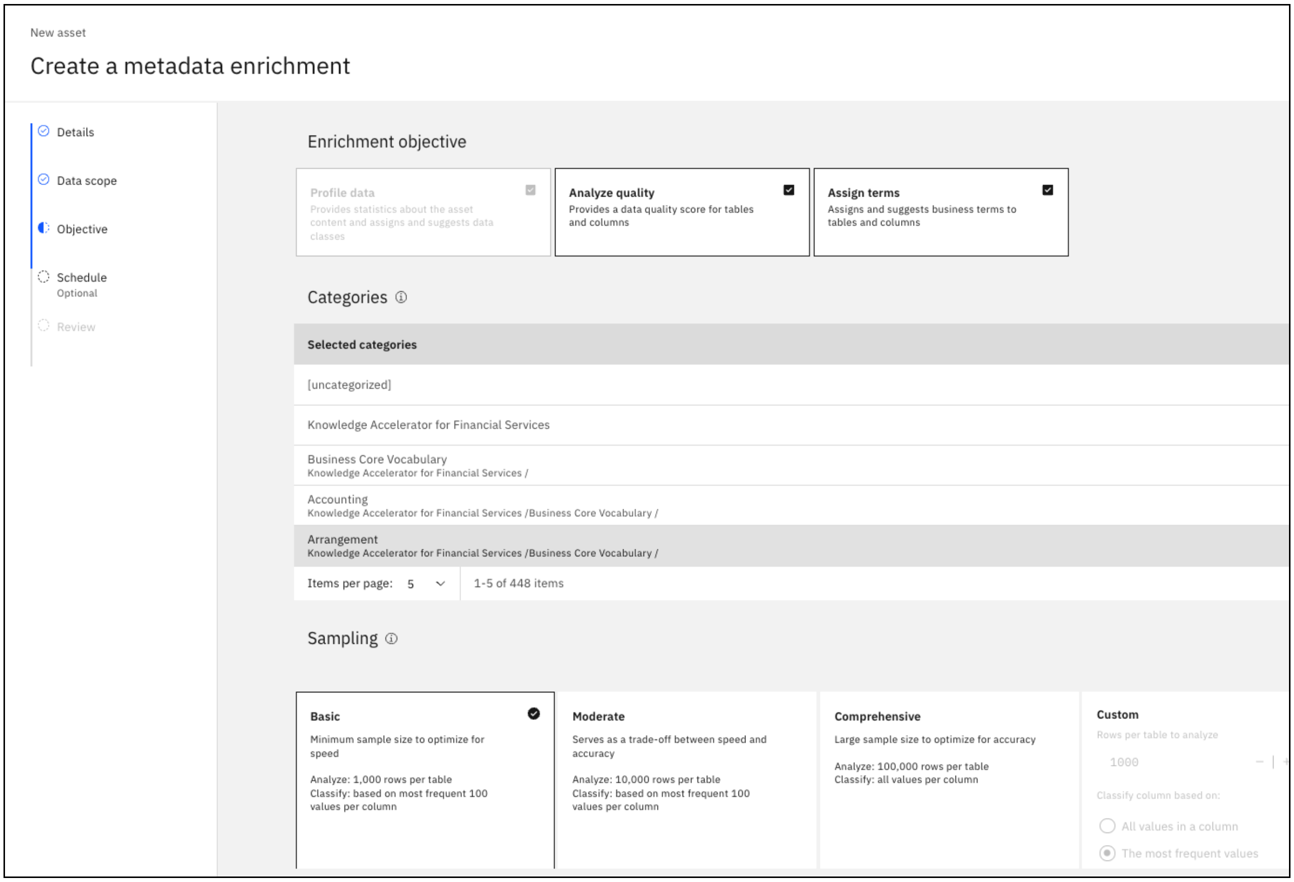

Figure 3-46 Metadata enrichment configuration

The more rows and values that are used, the more resources and time is needed for the metadata enrichment. For this use case, the Basic sampling is used.

6. After the configuration is complete, click **Create** to create the metadata enrichment.

   A job is created for the metadata enrichment that can be viewed in the Jobs tab of the test project. This view is used to track the progress of the enrichment. You also can remain on the Enrichment page and wait until the process it completes, which can take some time.

When the job is finished, the results are displayed (see Figure 3-47). The same job framework is used to create the metadata enrichment job and job run as was used for the metadata import. The jobs and runs can be accessed from the jobs tab of the project if more detailed information about the run is required.

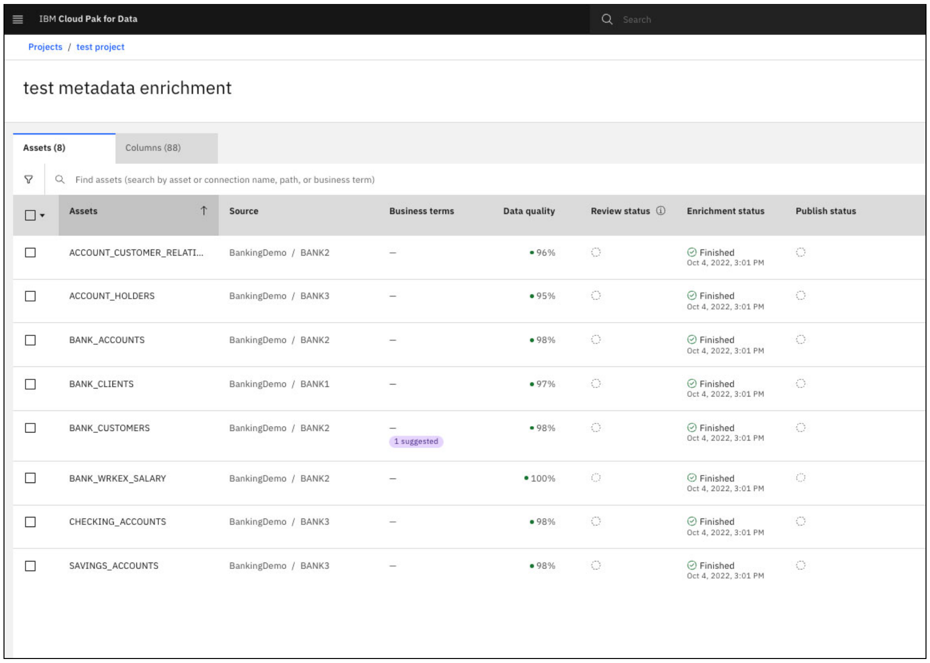

Figure 3-47 Metadata enrichment initial results

7. The Metadata Enrichment asset is the data stewards workspace. It allows a steward to curate data that is based on the recommendations from the enrichment process. In this workspace, the user also can work at the level of the assets or at the level of the columns within an asset. This second detailed level is where most of the work is done.

   The enrichment work area provides details about the terms and data classes that are assigned and provides links to review the quality and statistics of the data that is being enriched, as shown in Figure 3-48. When curation is complete, the steward can mark the assets as reviewed so that progress can be tracked.

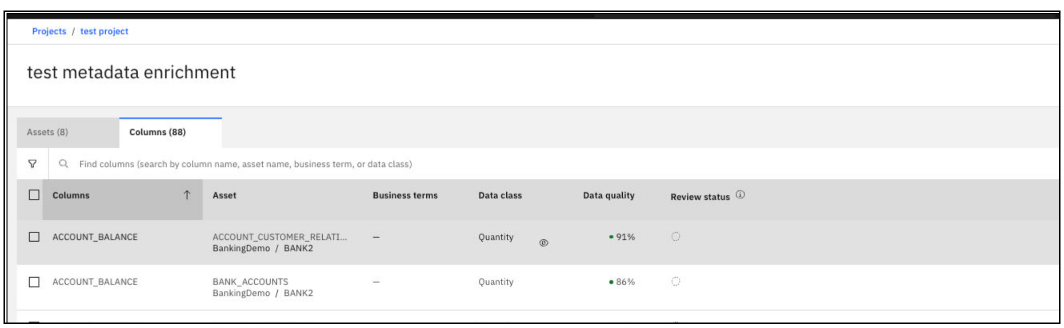

Figure 3-48 Metadata enrichment activities

8. The suggested business terms and data classes are just that: suggestions. It is the Data Steward who decides whether the suggestions are suitable. Edits of one or more assets to a catalog are possible for downstream use. Select the assets to publish and then, click those assets, as shown in Figure 3-49.

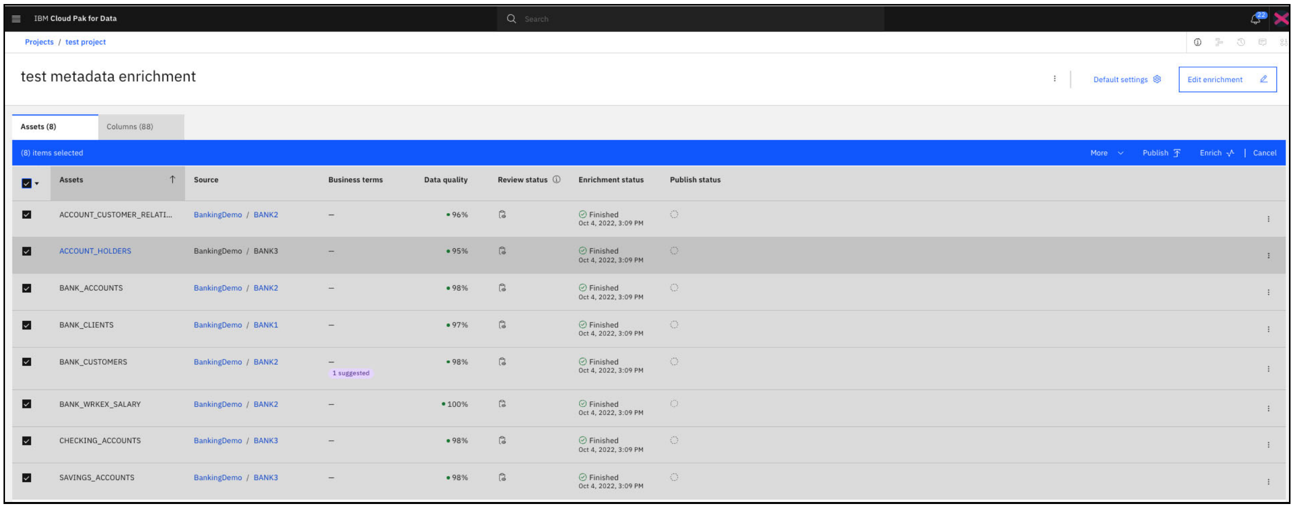

Figure 3-49 Metadata enrichment published to catalog

9. The list of catalogs that the user has permissions to write to is shown, as in Figure 3-50. Select the test catalog and then, click **Next**. A summary of the assets to be published is shown. Click **Publish** to publish the assets to the catalog.

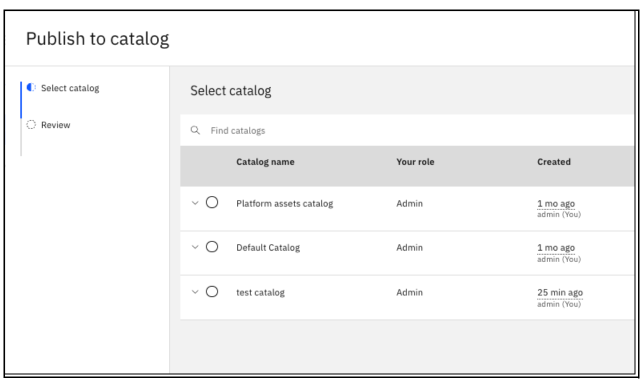

Figure 3-50 Selecting the catalog to which to publish

10. The results of the publish action can be seen in the catalog.

    From the navigation menu, select **Catalogs** and then, select the test catalog. From the test catalog, select the **BANK_CLIENTS** data asset. The Asset tab shows an overview of the asset, as shown in Figure 3-51 on page 146. The Email field of the asset includes the `Email Address` data class that is assigned to and is shown in the row that is under the column definition. Clicking **Yes** next to the name of the column also shows that the test email term was assigned to this column.

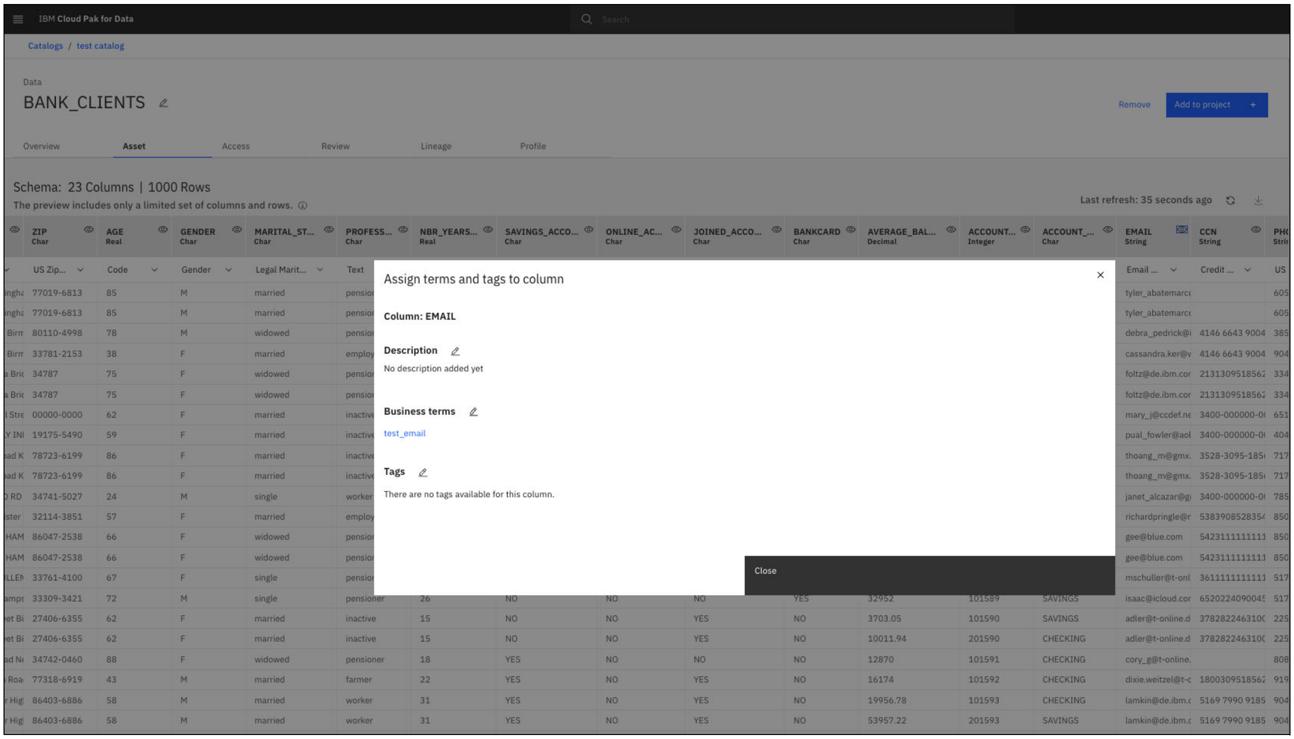

Figure 3-51 BANK_CLIENTS data asset with email address classification and term assigned

3.3.3 Data quality

Cloud Pak for Data version 4.5 introduced a new microservices-based data quality capability that is delivered with Watson Knowledge Catalog. By using this feature, you can measure, monitor, and maintain the quality of your data.

Note: The data quality feature is not installed by default when Watson Knowledge Catalog is installed. Therefore, an Red Hat OpenShift project Administrator must ensure that this feature is enabled before proceeding with the rest of this chapter. For more information about instructions for installing the data quality feature, see this IBM Documentation web page.

By using the data quality feature, a user with the Data quality analyst role can create rules and run data quality jobs. Before this part of the process is started, log in to the Cloud Pak for Data cluster as an administrator and assign the data quality analyst role to the test stewards group so that users in that group are enabled for quality management.

Note: With the release 4.5, the following permissions to the data quality analyst role were added:

- Access Data Asset Quality Types
- Access Data Quality

Therefore, edit the Data Quality role and add these permissions by clicking **Add permission**.

The following key components comprise the data quality feature:

- Data Rule Definitions: Define the logic for a rule. The rule definitions are created independent of any data asset
- Data Rules: Bind the rule defintions to specific data assets in preparation for analyzing data quality
- Data Quality Jobs: Run the Data Quality Rules to determine the quality of your data.

Complete the following steps:

1. Log on to the Cloud Pak for Data cluster as a member of the test stewards user group. Access the test project. From the Assets tab of the project, select **New asset** and then, select **Data quality definition** for the list of available assets, as shown in Figure 3-52.

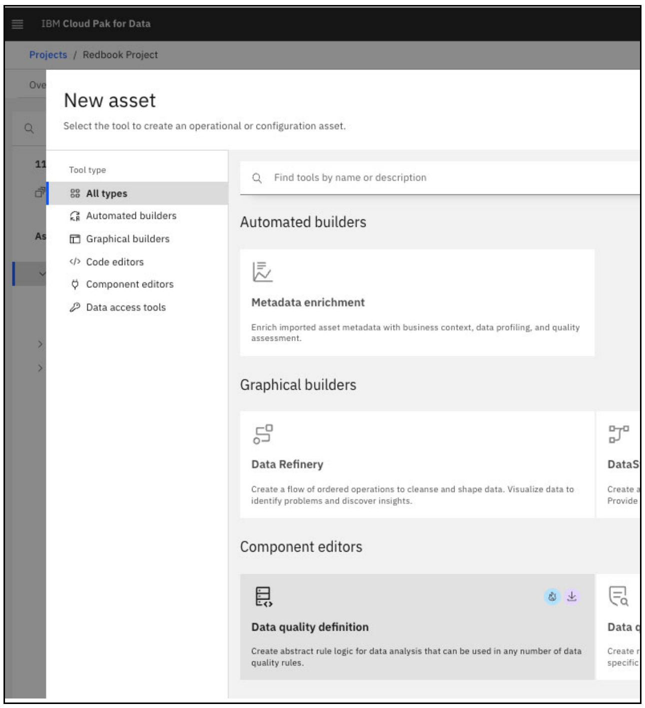

Figure 3-52 Creating a Data Quality Definition

a. A new window opens in which you enter the name for the rule and the rule logic (in this case, we create a simple rule, `Gender Complete`, that validates that a field-defining gender contains our organization's preferred set of values, as shown in Figure 3-53). The rule definition can be entered as text (the definition is validated as you type) or the graphical rule builder can be used to drag the rule definition elements.

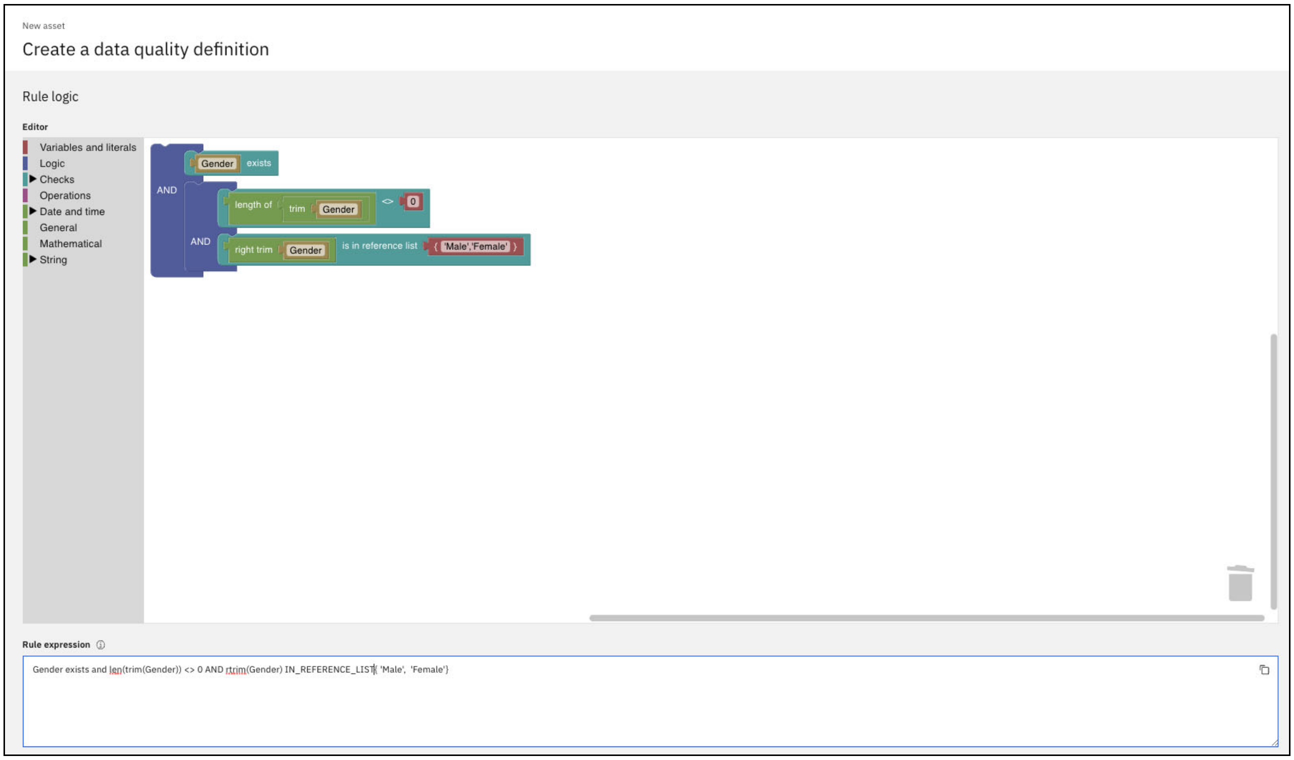

Figure 3-53 Logic for gender rule definition

2. When the rule logic is entered, click **Create** to create the rule definition. A summary of the rule is displayed (see Figure 3-54).

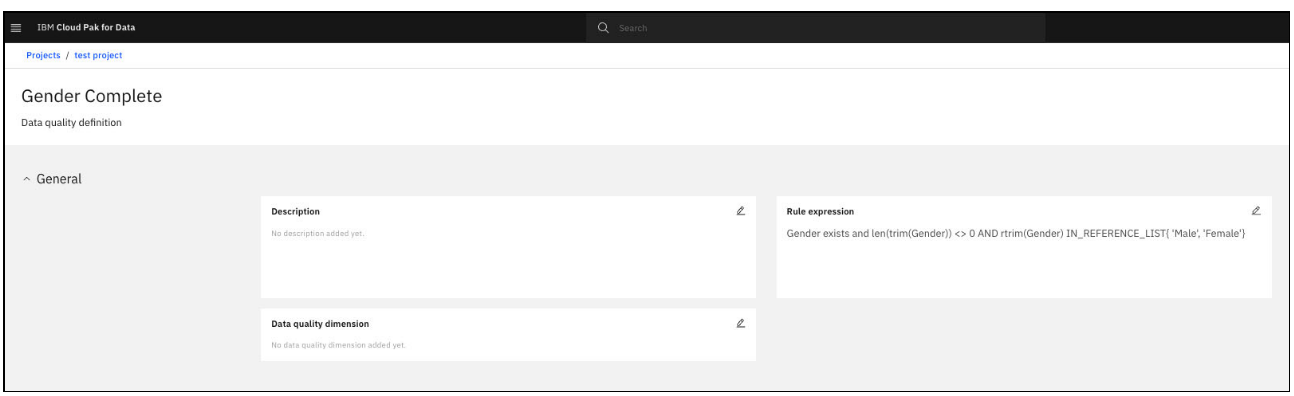

Figure 3-54 Summary of the gender complete rule definition

3. A Quality rule must be created that is based on the newly created rule definition. From the summary page, click **Create rule** in the upper right. A new dialog opens for creating the rule, as shown in Figure 3-55. Enter the name for the new Data Quality Rule and then, click **Next**.

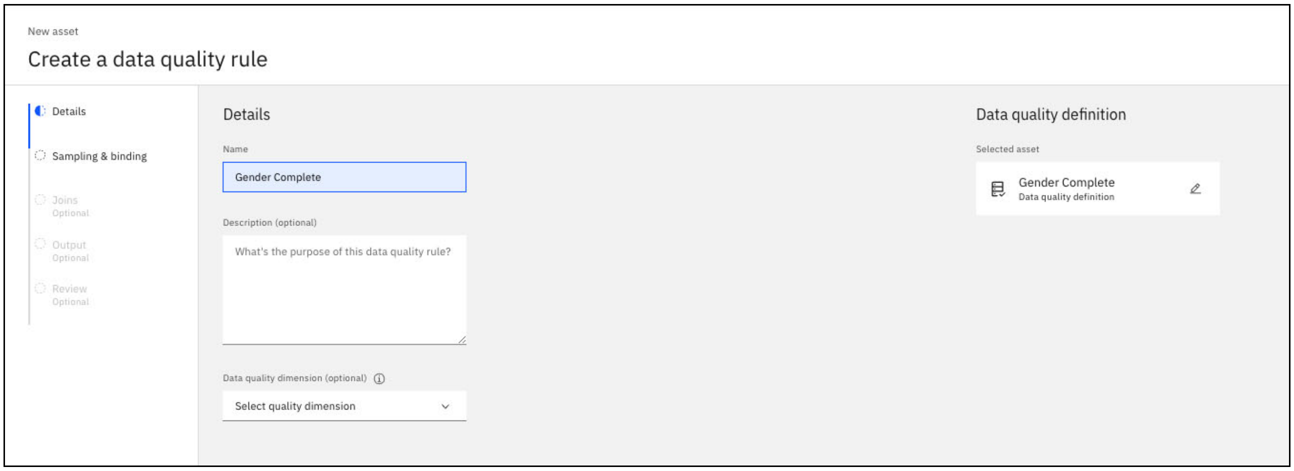

Figure 3-55 Data Quality Rule details

4. We use the default sampling method for the rule (which is sequential) starting at the top of the file and sampling the first 1000 rows. The rule must be bound to a column. In the bindings section, click **Select** in the Bind To column, as shown in Figure 3-56.

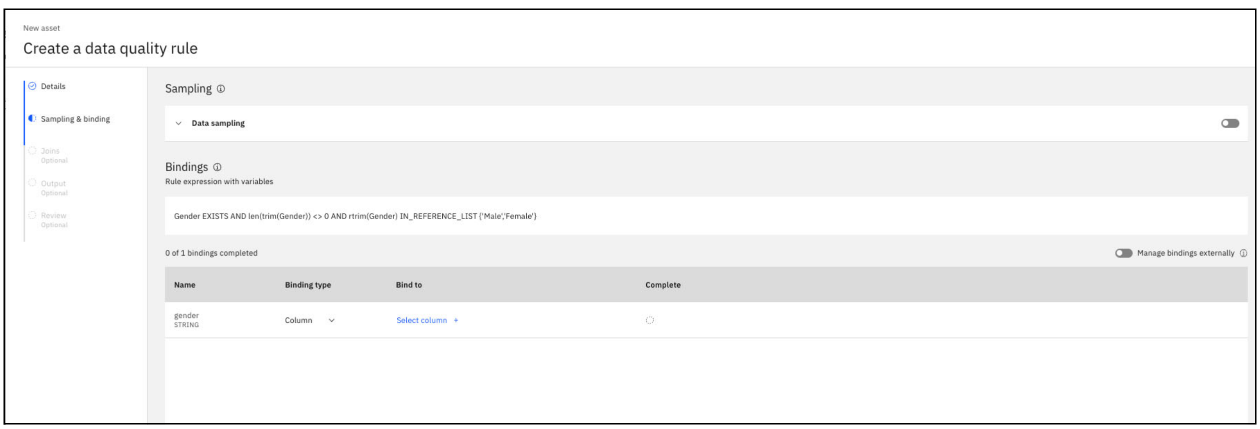

Figure 3-56 Data Quality Rule sampling and binding

5. A dialog opens in which you choose the column to which to bind. Select the **BANK_CUSTOMERS** data asset and then, select the **Gender** column, as shown in Figure 3-57.

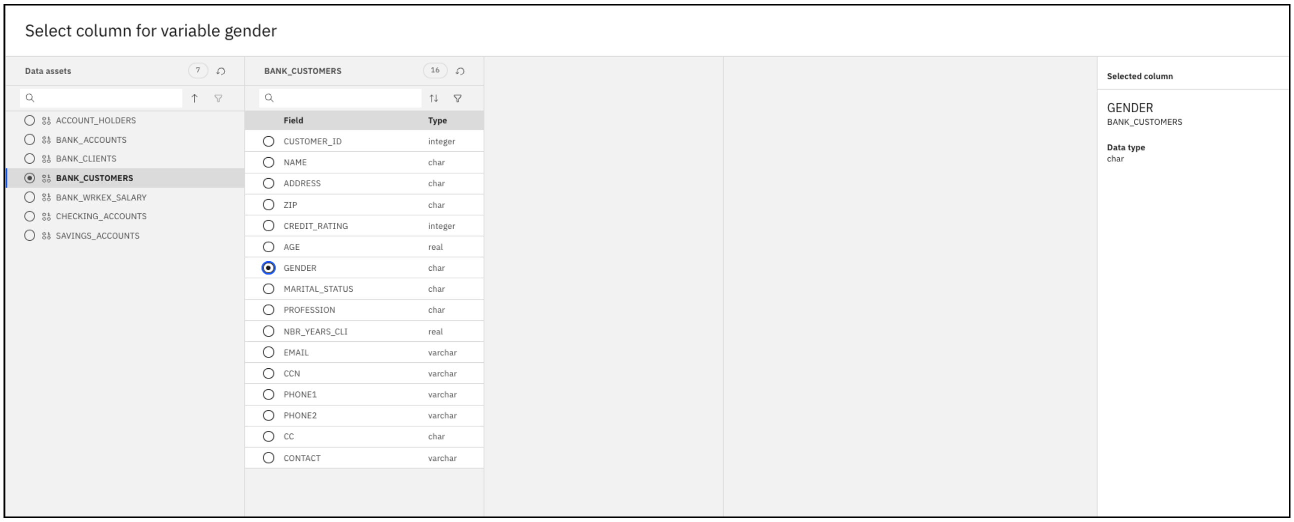

Figure 3-57 Data Quality Rule select column to which to bind

6. The column binding for the rule is shown as in Figure 3-58. Skip through the rest of the dialogs for the optional joins, output, and review to get to the point in the process where the rule can be created by using the Create option. For more information, see this IBM Documentation web page.

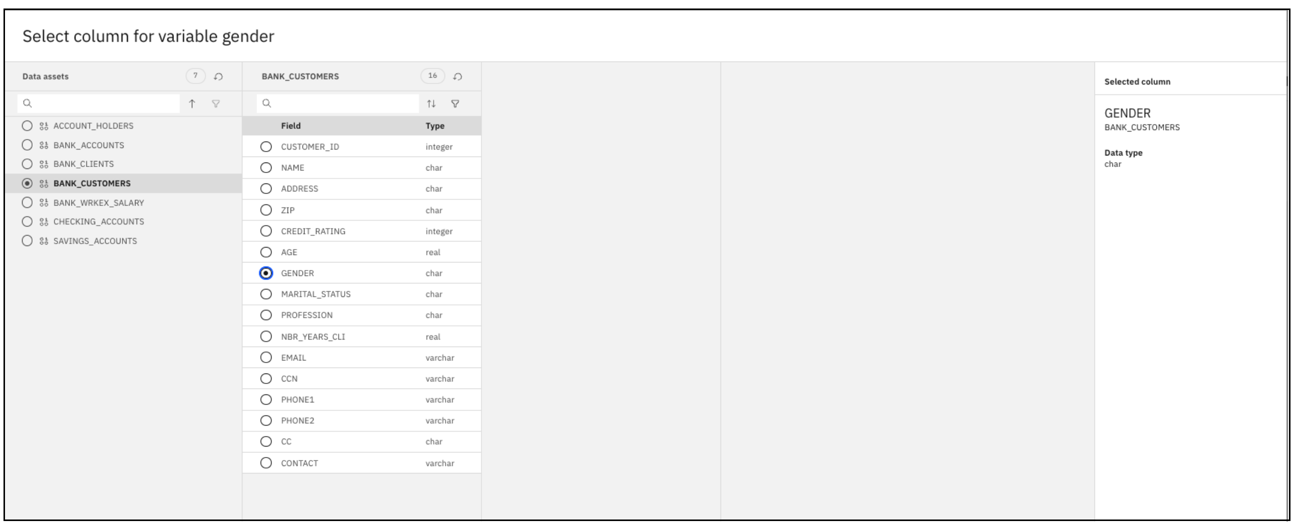

Figure 3-58 Summary of the Data Quality Rule column binding

7. After the rule is created, you can create a job to run the rule by clicking the three dots (**...**) that are in the upper right of the Quality Rule, and selecting **Create Job**, as shown in Figure 3-59

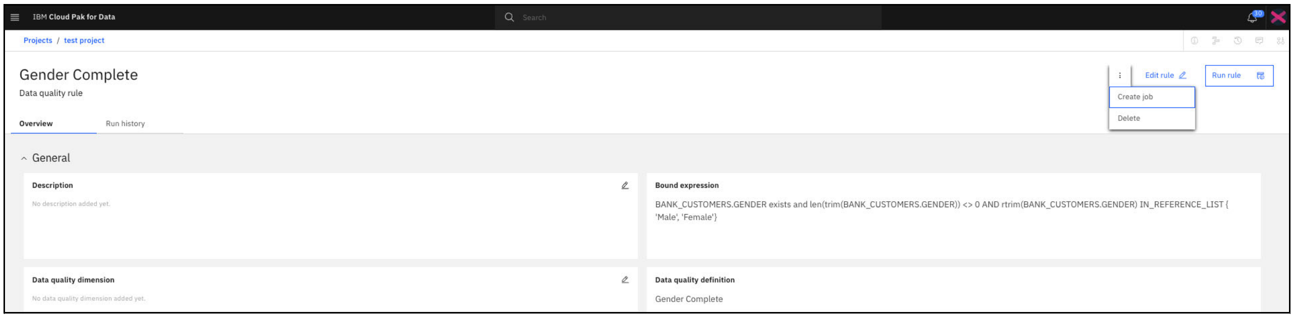

Figure 3-59 Creating a job to run the rule

8. Enter the details about the job and a name so that the job can be easily identified later, as shown in Figure 3-60. For our example, we use the default settings for the job, so we clicked through the other dialog windows until the **Create** and **Run job** options are selected.

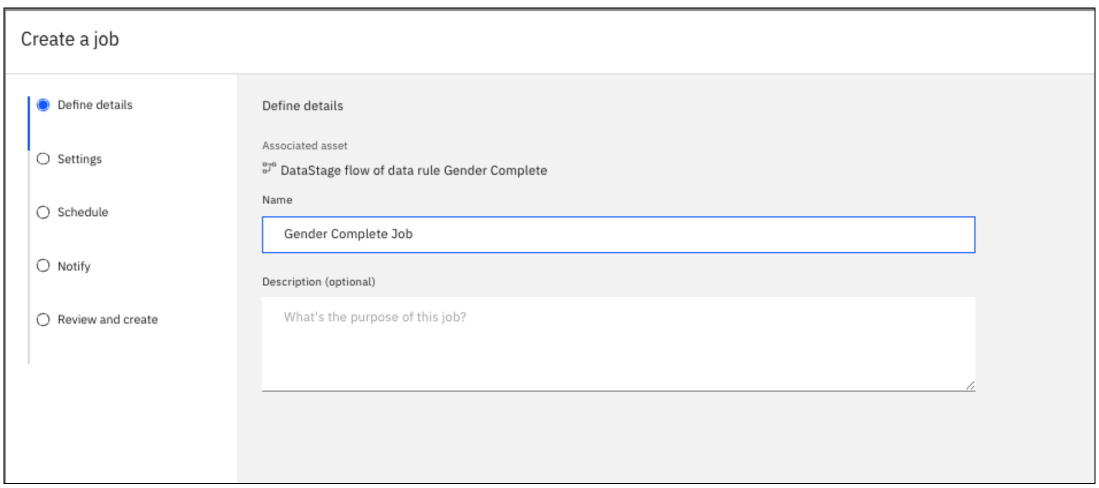

Figure 3-60 Configuring the job for the Data Quality Rule

The Job Details page shows when the job is complete (see Figure 3-61).

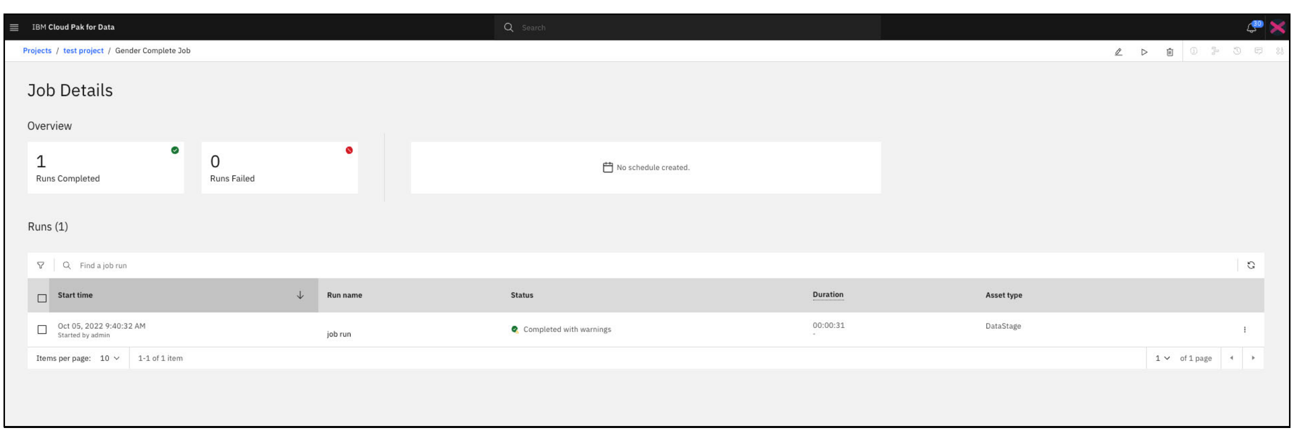

Figure 3-61 Data Quality Rule Job Complete

9. Click the job run to see more information about the job, as shown in Figure 3-62

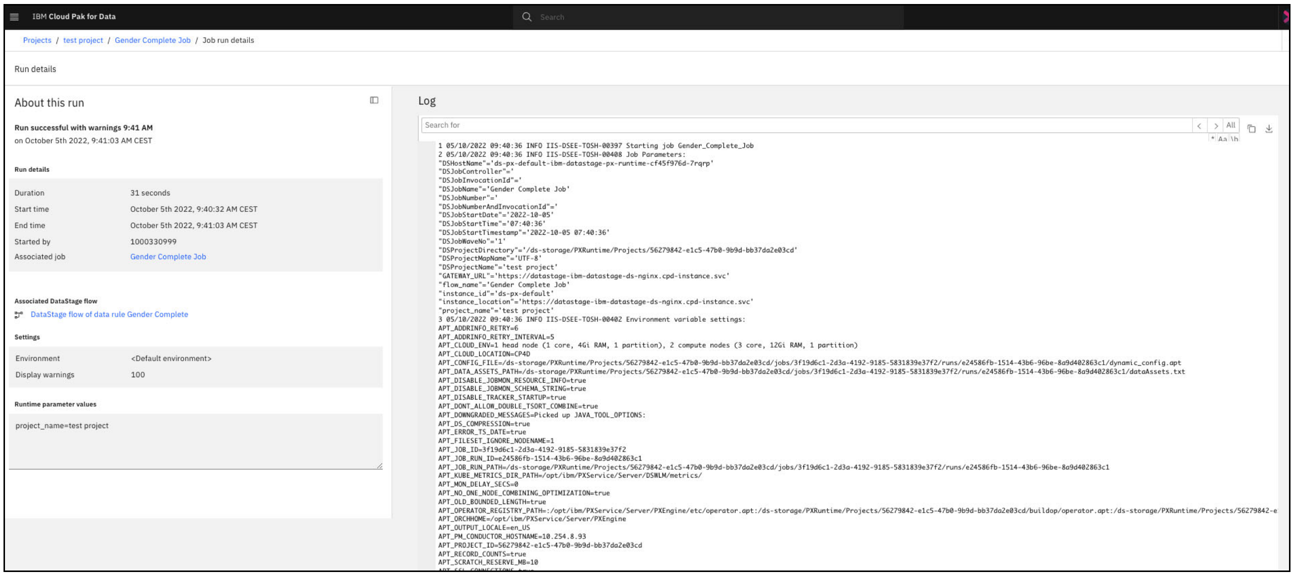

Figure 3-62 Data Quality Job Run details

3.3.4 Data lineage

Data lineage provides the capability to import data transformation assets, business reporting assets, and data assets. It includes the lineage that is contained within these assets. The lineage capability is important for several key business use cases, including the following examples:

- Impact analysis: If an element is changed within the lineage graph, what other assets are affected?
- Traceability: Where did this data come from and what transformations did it undergo?

With the advent of Cloud Pak for Data 4.5.1, lineage capabilities are provided through the integration between Cloud Pak for Data and MANTA.

To work through the examples in this section, a MANTA Automated Data Lineage for IBM Cloud Pak for Data license must be purchased separately and the license key that is provided must be installed. The MANTA lineage capability is not installed by default. A Cloud Pak for Data administrator must install the capability by following the instructions that are available at this IBM Documentation web page.

The sources that are to be used for lineage must be configured with the correct permissions so that MANTA lineage can work with the sources. For more information, see this IBM Support web page.

For this example, we use the Db2 instance that we used earlier in this use case. A script is provided in the GitHub repository that can be used to add the correct permissions to the database.

To set the permissions correctly, ensure that the **`configure-db-lineage-settings.sh`** shell script is executable. Then, run the script. The validity of the database can be validated by using the MANTA User Interface that is accessible by using the following URL:

```
https://<cloud pak for data url>/manta-admin-gui/app/#
```

Complete the following steps to validate the database:

1. To log on to the MANTA user interface by using the default credentials for MANTA, which are available by way of the Red Hat OpenShift console.
2. Browse to the project where Cloud Pak for Data is installed and look for the MANTA-credentials secret. The log on credentials are in the MANTA_USER and MANTA_PASSWORD fields of the secret.
3. After logging on to the MANTA user interface, add a connection by using the information for your BankingDemo database. (The test user and password that is created earlier are used.)
4. Click **Validate** in the lower left to ensure that all permissions are correctly set on the database for the MANTA lineage component to work correctly, as shown in Figure 3-63.

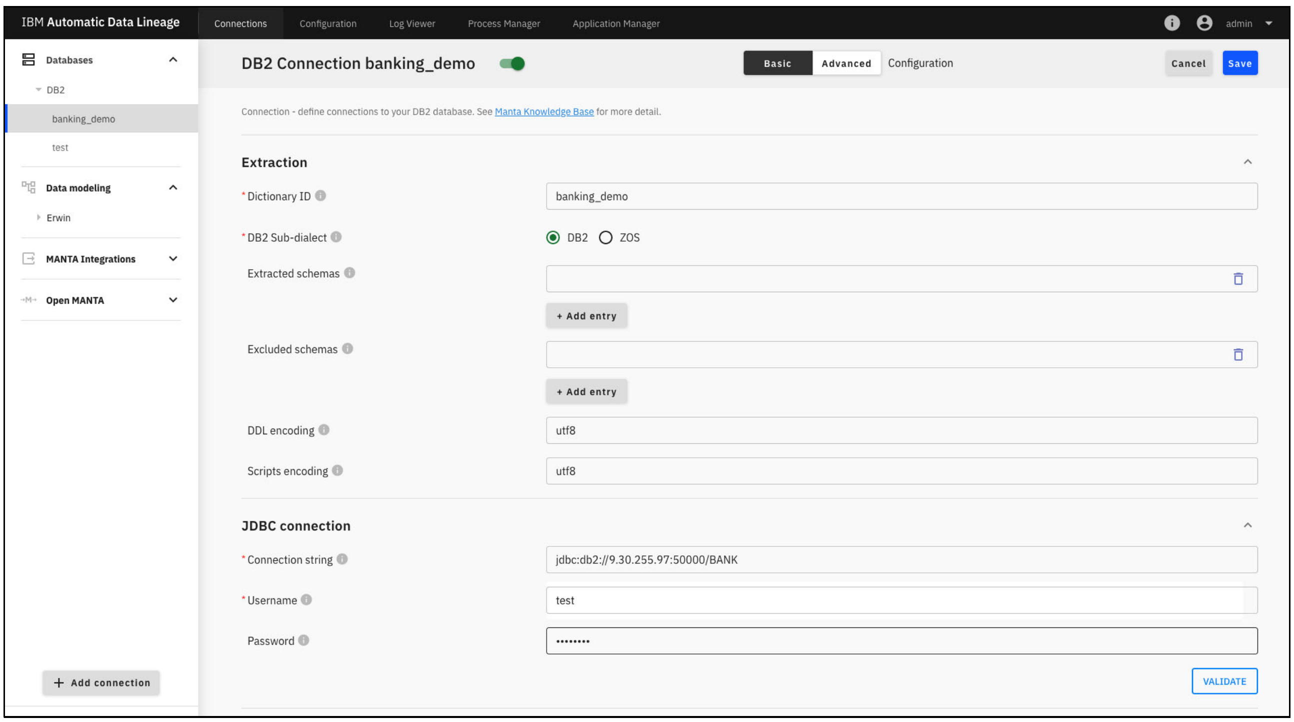

Figure 3-63 Validating the sample database for use with MANTA

In addition to setting up the permissions, we also must create some lineage information within the source. To create this information, we create a view that joins the BANK_ACCOUNTS and BANK_CUSTOMERS tables to create a new view ACCOUNT_CUSTOMER_RELATIONSHIP. To create this view, the GitHub repository contains a script (`configure-view.sh`) that ensures that the script is executable to run the script.

5. To begin importing a lineage, browse to the test project, select the **Assets** tab, and then, click **New Asset**. Click the **Metadata import tile** and then, select the **Get Lineage** tile, as shown in Figure 3-64. Then, click **Next**.

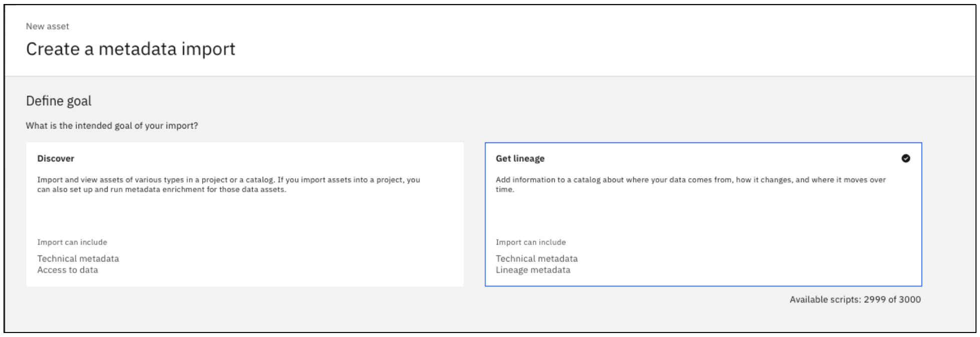

Figure 3-64 Starting the metadata import with lineage

6. Enter the name and description of the new metadata import, as shown in Figure 3-65.

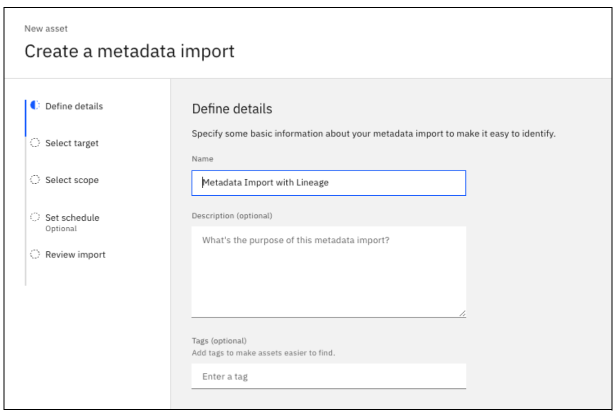

Figure 3-65 Entering a name and description for the metadata import

7. Select the test catalog as the target for the import, as shown in Figure 3-66. Then, click **Next**.

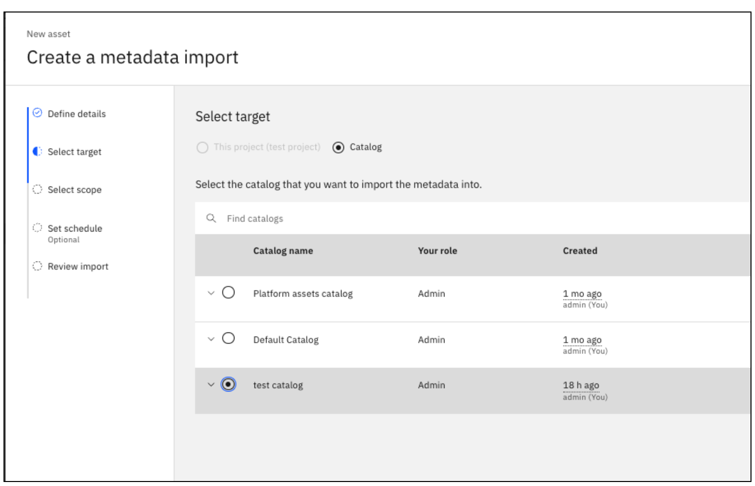

Figure 3-66 Selecting the test catalog as the target for the import

8. Set the scope of the import by selecting the **BankingDemo** database, as shown in Figure 3-67. Then, click **Next**.

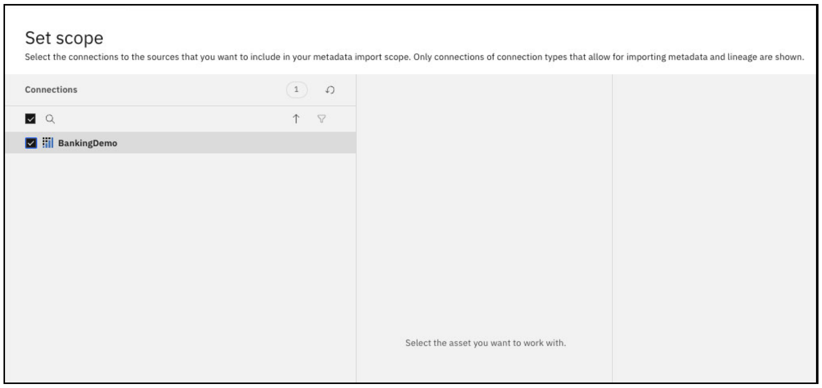

Figure 3-67 Select the scope for the metadata import.

9. Click **Next**. Again, you are presented with a window in which you set a schedule. For our example, we conduct a one-off import of the data. Click **Next**. Then, click **Create** to create the metadata import. The import can take some time to run. When the process completes, the imported assets are displayed (see Figure 3-68).

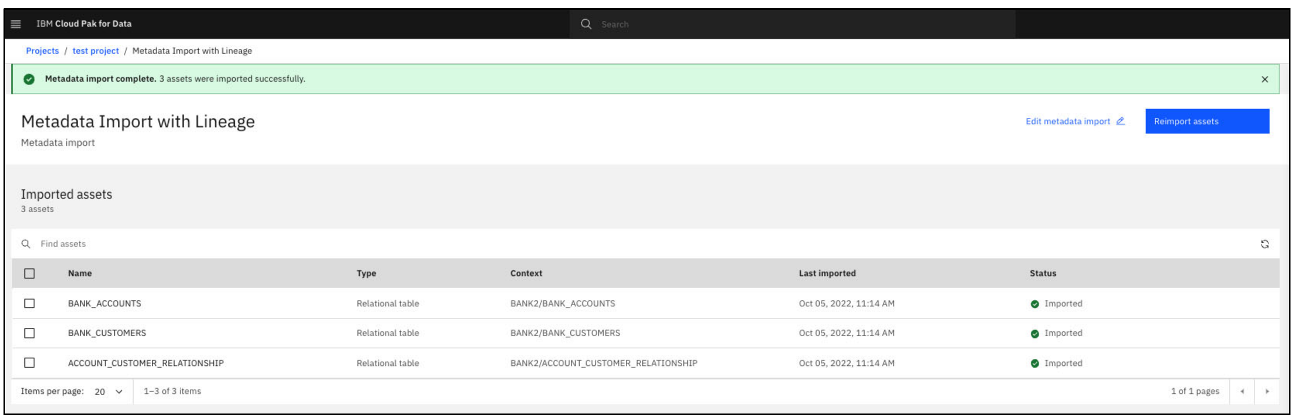

Figure 3-68 Metadata import with lineage results

10. To see the results of the import and the lineage, browse to the test catalog. Click the **ACCOUNT_CUSTOMER_RELATIONSHIP** table and then, select the **Lineage** tab. The lineage from the base tables to the view is displayed, as shown in Figure 3-69.

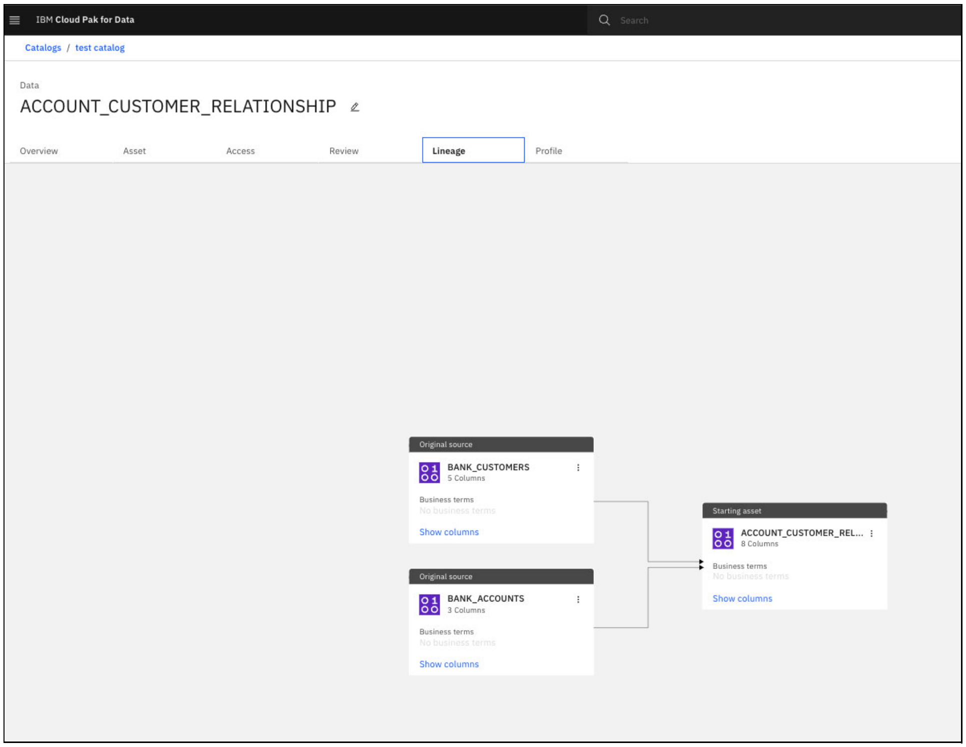

Figure 3-69 Lineage view

3.3.5 Metadata reporting

With the advent of Cloud Pak for Data 4.0.8, a data mart can be set up for metadata reporting from within Watson Knowledge Catalog. Metadata from categories, projects, and catalogs can be configured to replicate to this data mart. This data mart can be used for creating, among others, information governance reports and data quality reports.

To set up reporting, a user must have the following permissions:

- Reporting Administrator
- Manager

Before a user can configure Watson Knowledge Catalog for reporting, a database must be available and the connection to the database configured in the Platform Connections catalog of Cloud Pak for Data. At the time of this writing, the database types that are supported for the reporting data mart are Db2 11.x and Postgres.

The database can be created within the Cloud Pak for Data instance if the Db2 operators and instance were set up, or an external database can be used. The database must include a default schema to which the reporting data can be written. Regardless of which approach is used, the database must be accessible from the Cloud Pak for Data cluster. In our example, we use our sample database as the target for the reporting database.

The GitHub repository contains a shell script (**`configure-reporting-mart.sh`**) that creates the schema for reporting. Ensure that the script is executable and then, run the script.

A new schema, REPORTINGMART, is created in the samples database.

We now need to configure the reporting capability:

1. From the navigation menu, select **Administration** → **Catalogs**. The catalogs dialog box opens (see Figure 3-70). The Overview tab shows the catalogs that the logged in user can access.

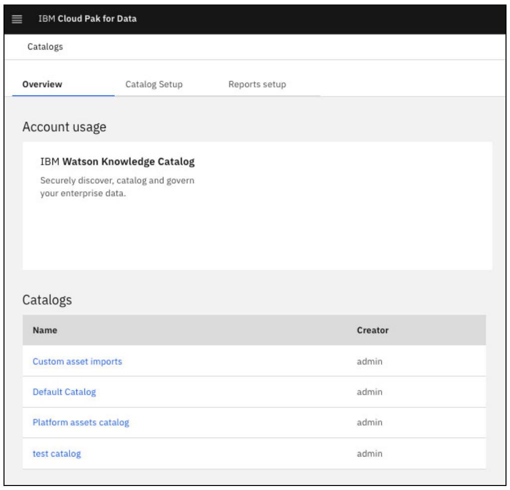

Figure 3-70 Catalog Administration

2. To set up reporting for a Watson Knowledge Catalog instance, select the **Reports setup** tab. In this tab, the categories, projects, catalogs, and rules that are to be synchronized to the reporting mart can be configured, as shown in Figure 3-71.

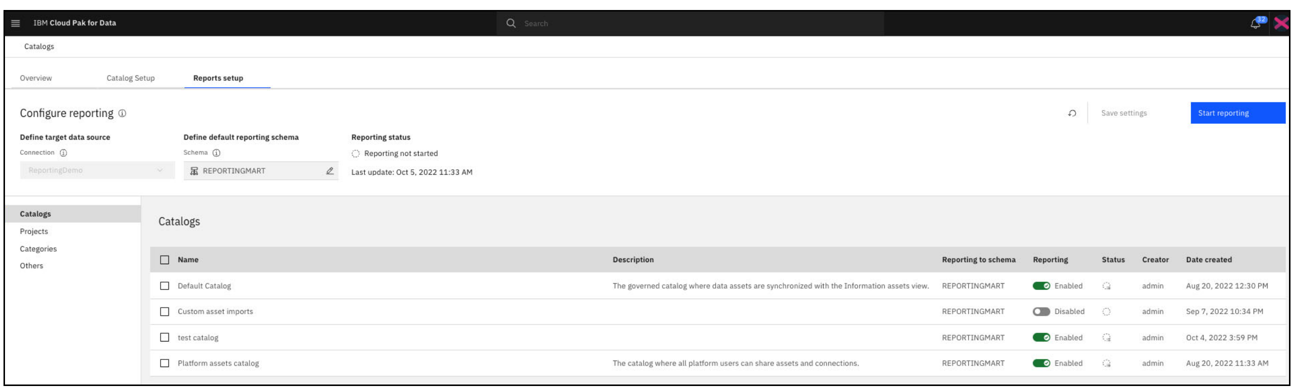

Figure 3-71 Establishing the reporting scope

Complete the following steps in the Reports setup tab:

a. Select the database connection to the sample database as the target data source.

b. Select the newly created REPORTING schema as the target schema.

c. Configure the scope of the reporting.

d. From the Catalogs window, enable synchronization for the platform assets catalog and the test catalog.

e. From the Projects window, enable synchronization for the test project.

f. From the Categories window, enable the synchronization of all categories.

3. When complete, click **Start reporting** to start the synchronization. The schema is populated with several tables and the data is synchronized to these tables. This process can take some time.

For more information about the Reporting mart tables and content, see this IBM Documentation web page.

A set of sample queries also is available at this the IBM Documentation web page.

3.3.6 Data privacy and data protection

Watson Knowledge Catalog features the concept of a data protection rule. By using a data protection rule, a user can control access to data within a governed catalog, and in some cases, projects, and data virtualization. A data protection rule contains a criteria and an action. One or more conditions can make up the criteria, and these conditions can use the user and asset metadata to form the condition.

The action block specifies the action to take if the rule is triggered. Access to the data can be blocked, or information can be redacted or obfuscated.

Advanced data privacy extends the action to allow for advanced data masking techniques. The techniques maintain the format and integrity of the data. In this example, we use the standard masking approach. For more information about advanced data masking, see this IBM Documentation web page.

> **Note:** Ensure that you are logged on to Cloud Pak for Data with a member of the administrator group to create the rule. The rule is *not* applied for the owner of the rule. Later, we swap users to see the effect of the rule for a different user.

Before we begin, we must ensure that our catalog includes the required settings to enforce the data protection rule. To confirm this inclusion, open the test catalog. From the settings page, check to ensure that governance protection rules are enabled (see Figure 3-72).

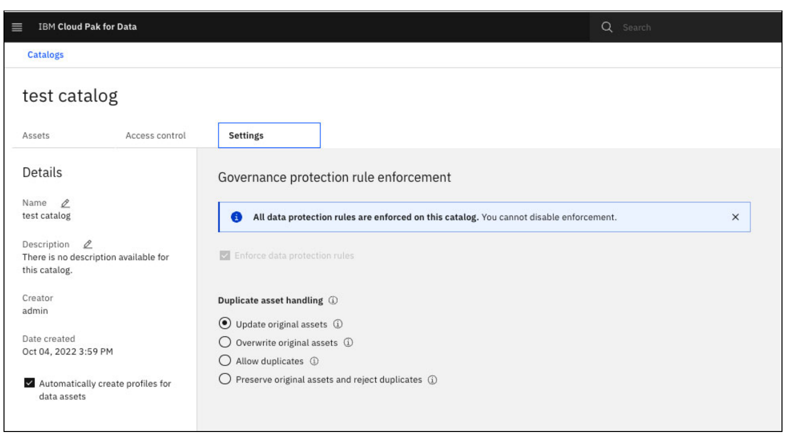

Figure 3-72 Data protection rules enabled

Complete the following steps to create a simple rule to demonstrate the capabilities:

1. From the navigation menu, select rules to take you to the Rules window, as shown in Figure 3-73.

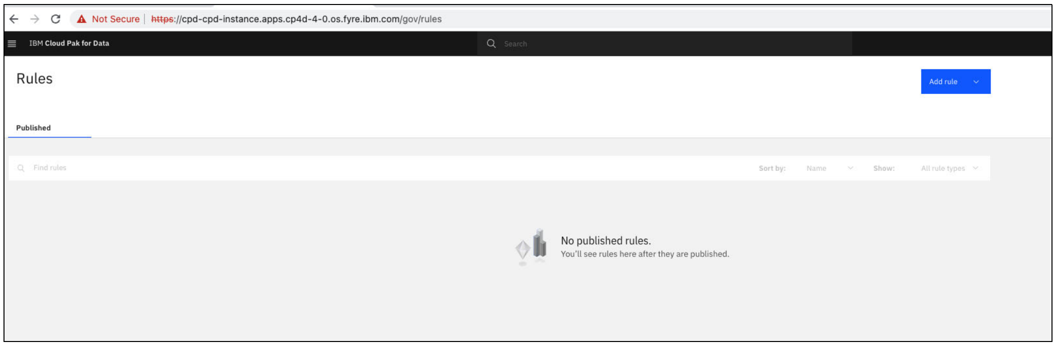

Figure 3-73 Initial rules page

2. Click **Add rule** to create a rule. An initial dialog is populated with two choices, as shown in Figure 3-74. Select **Data protection rule** and then, click **Next** to create the rule.

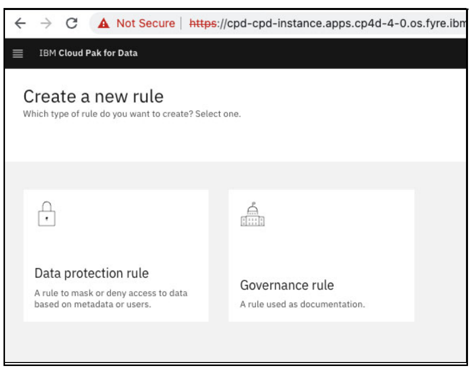

Figure 3-74 Creating a rule

3. A simple rule must be created that masks any information assets that are associated with the email data class. Complete the following steps:

   a. Create a single condition for the criteria that causes the rule to trigger if any data assets include an assigned data class of email address.

   b. In the Action section, click **Obfuscate**. The data is then modified. The rule is shown in Figure 3-75 on page 160.

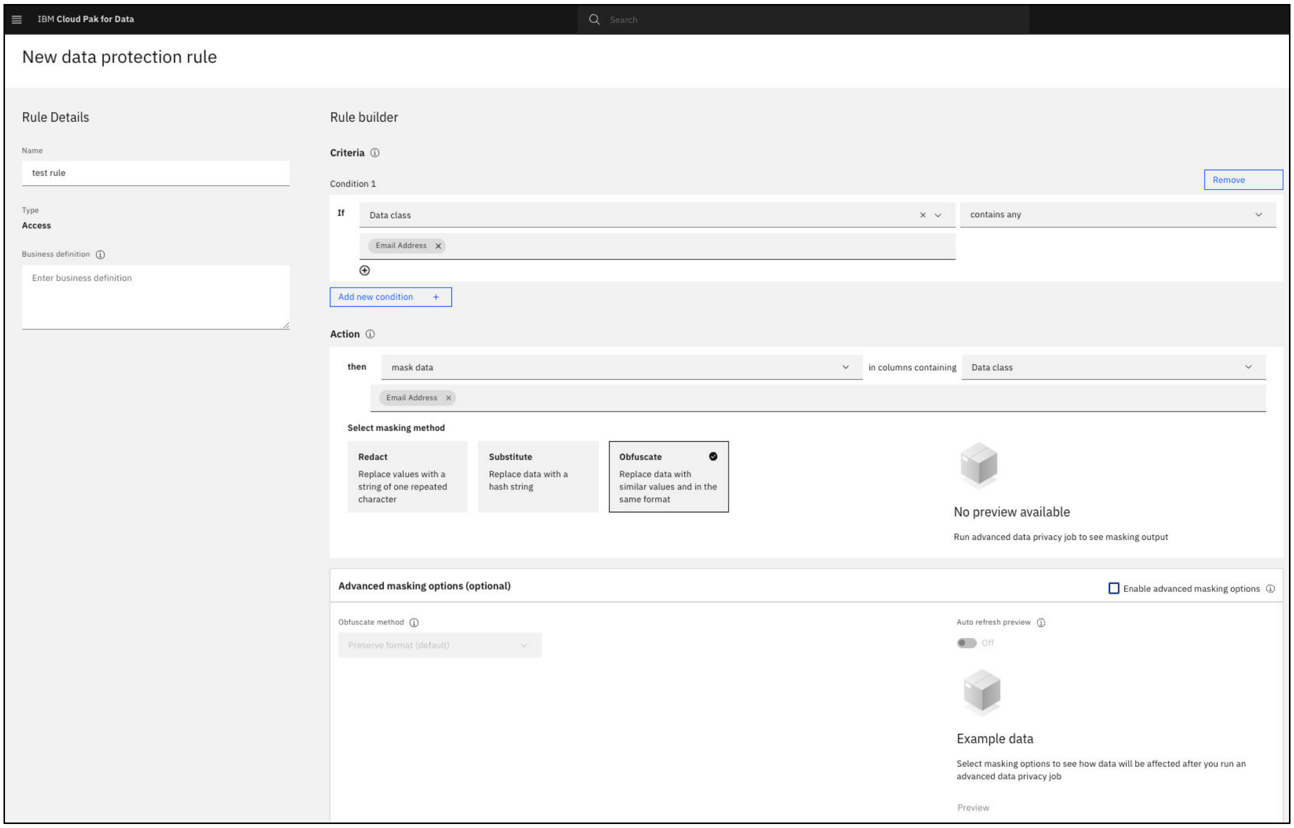

Figure 3-75 Create a simple rule to obfuscate email addresses

4. Click **Next** to create and activate the rule. The rule shows as Published, as shown in Figure 3-76.

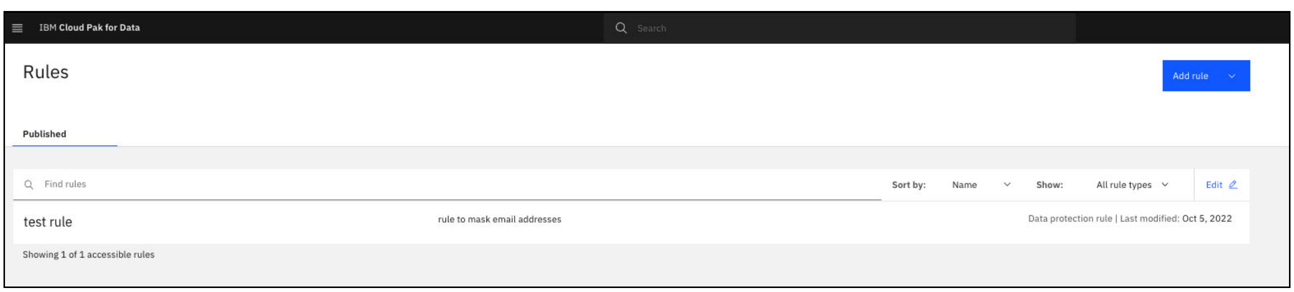

Figure 3-76 Published email rule

5. Log out of Cloud Pak for Data and log in as one of the members of the test stewards group.
6. Browse to the test catalog and open the **BANK_CUSTOMERS data asset**. Navigate to the **Assets** page (a delay in displaying the asset might occur because the data protection rules are applied). The Asset shows that the email addresses were obfuscated, as shown in Figure 3-77.

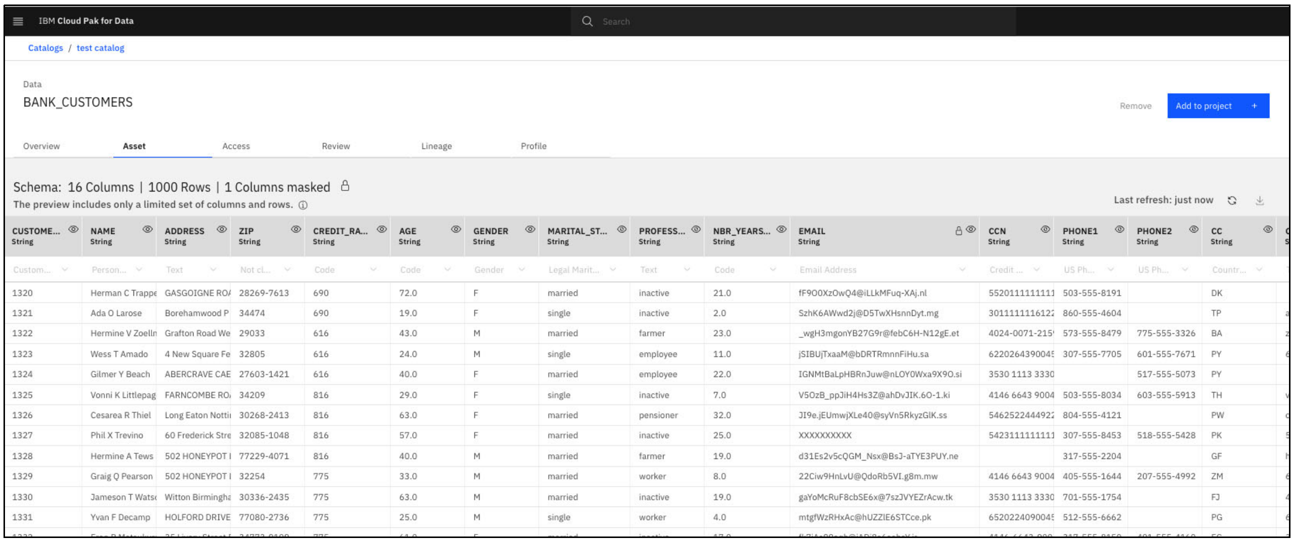

Figure 3-77 Data set with email data protection rule applied

4

Multicloud data integration

Data integration helps combine structured and unstructured data from disparate sources into meaningful and valuable data sets. Integration styles, such as bulk and batch integration (ETL or ELT) and data virtualization enable various integration use cases.

With enterprise data being distributed across many on-premises and cloud sources and different clouds, integrating that data, enabling continuous availability of mission-critical data, serving it to relevant consumers, and democratizing access to data is no mean feat.

IBM Cloud Pak for Data is a scalable modular Data Fabric platform that provides key integration and governance capabilities that enable multicloud data integration and data democratization.

This chapter includes the following topics:

4.1 Introduction

IBM Cloud Pak for Data is a Data Fabric solution that is designed with multicloud data integration in mind.

It features an advanced connectivity framework and ships with a wide range of pre-built standard connectors to various cloud-based and on-premises databases and applications with which you can connect to your data no matter where it lives.

For more information about supported sources, see this IBM Documentation web page.

The connectivity framework's design is centered around the concept of Connections and does not involve any data movement or copying at the point of Connection creation. A Connection acts as a reference and a pointer to the source system. The data stays at the source until and unless relevant analytical and data integration and movement activities are further started through the platform.

Figure 4-1 showcases some of the connection types that are available in IBM Cloud Pak for Data as standard.

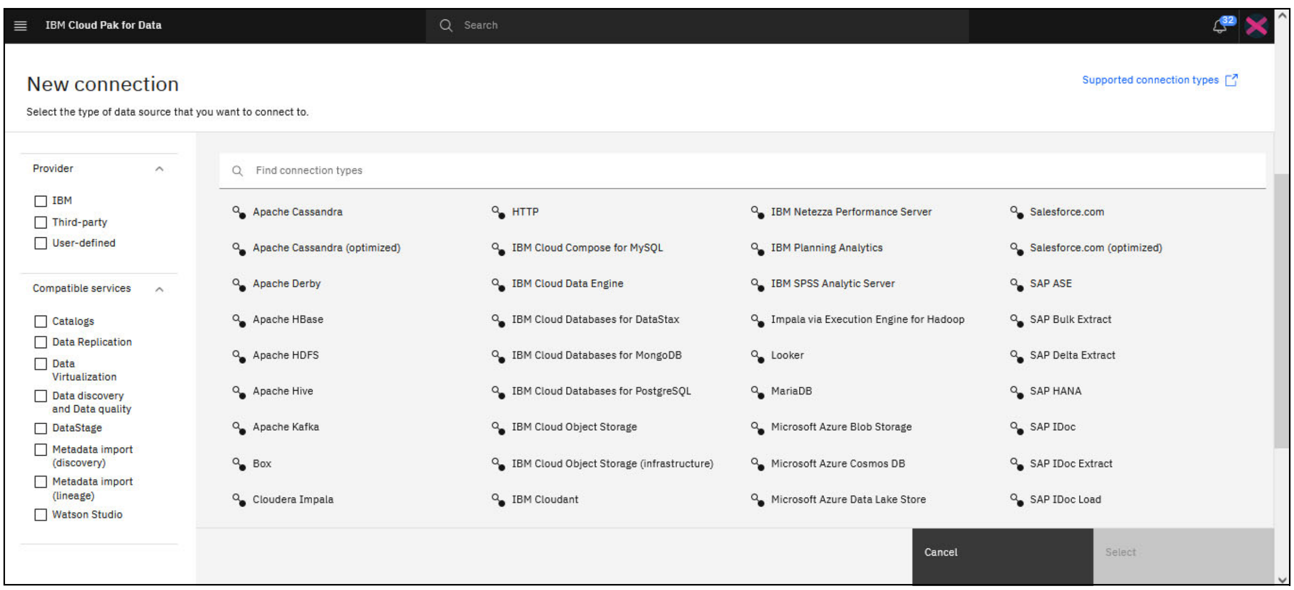

Figure 4-1 Connection types (partial list)

Where a standard connector is not provided, it also is possible to upload custom JDBC drivers for the target source to the platform. Then, establish a custom generic JDBC connection to that source (see Figure 4-2).

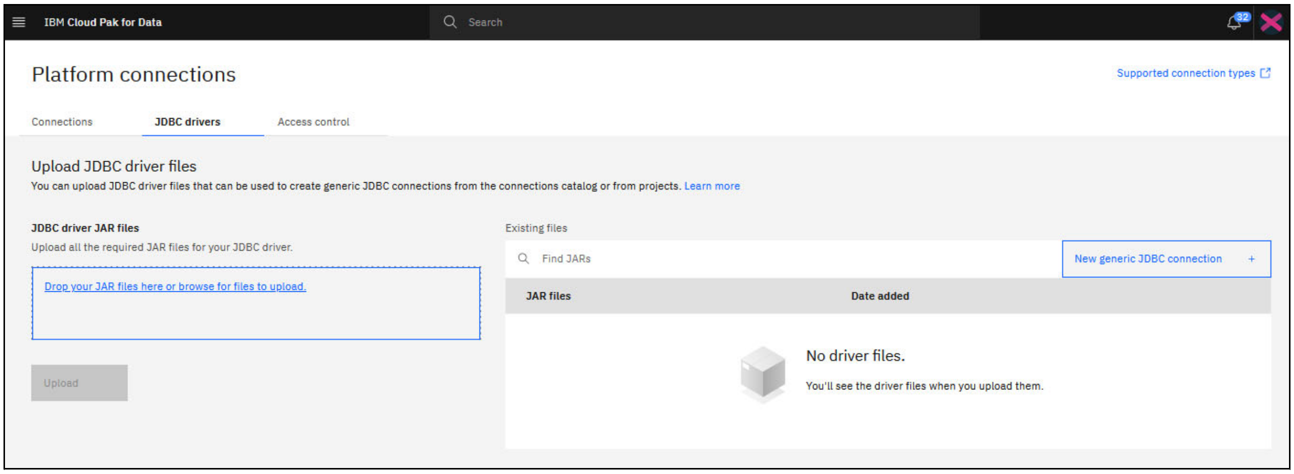

Figure 4-2 JDBC driver upload window

IBM Cloud Pak for Data further provides a range of data integration capabilities for data movement, transformation, virtualization, and preparation. These capabilities can be used independently or in concert with and to complement one another.

Table 4-1 lists the available key services and tools and maps those services to personas, users, and teams that most commonly prefer to work with them.

Table 4-1 Data integration and transformation tools available in IBM Cloud Pak for Data

Capability	Service	Editor type	Personas and teams
Data Virtualization	IBM Data Virtualization	No code UI-based view creation, with an option to work with SQL code if and as needed.	► Personas: Administrator, Data Engineer, and Data Steward ► Teams: Data Engineering, data integration, and IT teams
Extract-Transform-Load (ETL)	IBM DataStage	Low-code/no-code (DataStage visual flow designer)	► Personas: Administrator and Data Engineer ► Teams: Data Engineering, data integration, and IT teams
	Python and Spark development capabilities of IBM Watson Studio and Analytics Engine for Apache Spark	Code (Jupyter Notebooks or JupyterLab IDE)	► Personas: Administrator, Data Engineer, Data Scientist, and Developer ► Teams: Data Engineering, data integration, and IT teams; Developer teams
Data preparation, wrangling and cleansing	IBM SPSS Modeler IBM Data Refinery	Low-code/no-code (SPSS Modeler visual flow designer); no-code with an option to use more R code (Data Refinery)	► Personas: Business Analyst, Data Scientist, and Data Engineer ► Teams: Line of Business or any team

This chapter covers IBM Data Virtualization and IBM DataStage use in more detail.

4.2 Data integration and transformation with IBM DataStage

IBM DataStage allows you to integrate data from disparate on-cloud and on-premises sources, and transform, cleanse, analyze, and move data.

The service enables design, scheduling, and execution of Extract-Transform-Load (ETL) data flows, with no coding required. The service provides a graphical drag-and-drop flow designer interface with which you can easily compose transformation flows from reusable graphical units. As such, it enables technical (for example, data engineers, data integration specialists and IT teams) and nontechnical users to participate in ETL pipelines design and execution.

For this use case, a loan applicant’s qualification example from the Banking and Financial Services industry is used to illustrate the capabilities of this service.

Whenever our bank receives a loan or mortgage application from a customer, it must assess the customer’s situation and the requested loan amount to determine what interest rate it is prepared to offer the customer. The riskier the investment, the higher the interest rate.

As a data engineer of the bank, you can access a data warehouse that holds applicant’s and applications data, and a separate database that holds credit score information. You also have guidance about the rate that the bank often is prepared to offer a customer based on their credit score, which is held in a separate No-SQL data store.

You are tasked with building an integration flow that is based on all of that data, which underpins the decision-making process behind the interest rate calculation. The Loans department of the bank urgently needs this information, and requested it to be delivered weekly as a flat file to their object storage repository.

4.2.1 Designing the DataStage flow

DataStage flow and job design and execution is performed within projects. Projects are collaborative workspaces that provide access to various analytical, AI, data science, dash boarding, developer, integration, and other tools that help store and organize the related artifacts.

The available tools in projects depend on which services are deployed on your Cloud Pak for Data cluster. For example, the IBM DataStage service must be deployed on the cluster to enable and see the DataStage, DataStage Component, and Parameter Set tiles in the list of components that are available in your project.

Two versions of the service are available: IBM DataStage Enterprise and IBM DataStage Enterprise Plus. Both versions provide the same connector types and core processing stages. The Plus version enables more data quality-specific transformation stages for the following tasks:

- Identifying potential anomalies and metadata discrepancies
- Addressing data verification (parsing, standardization, validation, geocoding, and reverse geocoding)
- Identifying and handling duplicates (including probabilistic matching)

In this scenario, we are working with the IBM DataStage Enterprise Plus version of the service.

We completed the following steps to build the flow:

1. We created a Project that is named `MultiCloud Data Integration`.
2. To create the Project, we browsed to **Projects** → **All Projects** from the main navigation menu (see Figure 4-3).

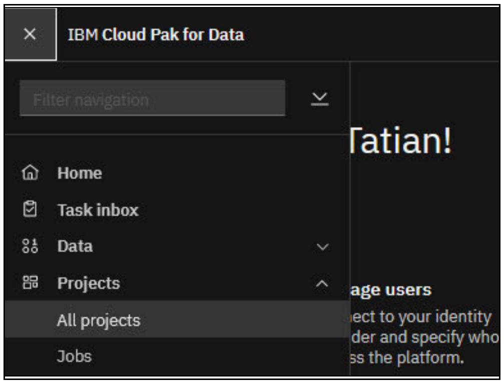

Figure 4-3 Accessing Projects from the main menu

3. The projects that were created and the projects that were added to as a collaborator by other users appear in this list. Different projects can be used to organize, group, and control access to different transformation and ETL activities and initiatives. To add a project, we clicked **New Project** (see Figure 4-4).

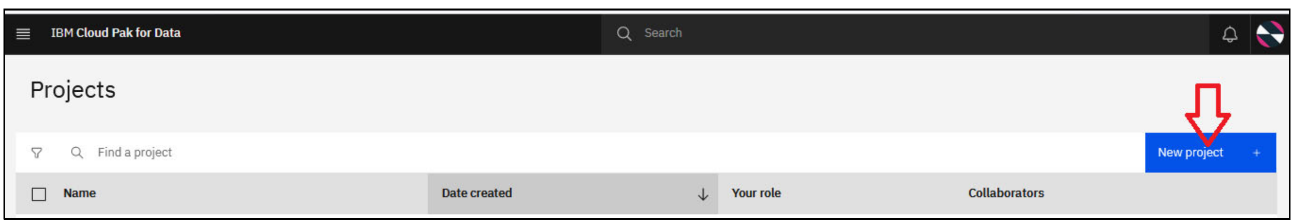

Figure 4-4 Creating a project

4. Projects can be created empty from scratch or by reusing previous work and assets by uploading a project file in `.zip` format. In this case, we use the first option (see Figure 4-5).

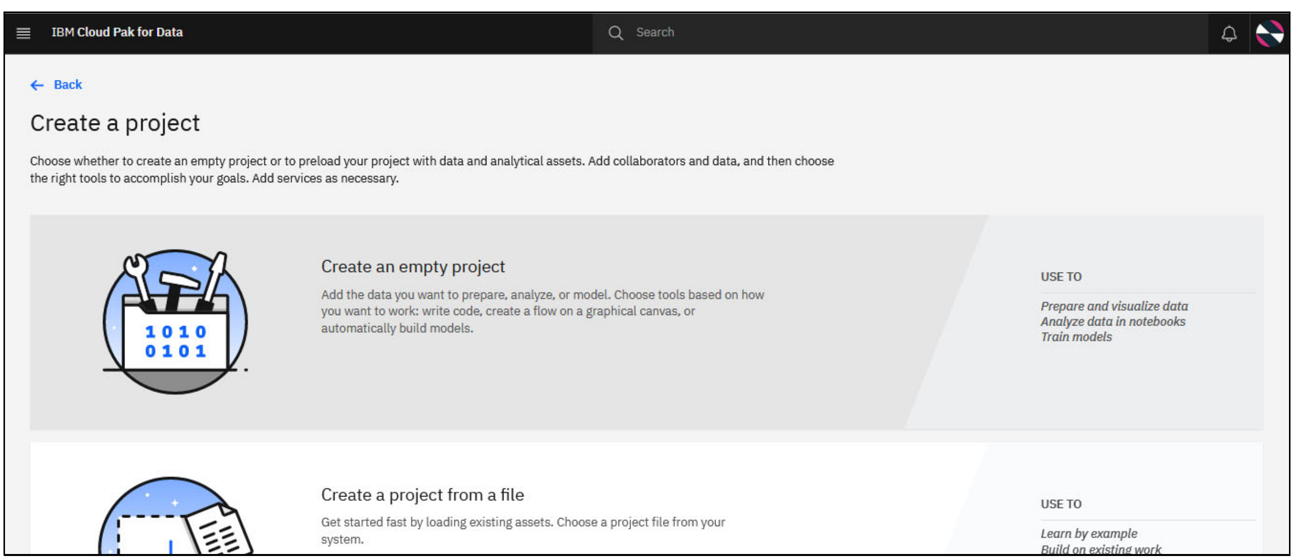

Figure 4-5 Creating an empty project versus creating a project from a file

5. Each project needs a name, and optionally, a description.

   Two other choices also must be made: whether to designate the project as Sensitive, and whether to log all project activities.

   Designating the project as Sensitive prevents data from being moved out of the project and adds a `Sensitive` tag to the project with an editable message as hover help. It also disables members of a project, including administrators, from downloading or exporting data assets, connections, or connected data from a project.

   This choice must be made at the beginning of the process and cannot be changed after the project is created.

   Choosing to log all project activities stores detailed information about project activities in a downloadable log. This choice can be changed after the project is created, if needed.

   For more information about these options, see the following IBM Documentation web pages:

   – Marking a project as sensitive
   – Logging project activity

   In our example, we do select either of these options and proceed with creating the project (see Figure 4-6) by clicking **Create**.

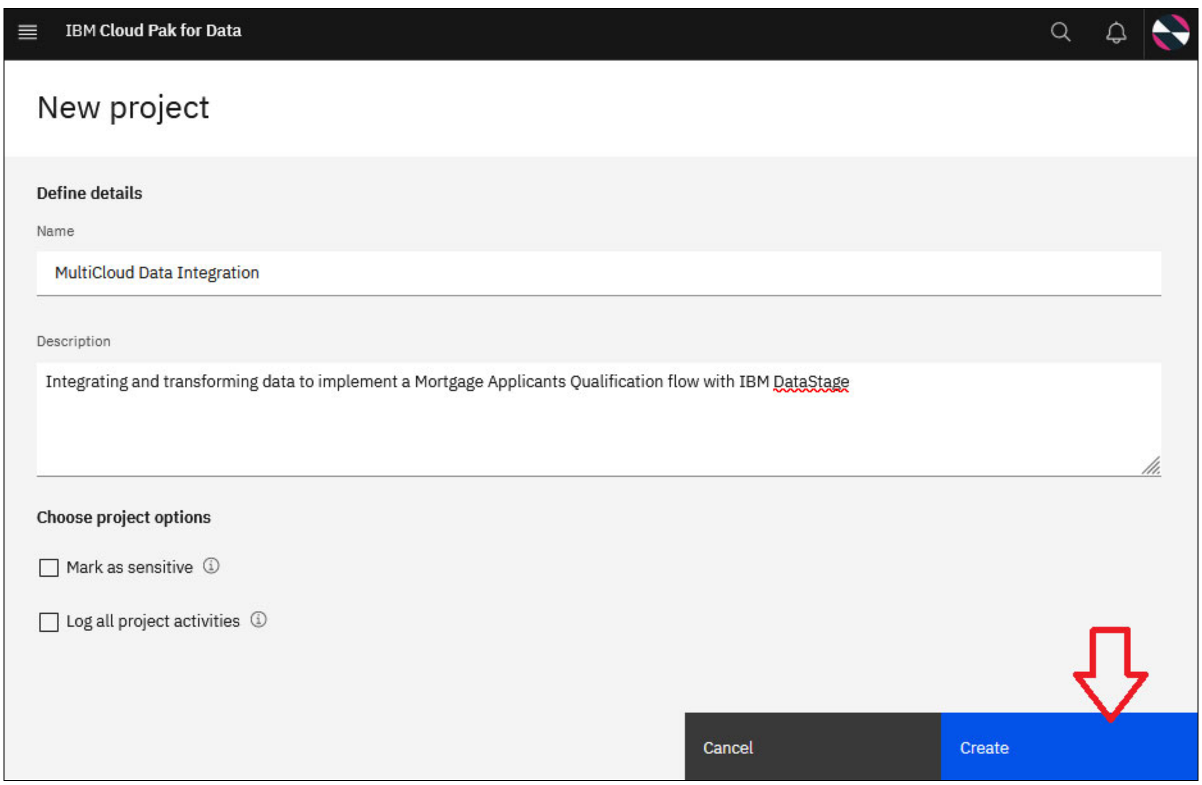

Figure 4-6 New project details

6. Determine which users are collaborating within this project workspace. The Manage Tab of each project includes project-level settings. The Access Control subsection is where collaborator setup is managed.

   Project roles (Admin, Editor, or Viewer) are assigned per project. They enable and control access to project data and assets. This step is mandatory in addition to assigning platform-level roles for each relevant user.

Platform-level roles enable general feature-function type access and permissions. Catalog, category, and project-level role assignment further enables and controls access to data and assets within.

A user who created a project is automatically assigned a project Admin role and can create any assets and artifacts within it. They also can assign other users as collaborators. Generally, if a user or a user group is not assigned a role within a project, they cannot contribute to it and are unaware that they exist. The only exception to this rule is for users with the platform-level Administrator role who have the Manage projects permission.

In our example, we add a User Group that is called `Rdb Admins` as Editors to the project, as shown in Figure 4-7 - Figure 4-11 on page 171.

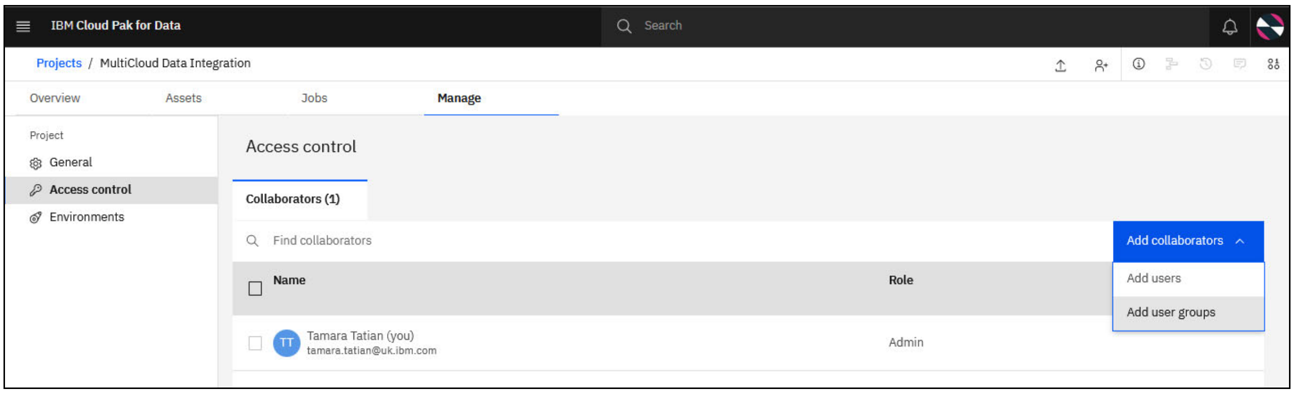

Figure 4-7 Adding collaborators to a Project

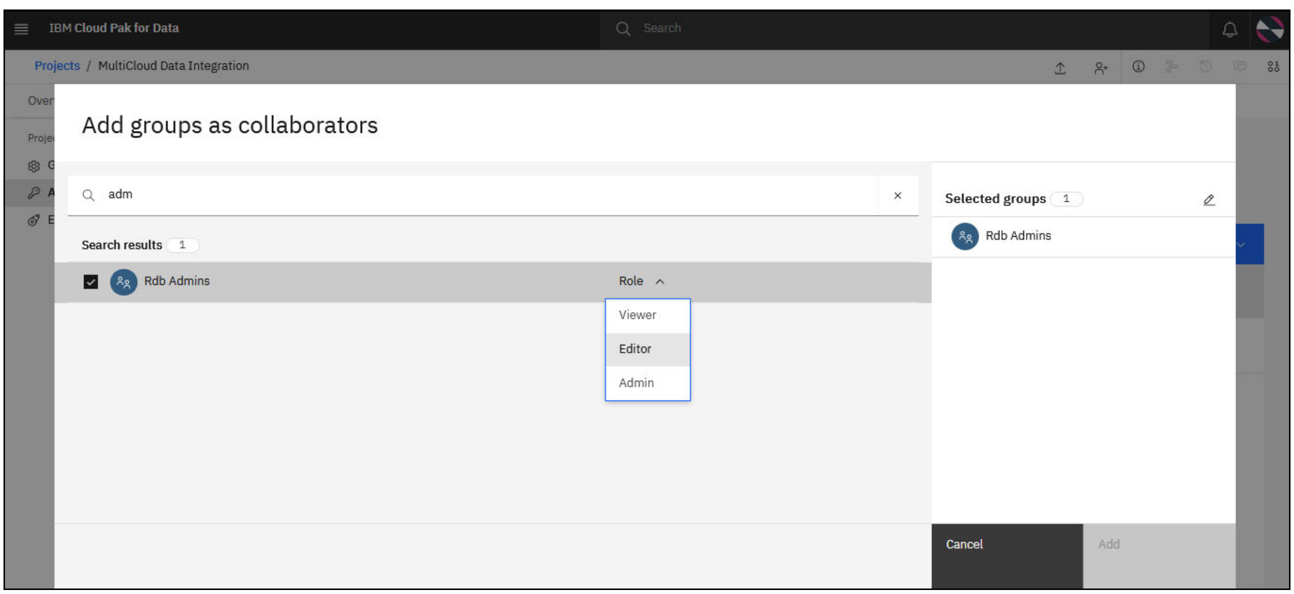

Figure 4-8 Assigning project-specific roles to chosen collaborators

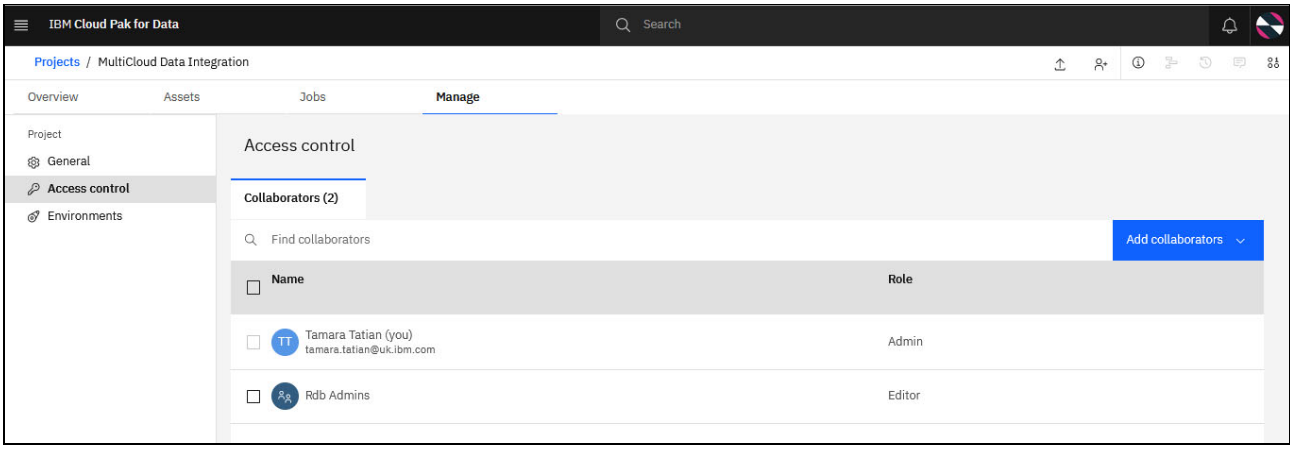

Figure 4-9 Access control tab - Project collaborator list

7. Project Editors can add various assets to the Project, but cannot add others as collaborators to it, or change Project settings.

   The relevant project setup is done and we can start creating project assets and generating project artifacts.

   For integration tasks and activities, the relevant related assets are:

   – Connections
   – Connected Data Assets
   – Local files
   – DataStage flow instances and definitions
   – DataStage Components and Parameter Sets
   – Job definitions

   As you start working with the flows and running the transformations that they define as jobs, the platform generates more artifacts, such as job run details and logs.

   Transformation flow design starts with capturing and defining the wanted source and target data sources, files, and applications.

   Click **New asset** on the project Asset tab to see the list of project asset types that can be created.

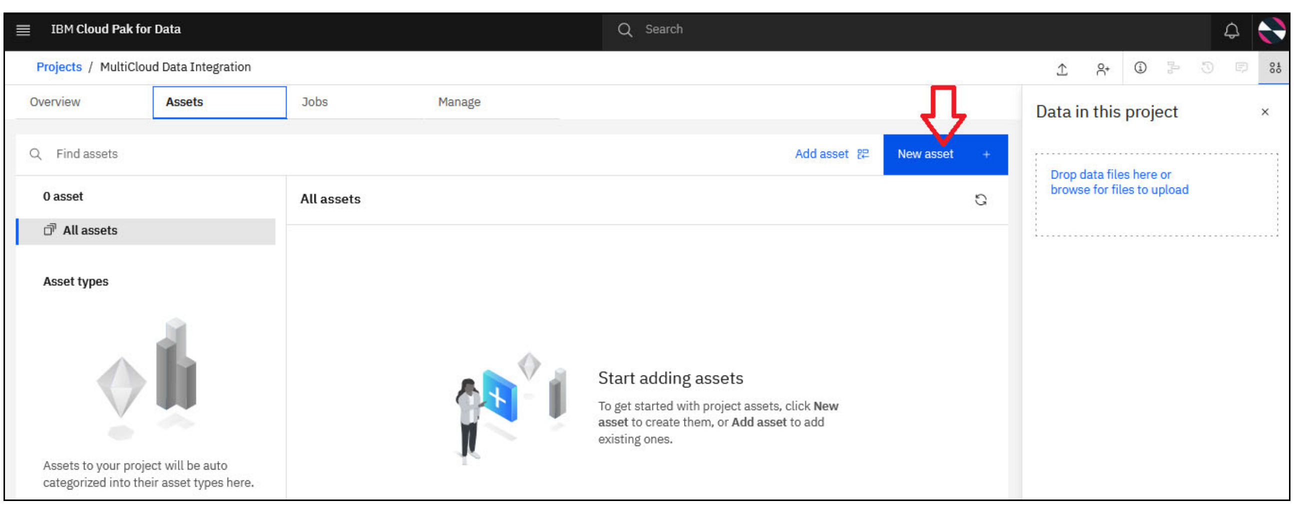

Figure 4-10 Adding new assets to a Project

8. Data that is required for the transformation flow we are building (tables, and so forth) is in several separate remote data sources and is referred to as *Connected data* in IBM Cloud Pak for Data. Connected data comes from the associated Connections. Connections assets help represent the data sources and contain the information that is required to connect to and access data in them (see Figure 4-11).

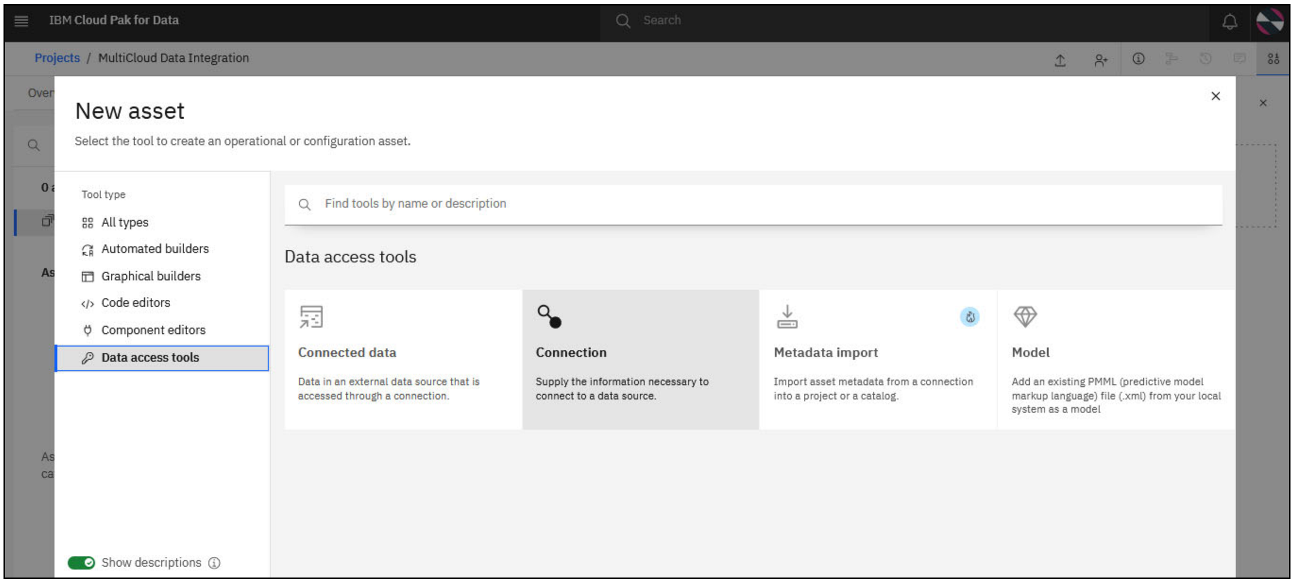

Figure 4-11 Adding data

IBM Cloud Pak for Data includes carious predefined standard connectors to common data source types. You can reduce the available connectors by service (in this case, DataStage) to see which are supported for that service.

The first connection that we add is an IBM Db2 Warehouse connection. Applicant and Applications details for our use case are in one of our organization's IBM Db2 Warehouse instances (see Figure 4-12).

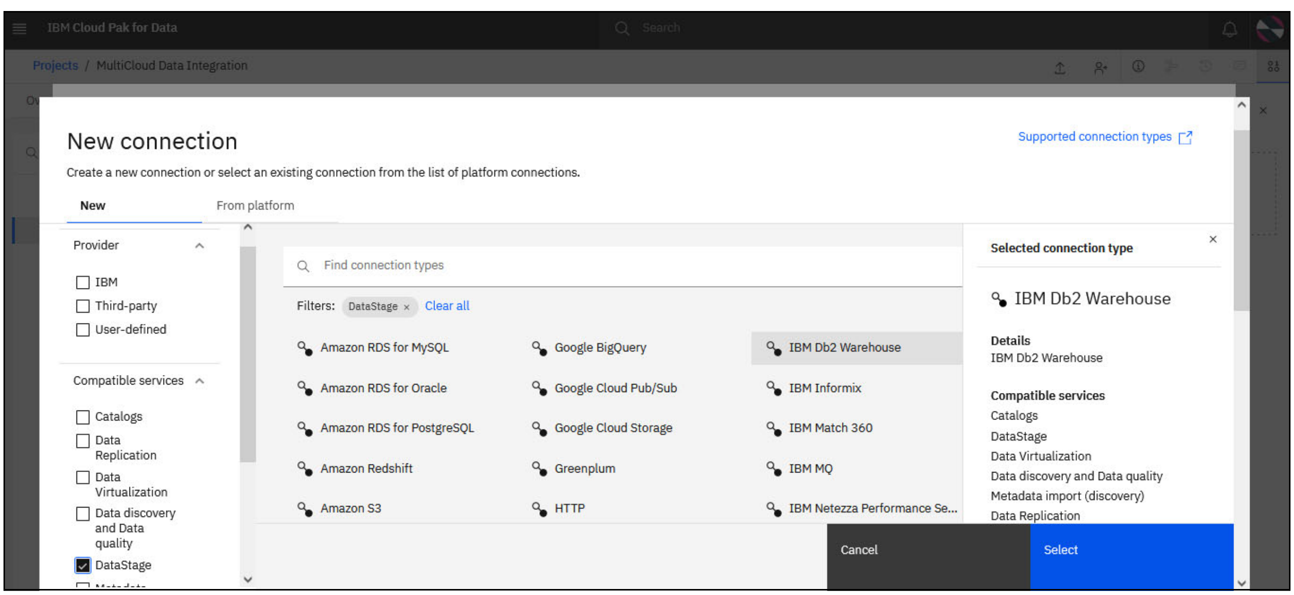

Figure 4-12 Standard connectors that are included in the solution

9. Enter a name for the connection setup and optionally, a relevant description to help easily distinguish the connections (see Figure 4-13).

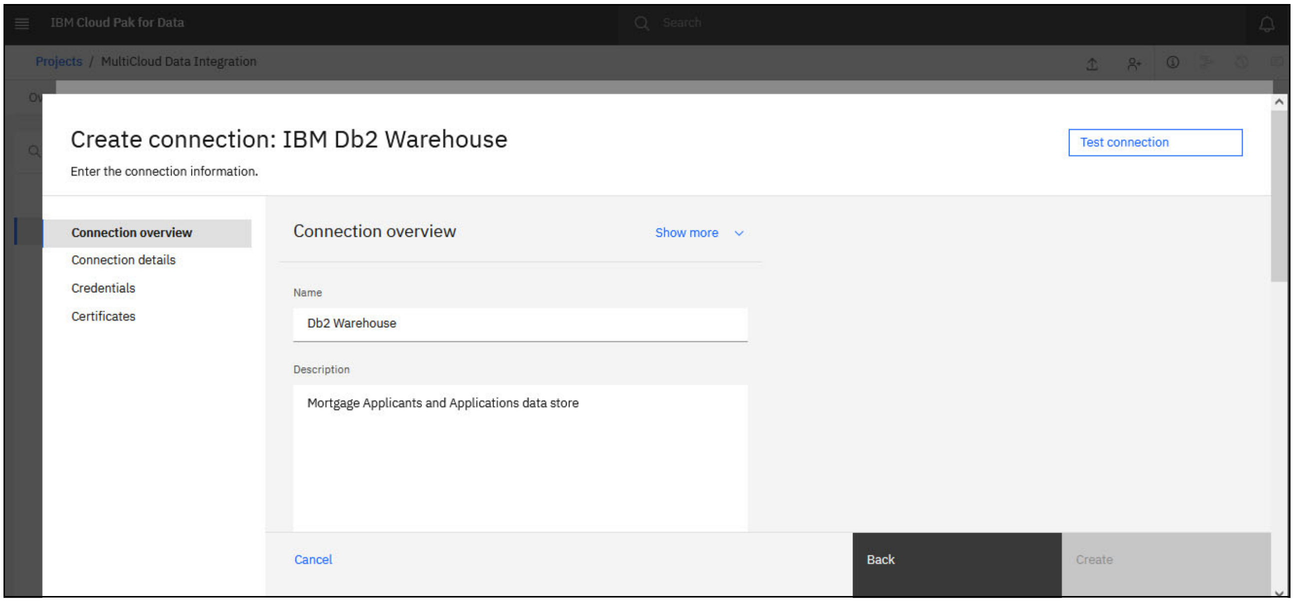

Figure 4-13 Creating a connection to IBM Db2 Warehouse: Part 1

10. Define the connection's details. The specific information varies according to connection type, and IBM Cloud Pak for Data clearly indicates which details must be specified.

    For the IBM Db2 Warehouse connection that we are working with, these details include database name, hostname, port, and authentication method (username and password in this case).

    During the setup, the connection can be designated as Shared (one set of credentials is reused by all users of the connection and connected data coming from it) or Personal (each user must enter their own credentials to the source system to work with the connection and connected data assets that are coming from it). The connection that we are setting up is designated as Shared (see Figure 4-14).

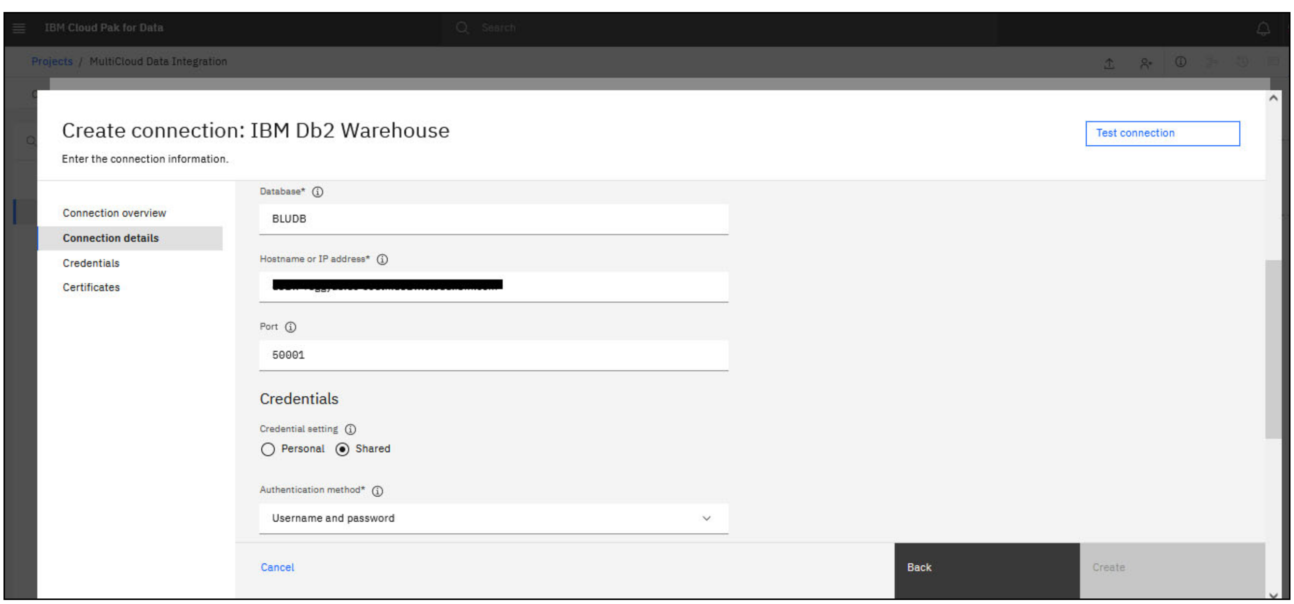

Figure 4-14 Creating a connection to IBM Db2 Warehouse: Part 2

11. After all of the required details are provided, click **Create** to finish creating the connection. Before proceeding to this final step, as a best practice it is recommended that you validate the setup first. The setup can be tested by clicking **Test connection** in the upper right of the window (see Figure 4-15).

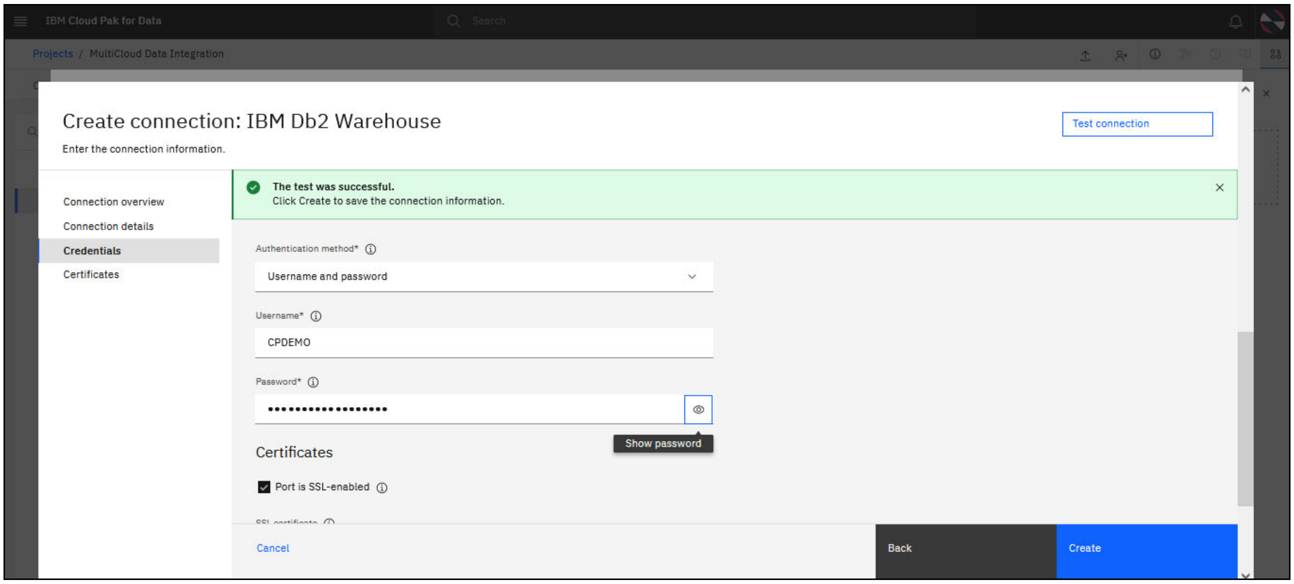

Figure 4-15 Creating a connection to IBM Db2 Warehouse - part 3

12. Because our test is successful, and we finish creating the connection by clicking **Create**.

> **Note:** Consider the following points:
>
> - The Test connection feature availability depends on the connector type that is selected.
> - If the test fails, the error details guide you as to what went wrong and must be fixed. The system still allows you to create your connection, but the errors must be addressed so that you can access data that is at the data source the connection represents.

13. Create a connection to the `PostgreSQL` database where the credit score information for the use case is stored. The PostgreSQL instance that is used by our organization is a managed instance on IBM Cloud; therefore, choose the **IBM Cloud Databases for PostgreSQL connector** for it and follow the same general setup process (see Figure 4-16).

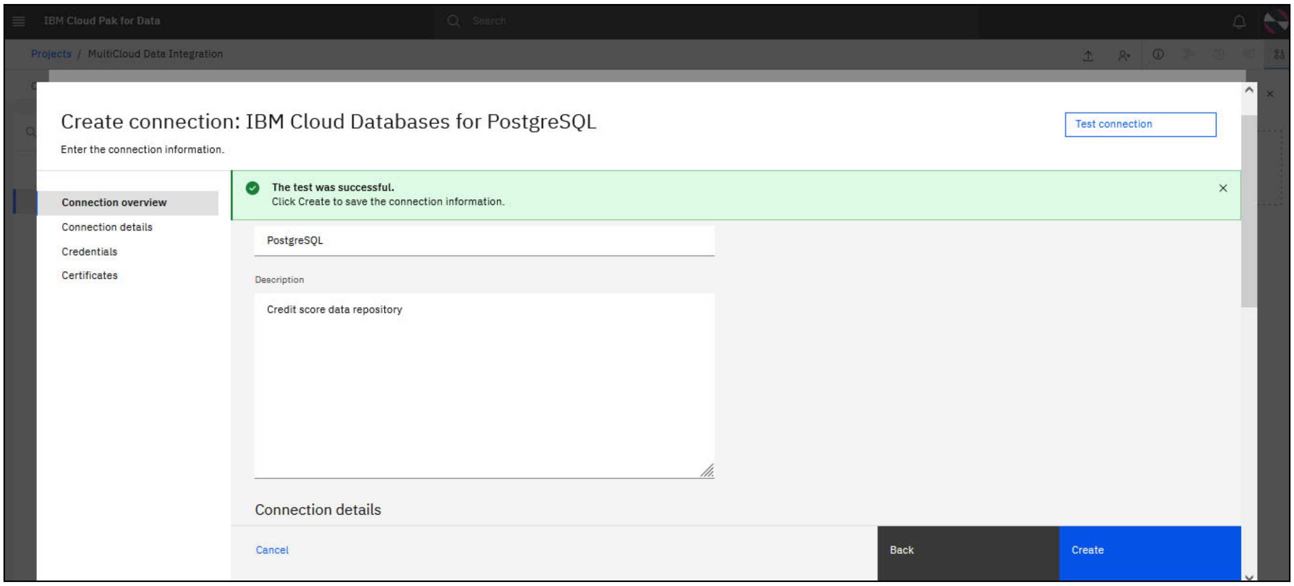

Figure 4-16 Creating a connection to IBM Cloud Databases for PostgreSQL

A connection to MongoDB is required to pull interest rate information. This connection is created next (see Figure 4-17).

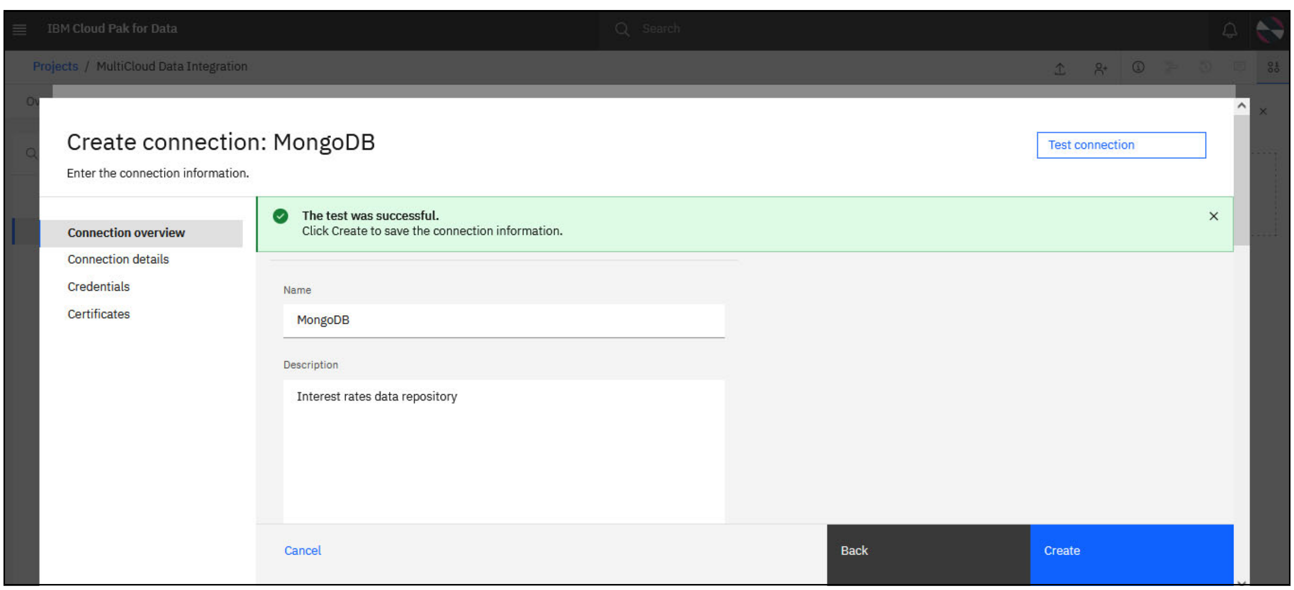

Figure 4-17 Creating a connection to MongoDB

14. We are prompted to copy the results of our transformation flow into the Loans department's Cloud Object Storage instance. A connection to that instance must be defined (see Figure 4-18).

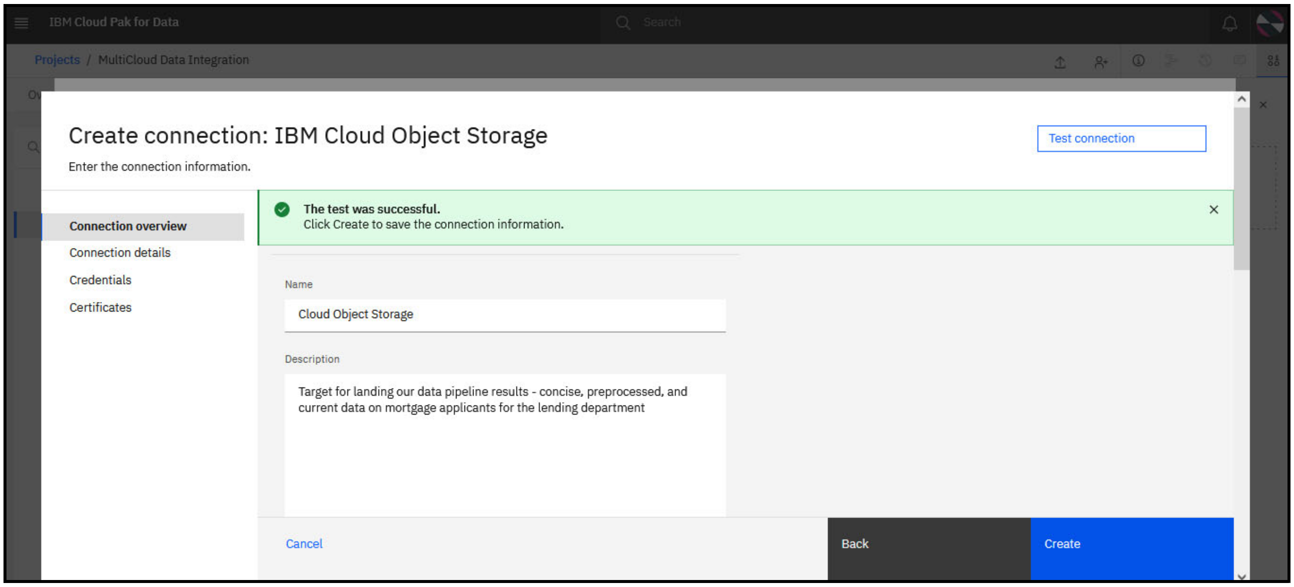

Figure 4-18 Creating a connection to IBM Cloud Object Storage

All the connections (source and target) for our transformation flow are now defined and we can proceed with the flow design and build (see Figure 4-19).

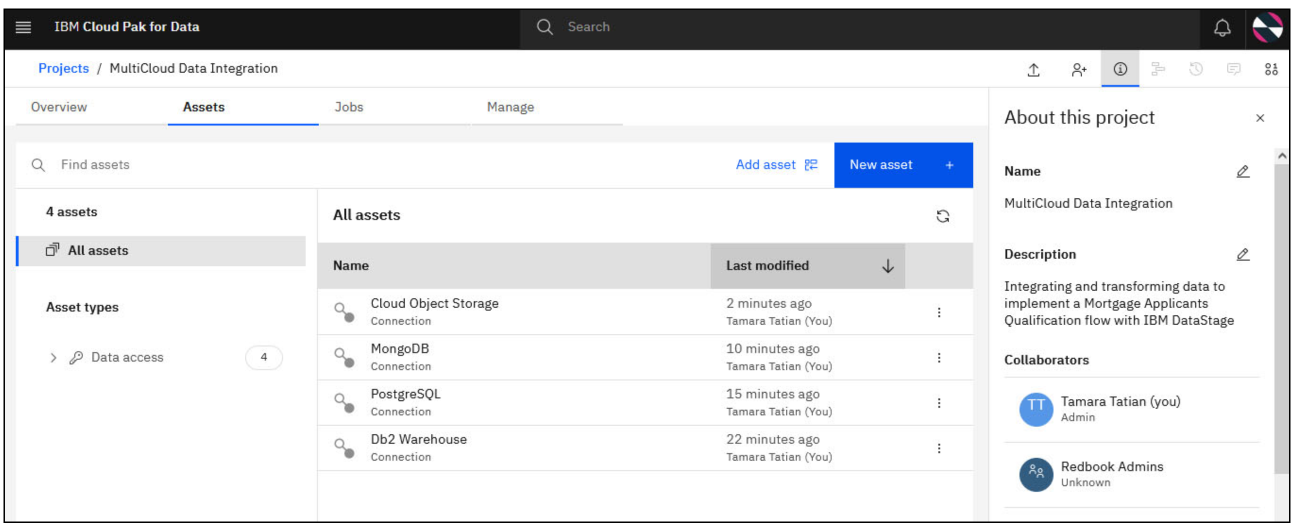

Figure 4-19 All connections created

The DataStage Flow Designer canvas can be started by adding a New Asset of type `DataStage`. The DataStage asset type is the main type of asset for the DataStage service. It starts a graphical canvas that provides a visual drag-and-drop flow designer interface with which you can easily compose transformation flows by using various pre-built standard data source connector and processing stage objects (see Figure 4-20).

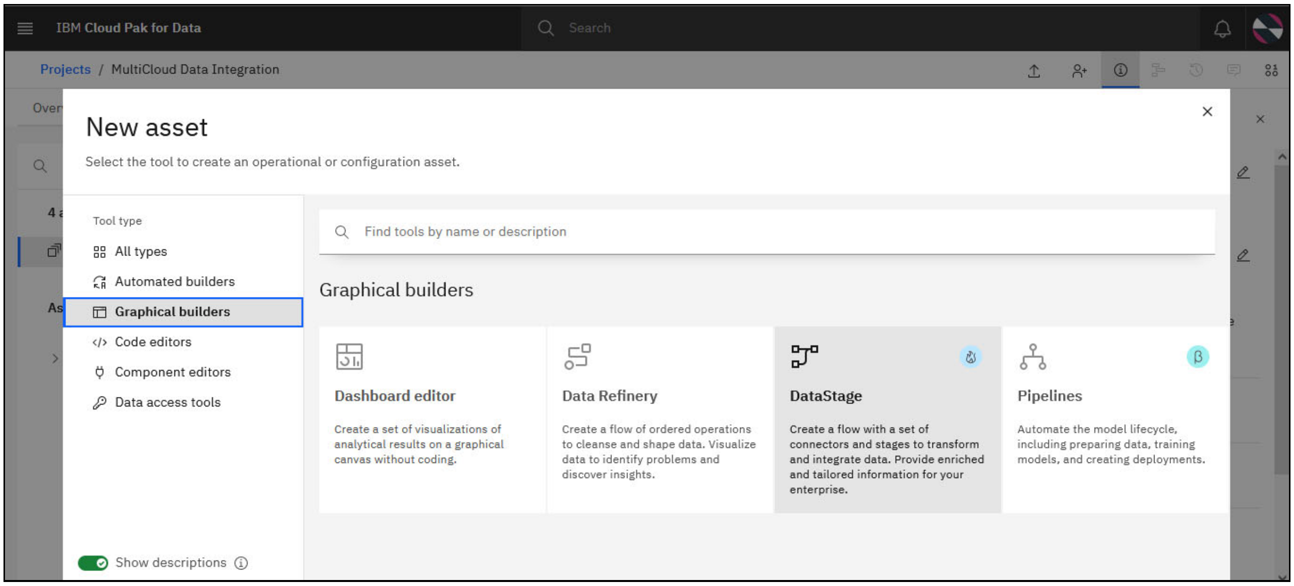

Figure 4-20 Asset types: Graphical builders

15. Other types of assets are specific to the service are DataStage components (subflows, libraries, data definitions, standardization rules, and custom stages) and Parameter Sets (see Figure 4-21).

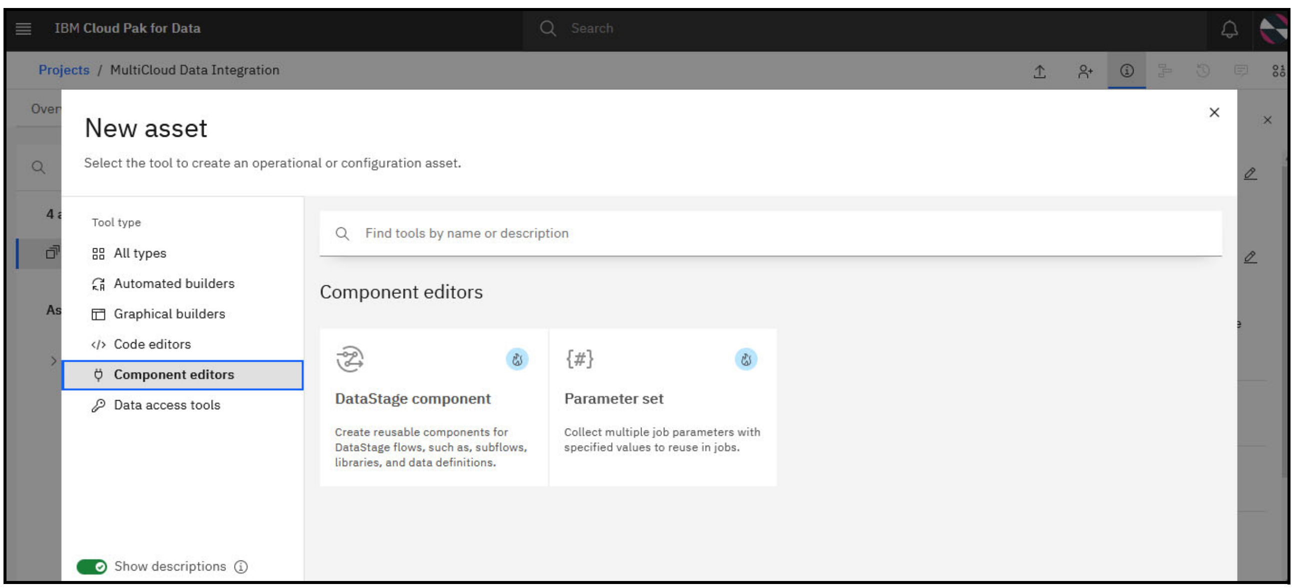

Figure 4-21 Asset types: Other relevant DataStage asset types

16.For our use case, we create only a New Asset of type DataStage; that is, a DataStage flow. Each flow must have a name and an optional description, and can be created from scratch or by uploading an ISX or compressed file that contains a created flow. In our case, we call our flow MultiCloud Data Integration, enter a relevant description for it, and click **Create** to create it without reusing any flows, as shown on Figure 4-22.

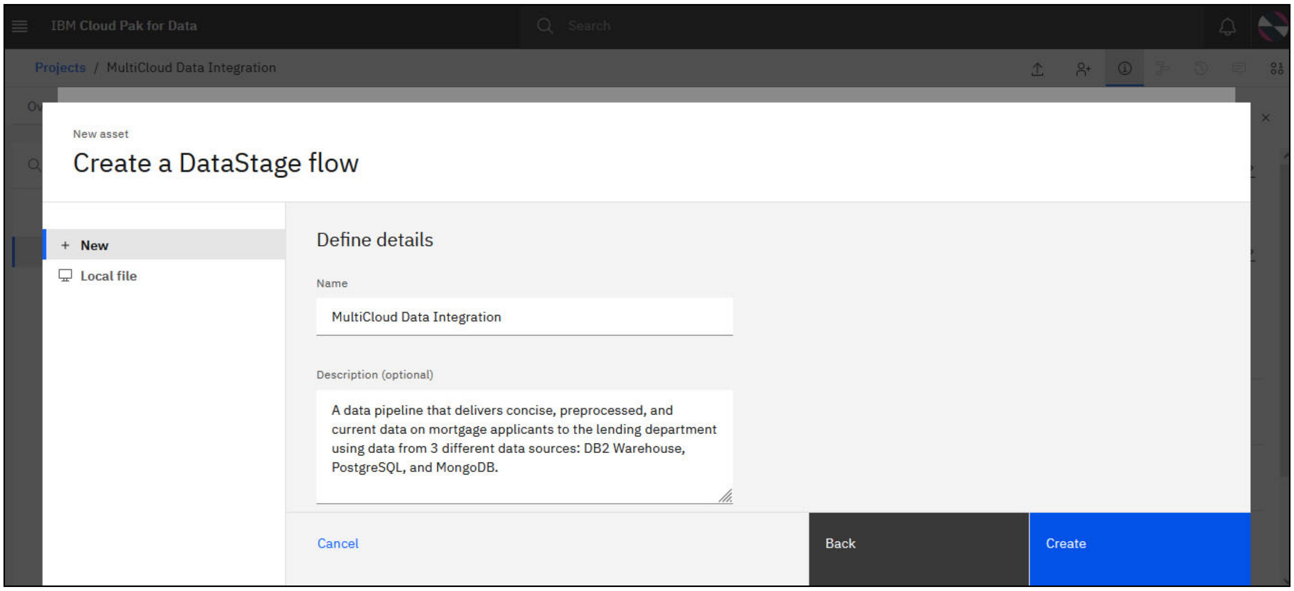

Figure 4-22 Creating a DataStage flow

17.After it is created, the DataStage flow asset allows you to browse to and work with the flow designer canvas. The flow designer helps you build your flow by selecting, dragging and dropping, connecting, and arranging relevant connectors and stages on the canvas. It also adds and edits relevant properties for the connectors and stages of the flow (see Figure 4-23).

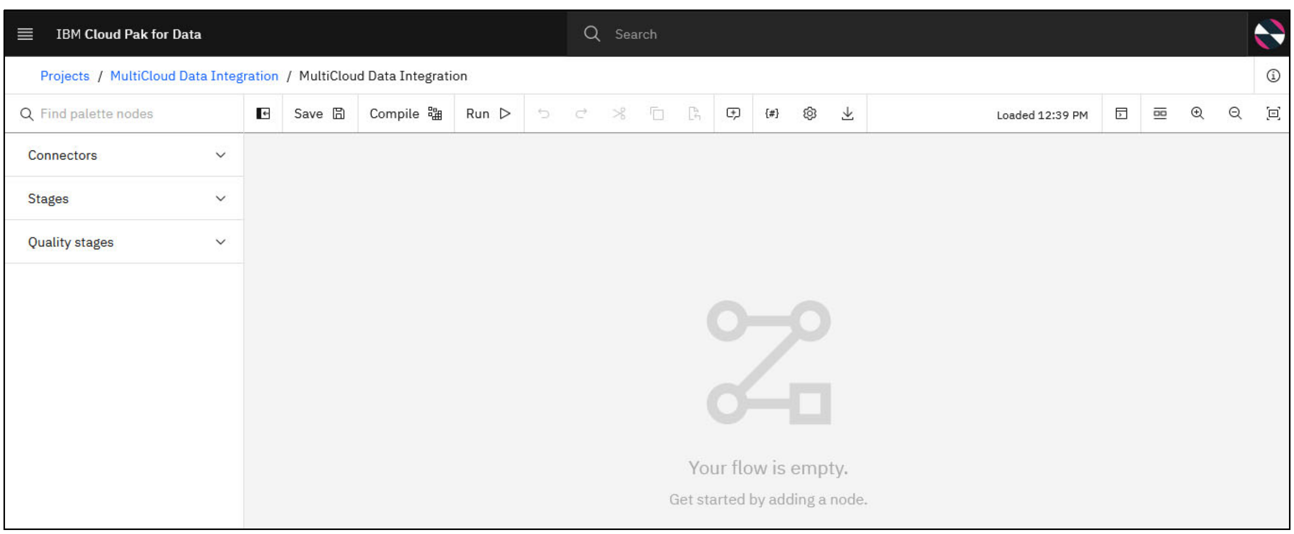

Figure 4-23 DataStage canvas view: A new DataStage flow

Figure 4-24 shows the connectors that are available and supported in IBM DataStage at the time of this writing (IBM Cloud Pak for Data version 4.5.2).

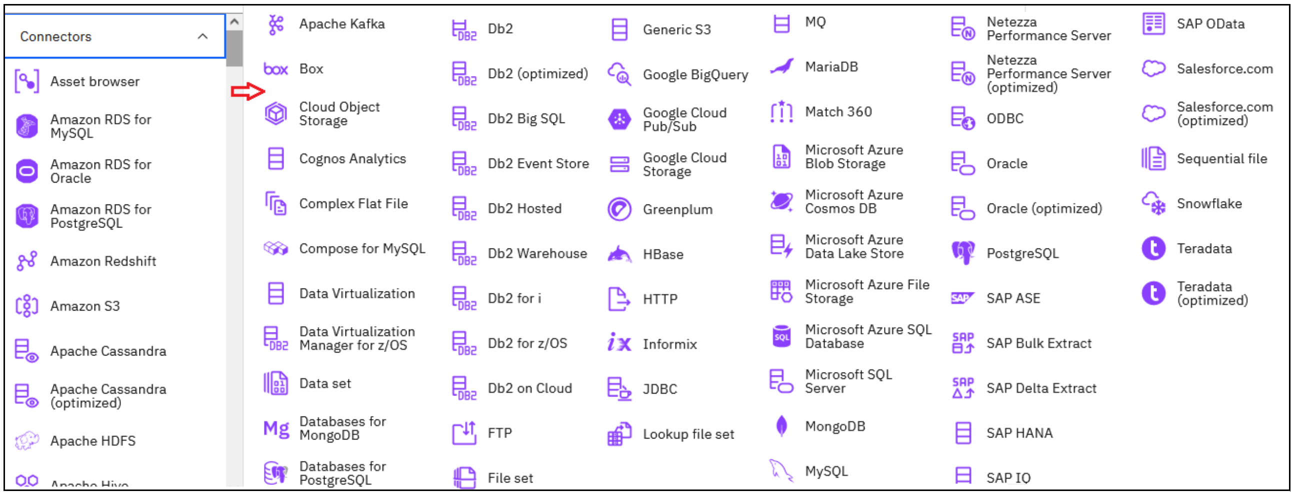

Figure 4-24 DataStage connectors list (IBM Cloud Pak for Data 4.5.2)

18. Dedicated data source connectors allow you to work with data from the corresponding connection types that were defined within your Project.

    By using the Asset browser connector, you can browse all Project data, including uploaded files, all of the connections of supported types that were defined in the Project, and all connected data in those projects. Depending on your final data selections, the connector then morphs into one or more data source-specific connector types on the canvas.

19. To start constructing the flow, we added our source Applicant and Applications data first. These tables are in the Db2 Warehouse connection types that were defined earlier. Dragging and dropping the Asset Browser connector unlocks the setup window, as shown in Figure 4-25.

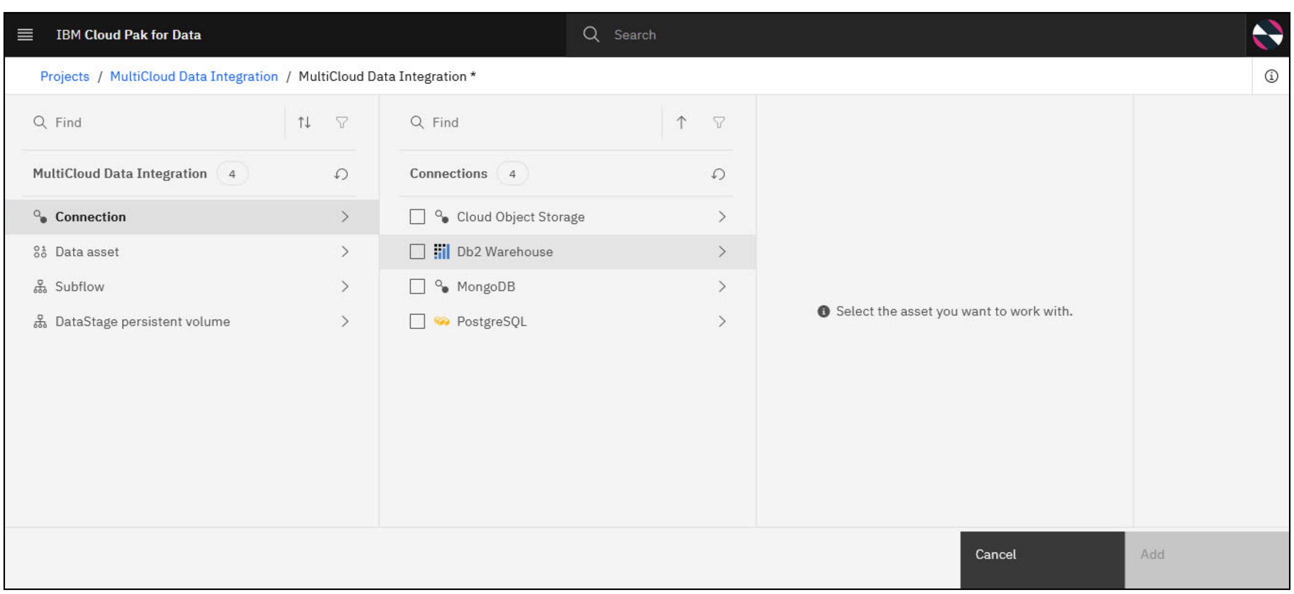

Figure 4-25 Connections browsing

20. Browsing through the available schemas, we selected **Mortgage_Applicants** and **Mortgage Applications** tables and then, added them to the canvas (see Figure 4-26).

Note: Asset Browser allows you to select multiple assets at once.

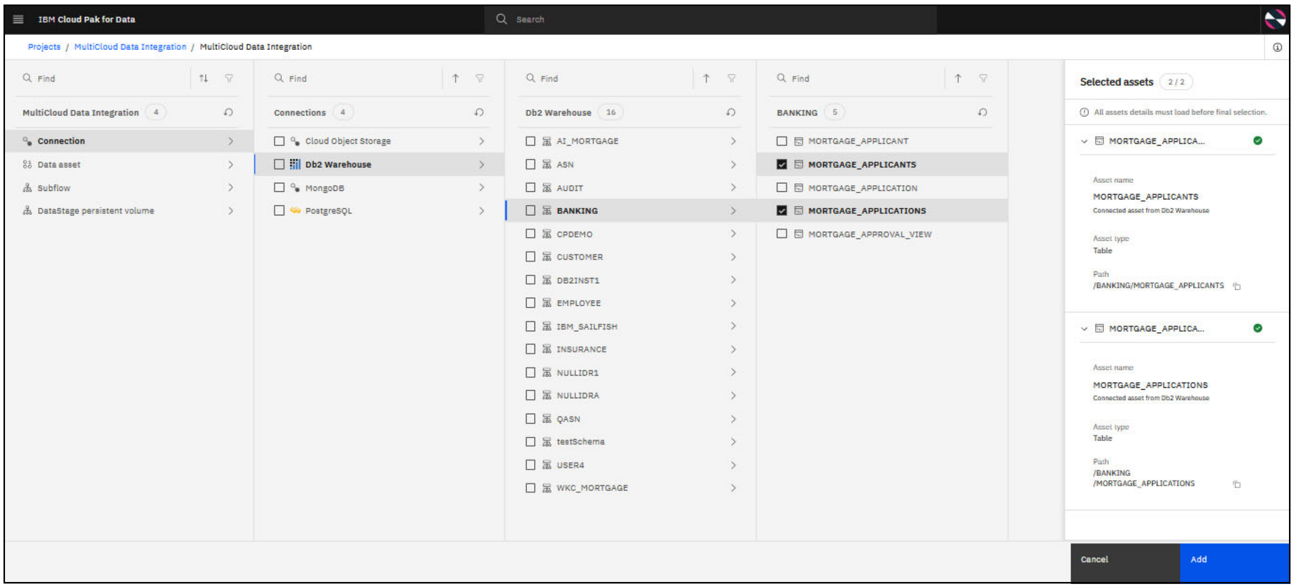

Figure 4-26 Selecting tables

The two tables now appear on the canvas, as shown in Figure 4-27.

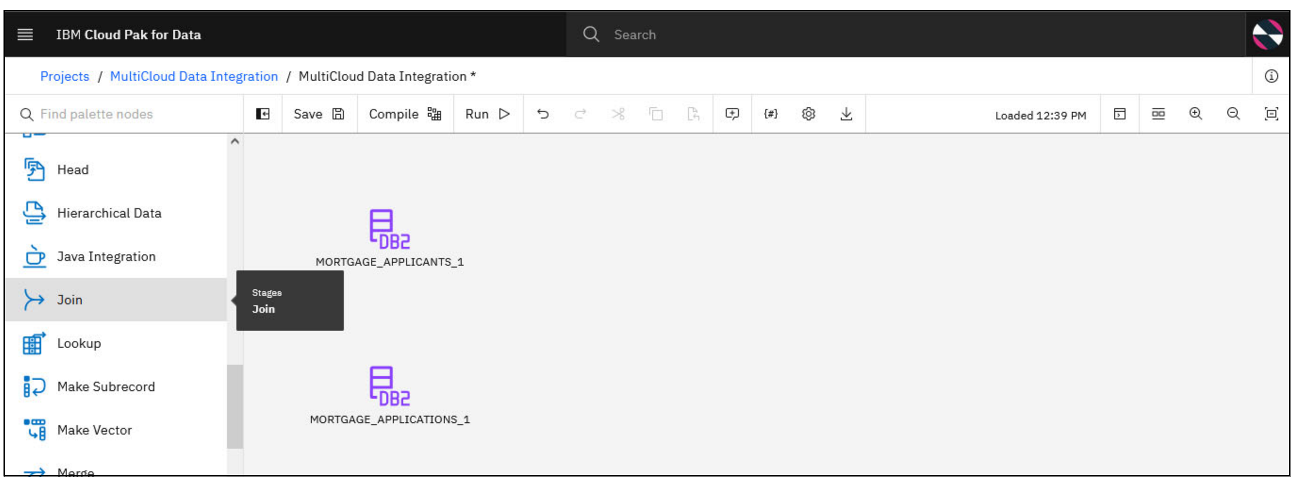

Figure 4-27 Connected data appearing on the canvas

21. We started building the transformation and integration logic of the flow. Figure 4-28 shows all the predefined stages that are included with DataStage Enterprise Plus version of the service in IBM Cloud Pak for Data 4.5.2.

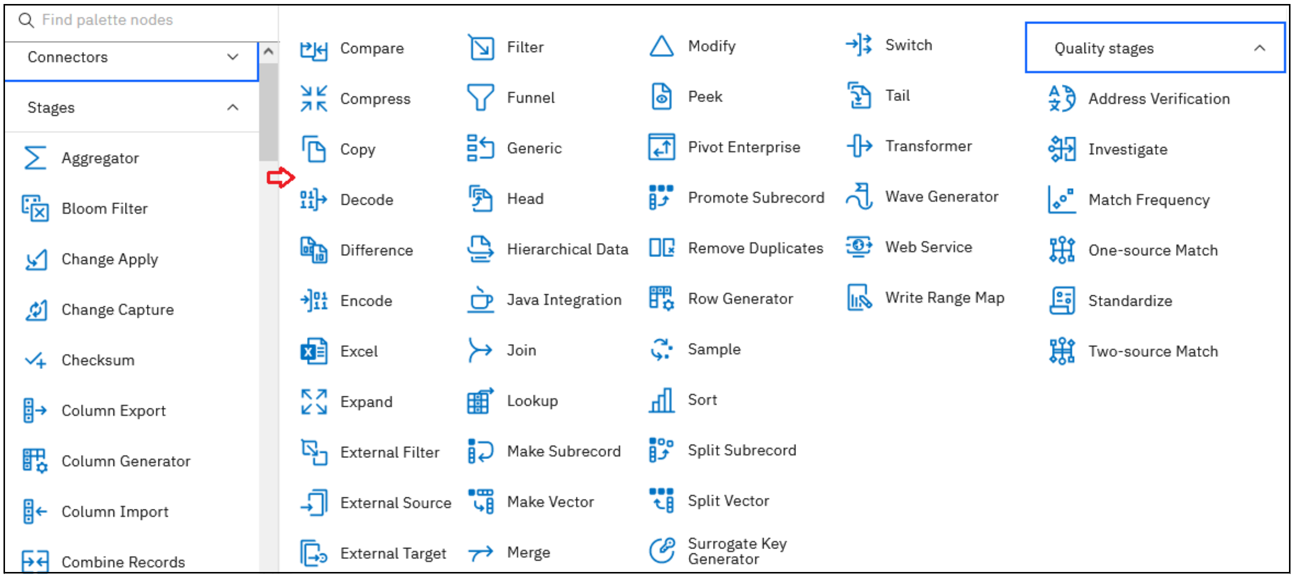

Figure 4-28 DataStage stages list (IBM Cloud Pak for Data 4.5.2, IBM DataStage Enterprise Plus service version)

A *stage* defines the processing logic that moves and transforms data from its input links to its output links. Stages often include at least one data input or one data output. However, some stages can accept more than one data input and output to more than one stage.

Table 4-2 lists the processing stages and their functions.

Table 4-2 DataStage Enterprise stages and their functions

Stage	Function
Aggregator	Classifies incoming data into groups, computes totals and other summary functions for each group, and passes them to another stage in the job.
Bloom Filter	Looks up incoming keys against previous values.
Build stage	Creates a custom operator that can be used in a DataStage flow. The code for a Build stage is specified in C++.
Change Apply	Applies encoded change operations to a before data set based on a changed data set. The before and after data sets come from the Change Capture stage.
Change Capture	Compares two data sets and makes a record of the differences.
Checksum	Generates a checksum value from the specified columns in a row and adds the checksum to the row.
Column Export	Exports data from a number of columns of different data types into a single column of data types ustring, string, or binary.
Column Import	Imports data from a single column and outputs it to one or more columns.
Column Generator	Adds columns to incoming data and generates mock data for these columns for each data row processed.
Combine Records	Combines records in which particular key-column values are identical into vectors of subrecords.

Stage	Function
Compare	Performs a column-by-column comparison of records in two presorted input data sets.
Compress	Uses the UNIX compress or GZIP utility to compress a data set. It converts a data set from a sequence of records into a stream of raw binary data.
Copy	Copies a single input data set to a number of output data sets.
Data Set	Reads data from or writes data to a data set.
Decode	Decodes a data set by using a UNIX decoding command that you supply.
Difference	Performs a record-by-record comparison of two input data sets, which are different versions of the same data set.
Encode	Encodes a data set by using a UNIX encoding command that you supply.
Excel	Extracts information from Excel files.
Expand	Uses the UNIX decompress or GZIP utility to expand a data set. It converts a previously compressed data set back into a sequence of records from a stream of raw binary data.
External Filter	Allows you to specify a UNIX command that acts as a filter on the data you are processing.
External Source	Reads data that is output from one or more source programs.
External Target	Writes data to one or more source programs.
File Set	Reads data from or writes data to a file set.
Filter	Transfers, unmodified, the records of the input data set that satisfies requirements that you specify and filters out all other records.
Funnel	Copies multiple input data sets to a single output data set.
Generic	Incorporates an IBM Orchestrate® Operator in your job.
Head	Selects the first N records from each partition of an input data set and copies the selected records to an output data set.
Hierarchical (XML)	Composes, parses, and transforms JSON and XML data.
Join	Performs join operations on two or more data sets input to the stage and then, outputs the resulting data set.
Lookup	Performs lookup operations on a data set that is read into memory from any other Parallel job stage that can output data or provide by one of the database stages that support reference output links. It also can perform a lookup on a lookup table that is in a Lookup File Set stage.
Make Subrecords	Combines specified vectors in an input data set into a vector of subrecords whose columns include the names and data types of the original vectors.
Make Vector	Combines specified columns of an input data record into a vector of columns.
Merge	Combines a sorted master data set with one or more sorted update data sets.
Modify	Alters the record schema of its input data set.
Peek	Prints record column values to the job log or a separate output link as the stage copies records from its input data set to one or more output data sets.
Pivot Enterprise	The Pivot Enterprise stage is a processing stage that pivots data horizontally and vertically: ► Horizontal pivoting maps a set of columns in an input row to a single column in multiple output rows. ► Vertical pivoting maps a set of rows in the input data to single or multiple output columns.

Stage	Function
Promote Sub-records	Promotes the columns of an input subrecord to top-level columns.
Remove Duplicates	Takes a single sorted data set as input, removes all duplicate records, and writes the results to an output data set.
Row Generator	Produces a set of mock data fitting the specified metadata.
Sample	Samples an input data set.
Sort	Sorts input columns.
Split Subrecord	Separates an input subrecord field into a set of top-level vector columns.
Split Vector	Promotes the elements of a fixed-length vector to a set of similarly named top-level columns.
Surrogate Key Generator stage	Generates surrogate key columns and maintains the key source.
Switch	Takes a single data set as input and assigns each input record to an output data set that is based on the value of a selector field.
Tail	Selects the last N records from each partition of an input data set and copies the selected records to an output data set.
Transformer	Handles extracted data, performs any conversions that are required, and passes data to another active stage or a stage that writes data to a target database or file.
Wave Generator	Monitors a stream of data and inserts end-of-wave markers where needed.
Wrapped	Specifies a UNIX command to be run by a DataStage stage.
Write Range Map	Writes data to a range map. The stage can have a single input link.

For more information about these stages, see this IBM Documentation web page.

DataStage Enterprise Plus version of the service adds IBM InfoSphere QualityStage® stages for investigating, cleansing, and managing your data. With these data quality stages, you can manipulate your data in the following ways:

- Resolve data conflicts and ambiguities.
- Uncover new or hidden attributes from free-form or loosely controlled source columns.
- Conform data by transforming data types into a standard format.

Table 4-3 lists the IBM InfoSphere QualityStage stages that are included with IBM Cloud Pak for Data 4.5.2.

Table 4-3 QualityStage stages and their functions

Stage	Function
Address Verification	Provides comprehensive address parsing, standardization, validation, geocoding, and reverse geocoding, which is available in selected packages against reference files for over 245 countries and territories.
Investigate	Shows the condition of source data and helps to identify and correct data problems before they corrupt new systems. Understanding your data is a necessary precursor to cleansing.
Match Frequency	Generates the frequency distribution of values for columns in the input data. You use the frequency distribution and the input data in match jobs.
One-source Match	Matches records from a single source file.
Two-source Match	Compares two sources of input data (reference records and data records) for matches.

Stage	Function
Standardize	Makes your source data internally consistent so that each data type features the same type of content and format.

For more information about these stages, see this IBM Documentation web page.

22. In our use case, we are primarily working with the Join, Transformer, and Lookup stages. First, we must join data in the two tables that we added. The Join stage from the menu on the left side of the canvas must be dropped next to the tables (see Figure 4-29).

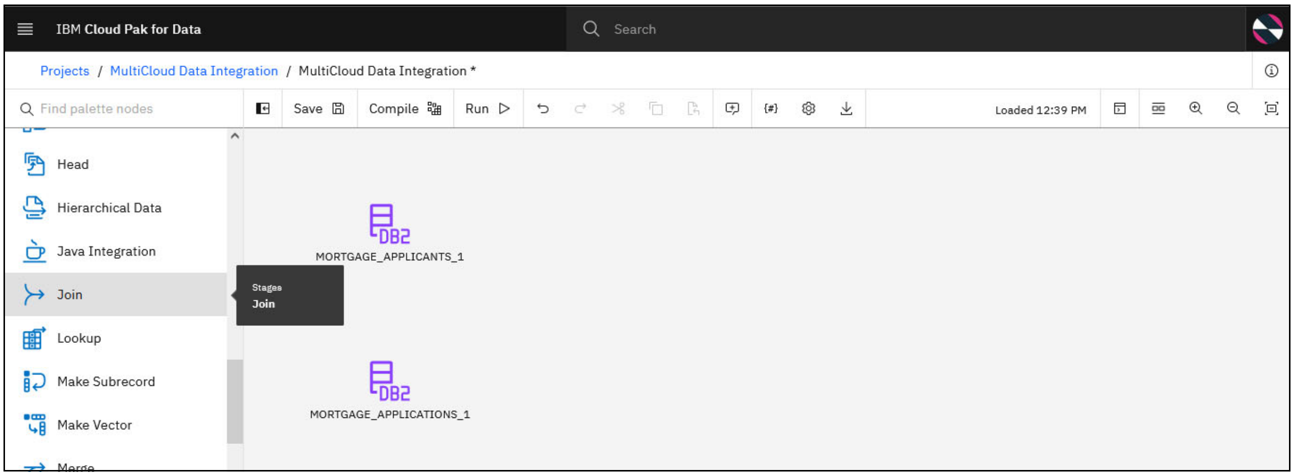

Figure 4-29   Adding a Join stage: Part 1

23. The tables must become input sources for the Join stage. Therefore, we click the icon that represents each table and then, create a link by dropping the **circled arrow sign** onto the Join stage.

    The name of the link can be edited by double-clicking it (see Figure 4-30).

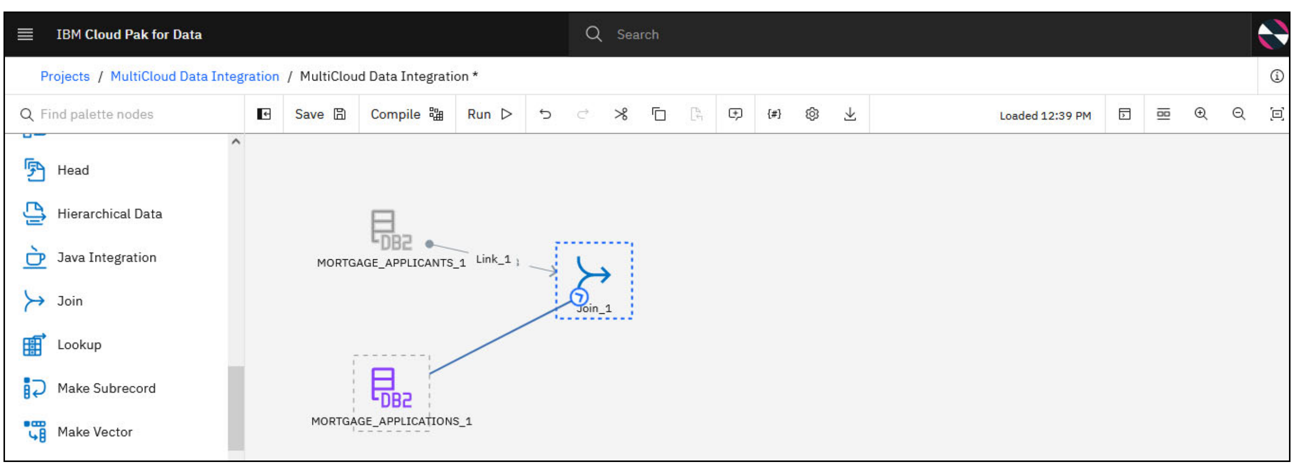

Figure 4-30   Adding a Join stage: Part 2

Similarly, by double-clicking the table and stage names on the canvas, the names can be edited. Double-clicking the connector or stage enables editing the settings of each component.

24. We set up the Join stage next. First, we renamed it to `Join_on_ID` to make the flow that is better documented and more readable. then, we added the join key (see Figure 4-31).

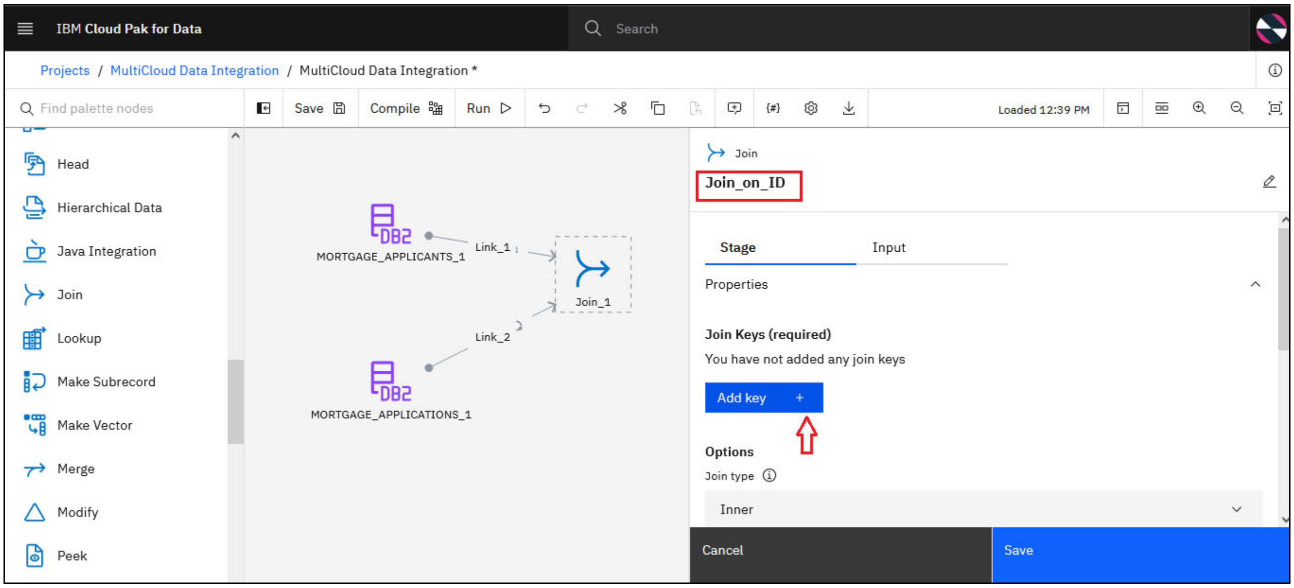

Figure 4-31 Join stage setup

25. We select the ID key because it is the ideal join key for our two tables (see Figure 4-32 and Figure 4-32 on page 184).

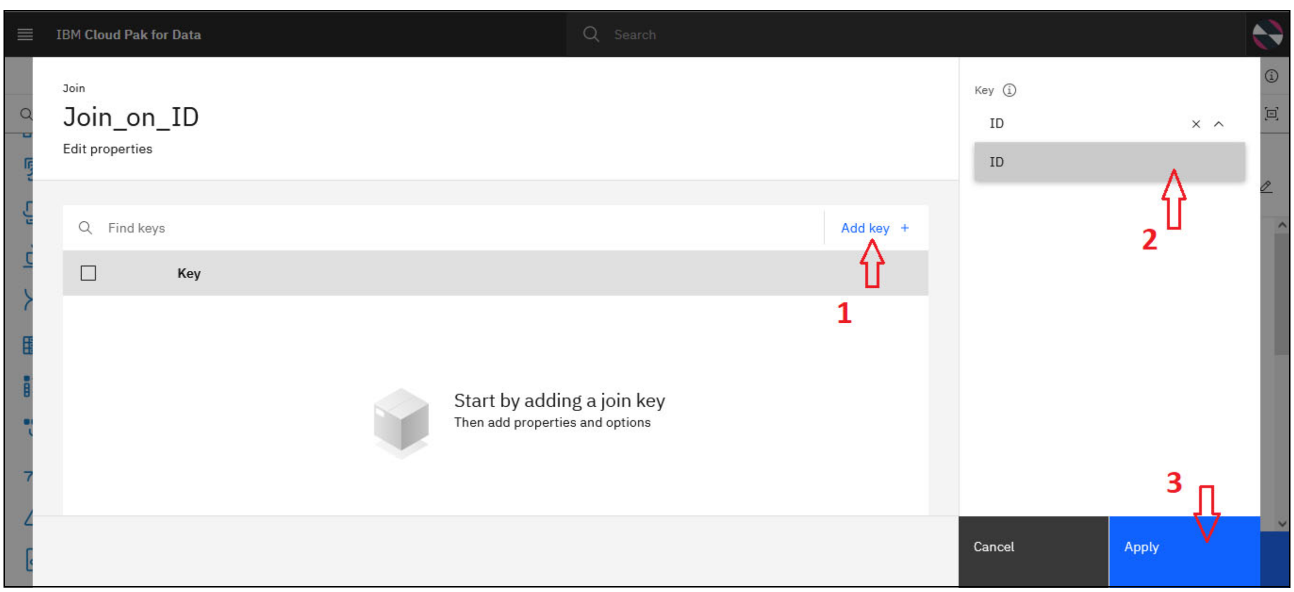

Figure 4-32 Adding join keys

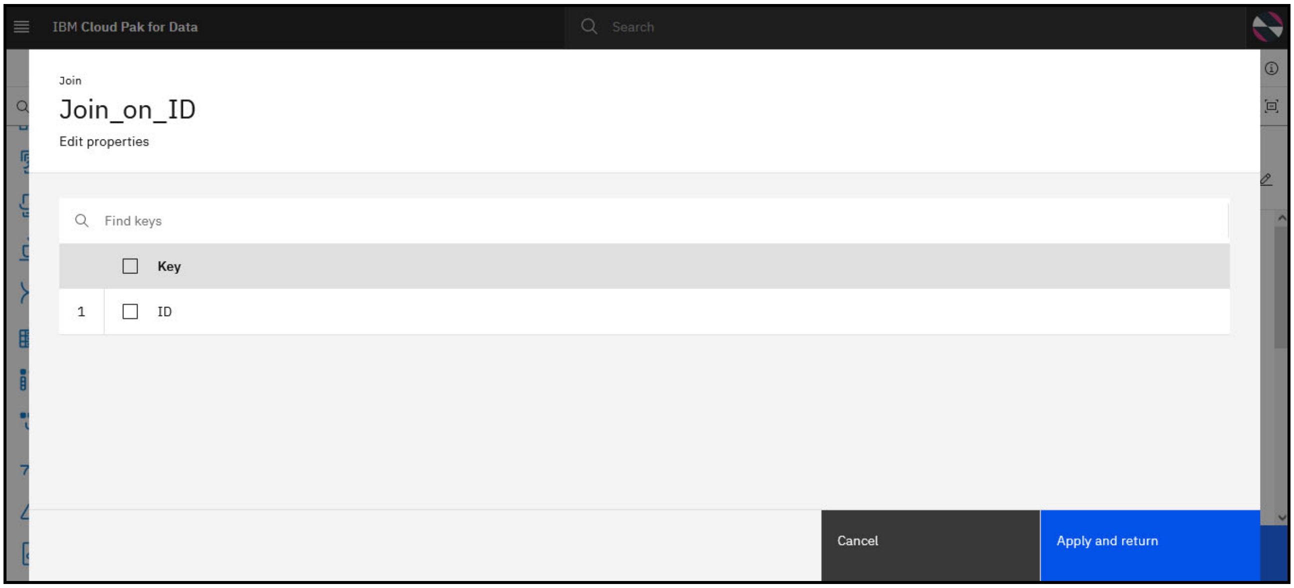

Figure 4-33 Adding join keys: Continued

26. After the correct key is selected, we click **Apply and return**. We are returned to the main setup window, as shown on Figure 4-34.

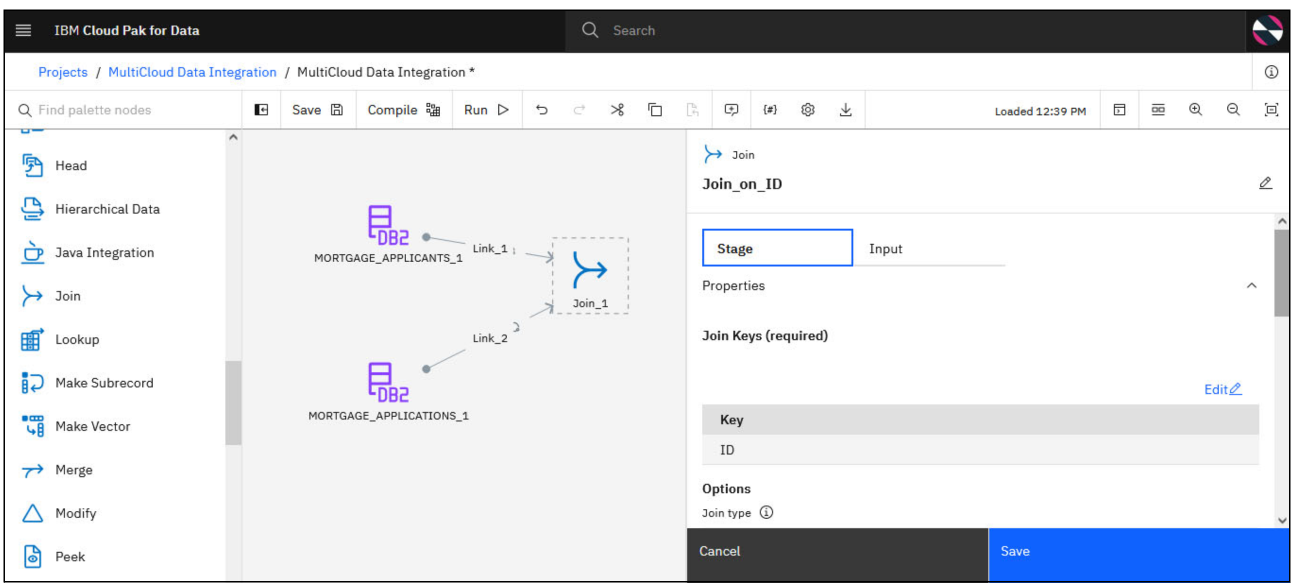

Figure 4-34 Join keys setup complete

27. The join type and reorder input links can be edited, if necessary. In our case, we proceed with the default Inner join option (see Figure 4-35).

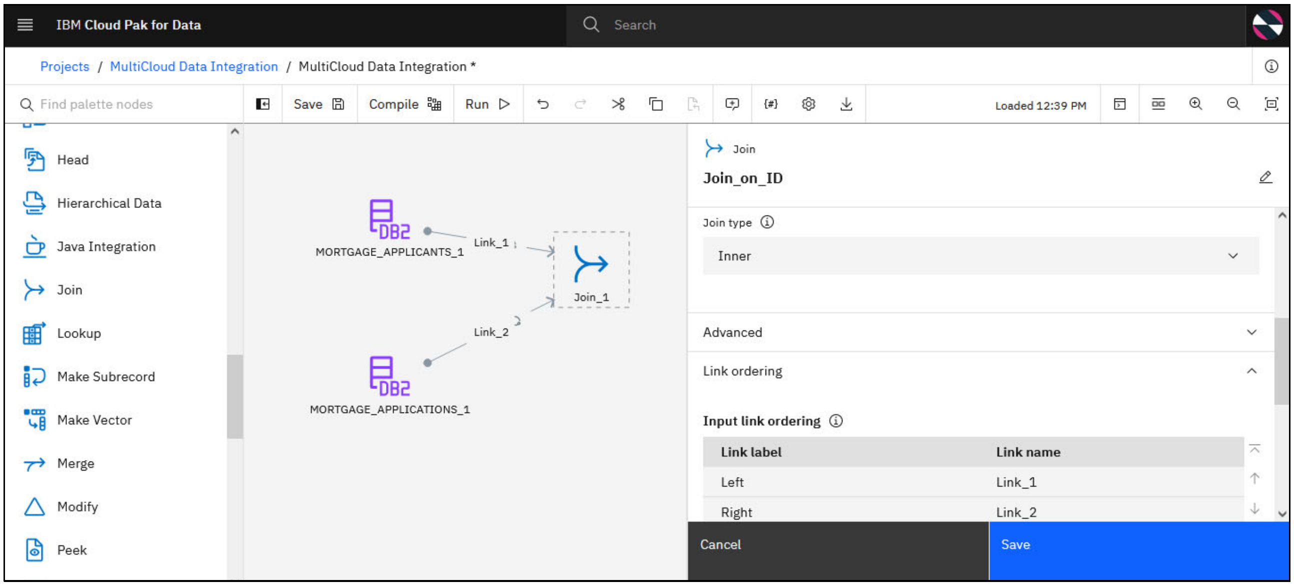

Figure 4-35 Join type settings

We also can change partitioning settings, if needed, and review columns that come from each of the input links and their associated data types. DataStage automatically deduced and imported that metadata from the source connected tables (see Figure 4-36).

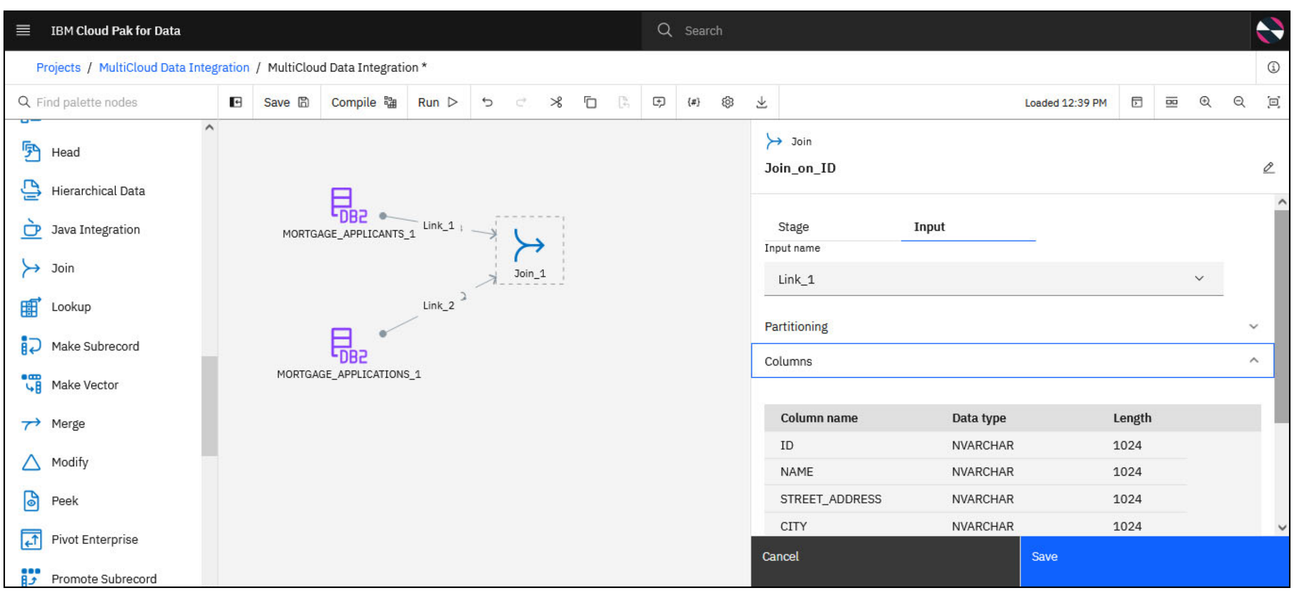

Figure 4-36 Viewing input columns coming from join source links

Because nothing was added to the flow after the Join, Output settings cannot be edited.

Also, the column details from the input links are read-only and cannot be edited. If we wanted to refine the column selection and link inputs, those changes can be made by editing the table assets that we added by way of the Asset Browser.

To explore how we can do that, we clicked **Save** the Join stage and then, double-clicked the **Mortgage_Applicants table** that is on the canvas.

The Output tab of the connector allows us to edit what is output into the next processing stage by way of the link through which the table and stage are connected (see Figure 4-37).

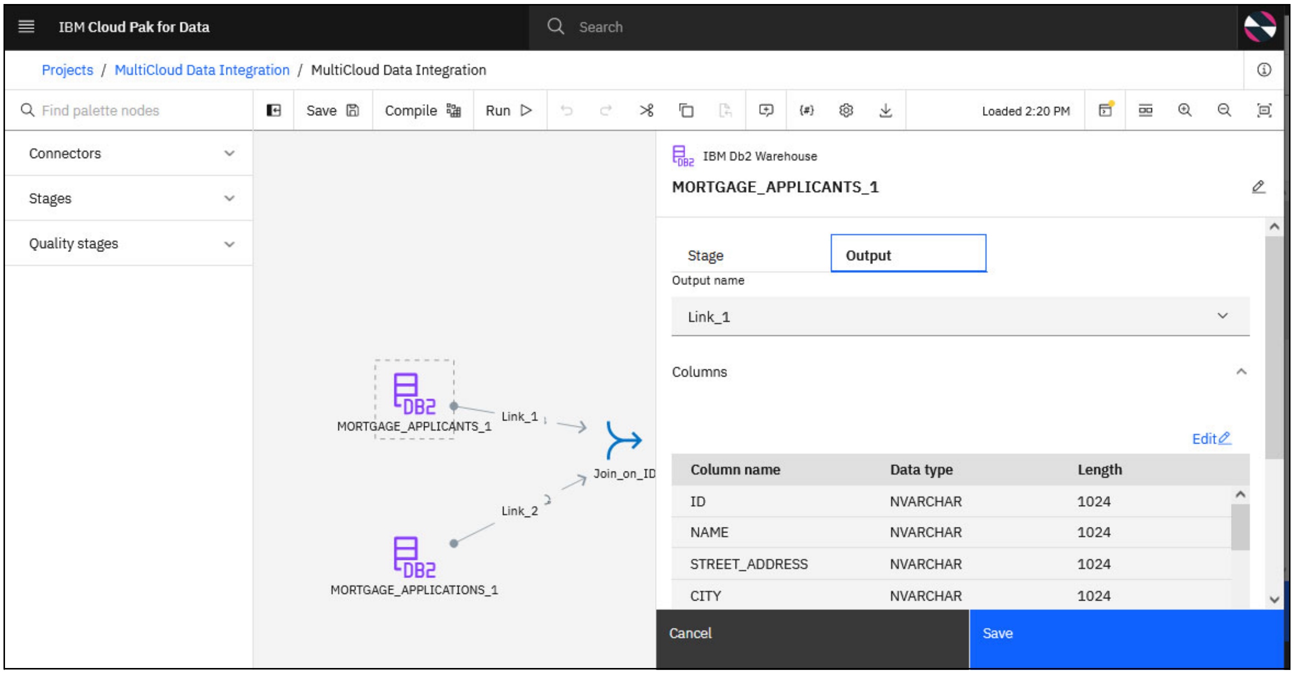

Figure 4-37 Viewing and modifying connector outputs

28. We clicked **Edit** as shown on Figure 4-37, which took us to the window in which we can add or exclude columns, import or export data definitions, and change nullability and key settings for the columns, reorder columns, change their names, descriptions, and so forth (see Figure 4-38).

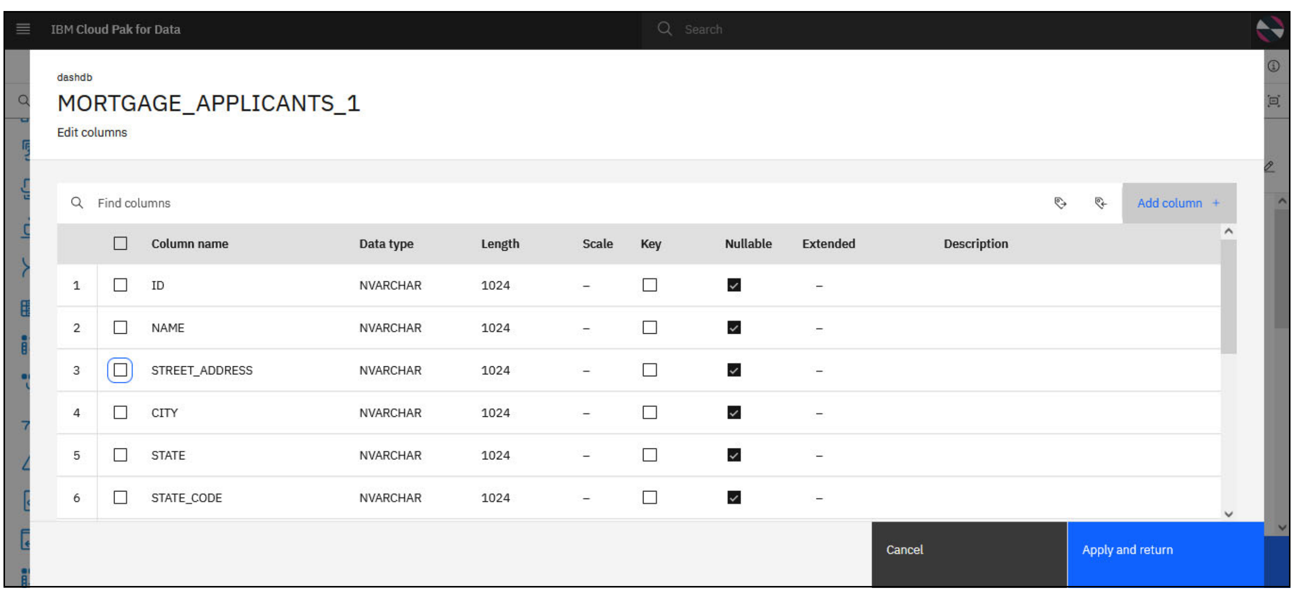

Figure 4-38 Editing column details

29. We now try to understand the data coming from this source. We clicked **Cancel** or **Apply and return** to return us to the main setup window in which we can switch from the Output tab to the Stage tab. By clicking the **Stage** tab, we can review, and, if necessary, edit the source details for the table (see Figure 4-39).

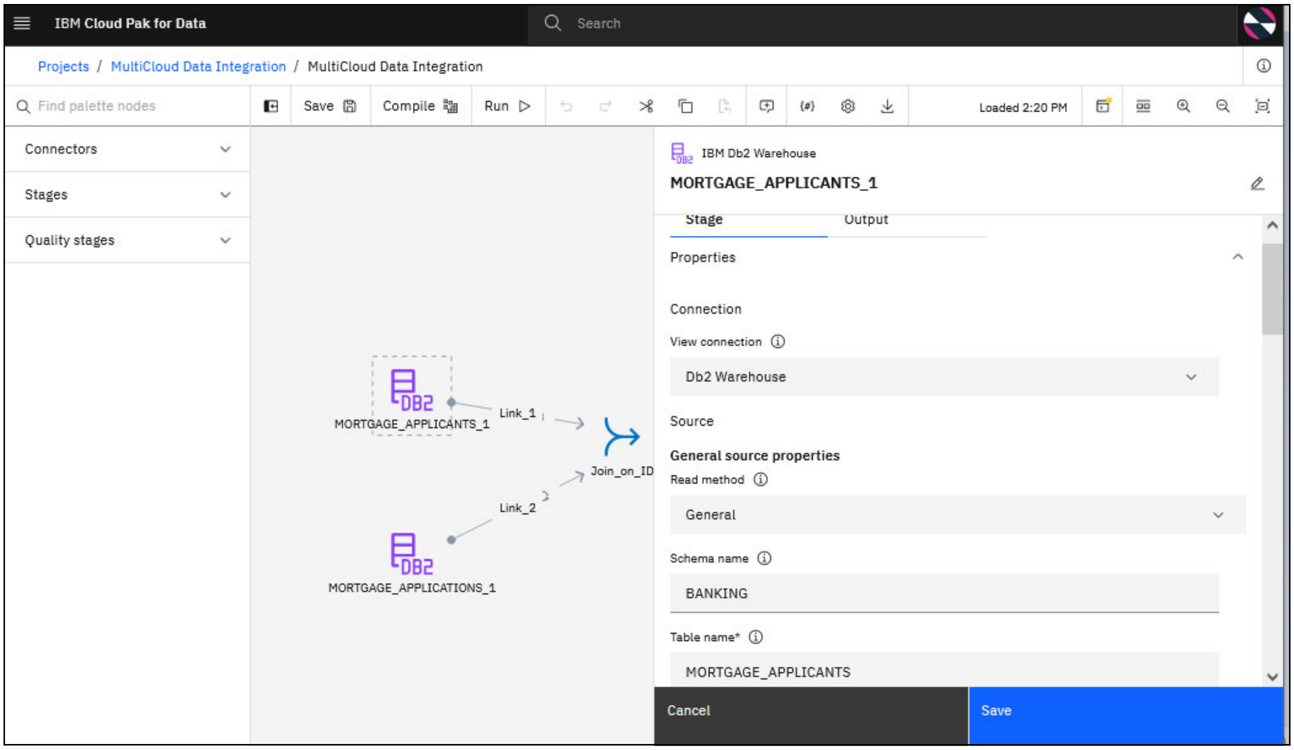

Figure 4-39 Stage tab

DataStage also enables you to run various Before and After SQL statements, specify flow termination conditions if any of the SQL queries fail, and preview the source data to make sense of it (see Figure 4-40).

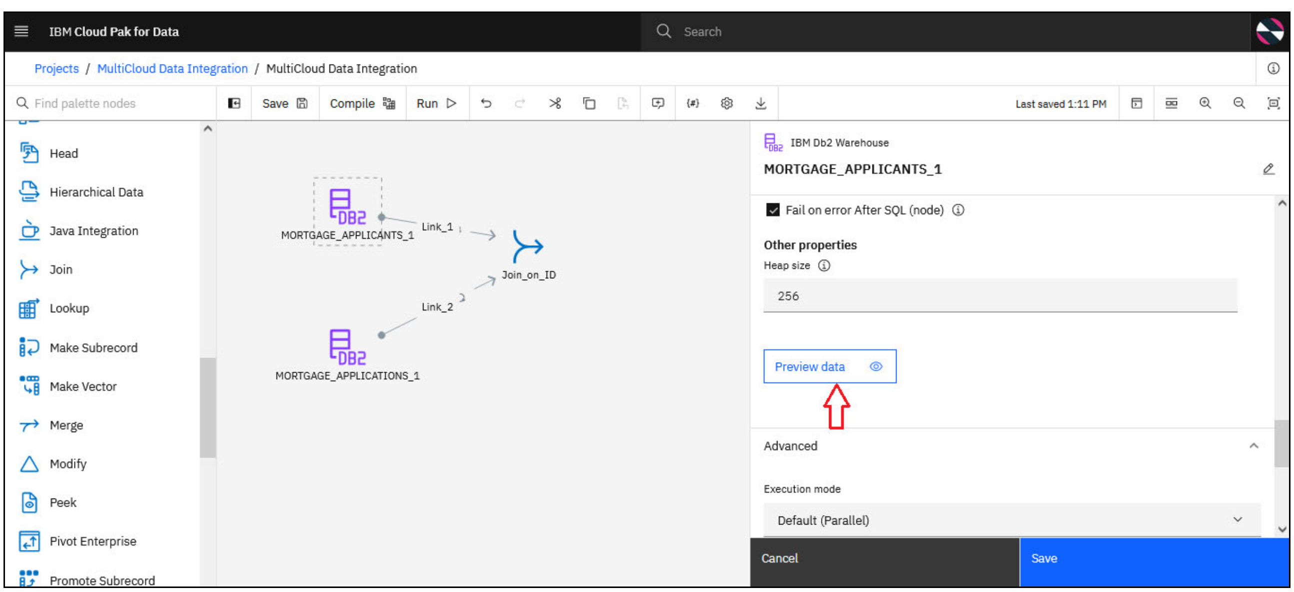

Figure 4-40 Previewing data

IBM DataStage includes with a rich set of capabilities for interactive data preview and trends analysis and visualization. Clicking the **Preview data** in any of the connector objects brings you to the following windows: Data, Chart, Profile, and Exploratory Analysis. These windows help you to explore and make sense of your data.

The Data tab shows you a preview of a sample of your data (see Figure 4-41).

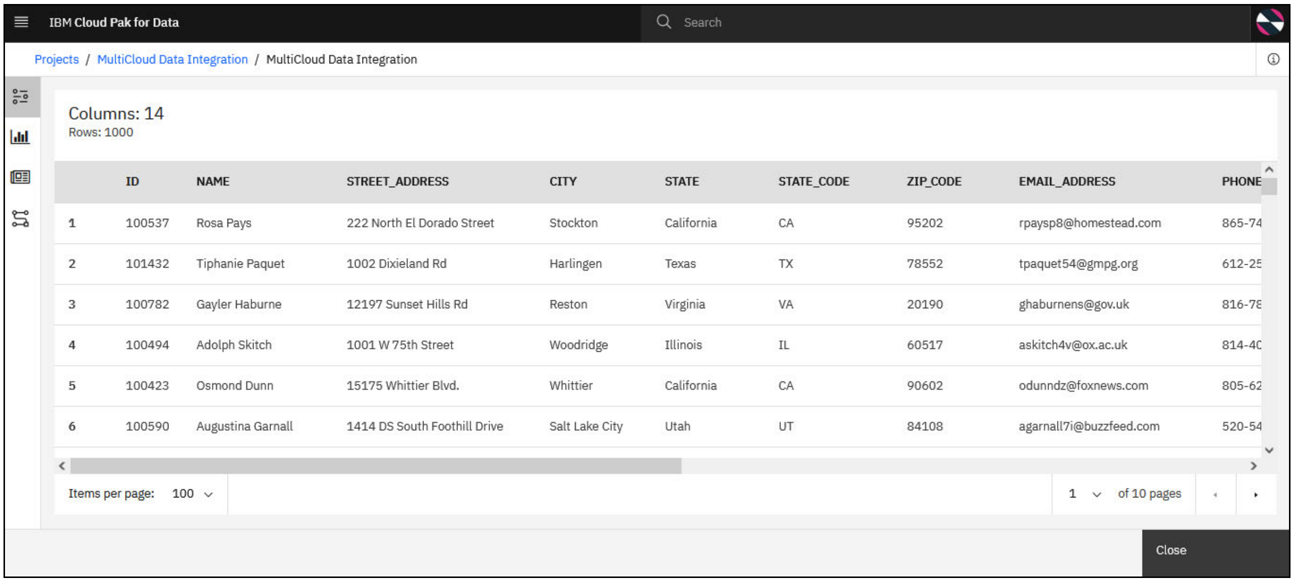

Figure 4-41 Previewing data

The Chart tab allows you to dynamically build various charts that are based on the preview sample of the data (see Figure 4-42).

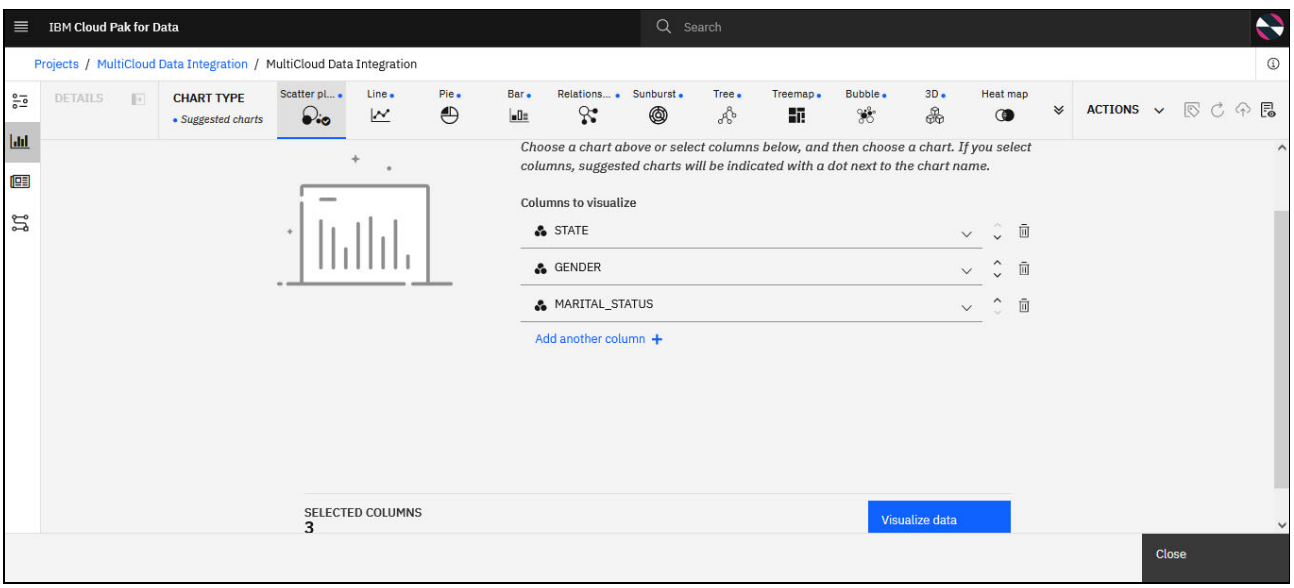

Figure 4-42 Previewing data - visualizing data

30. The solution automatically suggests the best fitting visualization types that are based on the types and number of columns you select. Alternatively, you can specify which visualization type you prefer. In our case, we built a Relationship graph that is based on `State`, `Gender` and `Marital_Status` columns in our data set (see Figure 4-43).

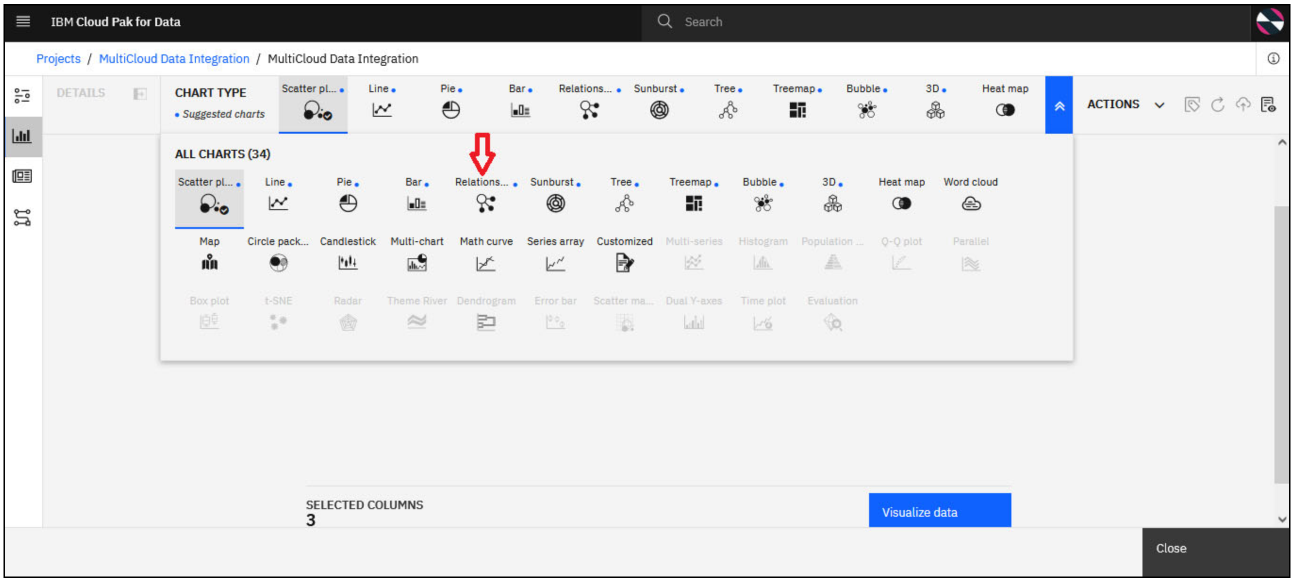

Figure 4-43   Relationship charts

The charts are interactive and you can highlight, click into, and zoom into specific data, depending on the chart type that is selected. They also be saved as an image or a visualization into the parent Project of the flow (see Figure 4-44).

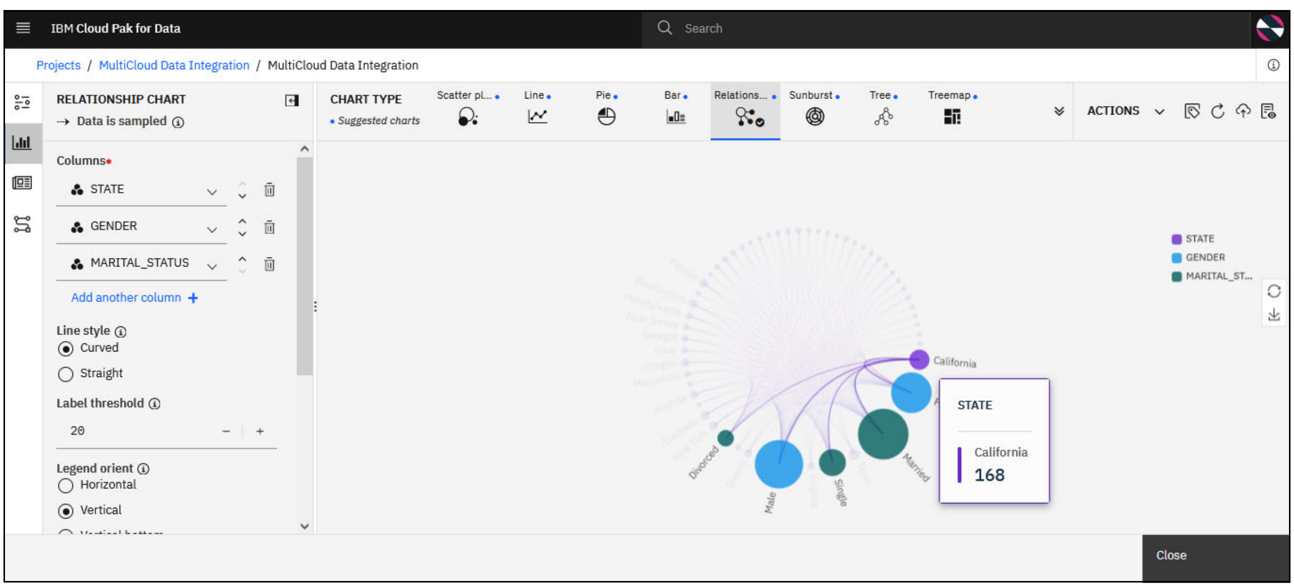

Figure 4-44   Relationship chart for our data

The Profile tab provides statistical and frequency analysis results for your data (Audit view, see Figure 4-45), and data quality insights (Quality view, see Figure 4-46).

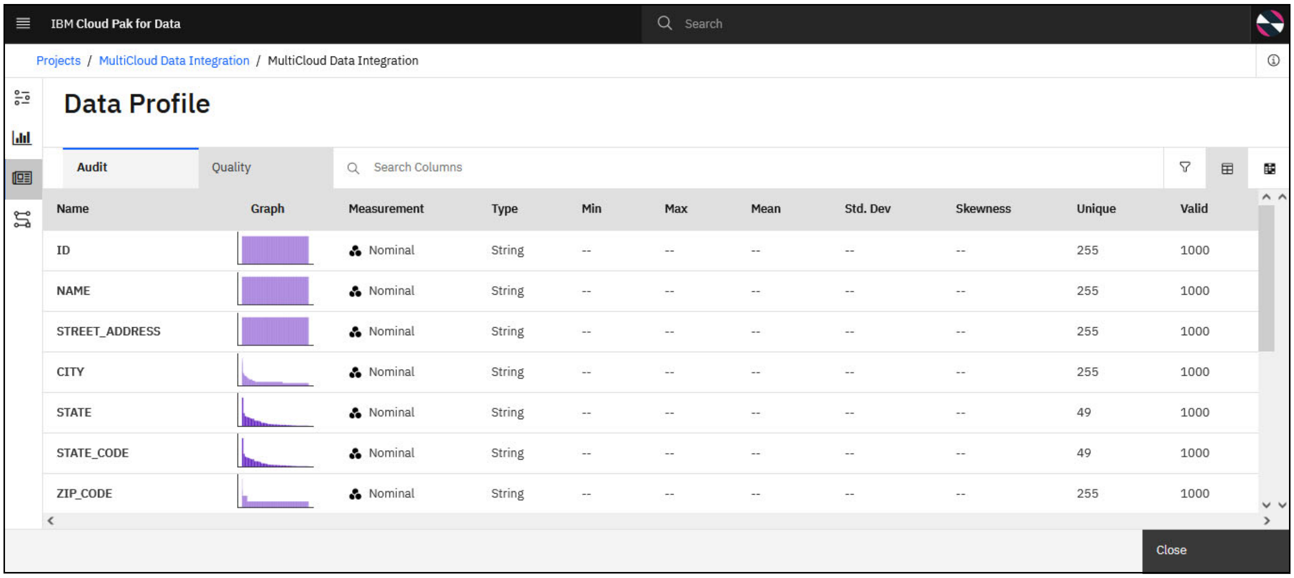

Figure 4-45 Data profile

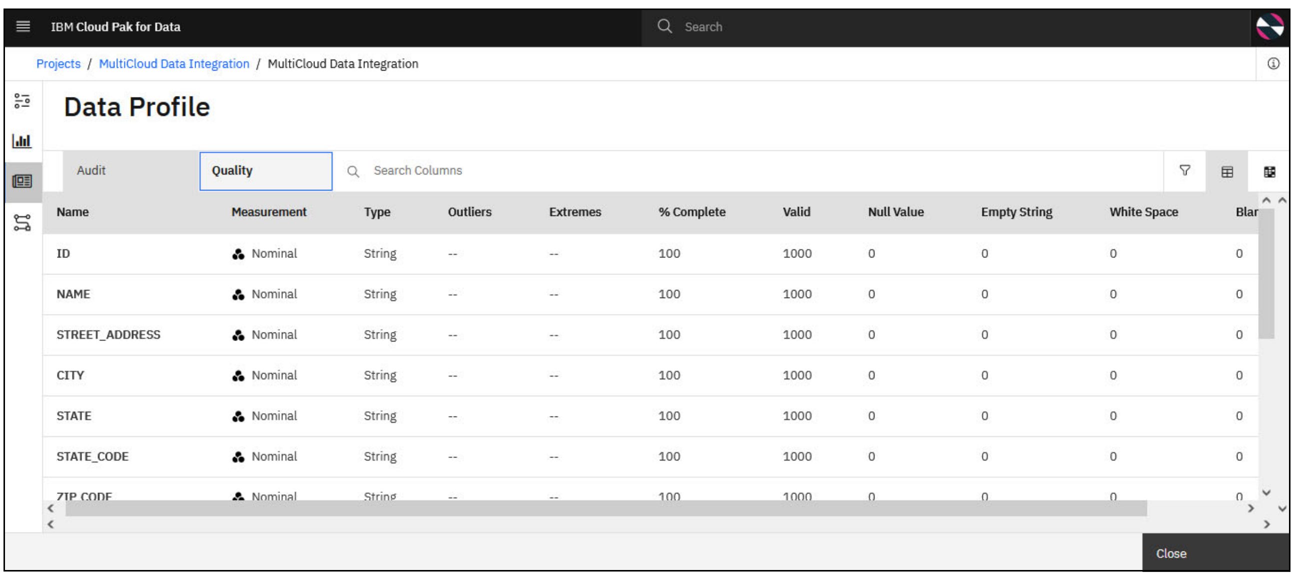

Figure 4-46 Data quality profile

31. By using the Exploratory Analysis tab, you can perform other descriptive statistics analyses (see Figure 4-47).

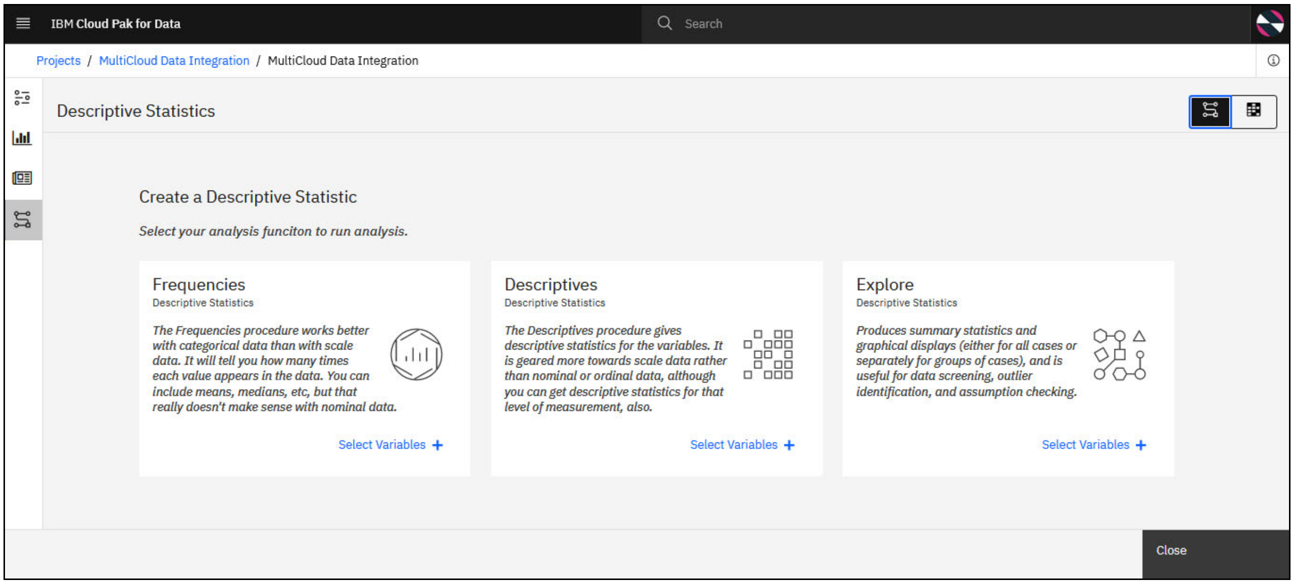

Figure 4-47 Previewing data - descriptive statistics analyses

32. Having explored and understood our source data better and joined the two source tables by the ID key, we now perform a quick test and check our progress.

    To successfully compile and test run our flow, it must be valid. In our case, the Join stage requires at least one output link. We cannot proceed until this condition is met.

    To complete our flow, we temporarily added a Peek stage element to the canvas. The Peek stage prints record column values to the job log or a separate output link (in our case, we use the former option).

    We clicked the **Join stage** on the canvas and then, double-clicked the **Peek stage** on the left side menu to easily create a Peek object on the canvas that is automatically connected to our Join object.

33. After the Peek stage is in place, we compiled and ran the flow by selecting **Compile** and then, **Run** on the menu bar. The run succeeded, which confirmed that we are on the correct track.

We can further validate by reviewing the job logs, if necessary. The logs and status messages that are generated during job run flag errors that are in the flow and its setup, and any other issues that the integration flow encountered (see Figure 4-48).

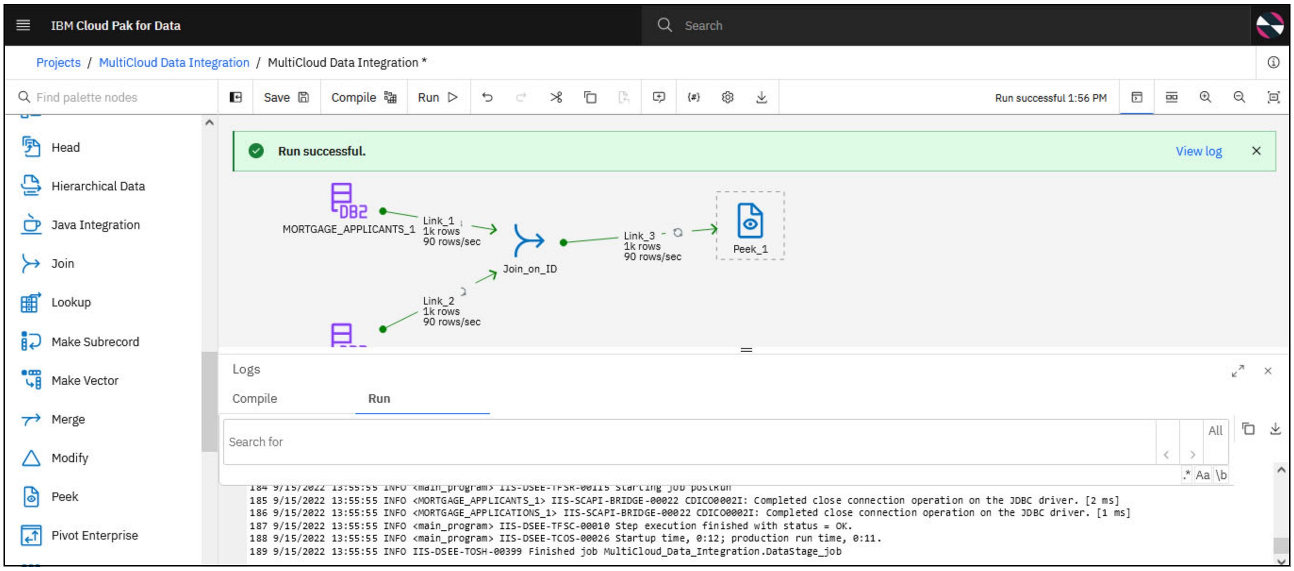

Figure 4-48 Interim testing and validation of your flow with Peak stage and job run logs

> **Note:** Consider the following points:
>
> - The menu that is at the top of the canvas provides dedicated Save, Compile, Run, Undo, Redo, Cut, Stop, Add Comment, and other functions.
> - Clicking Run automatically triggers Save and Compile, followed by Run.
> - Run creates and runs a job that runs the flow.

34. The Peek stage fulfilled its intended purpose and can now be removed. To do so the stage is selected on the canvas by clicking it and then, press the **Delete** button on your keyboard.

    However, in our example, we added a Join to the canvas after the first join and renamed it `Join_on_email` to annotate its intended purpose.

Then, we added our PostgreSQL data instead of using the Asset Browser. In this case, we added the dedicated IBM Cloud Databases for PostgreSQL from the connectors menu. After it is added, we double-clicked the connector and selected the required PostgreSQL connection (the connector searches for all connections of that type that are defined in the Project). Then, we specified the relevant schema and source table (see Figure 4-49 and Figure 4-50).

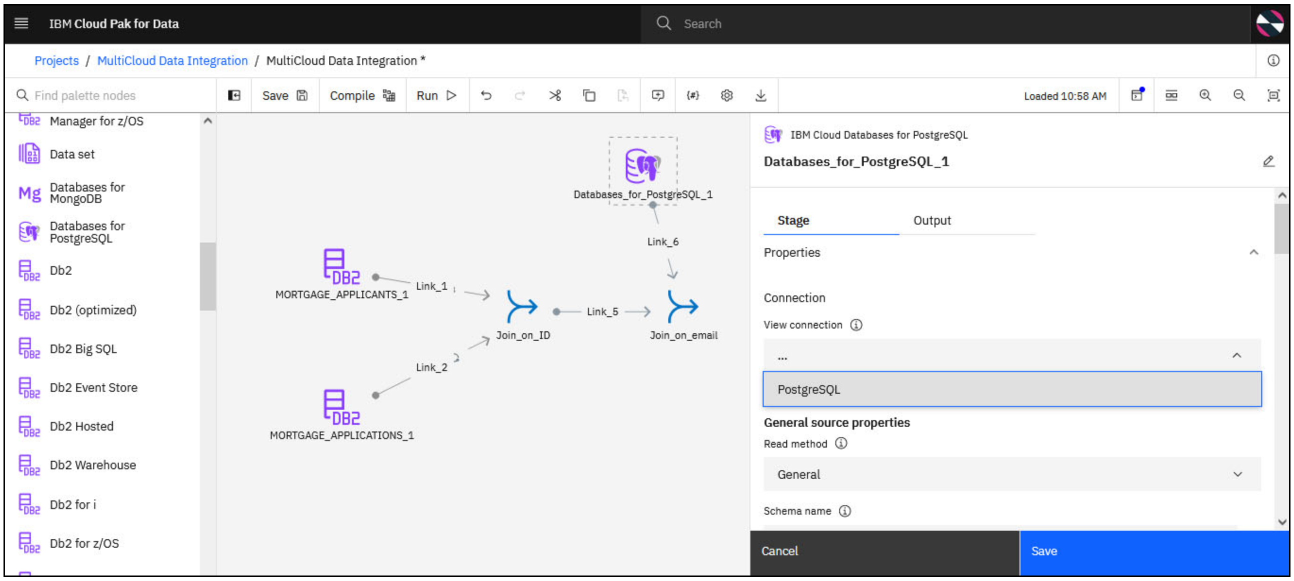

Figure 4-49 Adding PostgreSQL data to the canvas by way of the dedicated connector

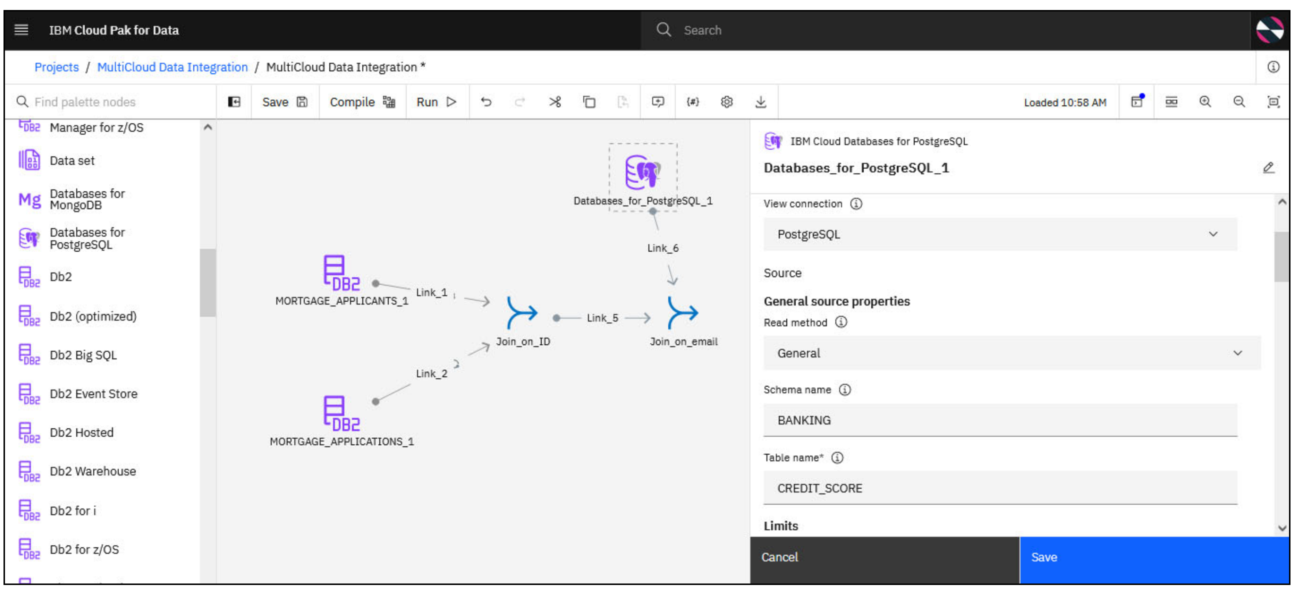

Figure 4-50 PostgreSQL connector setup

35. We can further edit the columns that the table is contributing to our flow and decide whether to enable Runtime column propagation.

IBM DataStage is flexible about metadata and can cope with the situation where metadata is not fully defined. You can define part of your schema and specify that, if your job encounters extra columns that are not defined in the metadata when it runs, it adopts these extra columns and propagates them through the rest of the job. This process is known as *runtime column propagation* (RCP) and can be set for individual links by way of the Output Page Columns tab for most stages, or in the Output page General tab for Transformer stages.

Always ensure that runtime column propagation is enabled if you want to use schema files to define column metadata.

In our case, we specifically enabled this feature for our `Join_on_ID` stage by selecting the dedicated Runtime column propagation option, as shown on Figure 4-51.

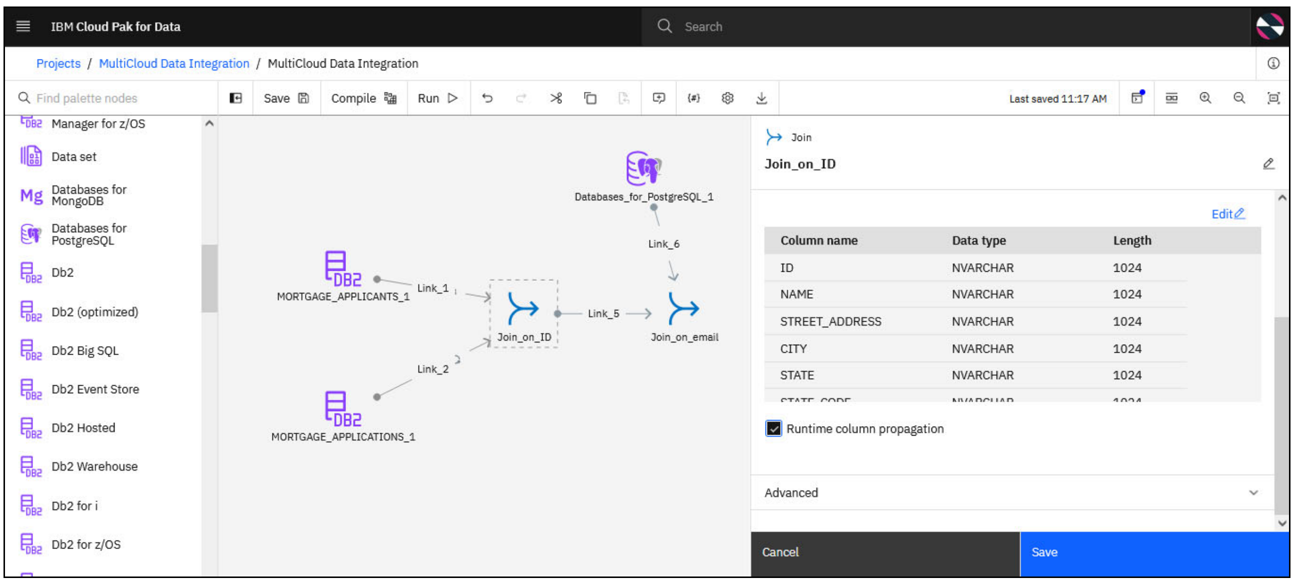

Figure 4-51 Runtime propagation setting

36. We added the MongoDB data set that contains Interest Rate data by way of the Asset Browser connector (see Figure 4-52).

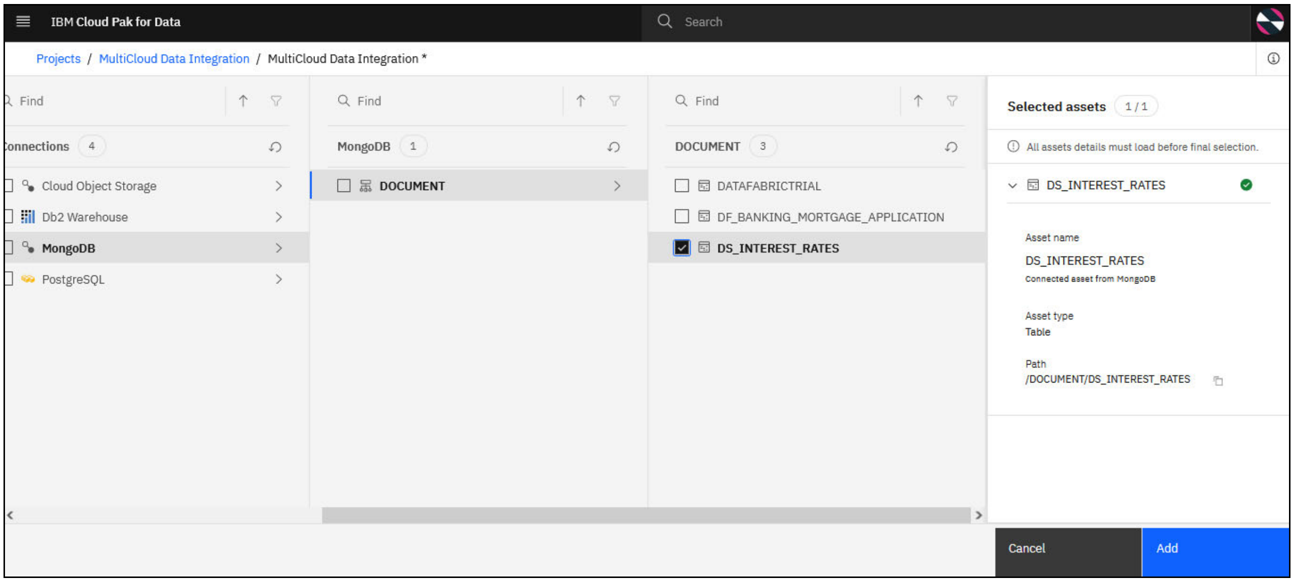

Figure 4-52 Adding MongoDB data by way of the Asset Browser Connector

37. We then added Transformer and Lookup stages, the target for our flow results (Cloud Object Storage), and connect them, as shown in Figure 4-53.

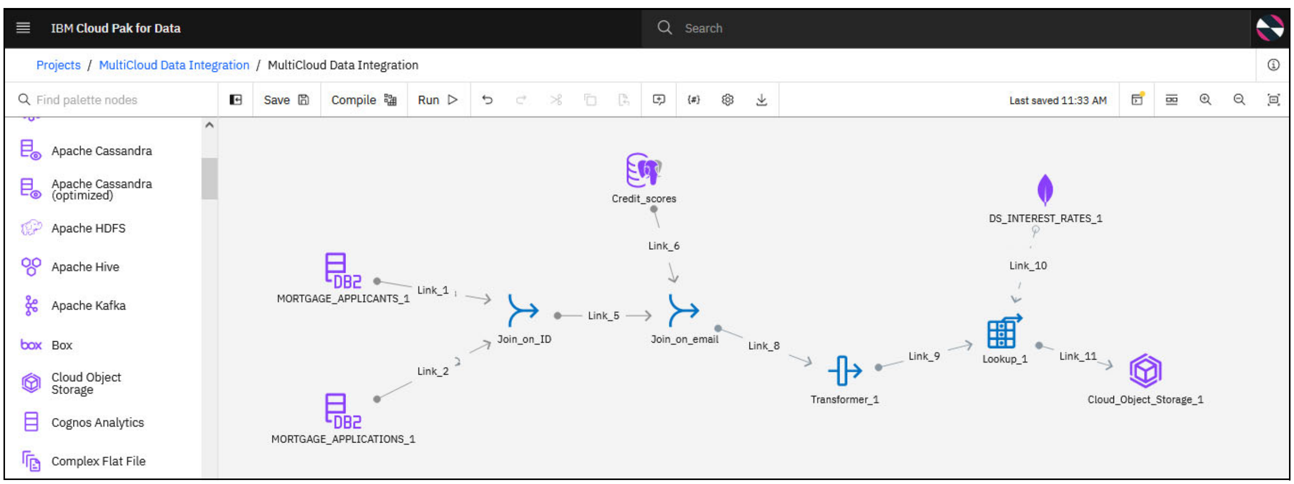

Figure 4-53 Final view of connectors and stages for the flow

Although the design of the flow is now complete, the setup of the newly added stages still must be finalized.

38. In the Cloud Object Storage connector, we specified that a new `.csv` file must be created in the bucket that we specified. Also, we added it as a new connected asset within our Project (see Figure 4-54).

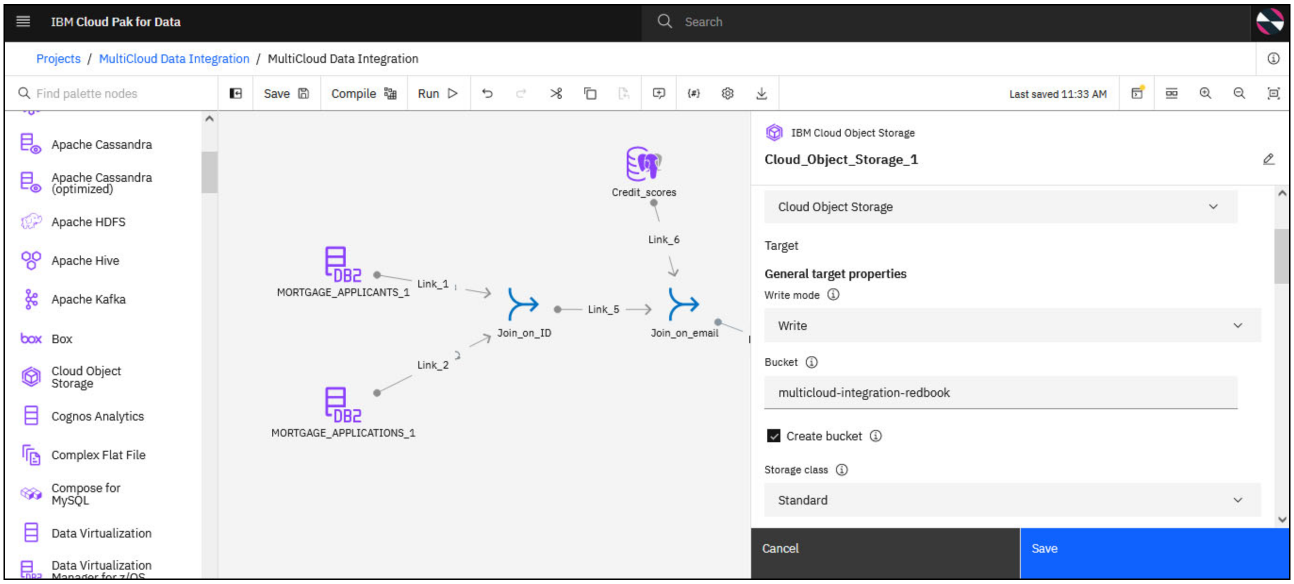

Figure 4-54 Cloud Object Storage connector setup

39. The `.csv` file is in the bucket and the newly created connected asset can have different names, which can be configured, as shown on Figure 4-55. Selecting the First line as a header option ensures that the column names are propagated from the metadata in our flow.

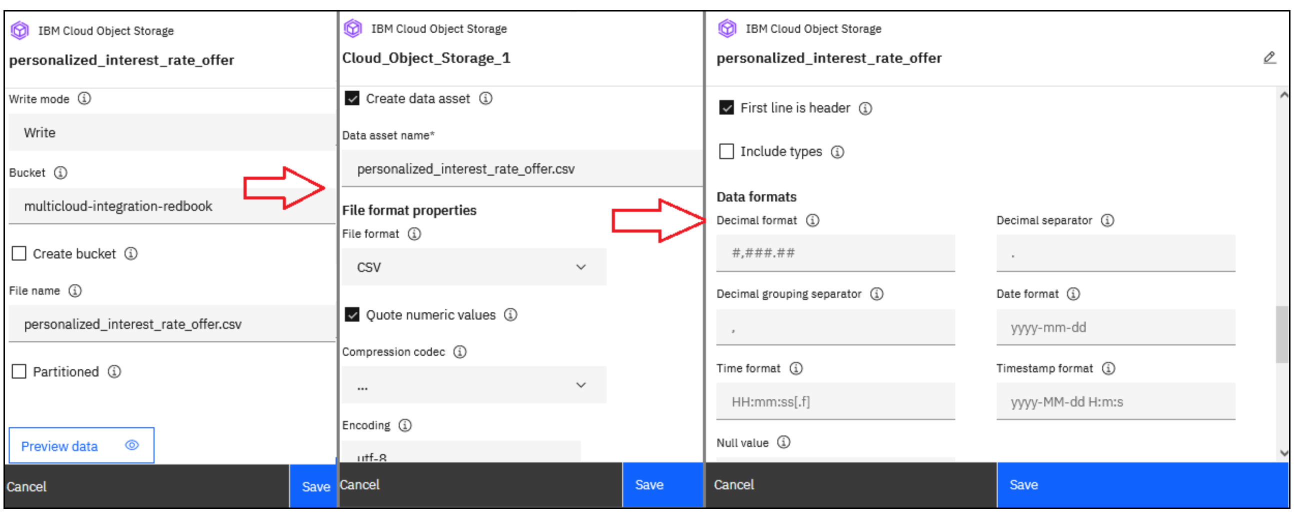

Figure 4-55 Cloud Object Storage connector setup - continued

40. We then edited the Transformer stage. This stage allows you to perform various conversions and add features and columns to your data by using calculations and formulas that range from the simple to the most complex.

    In our case, we created a feature in our data: a column that is called `TOTALDEBT`. We also specified the formula for calculating the values in that column that is based on the requested Loan Amount and preexisting Credit Card Debt for each of our mortgage applicants.

41. We double-clicked the **Transformer stage** icon that is on the canvas to enter its setup menu. Then, we switched to the Output tab (see Figure 4-56).

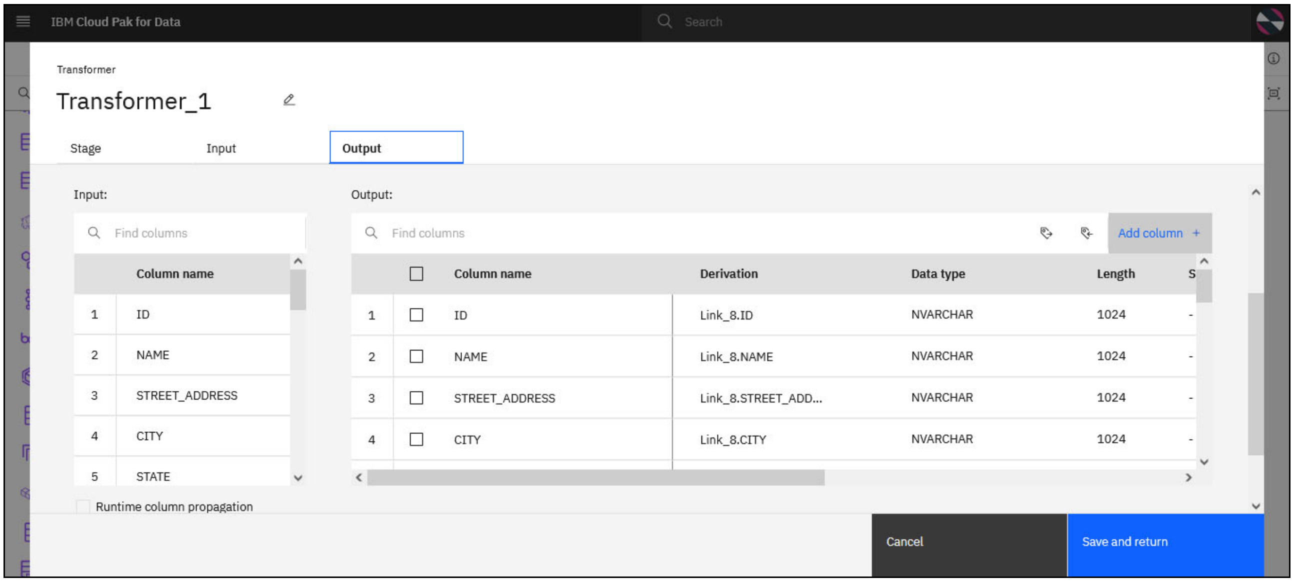

Figure 4-56 Setting up the Transformer stage - Adding columns

42. We clicked **Add Column** to add the new feature that we needed. We renamed it to `TOTALDEBT` and then, clicked the **Edit expression** icon in the Derivation column to enter the formula that we needed (see Figure 4-57 and Figure 4-58 on page 199).

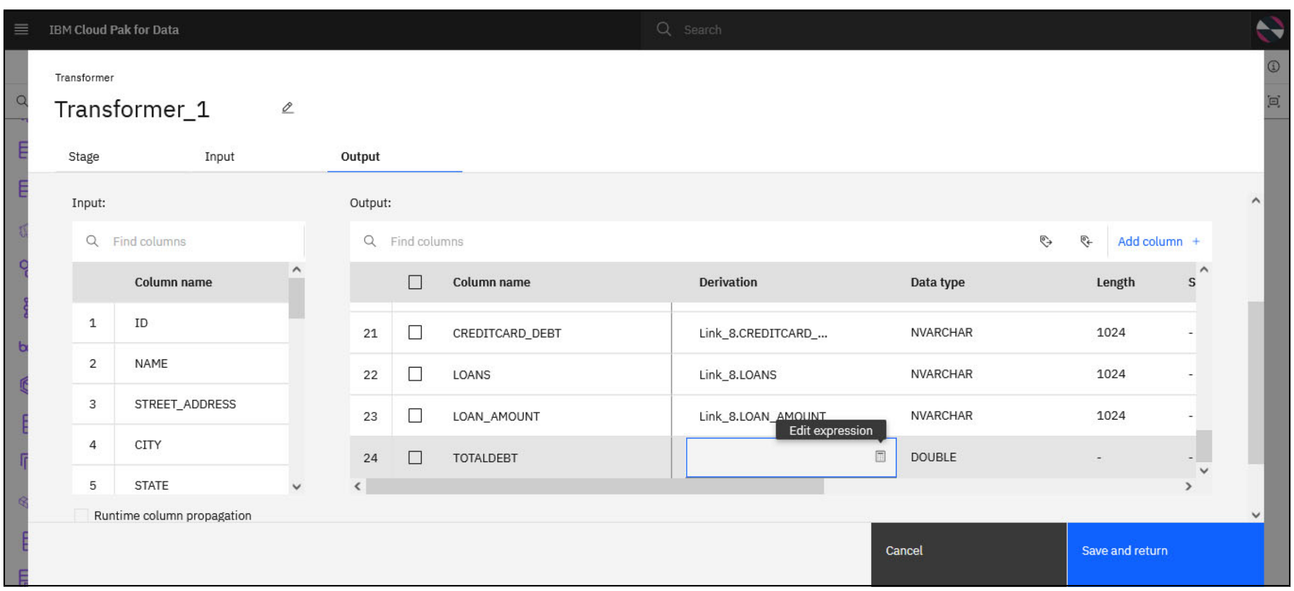

Figure 4-57 New column setup

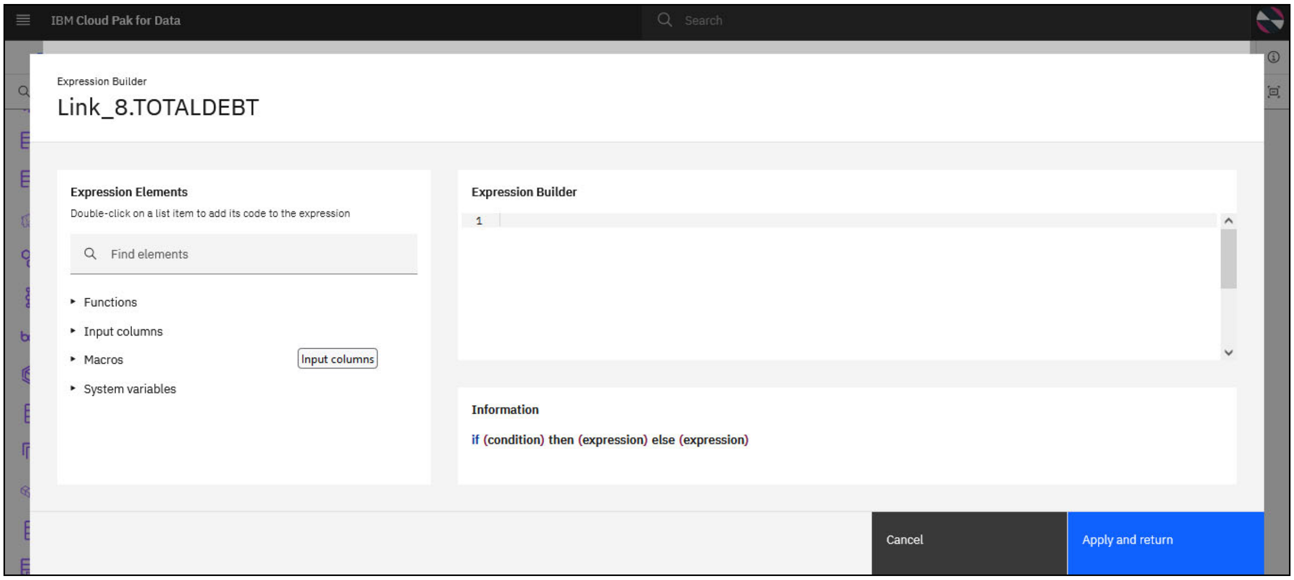

Figure 4-58 Setting up the calculation / formula for the new column

The Transformer stage is used build out your derivation that is based on system variables, macros, predefined functions, and input columns. For more information about this stage, see this IBM Documentation web page.

43. For our use example, we used a simple calculation that is based on the input columns `CREDIT_CARD_DEBT` and `LOAN_AMOUNT`. We double-clicked the names of the columns on the Input columns list on the left to add them to the expression builder. The addition can be specified by entering the "+" sign.

44. After the formula was completed, we clicked **Apply and return** to return to the main Transformer stage menu (see Figure 4-59).

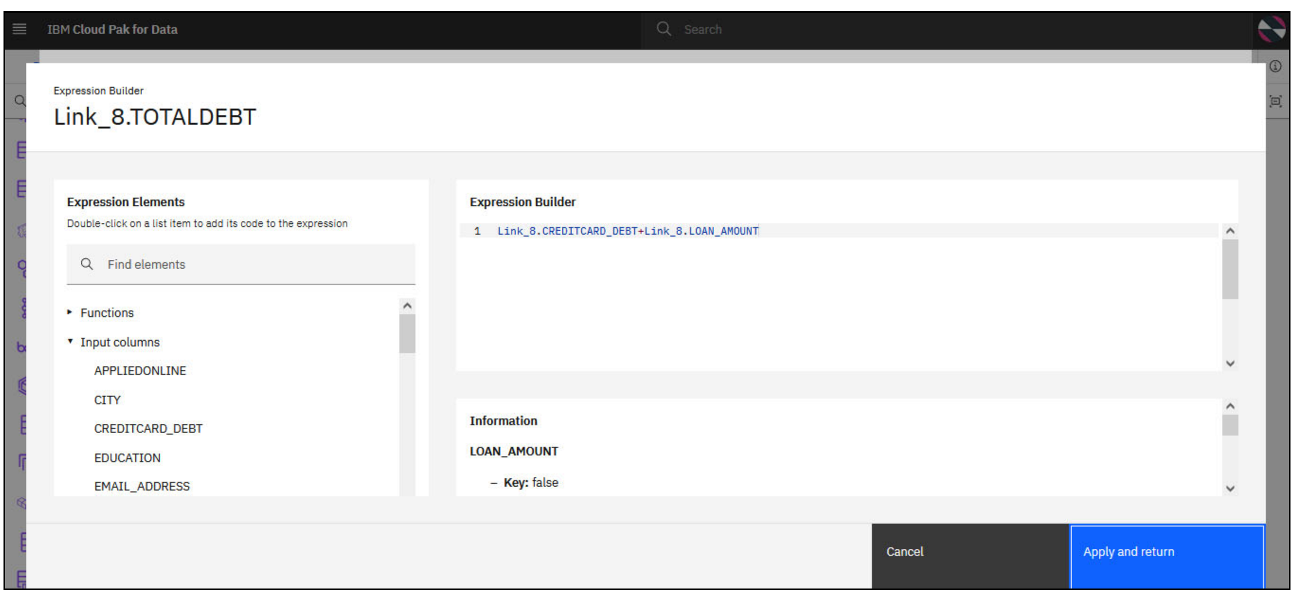

Figure 4-59 TOTALDEBT formula

45.The next change that we made in this window was the data type for our newly introduced column. We manually changed it to **Double** (see Figure 4-60).

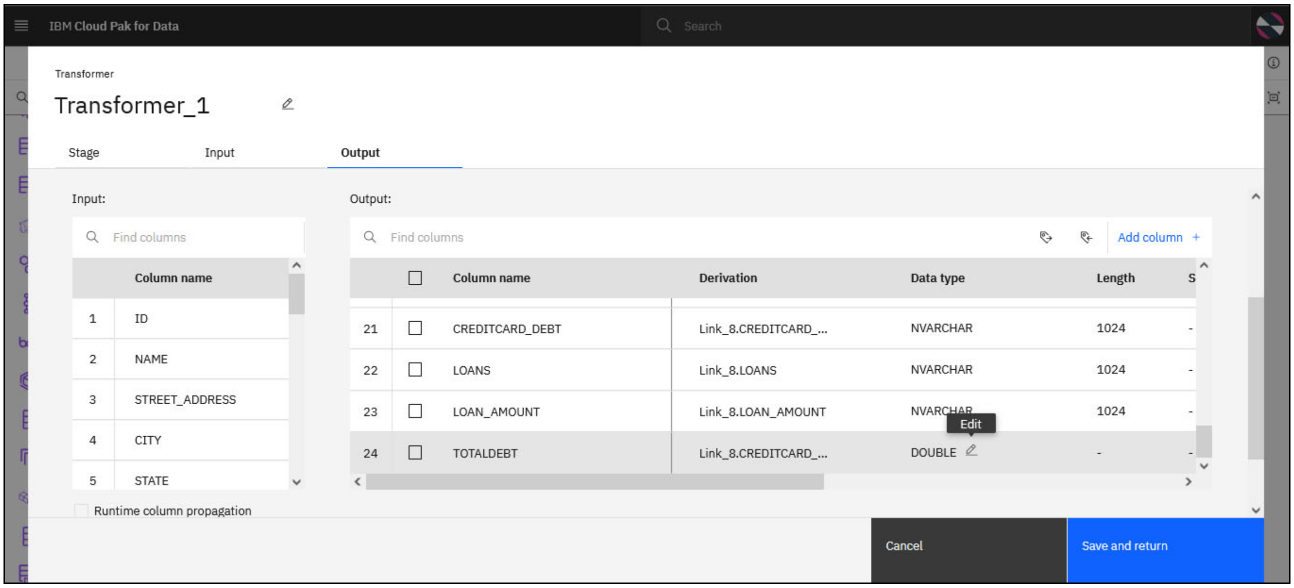

Figure 4-60 Finished column setup

46.We designated the `CREDIT_SCORE` column as our Key to enable the correct setup for the next step in our flow; that is, the Lookup stage that follows the Transformer. We clicked **Save and return** to return to the canvas (see Figure 4-61).

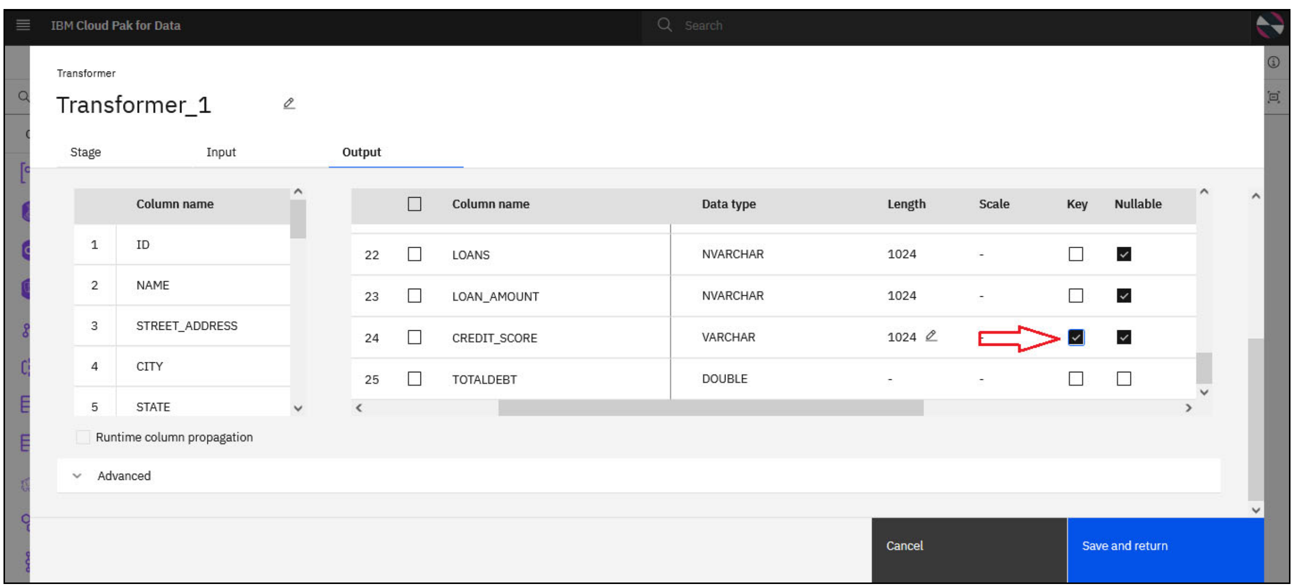

Figure 4-61 Extra column setup

47.The Lookup stage is used to perform lookup operations on a data set that is read into memory from any other Parallel job stage that can output data or provided by one of the database stages that support reference output links. It also can perform a lookup on a lookup table that is contained in a Lookup File Set stage.

In our example, we used the interest rates data that came from the MongoDB file as our lookup table. The flow looks up interest rates that are defined for different credit score ranges in that file. Then, it maps that information against individual applicant information and their credit score that are from the preceding steps of our flow.

Therefore, Link 10 in our example was our lookup (reference) link. Link 9 that is from the Transformer stage was our Primary link. Because we were looking up interest rates that are based on credit scores, the `CREDIT_SCORE` column must be selected as the `Apply range to columns` column for the Primary link (see Figure 4-62).

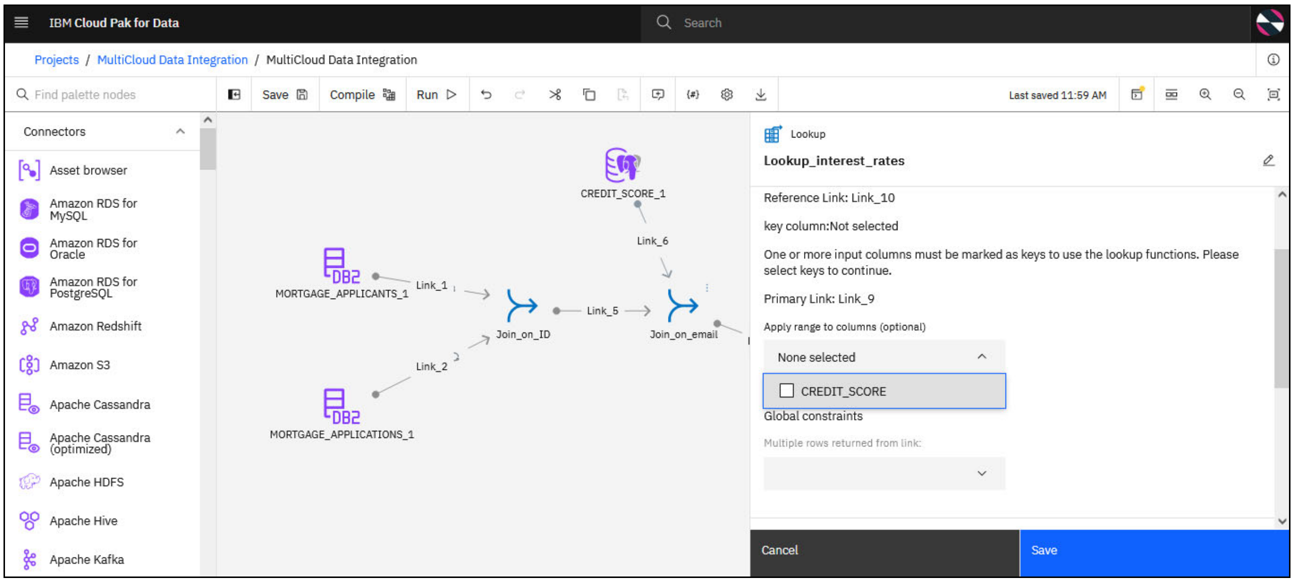

Figure 4-62 Lookup stage setup

48. We finalized range set up for the Lookup stage and the key column by selecting **CREDIT_SCORE**.

The values in this column must be less than the `ENDING_LIMIT` column value and more than the `STARTING_LIMIT` value. For example, for an applicant with the `CREDIT_SCORE` 400 - 499, our flow pulled and assigned the associated interest rate of 4.56% for our mortgage applicant (see Figure 4-63).

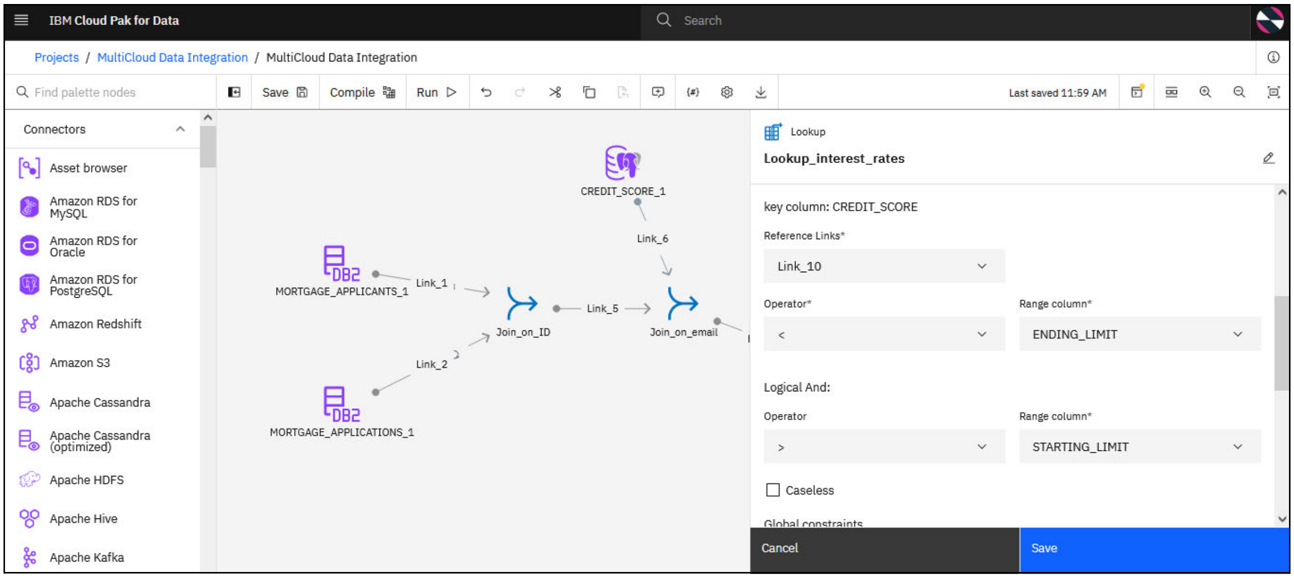

Figure 4-63 Lookup stage setup: Continued

49. We saved the stage, returned to the canvas, and tweaked the MongoDB connector settings.
50. In the MongoDB connector's Output tab, we edited the outputs that the connector is feeding into `Link_10` of the flow by clicking **Edit** (see Figure 4-64).

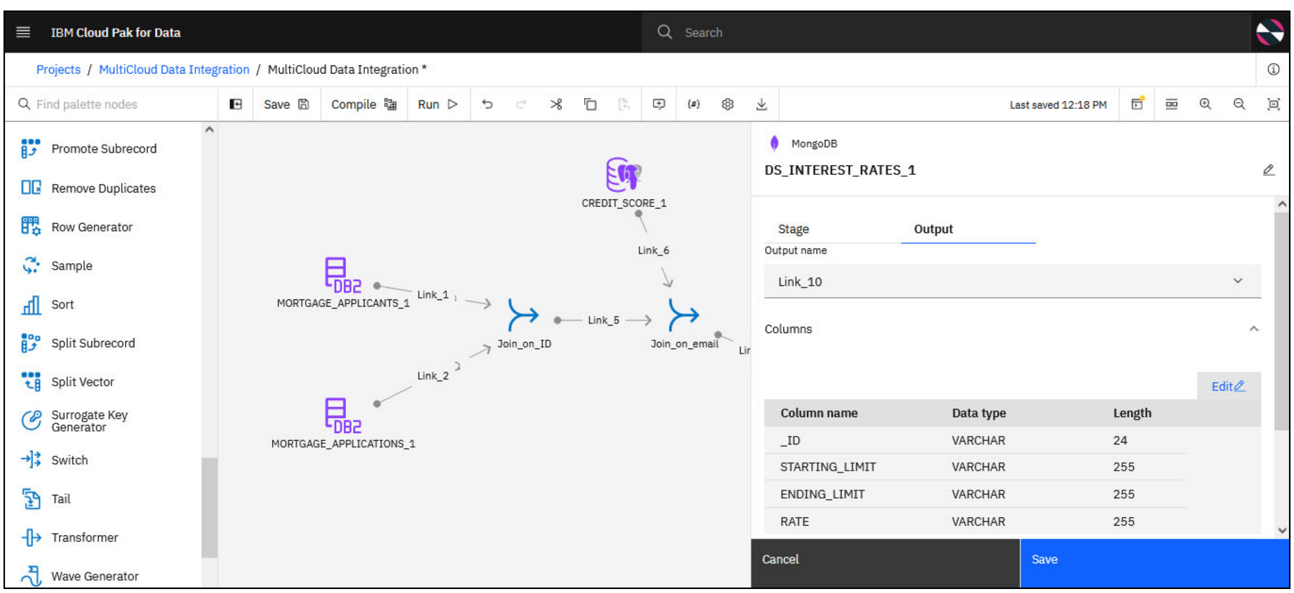

Figure 4-64 Lookup stage setup - continued

51. The ID column is not needed because it might confuse our Lookup stage; therefore, it can be deleted by selecting the column and then, clicking **Delete** in the blue ribbon menu (see Figure 4-65).

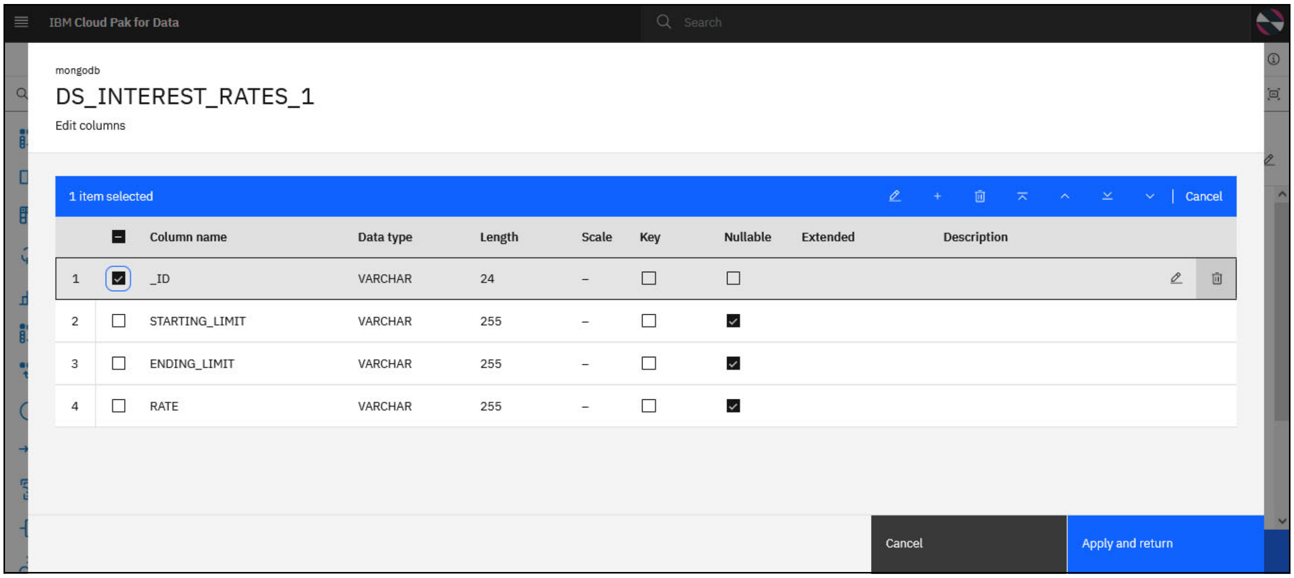

Figure 4-65 Removing input columns

52. We returned to the Lookup stage setup and edited its outputs. The STARTING_LIMIT and ENDING_LIMIT columns were not needed in our final flow output and were removed (see Figure 4-66).

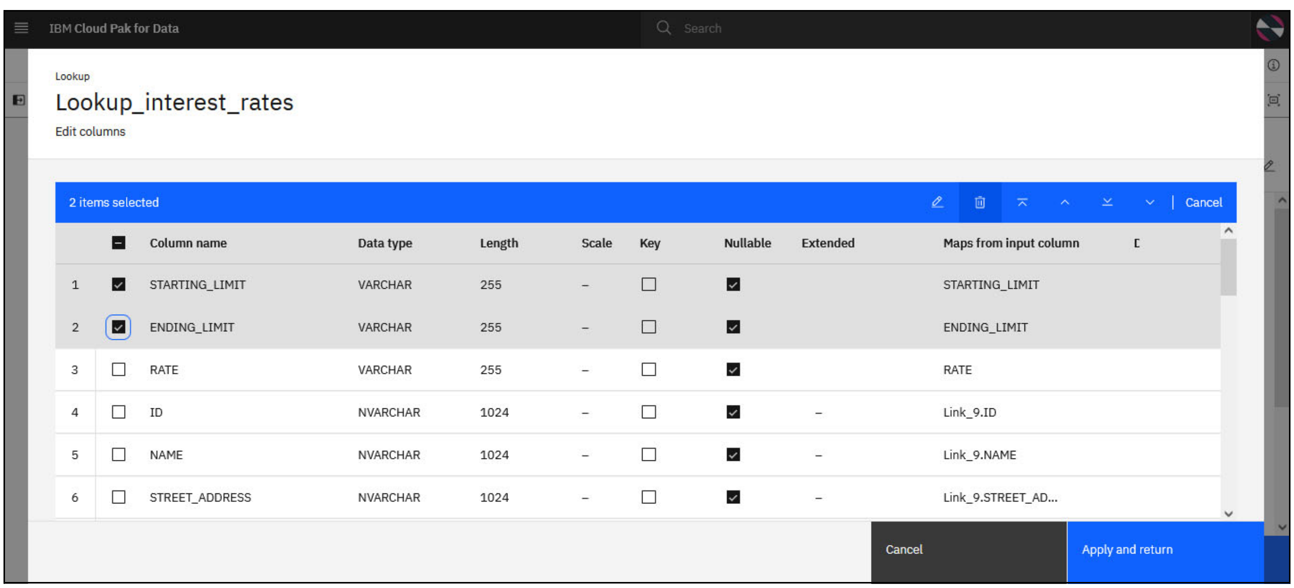

Figure 4-66 Removing output columns

53. We also reordered the columns by moving the Rate column to the end of our output data set (see Figure 4-67).

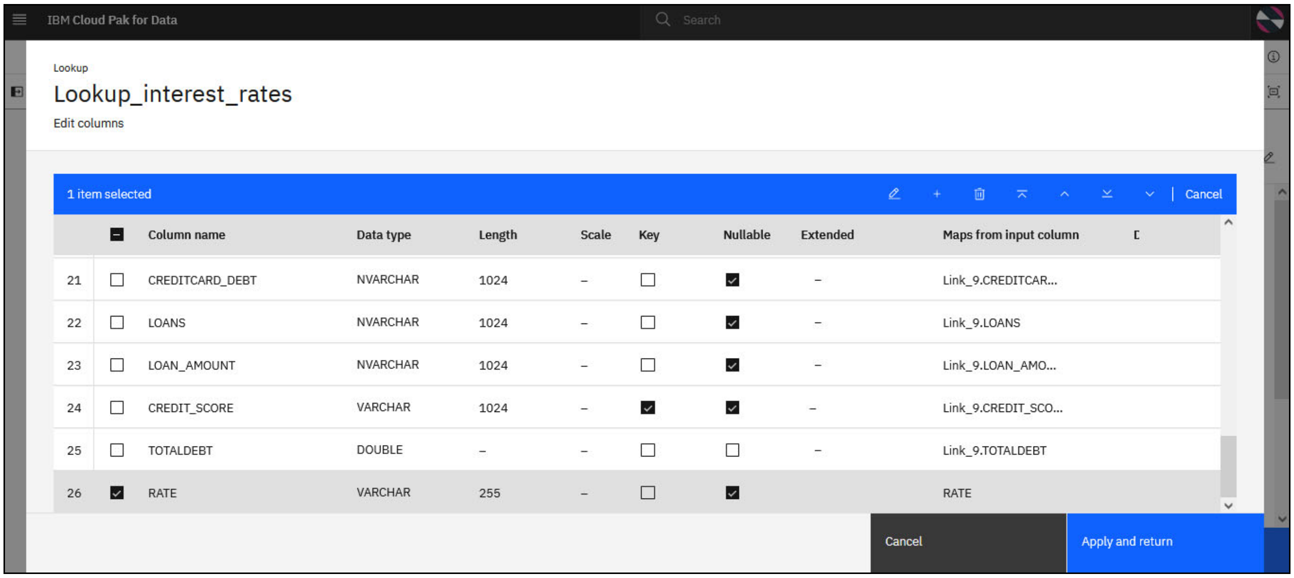

Figure 4-67 Reordering columns

54. To finalize the Lookup stage setup, we clicked **Apply and return** and enter a more descriptive name for the stage: `Lookup_interest_rates`.

55. Our flow was now ready to be run, which can be triggered by clicking **Run** at the top menu of the canvas. This action automatically triggered the **Save and Compile** flow as the prerequisite steps (see Figure 4-68).

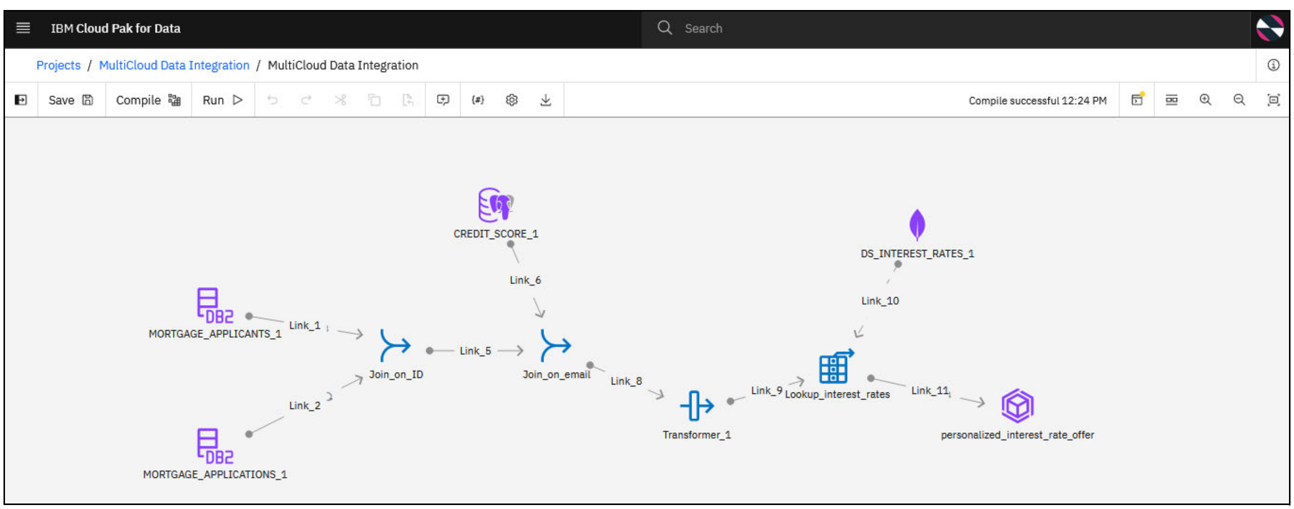

Figure 4-68 Finished flow

Our flow ran successfully, as shown on Figure 4-69.

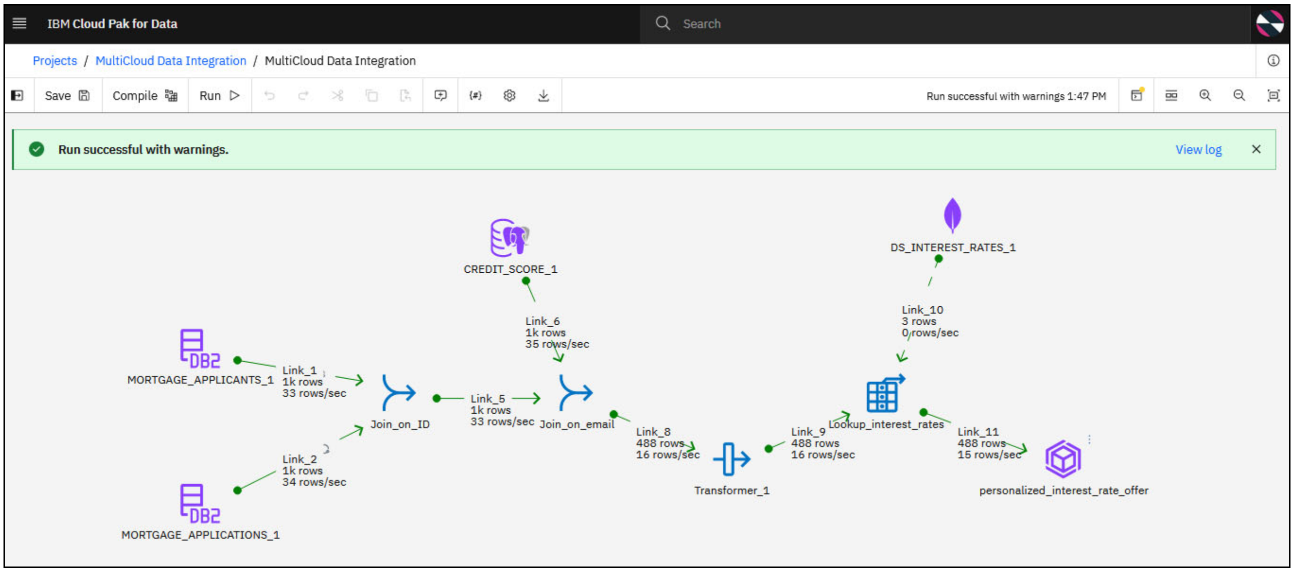

Figure 4-69 Run results

56. We then returned to the Project level by clicking the link with the name of our project (MultiCloud Data Integration) that is shown in the breadcrumbs menu in the upper left of the window. The Assets tab of our Project now includes a new asset of type Data listed (see Figure 4-70).

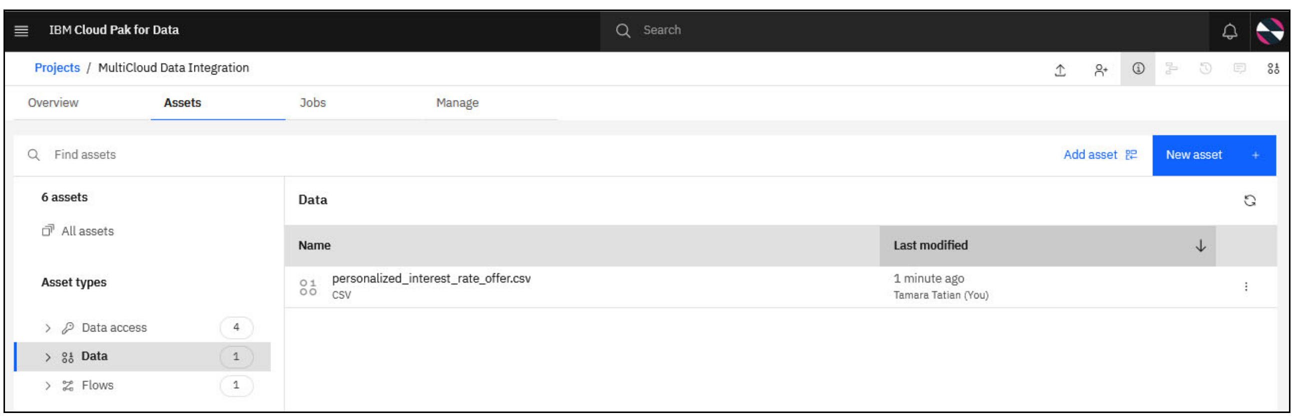

Figure 4-70 Flow output: A new asset and csv file

57. Clicking the asset, we can see its preview and verify that the resulting data set includes the joins, transformations, and data integration steps that we specified.

 We successfully used IBM DataStage to aggregate anonymized mortgage applications data from one data source with the mortgage applicants' personally identifiable information from another source, calculated total debt per applicant based on their current credit card debt and the requested loan amount, factored in risk score information that is stored in a third system to determine the mortgage rate our bank is prepared to offer the applicant. Then, we wrote the results into a `.csv` file that is stored in a Cloud Object Storage database (see Figure 4-71 on page 206).

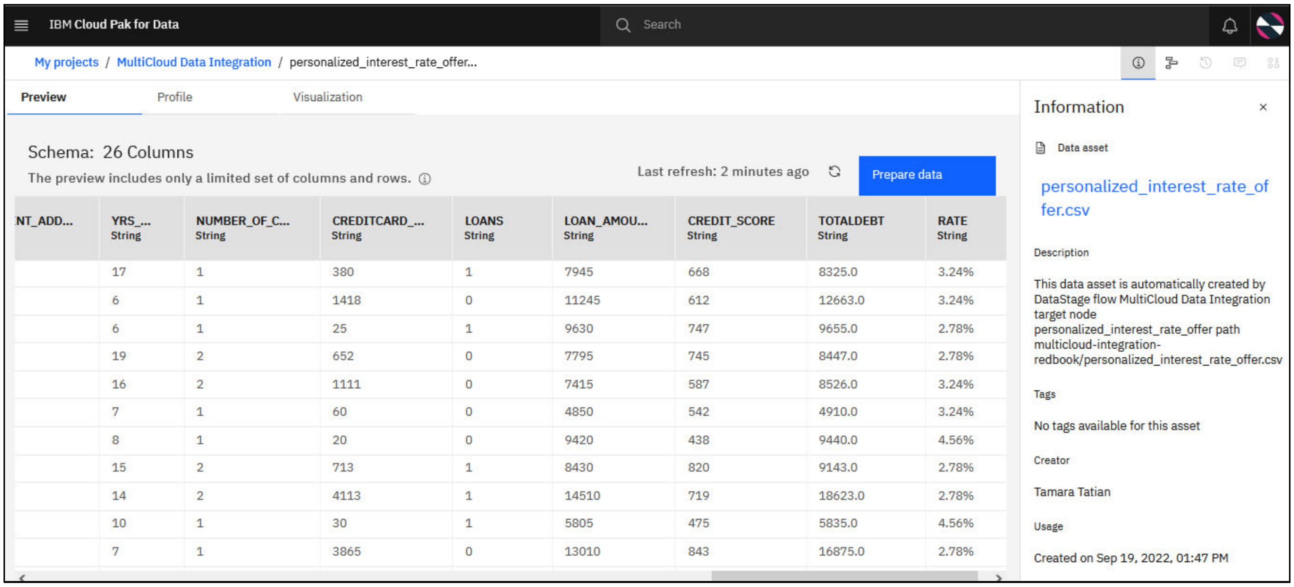

Figure 4-71 Flow output - personalized interest rates per mortgage applicant

4.2.2 Running DataStage jobs

In the previous section, we described how we delivered the dataset that contains personalized interest rate offers to our Loans department. However, a key part of their request was to receive the latest version of that dataset every week, based on the updates to the source applicant, applications, credit score, and interest rates data.

In this section, we describe how we can fulfill that requirement by reviewing the concept of DataStage jobs.

In DataStage, flow design and flow execution are logically separated. The flow acts as a blueprint for the steps that must be run; a job is an individual runtime instance that runs those steps.

After it is built, the flows are compiled, and a job instance that runs the steps of the flow is then created and runs on the in-built parallel processing and execution engine. This engine enables almost unlimited scalability, performant workload execution, built-in automatic workload balancing, and elastic scaling.

This design paradigm accelerates development, increases productivity, promotes reusability, and enables a consistent, standardized way of creating integration workloads to satisfy data integration tasks of any complexity. Different job runtime sizes and specifications can be chosen to suit your job complexity and data processing volumes.

DataStage jobs are highly scalable because of the implementation of parallel processing. DataStage can run jobs on multiple CPUs (nodes) in parallel. It also is fully scalable, which means that a correctly designed job can run across resources within a single machine or take advantage of parallel platforms, such as a cluster, GRID, or massively parallel processing (MPP) architecture.

With Red Hat OpenShift, which is the container orchestration platform that IBM Cloud Pak for Data runs on, resource-intensive data transformation jobs can be distributed across multiple compute nodes on the container cluster.

The jobs can be run on-demand and scheduled.

When we clicked **Run** in the flow designer canvas, we triggered on-demand job execution by using the default runtime settings that are specified for the flow and our DataStage setup on the IBM Cloud Pak for Data cluster.

The Jobs and their runs can be accessed by way of the Jobs tab of our Project. Our default job is the only job that is listed there (see Figure 4-72).

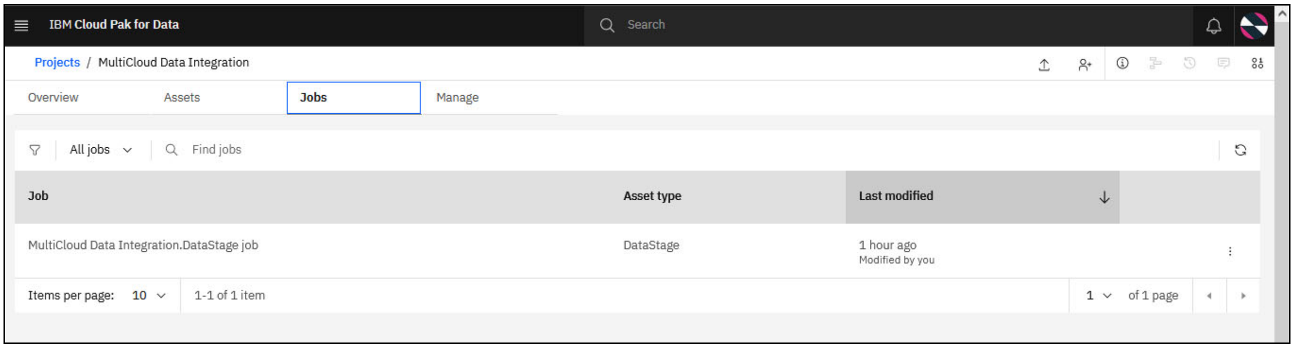

Figure 4-72 Jobs list

Clicking into that job, we can see more information about all its runs (successful and unsuccessful), see Figure 4-73.

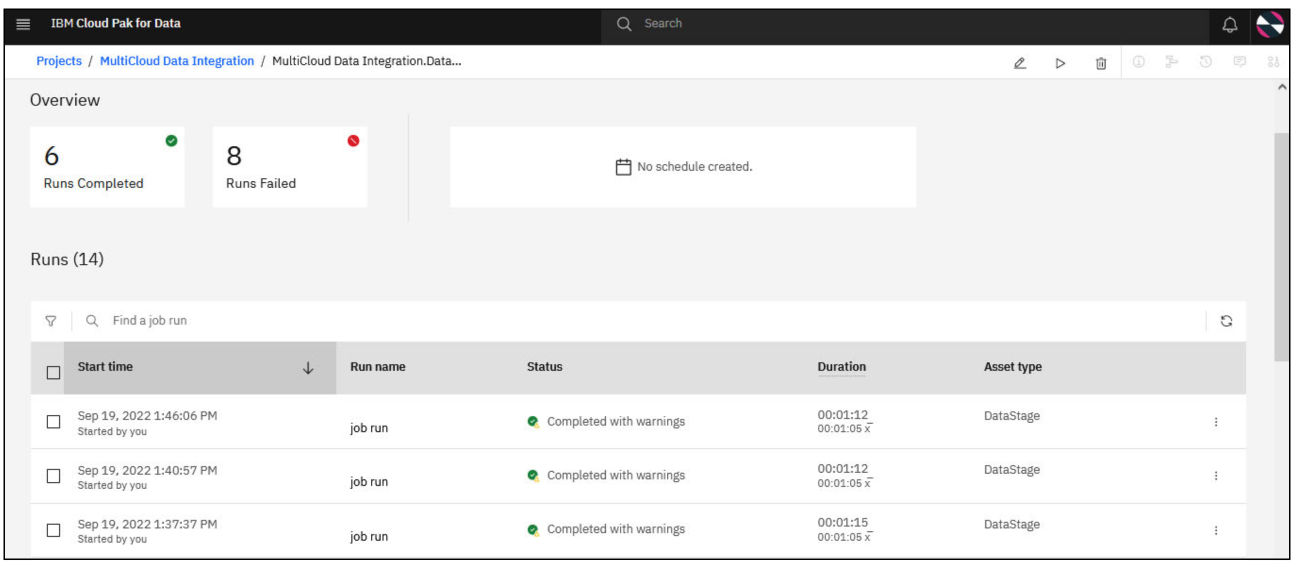

Figure 4-73 Job run statistics

Each job run generates a log that contains execution status, warnings, and errors. Clicking individual run entries takes you to the corresponding run details (see Figure 4-74).

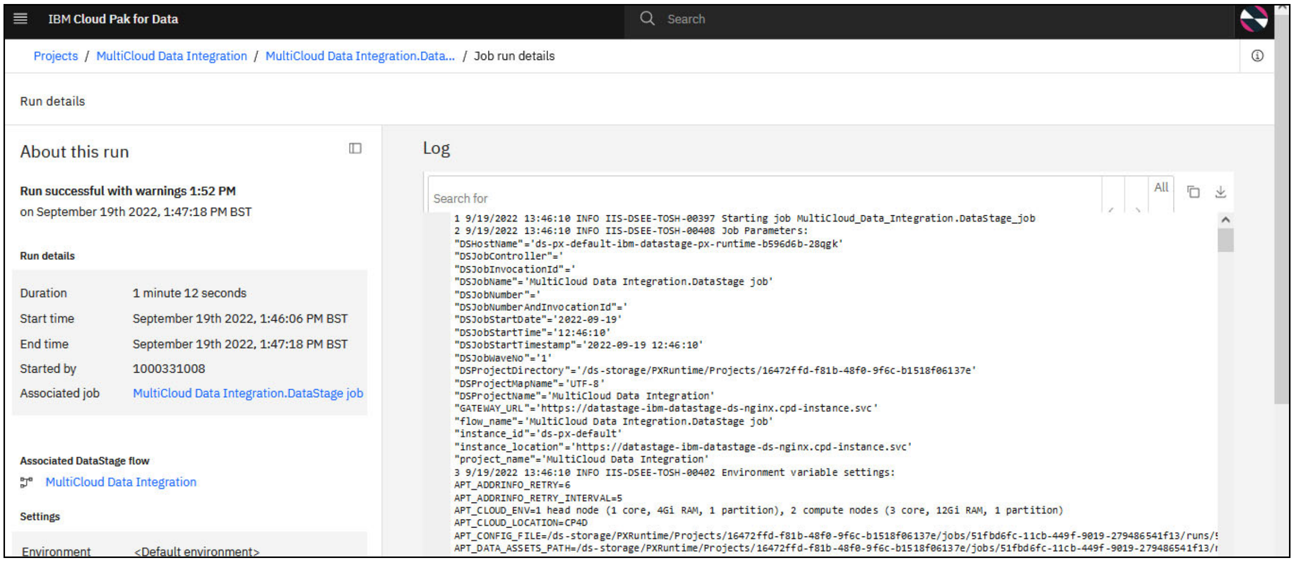

Figure 4-74 Job run logs

Because our Loans department requested that the latest results of our integration work are delivered to them on weekly, we created a schedule for our designed transformation flow.

Jobs that are running a flow can be created by browsing to the **Assets** tab, clicking the menu that is next to the relevant flow, and selecting **Create job** (see Figure 4-75).

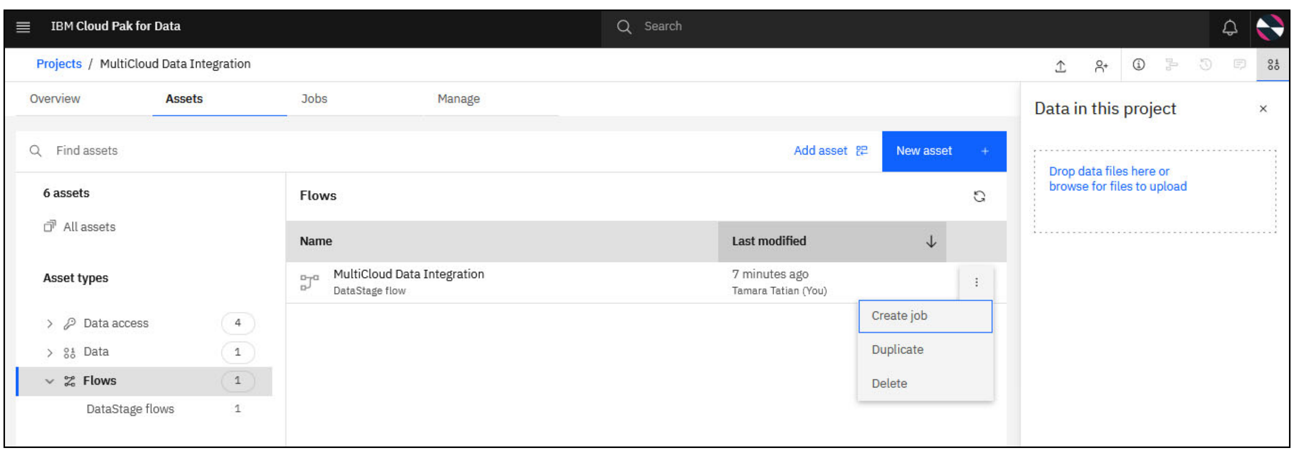

Figure 4-75 Creating new jobs

Each job requires name and optionally, a description (see Figure 4-76).

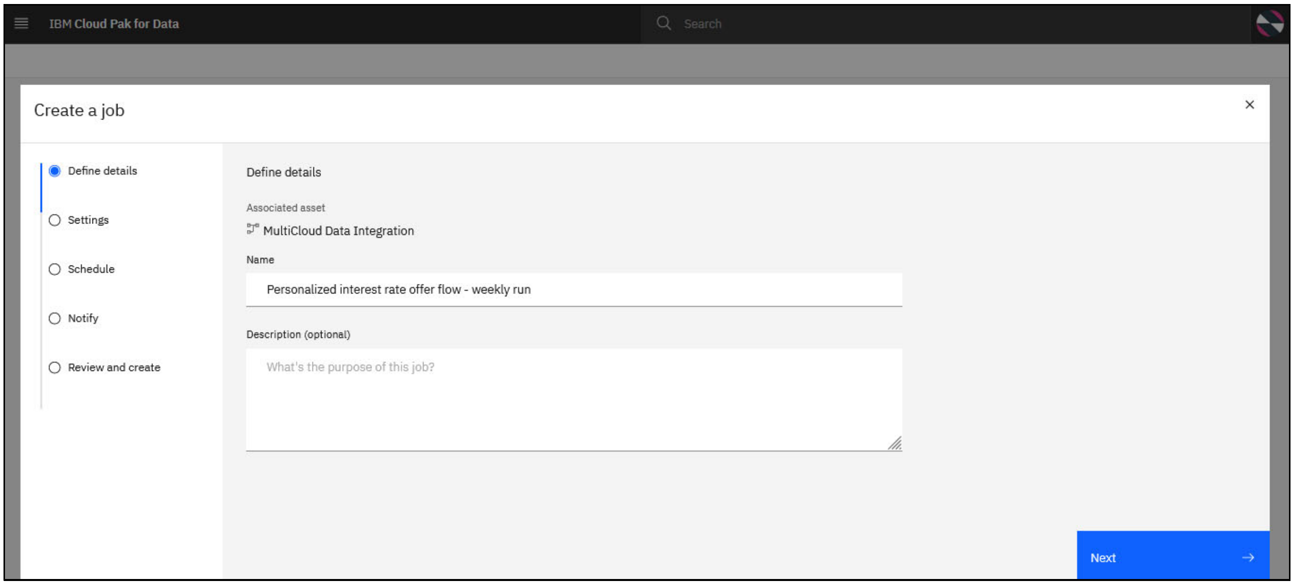

Figure 4-76 New job setup

The next step in setting up the job is to select the runtime environment specification that suits your needs best. Different run times can be chosen for different jobs. They can have different vCPU and memory specifications, parallelism settings, and more (see Figure 4-77 and Figure 4-78 on page 210).

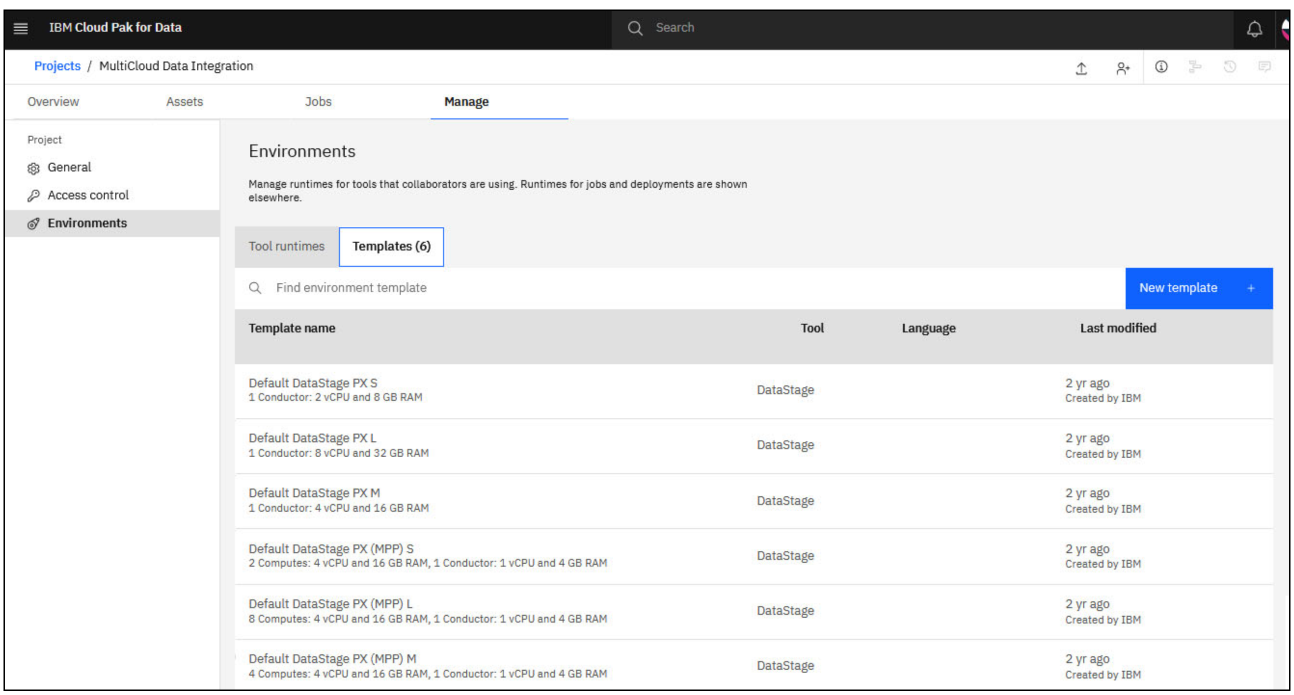

Figure 4-77 Job runtime environments: Example

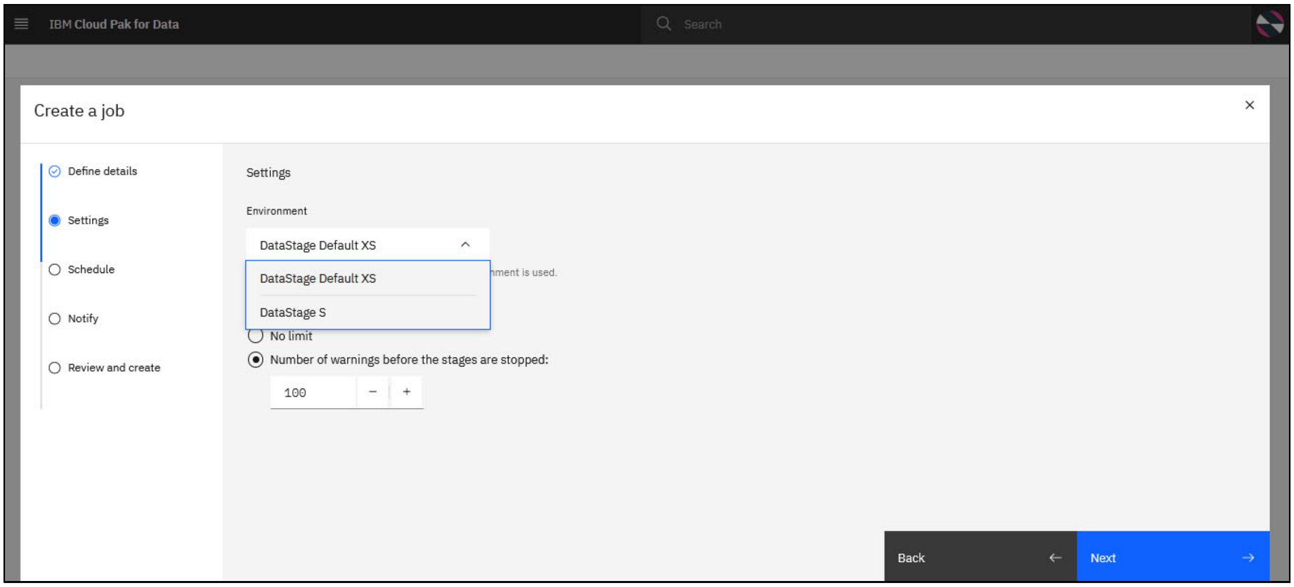

Figure 4-78 Choosing runtime environment specification and size

By using DataStage, you can define a schedule to suit your specific needs. In our example, we create a weekly execution schedule and choose to run the flow at night to minimize the effect on our source systems (see Figure 4-79).

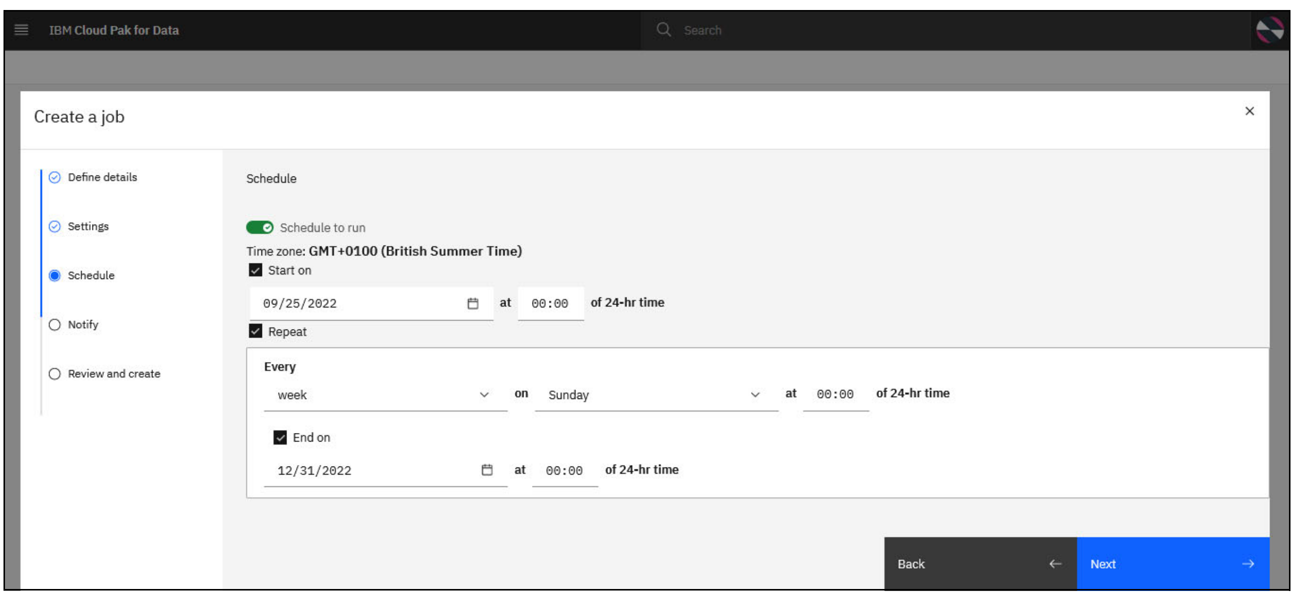

Figure 4-79 New job setup - scheduling

Each job can have different notification settings (see Figure 4-80 and Figure 4-81).

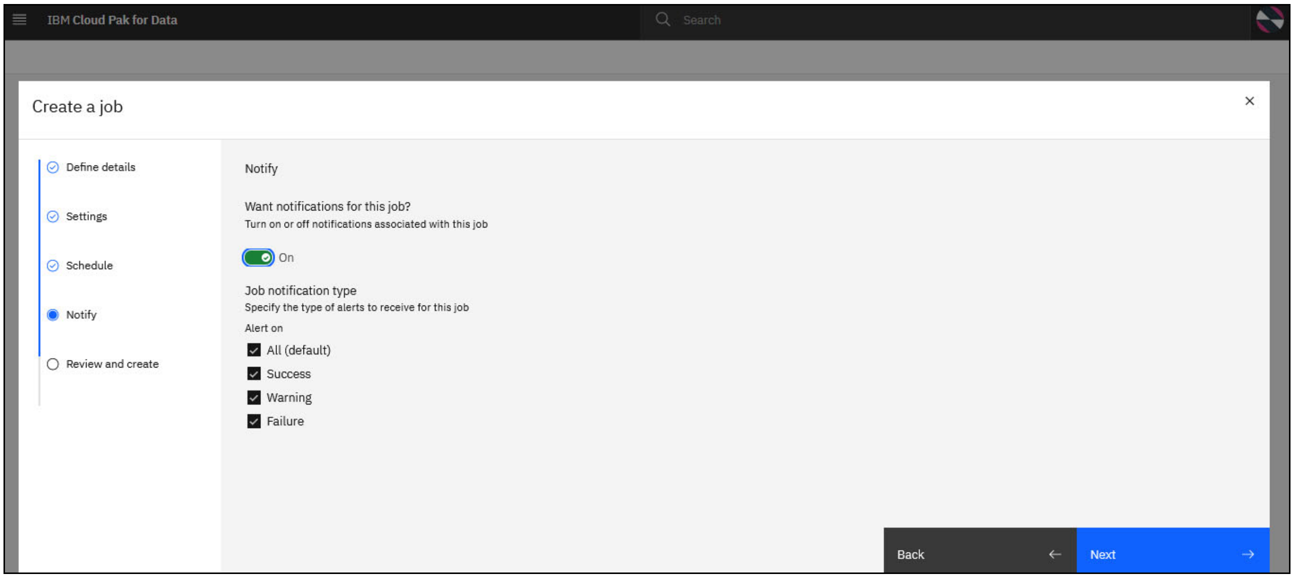

Figure 4-80 Notification setup

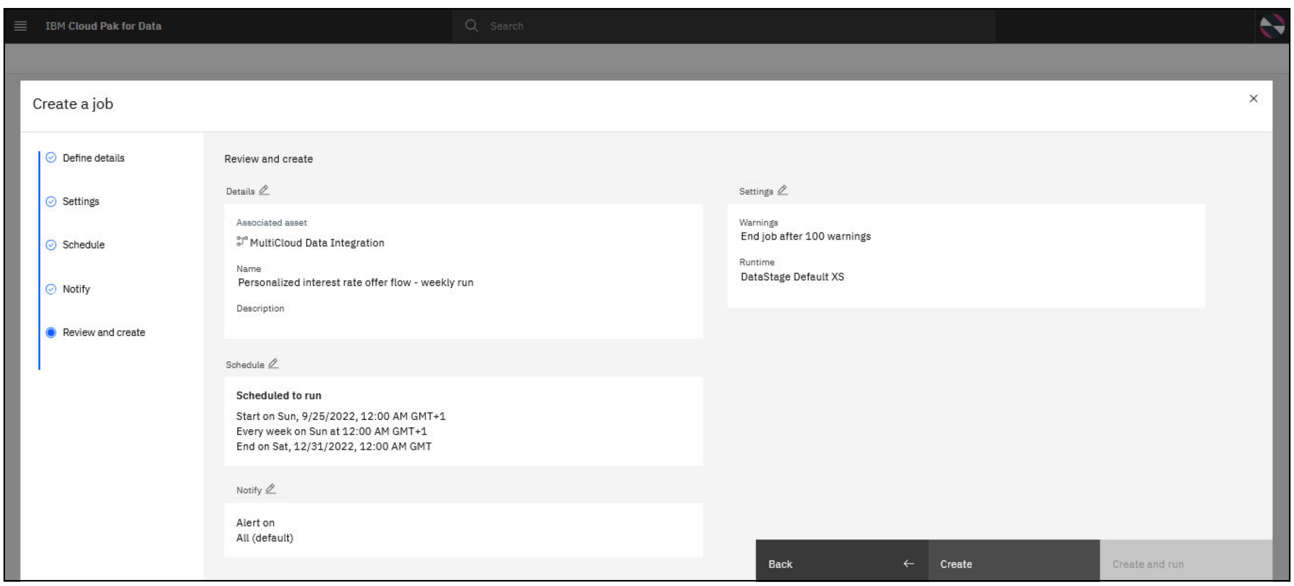

Figure 4-81 Set up review

After the setup is complete, clicking **Create** creates the job. Then, the in-built scheduling engine triggers and run it at the specified times and with the specified frequency.

4.2.3 Job sequencing

Flows are built by using the DataStage Flow Designer, and the jobs that are running them run in parallel.

Parallel jobs bring the power of parallel processing to your data extraction and transformation activities. They consist of individual stages, with each stage describing a specific process, such as accessing a database or transforming data in some way.

Although they provide a rich set of features, in some cases extra capabilities for triggering sequential execution with or without more logic and conditions might be required.

To build and run sequence jobs with which you can link multiple parallel job executions and incorporate branching, which is looping and other programming controls, the IBM Watson Studio Pipelines service is required. It must be installed on your IBM Cloud Pak for Data cluster.

The Watson Studio Pipelines editor, as shown in Figure 4-82, provides a graphical interface for orchestrating, assembling, and configuring pipelines. These pipelines can include the following components:

- DataStage flows
- Various bash scripts
- Data Refinery, notebook, AutoAI, and other flow runs
- Various execution logic and conditions (for example, Wait)

> **Important:** As of this writing, service is provided as a Beta solely for testing and providing feedback to IBM before general availability. It is not intended for production use. You can download this service from this web page (log in required).

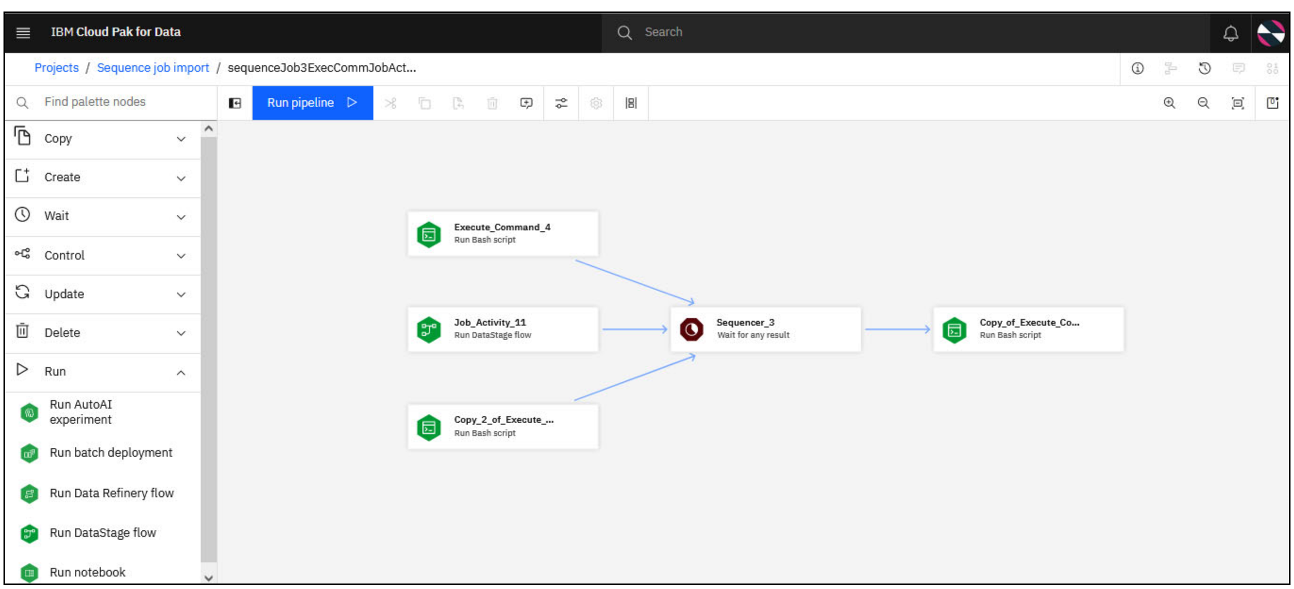

Figure 4-82 Sequence job example

For more information about the IBM Watson Pipelines service, see the following IBM Documentation web pages:

- Orchestrating flows Watson Pipelines
- Watson Studio Pipelines

4.2.4 DataStage components and parameter sets

In addition to DataStage flows and jobs, the DataStage service allows you to create DataStage components that can be reused across different flows and Parameter Sets that capture several job parameters with specified values for reuse in jobs.

DataStage components that you can create include subflows, schema library components, data definitions, standardization rules, and custom stages.

DataStage components enhance the reusability of your flow components (see Figure 4-83).

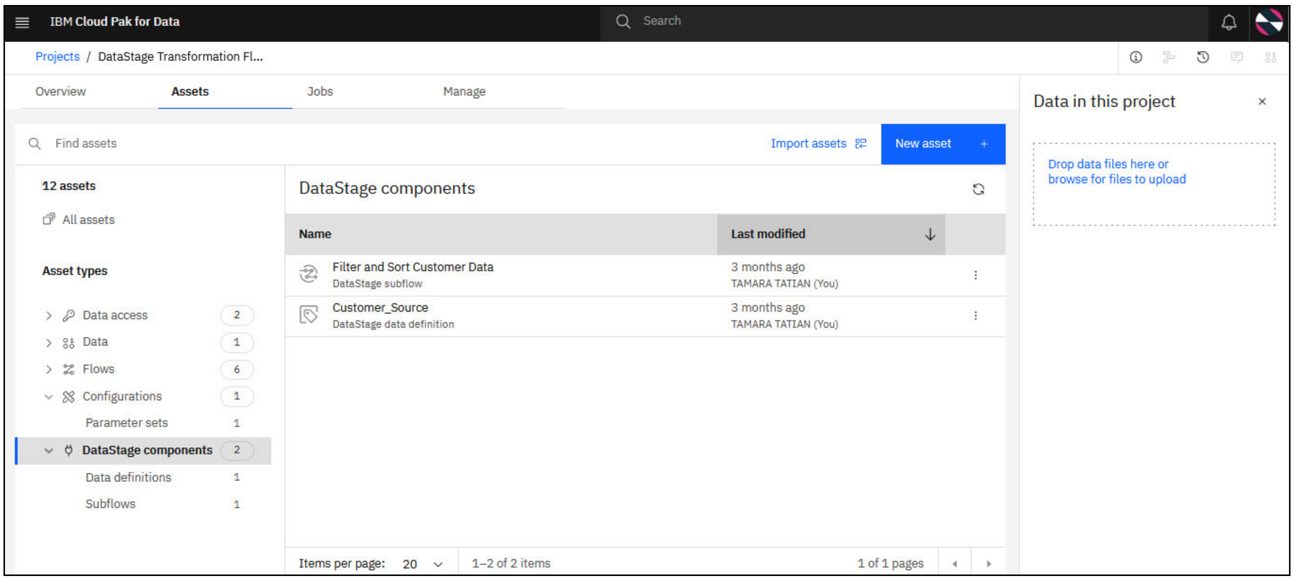

Figure 4-83 DataStage components example

For example, if a specific sequence of processing steps is frequently appearing in your flows, that sequence can be defined as a reusable subflow (see Figure 4-84).

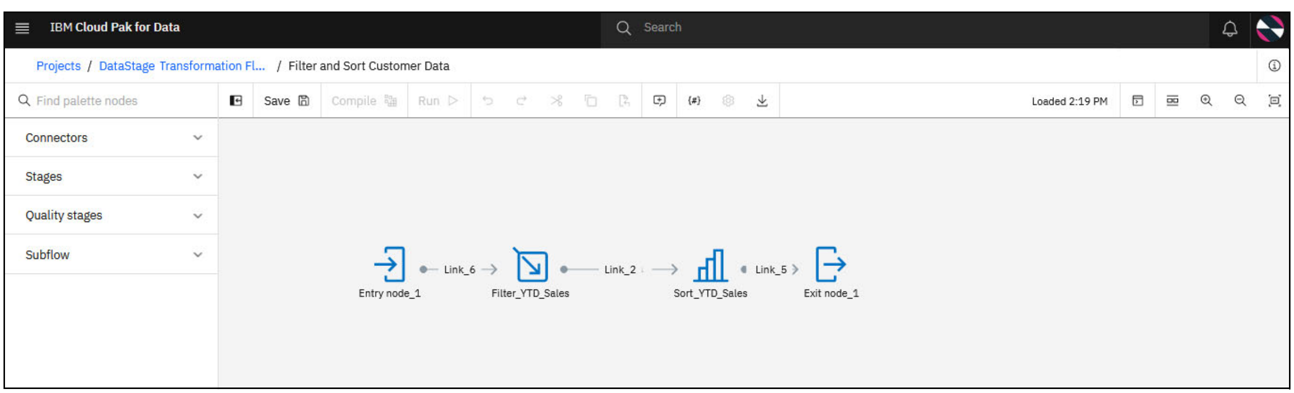

Figure 4-84 DataStage subflow example

The subflow asset can then be added to and reused by different DataStage flows, as shown in Figure 4-85.

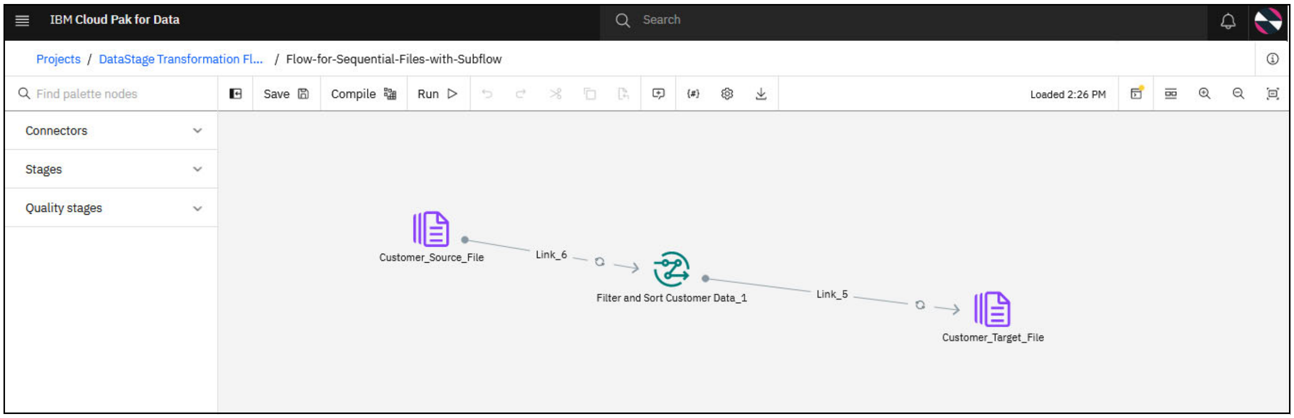

Figure 4-85 DataStage subflow component reuse example

Similarly, Parameter sets help you design flexible and reusable jobs. Without the use of parameters, a job might need to be redesigned and recompiled to accommodate those settings when a job must be run again for a different week or product for which it was designed. However, parameters sets use these settings to form part of your job design and eliminate the need to edit the design and recompile the job.

Instead of entering variable factors as part of the job design, you can create parameters that represent processing variables.

4.2.5 Summary

IBM DataStage is an industry-leading data integration tool that helps you design, develop, and run jobs that integrate, move, and transform data. At its core, the DataStage tool supports extract, transform, and load (ETL) and extract, load, and transform (ELT) patterns. Advanced data transformation can be applied in-flight (ETL) or post-load (ELT).

The service supports bulk and batch integration scenarios, with the in-built Apache Kafka, Google Pub/Sub, and IBM MQ connectors helping enable streaming integration scenarios.

The provided set of pre-built data transformation stages and connectors enable efficient transformation flow design and build of flows to support various integration use cases, from simpler use cases, including the one that is described in this chapter, to the most complex.

Data sources can be in many different repositories, including cloud-hosted sources, Hadoop sources and services, relational and NoSQL databases, enterprise and web applications, established systems or systems of record. They can be easily connected to use the range of standard connectors that are included with the solution.

The No-Code Design paradigm of IBM DataStage enables the fast and consistent creation of integration workload, even for nontechnical users. Reusability and flexibility of flow design and job execution easily can be achieved by using and reusing standard, pre-built graphical units (canvas objects), and creating reusable flow components and job parameter sets.

The parallel execution engine and workload balancing capabilities of DataStage help achieve fast and efficient job execution, and theoretically unlimited scalability.

For more information about IBM DataStage, see the following resources:

- IBM Documentation:
  - DataStage on Cloud Pak for Data
  - Transforming data (DataStage)
- Data fabric tutorials

4.3 Data integration and virtualization with IBM Data Virtualization

Although ETL is often the more popular and prevalent data integration approach, data virtualization is gaining in popularity as an integration style in its own right, and a complementary approach to ETL. It also is one of the key elements and capabilities of a Data Fabric architecture and approach.

IBM Cloud Pak for Data includes IBM Data Virtualization (also known as IBM Watson Query), which is the service that provides advanced data virtualization capabilities.

IBM Data Virtualization helps create virtual data lakes out of multiple, disparate, and siloed on-premises and on-Cloud sources. It enables real-time analytics and read-only access without moving data, copying or duplication, ETLs, or other storage requirements.

It also enables viewing, accessing, manipulating, and analyzing data without the need to know or understand its physical format or location, and query a multitude of sources as one.

Data Virtualization as a data integration approach lends itself well to use cases and scenarios that are driven by patterns that are characterized by low data latency and high flexibility with transient schemas, including the following examples:

- Creating on-demand virtual data marts instead of standing up new physical Enterprise Data Warehouses (EDWs) to help save time and cost.
- EDW prototyping and migration (mergers and acquisitions).
- EDW augmentation (workload offloading).
- Virtualization with big data (Hadoop, NoSQL, and Data Science).
- Data discovery for “what if” scenarios across hybrid platforms.
- Unification of hybrid data sources.
- Combining Master Data Management (MDM) with IoT for Systems of Insight (IT/OT).
- Master data hub extension to enrich 360 View (for example, multi-channel CRM).

Data Virtualization can help reduce ETL effort and the number of ETL pipelines that are built within your enterprise. It also augments and complements the more traditional ETL-based data engineering approach.

In this section, we use IBM Data Virtualization to implement portions of the integration use case that is discussed in 4.2, “Data integration and transformation with IBM DataStage” on page 166.

Our bank’s data engineers use IBM Data Virtualization to join mortgage applicants and applications data from an IBM Db2 Warehouse. They also explore other capabilities of the service for segmenting, combining, and querying data, and its administration and governance integration capabilities.

4.3.1 Service overview. Working with data sources. Creating a constellation

Following the installation of the IBM Data Virtualization service on your IBM Cloud Pak for Data cluster, one or more instances of the service must be provisioned by a user with Create Service Instances permission that is included in the permissions scope of their assigned platform (functional) roles; for example, the administrator or the data engineer roles that are included with the solution as standard. The user who provisioned the Data Virtualization instance automatically is assigned the service-specific Data Virtualization Admin role for that instance.

Following instance provisioning, they can manage users, connect to multiple data sources, create and govern virtual assets and then, use the virtualized data.

Each of the IBM Data Virtualization instances that are provisioned on the cluster can represent a separate virtual data lake or a virtual EDW. They also can be purpose-built for one or more specific use cases or lines-of-business. They can have their own set up in terms of the sources that are plugged into the virtualization layer and the objects from those sources that are surfaced within in, access controls, and users allowed access to the instance and the virtualized objects it contains, and more settings (including governance enforcement).

Figure 4-86 shows the three key steps that are involved in working with the service.

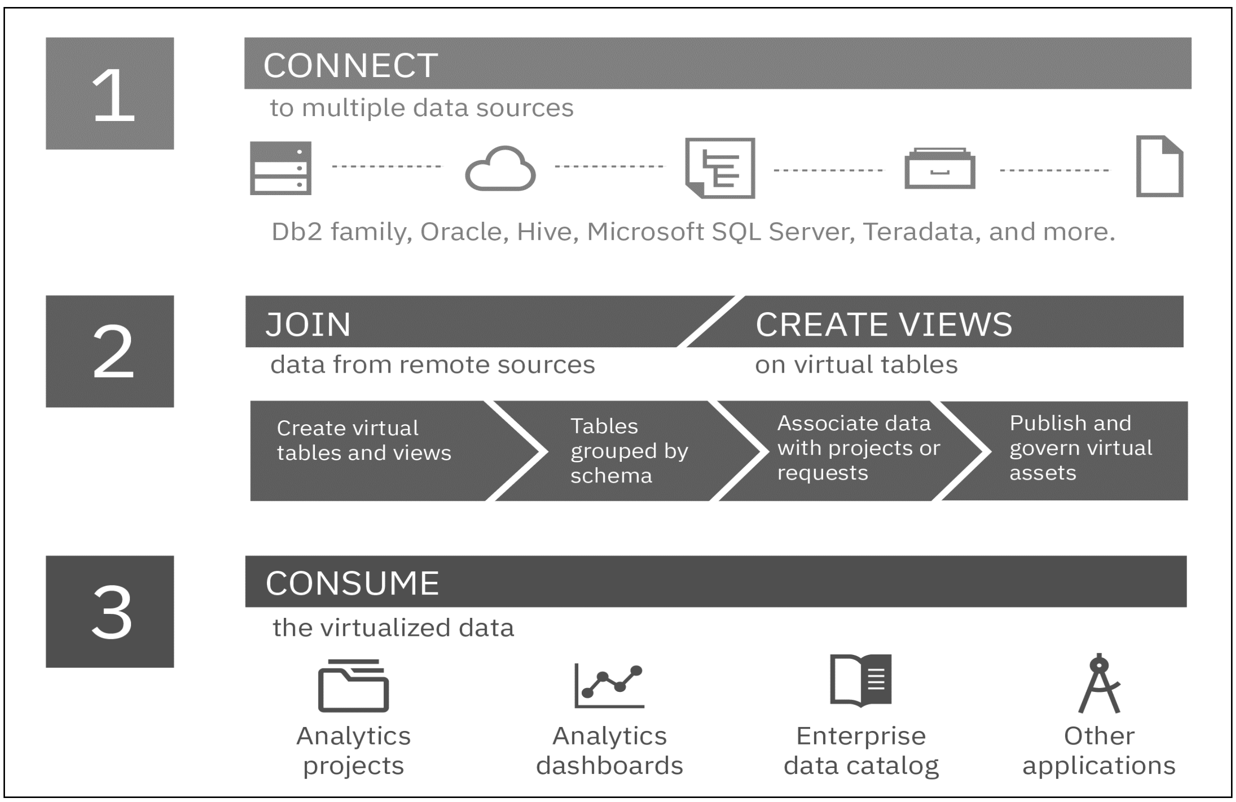

Figure 4-86 High-level overview of IBM Data Virtualization capabilities and setup and usage process

For the purposes of illustrating the capabilities of the service, we concentrate on the functional setup and overview first and delve into the administration of the service later. In a real-life scenario, administration and security setup of the service often must be taken care of up front.

Data Virtualization capabilities can be accessed from the main Cloud Pak for Data menu, as shown in Figure 4-87.

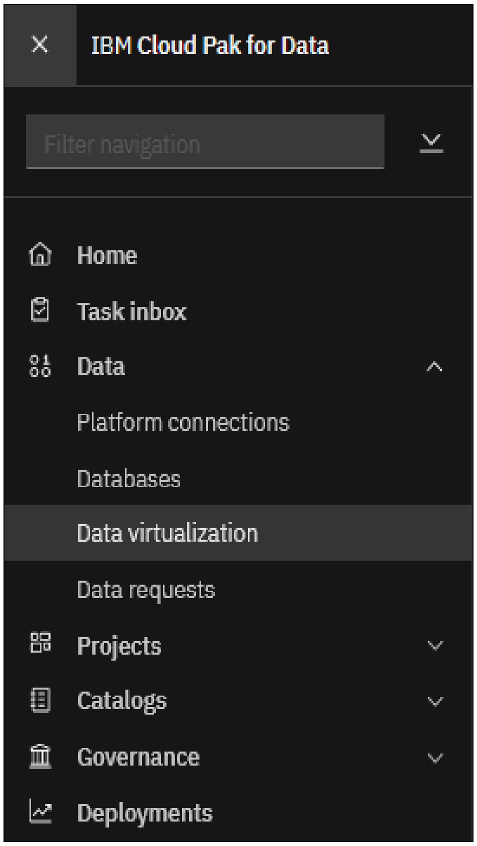

Figure 4-87 Accessing Data Virtualization

The menu entry leads to the default landing page for the service: the Constellation view window, where the list of sources that are connected to the virtualization layer can be viewed (Table view) or as a graph. Figure 4-88 shows this window for a newly provisioned Data Virtualization instance.

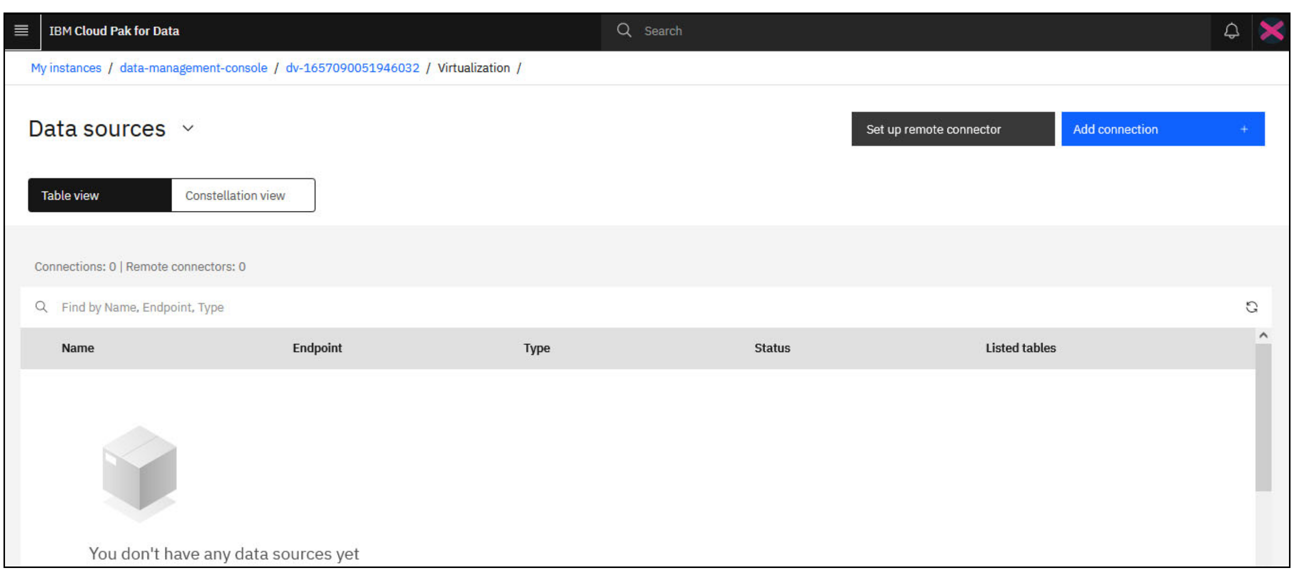

Figure 4-88 Constellation view page of a newly provisioned Data Virtualization service

IBM Data Virtualization automatically creates a self-organizing computational mesh (that is, a constellation) from the data sources (connections) that were added to the Data Virtualization instance. Also included are optional agents that are deployed and running on edge nodes to increase the parallelism and the processing power of the mesh.

When a query is issued against the Data Virtualization layer, the query execution is pushed down to the constellation mesh. The coordinator node of the service receives the request and relays it to the mesh, where its nodes then collaborate with several their peers to perform almost all of the required analytics, and not only the analytics on their own data. The coordinator node then receives mostly finalized results from a fraction of nodes and returns the query result to the requester.

This architectural model delivers improved performance and scalability compared to the more traditional data federation approaches.

Therefore, Data Virtualization allows you to scale the constellation and the computational mesh by adding sources and edge nodes. You also can right-size and scale the coordinator layer and the service instances by adding resources and processing capacity to cater to your expected query workloads, performance, and high availability requirements.

For more information about scaling best practices for the service, see this IBM Documentation web page.

Creating the constellation often starts with adding relevant sources to it. Data Virtualization allows you to reuse connections that were set up at platform level (platform connections). You also can add connections to new sources locally within each Data Virtualization service instance. You can choose either or both of those connection setup options, depending on your use case and access and data separation requirements.

If the data you want to add to virtualize is in a remote file system or within a database on a private server, an extra Remote data source connectivity option is provided. To use this option, you must install a remote connector (edge agent) on the required remote system.

Figure 4-89 shows the options that are provided to you for constellation nodes setup.

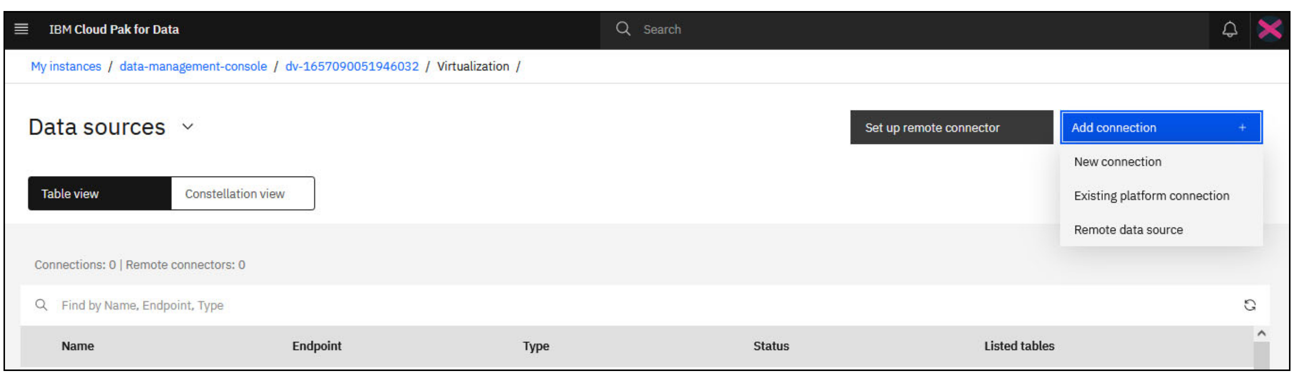

Figure 4-89 Add connection menu

The edge agents do not always have to service a specific file system or facilitate access to a remote data source on a private server. In addition to facilitating access to data and filtering data at the source when dealing with large data sets, remote connectors help improve performance by enhancing parallel processing. Therefore, they can be added for only that purpose. The Data Virtualization service provides another, separate menu option that is called the s*et up remote connector*.

An example of a constellation that is formed from connections, connected remote data sources, and linked remote connectors is shown in Figure 4-90.

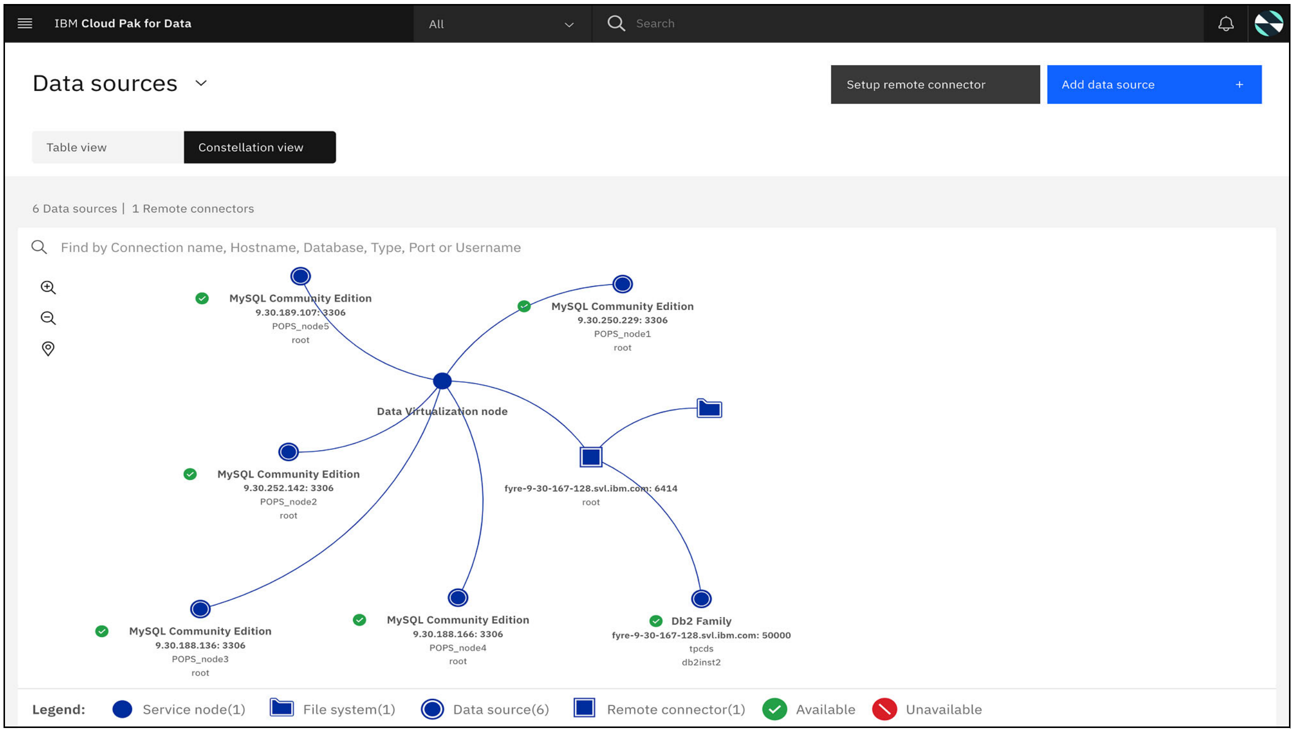

Figure 4-90 A constellation containing file system connections and deployed remote connectors

For more information about the connector setup and remote data source connectivity, see the following IBM Documentation web pages:

- Accessing data sources by using remote connectors in Data Virtualization
- Tutorial: Improve performance for your data virtualization data sources with remote connectors

For our use case, we reuse platform connections and add connections within our service instance. We do *not* set up a remote connector.

To reuse a platform-level connection that was set up in the system, the **Existing platform connection** option is chosen from the Add connection menu. IBM Cloud Pak for Data lists all of the platform connections to which you have access.

Figure 4-91 shows this step of the platform connection reuse process.

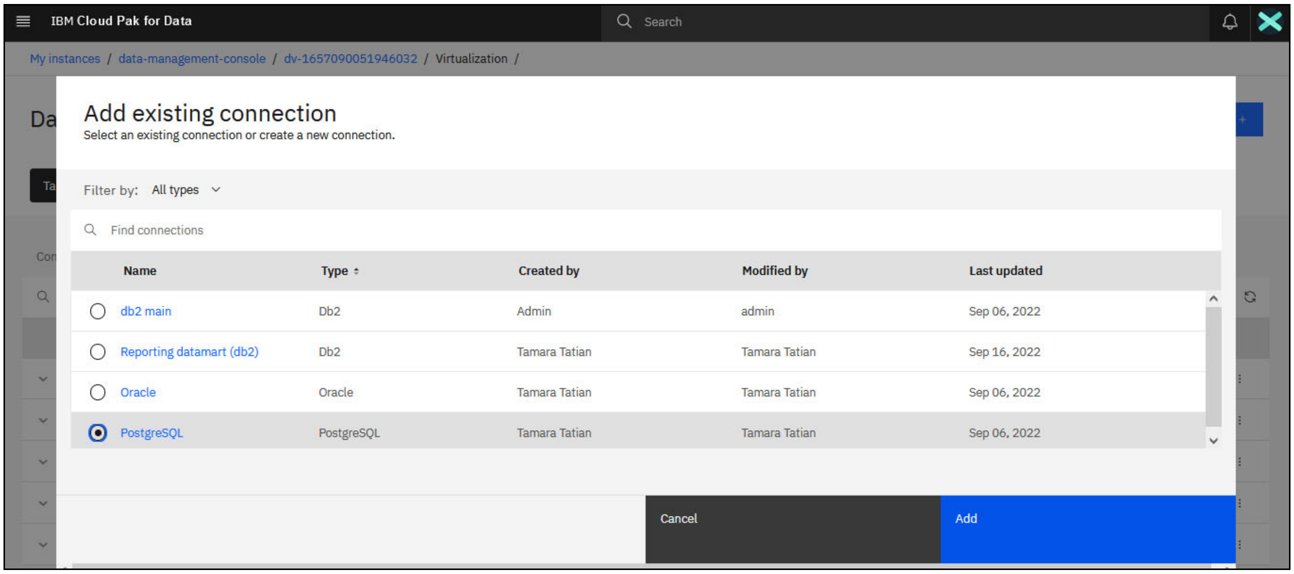

Figure 4-91 Reusing a platform connection in Data Virtualization

After the relevant connection is chosen (in our example, a PostgreSQL), we click **Add** and are taken to the next step of the setup process: setting up the optional remote connector association.

Figure 4-92 shows the choices that are available.

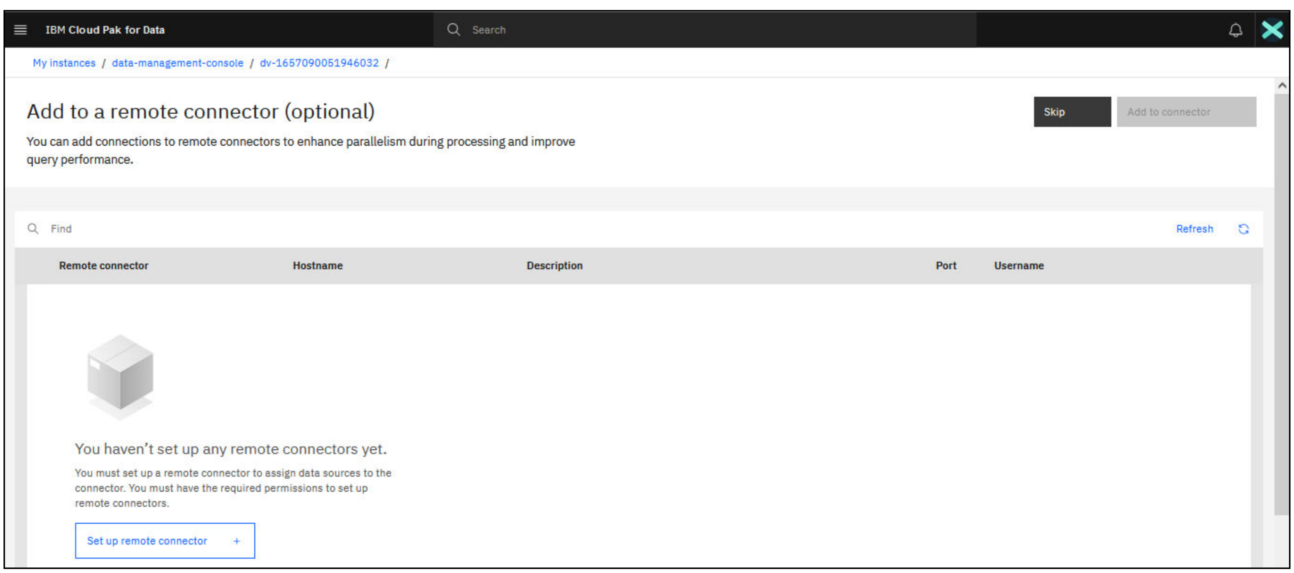

Figure 4-92 Adding a remote connector as part of data source setup

Here, we choose to skip this step. The setup process is now complete. Our platform connection is successfully connected into the virtualization layer.

Repeating this process of adding connections for our Oracle and Db2 platform connections, we now have three different data sources in physically different locations that are connected to our Data Virtualization service instance and transformed into a single virtual data lake.

Figure 4-93 shows the table view of our connected data sources.

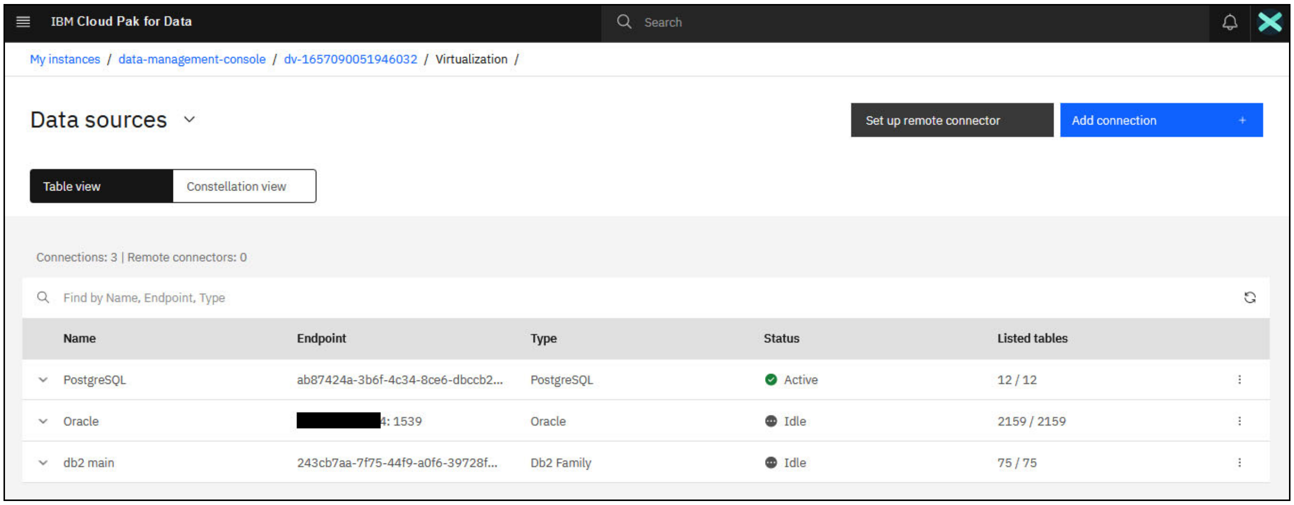

Figure 4-93   Table view of the data sources that are connected to the virtualization layer

Figure 4-94 shows the corresponding constellation view (graph).

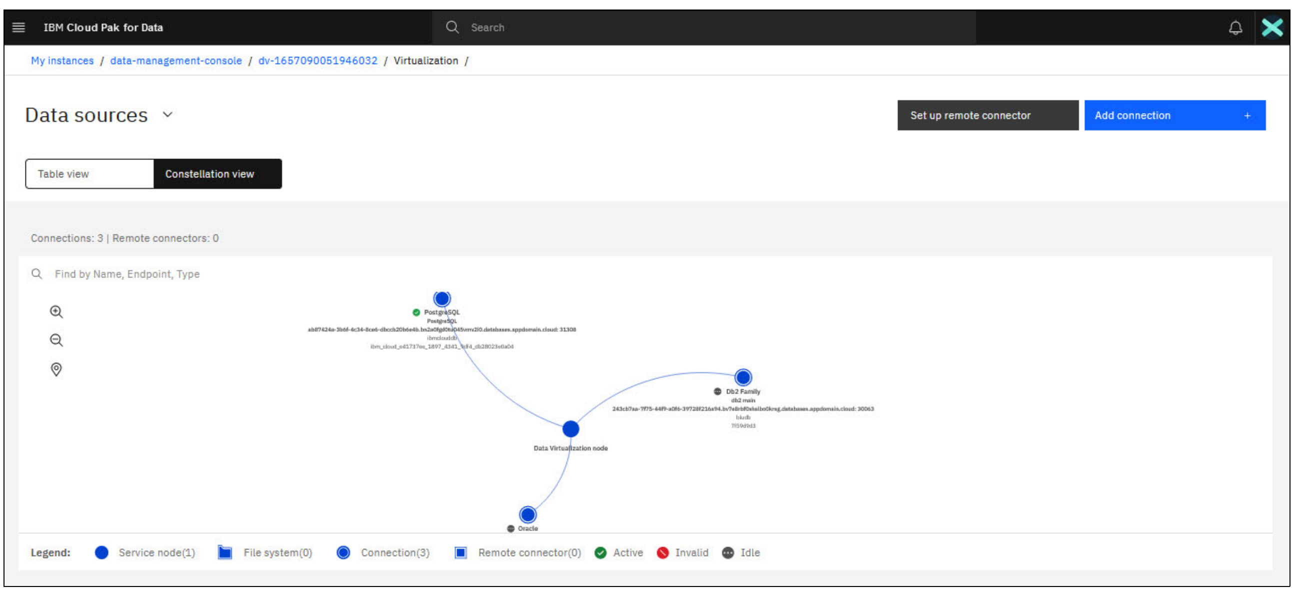

Figure 4-94   Constellation view of the data sources that are connected to the virtualization layer

Next, we add data sources to the constellation. Data Virtualization supports many relational and nonrelational data sources as standard.

Figure 4-95 lists all the data source types that can be used to create connections within the Data Virtualization service in IBM Cloud Pak for Data 4.5.2.

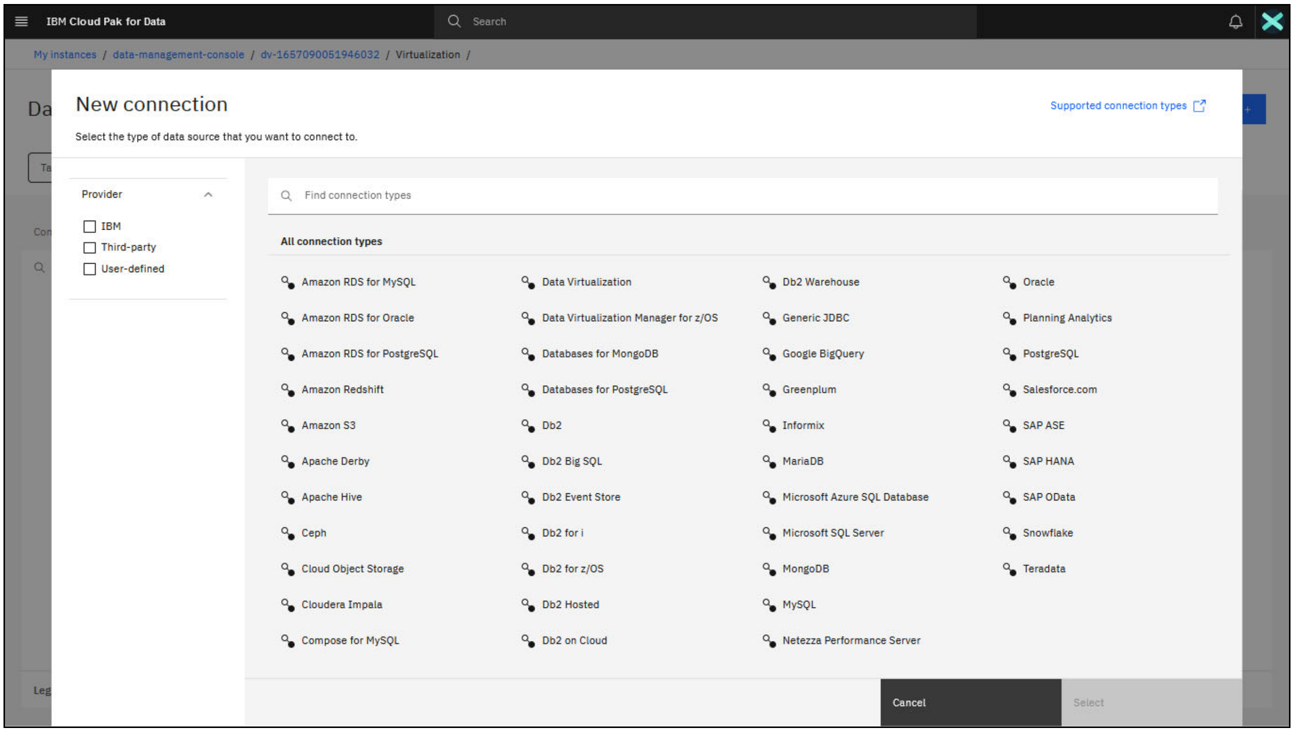

Figure 4-95 Data source types support by the Data Virtualization service in Cloud Pak for Data 4.5.2

For more information about the supported data sources and their setup prerequisites, see this IBM Documentation web page.

We add a MongoDB connection first. Our MongoDB instance that holds interest rate information for our bank is a managed service on IBM Cloud. Therefore, `Databases for MongoDB` is the most suitable connector type choice.

Figure 4-96 shows the first step in the connector setup process. Each connection requires a name and optionally, a description.

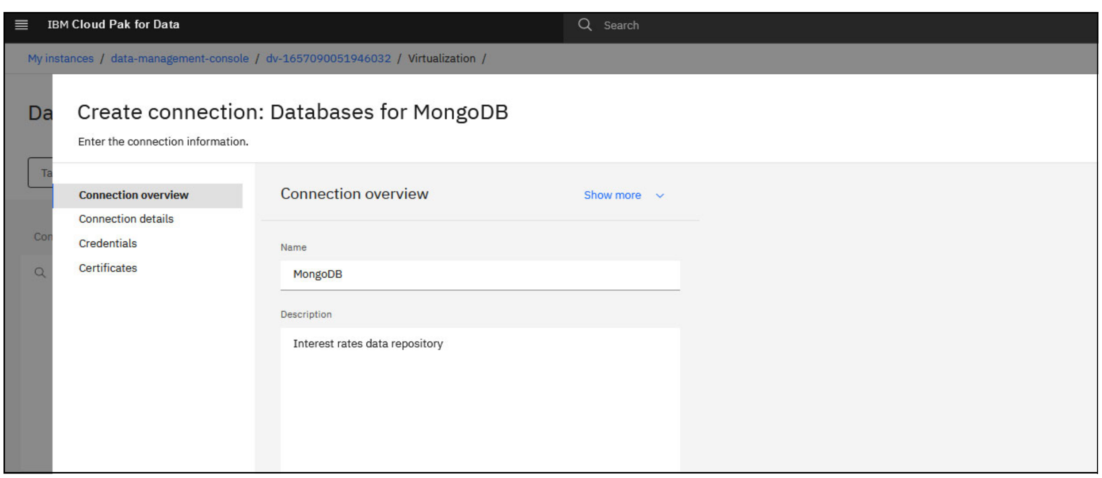

Figure 4-96 Setting up Databases for MongoDB connector

Database, host, and port details must be provided next.

Figure 4-97 shows the relevant setup of these parameters for our MongoDB instance.

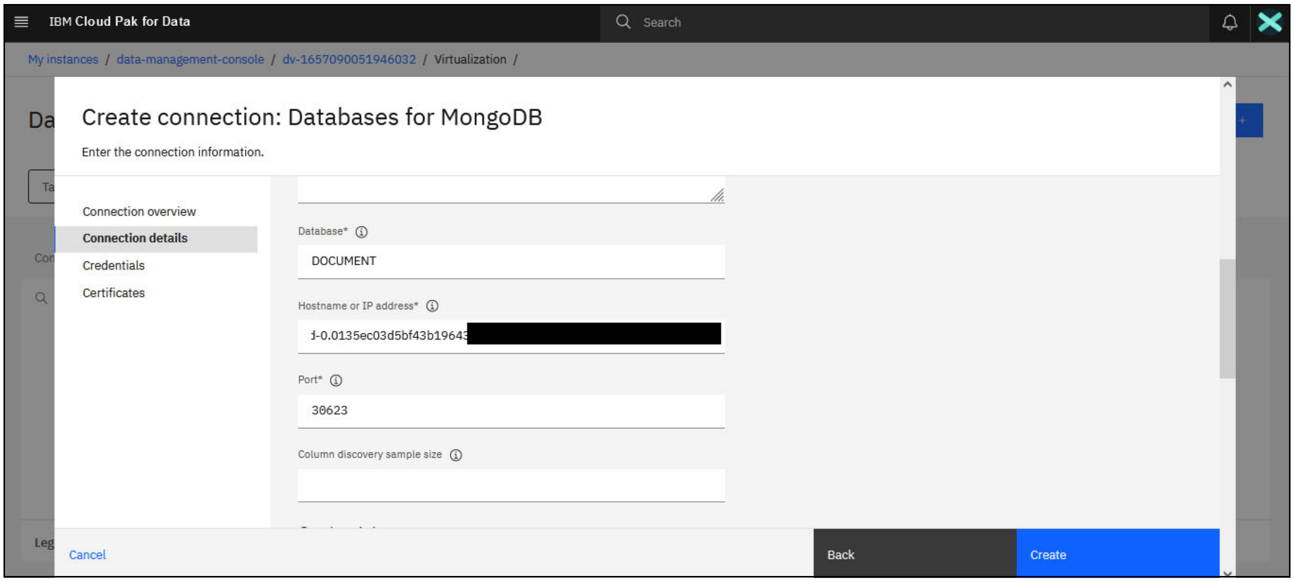

Figure 4-97 Setting up parameters for MongoDB instance

Finally, credentials to the source must be added.

As seen in Figure 4-98, similar to other parts of Cloud Pak for Data, connections in Data Virtualization can be created with Shared setup (one set of credentials that are provided during setup is further reused by other users when accessing data from the connection). A Personal setup also can be used (each user must provide their own credentials to the source to work with the data that comes from it).

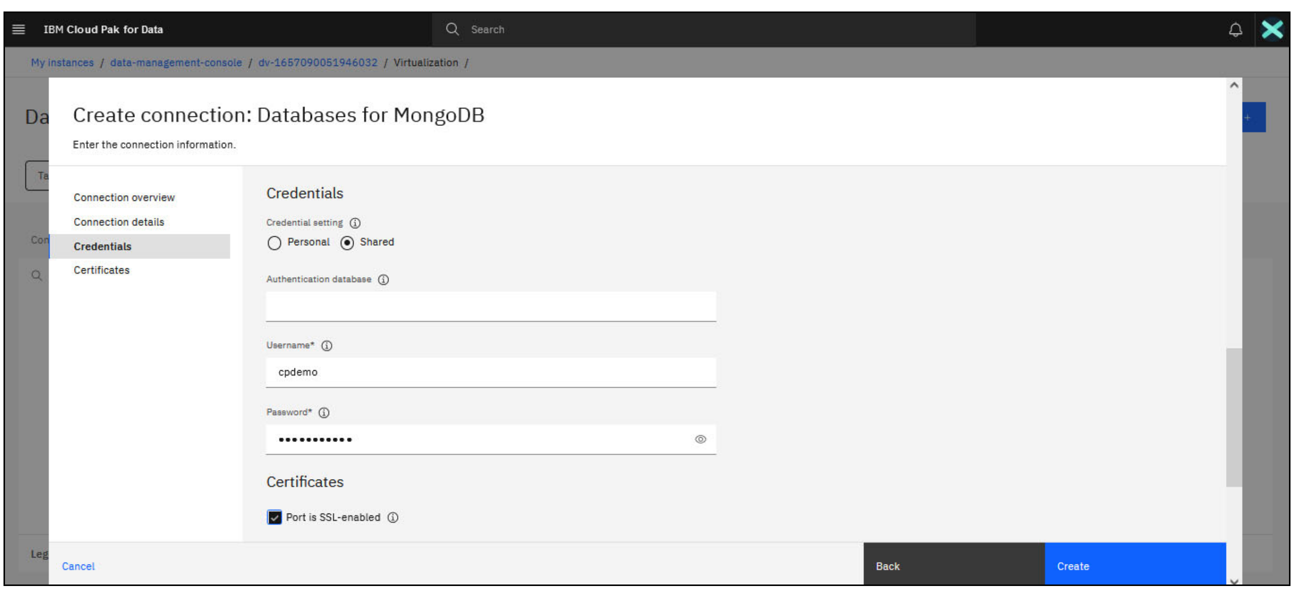

Figure 4-98 Creating connections in Data Virtualization

After all of the required details are provided, we click **Create** to finalize the setup process. Then, the new connection appears in our constellation.

Figure 4-99 shows the updated data source view list with the MongoDB source now added.

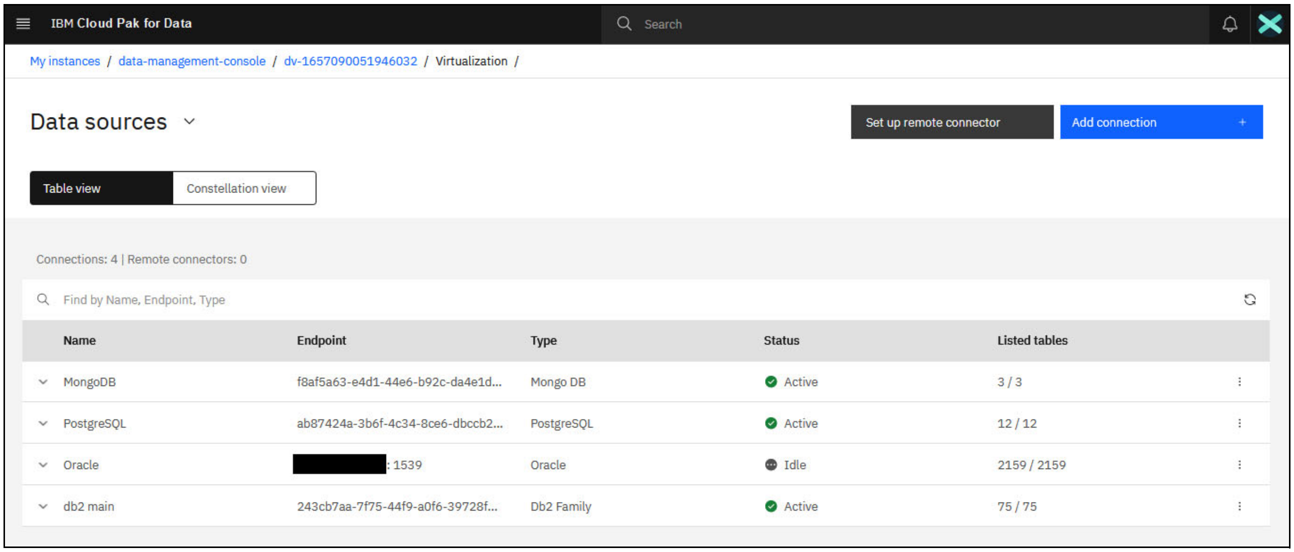

Figure 4-99 New MongoDB connection added to the constellation

We finalize our constellation setup by adding a Db2 Warehouse connection to it. That connection is the Applicants and Applications data that we must join and virtualize in our bank's Db2 Warehouse.

Figure 4-100 shows the first set up window for the connector.

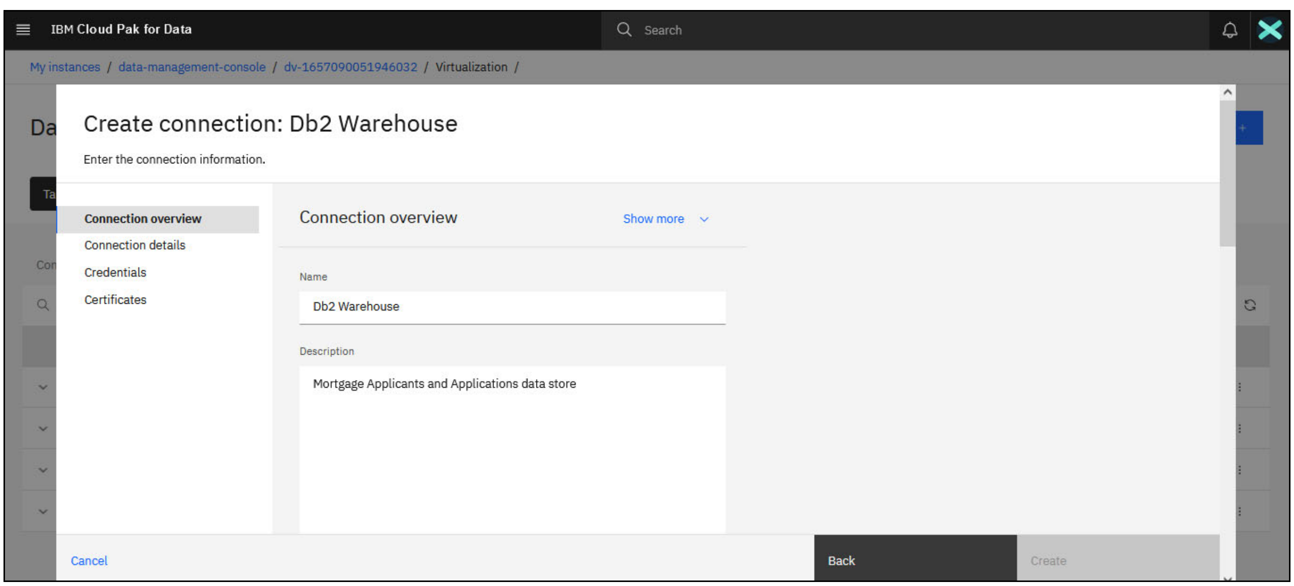

Figure 4-100 Db2 Warehouse connection setup - step 1

Figure 4-101 shows the final setup step. In our example, we connect to the source by using the username and password option rather than the API key option.

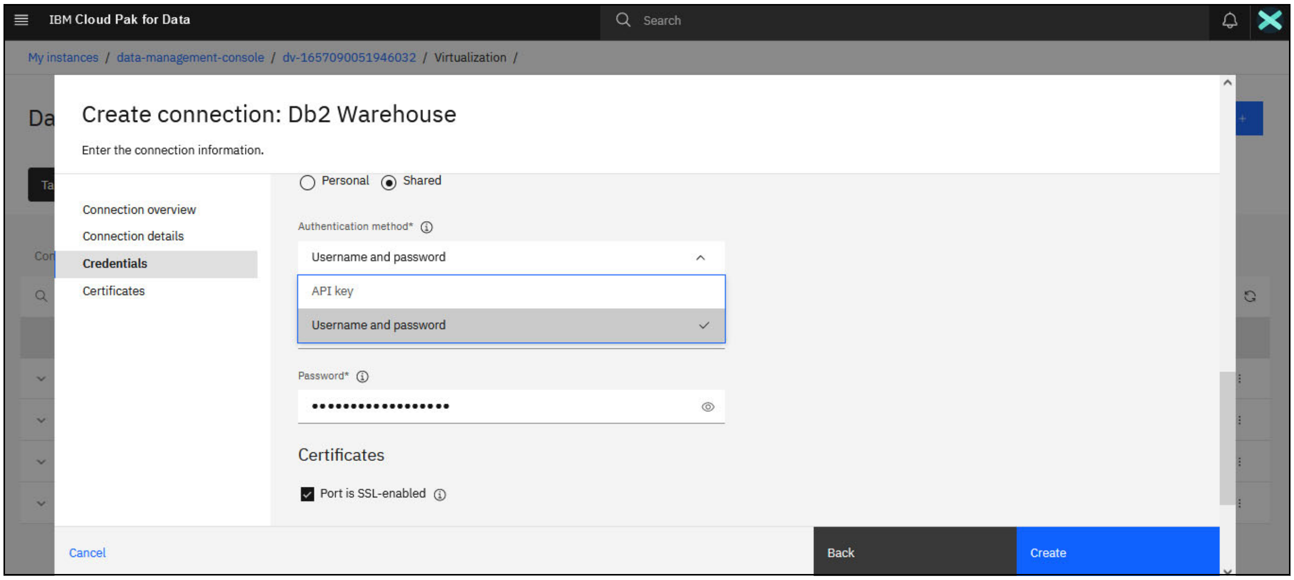

Figure 4-101 Creating a connection in Db2 Warehouse

Figure 4-102 shows the completed constellation setup.

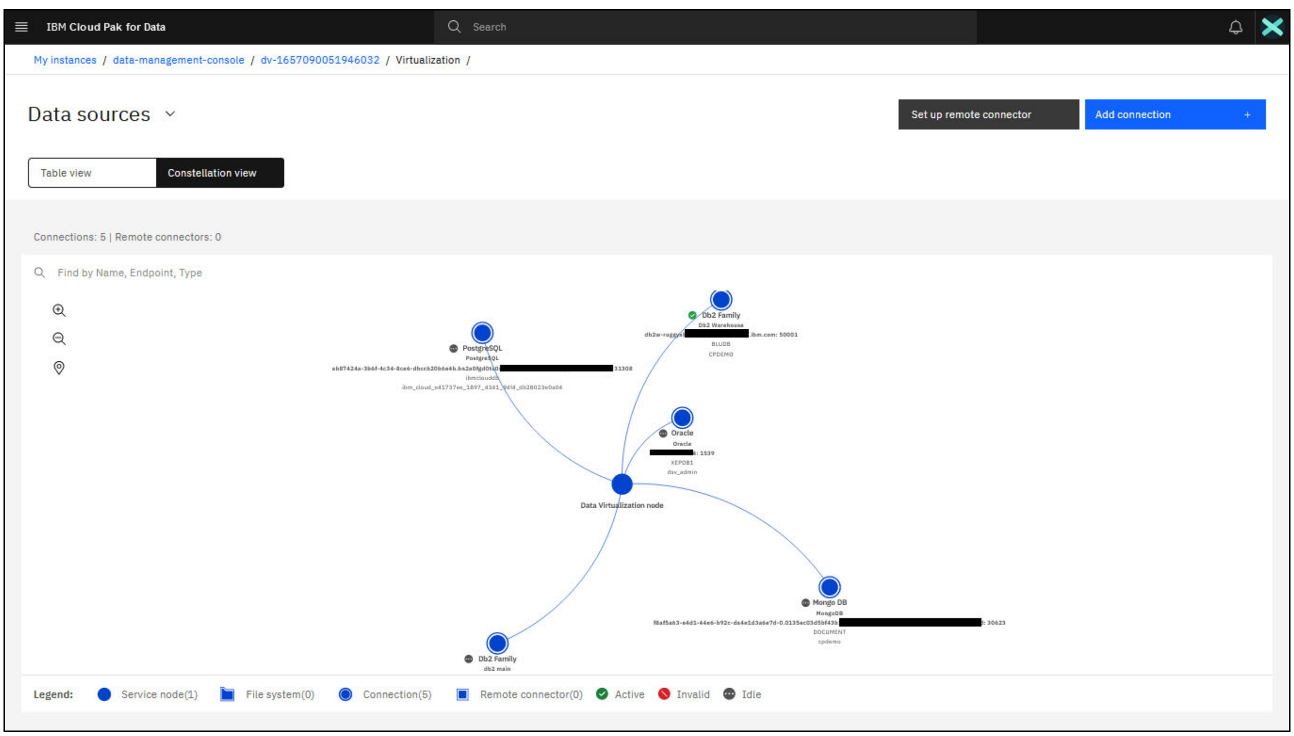

Figure 4-102 Final constellation view

Our virtual data lake is built out of siloed data sources of different types: a NoSQL MongoDB, and relational Db2, Db2 Warehouse, PostgreSQL, and Oracle sources. All of these sources now can be used and queried through the virtualization node as one source. No data was moved or copied out of those sources at any point of this process (see Figure 4-103).

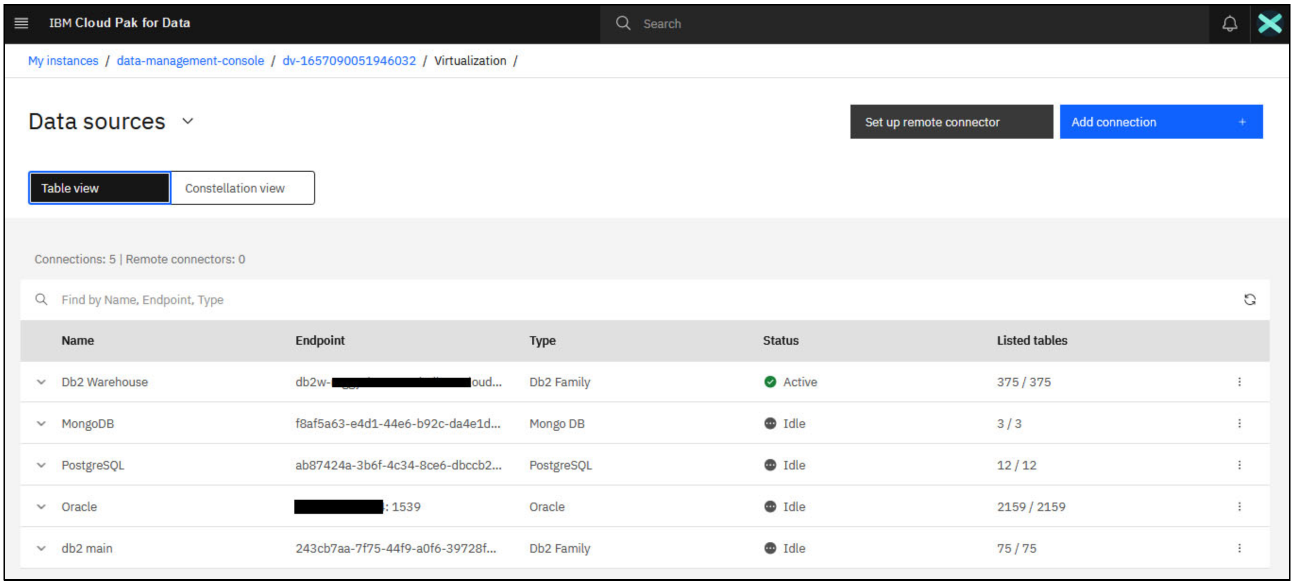

Figure 4-103 Diverse data source types forming our virtual data lake

Also, no data is automatically surfaced within the virtualization layer. The choice of which tables and files to expose into the layer and make available to your enterprise (if any) is up to you. Choice of data selection includes the choice of individual tables, which is described in 4.3.2, “Working with virtual views and virtualized objects” on page 227. Optional extra pre-filtering can be set up by the service administrator for each source, see Figure 4-104.

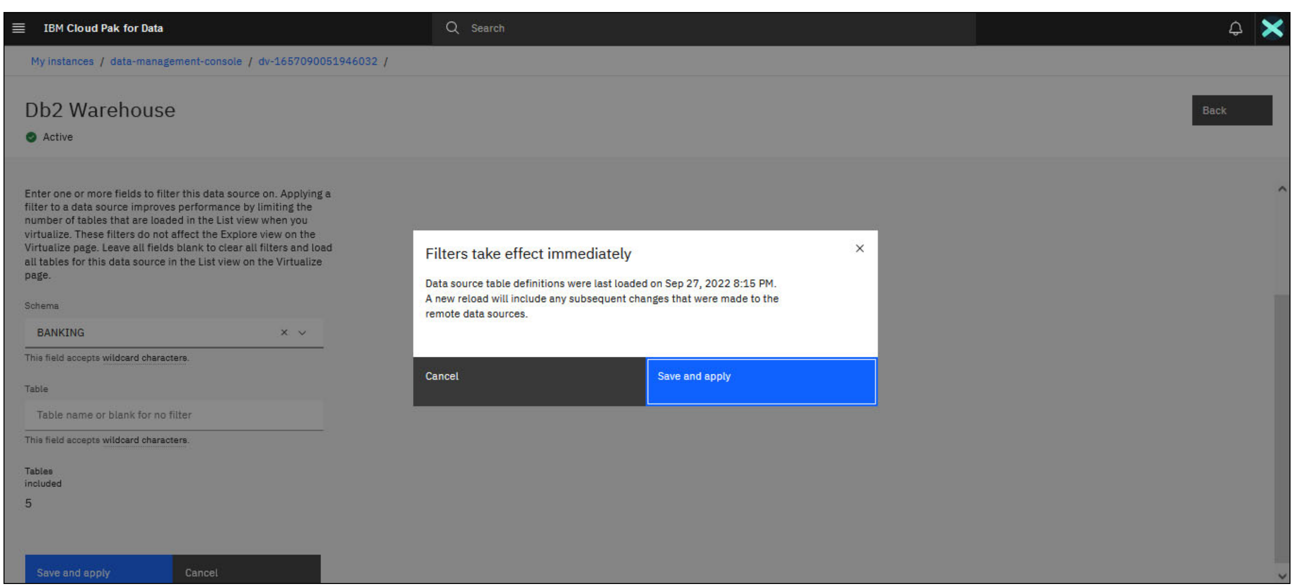

Figure 4-104 Applying filters to a data source (optional)

4.3.2 Working with virtual views and virtualized objects

Connecting relevant source systems to the virtualization layer is the first step of the data virtualization process. The next step is deciding which data from those systems must be virtualized and made available for integration, reuse, and as part of the data democratization initiatives within your enterprise. This task is completed by using the virtualize entry of the Data Virtualization service menu (see Figure 4-105).

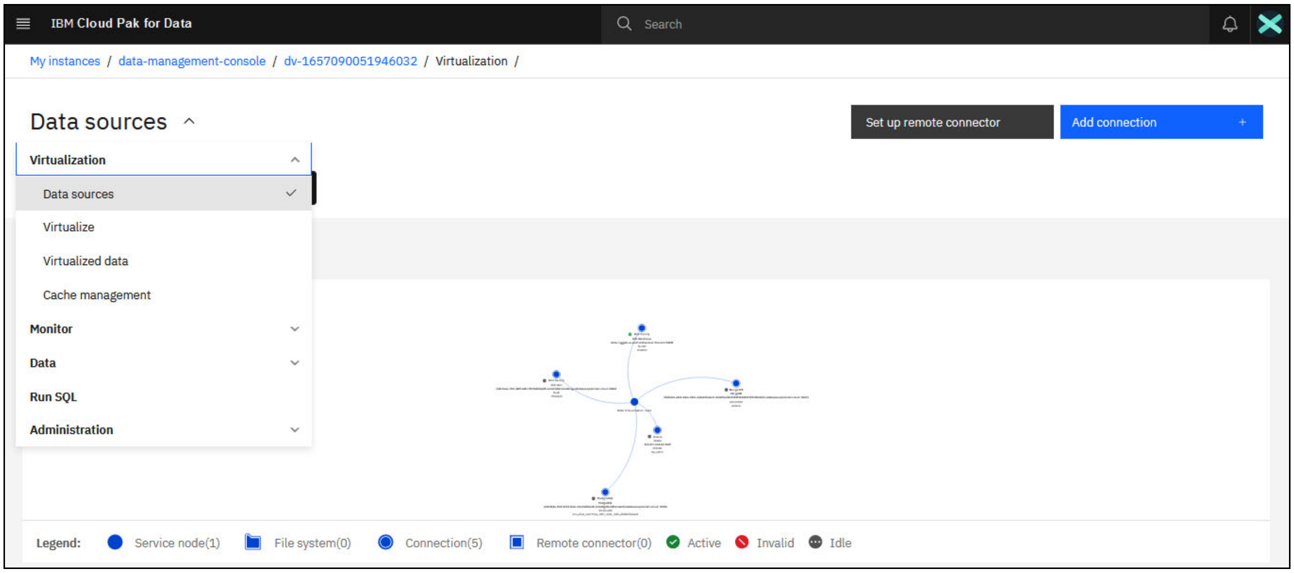

Figure 4-105 Virtualization menu - Virtualize step

The service enables browsing through the connections that are added to the virtualization layer hierarchically, as shown on Figure 4-106, and by searching for relevant objects by way of a list view.

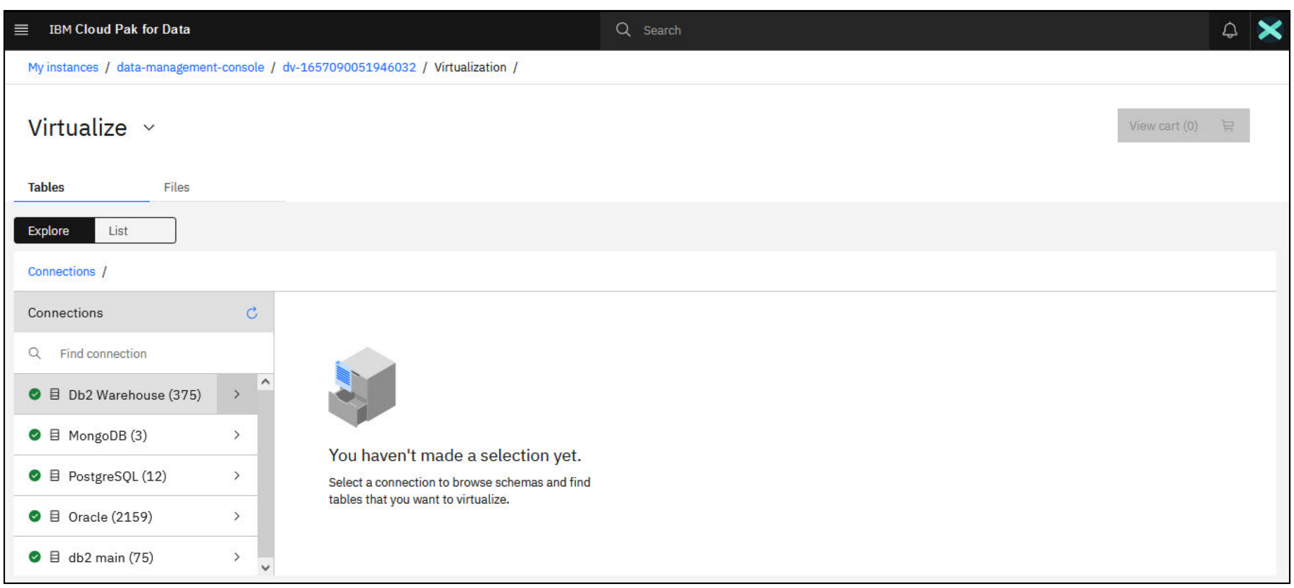

Figure 4-106 Browsing through connections

Relevant tables and files from one or more sources can be added to the list of objects you want to virtualize by clicking the **Add to cart** icon that is next to each object. You also can select one or more objects and click the **Add to Cart link** (see Figure 4-107).

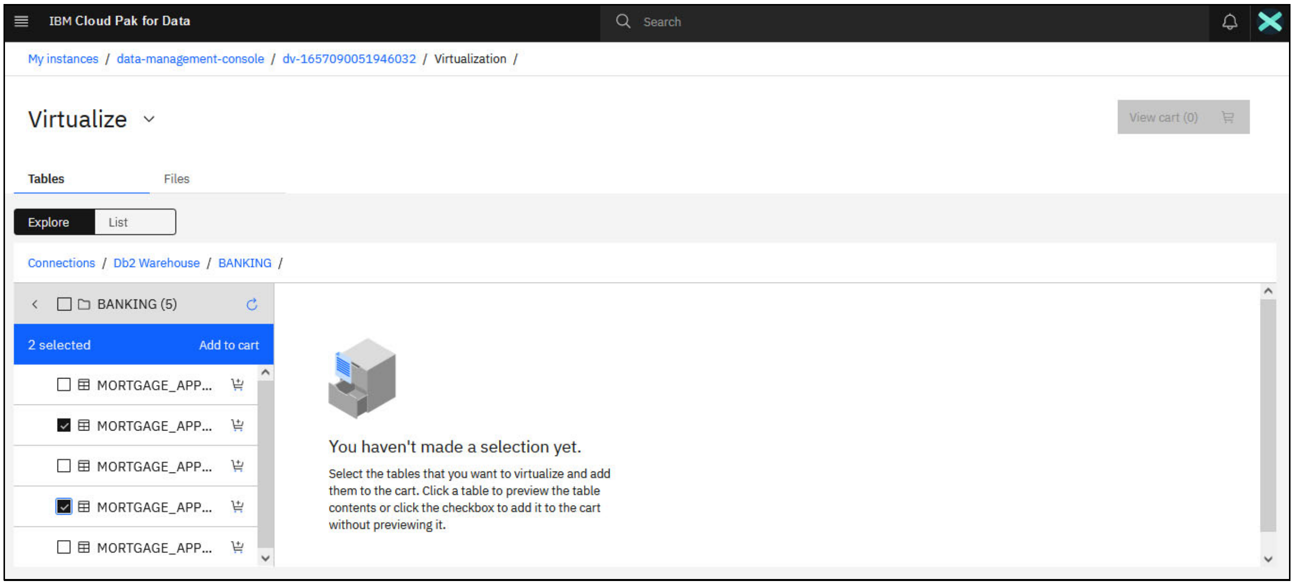

Figure 4-107 Selecting candidate objects

After all relevant objects are selected, click **View cart,** and the Review window opens in which you can further refine your selections. You also can decide how and where to assign and publish the objects, preview source object data, change column selections for each object, and more.

Figure 4-108 shows the Review window and the selection of five objects from three different sources (Db2, Db2 Warehouse, and MongoDB) that we added to the cart.

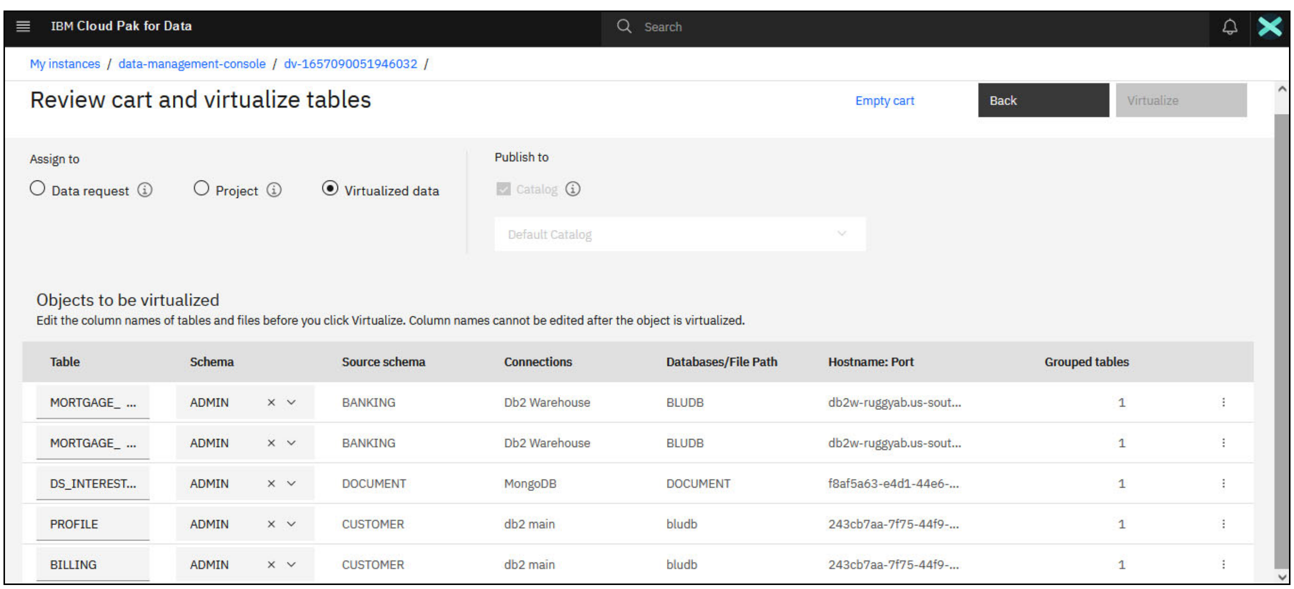

Figure 4-108 Reviewing the virtualization cart

Figure 4-109 shows the preview of the `DS_INTEREST_RATES` document from our MongoDB connection that can be accessed by clicking the **three dots menu** that is next to the corresponding object in our virtualization cart.

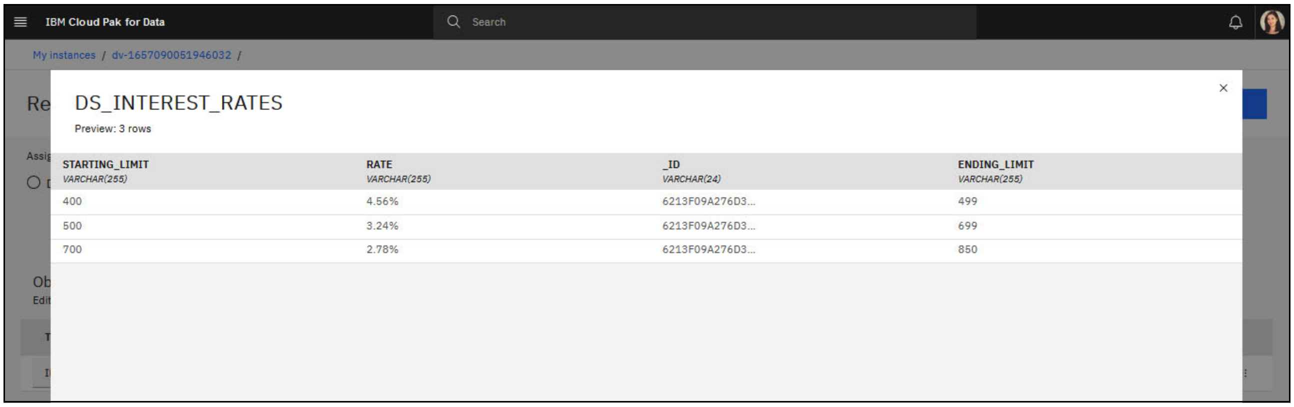

Figure 4-109 Object preview

When virtualizing objects, the resulting virtualized assets can be assigned to different Projects within IBM Cloud Pak for Data, to different Data Requests, or added to the general pool of virtualized data.

For example, a data engineer, a business analyst, or a data scientist might want to use a virtualized view in an ETL flow to build a dashboard, in a Jupyter Notebook to build a data science model, or for other analytical purposes. In this case, you can assign the virtualized object directly to the Project with which those users are collaborating and performing their work.

If you are fulfilling a data request by virtualizing the object, you might want to choose the assign to Data Request option. Cloud Pak for Data users can submit data requests if they cannot find the data that they need, or if they cannot access a data source where the data is stored. The requests can be fulfilled; for example, by virtualizing the relevant data with IBM Data Virtualization or Transforming the relevant data with IBM DataStage.

In our case, we place all our virtualized objects into the shared virtualized data pool.

An IBM Data Virtualization instance presents itself to external systems and internally within IBM Cloud Pak for Data as an IBM Db2 database. The database can be connected to by external systems by way of JDBC, queried by using ANSI-standard SQL or APIs, and monitored, administered, and used within Cloud Pak for Data.

The service organizes data that surfaced within the virtualization layer into schemas of this database, with directly virtualized objects appearing as Nicknames within the schemas. Views are created by using the Data Virtualization service as Views. Therefore, schema selection forms part of virtualized objects setup.

When you first see the Review Cart object virtualization window, the default private schema that was created by the system for your user shows as auto assigned to all objects in your cart. A user’s default schema name is the same as their user name in the system.

Users that are assigned the Data Virtualization Admin role within each Data Virtualization instance also can choose to create schemas. They also and must grant suitable permissions to other users or roles to enable them to use those schemas to create virtual objects.

Figure 4-110 shows how a new schema that is called `banking` can be easily created by our Administrator user in the Review cart window.

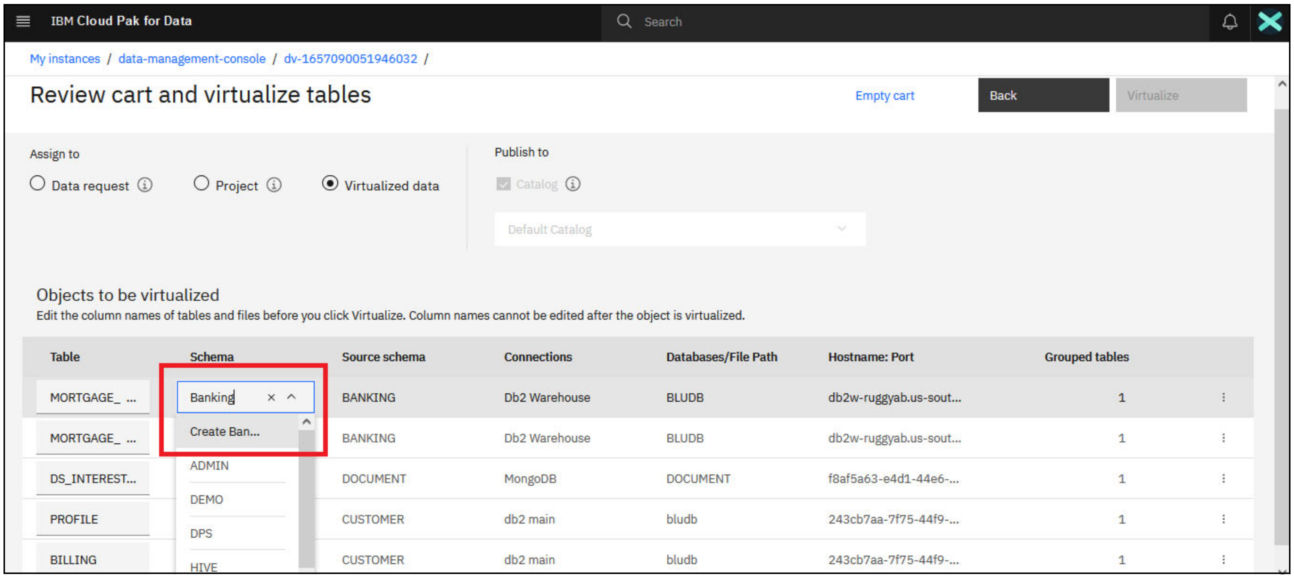

Figure 4-110 Creating a new schema for the virtualized object

Let us create another schema that is called `Customer`, and assign those new schemas to the objects that we are planning to virtualize. We also edit virtualized object names to reflect how we want them to appear in the system (see Figure 4-111).

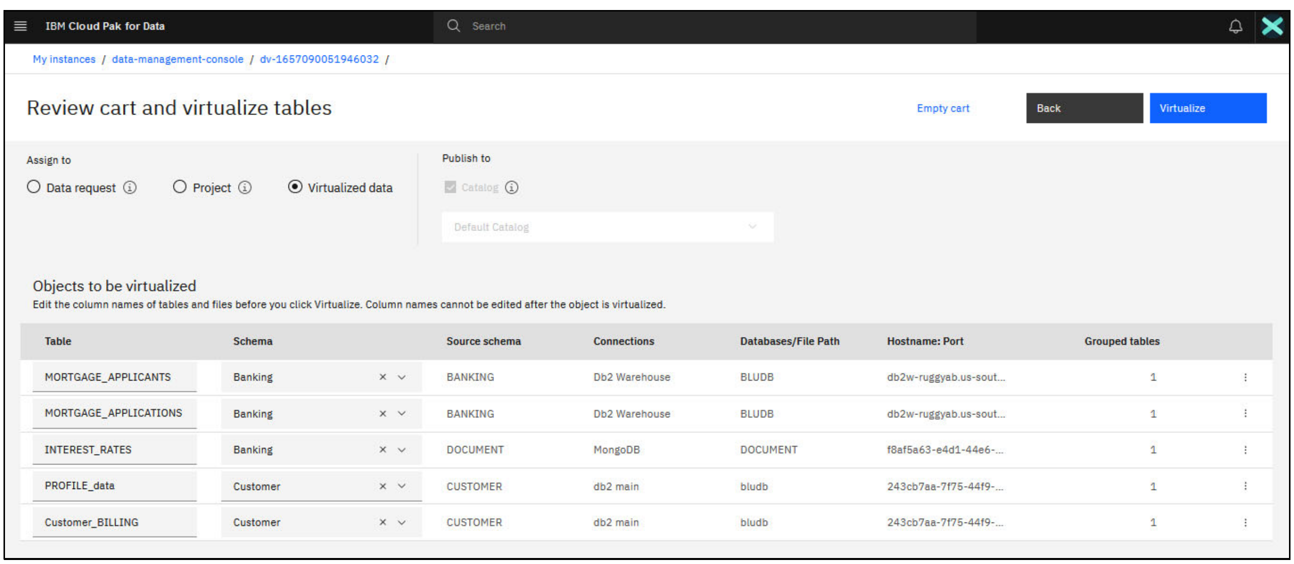

Figure 4-111 Assigning custom schemas to objects

By using the platform, you also can edit columns and refine column selection for any of the objects in your cart. The three dots menu that is next to the object that is shown in Figure 4-111 provides the corresponding option.

The editing process is shown in Figure 4-112.

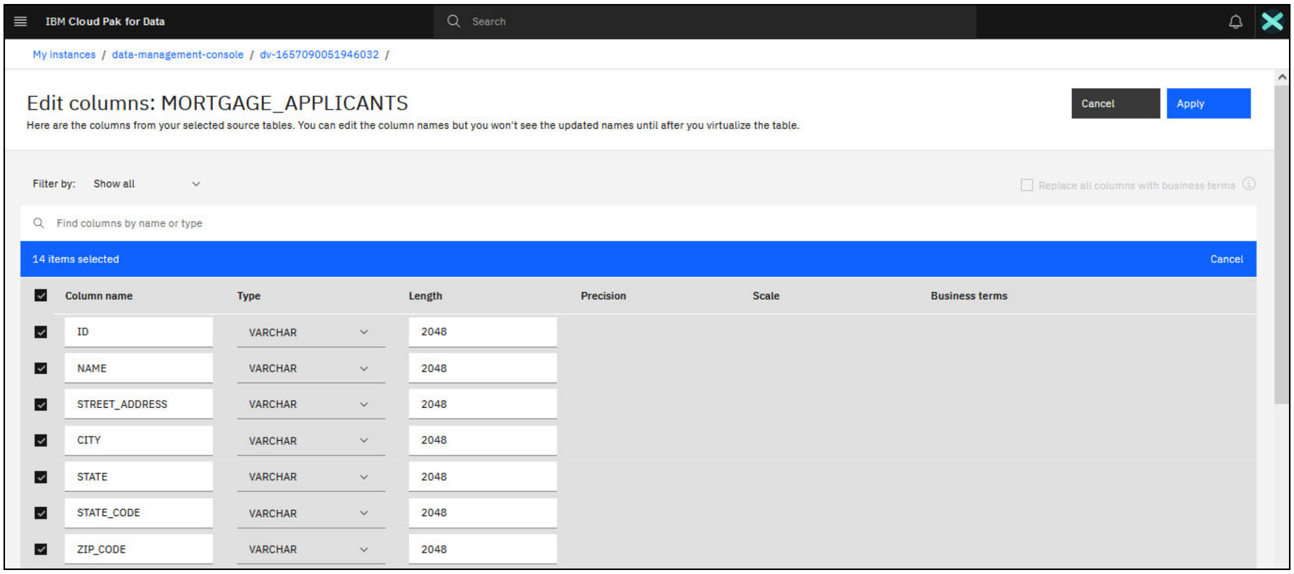

Figure 4-112 Refining column selection for the virtual object

After all of the relevant selections and changes are made, clicking **Virtualize** triggers object virtualization (see Figure 4-113).

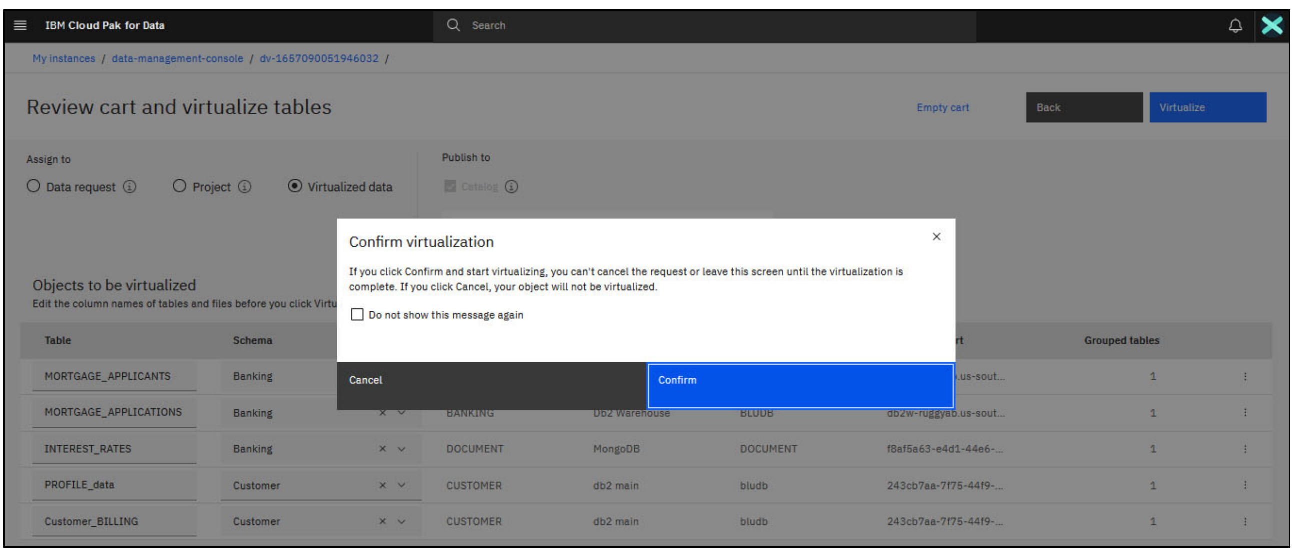

Figure 4-113 Triggering the Virtualize process

Depending on the setup, the system can automatically publish the resulting assets to a governed catalog within the solution. You also can choose which catalog to publish the assets to, if at all. In our case, we choose to publish the resulting assets.

After the process is complete, you can browse to your newly virtualized assets directly from the confirmation window, or choose to virtualize more assets (see Figure 4-114).

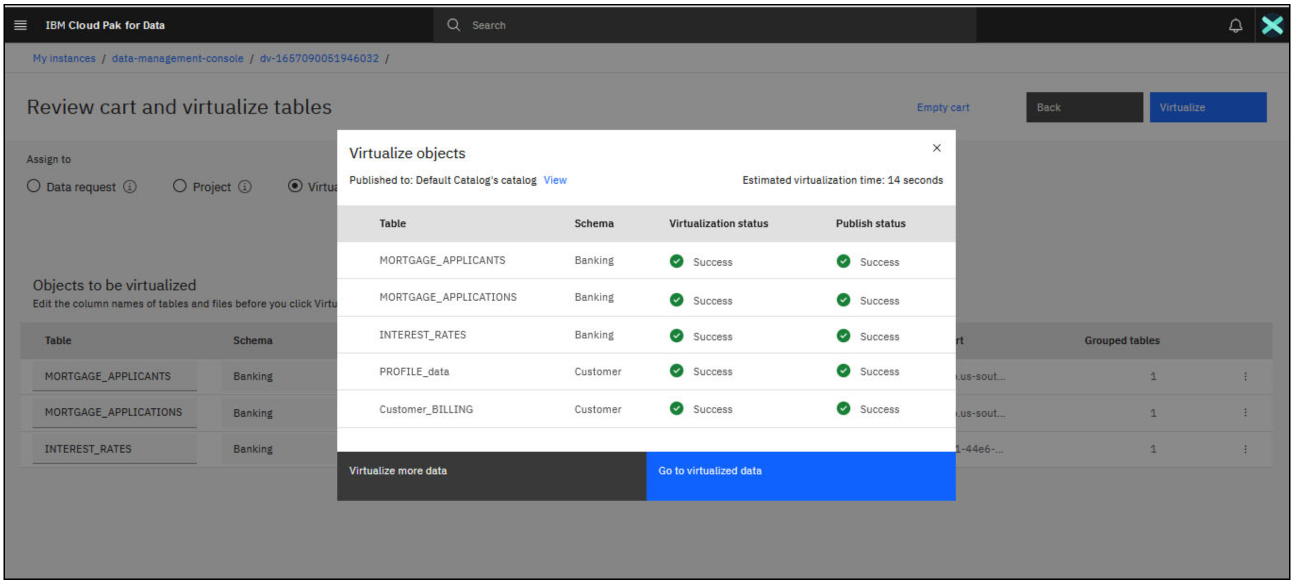

Figure 4-114 Successfully completed virtualization

Now, we can browse to our newly virtualized assets list.

Figure 4-115 shows the new assets that we created. You also can browse to this page by selecting the Virtualized data option from the Virtualization menu of the service. Each object also includes a list of other actions that can be performed. Next, we preview the `MORTGAGE_APPLICANTS` object.

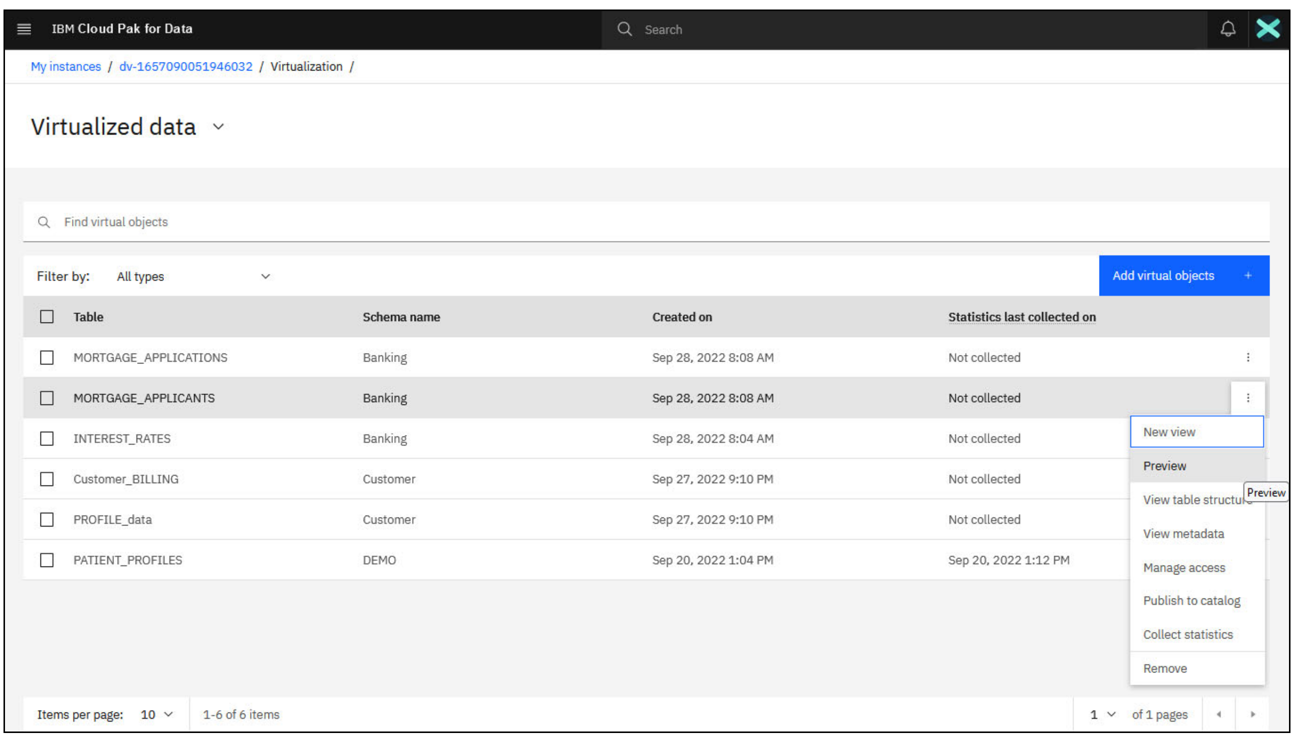

Figure 4-115 Virtualized objects list and object menu

The preview contains a sample of the data that is coming from the source system. Beyond the captured metadata and a small data sample that is pulled for data preview purposes, we are not moving or copying any data from the source system; the data stays at source. We can also view our virtual table structure and the associated metadata in this window (see Figure 4-116).

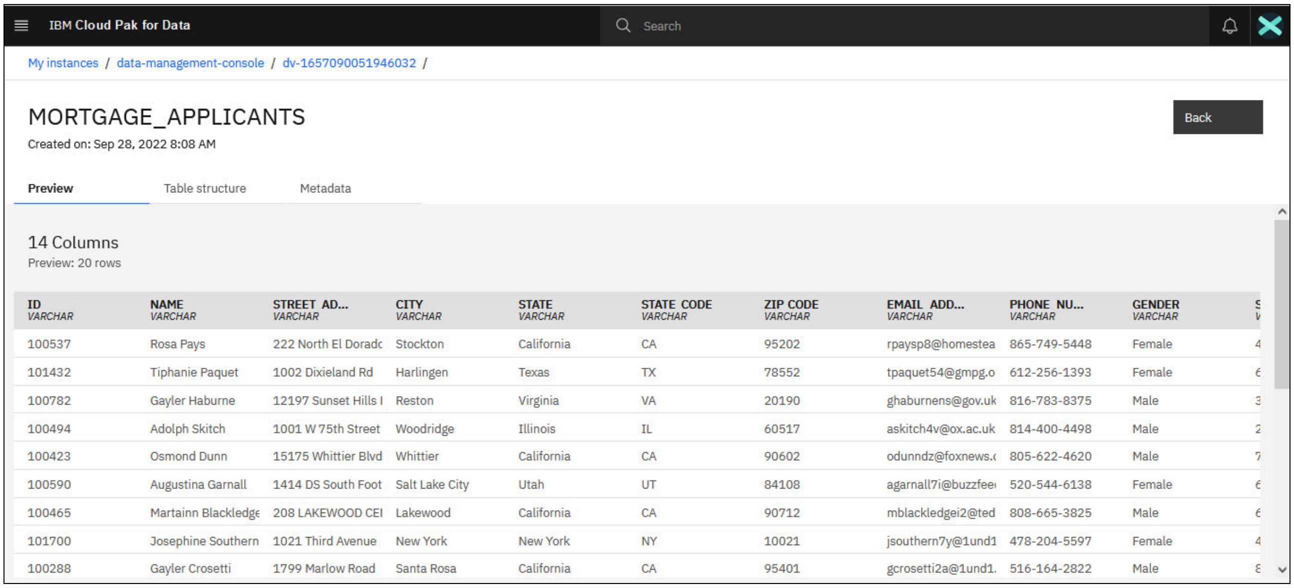

Figure 4-116 Virtualized asset preview

IBM Data Virtualization is tightly integrated with the IBM Watson Knowledge Catalog service of Cloud Pak for Data. As a result, data protection rules that are defined within Watson Knowledge Catalog can be enforced within Data Virtualization for relevant users and user groups.

Owners (creators) of the object are the only users that are automatically exempt from this enforced rule application. An example of this governance rule enforcement for the same `MORTGAGE_APPLICANTS` virtualized view (as seen by a different user) is shown in Figure 4-117.

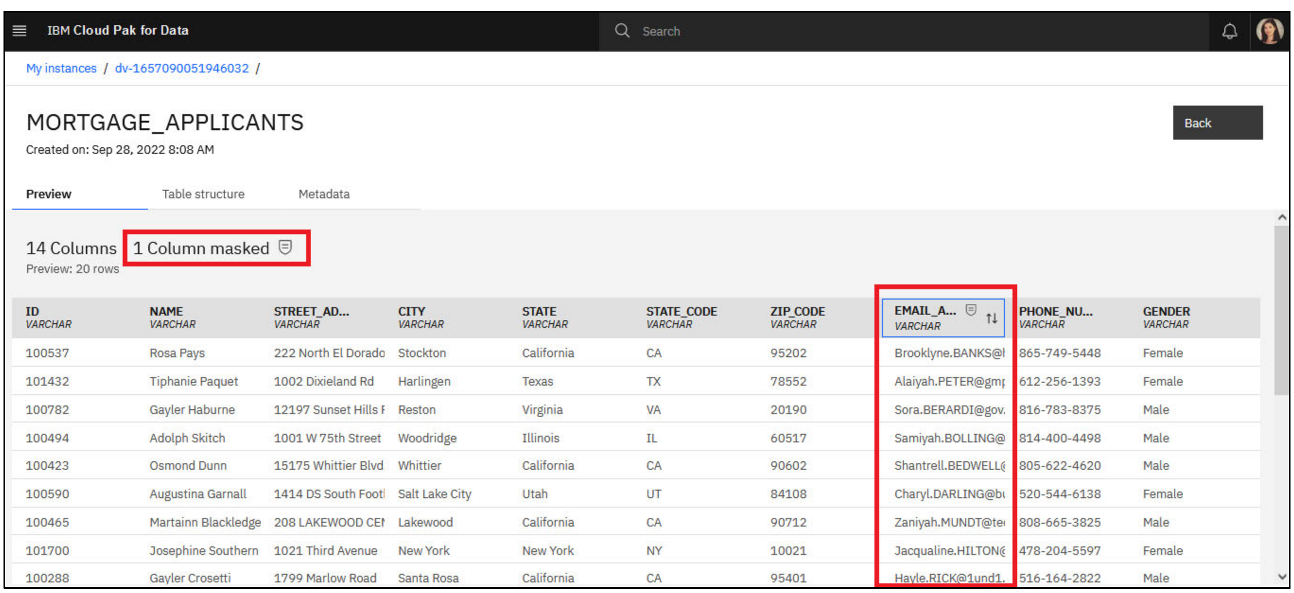

Figure 4-117 Data masking rule enforcement in a virtualized view

Another useful feature that can be started from the object menus on the Virtualized data window is the Collect Statistics feature, as shown in Figure 4-115 on page 232.

IBM Data Virtualization features a built-in query optimizer engine that uses statistics about the data that is queried to optimize query performance. Accurate and up-to-date statistics ensure optimal query performance. It is recommended that you collect statistics whenever the following conditions apply:

- A table is created and populated with data.
- A table's data undergoes significant changes, such as the following conditions:
  - New data is added
  - Old data is removed
  - Data is updated

You can choose to collect statistics as a one-off exercise, or automate the process by creating a schedule.

Figure 4-118 shows the first step in the Collect statistics setup process: selecting a relevant column.

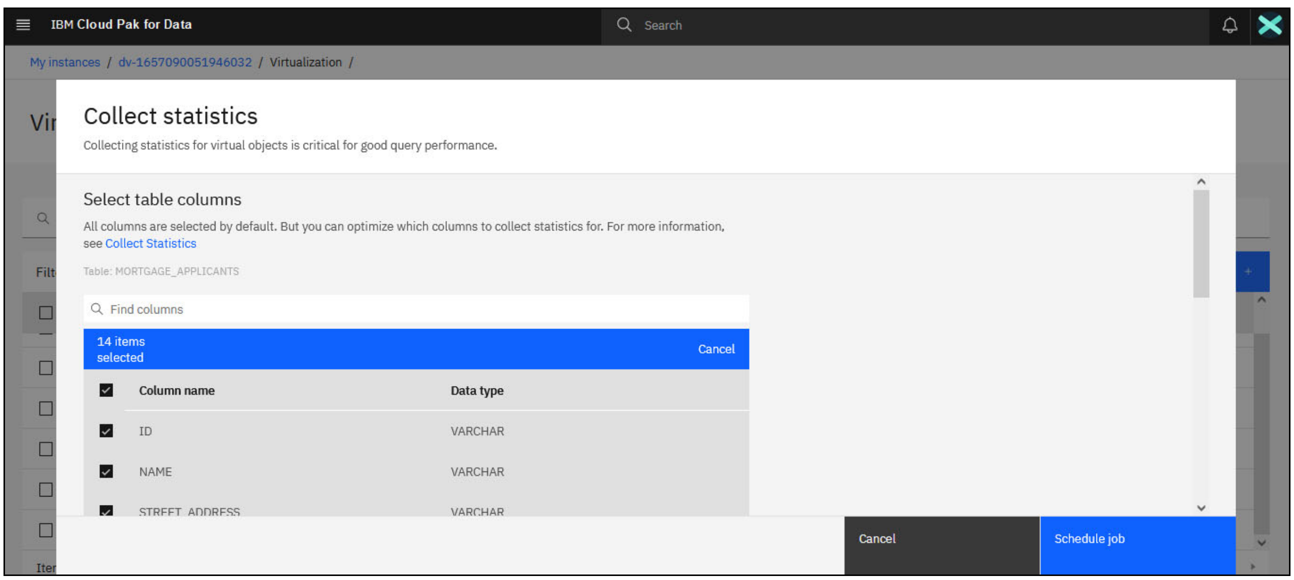

Figure 4-118 Selecting columns

The next step is to create and set up the corresponding job performing the collection, including the relevant schedule (see Figure 4-119).

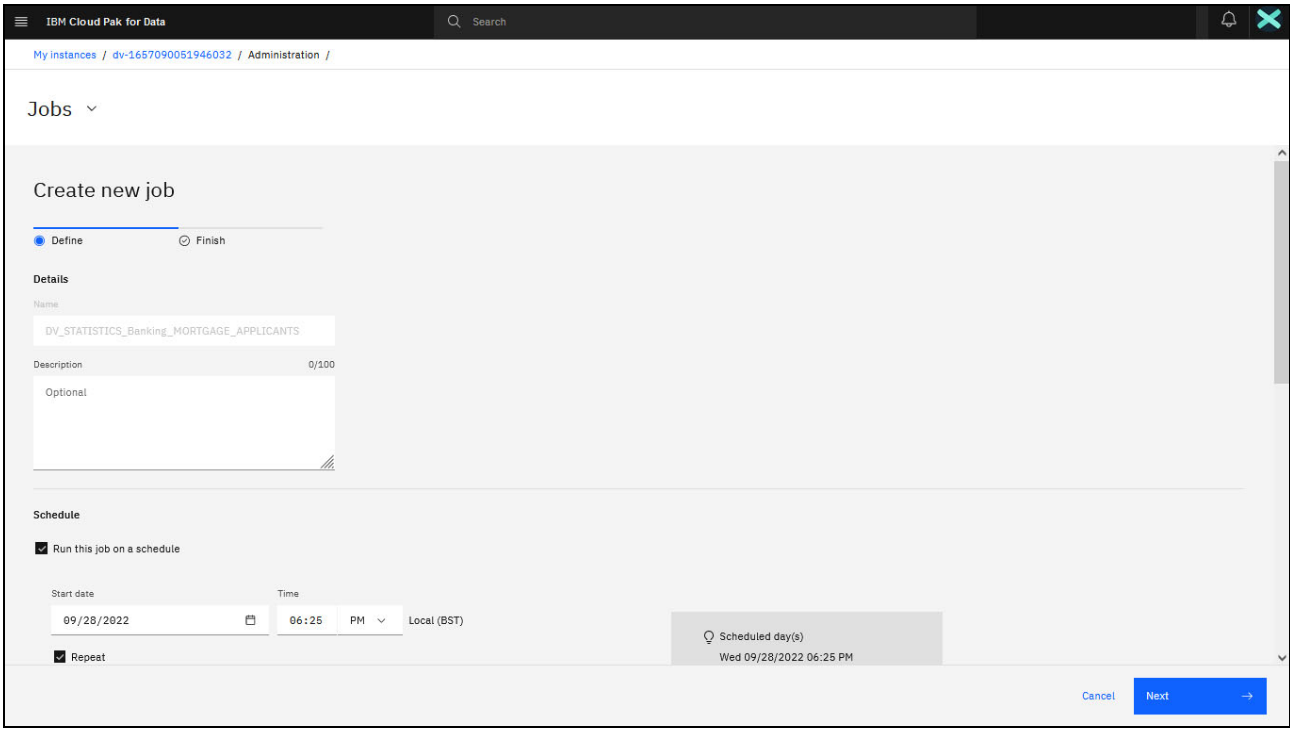

Figure 4-119 Setting up the job

The job schedule can be customized to your needs and preferences and to reflect how frequently the data is expected to change at the source (see Figure 4-120).

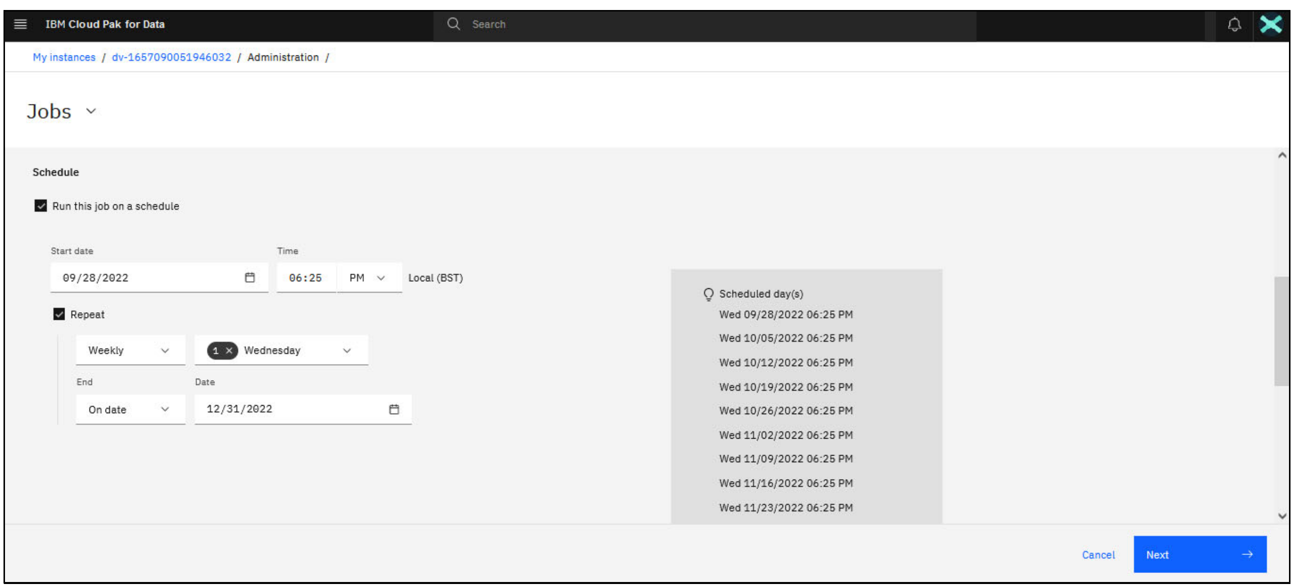

Figure 4-120 Setting a schedule

You also can set up your notification preferences (see Figure 4-121).

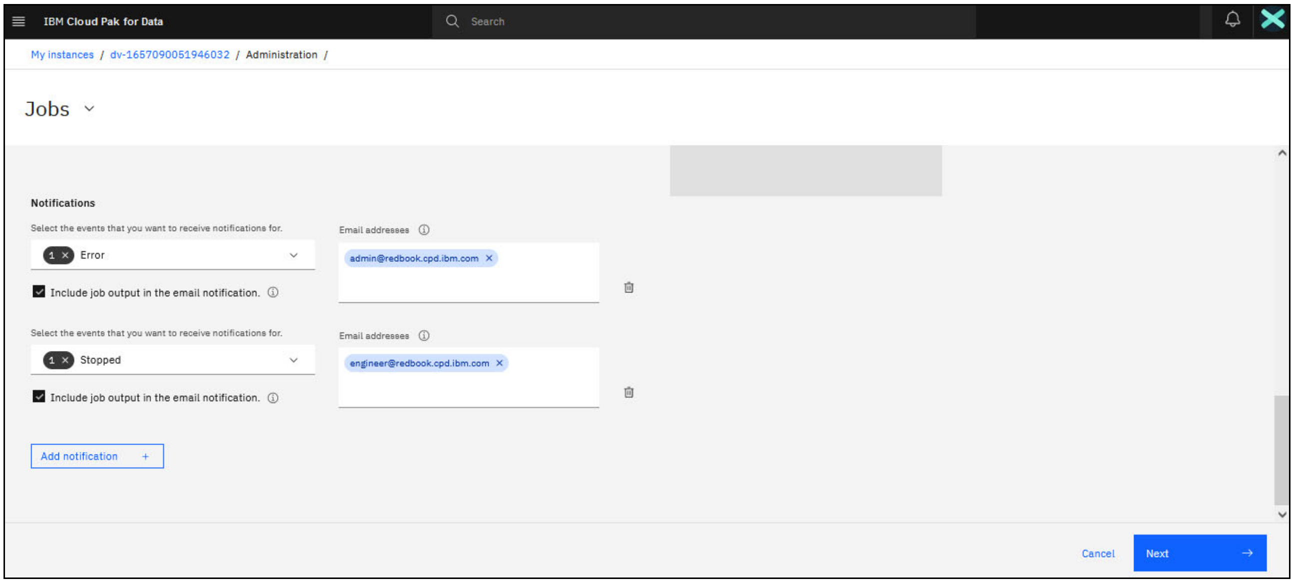

Figure 4-121 Setting up job notifications

Figure 4-122 shows the final setup for our job, which we commit to by clicking **Schedule**.

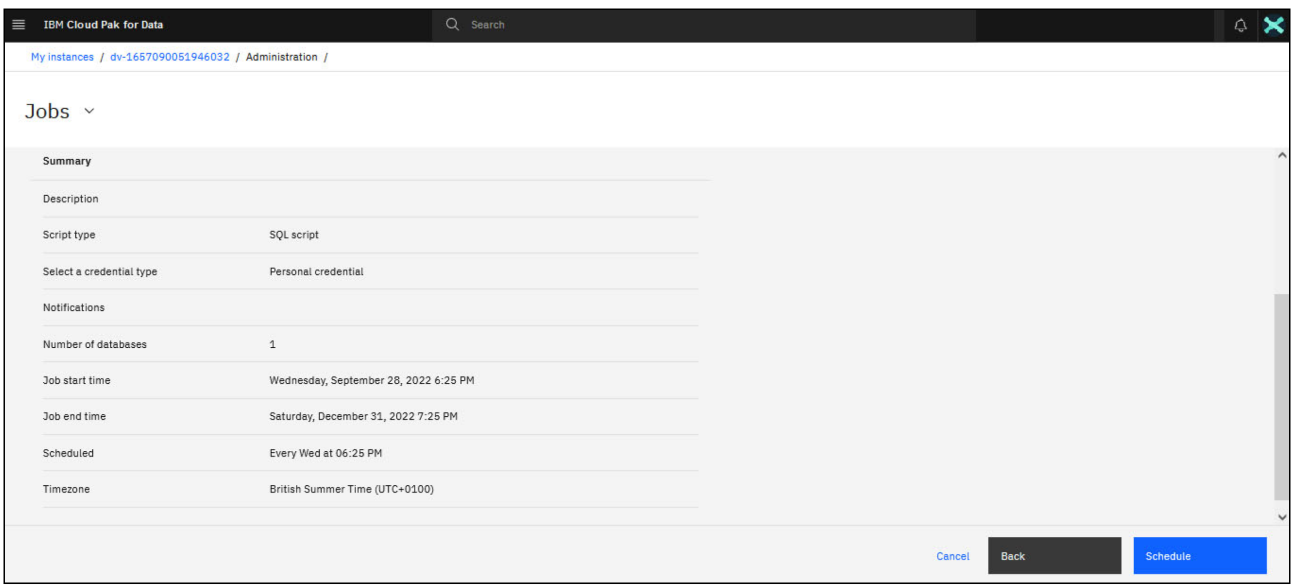

Figure 4-122 Completed job setup

The job now runs weekly according to our schedule. The statistics for all of the runs and the run history details are logged by the system and available for review and analysis (see Figure 4-123).

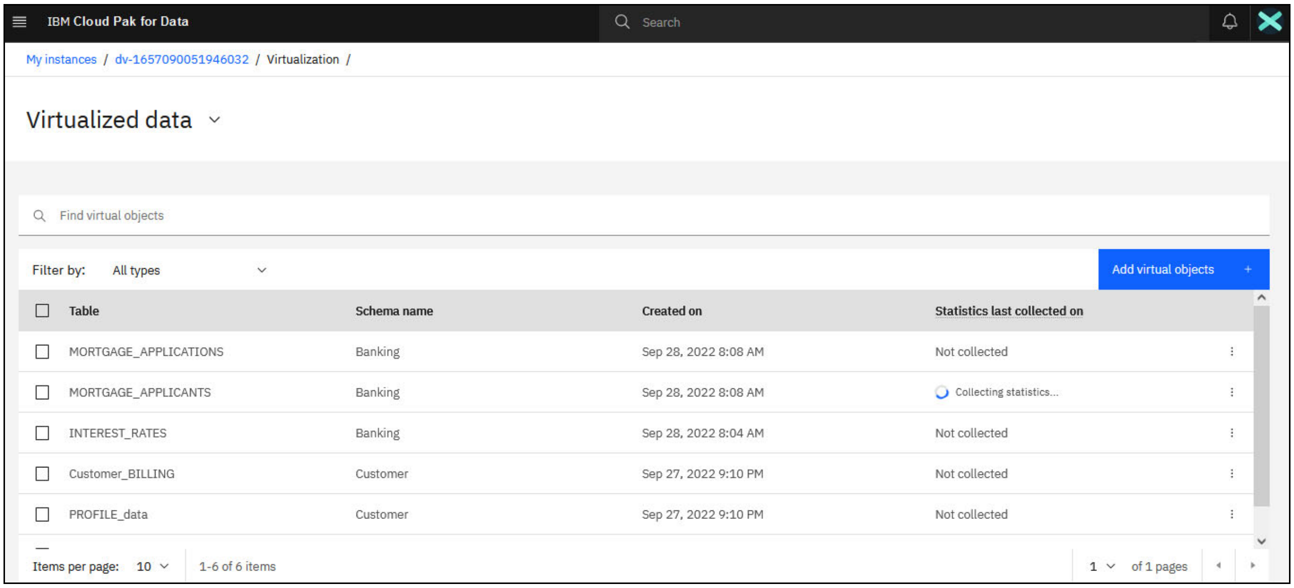

Figure 4-123 An in-progress statistics collection job

Creating a virtual view from virtualized objects

Next, we create a virtual view from several objects we virtualized previously.

As part of the ETL use case in this chapter, we used IBM DataStage to physically join data from our `MORTGAGE_APPLICANTS` and `MORTGAGE_APPLICATIONS` tables as part of a transformation flow integrating our data. This process involved physically extracting the data from the source Db2 Warehouse tables.

We now achieve the same result of joining data from those two tables, but without extracting or moving any data from our source systems or building any ETL jobs. The resulting joined virtual asset then can be queried directly and reused by different users and in relevant processes as needed.

Data Virtualization offers an intuitive graphical editor for joining pairs of data sets.

Selecting an object from your virtualized objects list unlocks another menu, as shown in Figure 4-124. To unlock the graphical editor-based Join option in that menu, two objects must be selected.

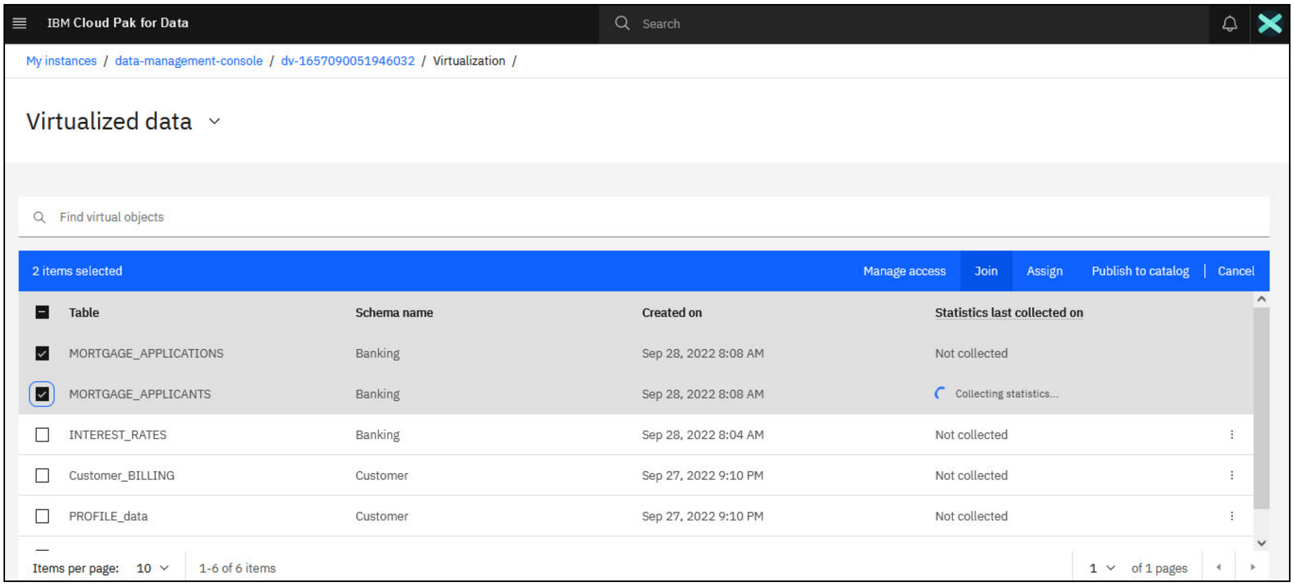

Figure 4-124 Manage access, Join, Assign, and Publish to catalog

Clicking **Join** takes us to the editor, where the join easily can be performed by dropping the relevant column name from the designated Table 1 in your selection to the relevant column name in the object that is designated as Table 2.

You also can choose to de-select the columns that you do not need in the resulting joined view. In our example, we de-select two columns from our Table 4-1 on page 165.

Figure 4-125 shows what a successful join. The key icon appears next to the names of the columns that were chosen as join keys.

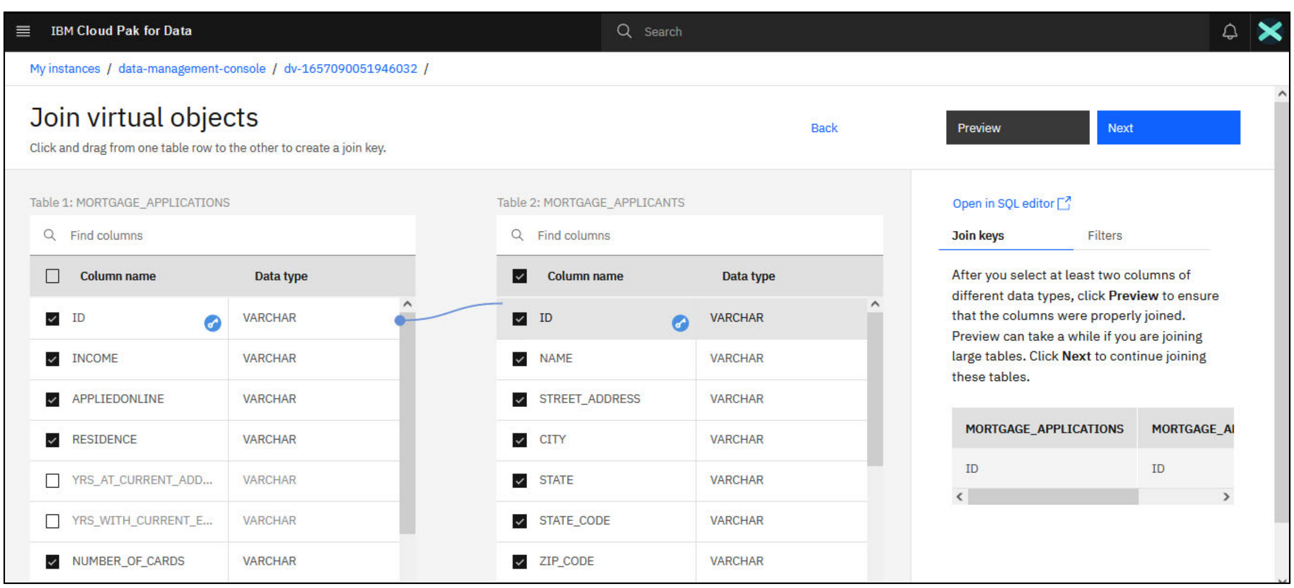

Figure 4-125 Joining virtual objects using the graphical editor

You can choose to apply more filters to your view. By switching from the Join keys tab to the Filters tab on the right side of the window, you can specify the required filtering conditions, as shown in Figure 4-126.

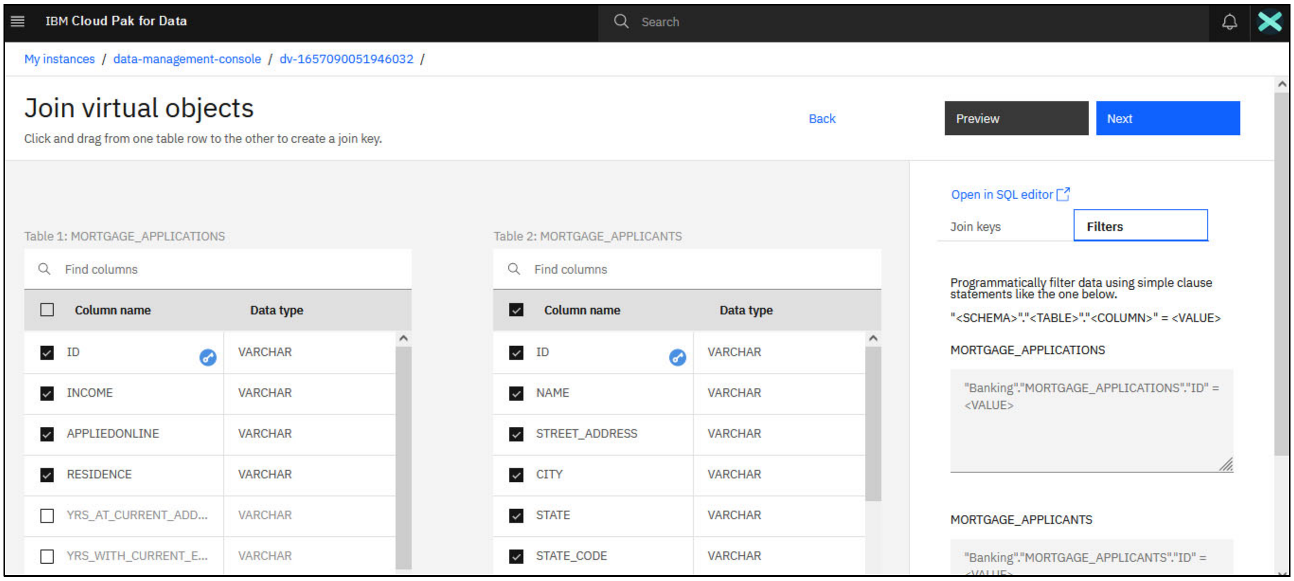

Figure 4-126 Joining virtual objects - Filters tab

In our example, we do not add any filters. We proceed to editing column names in our joined view by clicking **Next** (see Figure 4-127).

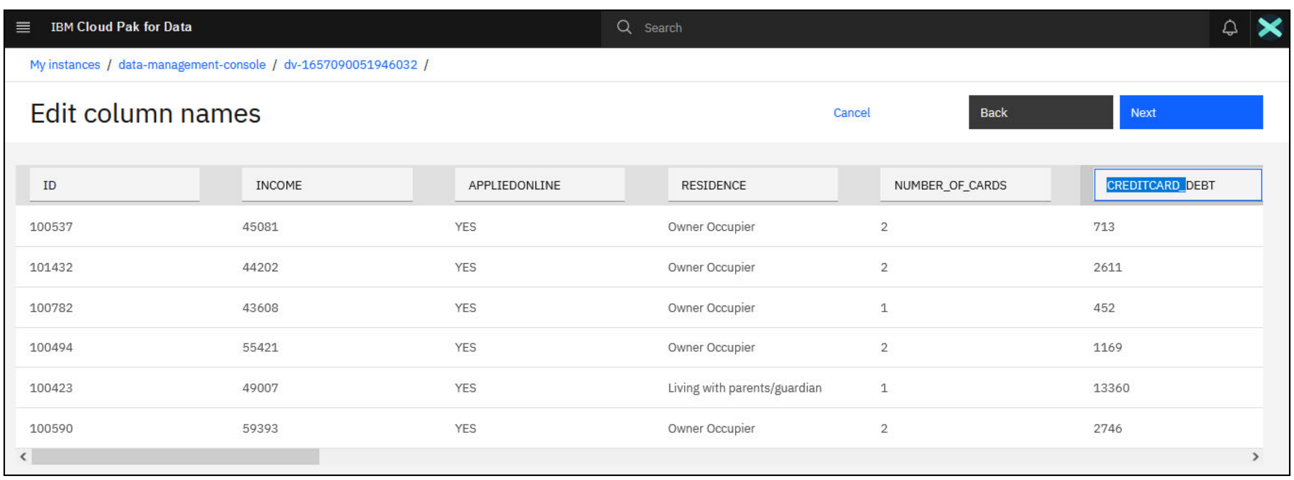

Figure 4-127 Editing column names in the joined view

After all the required edits were completed, clicking **Next** takes us to the final view setup window. As with the virtual objects that we created by virtualizing individual tables from source systems, we can select where to assign the resulting joined view in this window. We also can enter its name and the schema it is placed into within the data virtualization node, and choose whether and where to publish our new object.

Figure 4-128 shows the choices that we made. In our example, we create the view in the Customer schema and do not publish it to a catalog.

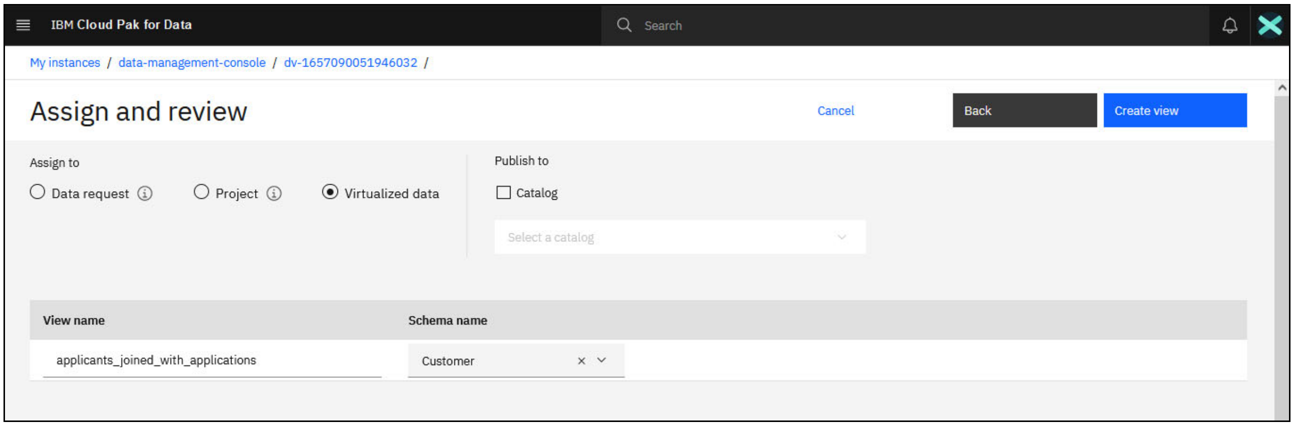

Figure 4-128 Finalizing the setup of the joined view

Clicking **Create view** creates the view based on the choices that we made (see Figure 4-129).

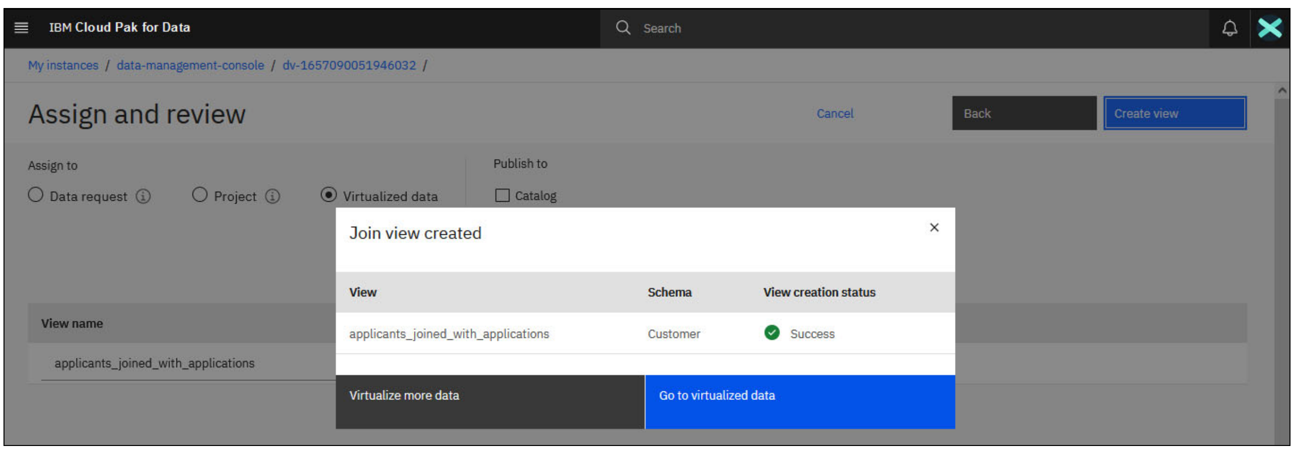

Figure 4-129 Successfully created join view

We can review and preview our newly created joined view to ensure that we successfully joined our loan applicants and loan applications data correctly (see Figure 4-130).

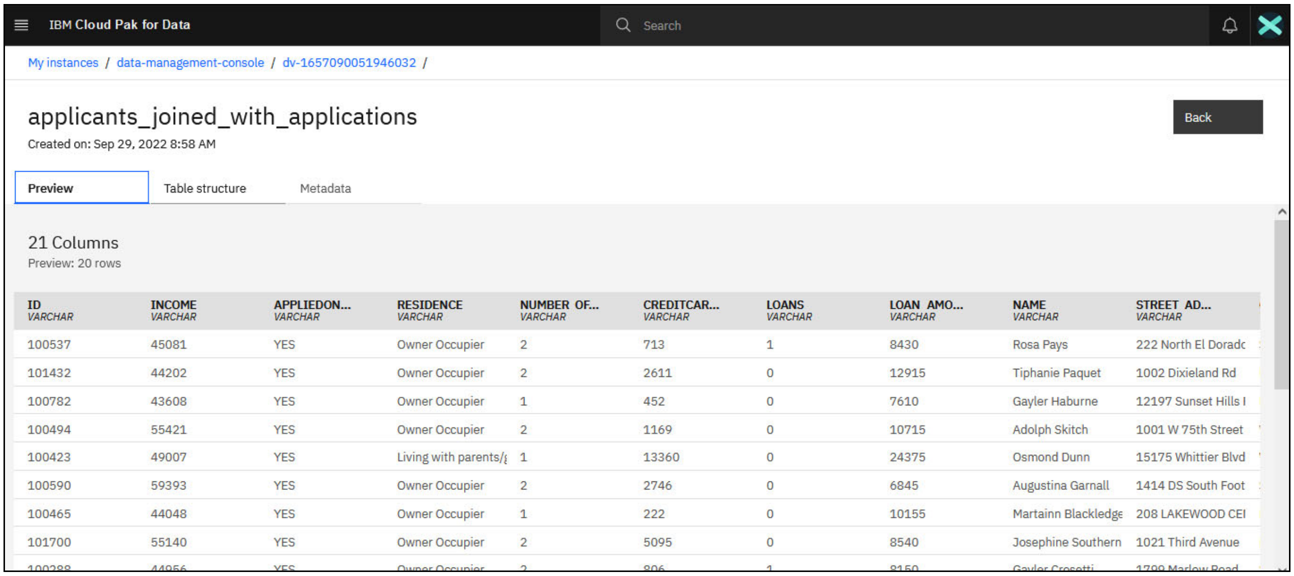

Figure 4-130 Previewing the joined view

Switching to the Metadata tab of the view shows the origin of the view. It captures the fact that the view was created from two source objects, by the user ADMIN, the time and date of its creation, source object names, and other information, including the SQL statement that was used to create the view.

The system automatically converted our selections in the GUI-based editor to an SQL query that was then used to create our resulting view in the background, see Figure 4-131.

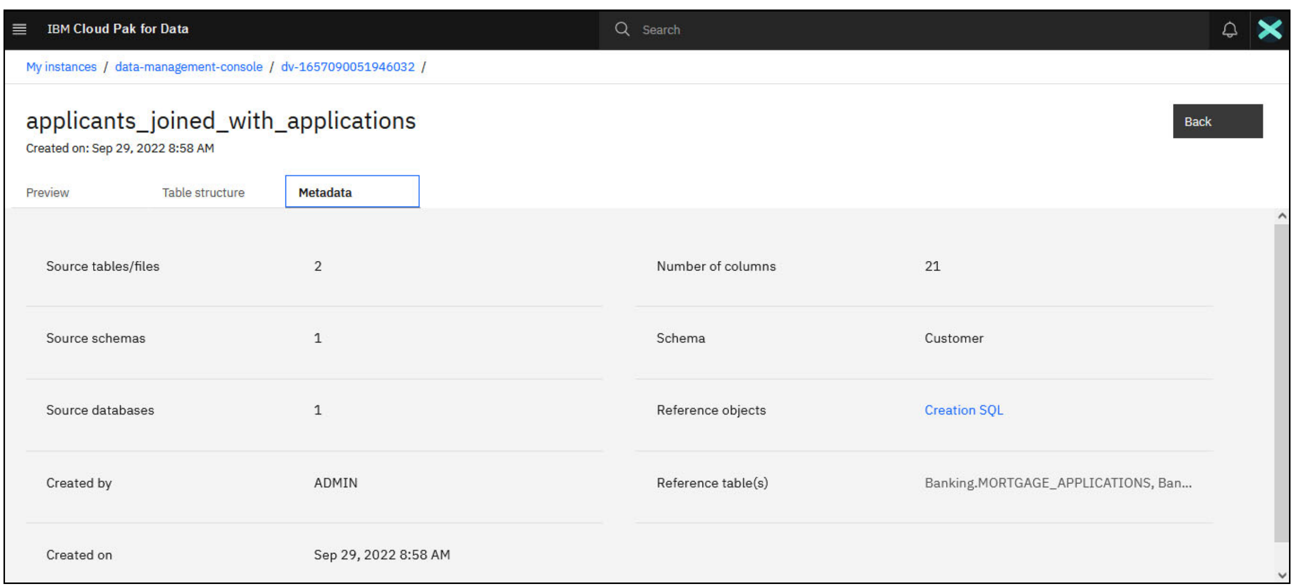

Figure 4-131 Joined view - Metadata tab view

The GUI-based virtual object creation process and the Join graphical editor that is used for the new view creation offer an easy and intuitive way of creating virtual objects that can be used by any approved user, regardless of their SQL skills.

However, IBM Data Virtualization caters for the less technically inclined users and the simple use cases that are covered in this chapter. It also allows users with more advanced data engineering and database administration skills to write and run SQL queries of any required complexity by using ANSI-standard SQL.

On Virtualized data landing page, select the same source objects (`MORTGAGE_APPLICANTS` and `MORTGAGE_APPLICATIONS`) and start the Join editor again. However, after joining the objects by the ID key, we start the SQL editor by clicking **Open in SQL editor** link from the right side window (see Figure 4-132).

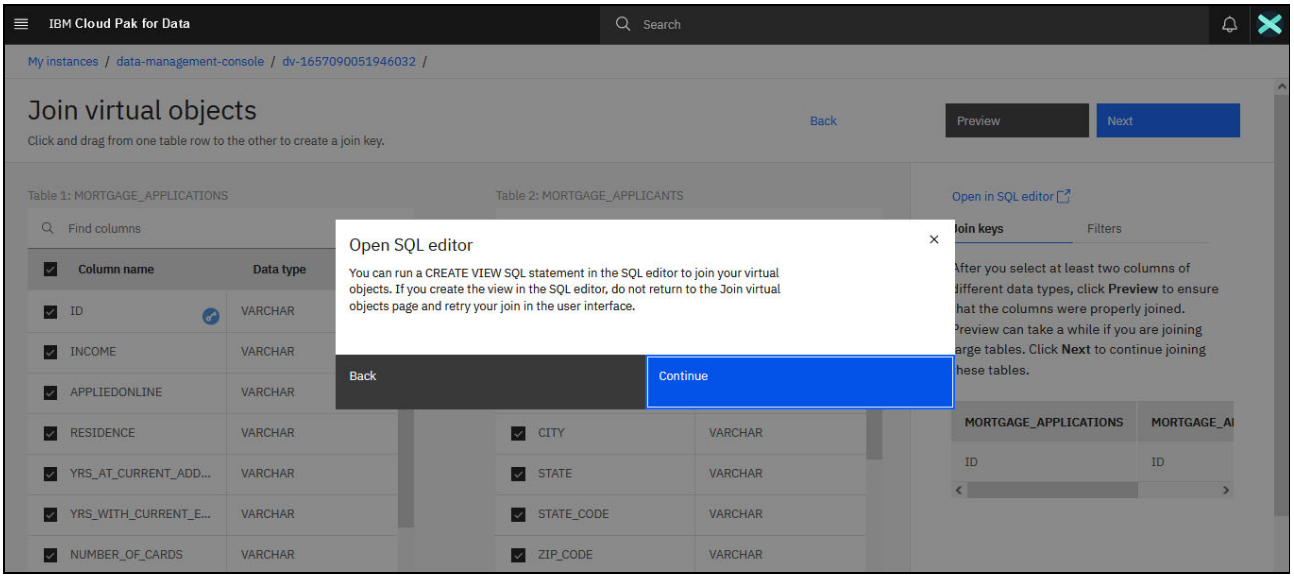

Figure 4-132 Switching from the graphical Join editor to the SQL editor

By using the SQL editor, you can edit the SQL statement that the Join editor automatically generated for us; validate the syntax; use a range of explain, export, and save features; and more.

Figure 4-133 shows the SQL statement that is pre-generated by the Join editor for our join.

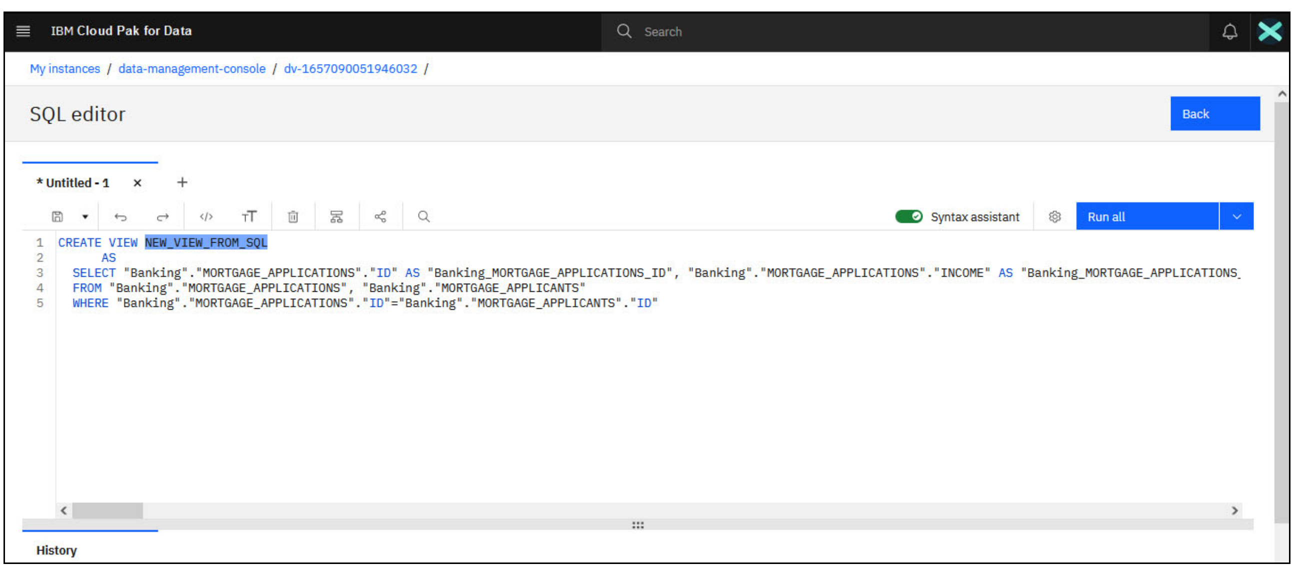

Figure 4-133 Viewing the auto-generated SQL query in the SQL editor

To show the capabilities of the SQL editor, we customize our view by adding conditions to our WHERE statement. We join our applicants and applications data by ID and then, narrow our result set to only the applicants from the state of California that earn more than $50,000 annually.

Figure 4-134 shows the addition of those conditions to the query.

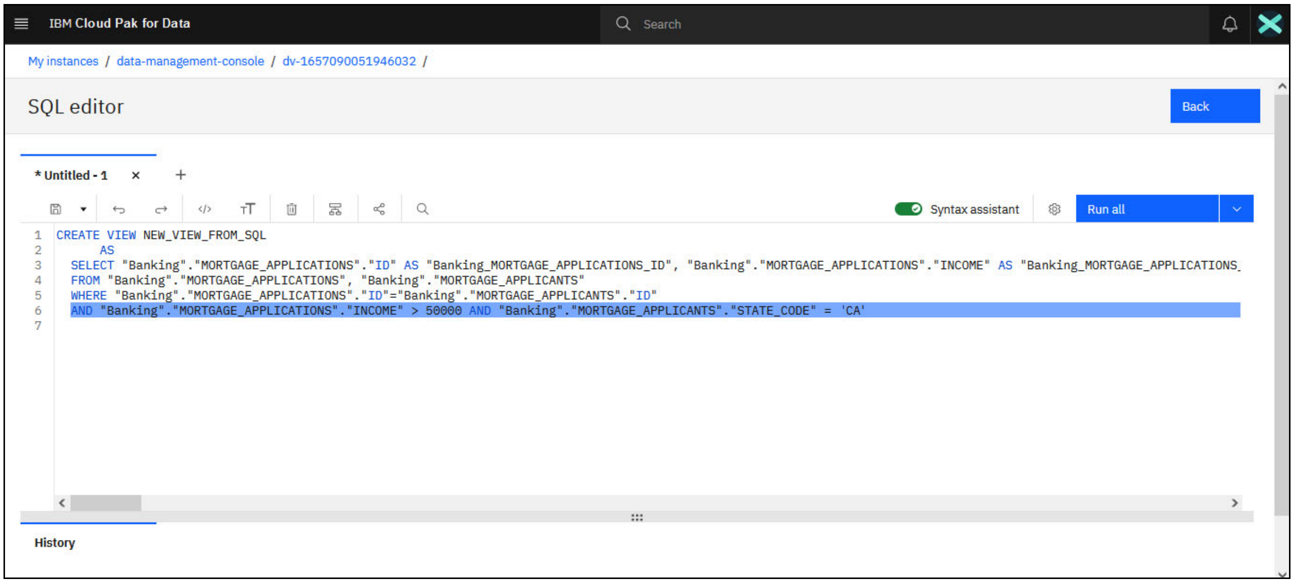

Figure 4-134 Adding conditions to the WHERE statement

Next, we select the entire SQL statement we now have and format the selection. Right-clicking the relevant selected portion of the statement opens a menu that includes the option (see Figure 4-135).

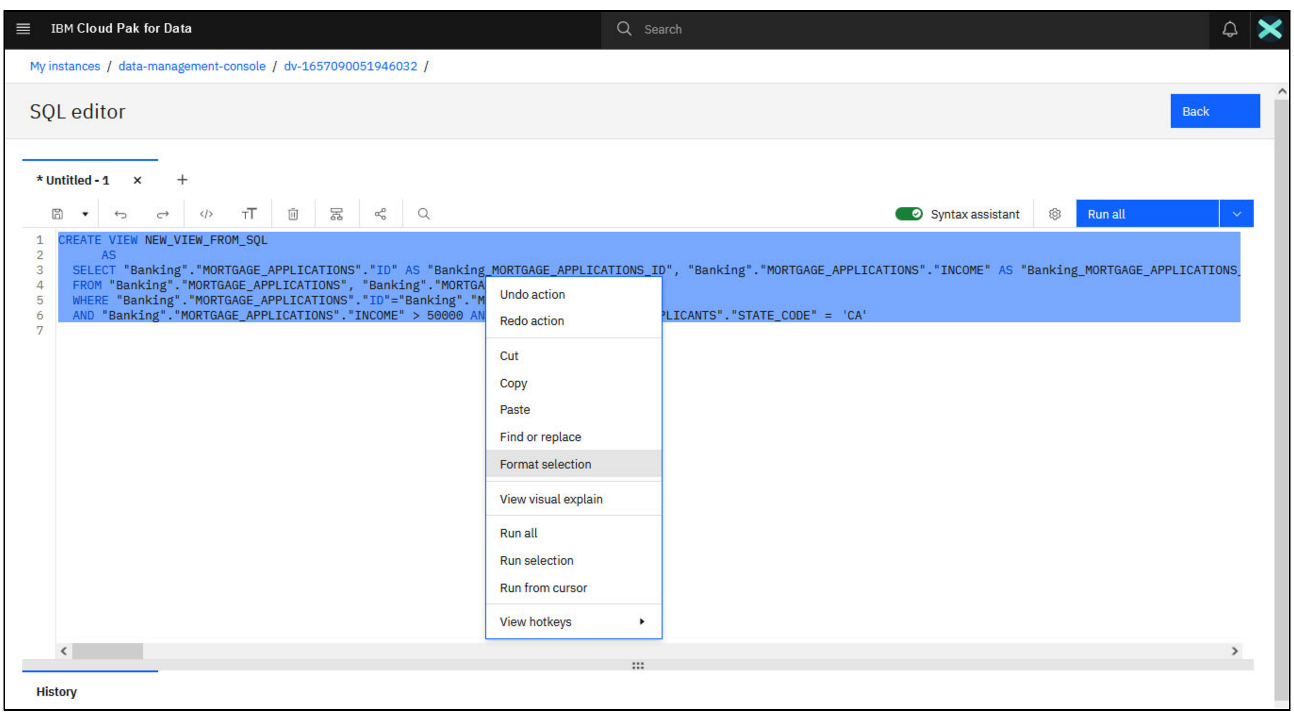

Figure 4-135 Starting the menu for the statement in the SQL editor

Our SQL statement is now formatted to improve readability (see Figure 4-136).

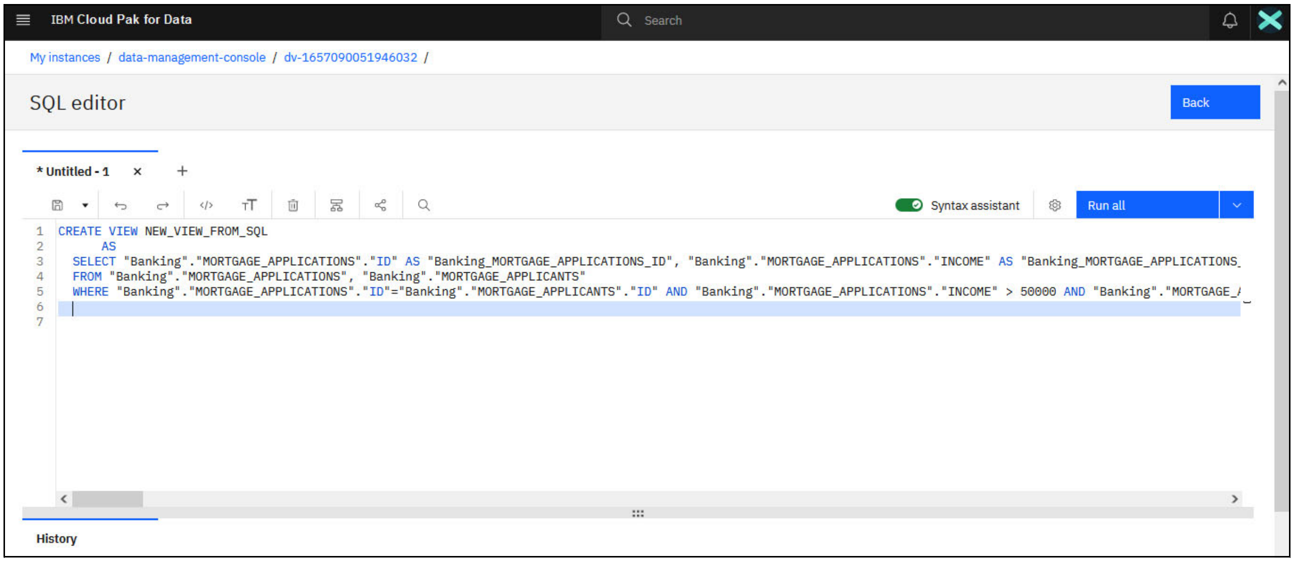

Figure 4-136 Formatted SQL in the SQL editor

Because the Syntax assistant did not identify any errors in our SQL and the code reflects what we are intending to achieve with this new view that is named `NEW_VIEW_FROM SQL`, we can now run the statement to generate the view.

Figure 4-137 shoes the available options in the Run all drop-down menu. In this case, we click **Run all**.

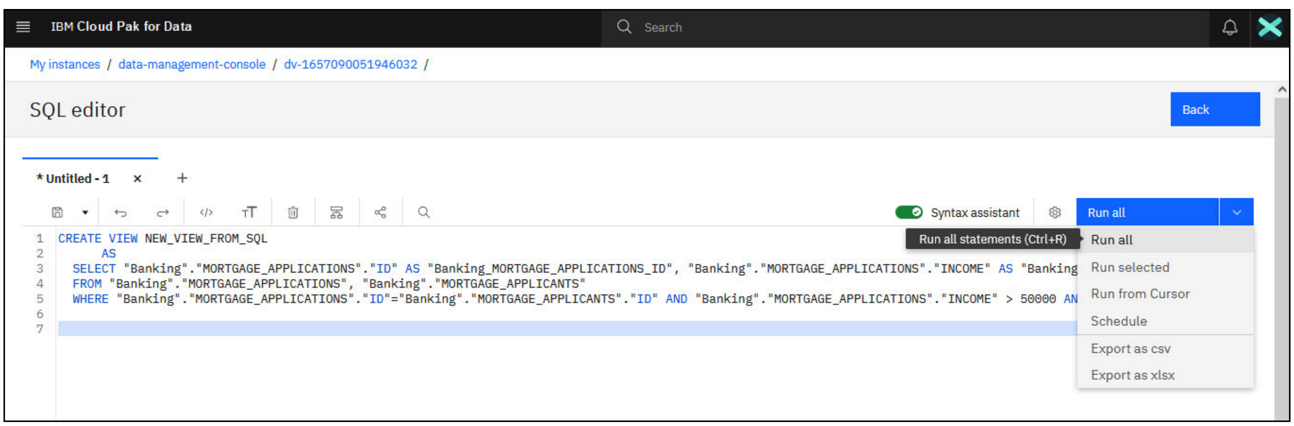

Figure 4-137 Running the statement from the Run all drop-down menu

The system runs the statement and the view is created successfully. Now, we can see the run results, log, and statistics at the bottom of the window (see Figure 4-138).

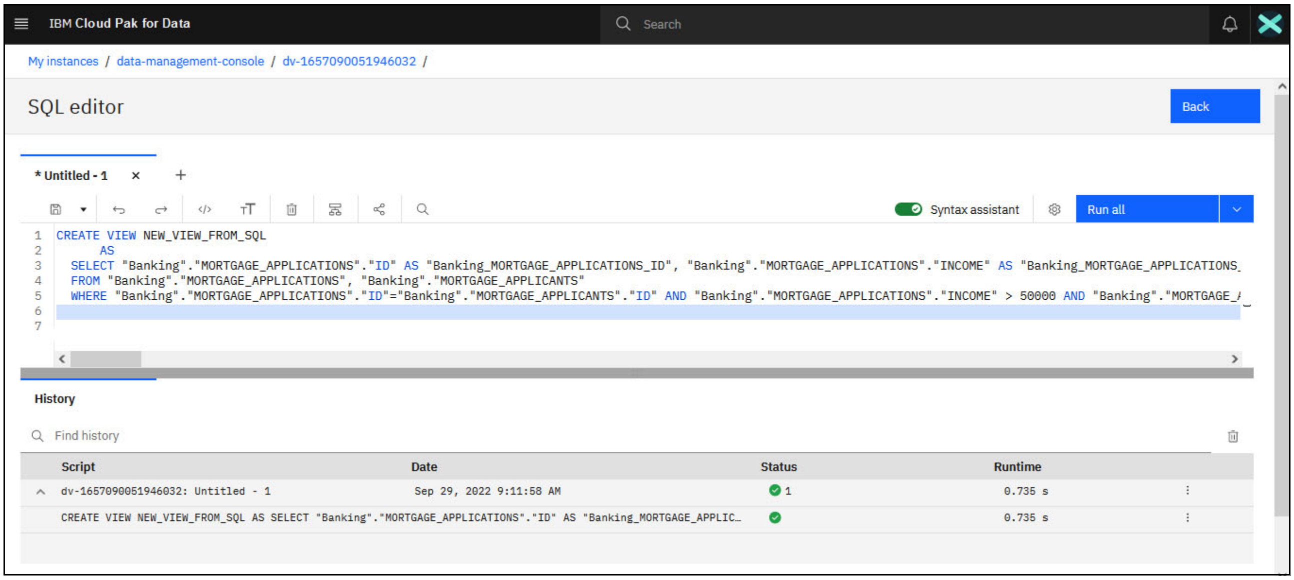

Figure 4-138 Viewing the results of a successful SQL statement execution

Now, we want to save our query for reuse. Figure 4-139 shows how to trigger the Save option for the SQL statement that was composed in the editor.

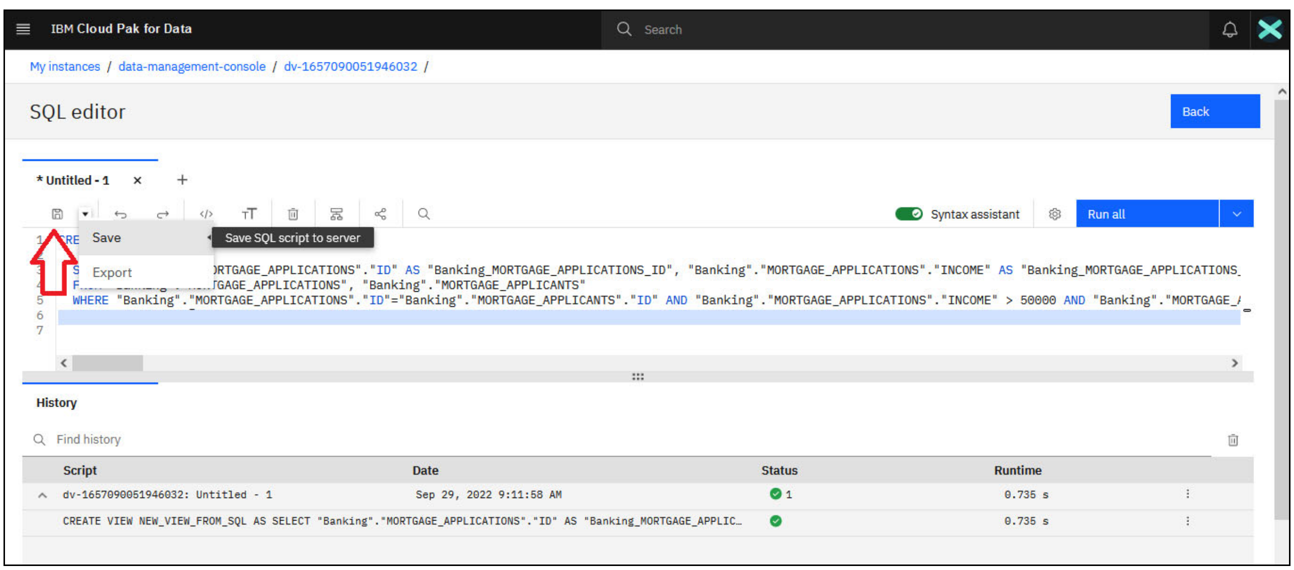

Figure 4-139 Saving the SQL statement as a script

Figure 4-140 shows the final setup step.

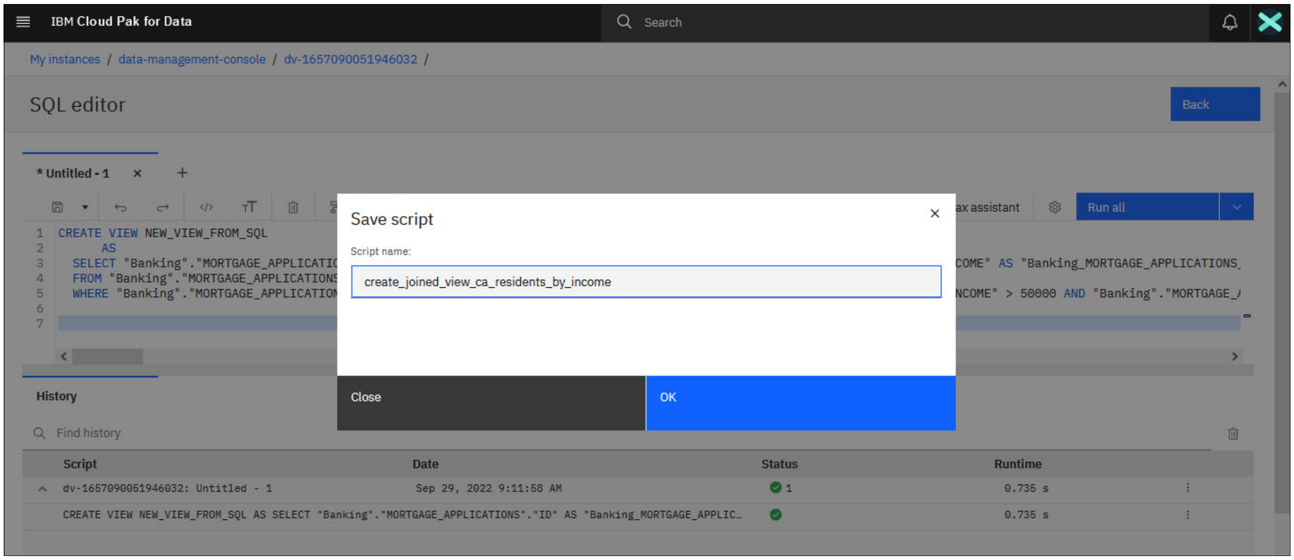

Figure 4-140 Saving the script - final setup step

With the script now saved, we browse to the Run SQL function from the main Data Virtualization drop-down menu and review the in-built Run SQL editor capabilities.

4.3.3 Run SQL interface

The Run SQL editor provides another advanced SQL writing and editing features with which you can create and reuse scripts and script templates, automatically generate SELECT, INSERT, UPDATE, and other statements for your chosen virtual objects, and more.

Figure 4-141 shows the main Run SQL landing page.

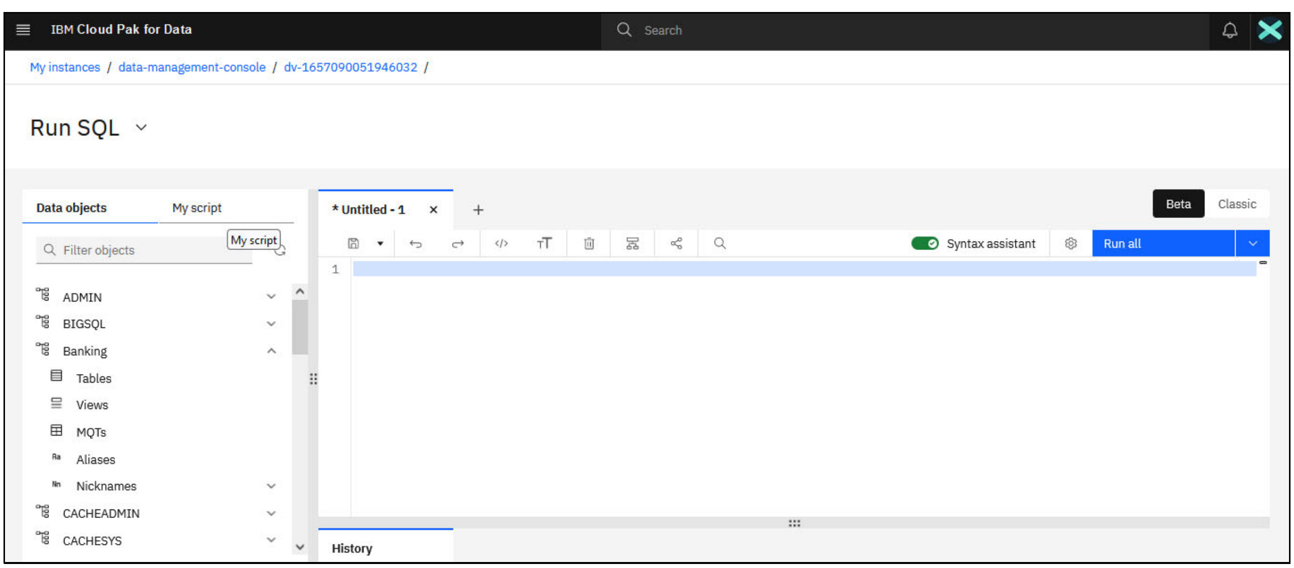

Figure 4-141 Run SQL tools

We browse the Banking schema in our Data Virtualization node and expand the Nicknames list, where all our objects that were virtualized directly from the source systems are now stored. By right-clicking the **INTEREST_RATES** object, a menu opens. From there, we click **Generate DML** → **Select** (see Figure 4-142).

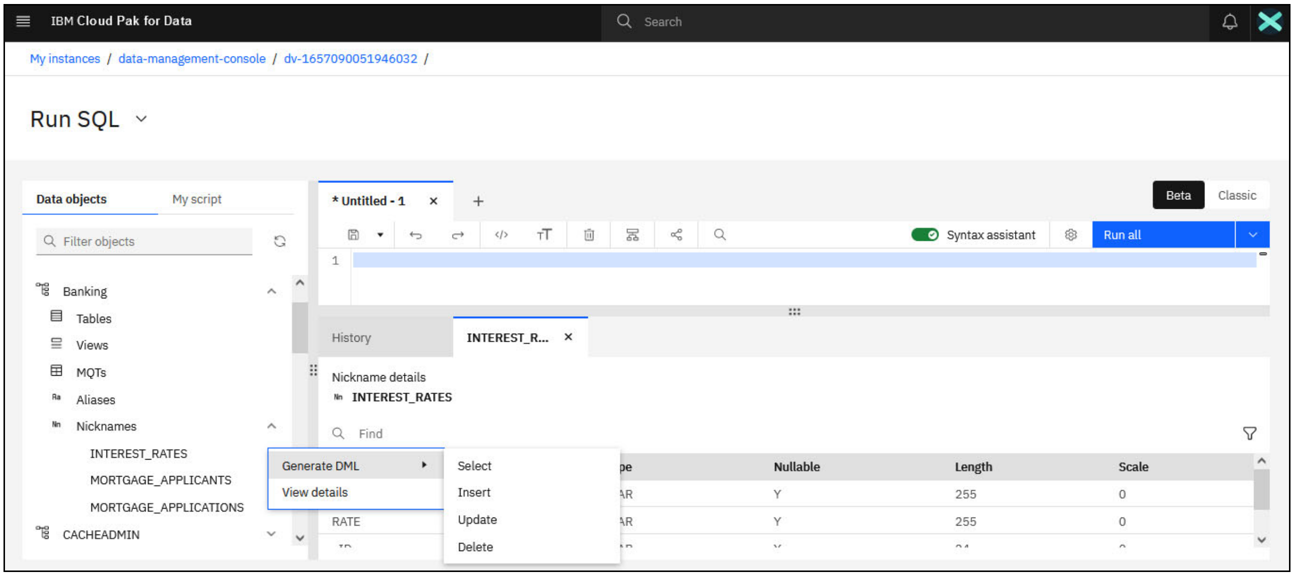

Figure 4-142 Generating a DML Select statement for the selected view

Cloud Pak for Data automatically generates the statement for us. We can use it as-is or as a starting point for a more complex SQL query (see Figure 4-143).

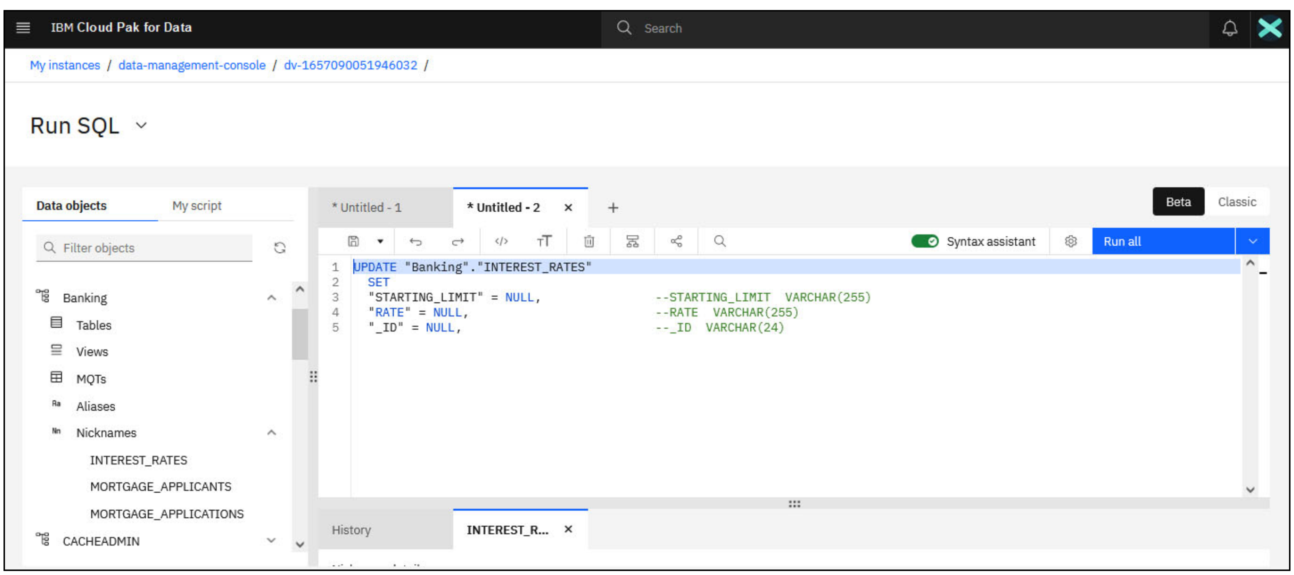

Figure 4-143 DML statement generation results

Next, we switch to the My scripts view, as shown in Figure 4-144.

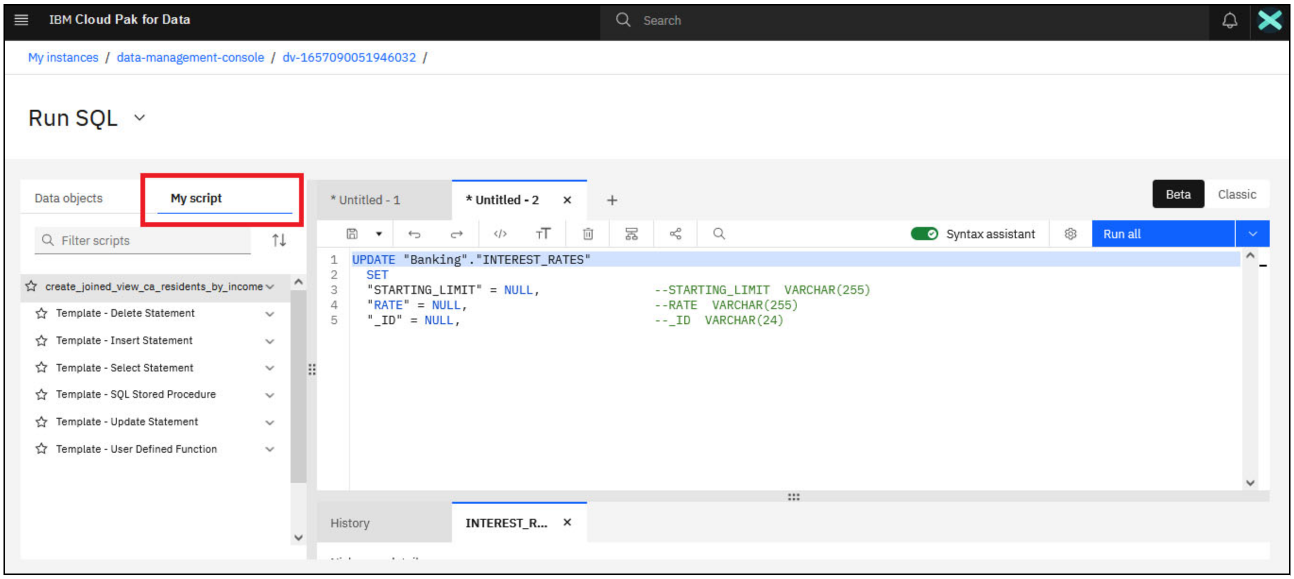

Figure 4-144 My Scripts view

The scripts view lists a selection of templated scripts that can be adapted and reused. Also shown is the custom joined view creation script that we created earlier (see Figure 4-145).

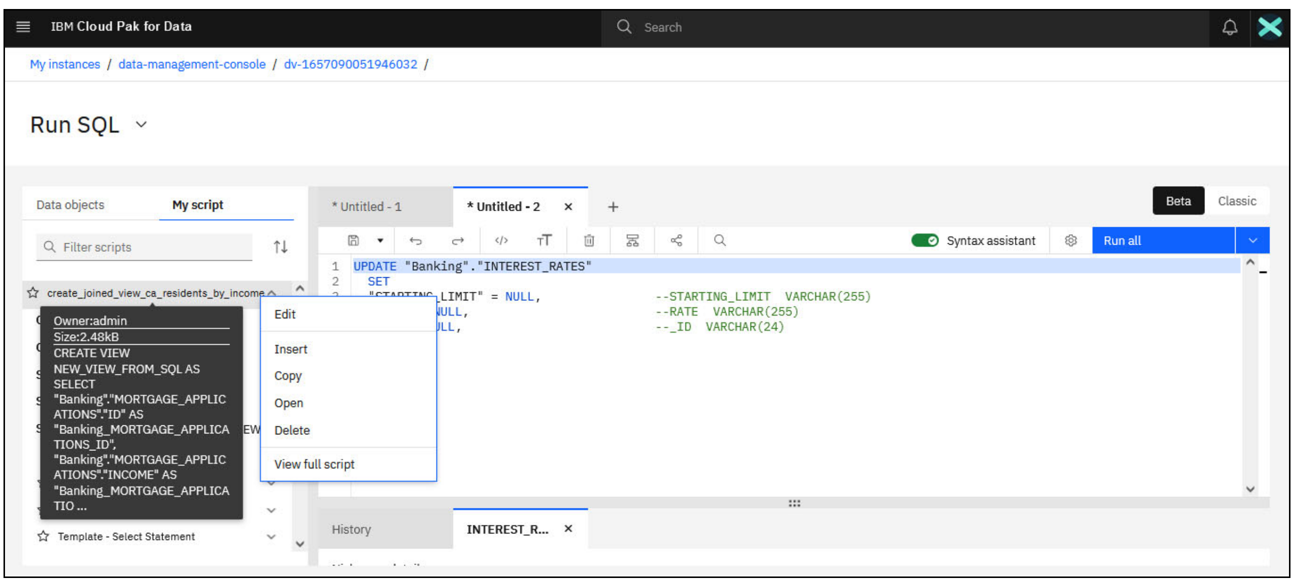

Figure 4-145 Saved custom script menu

The custom script can be copied, edited, viewed, inserted, and deleted as needed.

> **Note:** You are not limited to the in-built SQL Editor and Run SQL tools that are included with IBM Cloud Pak for Data. Equally, you can use your preferred external editor or tool to author and run SQL queries against IBM Data Virtualization instances. In this case, you work with the instances and connect to them in the same way you do so for IBM Db2 databases.

We conclude this section with reviewing the custom view that we created earlier by running the save script.

We switch to the Virtualized data list view from Data Virtualization main drop-down menu. Then, we expand the menu for our view and select **Preview** (see Figure 4-146).

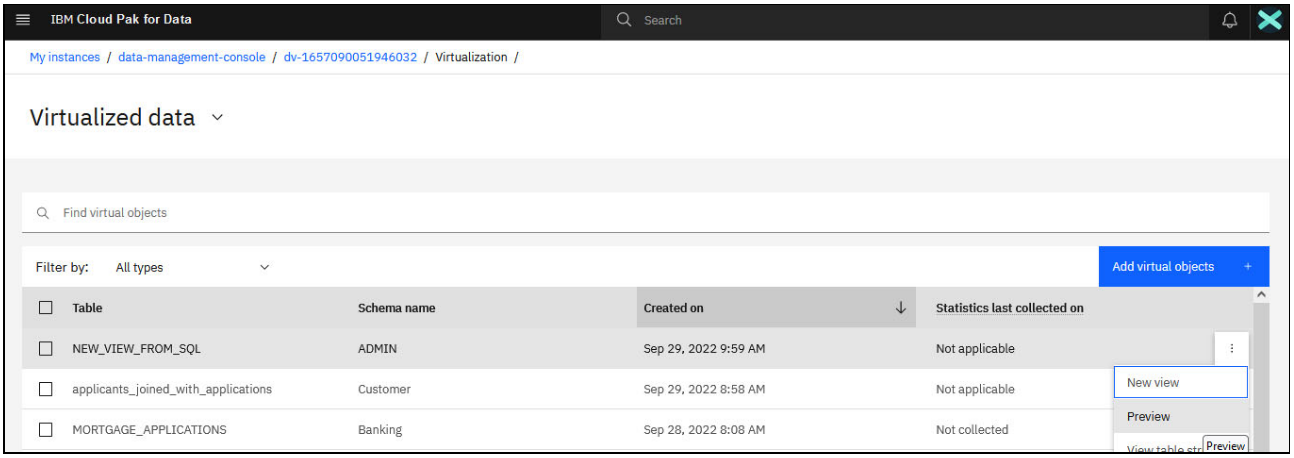

Figure 4-146 Previewing the custom view that is created from the saved script

The preview shows that the view contains the correct data, which is filtered in line with our state and income selections (see Figure 4-147).

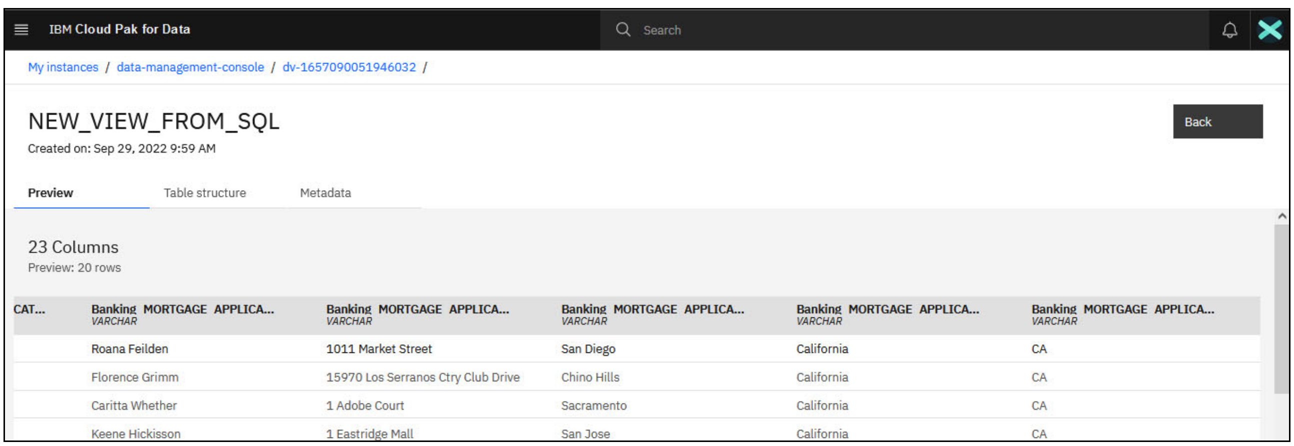

Figure 4-147 Filtered custom view preview

The metadata also correctly reflects view creation specifics (see Figure 4-148).

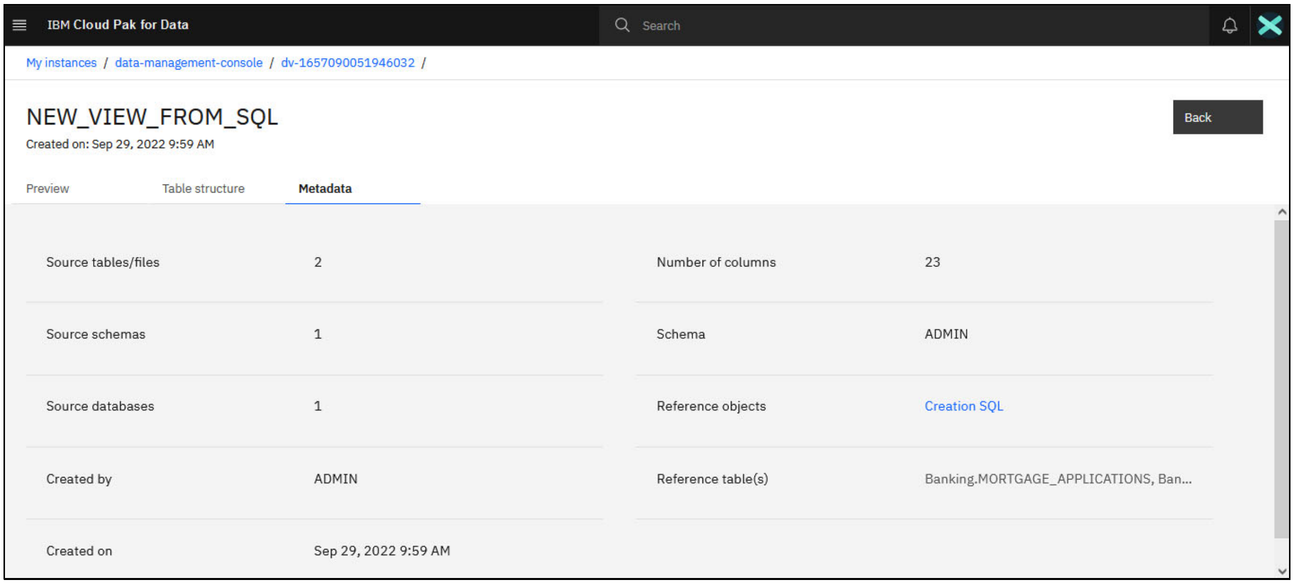

Figure 4-148 Filtered custom view metadata

By clicking **Creation SQL,** the source script that was used to create the view is shown (see Figure 4-149).

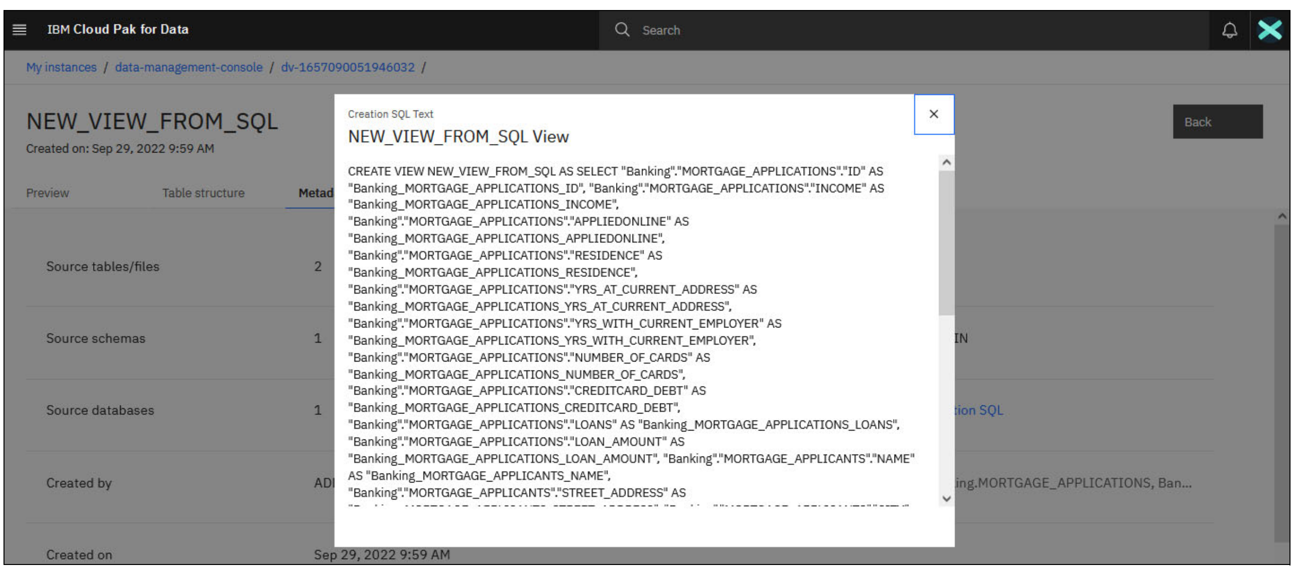

Figure 4-149 Previewing the creation SQL for the custom view

By using the Virtualized data view, you can manually publish the virtual object to a catalog for reuse and browse by approved relevant users (see Figure 4-150).

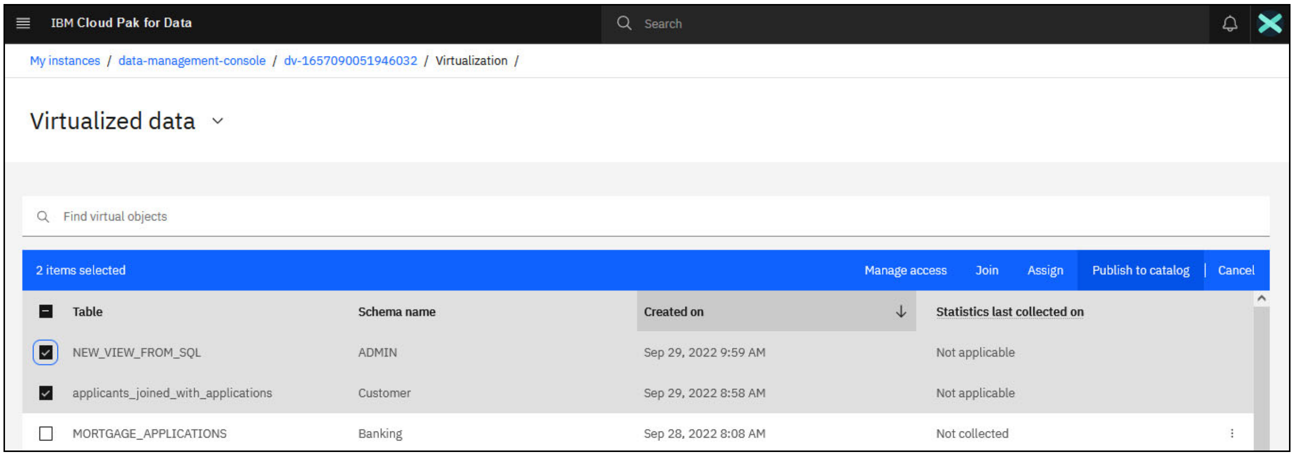

Figure 4-150 Publishing virtual assets to a catalog

4.3.4 Service administration, security, and governance setup

IBM Data Virtualization features its own security model. The platform applies this model on top of the security scope that is defined by the platform role of a user that is controlling general feature-function type access.

Also applied is the security scope for each data source that is connected to the Data Virtualization that is enforced for the user profile that is used when the corresponding Connection to the source is created (that is, the specific username and password or other types of credentials that are used).

The user who provisioned the Data Virtualization instance is automatically assigned the Data Virtualization Admin role and must further grant other relevant users access to it.

Figure 4-151 shows how to browse to the User Management function for your chosen Data Virtualization instance.

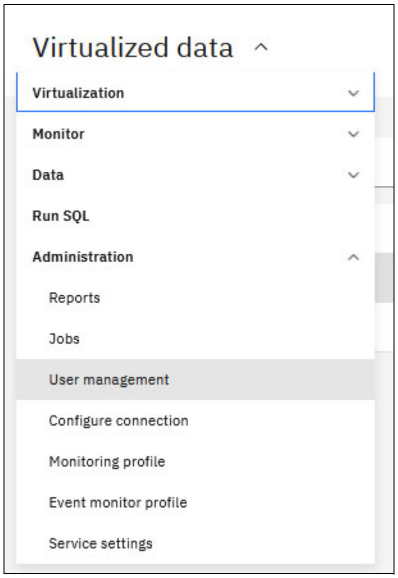

Figure 4-151 Virtualized data

The Data Virtualization service provides the following service-specific roles:

- The Data Virtualization administrator is considered to be the manager of the Data Virtualization instance. They can configure and virtualize data and manage access to virtualized objects. Only users with this role can add other users to the instance and assign suitable Data Virtualization roles to Cloud Pak for Data users or groups.
- Data Virtualization engineers typically configure the data sources, virtualize data, and manage access to virtual objects. Users or groups with this role can create a virtual table or view and grant access to it to users or groups with any Data Virtualization role. Data source administrators are expected to provide access to a user or group with a Data Virtualization Engineer role before that user or group can add a data source. Users or groups with this role service and fulfill data requests from Data Virtualization users.
- Data Virtualization stewards can access data in all user tables and views.
- Data Virtualization users can request access to virtualized data or data in general by initiating a data request. Users with this role also can create views of virtual tables to which they can access.

For more information about the security model and roles setup used within the Data Virtualization service, see this IBM Documentation web page.

Figure 4-152 shows how these roles can be assigned to individual users or user groups that are defined at the platform level.

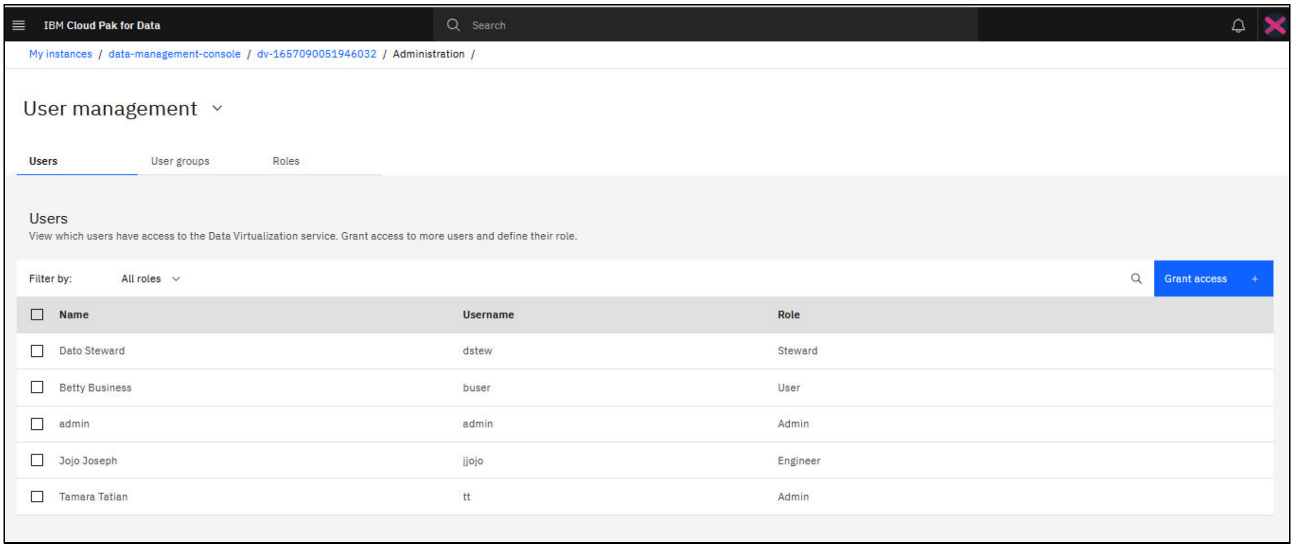

Figure 4-152 Assigning Data Virtualization roles to users and groups

In addition to the roles setup, other security setup steps are performed for each virtual object.

By default, every virtual object that is created in Data Virtualization is private. This privacy means that for a virtual object to be accessed by a user or group other than its creator, access to the virtual object must be granted.

> **Note:** As part of the overall instance setup, you can determine whether Data Virtualization users can view virtual objects that they cannot access. The Enable/Disable the Restrict visibility option in the **Service settings** → **Advanced options** section of the main drop-down menu of your Data Virtualization instance controls this behavior:
>
> - If the Restrict visibility option is Enabled, the Disable feature is visible in the UI, and users are prevented from seeing column and table names of virtual objects that they cannot access. This option is enabled by default.
> - If the option is Disabled, the Enable feature is visible, and users can see column and table names of virtual objects that they cannot access. Also, when you clear the Restrict visibility setting, the Virtualized data page becomes All virtualized data.

Figure 4-153 shows the Restrict Visibility setting of the instance we use in this chapter (it was set to `Enabled`).

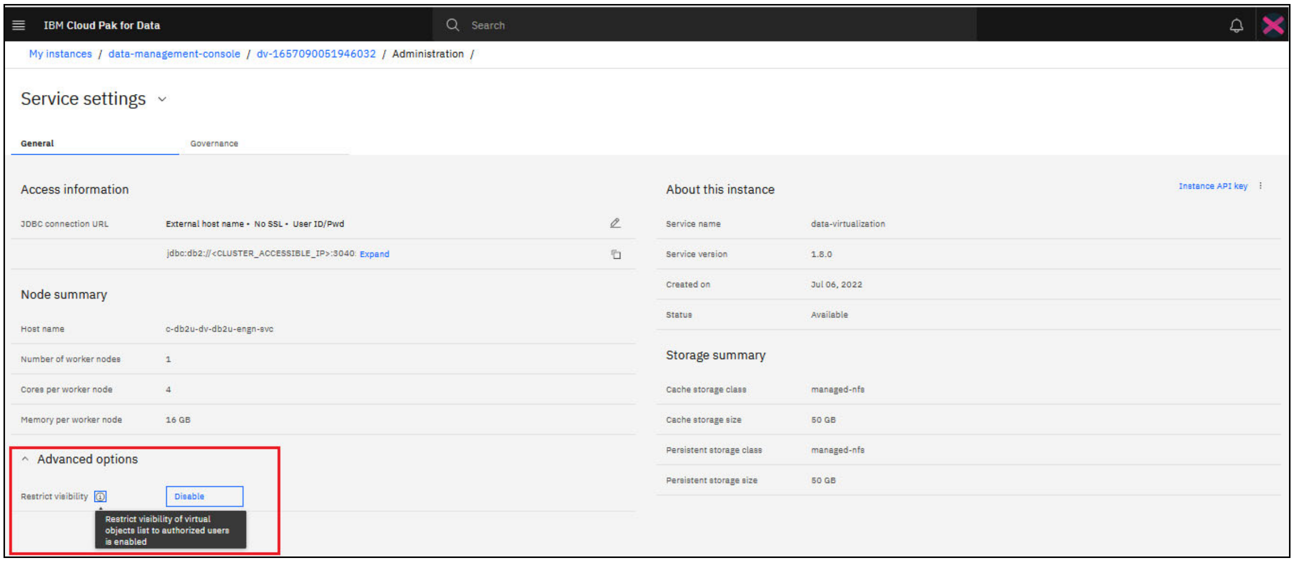

Figure 4-153 Controlling virtual asset visibility - system-wide settings

Controlling object access

In this section, we describe how we can control access to the objects we created earlier in the chapter.

In the Virtualized data view, we select one or more objects and then, click **Manage access** from the menu. We can now grant relevant access to the selected objects to relevant users and groups, as shown on Figure 4-154.

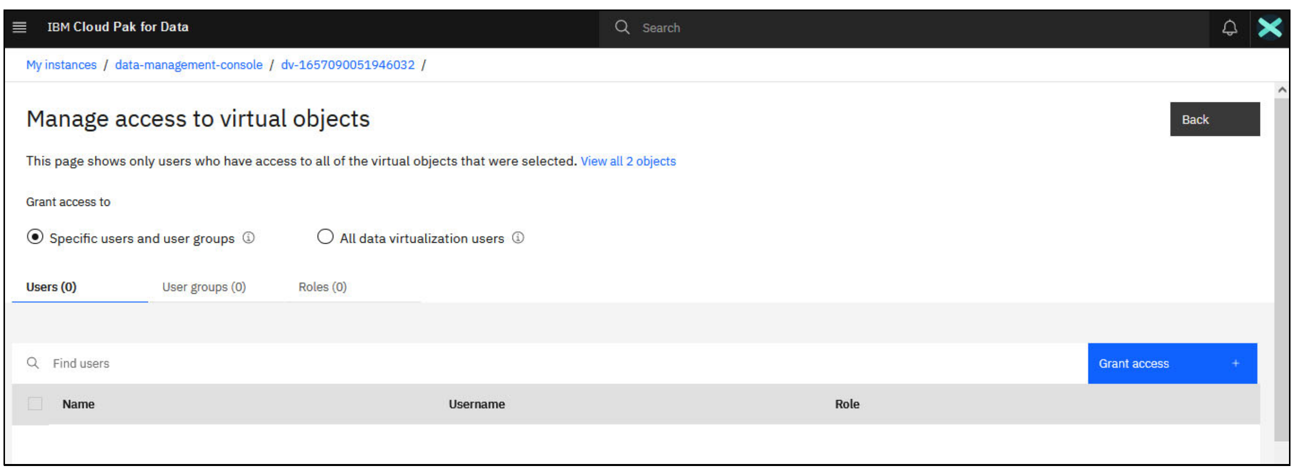

Figure 4-154 Granting access to objects to relevant users

Access can be granted to all data virtualization users, or to specific users and user groups only.

Figure 4-155 shows how individual users can be chosen and access to the objects granted to them.

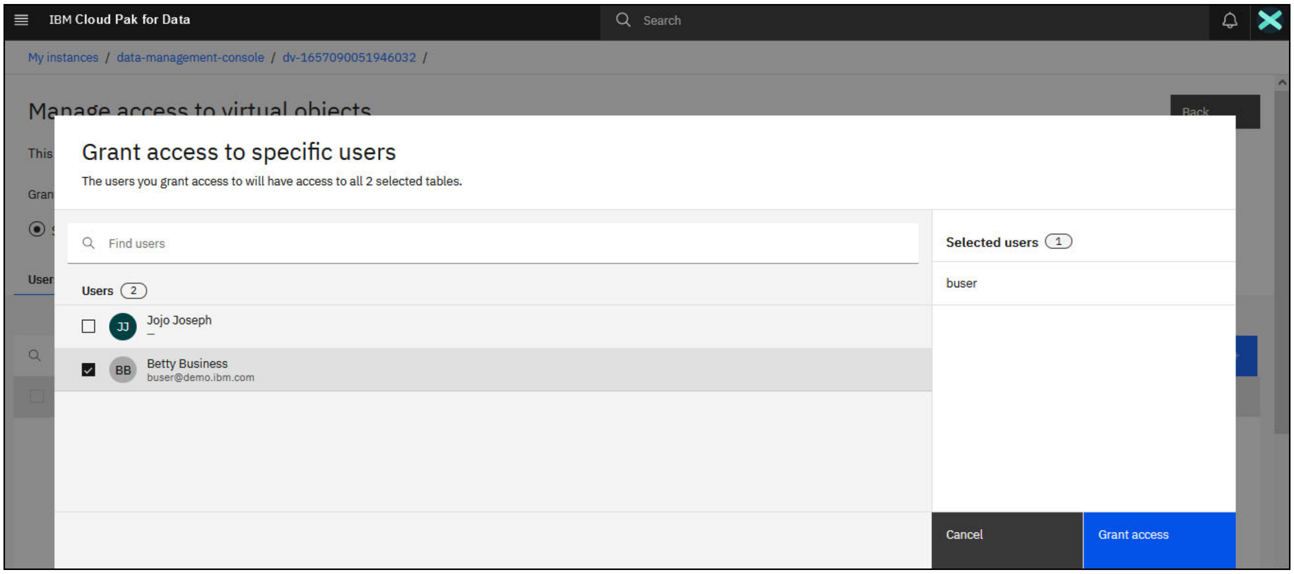

Figure 4-155 Granting access to the objects to individual users

Figure 4-156 shows the alternative; that is, granting access to the objects to all users.

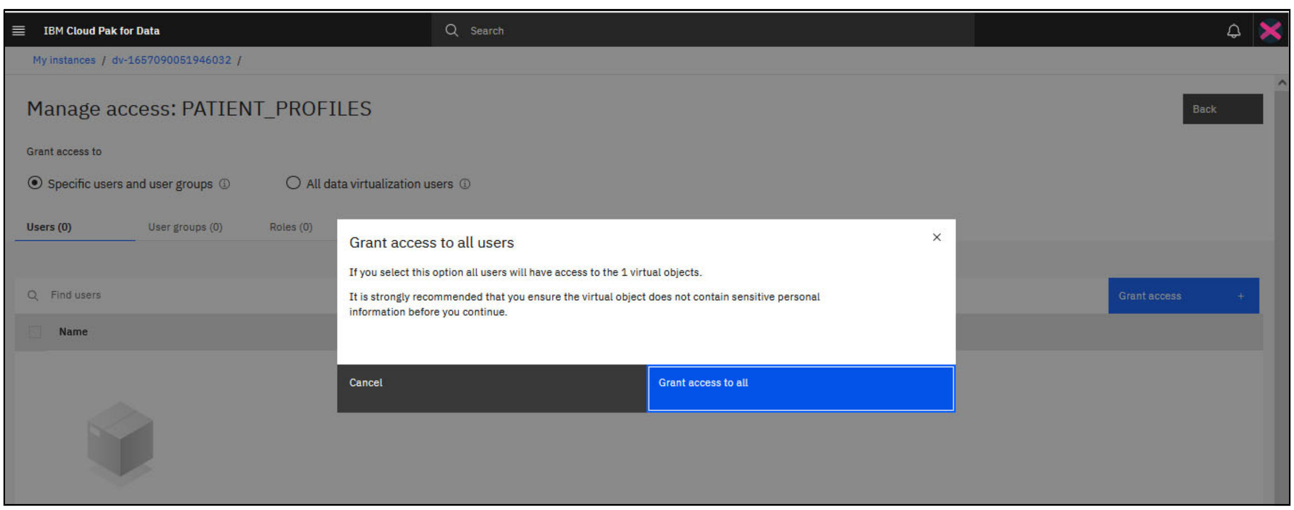

Figure 4-156 Granting access to the objects to all users

For more information about object access management, see this IBM Documentation web page.

Another important choice that must be considered when setting up the service is the Governance enforcement (click **Service settings** → **Advanced Settings** → **Governance tab**).

If the IBM Watson Knowledge Catalog service and the IBM Data Virtualization service of IBM Cloud Pak for Data are deployed on the same cluster, you can take advantage of the advanced features that the integration between the two services adds to your Data Virtualization instance.

Cloud Pak for Data can enforce masking, row filtering, and denial of access (data protection) rules that are defined in Watson Knowledge Catalog in the objects and views that are defined in the virtualization layer. This enforcement applies within the Cloud Pak for Data context and externally in situations where an external application or service is querying the virtualization layer. Cloud Pak for Data can consistently enforce those rules in-flight.

Figure 4-157 shows the governance settings of the service.

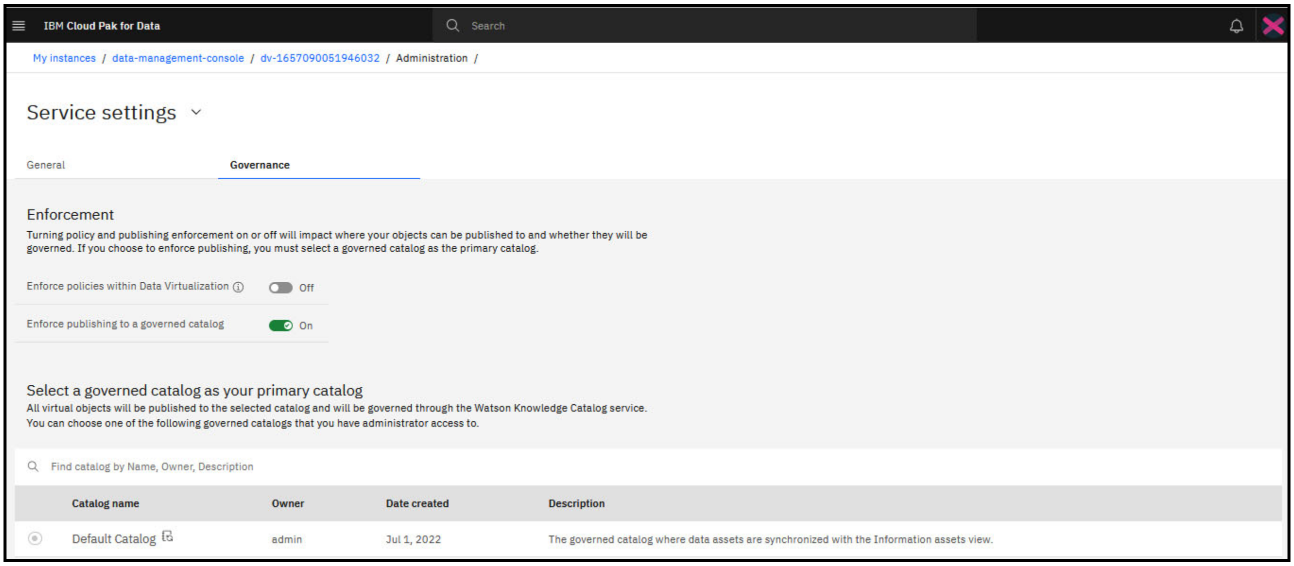

Figure 4-157 Governance setup in Data Virtualization

Figure 4-158 shows the results of governance enforcement when the Enforce policies system setting is enabled. An example from the Healthcare industry is used to show data protection rule enforcement for patients' date of birth and email details in a virtualized object.

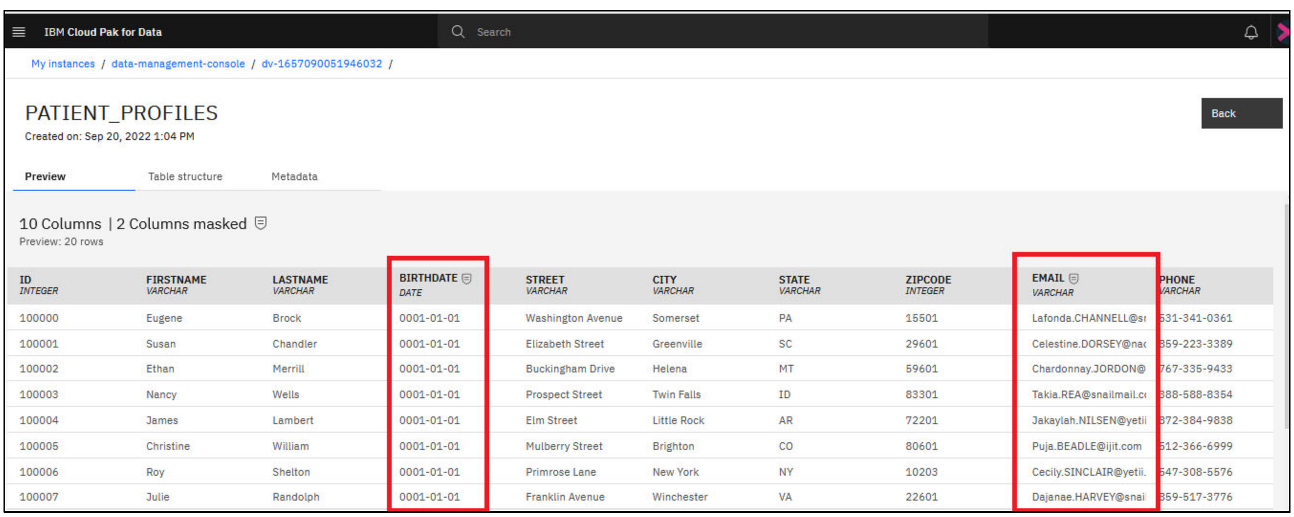

Figure 4-158 Governance policies enforcement in a virtualized view

For more information, see this IBM Documentation web page.

Next, we briefly review some of the other monitoring, administration, and management capabilities of the service.

Figure 4-159 shows all of the entries that are available in the main drop-down menu of your Data Virtualization instance.

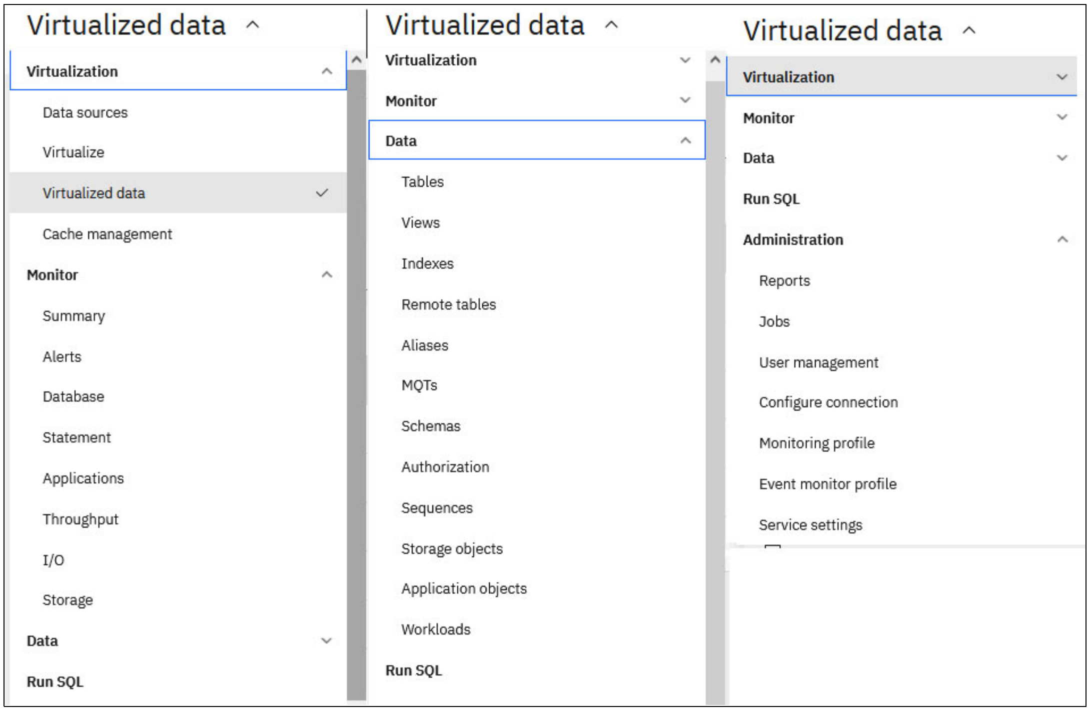

Figure 4-159 Expanded view of Data Virtualization main menu

Thus far, we briefly reviewed the User Management and Service settings features from the Administration section of the menu.

Figure 4-160 shows the Monitoring summary window example for our instance (click **Monitor** → **Summary** in the menu).

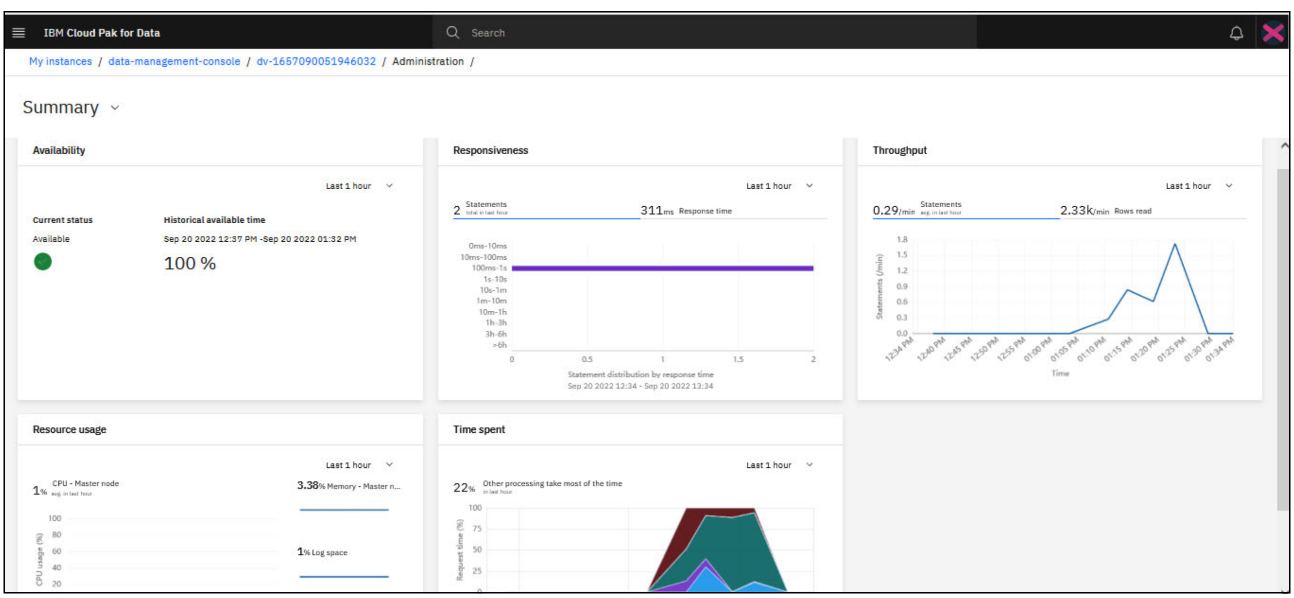

Figure 4-160 Virtualization instance monitoring summary

Figure 4-161 shows the **Monitor** → **Database** view.

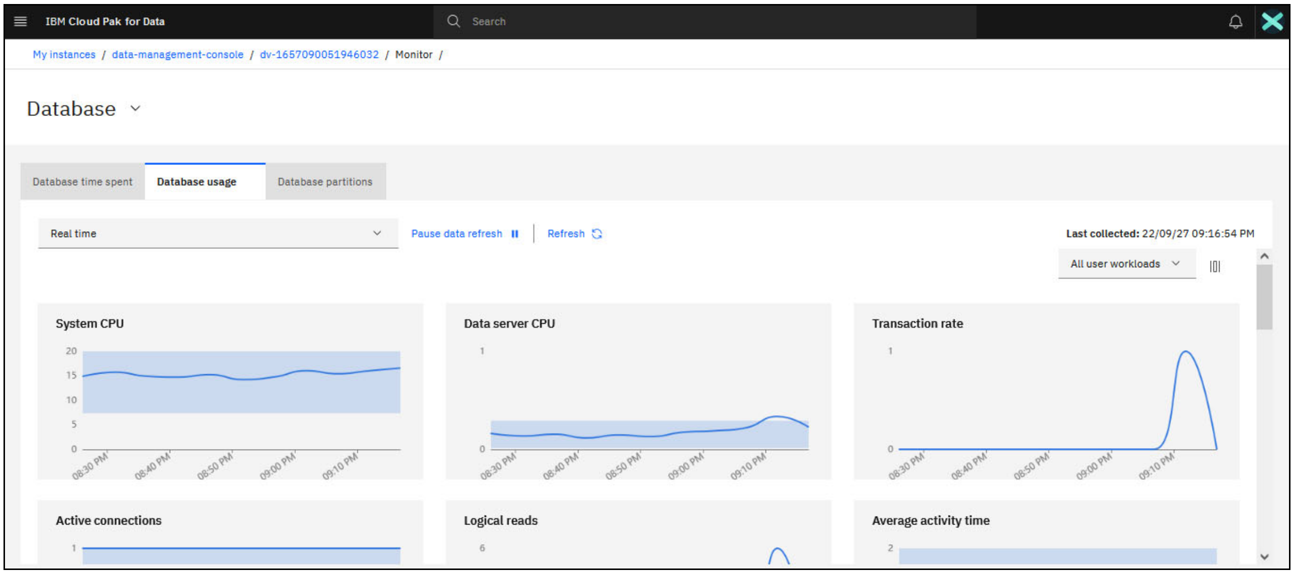

Figure 4-161   Database monitoring in Data Virtualization

These examples are only some of the examples of tools and dashboards that are provided with IBM Data Virtualization with which you can monitor and manage its usage, availability, and performance.

The caching capabilities of the solution are described next.

4.3.5 Working with caches

By default, queries that are issued against Data Virtualization are run “live”; that is, the query is submitted to the constellation and the computational mesh breaks down and routes the query to the specific source data sources.

The relevant data is then pulled from those sources “live”, the necessary extra calculations and analysis are performed by the constellation and finalized by the coordinator node of the service, which returns the results back to the query submitter.

Although this approach always ensures that the freshest data is pulled from the sources, it also can affect system performance and query execution time.

To help address this issue, Data Virtualization administrators can create a cache entry to save query data, and results and optimize query performance. Examples of use cases where this approach might prove beneficial include complex and long-running queries, queries that are issued against systems with a slow rate of changes or a known change frequency, and when a need exists to minimize the effect on the underlying source systems for operational or other reasons.

After a cache is created for a specific query, the coordinator node uses the most recent version of the cache to run the query instead of reaching directly to the source systems.

Cache management capabilities are accessed from the main menu, and are part of the Virtualization section.

Figure 4-162 shows the cache management landing page.

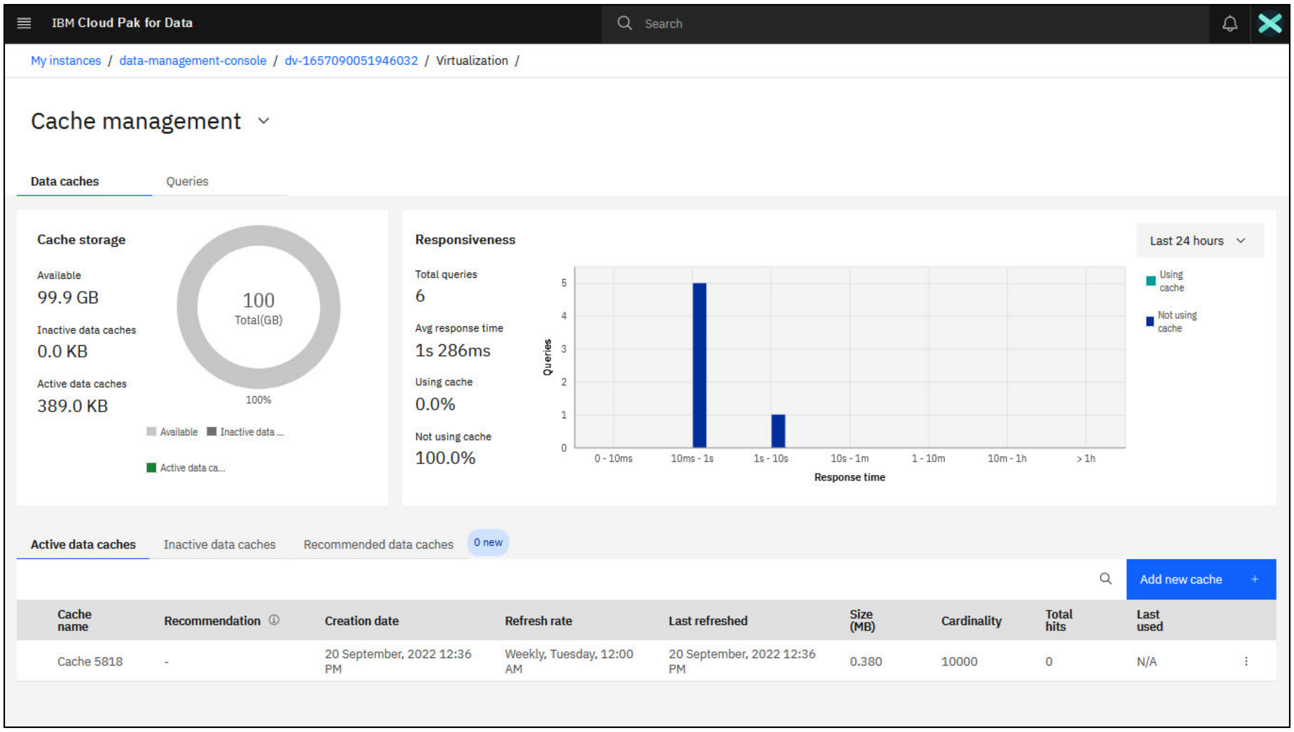

Figure 4-162 Cache management landing page

Data Virtualization tracks all the queries that are run. Those queries can be accessed on the Queries tab (see Figure 4-163). In this example, we select the query that took the longest to run and create a cache that is based on it.

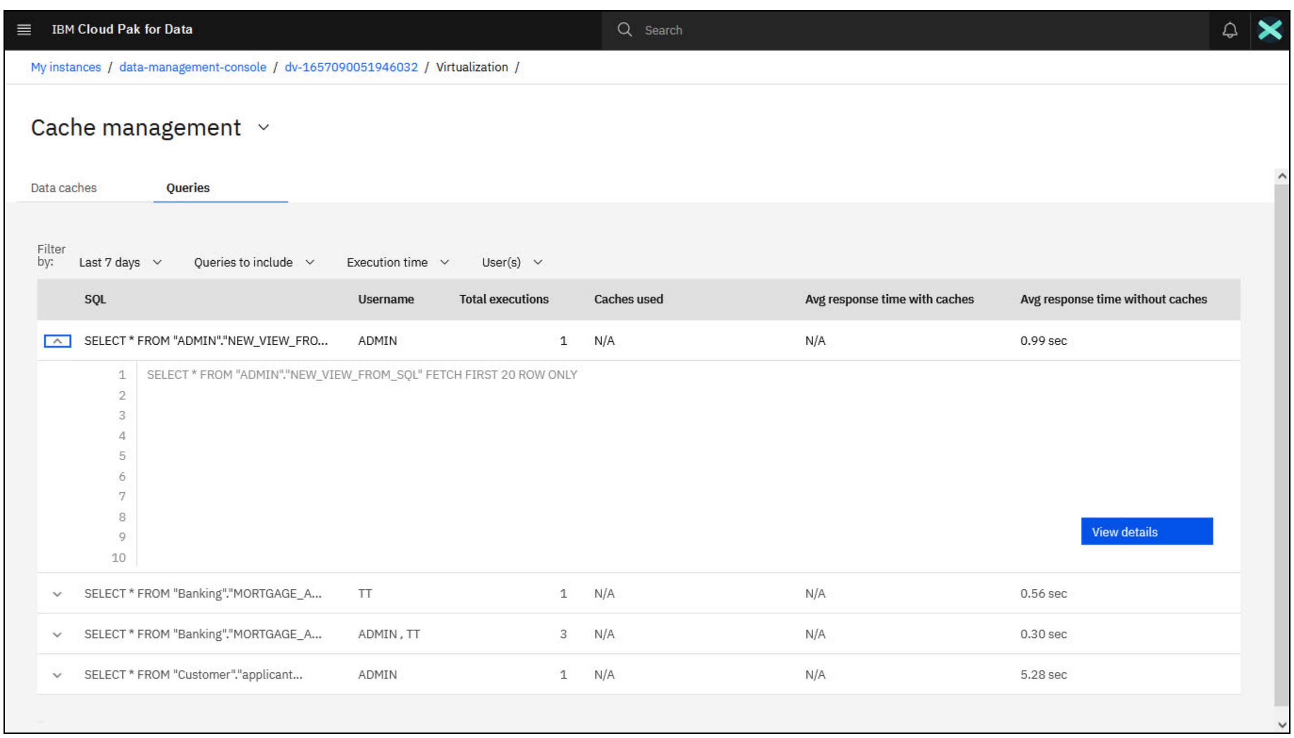

Figure 4-163 Viewing cache creation candidate queries

The query can be used as-is, or amended as and where needed. In this case, we rewrite the original query to make it a bit more complex (see Figure 4-164).

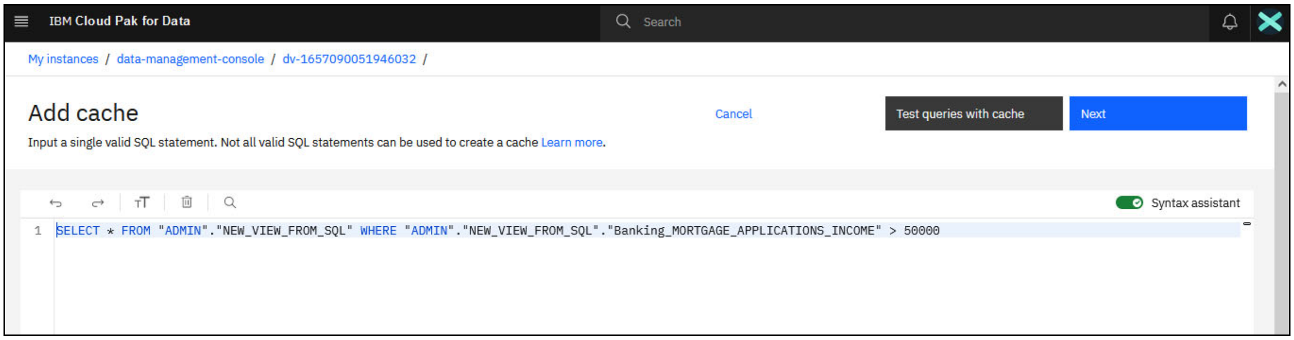

Figure 4-164 Editing the query that is used in the cache setup

Then, we can test what difference this new cache might make to the queries execution and select which queries we want the system to use this cache for going forward (see Figure 4-165).

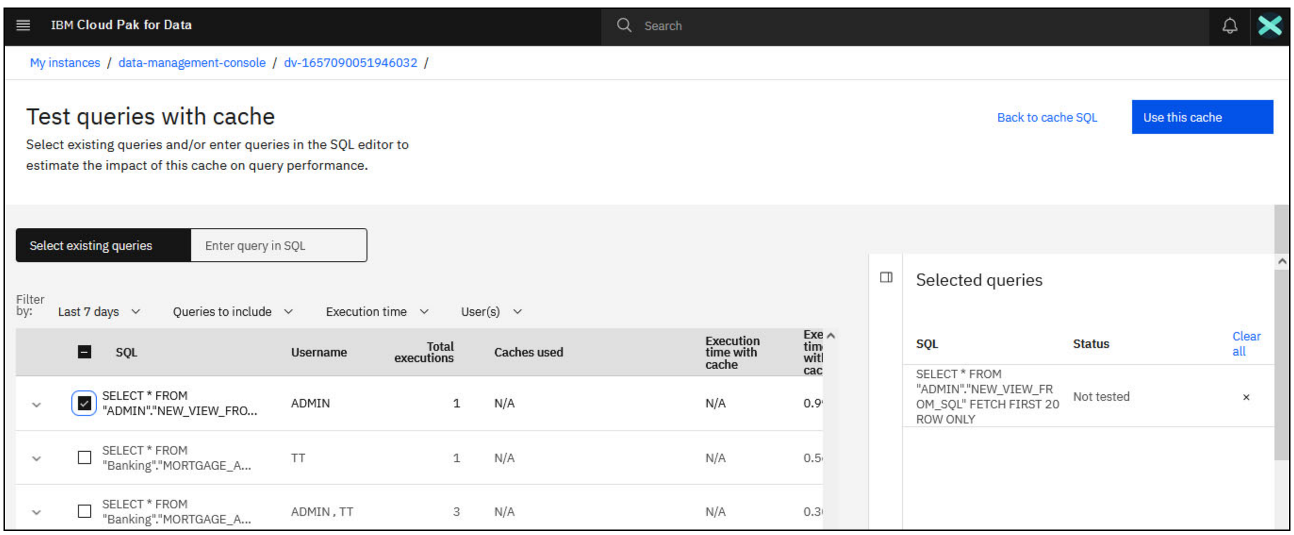

Figure 4-165 Testing queries with cache

The cache refresh rate schedule is set up next (see Figure 4-166).

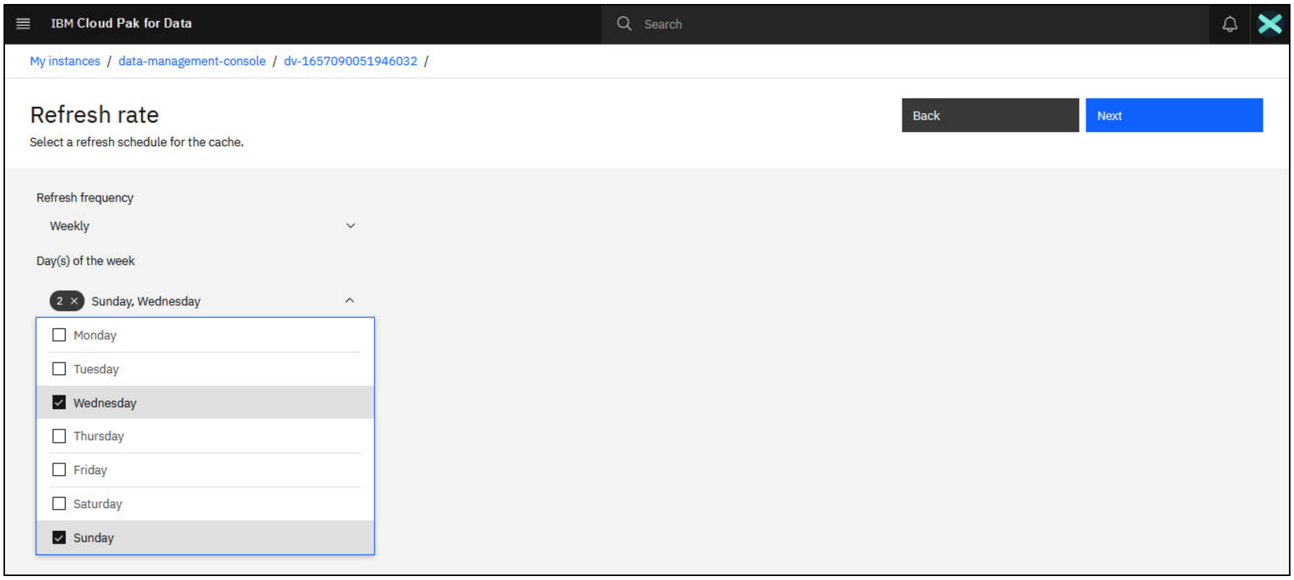

Figure 4-166 Cache refresh rate schedule setup

In this case, the cache must be refreshed twice a week (on Wednesdays and Sundays), as shown in Figure 4-167.

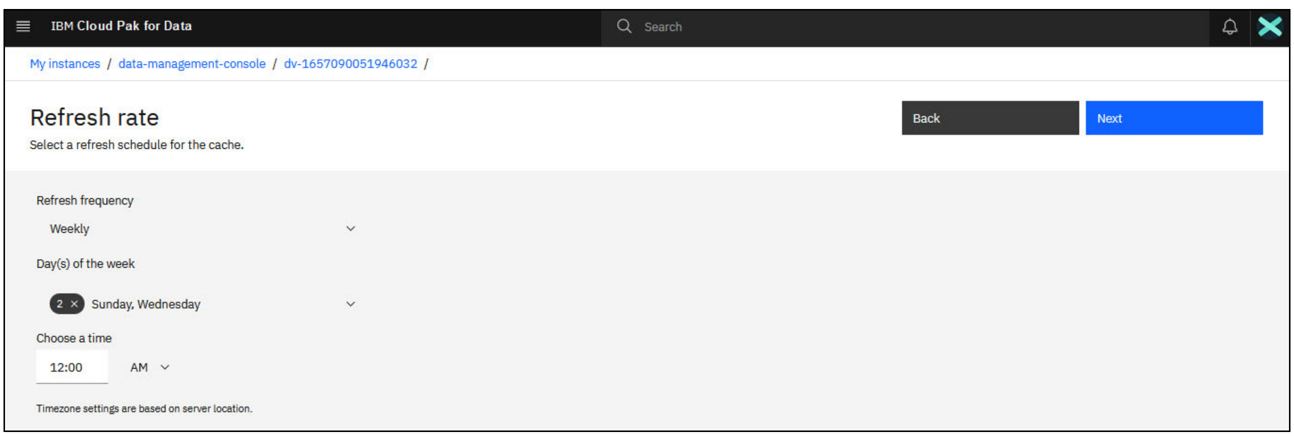

Figure 4-167 Bi-weekly refresh schedule setup

Then, review and finalize the cache setup, as shown on Figure 4-168.

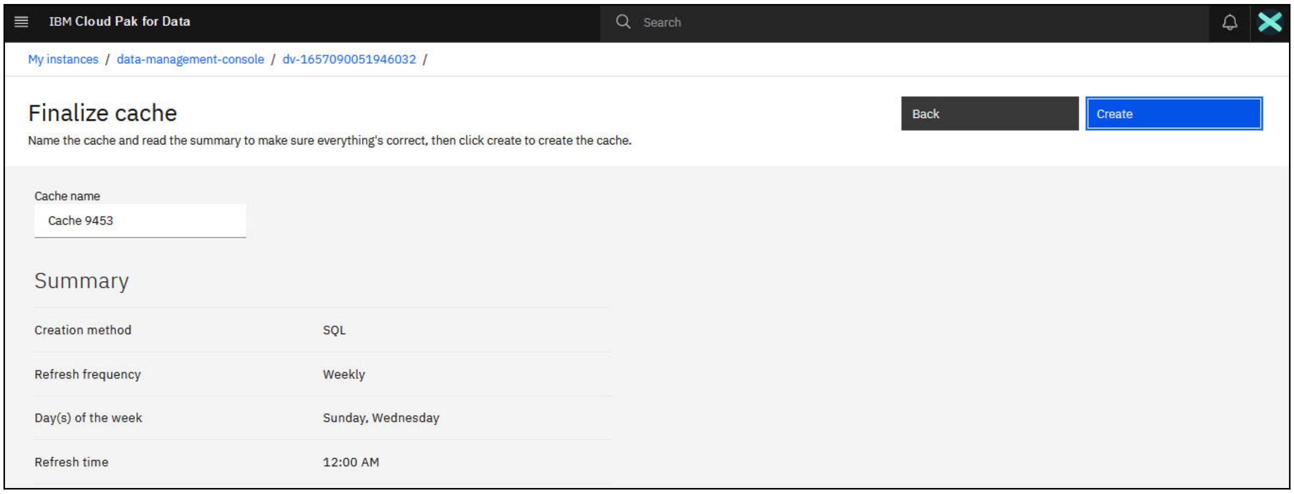

Figure 4-168 Finalizing cache setup

The new cache is now active and can be used by the system whenever the corresponding query is issued against it.

Figure 4-169 shows the statistics and the difference that the new cache made to our setup.

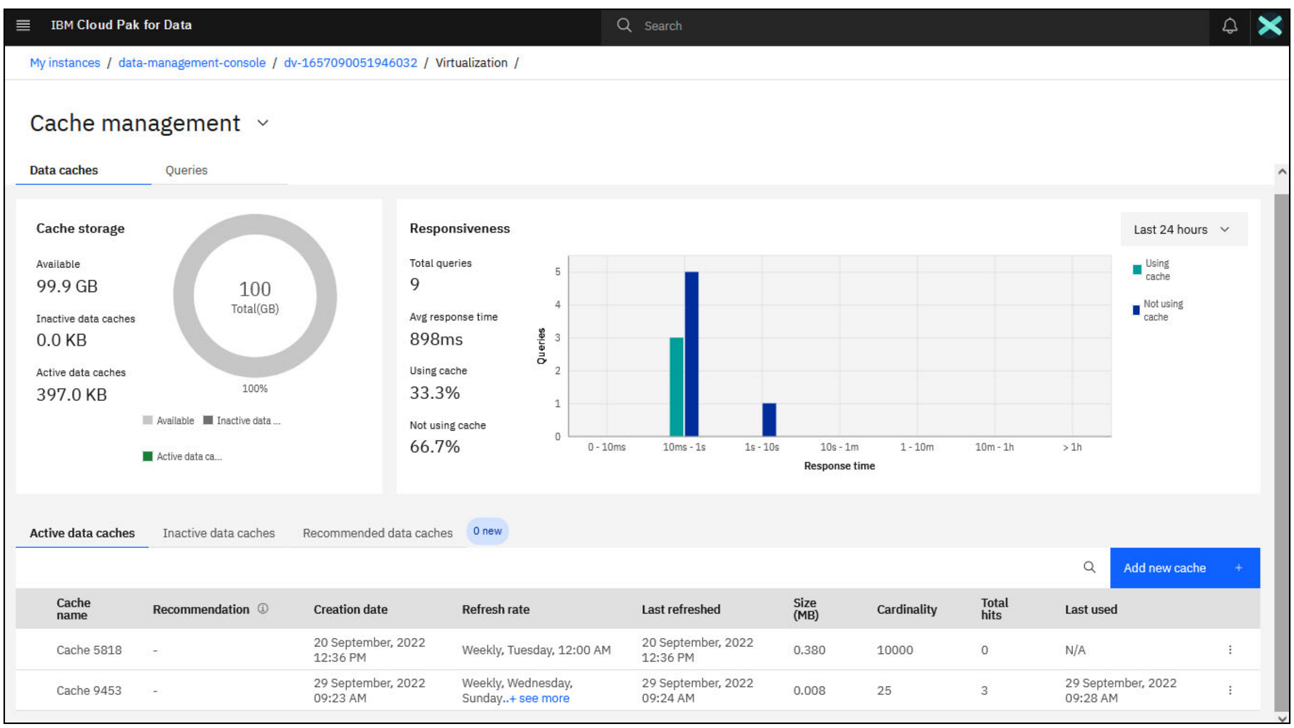

Figure 4-169 Using the new cache

Machine-learning powered cache recommendation engine

In addition to the manual cache creation capabilities, IBM Cloud Pak for Data offers a machine-learning powered cache recommendation engine to help simplify and automate the decision-making process behind best cache candidate selection.

By using an input set of queries, Data Virtualization can recommend a ranked list of data caches that can improve the performance of the input queries and potentially help future query workloads (see Figure 4-170).

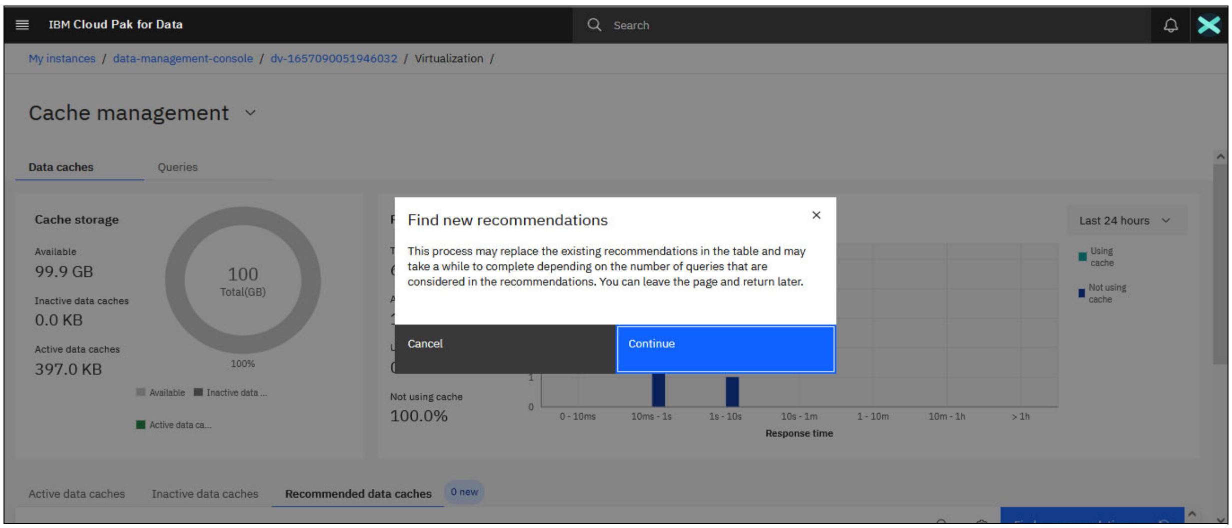

Figure 4-170 Using the in-built cache recommendation engine

The cache recommendation engine uses the following models to generate recommendations:

- Rule-based: Uses sophisticated heuristics to determine which cache candidates help the input query workload.
- Machine learning-based: Uses a trained machine learning model that detects underlying query patterns and predicts caches that help a potential future query workload.

For more information about this feature, see this IBM Documentation web page.

Figure 4-171 shows the cache management landing page of a Data Virtualization instance servicing production-level workloads

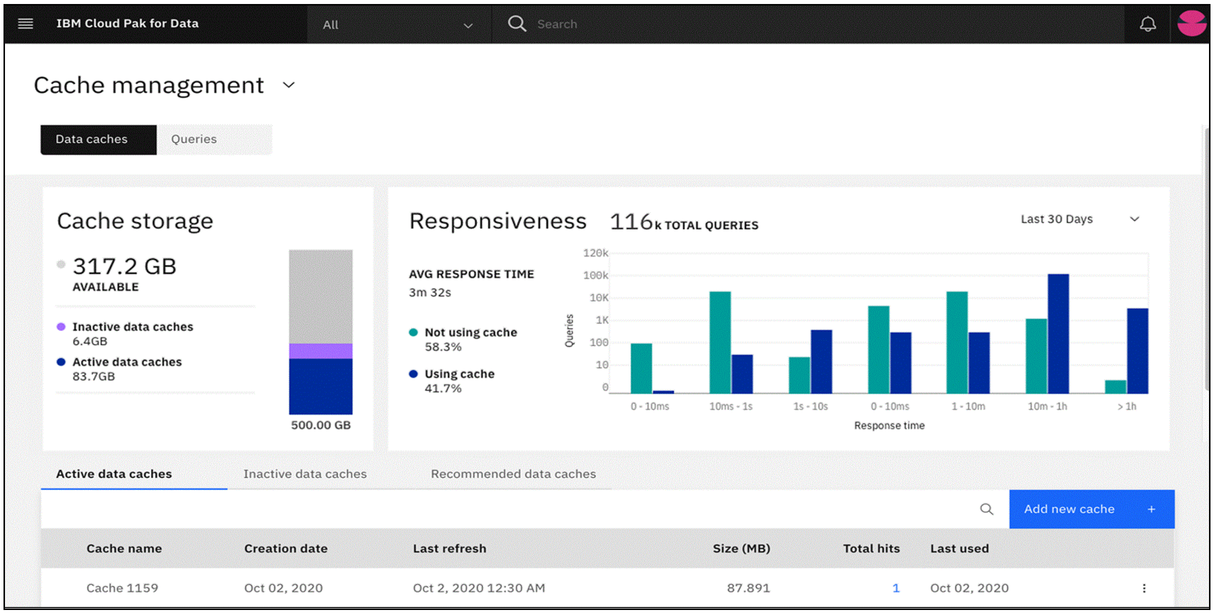

Figure 4-171 Cache management in a production environment

4.3.6 Summary

Enterprise data often is spread across a diverse system of data sources. Companies often seek to break down these silos by copying all of the data into a centralized data lake for analysis.

However, this approach is heavily reliant on ETL, involves data duplication, can lead to stale data and data quality issues, and introduces more storage and other costs that are associated with managing this central data store.

Data virtualization provides an alternative solution that allows you to query your data silos without copying or replicating the data in them in a secure, controlled manner. The solution also provides the ability to view, access, manipulate, and analyze data through a single access point, without the need to know or understand its physical format, size, type, or location.

As a result, data virtualization can help reduce costs, simplify analytics, and ensure that each user is accessing the latest data because the data is accessed directly from the source and not from a copy.

It also is one of the cornerstones of the data fabric solutions and the related architectural approach.

IBM Data Virtualization, also known as IBM Watson Query, is an enterprise-grade data virtualization solution that is available within IBM Cloud Pak for Data, which is the data fabric offering of IBM.

In this chapter, we reviewed the key capabilities of the service by using a simple integration scenario as an example. We also explained that IBM Data Virtualization can complement your ETL workloads (for example, IBM DataStage), and be used as an integration approach in its own right.

For more information about IBM Data Virtualization, see the following resources:

- IBM Documentation: Data Virtualization on Cloud Pak for Data
- Tutorials:
  - Data virtualization on IBM Cloud Pak for Data
  - Improve performance for your data virtualization data sources with remote connectors
- IBM Watson Query on Cloud tutorials:
  - Multicloud data integration tutorial: Virtualize external data
  - Data governance and privacy tutorial: Govern virtualized data

5

Trustworthy artificial intelligence concepts

This chapter explains the trustworthy artificial intelligence (AI) concepts and provides several practical example scenarios.

This chapter includes the following topics:

5.1 Use case description

In an era where trust gained tremendous importance to customers, every organization has a responsibility to adhere to ethical, explainable AI, and respect individual rights, privacy, and non-discrimination.

When trust in AI is established, revenues and customer satisfaction increase, time to market shrinks, and competitive positioning improves. However, eroding trust can lead to failed audits, regulatory fines, a loss of brand reputation, and ultimately to declining revenues.

Success in building, deploying, and managing AI/ML models is based on trusted data and automated data science tools and processes, which requires a technology platform that can orchestrate data of many types and sources within hybrid multicloud environments. Data fabric is a technology architecture approach that helps ensure that quality data can be accessed by the right people at the right time no matter where it is stored.

Well-planned, run, and controlled AI must be built to mitigate risks and drive wanted analytic outcomes. It revolves around the following main imperatives (see Figure 5-1):

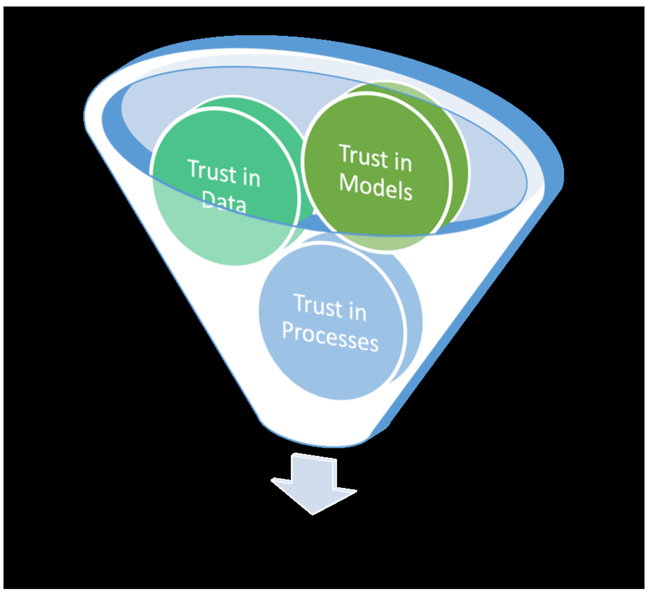

Figure 5-1 Trustworthy AI imperatives

- Trust in data

  Strength and trust in AI outcomes require accurate, high-quality data connections that are ready for self-service use by the correct stakeholders. AI model strength depends on the ability to aggregate structured and unstructured data from disparate internal and external sources, and from on-premises or from public or private clouds.

  Successful data collection and use facilitates fairness in training data, tracking lineage, and ensuring data privacy when offering self-service analysis by multiple personas.

- Trust in models

  To ensure transparency and accountability at each stage of the model lifecycle, Machine Learning Operations (MLOps) automations and integrated data science tools help to operationalize building, deploying, and monitoring AI models.

MLOps increases the efficiency for continuous integration, delivery, and deployment of workflows to help mitigate bias, risk, and drift for more accurate, data-driven decisions. Some unique MLOps implementations also infuse the AI model process with fairness, explainability, and robustness.

- Trust in processes

  Across the model lifecycle, lack of automated processes can lead to inconsistency, inefficiency, and lack of transparency. AI Governance provides automation that drives consistent, repeatable processes to decrease time to production, increase model transparency, ensure traceability and drive AI at scale.

IDC predicts that by 2025, 60% of enterprises will have operationalized their ML workflows through MLOps/ModelOps capabilities and AI-infused their IT infrastructure operations through AIOps capabilities[1].

5.2 Trustworthy AI lifecycle

In this section, we discuss Trustworthy AI at IBM by reviewing the full lifecycle, as shown in Figure 5-2.

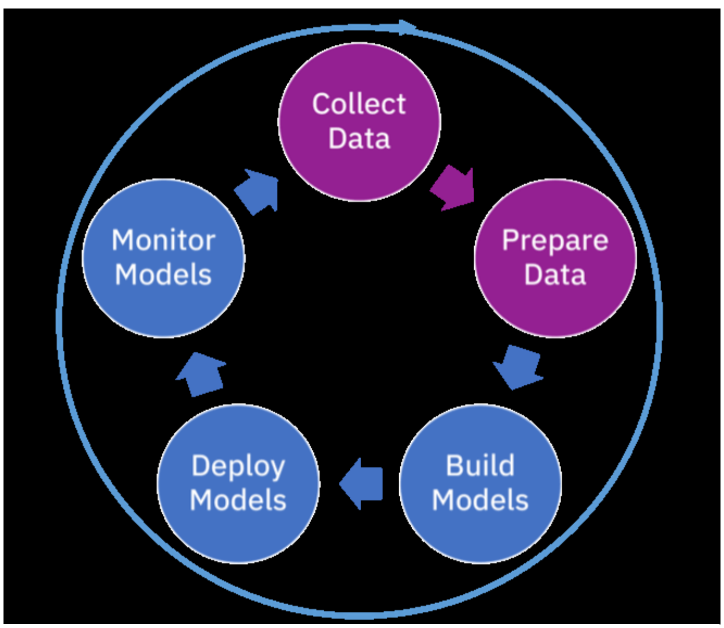

Figure 5-2 Trustworthy AI lifecycle

Typically, when looking from the data scientist's perspective, AI starts at building and managing the models. However, the process starts much earlier. We argue that the most important tasks include collecting and preparing the data to create a trusted data foundation, creating a full view of where the data comes from, what we have done to it, and how to continue to use it. That is, build trust in data.

After data is prepared, the goal is to build models, and eventually deploy them. During this deployment process, many things can occur: the model can be unfair or biased, which can become an issue for the organization that wants to operationalize the model.

[1] Source: IDC Future Scape: Worldwide Artificial Intelligence and Automation 2022 Predictions—European Implications IDC, Jan 14, 2022, https://www.idc.com/getdoc.jsp?containerId=EUR148675522

For example, if the model automatically determines whether a mortgage is approved for a specific customer, it cannot discriminate against age or gender.

Furthermore, models that were infused into business processes must be continuously monitored for performance and a feedback loop created to ensure that they are still suitable and render the correct results.

Cloud Pak for Data implements Trustworthy AI building on a trusted data foundation, which is complemented by a consistent process to build and deploy models so they can be integrated into business processes. Finally, these models must be monitored so they can be validated by anybody in the company without requiring deep technical or data science skills.

The consistent process with a well-described lifecycle and product capabilities supports customers with a continuously updated state of AI implemented by the organization. This AI is then used to establish trust by business users, internal auditors, and compliance by regulators.

Figure 5-3 shows Cloud Pak for Data product capabilities.

Figure 5-3 Product capabilities

Cloud Pak for Data offers many options to collect and connect to data, often referred to as the first rung on the ladder to AI. In many organizations, data is managed in silos, close to where the application is stored, and owned by a certain part of the business.

For example, slivers of customer data might be held in the Customer Relationship Management (CRM) system, which is owned by the sales department. However, other customer-related data, such as shipping letters and invoices, might be held in different applications that are owned by different departments.

When analyzing data and creating analytical models, it is often important to have a broad view of customer data across different applications. Data Virtualization accelerates access to data by virtualizing data that is in different parts of the organization, or even across different public or private clouds.

Another aspect of accelerating the time to value and creating a trusted data foundation is having the ability to find data, especially finding the correct data. If a data consumer (business analyst, data engineer, or data scientist) is looking for customer data and reaching out to many departments, they might get different answers and different data sets, even misleading responses at times.

Establishing an organization-wide or subject area-focused catalog of data assets is imperative to accelerate analytics and establish trust in data. The catalog must consist of data asset names and locations and have metadata to establish data meaning.

For example, if a table column is named SPEED, it is important to have a business term that further defines the column. Is it the speed of a car, the number of rotations per minute of a fan, or wind speed or a hurricane measured in somewhere?

Context and business definition is critical here; without it, data consumers might and likely will make false assumptions that lead to lack of trust in the models that they publish.

Building models is based on capabilities within IBM Watson Studio that allow data engineers, data scientists, and other “data citizens” to collaborate. Consistency, flexibility, and repeatability are key factors in this part of the process.

Data scientists must be able to collaborate with business subject matter experts (SMEs) to understand the data and then, prepare data and build the models by using visual “no or low code” tools or coding in Python, R, or Scala. The building and validation of models must be automated through orchestration so that SMEs and data scientists can focus on their core activities.

Every model that is a candidate to support business decisions must be accompanied by a fact sheet. A *fact sheet* contains metadata that describes key attributes of the model, such as features that are used to train it, data preprocessing facts, model performance, bias, and other attributes to allow business users to understand the contents, much like a food label that provides nutrition facts.

After a model is put in production, it must be continually validated because the business, users, or data might change, which can cause the model to drift or bias might be affected. Linking the model to a business owner and processing the continuous validation of the model ensures that the model still meets the requirements of the business. It also creates confidence when making decisions that are supported by the model.

Models that drift can be automatically retrained, or the business owner can make a decision not to use the model anymore or to replace it with a different model or augment it with a manual activity.

This lifecycle and imperatives are easily managed by organizations that have only a few AI models that are infused into business processes. However, when scaling out to tens or even hundreds of models, a well-oiled process and lifecycle with checkpoints becomes a necessity.

In the sections to follow, we guide you through a practical, hands-on exercise that captures each of the phases in the lifecycle, from collecting and organizing the data to create a trusted data foundation, building and managing models to build trust in models and finally, monitoring models to establish the trust in the end-to-end process.

5.3 Architecture

Figure 5-4 a high-level overview of the IBM Data Fabric and Trusted AI platform reference architecture and its components. These components form the foundation that helps customers along in their Trustworthy AI journey. In this architectural diagram, we ignore data integration tools and focus on accessing data that is in place.

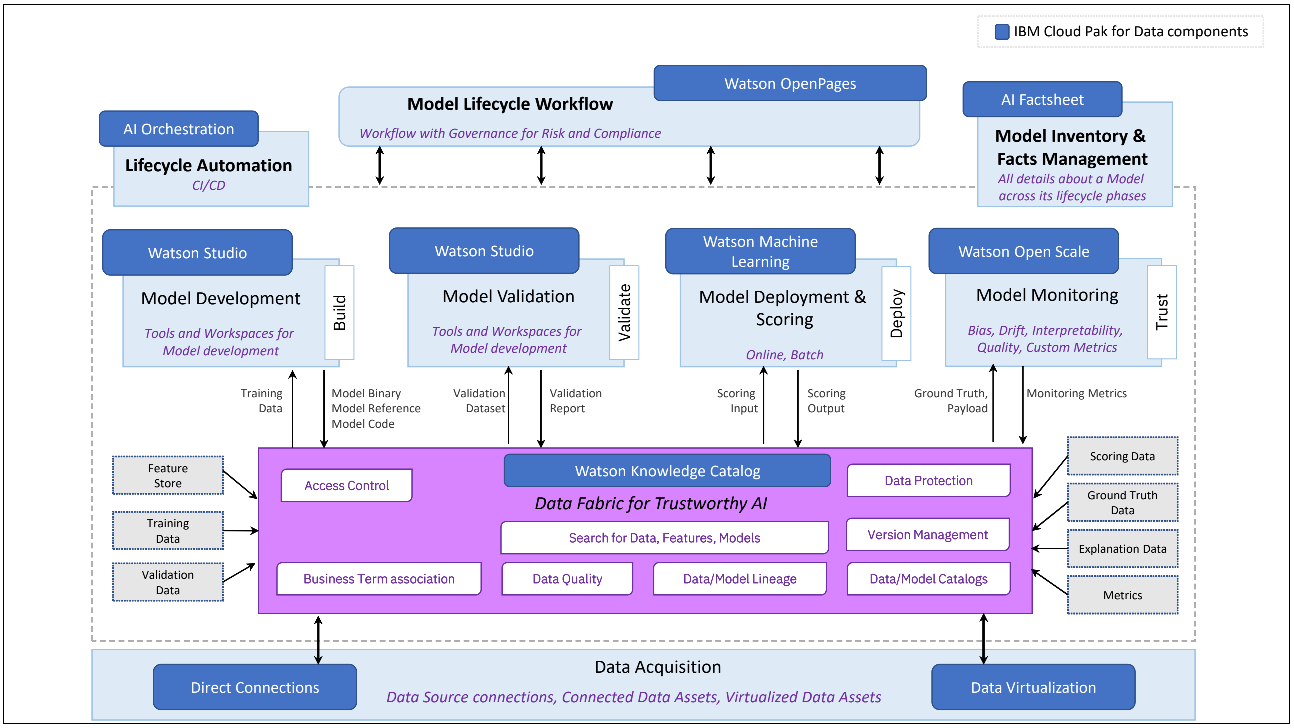

Figure 5-4 Reference Architecture of Trustworthy AI - Cloud Pak for Data components

The following IBM Cloud Pak for Data services enable the end-to-end Trustworthy AI platform capabilities:

- IBM Watson Knowledge Catalog: A data governance service that is used to find the right data assets fast. Discover relevant, curated data assets by using intelligent recommendations and user reviews.
- Watson Studio: A service for building custom models and infusing business services with AI and machine learning.
- Watson Machine Learning: Build and train machine learning models by using tools. Deploy and manage machine learning models at scale.
- Watson Open Scale: AI services to understand model drifts, and how your AI models make decisions to detect and mitigate bias.
- Watson OpenPages: Model Risk governance service that identifies, manages, monitors, and reports\ on risk and compliance initiatives in an organization.
- Data Virtualization: Connects multiple data sources across locations and turns all of this data into one virtual data view.

5.4 Trustworthy AI workshop

In this section, we will discuss a Trustworthy AI use case.

5.4.1 Downloading assets for the workshop

Throughout the workshop, you must download the assets that are available in this GitHub repository. Then, complete the following steps:

1. At the GitHub repository, click **Code** and then, **Download ZIP**.

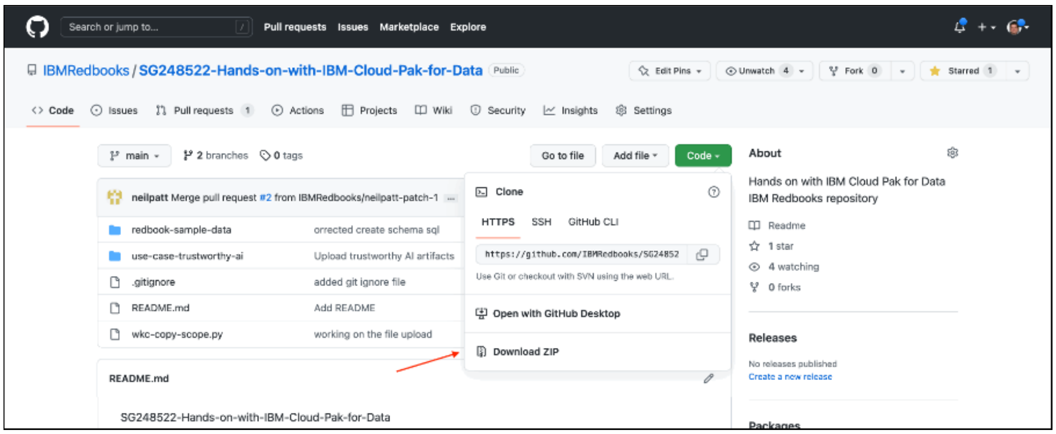

Figure 5-5 Download zip

2. After the assets are downloaded, decompress the downloaded file. You use these files throughout the workshop.

5.4.2 Creating the required users in Cloud Pak for Data

As an enterprise data and AI platform, Cloud Pak for Data supports separating roles and responsibilities. In this workshop, you run different tasks while assuming different user personas with different roles and permissions.

Specifically, you run tasks as one of the following users:

- `datasteward` (Data Steward role and permissions)
- `dataengineer` (Data Engineer role and permissions)
- `dqanalyst` (Data Quality Analyst role and permissions)
- `datascientist` (Data Scientist role and permissions)
- `dslead` (Data Scientist and Data Engineer role and permissions)
- `uatops` (Data Scientist role and permissions)
- `prdops` (Data Scientist role and permissions)

The Cloud Pak for Data environment initially has the admin user only. You log in as this admin user and create the different users and associate the corresponding roles.

Complete the following steps:

1. Log in to Cloud Pak for Data as the `admin` user.
2. Browse to **Access control** by clicking the top-left navigation menu and then, clicking **Administration → Access control**.

3. Click **Add user** and create the `datasteward` user. Enter a full name, user name, email, and password (twice). Click **Next**.
4. Click **Next** to accept the default of **Assign roles directly**.
5. In the next window, select the **Data Steward** role and then, click **Next**.
6. Review all the values you provided and then, click **Add**
7. Repeat these steps and create the following users:
   - `dataengineer` with the role of Data Engineer
   - `dqanalyst` with the role of Data Quality Analyst
   - `datascientist` with the role of Data Scientist
   - `dslead` with the roles of Data Scientist and Data Engineer
   - `uatops` with the role of Data Scientist
   - `prdops` with the role of Data Scientist

It is worthwhile spending some time reviewing the predefined roles and associated permissions for each role because this information influences what each user can and cannot do on the Cloud Pak for Data platform.

5.4.3 Collecting data

Collecting and organizing data is an important step in building an automated AI pipeline. Data scientists create projects, and data engineers collect data and add it to the projects so it can be organized and refined.

Models and algorithms are only as good as the data that is used to create them. Data can be incomplete, or biased, which can lead to inaccurate algorithms and analytic outcomes.

Data scientists, analysts, and developers need self-service data access to the most suitable data for their project. Setting up privacy controls for multiple personas, providing real-time access to “right” data users, and tracking data lineage can be challenging without the right tools.

By using Cloud Pak for Data, you can collect data from multiple sources and ensure that it is secure and accessible for use by the tools and services that support your ModelOps AI lifecycle. You can address policy, security, and compliance issues to help you govern the data that is collected before you analyze the data and use it in your AI models. A method of tying together these sources is vital.

Collected data assets are governed by using IBM Watson Knowledge Catalog capabilities to ensure that enterprise governance and compliance rules are enforced. Use cases exist in which it makes sense to connect your Python, R, or Scala code to data assets and sources directly and using such data for extracting insight and training AI models.

However, in general, it is recommended to catalog all data assets and use only cataloged data for subsequent tasks, such as business intelligence insights and training AI models. Cataloging the data ensures that the enterprise’s governance and compliance rules are enforced and trusted data is delivered.

Data Virtualization

Data Virtualization can be used for data collection when you need to combine live data from multiple sources to generate views for input for projects. For example, you can use the combined live data to feed dashboards, notebooks, and flows so that the data can be explored.

With Data Virtualization, you can quickly view all your organization's data after connections are created to your data sources. This virtual data view enables real-time analytics without data movement, duplication, ETLs, or other storage requirements, so data availability times are greatly accelerated. You can bring real-time insightful results to decision-making applications or analysts more quickly and dependably than methods without virtualization.

IBM Watson Knowledge Catalog

IBM Watson Knowledge Catalog is an intelligent data catalog that powers self-service discovery of data, models and more. With IBM Watson Knowledge Catalog, users can access, curate and share data, knowledge assets, and their relationships wherever they are stored.

Governance is the process of curating, enriching, and controlling your data. The task of setting up a governance foundation for an enterprise is not specific to one use case, but rather aligned to the governance and compliance requirements of the enterprise.

IBM Watson Knowledge Catalog supports creating governance artifacts, such as business terms, classifications, data classes, reference data sets, policies, and rules, which can be used throughout IBM Cloud Pak for Data. Governance artifacts can be organized by using categories.

IBM Watson Knowledge Catalog also supports workflows for governance artifacts, which enforce automatable task-based processes to control creating, updating, deleting, and importing governance artifacts.

During the following use case, you create the categories and the governance artifacts in IBM Cloud Pak for Data that are required for a Telco Churn use case. The governance artifacts that are created in this tutorial are derived from the Telco Churn Industry Accelerator.

Categories

IBM Watson Knowledge Catalog supports organizing governance artifacts by using categories. A *category* is similar a folder or directory that organizes your governance artifacts and administers the users who can view and manage those artifacts.

Categories provide the logical structure for all types of governance artifacts, except data protection rules. You group your governance artifacts in categories to make them easier to find, control their visibility, and manage.

Categories can be organized in a hierarchy that is based on their meaning and relationships to one another. A category can have subcategories, but a subcategory can have only one direct parent category.

Complete the following steps to import a category from a file:

1. Log in to your Cloud Pak for Data as the `admin` user (see Figure 5-6).

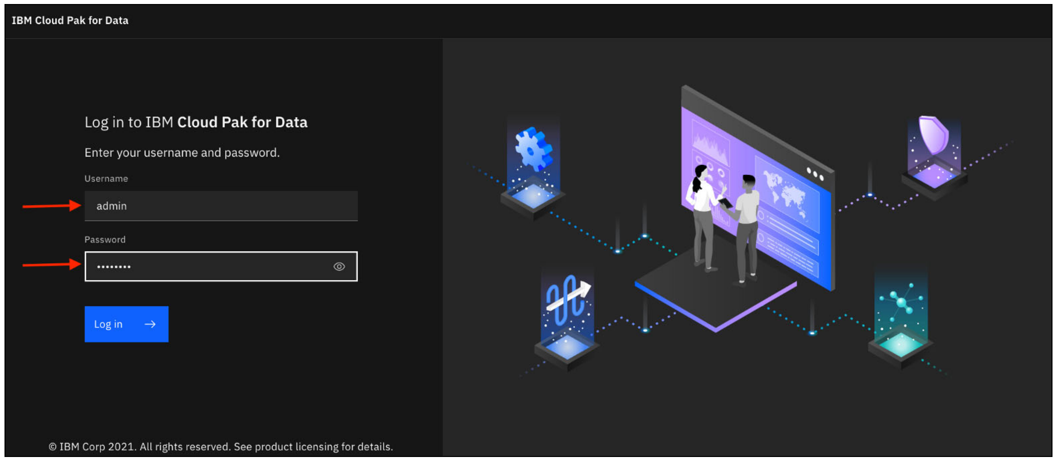

Figure 5-6 Logging in to Cloud Pak for Data

2. From the navigation menu in the upper-left, expand **Governance** and then, click **Categories** (see Figure 5-7).

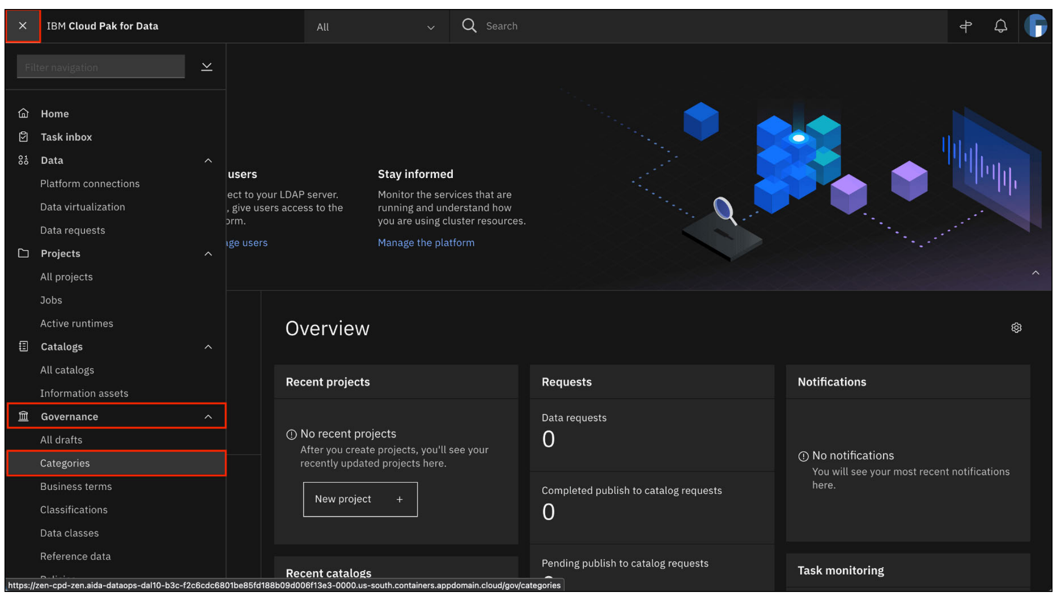

Figure 5-7 Selecting Governance Categories

3. Click **Add category** → **Import from file**.
4. Click **Add file** and then, select the **TelcoChurn-glossary-categories.csv** file from the **governance-artifacts** folder that you downloaded earlier. Click **Next**.
5. Select **Replace all values** and then, click **Import**.
6. The categories are imported from the file. After the import process is successful, you see the Import summary modal that says, “The import completed successfully.” It also shows you that three categories were created and no errors were encountered. Click **Close**.

The categories that were created by the import process is the Telco Churn category and its subcategories: Personal Info and Customer Data.

For more information about the overall solution for Telco Churn prediction, see this web page.

Adding collaborators to categories

Complete the following steps to add collaborators to categories:

1. Refresh the web page to see the newly imported categories. Click the **Telco Churn** category (see Figure 5-8).

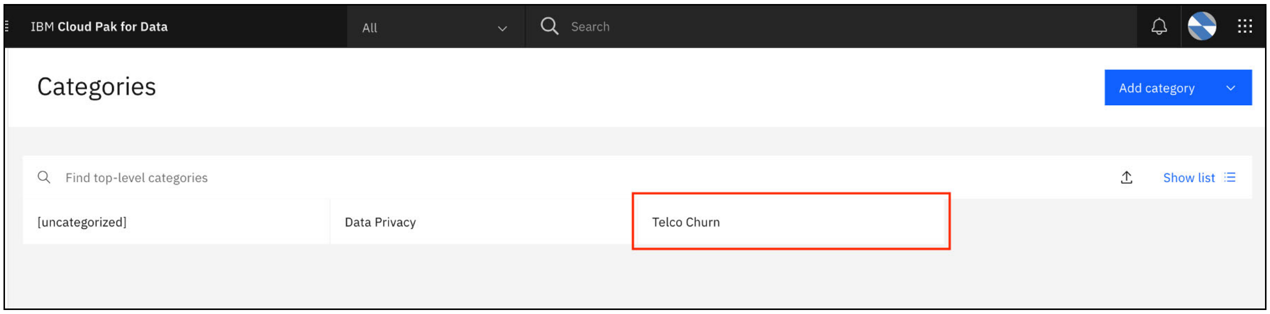

Figure 5-8 Select created Category

2. When first created, the categories include only a Name, Description, and a default set of Collaborators. To add collaborators to the Telco Churn category, click the **Access control** tab and then, click **Add collaborators** → **Add user groups** (see Figure 5-9).

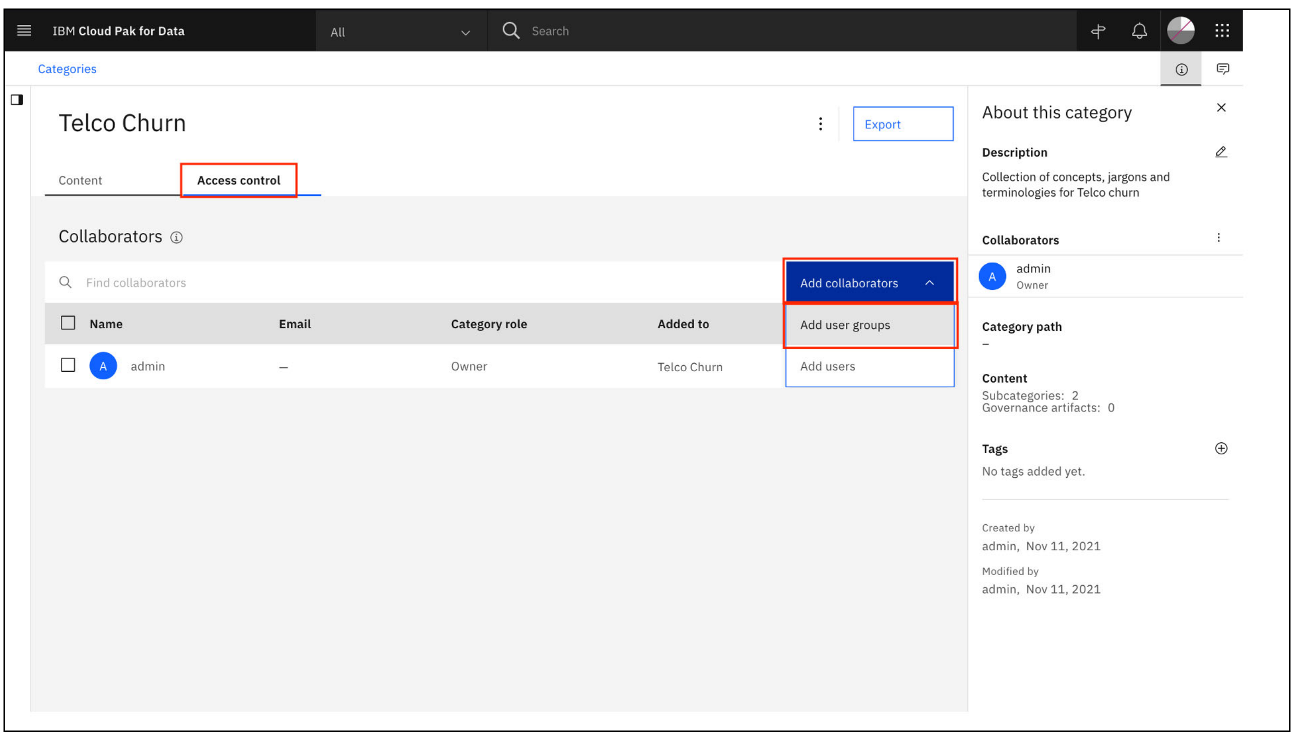

Figure 5-9 Adding user groups

> **Note:** To add a user or user group as a collaborator to a category, that user or user group must have the Cloud Pak for Data platform permission to “Access governance artifacts”.

3. Add the All users group as Viewers of the Telco Churn category. Select the **All users** users group. Click **Roles** and select one or more roles that you want to add (in this case, click **Viewer**). Click **Add**.
4. Return to the Access Control tab of the Telco Churn category and click **Add collaborators** → **Add users**.
5. Add the user `datasteward` as the Owner and Editor of the Telco Churn category. Select the `datasteward` user. Click **Roles** and select the roles that you want to add (**Owner** and **Editor**). Click **Add**.
6. When you add users to a category, the users are added with the same permissions to all the subcategories of that category. In the **Content** tab, click the name of one of the subcategories of the Telco Churn category to verify the permissions.
7. Click **Categories** in the breadcrumbs to return to the Categories main page (see Figure 5-10).

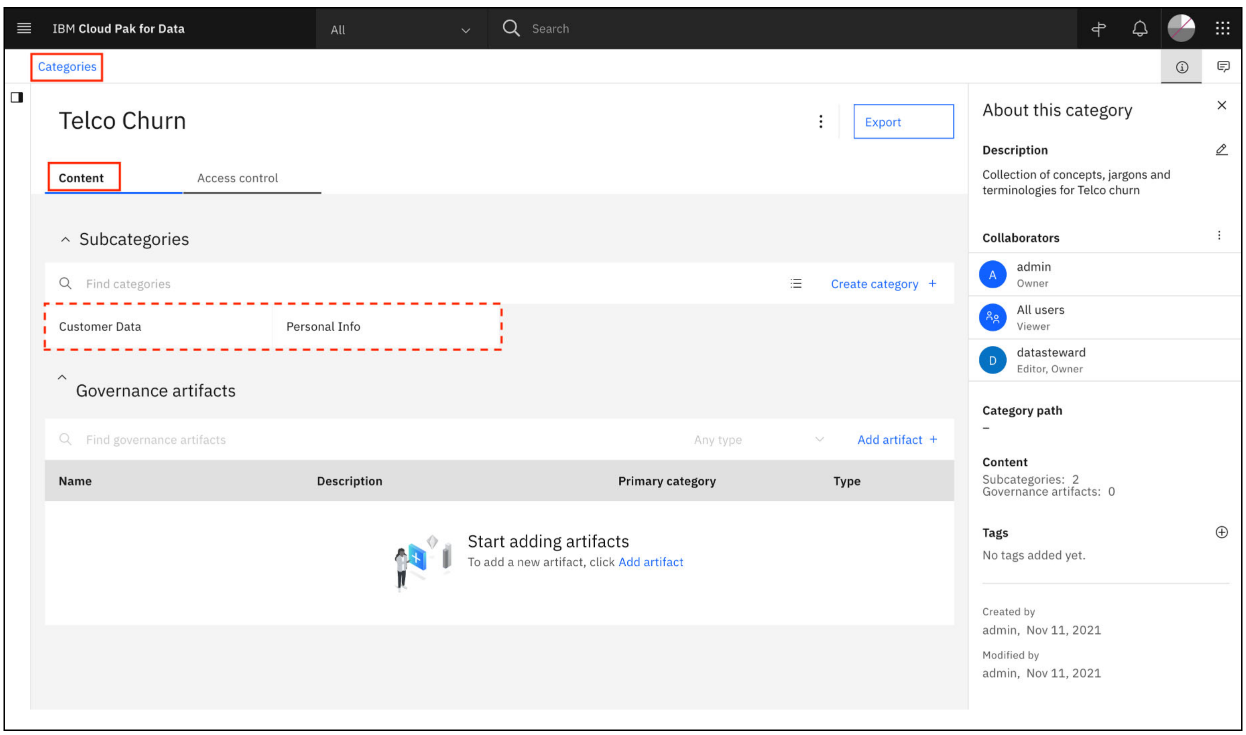

Figure 5-10 Permissions added to sub Categories

8. Repeat these steps to add the `datasteward` user as the Owner/Editor for the remaining category Data Privacy. The All users group was provided with the Viewer role for the Data Privacy category.

Business terms

Business terms are used to create an ontology around the business. They are used to characterize other artifacts and assets and columns in the catalog. The use of business terms for business concepts ensures that the enterprise data is described in a uniform manner and is easily understood across the entire enterprise.

Business terms can be used to describe the contents, sensitivity, or other aspects of the data, such as the subject or purpose. You can assign one or more business terms to individual columns in relational data sets, to other governance artifacts, or to data assets.

Creating business terms by importing a file

Complete the following steps to upload business terms under the Telco Churn category to fulfill the request that you approved:

1. Log in to Cloud Pak for Data as the `dataengineer` user.
2. From the navigation menu in the upper-left corner, expand **Governance** and then, click **Business terms**.
3. Expand the **Add business term** drop-down menu and click **Import from file**. (Business terms also can be added manually.)
4. In the pop-up window, click **Add file** and browse to select the downloaded **TelcoChurn-glossary-terms.csv** file from the **governance-artifacts** folder and then, click **Next**.
5. Select the **Replace all values** option and click **Import**.
6. After the import process is complete, you see a message: "The import completed successfully". Click **Go to task** to review and approve the imported business terms. All governance artifacts are first saved as draft to allow for edits and updates as needed before publishing.
7. In the **Assigned to you** tab, you see a new task to "Publish Business terms". Click **Claim task** so that other assignees know that you handle this task.

   The "Description" field indicates that the task is to review each business term and then publish or delete all of the business terms together. It also explains that the `datasteward` imported these business terms.

   Review the imported business terms and click **Publish**. It is also a good practice to add a comment. Enter a comment (for example, "Business terms imported and published.") and click **Publish** (see Figure 5-11).

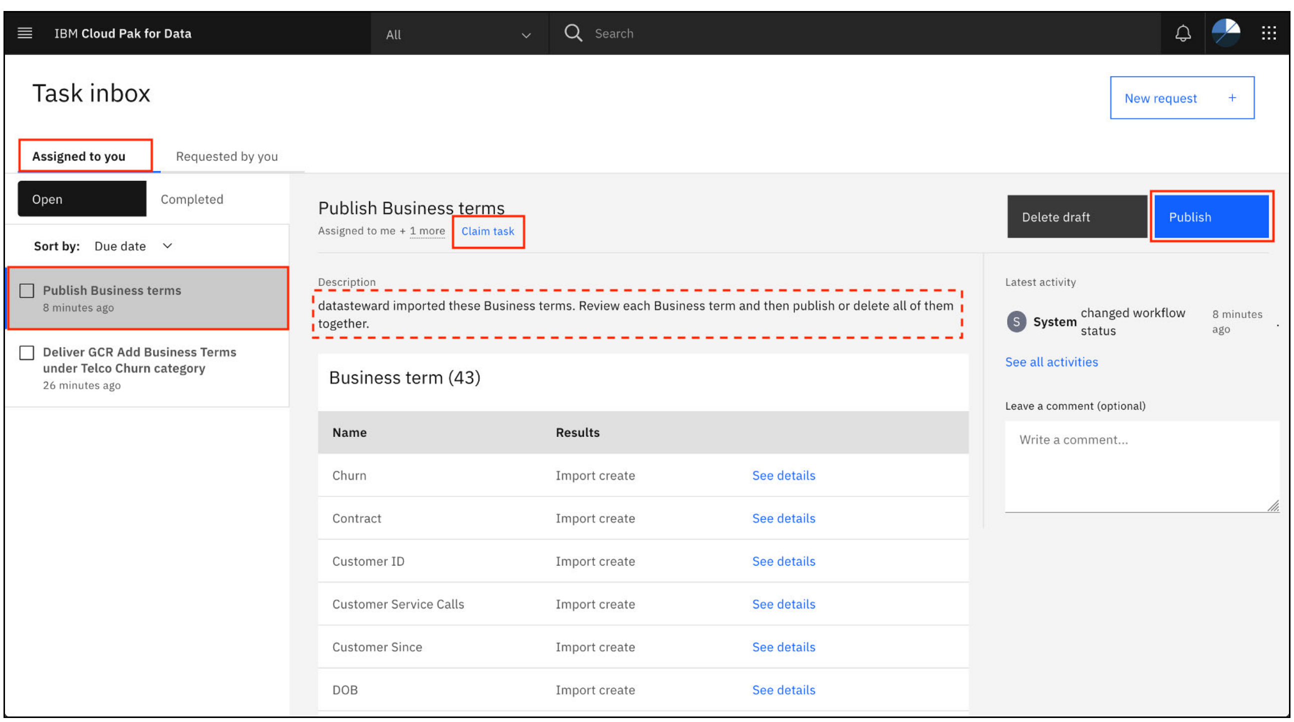

Figure 5-11 Claiming and publishing a task

8. Return to the Task inbox window Deliver GCR Add Business Terms under Telco Churn category. Click **Done** to mark the task as Completed. Optionally, enter a description in the confirmation window and then, click **Done** (see Figure 5-12).

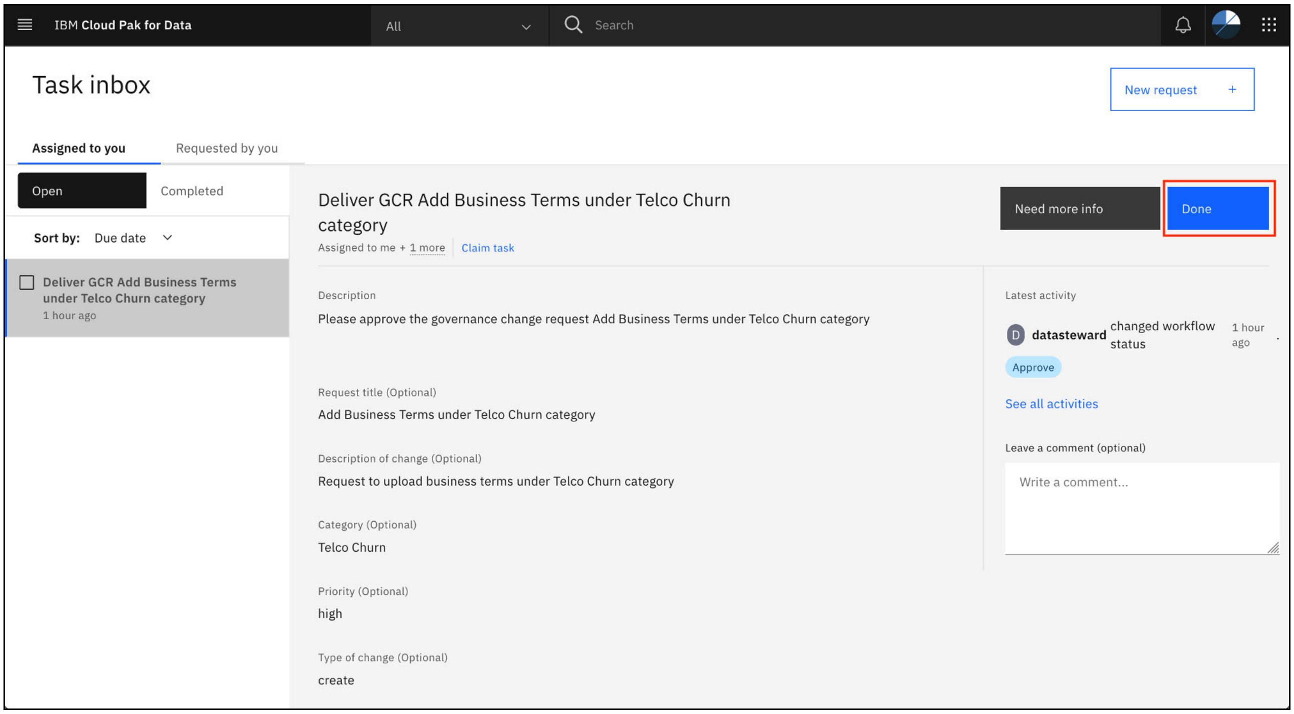

Figure 5-12 Completed task

Governance policies

Policies capture important initiatives that are driven by the organization.

In this section, you create a policy that requires all analytics teams to follow the same standards for US State and County codes.

Creating a policy by importing a file

Complete the following steps to create a policy by importing a file:

1. From the navigation menu in the upper-left corner, expand **Governance** and then, click **Policies** (see Figure 5-13).

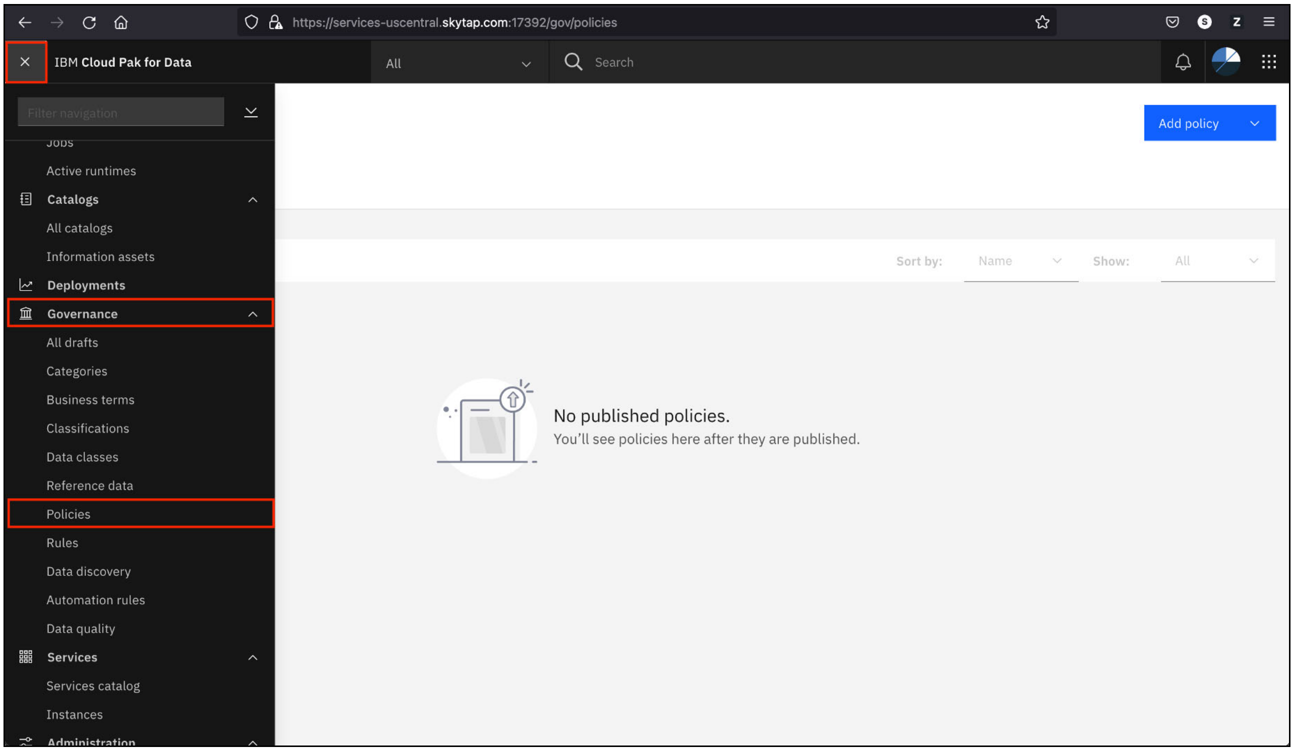

Figure 5-13 Selecting Governance Policies

2. Expand the **Add policy** drop-down menu and click **Import from file**.
3. Click **Add file** and browse to select the **TelcoChurn-policies.csv** file from the downloaded **governance-artifacts** folder and then, click **Next**.
4. Select the **Replace all values** option and click **Import**.
5. After the import process is complete, you see a message that confirms that the import completed successfully. Click **Go to task** to review and publish the policies by following the same process that is used for business terms.

Governance rules

IBM Watson Knowledge Catalog supports two types of rules that can be added to policies:

- Data protection rules automatically mask or restrict access to data assets.
- Governance rules explain specific requirements to follow a policy.

Creating governance rules by importing a file

Complete the following steps to define more governance rules by importing a file:

1. From the navigation menu in the upper-left corner, expand **Governance** and click **Rules** (see Figure 5-14).

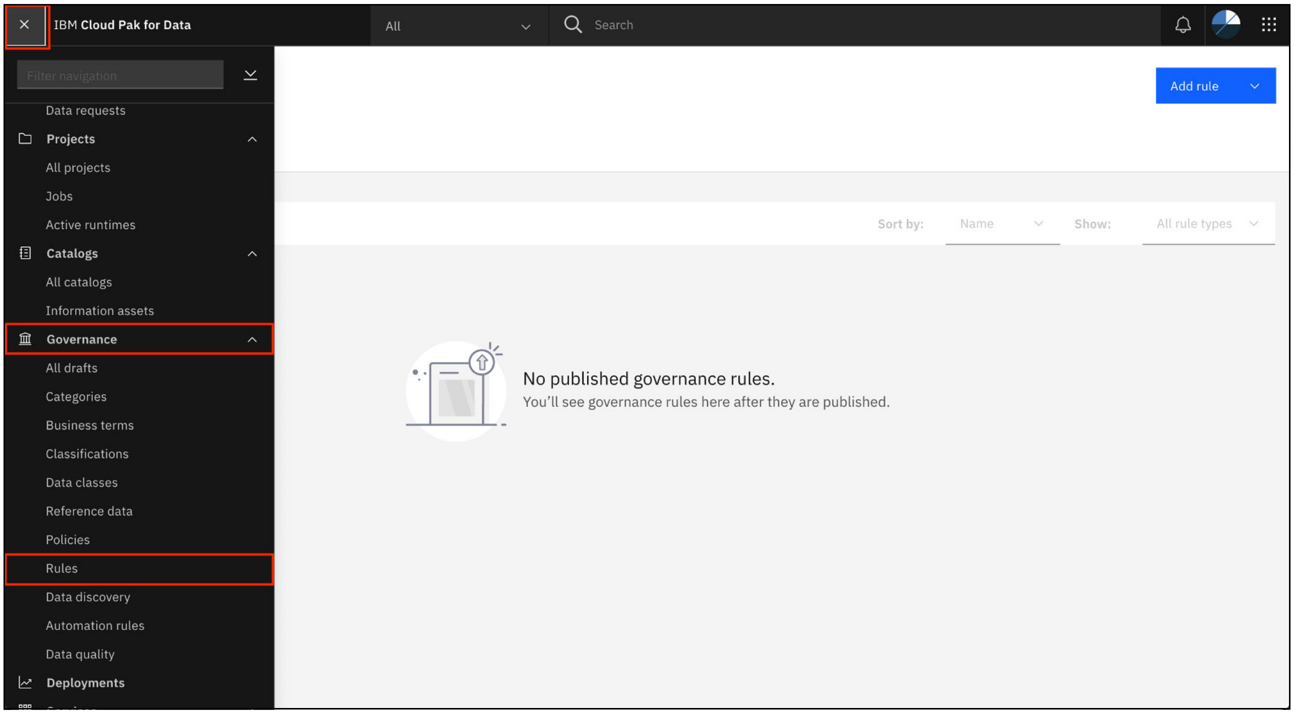

Figure 5-14 Selecting Governance Rules

2. Expand the **Add rule** drop-down menu and select **Import from file**.
3. Click **Add file** and browse to select the **TelcoChurn-rules.csv** file from the downloaded **governance-artifacts** folder and then, click **Next**.
4. Select the **Replace all values** option and click **Import**.
5. After the import process is complete, you see a message that confirms that the import completed successfully. Click **Go to task** to review and publish the rules by following the same process that is used for business terms.

Associating a governance rule with a policy

After the rule is published, you can associate it with the policy created earlier using one of the two methods: by adding the parent policy to the rule, or by adding the rule to the policy.

Complete the following steps to add the parent policy to the rule:

1. In the rule page, click **Add policy +** in the Parent Policies section (see Figure 5-15).

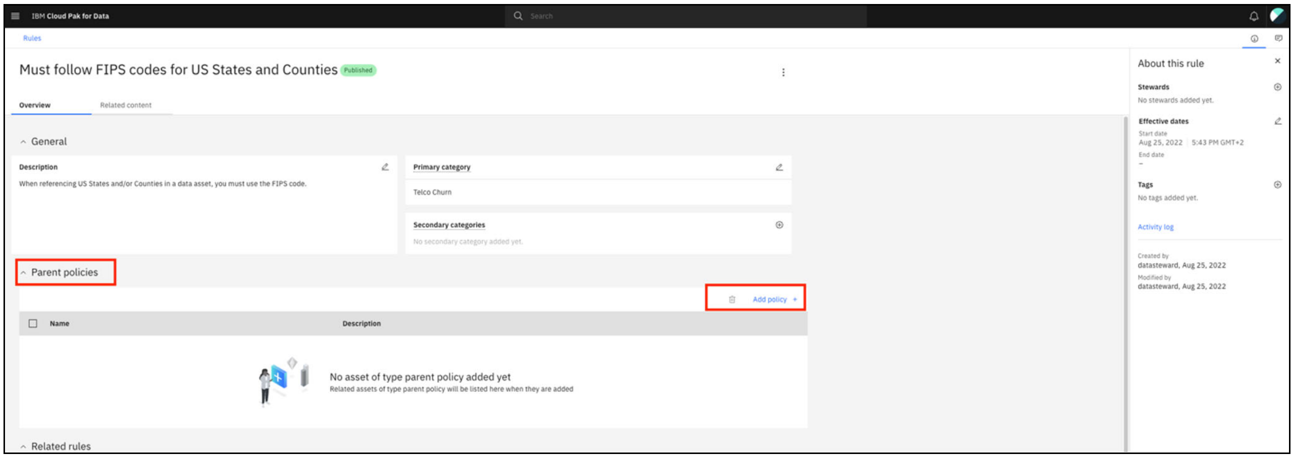

Figure 5-15 Adding Parent Policy to Rule

2. In the pop-up menu, select the **Use of US State and County codes** policy and click **Add**.
3. Publish the rule again to publish this association between the policy and the rule.
4. Alternatively, you can update the policy and add the rule to it:
   a. Return to the governance policies by clicking the navigation menu in the upper-left corner. Expanding **Governance** and clicking **Policies**.
   b. In the Policies window, click the **Use of US State and County codes** policy (see Figure 5-16).

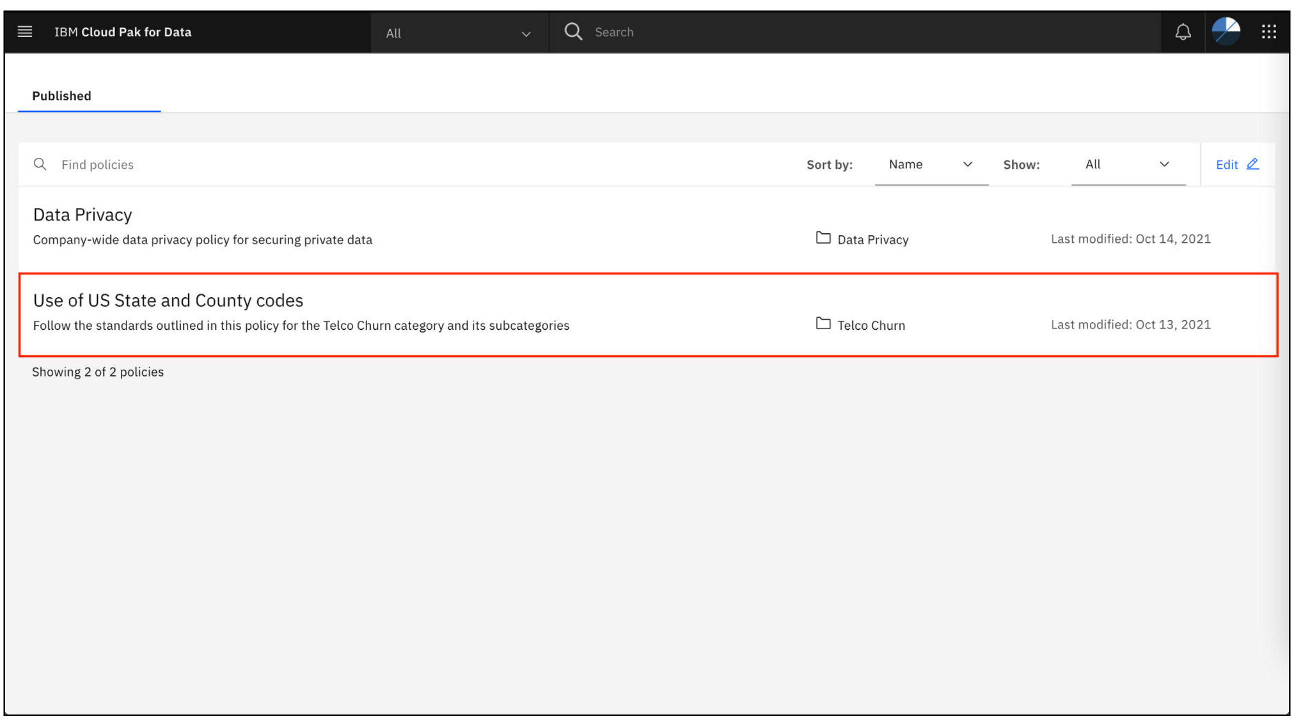

Figure 5-16 Adding Rule to Policy

5. Scroll down to the Rules section and click **Add rules +**.
6. Select the **Must follow FIPS codes for US States and Counties** governance rule and then, click **Add**.
7. After the policy is updated, you must publish the updated policy (see Figure 5-17).

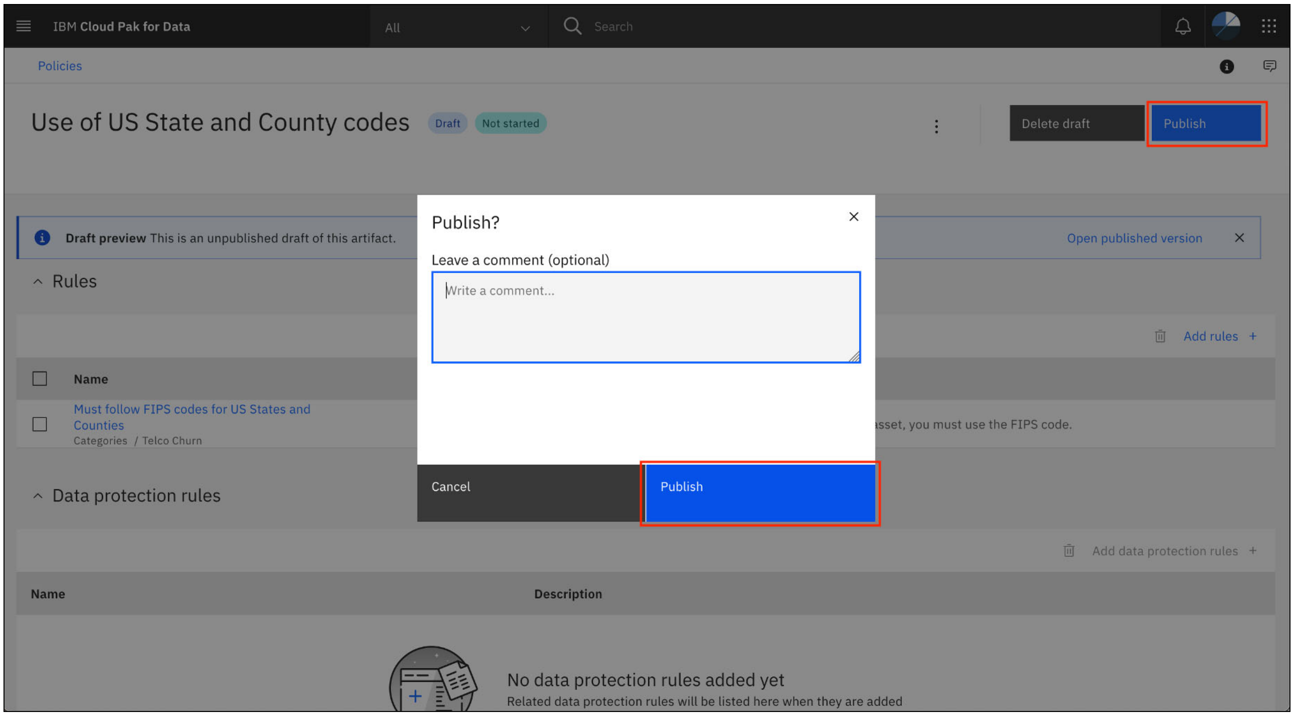

Figure 5-17 Publishing the updated policy

Referencing data sets and hierarchies

Reference data sets are common codes and values that are used across the organization. It is important to centrally manage reference data so that the organization can use codes and values consistently in different systems and track how reference data sets used in different systems relate to each other.

Creating reference data sets

Complete the following steps to create reference data sets:

1. From the navigation menu in the upper-left corner, expand **Governance** and click **Reference data** (see Figure 5-18).

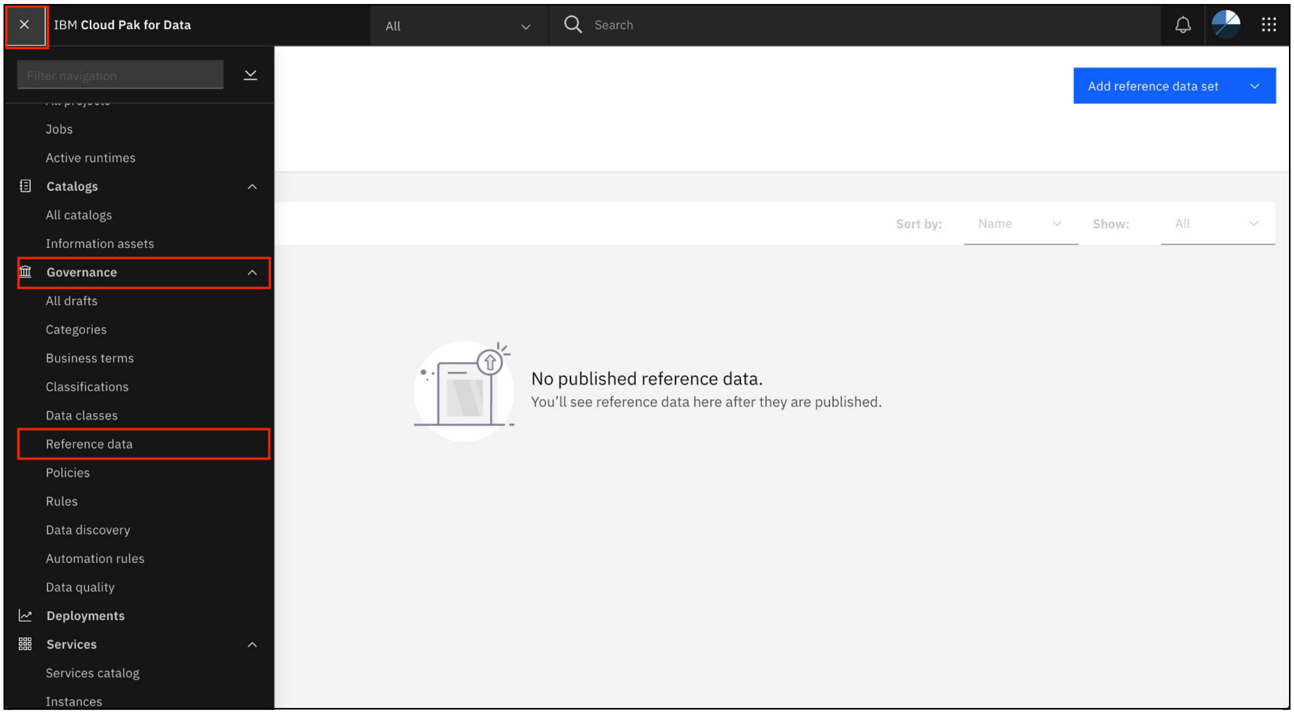

Figure 5-18 Selecting Governance Reference Data

2. Expand the **Add reference data set** drop-down menu and select **New reference data set**.

3. Upload the `USStateCodes.csv` file from the downloaded **governance-artifacts** folder, which includes the codes for the US State Codes reference data set. Enter a Name (US State Codes) and an optional Description (FIPS codes for US States) for the reference data set. Click **Change** to select the **Telco Churn** category and then, click **Next** (see Figure 5-19.

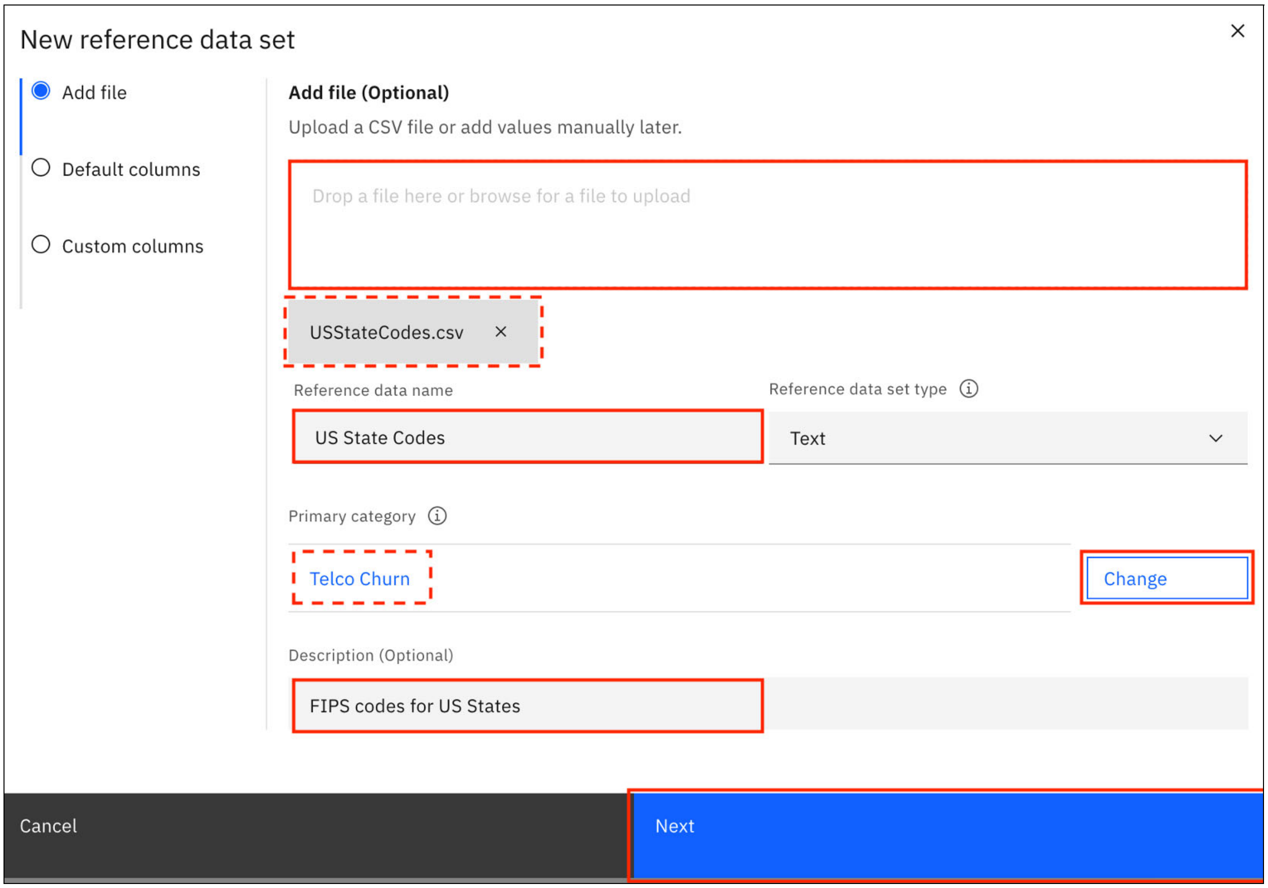

Figure 5-19 Adding a reference data set

4. Ensure that the **First row as column header** option is set to `On`, so that it picks up the titles of the columns in your file. Set the Name column to be `Value` and the FIPS Code column to be `Code` and then, click **Next** (see Figure 5-20).

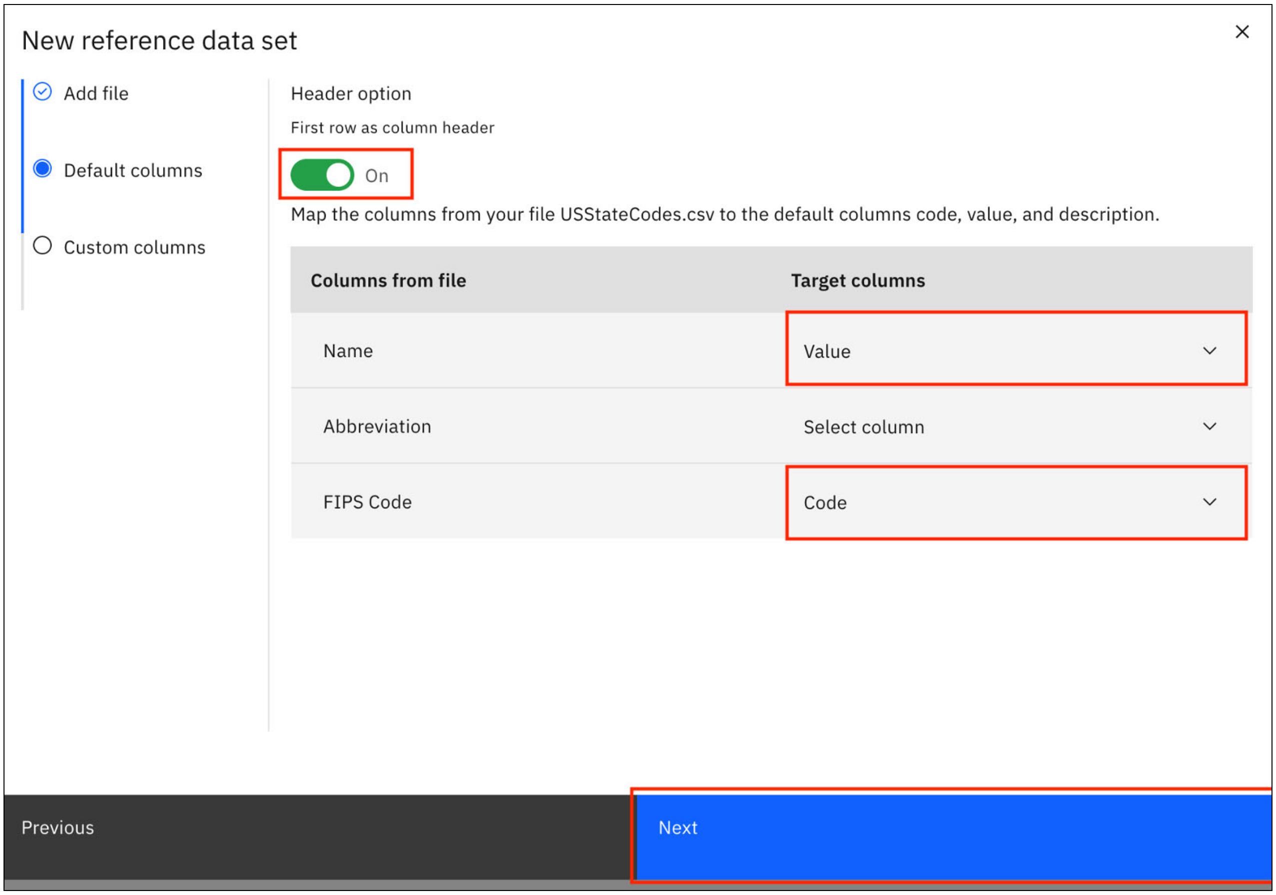

Figure 5-20 Reference data set default columns

Note: IBM Watson Knowledge Catalog offers the following default columns:

- Code
- Value
- Description
- Parent (when creating a set-level hierarchy)
- Related terms (when relating values to Business Terms)

5. Because “Abbreviation” is not one of the default columns that is supported in IBM Watson Knowledge Catalog, you must create a custom column:
   a. Select the **Abbreviation** option.
   b. Enter a name (ABBREVIATION) for the column. The name of the column is shown in the list with a plus sign.
   c. Click **ABBREVIATION** to add the custom column (see Figure 5-21).

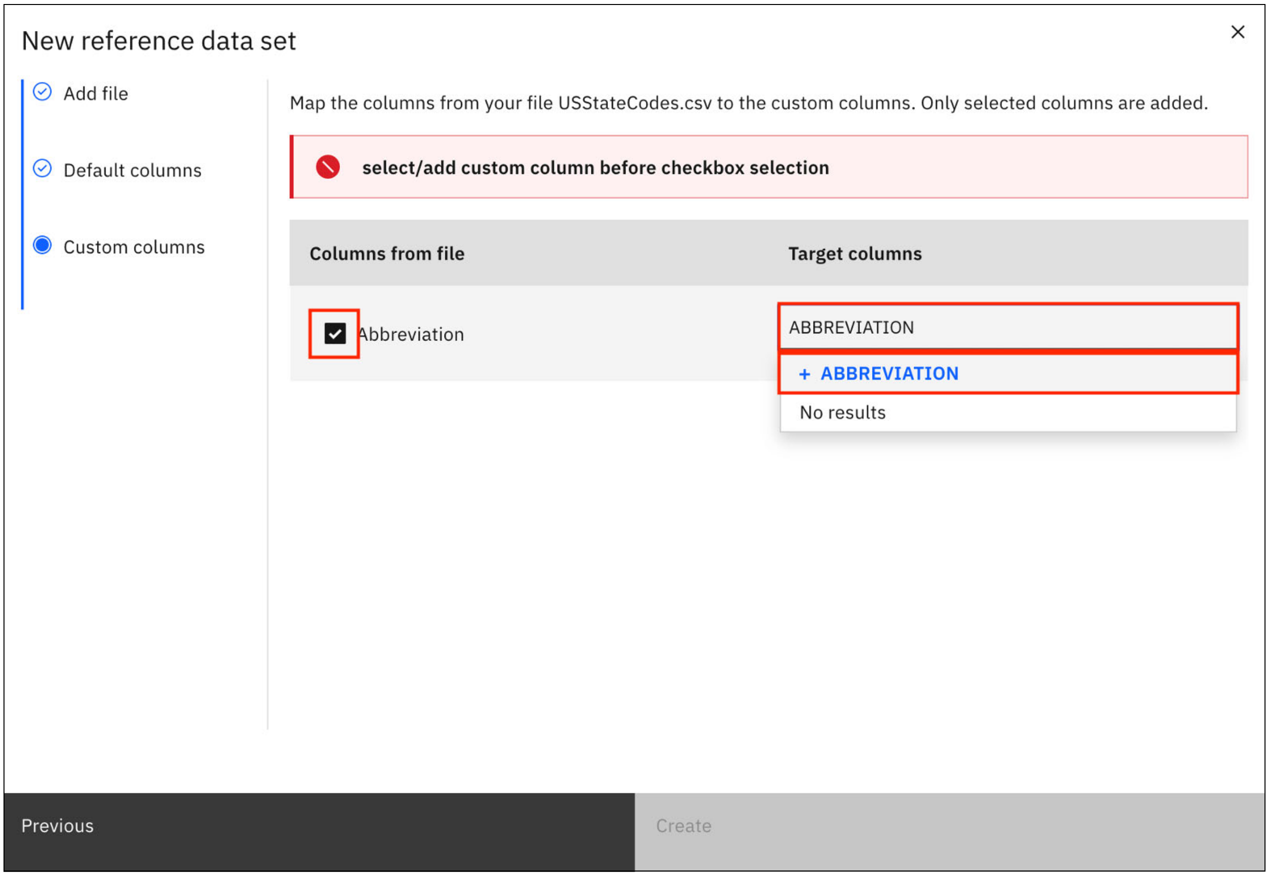

Figure 5-21   Reference data set custom columns

   d. Complete the fields in the next prompt by specifying the Type as Text and providing an optional description and then, click **Save** to save that column as a re-usable column (this process helps the organization be more consistent in how they name columns that are used in reference data sets).
   e. Click **Create**.

6. You are returned to the reference data set that you created. For each value, you see the default columns plus the new column that you created for Abbreviation. Add `datasteward` as the steward and add `churn` as a tag (see Figure 5-22).

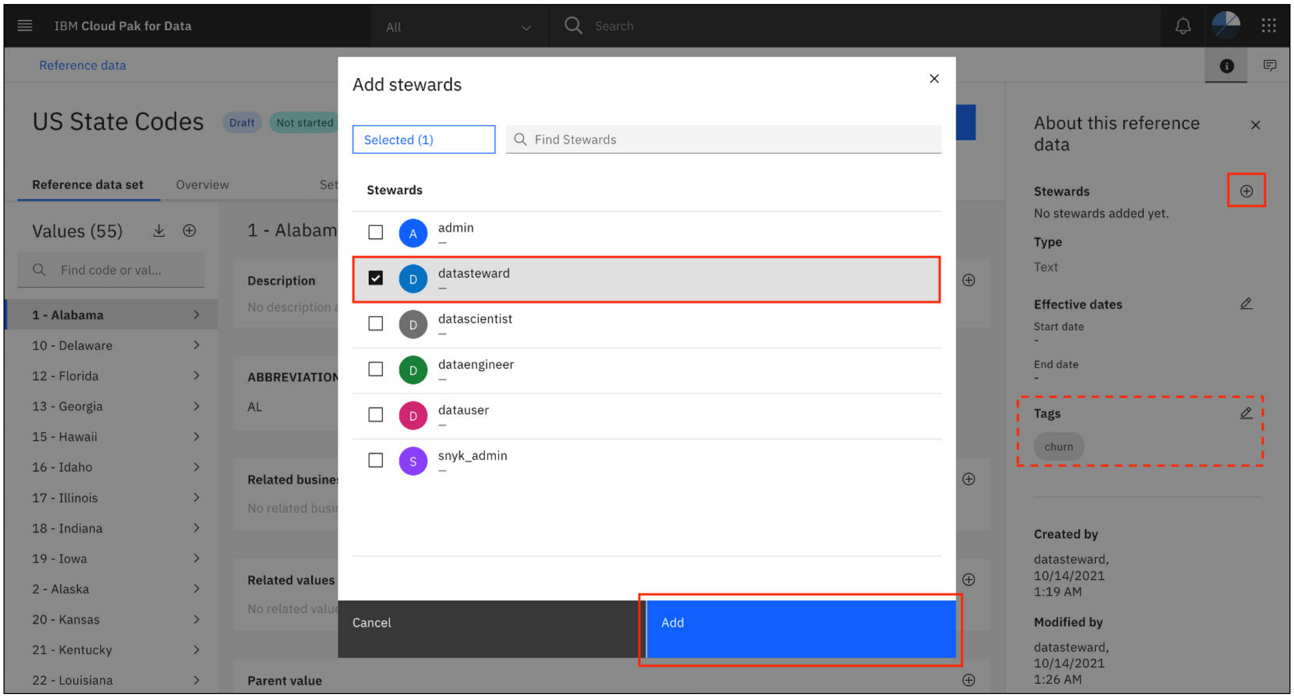

Figure 5-22 Adding a steward

7. Click **Publish** and then, click **Publish** again to publish this reference data set.

8. Repeat these steps to import the `USCountyCodes.csv` file into a new US County Codes reference data set. Select the target column Code for the FIPS column and Value for the Name column. As in with the Abbreviation column in the USStateCodes reference data set, you create a custom column (STATE) for the State column in the USCountyCodes reference data set.

> **Note:** Importing the `USCountyCodes.csv` file takes a few minutes because it is a larger data set (see Figure 5-23 on page 290).

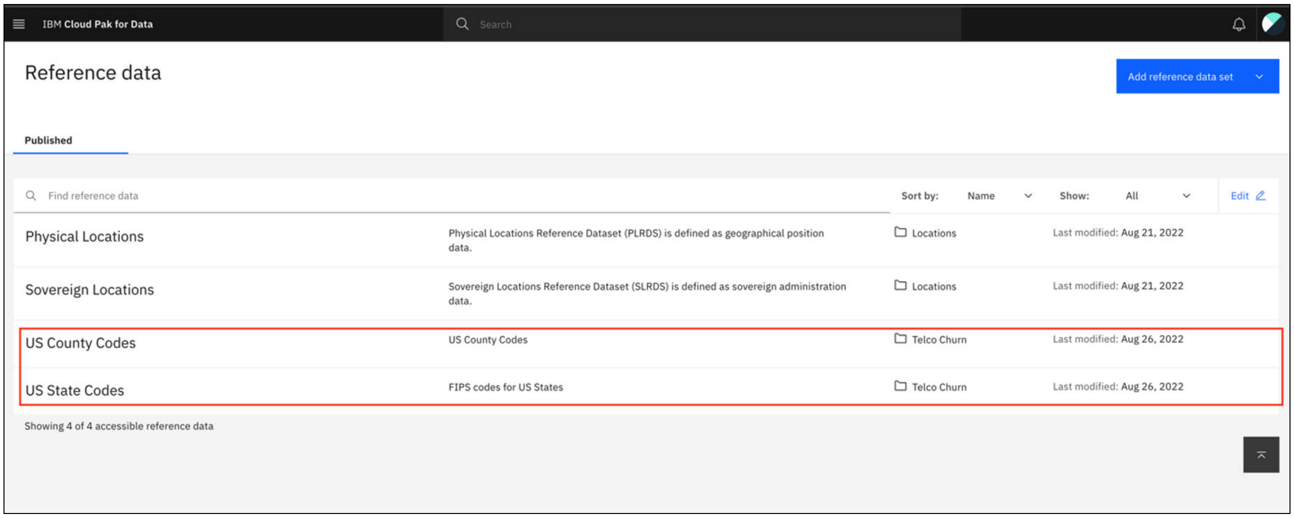

Figure 5-23 Add Second reference data set

9. After it is imported, add `datasteward` as the Steward and `churn` as the tag for the reference data set. Finally, click **Publish**.

When returning to the Reference data sets (click the Reference data breadcrumb), you see the US County Codes and US State both US County Codes are available as reference data sets.

Mapping related values

Related values are used to map reference data values to other reference data values within the same reference data set or in different reference data sets.

In this section, you see how to map a county to a US state, which shows how to map one reference data set (in this case the US County Codes) to another reference data set (US State Codes).

Complete the following steps:

1. In the US County Codes Reference data set, go to the first reference data value. Click the **+** sign that is next to **Related values** (see Figure 5-24).

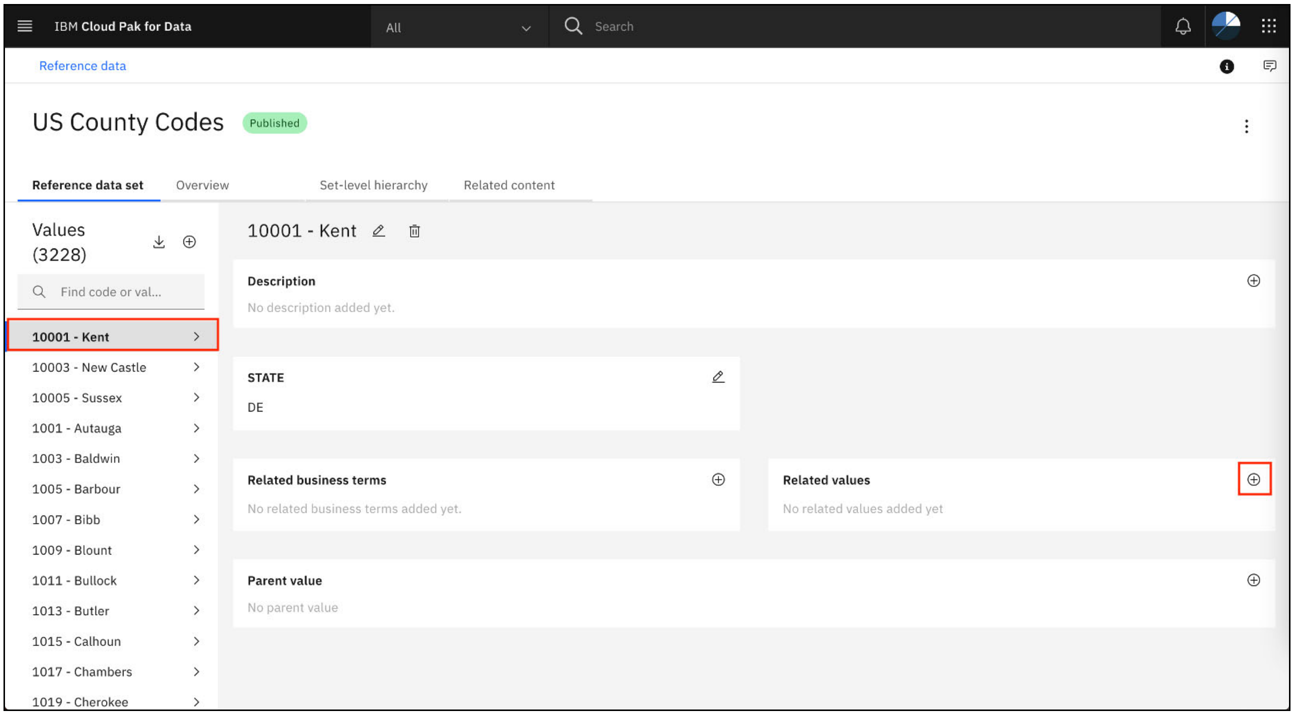

Figure 5-24 Related values

2. Select the **US State Codes** reference data set and then, click **Next**.
3. Select **One-to-one** mapping. One-to-one mapping means that one value in the reference data set (US County codes) can map to only one value in the related reference data set (US State codes).
4. Find the value in the related dataset. You can also use the search bar to filter results. Select **10 - Delaware** as the US State to map Kent County to and then, click **Add**.

You can add related values for all the other US County Codes reference data values in a similar manner, if wanted. In practice, you use the IBM Watson Knowledge Catalog API to populate the related values automatically.

Note: Although you mapped the related values, these changes are still in draft mode. They are published until you publish the reference data set, as described next.

Setting reference data hierarchy

Complete the following steps:

1. Select the **Set-level hierarchy** tab in the US County Codes reference data set to create a child-parent relation between the US County Codes and US State Codes reference data sets. Click **Add parent set** (see Figure 5-25).

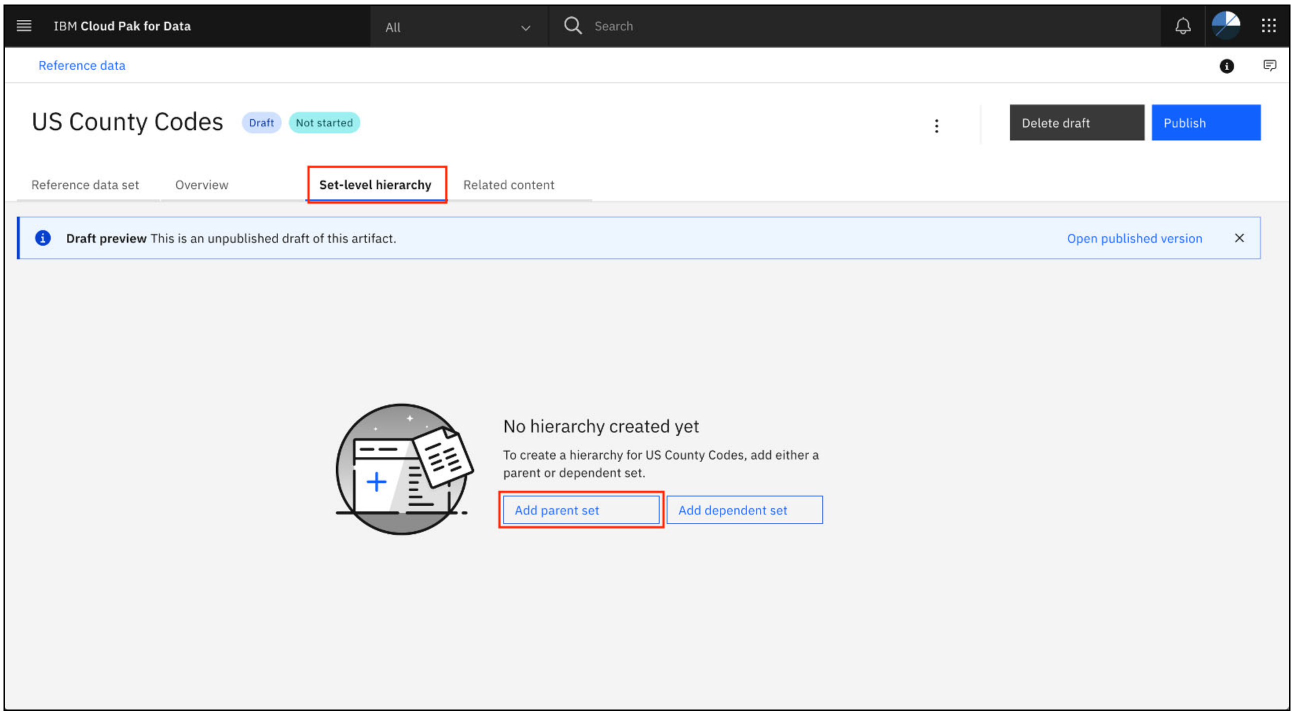

Figure 5-25 Mapping parent data set

2. Select the **US State Codes** reference data set by clicking the check box next to it and then, click **Save**. The child-parent relationship is now set between US County Codes and US State Codes.
3. Click **Publish** and then, click **Publish** again in the pop-up window to publish the US County Codes reference data set.

Reference data sets are now created that the analytics team needs to follow as standards. They can export these reference data sets as a `.csv` file or use the IBM Watson Knowledge Catalog APIs to connect to and use the reference data values.

Data classes

Data classes are governance artifacts that are used to automatically profile catalog assets during quick scan and discovery processes. They can be created by using a regular expression, Java class, column name, data type, list of values, or reference data set.

Automatically finding key data elements and understanding where PII (Personal Identifiable Information) is stored saves data stewards much time. Also, data quality analysis uses data classes to find anomalies that do not follow that regular expression, Java class logic, list of values, or reference data set codes.

In the following use case, you create a data class by using the US State Codes reference data set. By doing so, data stewards can profile data assets to find columns that use the values in that reference data set and find data quality issues where fields in those columns do not completely follow that standard.

In this way, the data stewards that are curating a catalog for Telco Churn analysis can ensure that all columns that are referencing US State Codes are following the standard and fix any that do not.

Complete the following steps:

1. From the navigation menu in the upper-left corner, expand **Governance** and then, click **Data classes** (see Figure 5-26).

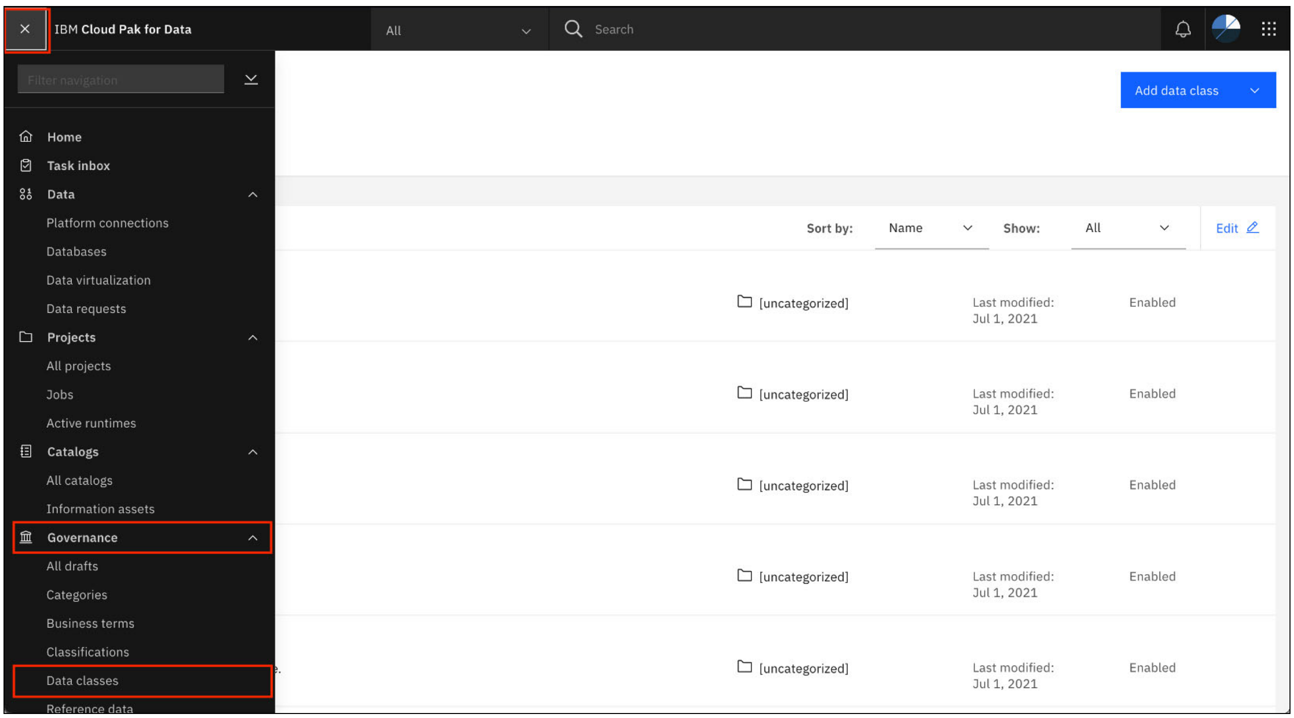

Figure 5-26 Selecting Governance Data classes

2. IBM Watson Knowledge Catalog includes over 120 data classes, such as Address Line 1, State Driver's License (for multiple states), Credit Card Number, and US Social Security Number. Create a data class by expanding the **Add data class** drop-down menu and selecting **New data class** (see Figure 5-27).

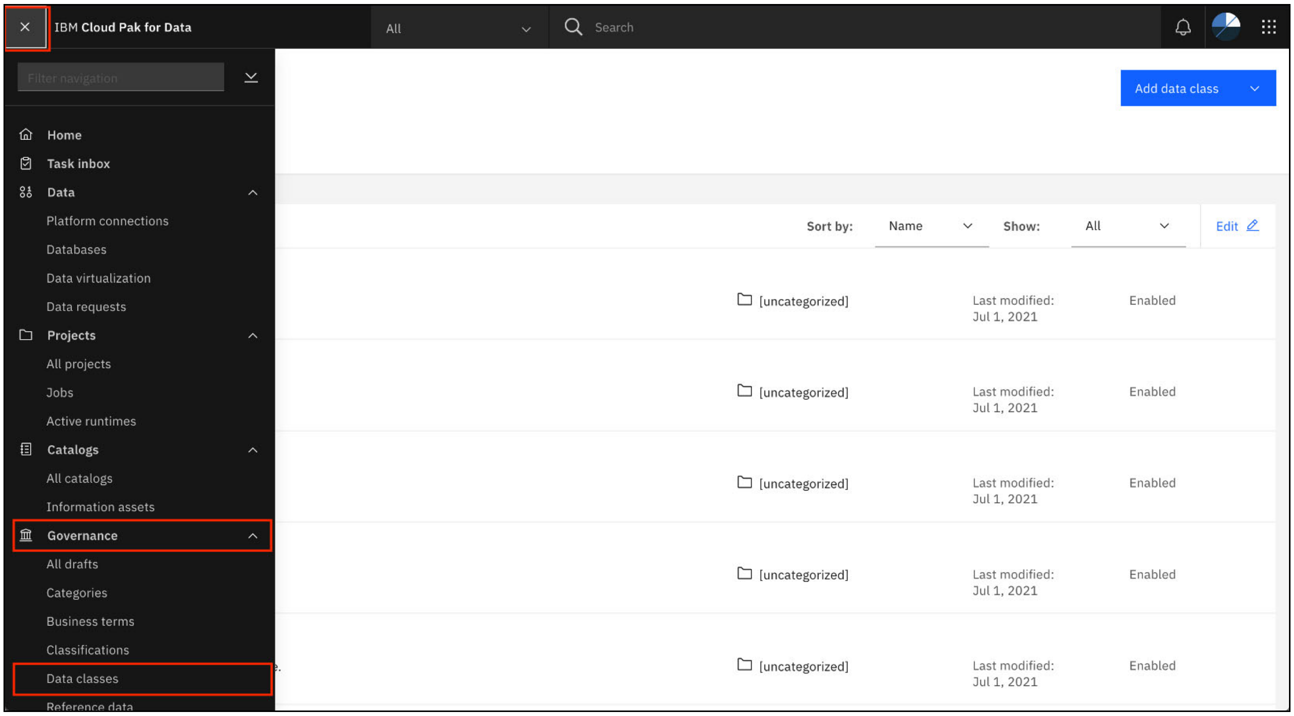

Figure 5-27 Add Data class

3. Enter a Name (US State Code) and an optional Description (Data class to identify US State Codes) for the new data class. Click **Change** and select the **Telco Churn** category to associate this data class with this category. Click **Save as draft** (see Figure 5-28).

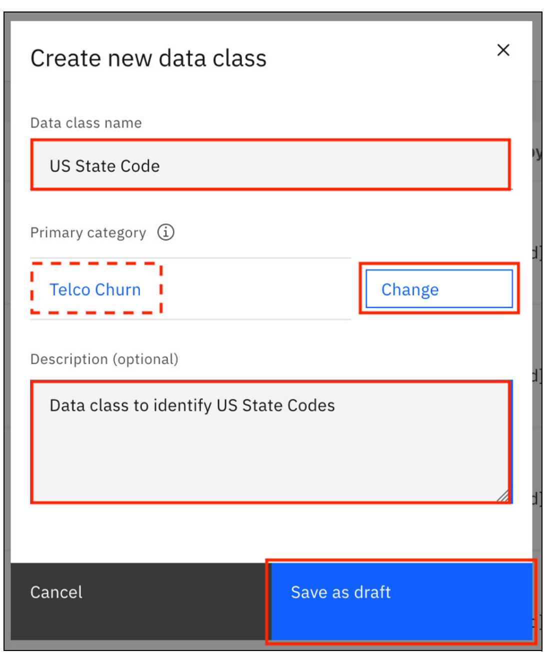

Figure 5-28 Create new data class window

4. Click the **+** that is next to **Matching method** to select how to match this data class (see Figure 5-29).

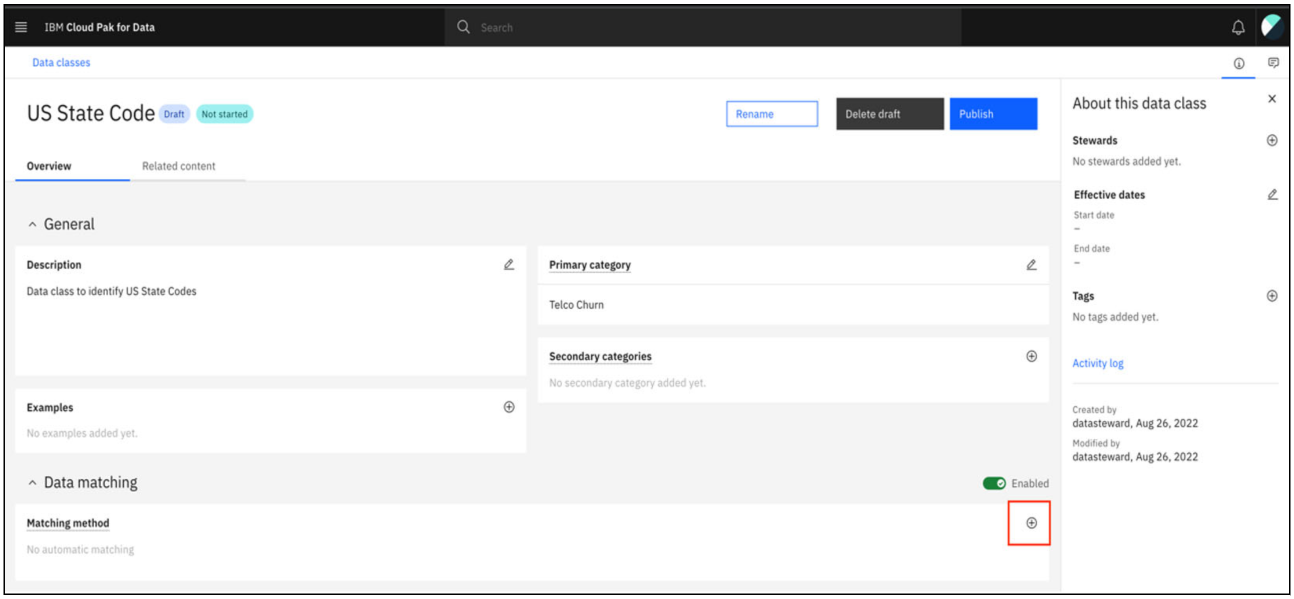

Figure 5-29 Matching to Reference Data

5. From the available matching methods, select **Match to reference data**. Click **Next**. The following matching methods are available:
   - No automatic matching: The data class can be manually selected to assign to a column, but the system does not automatically profile any columns as matching that data class.
   - Match to list of valid values: You can provide a list of values to which to match. It is recommended to save these values as reference data sets instead, but this option is good if you prefer not to have so much formal management of that list.
   - Match to reference data: While profiling columns, the system evaluates against values in a reference data set.
   - Match to criteria in a regular expression: A regular expression is used to determine whether values match a data class.
   - Match to criteria in deployed Java class: The logic that is specified in a Java class determines whether each value of a database column or the whole database column belongs to the data class.
   - Other matching criteria: Matching is based on criteria about the name or the data type of the column only. No other criteria is used to evaluate the values of the column.
6. Select the **US State Codes** reference data set. Keep the default percentage match threshold of **80%**. This threshold means that, if 80% of the values in a column match the values in that reference data set, that column is classified as the US State Code data class. Click **Next**.
7. IBM Watson Knowledge Catalog supports more matching criteria that is based on column name or data types. However, for now, you can leave those criteria empty (default values) and use only the reference data set to determine the data class. Click **Next**.
8. Change the priority of this rule that is related to other matching rules, which provides more granular control over to which data class a column is mapped.

9. Add the **`datasteward`** user as Steward, and add `churn tag` to the Tags. Click **Publish** and then, click **Publish** again in the pop-up window to publish the data class (see Figure 5-30).

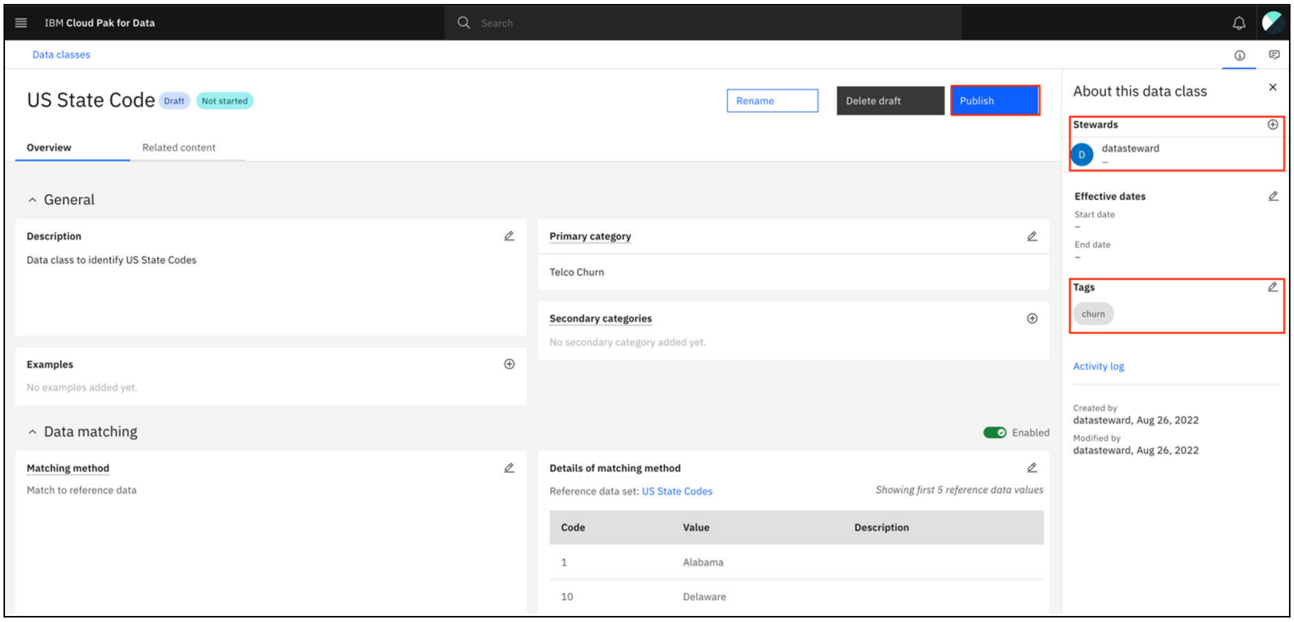

Figure 5-30 Adding datasteward data class

Summary

In this tutorial, you learned how to set up data governance for your enterprise. The enterprise teams can now efficiently apply their data analysis and data science models for improved Telco Customer Churn prediction while meeting the governance and compliance requirements for the enterprise.

You also learned how to us IBM Watson Knowledge Catalog to perform the following tasks:

- Create categories to define logical structure of governance artifacts.
- Configure workflows for governance artifact requests and assign user permissions.
- Create governance artifacts, such as policies, rules, business terms, data classes, and reference data sets. You learned how to create these artifacts manually by using the GUI and by importing CSV files.
- Update workflow requests to publish and deliver requests as you perform the various tasks that are associated with the governance artifact requests.

5.4.4 Preparing and understanding data

In this section, we discuss how to prepare and understand the data.

Creating a project

To build a customer churn prediction asset, the data engineer and the data scientist must collaborate to prepare the data and then build a model. The project in IBM Cloud Pak for Data enables people to work together on data, scripts, notebooks, and other assets.

Complete the following steps top create a project:

1. Log in to Cloud Pak for Data as the `dslead` user, who has the permission to create projects.
2. Select **All projects** by clicking the navigation icon in the upper left, or by selecting **All projects** from the main page.
3. Click **New project** to create a project (see Figure 5-31).

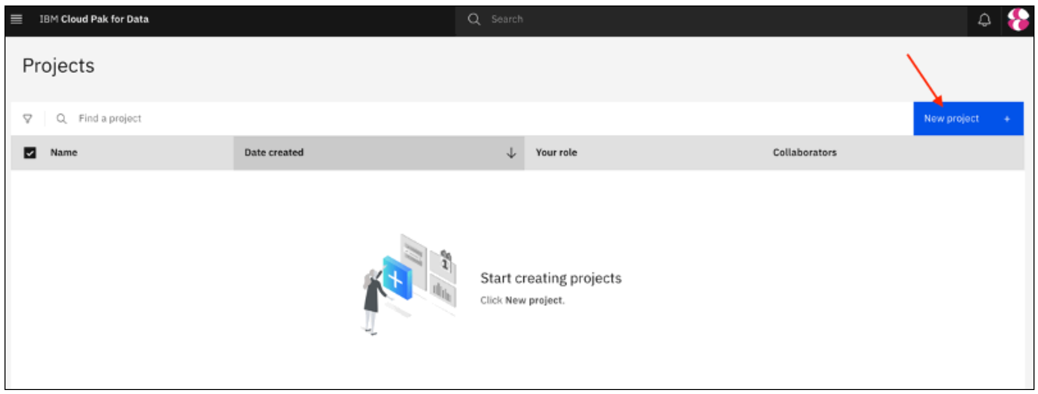

Figure 5-31 Create project

4. Select the **Create an empty project** option, as shown in Figure 5-32).

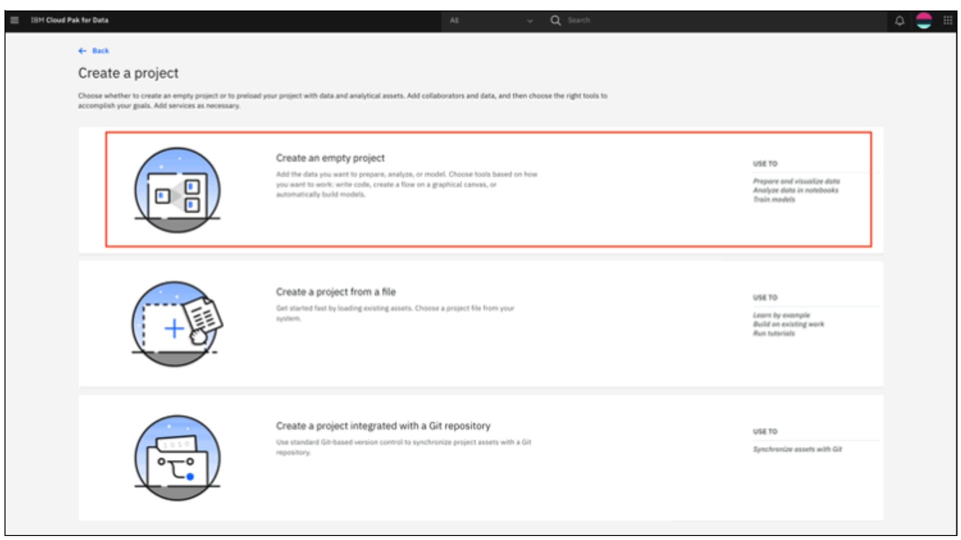

Figure 5-32 Empty project

5. On the New project page, enter a Name (for example, Customer Churn Prediction) and an optional Description, such as “Data Science Project for training AI models for customer churn prediction”. Also, select the check box that is next to **Log all project activities** to help the lead data scientist and management track project activities (see Figure 5-33).

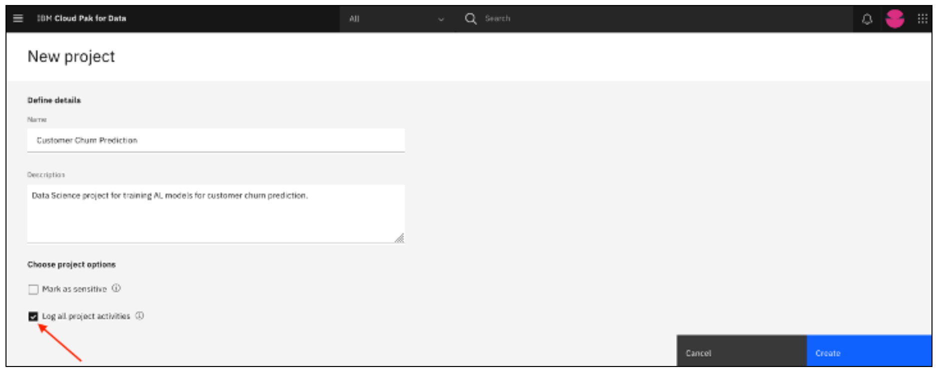

Figure 5-33 Track project activities

6. After the project is created, click the **Manage** tab to grant access to other collaborators. In our example, we provide Editor access to the `datascientist` and `dataengineer` users (see Figure 5-34).

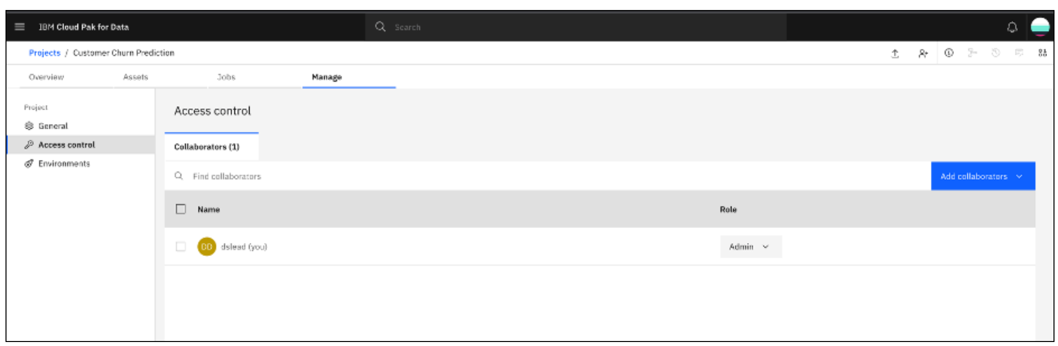

Figure 5-34 Access control

7. Click **Add collaborators** and then, click **Add users**. To find the users that you want to add as collaborators, enter `data` in the search field, which then display all users that include “`data`” in their names (see Figure 5-35).

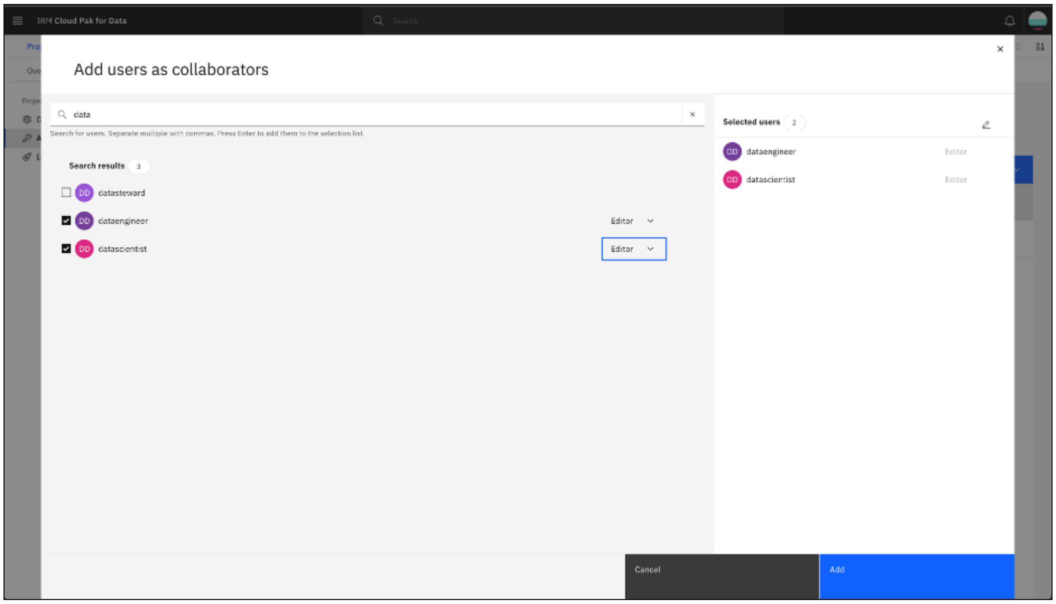

Figure 5-35 Add users as collaborators

8. Select **dataengineer** and **datascientist** and assign the Editor role to the users. Then, click **Add** to finish granting project access. You see the collaborators when you return to the project (see Figure 5-36).

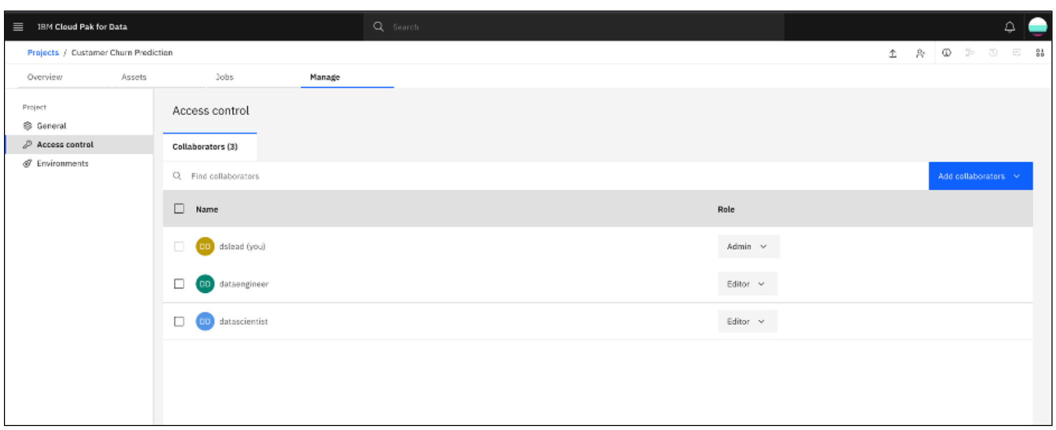

Figure 5-36 List of collaborators

Preparing data by using Data Refinery

The first task that you run is preparing the data by using *Data Refinery*, which is a visual UI-based tool that enables users to interactively discover, cleanse, and transform data with over 100 built-in operations.

Complete the following steps:

1. Log in to Cloud Pak for Data as the `dataengineer` user.
2. Select **All projects** by clicking the navigation icon in the upper left, or by selecting **All projects** from the main page.

3. The `dataengineer` user searches for relevant data assets that can help them with training this AI model for customer churn prediction. In practice, a `dataengineer` user searches for relevant assets in one or more catalogs or by using global search. However, for this exercise, we use the three data sets that were referenced in “Creating reference data sets” on page 285. Click the **Find and add data** icon (see Figure 5-37).

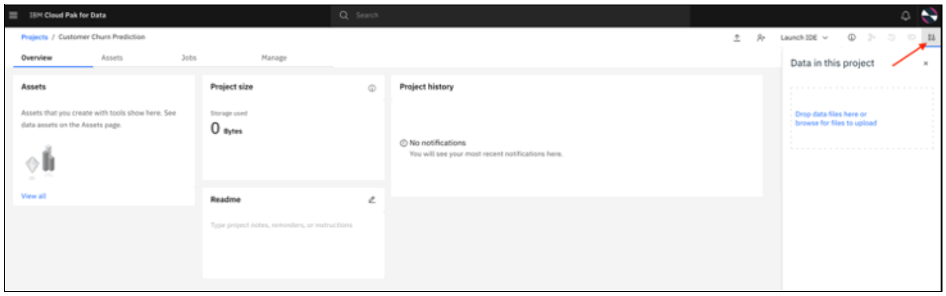

Figure 5-37 Finding and adding data

4. Drop the `customer_personal_info_simplified.csv` from the downloaded `datasets` folder into the Drop files here box or click **Browse** to search for the file to upload. Wait for the file to upload to the project. Repeat this step for the other `.csv` files: `customer_data_transactions.csv` and `customer_churn_labels.csv`, see Figure 5-38.

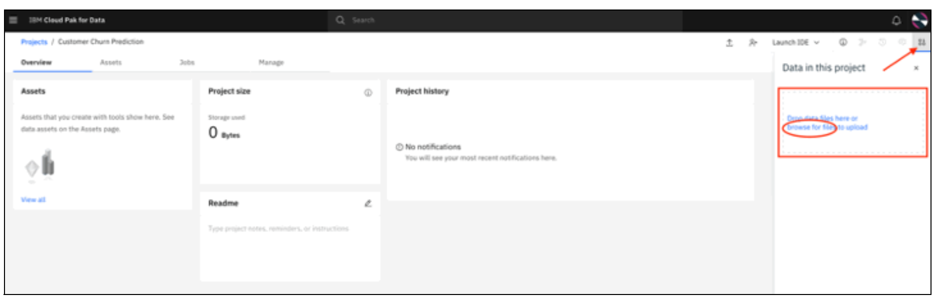

Figure 5-38 Adding data files

5. Verify that the three `.csv` files are uploaded to your project by clicking the **Assets** tab. You should see all three files and that the only Asset type in the project at this time is Data. As you proceed through next steps, you add other asset types to the project to perform various tasks. (see Figure 5-39).

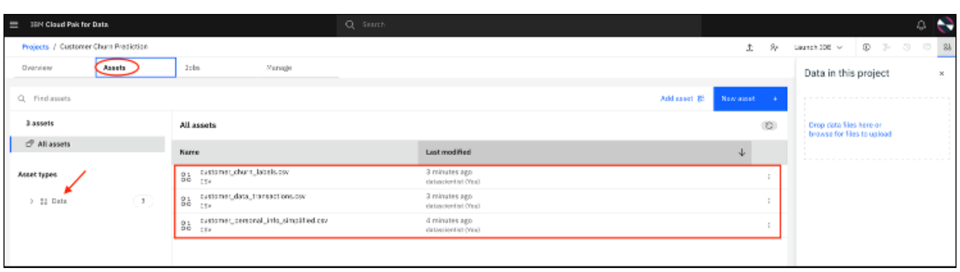

Figure 5-39 Project assets

6. Shape the data to get it ready to be used for training ML models. Cloud Pak for Data supports multiple approaches for data wrangling and transformation. In this lab, you use Data Refinery to cleanse and shape the data by using a graphical flow editor and create a joined data set of the Customer Data asset and the labeled churn data set.

   Click **New asset** + (see Figure 5-40).

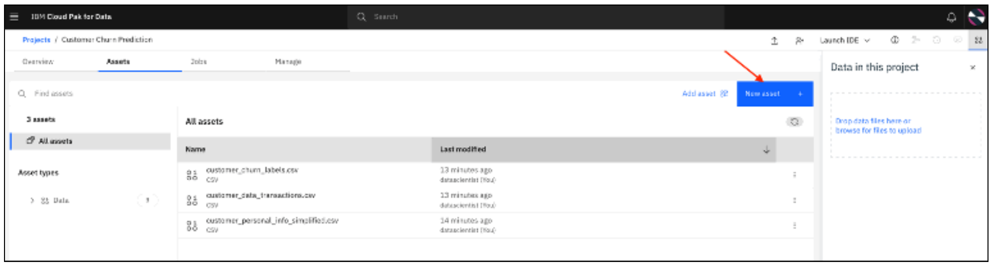

Figure 5-40 New asset

7. Scroll down and click **Data Refinery** tile. You also can filter the Tool type to Graphical builders for quicker access to such tools (see Figure 5-41).

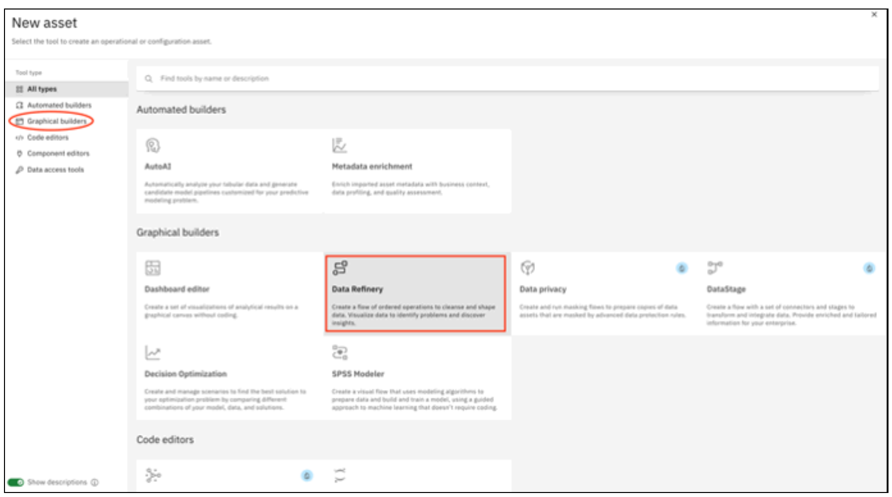

Figure 5-41 List of asset types

> **Note:** Alternatively, you can access Data Refinery by clicking the open and close list of options menu (three vertical dots) that is next to a data asset in your project and then, selecting **Refine**.

8. Select **Data asset** and then, select the `customer_data_transactions.csv` data set. Click **Select**. (see Figure 5-42).

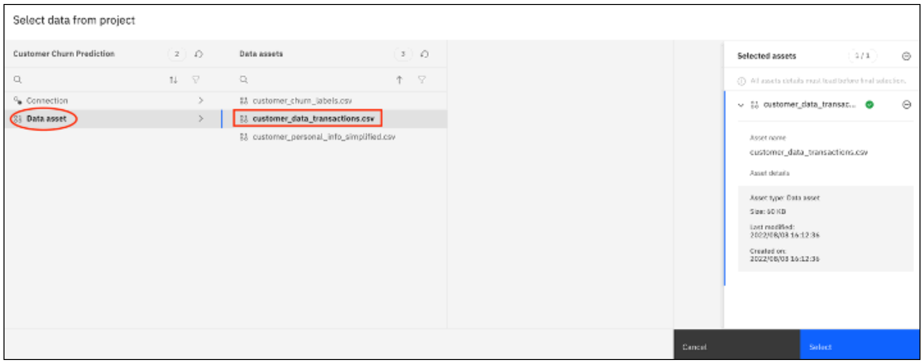

Figure 5-42 Selecting data asset

9. The dataset is then loaded into the Data Refinery. In the upper right corner of the window, you see a status message that indicates that the data set is being loaded and only the first 50 rows are shown in the Data tab. After some time, the Profile and Visualizations tabs appear (see Figure 5-43).

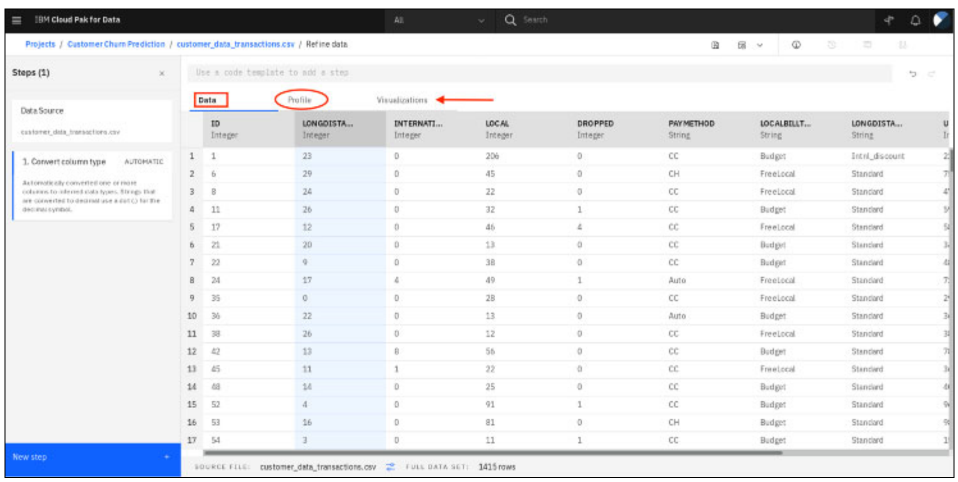

Figure 5-43 Visualize data

The following tabs are available:

– The Data tab displays the data and enables you to apply several common operations to cleanse and shape the data in a graphical editor. It also supports deploying R library operations, functions, and logical operators by using the command line.

– The Profile tab shows useful summary statistics, including a histogram of each of the data fields in the entire data set. This information is useful to understand the statistical distribution of the data and skew that might exist.

– The Visualizations tab provides over 20 customizable charts to provide perspective and insights into the data. (see Figure 5-43).

10. Change the ID column type from `Integer` to `String`. This change is needed because in the next step when you apply a join of this data and the other data sets, the column types must match. To do so, click the options menu (three vertical dots) that is next to the ID column, select **CONVERT COLUMN TYPE** and then, select **String** type (see Figure 5-44).

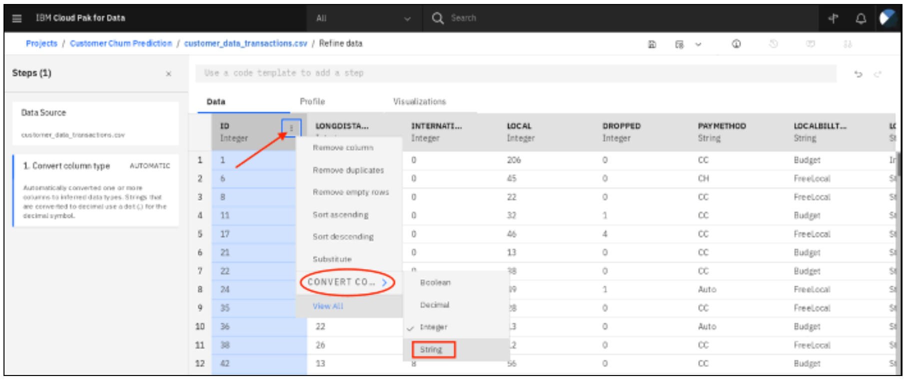

Figure 5-44 Converting column type

11. Add a step to join this data set with another data set to capture more features that can affect the likelihood of a customer to churn:

a. Click **New Step**, which opens the operations column.

b. Scroll down to find the Join operation and click **Join**. You also cab enter `Join` in the Search operations field and it filter the list of operations to find "Join" (see Figure 5-45).

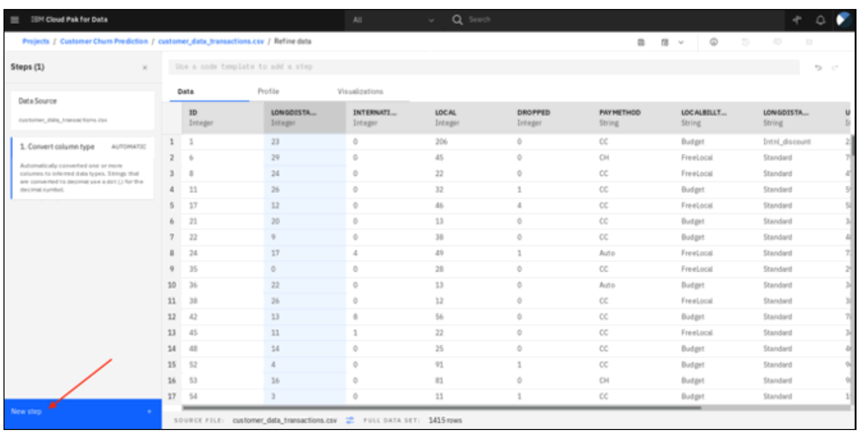

Figure 5-45 New refinery step

c. In the Join operations window, keep the type of join as “Left join” and then, click **Add data set** (see Figure 5-46).

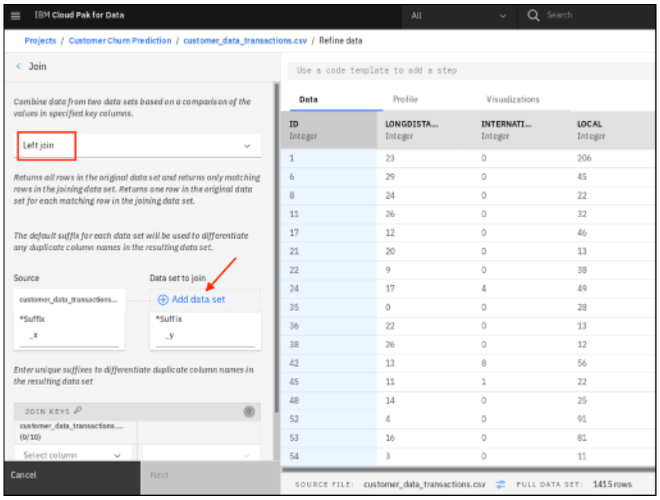

Figure 5-46 Add data set

12.On the Data set page, click **Data asset**, select the following data set:

```
customer_personal_info_simplified.csv
```

Then, click **Apply**. If you are not familiar with the Left Join operation, Data Refinery provides an explanation of the operation. Specifically, a Left Join returns all rows in the original data set and returns only the matching rows in the joining data set.

Data Refinery also supports creating other joins, such as Right Join and Inner Join (see Figure 5-47).

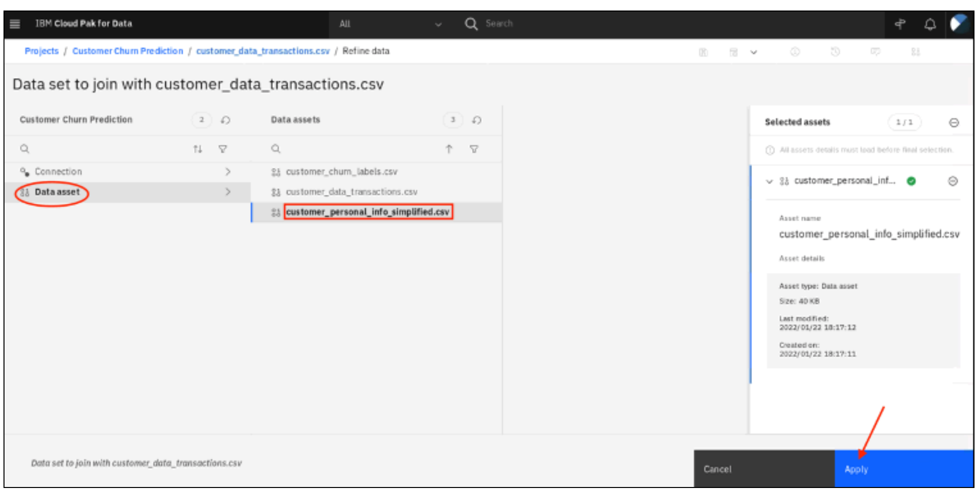

Figure 5-47 Select data set

13. In the Join operation window, click **Select column** to specify `ID` as the field to use for joining the two data sets. Then, click **Next** (see Figure 5-48).

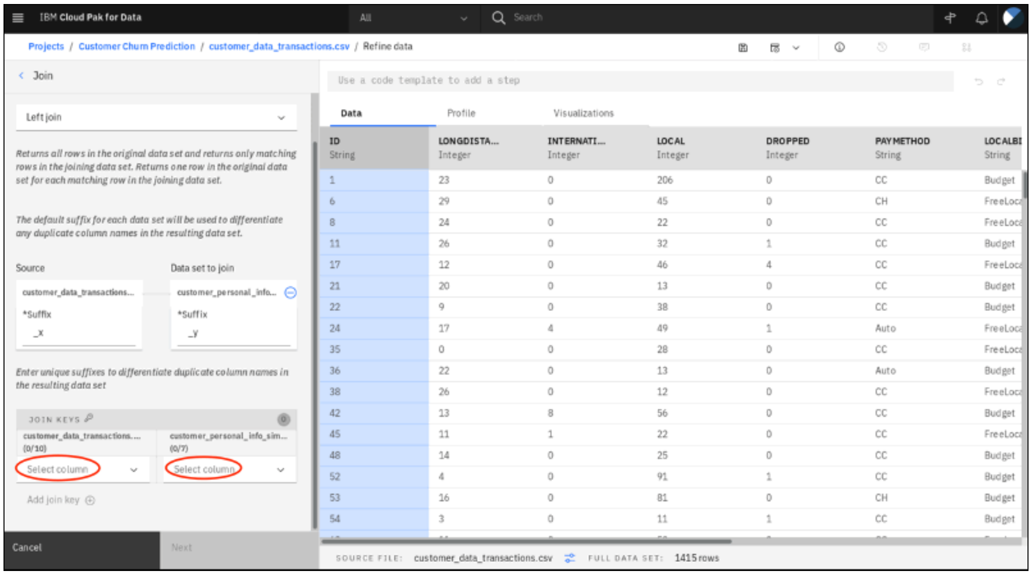

Figure 5-48   Join columns

14. In the next window, all the fields from both files as a result of the join operation are shown. Now, you can remove fields that you do not want to include in the final data set. For this lab, keep all the fields selected and click **Apply**.

15. Repeat steps 11 - 14) to apply a join operation on the resulting data set and the `customer_churn_labels.csv` dataset. The join field is `ID` and the data set to join is `customer_churn_labels.csv`, see Figure 5-49.

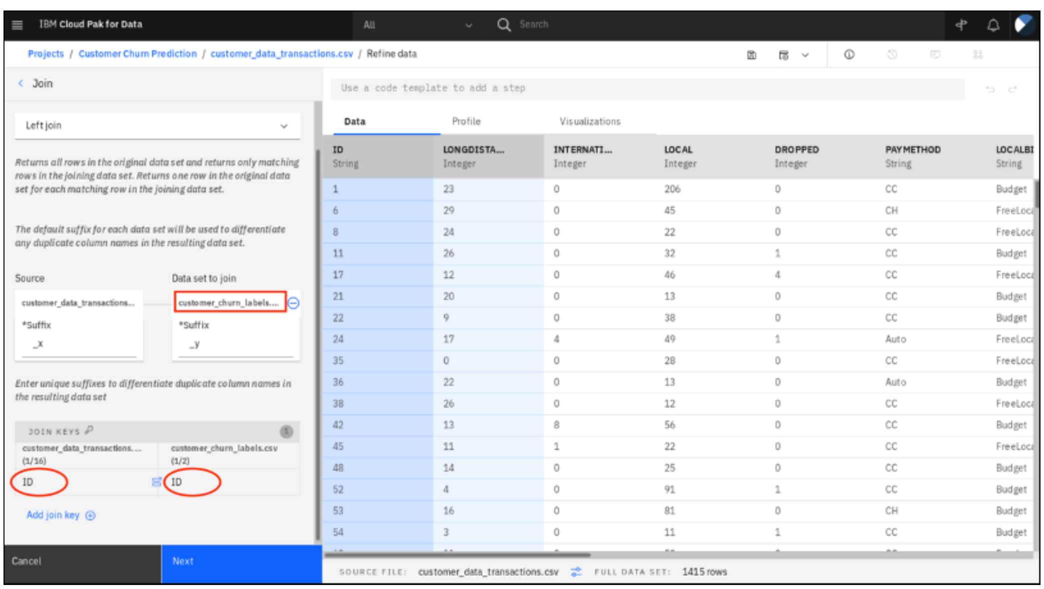

Figure 5-49   Add new data set and join

> **Note:** The Data Refinery flow was augmented with all of the operations that were run. As you perform more operations to shape the data, they are added to the Data Refinery flow. For this lab, we perform only the Join operations; however, typically you perform several other operations to transform the data and ready it for analytics insights and training machine learning models (see Figure 5-50).

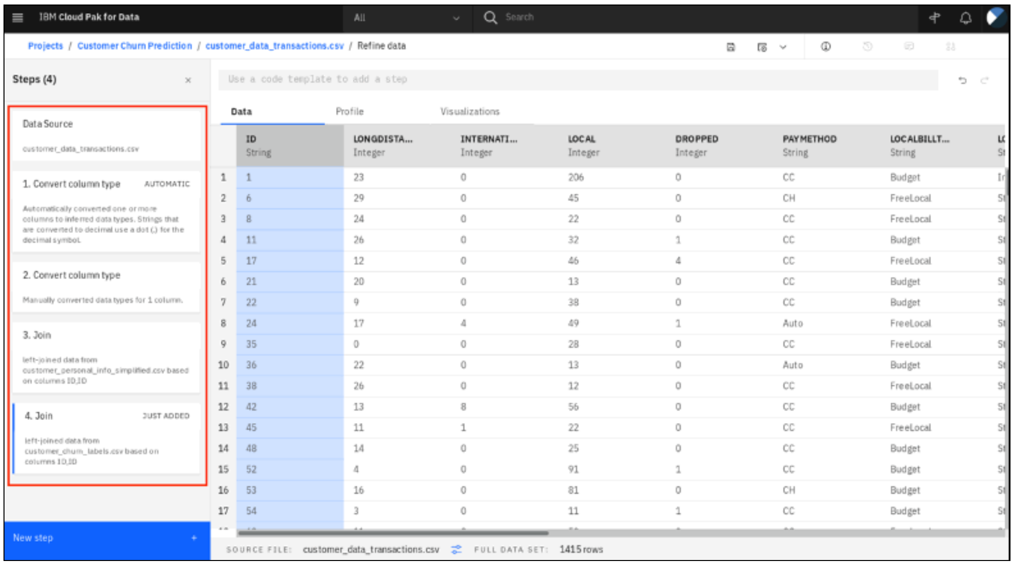

Figure 5-50 List of steps

In practice, the data typically requires several more operations to cleanse by removing nulls, filtering rows with missing data, aggregating data across fields, or applying several different operations.

In this lab, the dataset we use is ready and the only operations you apply is to join the customer data (which was a join of customer personal information and transaction data) and labeled `churn data set`.

16.After all the operations are applied to transform the data, click **Save and create a job** at the top of the window to apply this Data Refinery flow against the complete data set (see Figure 5-51).

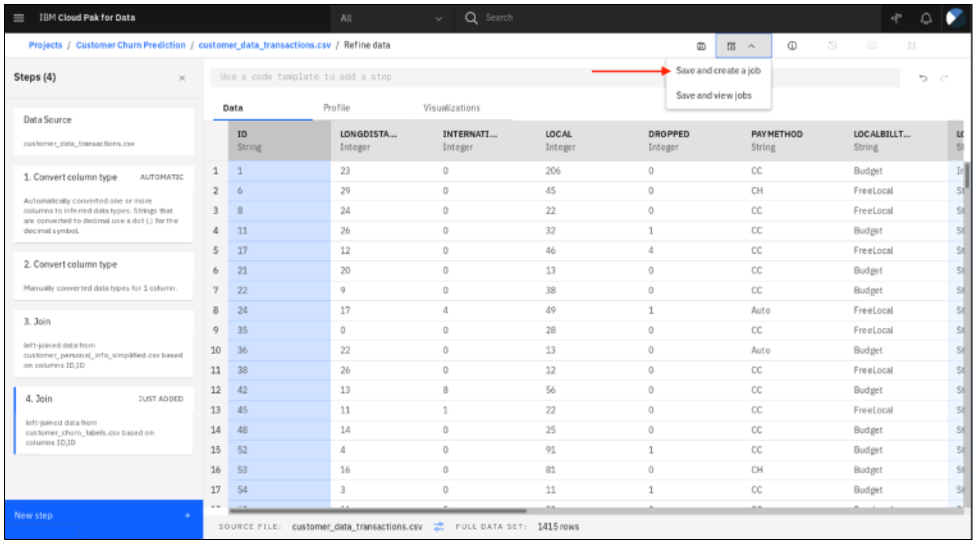

Figure 5-51 Saving and creating a job

17.Enter a name for the job, such as `drjob`, add an optional description (such as, "simple data refinery job to join multiple tables to get ready for training AI models for churn prediction") and click **Next** (see Figure 5-52).

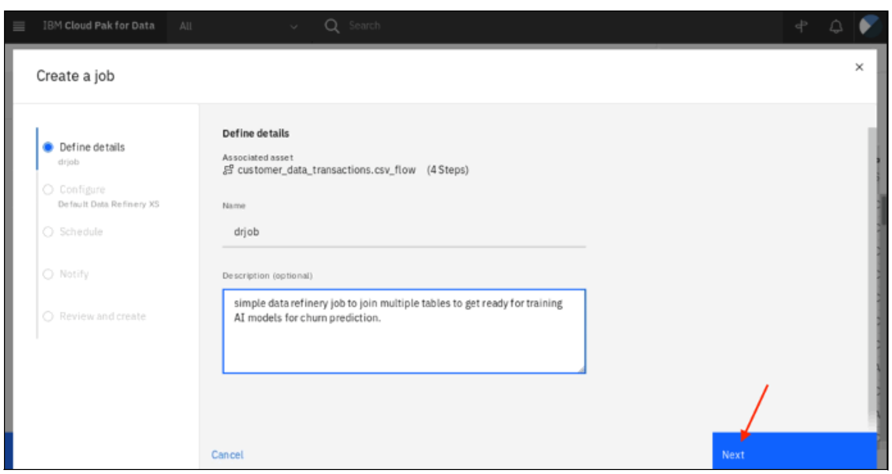

Figure 5-52 Define details

18. On the **Configure** tab, review the Environment and keep the default selection and then, click **Next**. For jobs that require more resources, you can select a larger environment to run the job (see Figure 5-53).

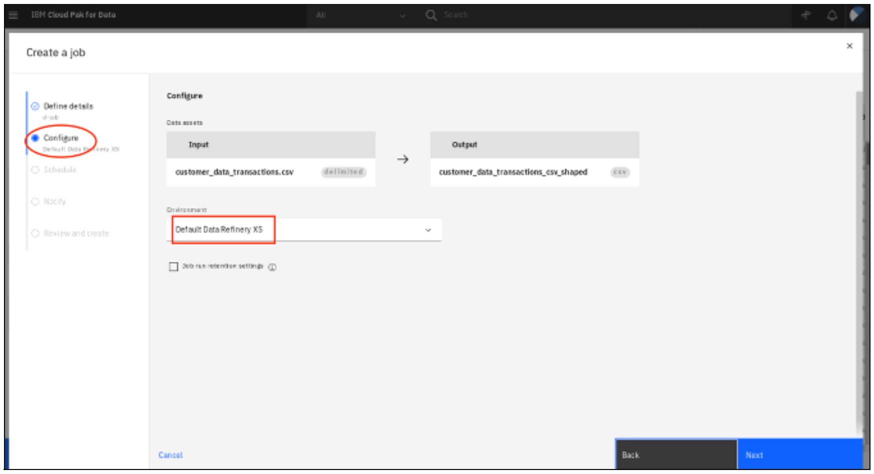

Figure 5-53 Environment

19. On the **Schedule** tab, keep the Schedule slider set to off (default), and click **Next**. In this lab, we do not need to run the Data Refinery job on a specific schedule; instead. we manually run it as needed, which is why we used the default setting (see Figure 5-54).

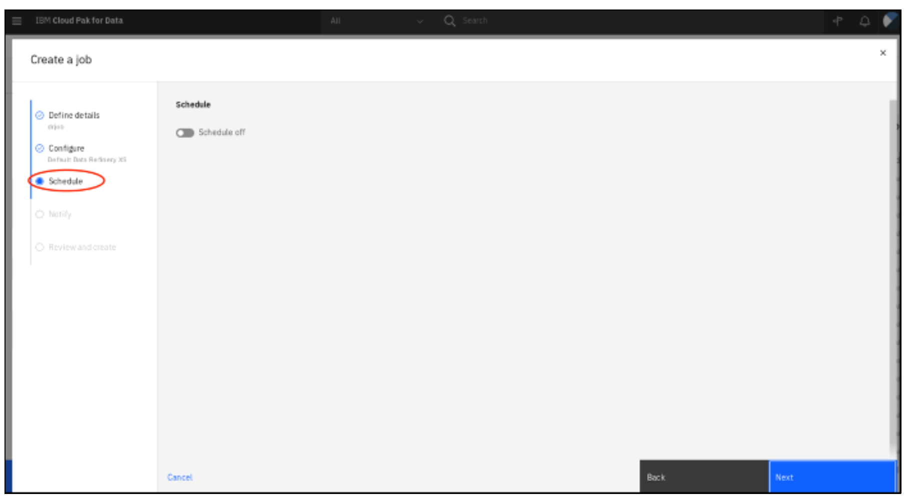

Figure 5-54 Schedule

20.On the Notify tab, keep the Notification off as default. Click **Next** (see Figure 5-55).

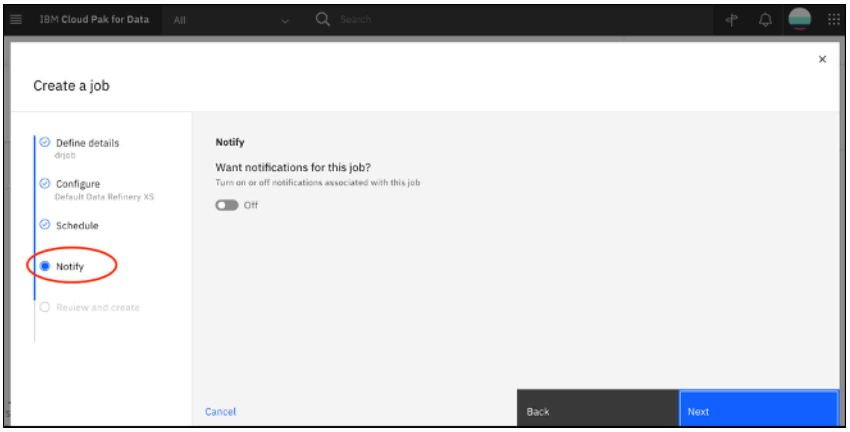

Figure 5-55 Notify

21.On the **Review and create** tab, review the job details and click **Create and run** (see Figure 5-56).

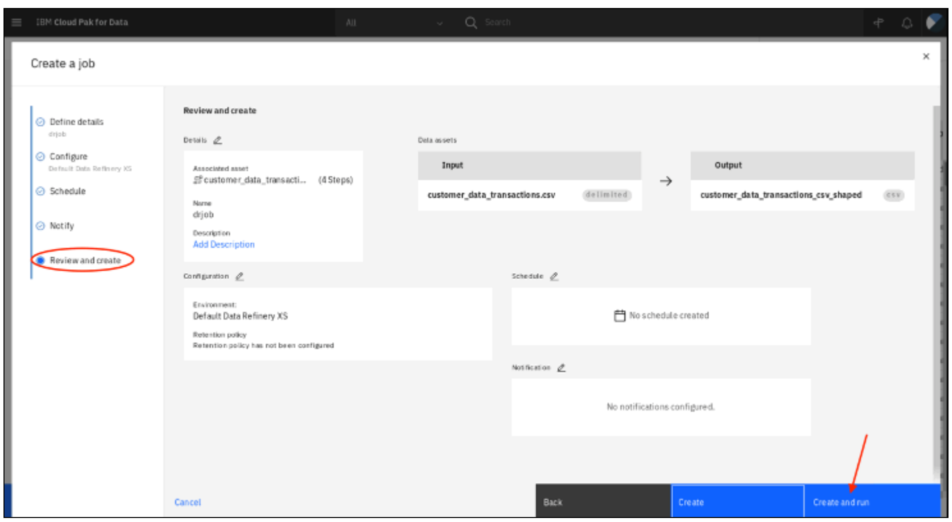

Figure 5-56 Create and run

22.Browse to the jobs view to monitor the progress by clicking the navigation menu and selecting **Jobs** (see Figure 5-57).

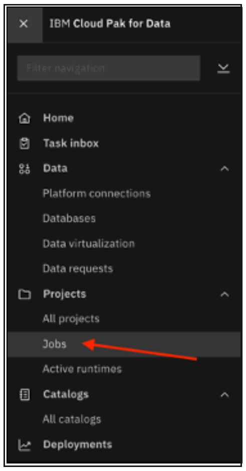

Figure 5-57 Select jobs

23.On the Jobs page, filter the view by selecting whether you want to review Active runs, Jobs with active runs, Jobs with finished runs, or Finished runs. Initially, the job appears in the Jobs with active runs view and when it completes, the job appears in the Jobs with finished runs view (see Figure 5-58).

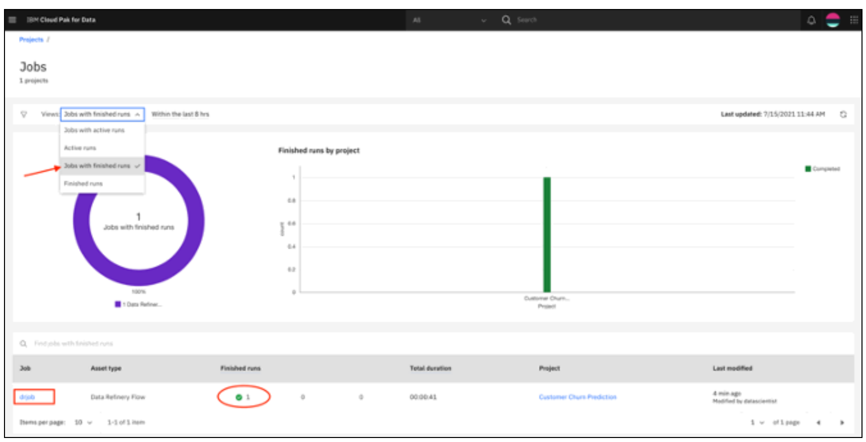

Figure 5-58 List of jobs

24. Browse back to the project and click the **Assets** tab. The `customer_data_transactions.csv_flow` Data Refinery flow becomes a project asset (see Figure 5-59).

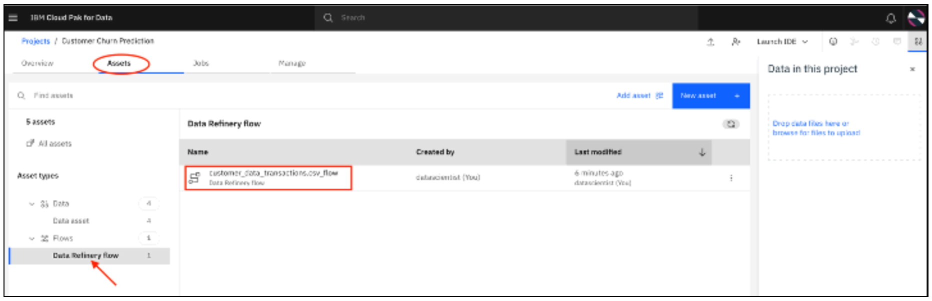

Figure 5-59 Browsing to the project and clicking the Assets tab

The list of project assets is shown in Figure 5-60.

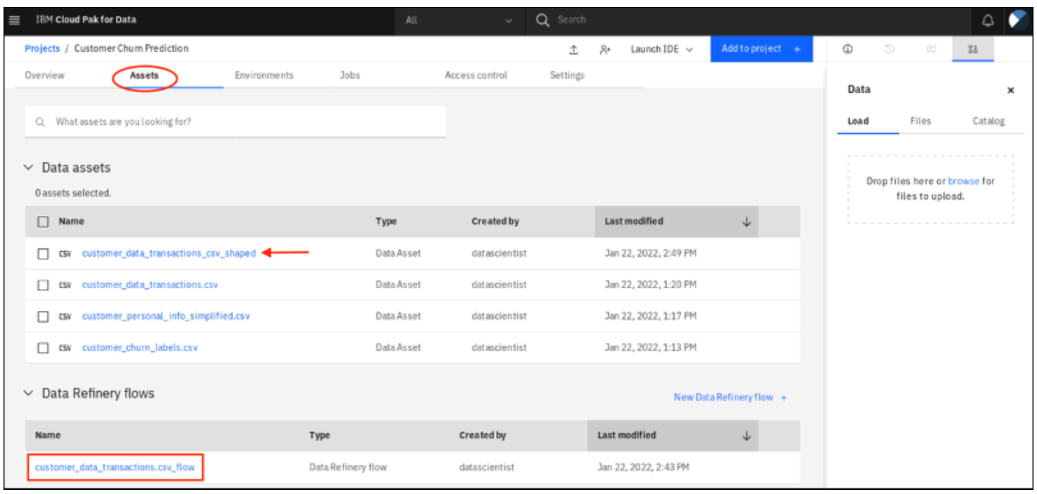

Figure 5-60 List of project assets

25. Under the Data asset type, click **Data asset** and then, click the newly created data asset, **customer_data_transactions.csv_shaped** to review the resulting data. This data asset was created by running the Data Refinery flow that joined the customer data transactions with the customer personal information and churn labels data sets (see Figure 5-61).

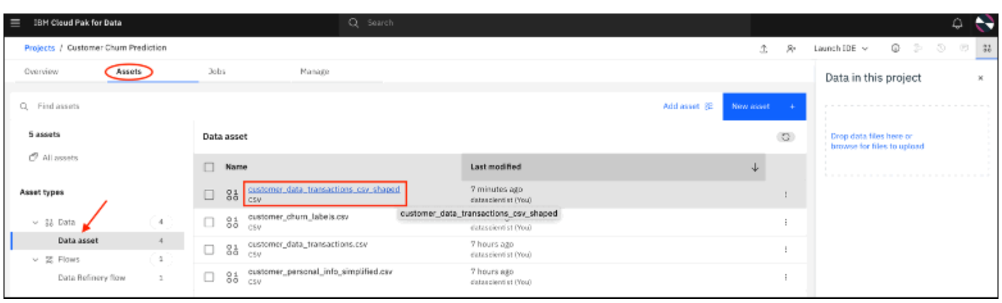

Figure 5-61 Resulting asset

26. Rename the dataset to `CUSTOMER_DATA_ready.csv` by clicking the pencil icon that is next to data asset name and then, clicking **Apply**. If the Information panel is hidden, click **View data asset information** to show that window (see Figure 5-62).

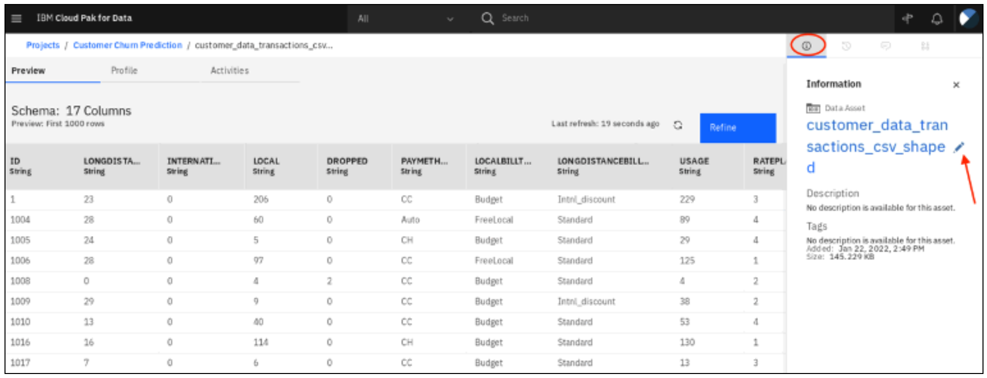

Figure 5-62 Rename dataset

Now, you collected data from various sources and used Data Refinery to shape the data by using a graphical editor. The data is ready to be used for training a machine learning model for predicting the likelihood of a customer to churn based on demographic and transaction data.

5.4.5 Building models

The previous stages focused on collecting and preparing the data, so we assume that the relevant data assets were identified, cleansed, quality-verified, and stored in data stores or catalogs, ready to be used.

You now assume the role of the `datascientist` user, who typically trains and evaluates AI models, and uses the data set that was built in the Data Refinery exercise.

If you skipped the previous exercise, upload the following file:

```
customer_data_transactions_csv_shaped.csv
```

Rename it to `CUSTOMER_DATA_ready.csv` within the project.

In this section, we demonstrate two methods that can be used to train models: by using AutoAI, and by using a Jupyter notebook. Both methods typically have different users: AutoAI is more suitable for business users who are not comfortable with coding; Jupyter notebooks are popular with data scientists.

Training AutoAI Model for churn prediction

In this section, we describe how to use AutoAI to quickly train multiple AI models for churn prediction and select the pipeline that delivers best performance. This process starts with selecting the data set on which the model is to be trained and then, setting the options for running the experiments, such as types of models and feature selection parameters.

Finally, the AutoAI process starts to build a leaderboard with the best performing models.

Complete the following steps:

1. Browse to your project and click the **Assets** tab and then, click **New asset** + (see Figure 5-63).

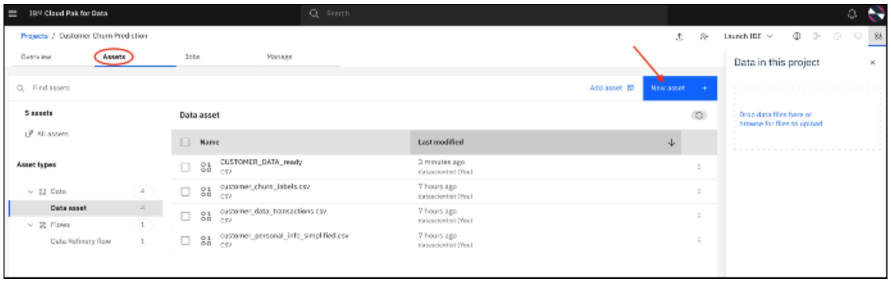

Figure 5-63 New asset

2. Scroll down and click the **AutoAI** tile. You also can filter the Tool type to `Automated builders` for quicker access to such tools (see Figure 5-64).

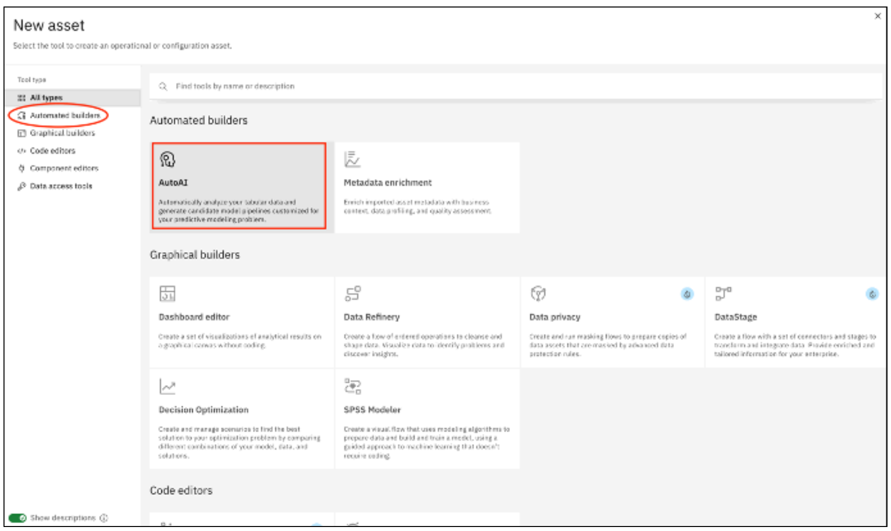

Figure 5-64 Select AutoAI

3. Enter a Name (`autoai_churn_prediction`) and an optional description ("AutoAI experiment for predicting the likelihood of a customer to churn") for your AutoAI experiment. Keep the default Compute configuration and click **Create**. You can select a different configuration if you must assign more resources for your AutoAI experiment (see Figure 5-65).

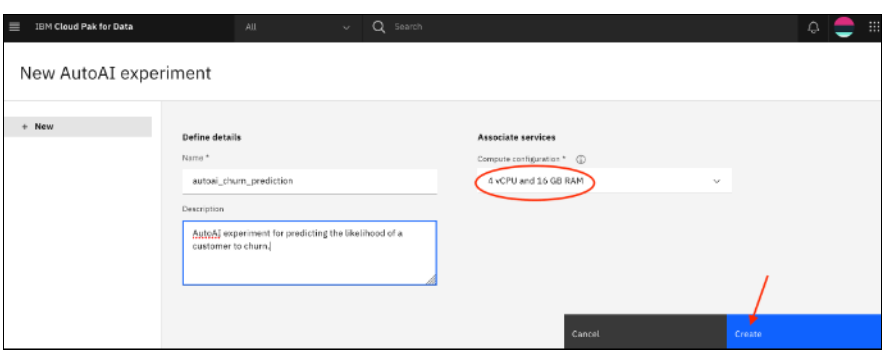

Figure 5-65 Type experiment name

4. In the AutoAI add data sources window, click **Select from project** because you use the dataset that you created earlier by using Data Refinery. You also can click **Browse** to upload data from your local machine (see Figure 5-66).

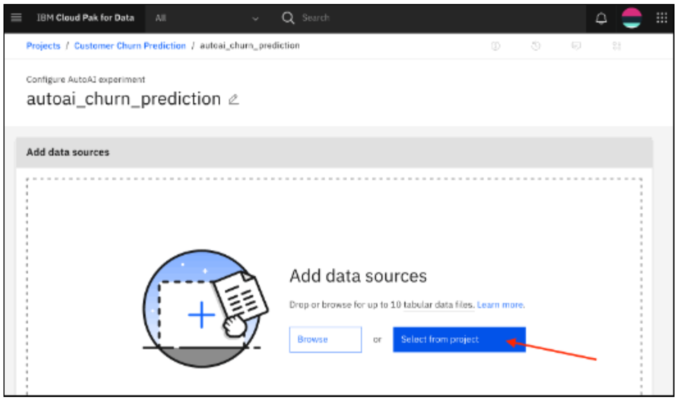

Figure 5-66 Clicking Select from project

5. Click **Data asset** and then, select the `CUSTOMER_DATA_ready.csv` dataset. Then, click **Select asset** (see Figure 5-67).

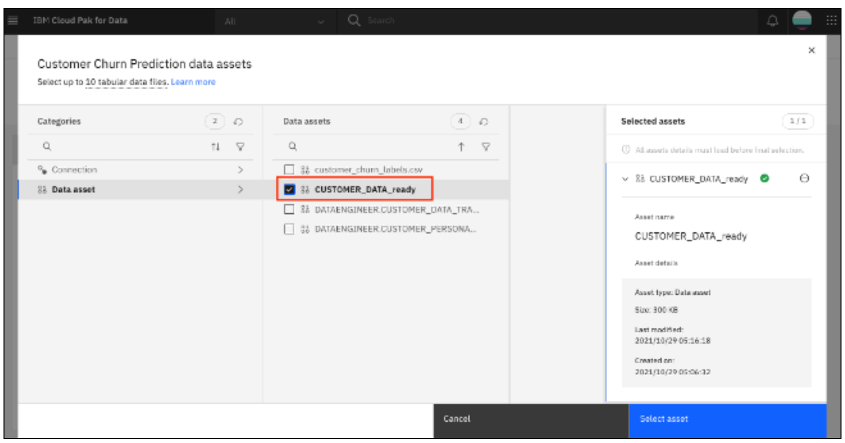

Figure 5-67 Select asset for AutoAI

6. In the next window, you see the selected dataset and you are prompted to select whether you want to create a time series forecast, which is supported by AutoAI. Click **No** because the customer churn prediction is a classification use case and not a time series forecasting use case. Then, you see the option to select which column to predict. Scroll down the list to select the **CHURN** column.

   Now, we provided the data set, indicated it is a classification use case, and selected the prediction column.

   Click **Run experiment** to begin the AutoAI run. You can click **Experiment settings** to review the default settings and change some of the configurations, if needed. Review those settings because they are informative (see Figure 5-68 on page 315).

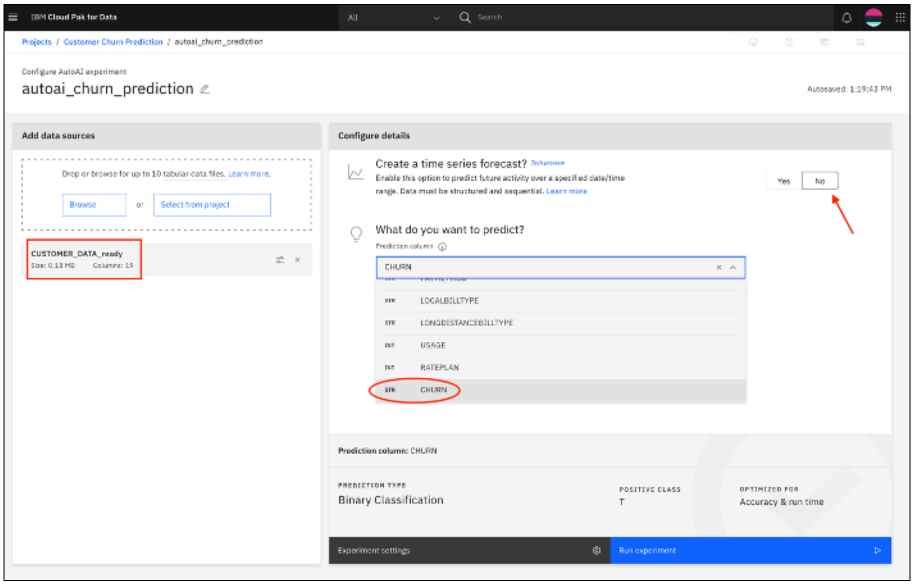

Figure 5-68 Specifying as target

7. AutoAI runs for a few minutes on this dataset and produces several pipelines (as shown in Figure 5-69) including training/test data split, data preprocessing, feature engineering, model selection, and hyperparameter optimization. You can delve into any of the pipelines to better understand feature importance, the resulting metrics, selected model, and any applied feature transformation. While waiting for AutoAI's run to complete, review this IBM Documentation web page.

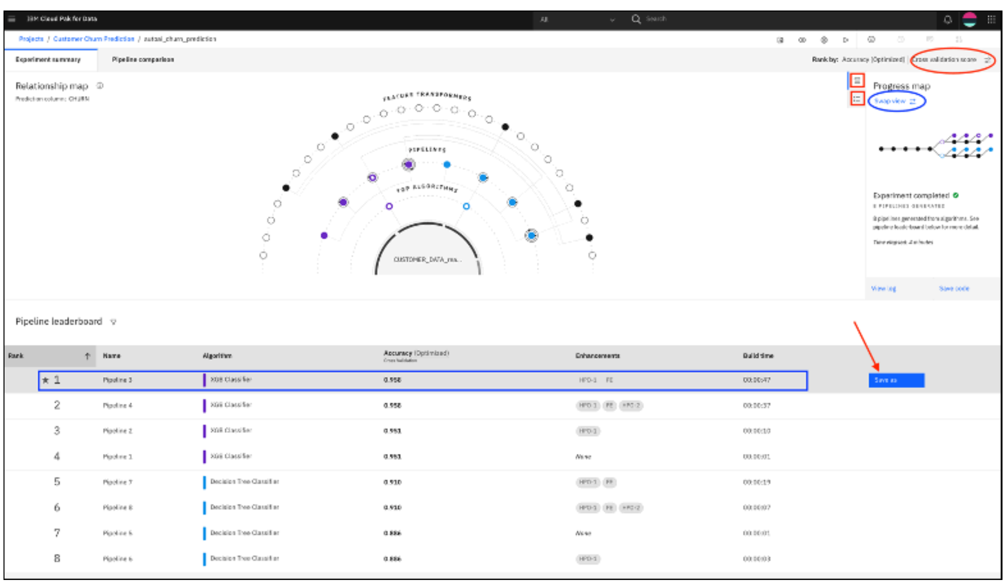

Figure 5-69 Run AutoAI

Specifically, review the AutoAI implementation that is available here to understand which algorithms are supported, which data transformations are applied, and which metrics can be optimized.

8. The AutoAI run takes 4 - 5 minutes to complete. After the process completes, spend a few minutes exploring the dashboard:
   - Switch between the Experiment details and the Legend information to better understand the generated Relationship map.
   - Switch between Cross Validation and Holdout results by clicking the icon that is next to Cross Validation to see how the pipeline ranking changed, depending on which data is being evaluated.
   - Swap the view between the Relationship and the Progress maps to see the different views of the AutoAI pipeline creation process.
   - Click the top pipeline to review the details for that pipeline. AutoAI reports several valuable evaluation criteria, such as several performance metrics (Accuracy, Area under ROC, Precision, Recall, and F1) and the confusion matrix, Precision Recall Curve, and feature importance.

     If the pipeline also included feature engineering (or feature transformation), the pipeline details explains what transformations were applied.

     Close the pipeline details window by clicking the **X** in upper right of the window. After reviewing the trained pipelines, you can decide which one you want save as a model to deploy. Assuming you select the first pipeline, mouse over the first pipeline and click **Save as**.
9. On the Save As page, select **Model**, update the default Name if you want, add an optional Description and Tags and then, click **Create**. You also can save the pipeline as a notebook, which can be customized (see Figure 5-70).

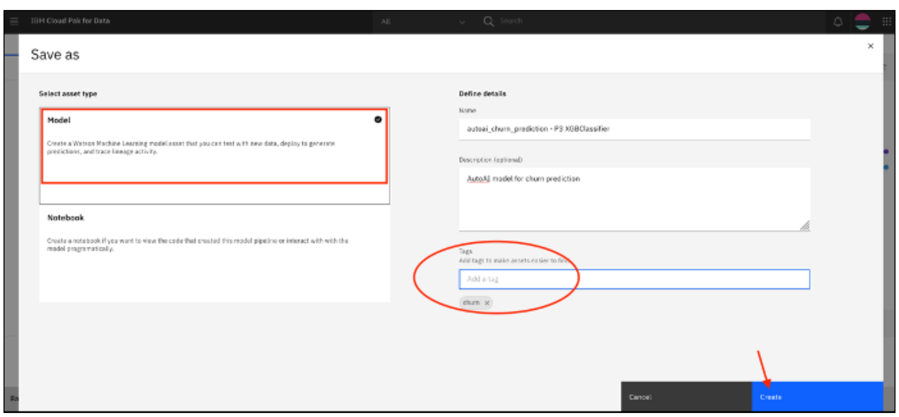

Figure 5-70 Save as model

10. Browse the project assets by clicking your project and then, clicking the **Assets** tab.

The new AutoAI experiment and the new model that you created under the Saved models section are available by clicking the **Actions** menu (three vertical dots) that is next to the model and selecting **Publish to Catalog**. Click the saved model, **autoai_churn_prediction – P3 XGB Classifier** (the name of your model might be different), as shown in Figure 5-71.

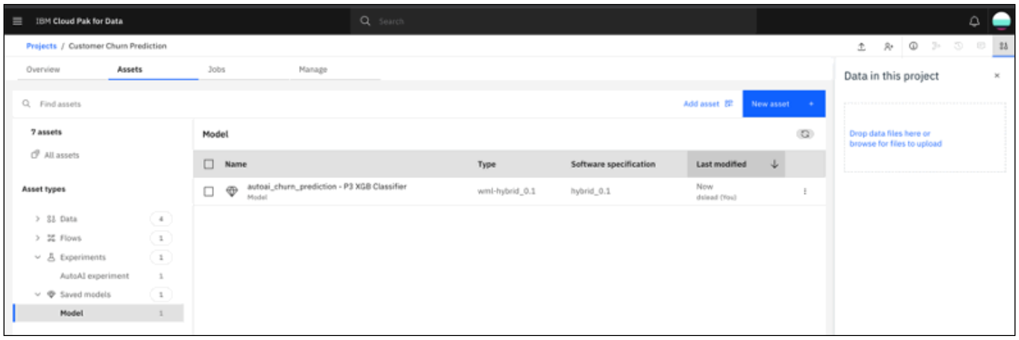

Figure 5-71 Publish to catalog

11.On the model page, click **Promote to deployment space** (see Figure 5-72).

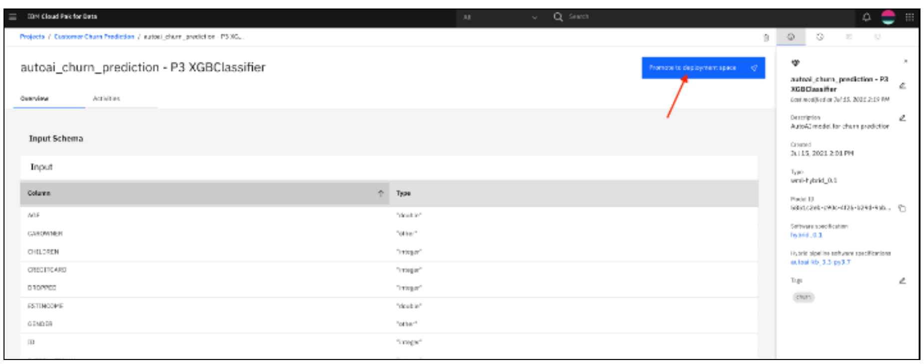

Figure 5-72 Promoting to deployment space

Deployment spaces allow you to create deployments for machine learning models and functions and view and manage all the activity and assets for the deployments, including data connections and connected data assets.

A deployment space is not associated with a project. You can deploy assets from multiple projects to a space, and you can deploy assets to more than one space. For example, you might have a test space for evaluating deployments, and a production space for deployments you want to deploy in business applications.

12. On the Promote to space page, select a target deployment space from the drop-down menu if you created spaces before. If not, click **Create a new deployment space** to create a space (see Figure 5-73).

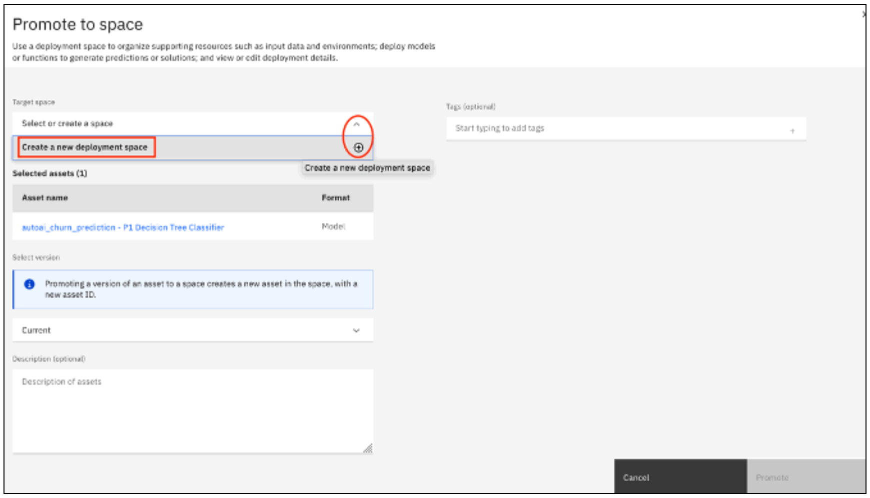

Figure 5-73 Creating a deployment space

13. Enter a Name (`dev`) and an optional description (for example, "Deployment space for collecting assets, such as data and model during development phase") for the deployment space and then, click **Create**. You also can add tags to the space (see Figure 5-74).

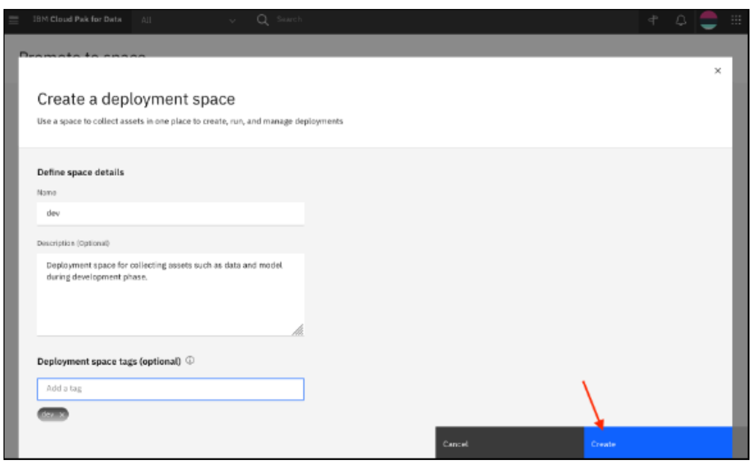

Figure 5-74 Deployment space name

14.In the Promote to space window, keep the default selected version (Current), optionally add a description and tags, and then, click **Promote** (see Figure 5-75).

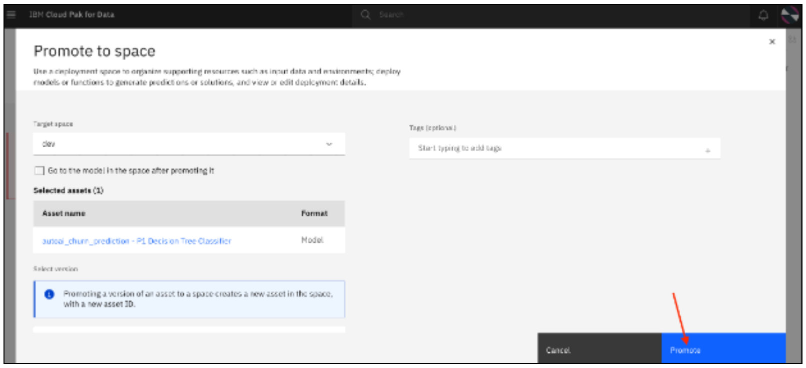

Figure 5-75 Promoting the model

15.After the model is successfully promoted to the deployment space, you see a notification message. Click the **deployment space** link to browse to the deployment space (see Figure 5-76).

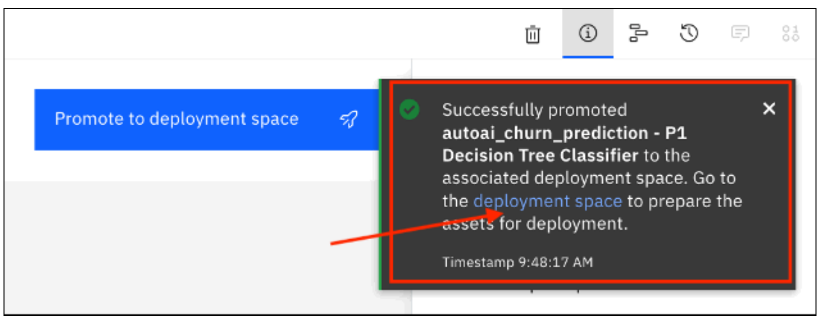

Figure 5-76 Deployment space hyperlink

16.In the Dev deployment space window, select the **Assets** tab and you see the AutoAI model, `autoai_churn_prediction`. Click the **Deploy** icon (a rocket) that is next to the model (see Figure 5-77).

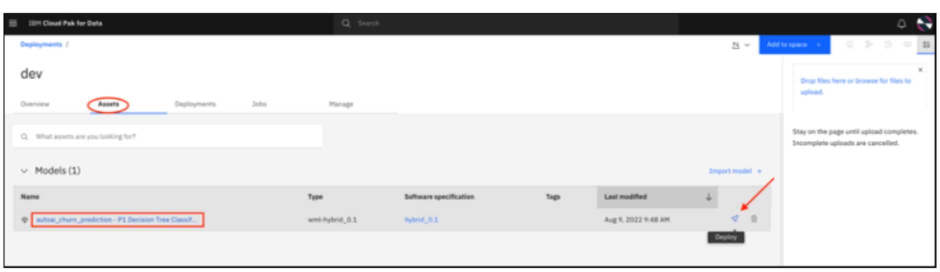

Figure 5-77 Selecting the Deploy option

17. In the Create a deployment window, select **Online**, add a Name (`autoaichurn`) and optional a description and any tags and then, click **Create** (see Figure 5-78).

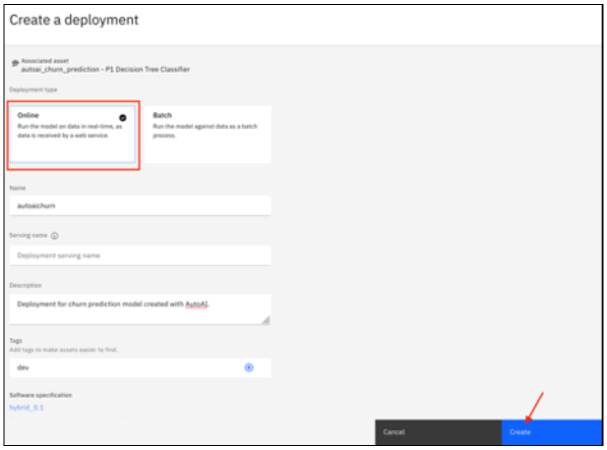

Figure 5-78 Online deployment

18. Click the **Deployments** tab and wait until the deployment status changes to `Deployed`. Then, click the deployed model name **autoaichurn** (see Figure 5-79).

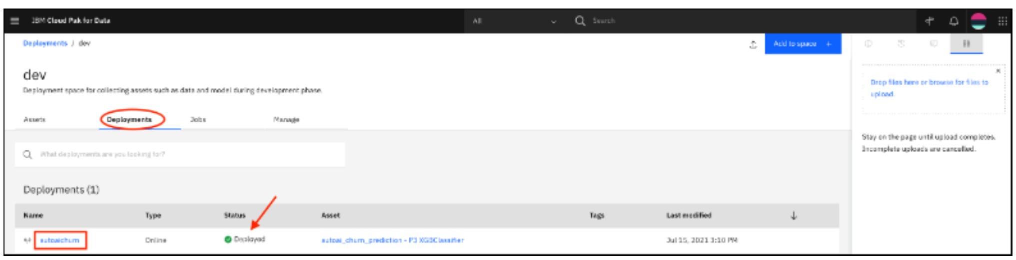

Figure 5-79 List of deployments

19. In the model page API reference tab, review the model endpoint (the URL that serves the model) and the various code snippets in different languages to illustrate how to make an API call to the deployed model.

You also can select a more user friendly serving name, which makes it easier to understand to whomever uses the endpoint (see Figure 5-80).

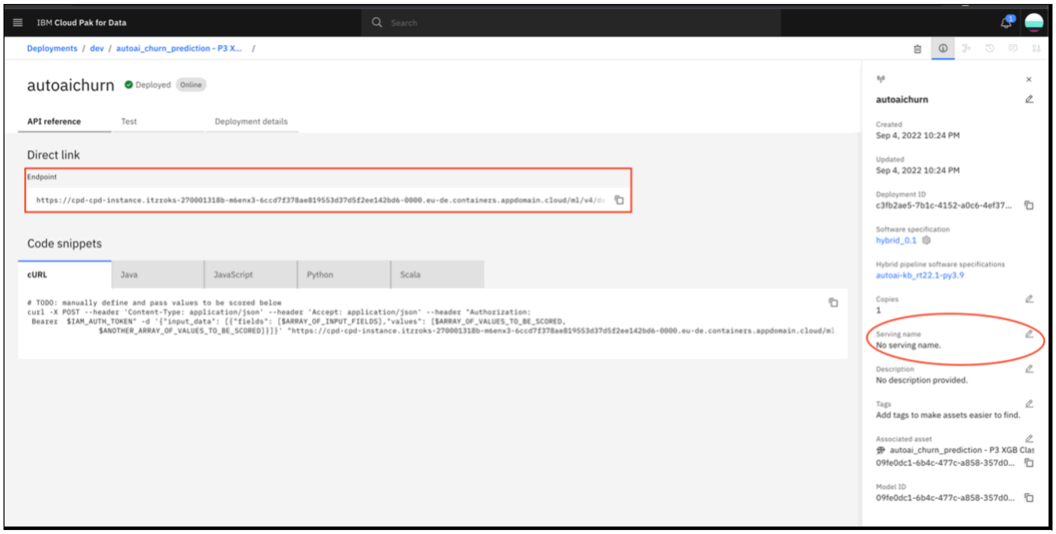

Figure 5-80   Scoring endpoint

20. Click the **edit** icon that is next to “No serving name” and change the name to `autoai_churn`. You notice that a second endpoint is added to the deployment. Then, select the **Test** tab, click the **Provide input data as JSON** icon, paste the JSON sample that is shown in Example 5-1 in the Enter input data window and click **Predict**. The scoring result is shown in Figure 5-81 on page 322.

> **Note:** Be careful as you paste the sample that is shown in Example 5-1. Special characters, such as double quotes, might not copy correctly, which can cause the prediction to generate an error.

Example 5-1   JSON sample

```
{ "input_data": [
{
"fields":["ID","LONGDISTANCE","INTERNATIONAL","LOCAL","DROPPED","PAYMETHOD","LOCALBILLTYPE","LON
GDISTANCEBILLTYPE","USAGE","RATEPLAN","GENDER","STATUS","CHILDREN","ESTINCOME","CAROWNER","AGE"]
,
"values":[[1,28,0,60,0,"Auto","FreeLocal","Standard",89,4,"F","M",1,23000,"N",45]]
}
] }
```

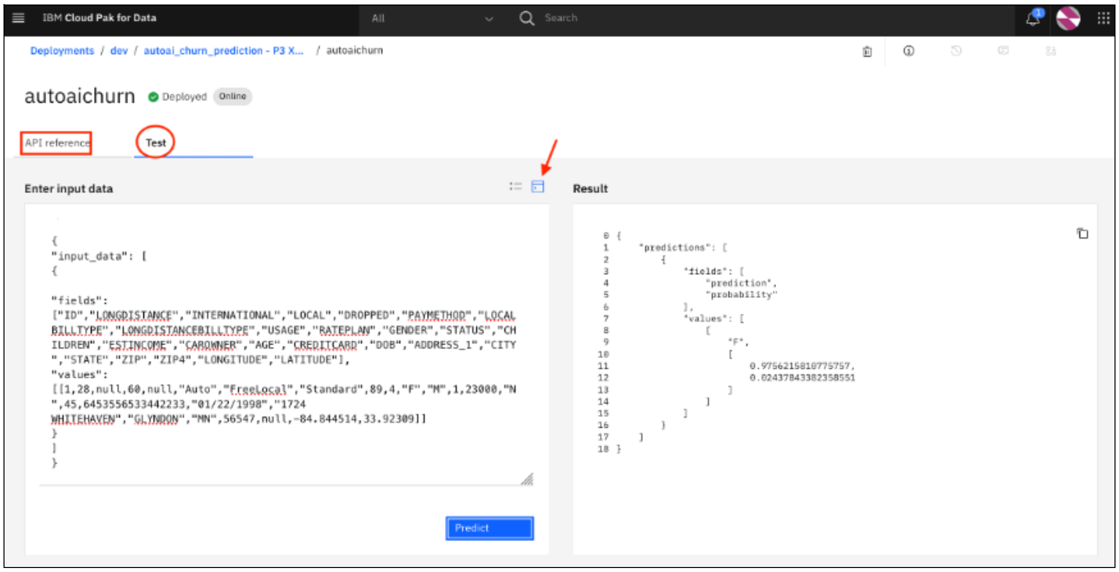

Figure 5-81 Scoring result

The deployed model predict the likelihood of the user to churn because of the specific values for the various features. The model returns the predicted churn label as `T` (true) or `F` (false) and the probability of that prediction, which effectively expresses the likelihood of that user to churn.

A `T` label that is returned by the model indicates the user is likely to churn and the corresponding probability. These probabilities can be used with the predicted label to better serve customers on a more granular basis. Your application can be customized to make decisions based on the predicted label and the probabilities of that prediction.

Training and deploying a churn prediction model with Jupyter notebook

In this section, we present an alternative method for training AI models in Cloud Pak for Data, namely by using Jupyter notebook and open-source libraries. This common and mostly preferred method by data scientists provides them with the utmost flexibility in exploring different algorithms for training best performing AI models.

Complete the following steps:

1. Log into Cloud Pak for Data as `datascientist` user.
2. Browse to your Customer Churn Prediction project.
3. Click the **Assets** tab and then, click **New asset +** (see Figure 5-82).

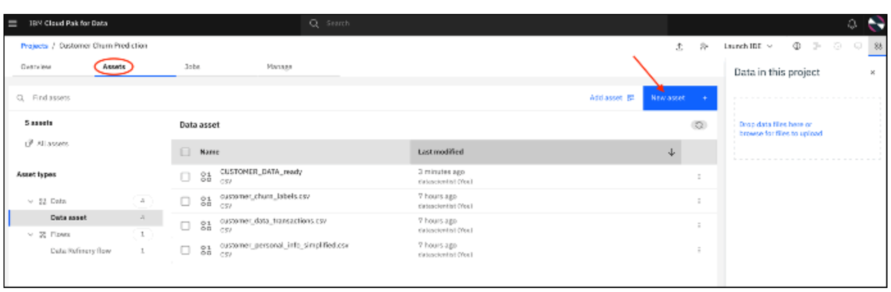

Figure 5-82 New asset

4. Scroll down and select the **Jupyter notebook editor**. Note that you can filter asset types by selecting the **Code editors** to quickly find Jupyter notebook editor (see Figure 5-83).

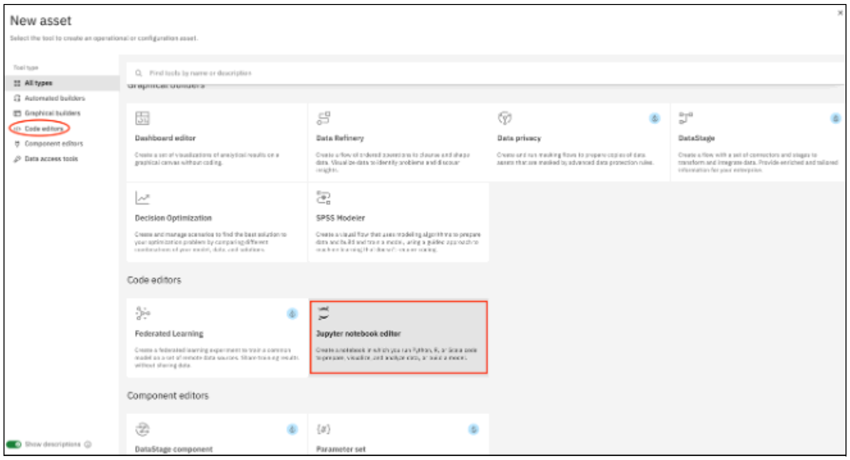

Figure 5-83 Jupyter notebook editor

5. In the New notebook window, click the **From file** tab and then, click the **Drag and drop files here or upload** option. Select the `churn_prediction_pyspark.ipynb` notebook from the downloaded notebooks folder. Click **Create** (see Figure 5-84).

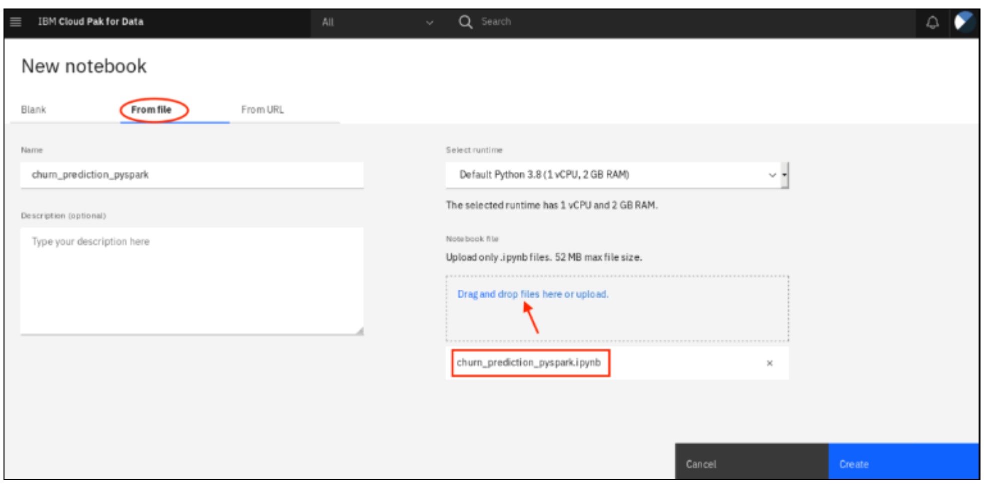

Figure 5-84 Creating notebook from file

6. Review the documentation and the steps in the notebook step by step by repetitively clicking the **Run** icon or by using **Cell Run All** from the menu. Running the entire notebook takes approximately 5 minutes, depending on the speed of your cluster. When the steps are complete, the notebook performed the following steps:
   a. Accessed the `CUSTOMER_DATA_ready.csv` data from your project.
   b. Processed the data to prepare features relevant for the prediction.
   c. Trained a Random Forest ML model to predict the likelihood of customers to churn by using a sample of the data.
   d. Evaluated the model against test data that was not used in training.
   e. Associated Watson Machine Learning with a deployment space.
   f. Created `churnUATspace` deployment space.

   g. Stored the model in the `churnUATspace` deployment space.
   h. Deployed the ML model to the `churnUATspace` deployment space.
   i. Ran a test to validate the online deployment of the model.

7. After the notebook completes, save the notebook and then, stop the run time to save resources because this environment is shared. To do so, navigate to the **Customer Churn Prediction** project, select the **Manage** tab, and click **Environments**. Then, select the check box that is next to the active environment and click **Stop runtime**. Click **Stop** on the confirmation pop-up window (see Figure 5-85).

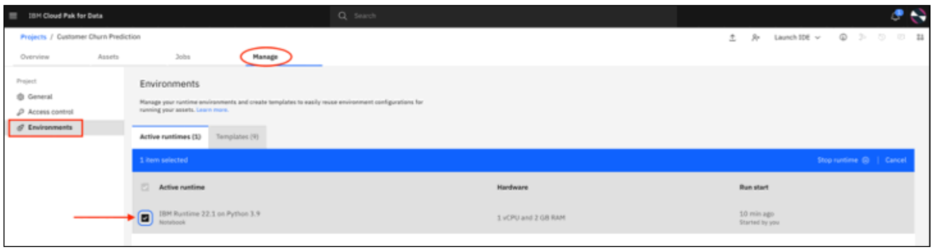

Figure 5-85 Stopping the run time

5.4.6 Deploying models

In the previous exercises, you used two different methods to deploy models: manually in the case of the AutoAI model, and programmatically by using API calls in the notebook.

IBM Watson Studio offers various methods to train and deploy models and make the models available to be infused in business processes. Models are always deployed in Cloud Pak for Data deployment spaces.

While in development, data scientists and application development teams experiment with different model parameters (features) and test integration with business applications and processes. After a model is ready to be taken to the next stage in its lifecycle; for example, to test integration with other applications, often it is migrated to a different deployment space to isolate it from development work.

Deployment spaces can exist in the same Cloud Pak for Data instance or other instances, potentially on different clusters or even different cloud infrastructures.

In most organizations, the lifecycle management of models is automated. This automation is used to avoid manual activities that are subject to errors, and it contributes to standardization processes. As such, when run correctly, trust in models and trust in processes (recall that the main imperative is to trustworthy AI).

For more information about the automation of the machine learning lifecycle, see 5.4.10, “Automating the ML lifecycle” on page 366. For now, it is assumed that you understand the concepts of training a model and deploying it into a space.

5.4.7 Monitoring Machine Learning models

Continuous monitoring and management of deployed AI models (see this web page) is critical for businesses to trust model predictions. Analysts reported that lack of trust in the AI models is one of the main reasons that inhibits AI adoption in enterprises in critical applications, which includes tracking several performances and business KPIs (key performance indicators). The lack of explainability and the potential concerns with fairness or bias expose an enterprise to significant risk to reputation and financial loss.

IBM Watson OpenScale is an integrated offering in IBM Cloud Pak for Data. It provides a powerful operations console for business users to track and measure AI outcomes.

5.4.8 IBM Watson OpenScale

IBM Watson OpenScale provides visibility and explainability into AI outcomes, which helps to ensure fair outcomes while giving business process owners greater trust in the ability of AI to augment decision making, and the confidence to scale it across their workflows.

At the same time, the solution provides a robust framework to ensure AI maintains compliance with corporate policies and regulatory requirements. It also helps remove barriers to AI adoption by empowering users to deploy and manage models across projects, at whatever scale the business requires.

IBM Watson OpenScale tracks and measures outcomes from your AI models, and helps ensure that they remain fair, explainable, and compliant no matter where your models were built or are running. IBM Watson OpenScale also detects and helps correct the drift in accuracy when an AI model is in production.

Figure 5-86 shows the IBM Watson OpenScale monitoring AI models.

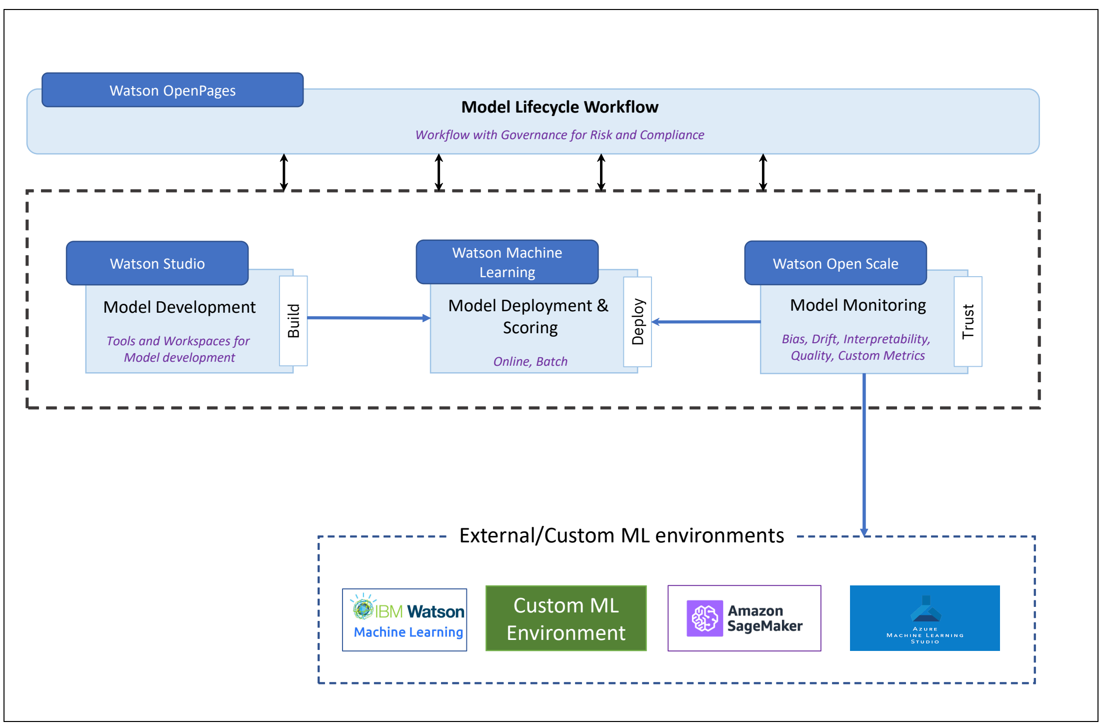

Figure 5-86 Watson OpenScale monitoring AI models

Enterprises use model evaluation to automate and put into service an AI lifecycle in business applications. This approach ensures that AI models are free from bias, can be easily explained and understood by business users, and are auditable in business transactions. Model evaluation supports AI models that are built and run with the tools and model-serve frameworks.

Figure 5-87 shows the following components of IBM Watson OpenScale:

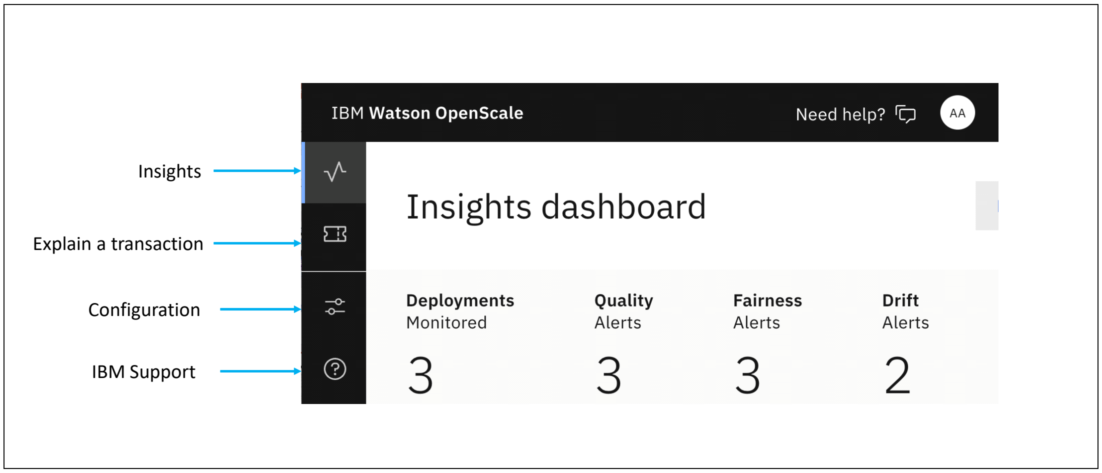

Figure 5-87 Components of IBM Watson OpenScale

- Insights: The Insights dashboard displays the models that you are monitoring and provides status on the results of model evaluations.
- Explain a transaction: Explanations describe how the model determined a prediction. It lists some of the most important factors that led to the predictions so that you can be confident in the process.
- Configuration: Use the Configuration tab to select a database, set up a machine learning provider, and optionally add integrated services.
- Support: The Support tab provides you with resources to get the help you need with IBM Watson OpenScale.

IBM Watson OpenScale monitors

Monitors in IBM Watson OpenScale evaluate the deployments against specified metrics. Alerts can be configured to indicate when a threshold is crossed for a metric. IBM Watson OpenScale evaluates your deployments based on the following default monitors:

- Quality describes the model's ability to provide correct outcomes that are based on labeled test data that is called Feedback data.
- Fairness describes how evenly the model delivers favorable outcomes between groups. The fairness monitor looks for biased outcomes in your model.
- Drift warns you of a drop in accuracy or data consistency.

IBM Watson OpenScale also supports Custom Monitors and metrics. This support enables the users to define custom metrics and use them alongside the standard metrics by way of a programmatic interface that uses Python SDK.

IBM Watson OpenScale Quality Metrics

The quality monitor (or accuracy monitor) reports how well the AI model predicts outcomes and does so by comparing the model predictions to ground truth data (see Figure 5-88). It provides several quality measurements that are suitable for different types of AI models, such as:

- Area under ROC curve (AUC)
- Precision
- Recall
- F1-Measure for binary classification models
- Weighted true positive rate
- Weighted recall
- Weighted precision
- Weighted F1-Measure for multi-class classification models and mean absolute error (MAE)
- Mean squared error (MSE)
- R squared
- Root of mean squared error (RMSE) for regression models

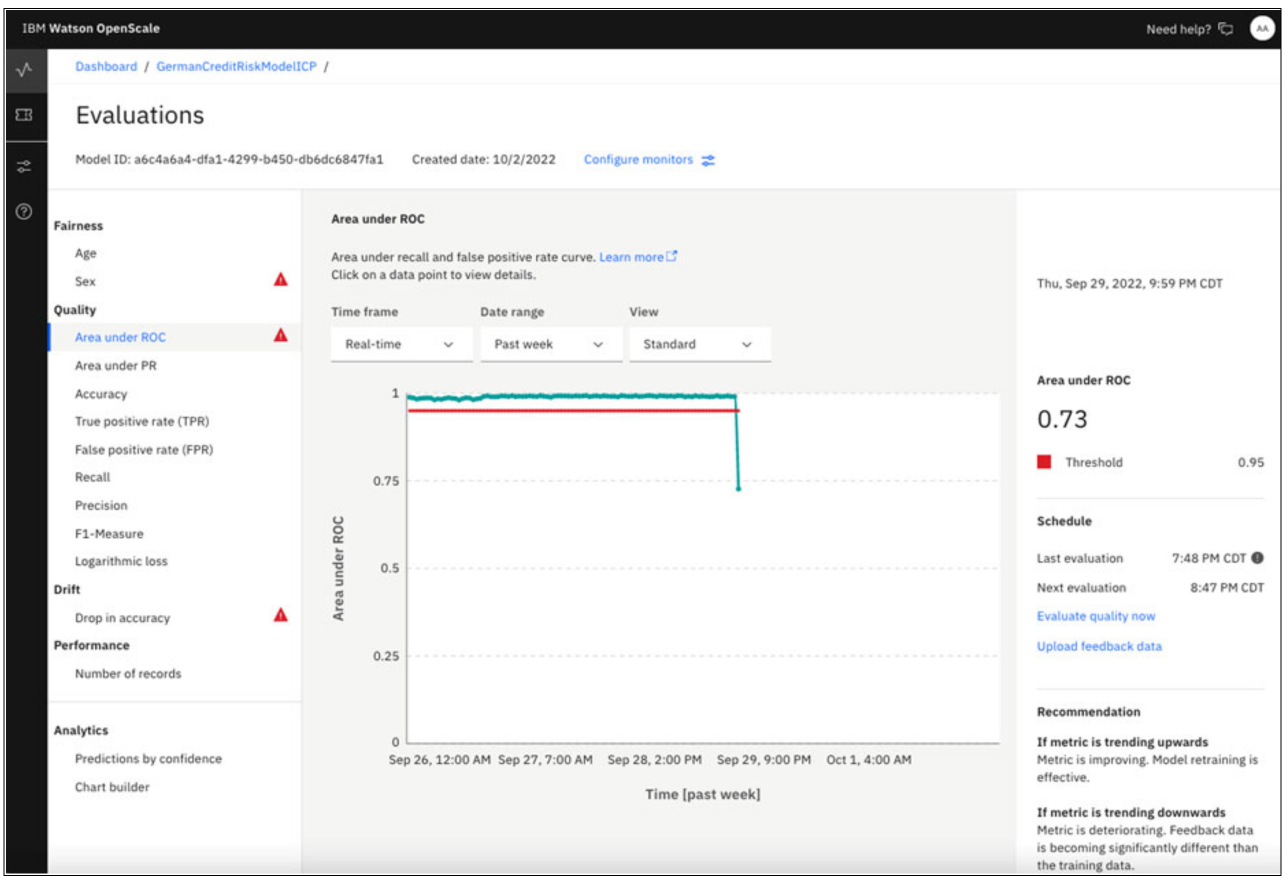

Figure 5-88 Quality Monitoring Area under ROC in IBM Watson OpenScale

IBM Watson OpenScale fairness metrics

By using the IBM Watson OpenScale fairness monitor, you can generate a set of metrics to evaluate the fairness of your model. These fairness metrics can be used to determine whether the model produces a biased outcome. This fairness monitor scans the deployment for bias to ensure a fair outcome across different populations.

IBM Watson OpenScale drift monitoring

In the machine learning lifecycle, *drift* refers to the degradation of predictive performance over time. IBM Watson OpenScale detects and highlights drift so that you can take action to prevent this model degradation over time.

IBM Watson OpenScale drift monitor analyzes the behavior of your model and builds its own model to predict whether your model generates an accurate prediction for a data point. The drift detection model processes the payload data to identify the number of records for which your model makes inaccurate predictions and generates the predicted accuracy of your model during training to identify the drop in accuracy.

IBM Watson OpenScale can be used with AI that is deployed in any runtime environment, including the non-IBM platforms, such as Amazon Sage Maker, Microsoft Azure ML Studio, Microsoft Azure ML Service, and custom runtime environments that are behind the enterprise firewall. It supports machine learning models and deep learning models that are developed in any open source, model-building and training IDE, including in TensorFlow, Scikitlearn, Keras, SparkML, and PMML.

Configuring IBM Watson OpenScale instance with local Db2 instance

A key component of a Governed MLOps solution is the ability to monitor AI models for accuracy, fairness, explainability, and drift. These capabilities deliver trustworthy AI that business leaders can safely adopt in their business processes and customer engagements.

With multiple customer engagements, we found that having the confidence to trust AI models is just as important, and sometimes even more important, than the performance of the AI models.

Requirements and integrated services

The following requirements and integrated services are featured:

- IBM Watson OpenScale must be installed in this Cloud Pak for Data environment.
- Db2 for storing model metrics and metadata, and any test or training data.
- IBM Watson Studio for building machine learning model.
- IBM Watson Machine Learning services for deploying and running the machine learning model.
- Jupyter notebook for model building and automating the model deployment and training.
- IBM OpenPages for model governance and risk management.
- Cloud Object Storage for storing training data.

Complete the following steps to the create the IBM Watson OpenScale instance in Cloud Pak for Data and configure it with the Db2 database:

1. Log in to Cloud Pak for Data as an admin user.
2. Browse to the service instance on your Cloud Pak for Data cluster by clicking the navigation menu and selecting **Services** → **Instances** (see Figure 5-89).

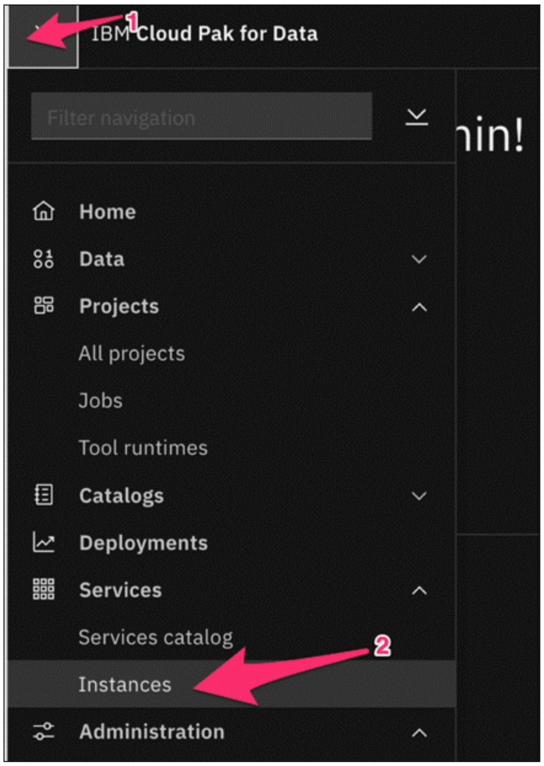

Figure 5-89 Cloud Pak for Data Console menu

3. On the Instances page, find the OpenScale instance. If it does not exist, create an OpenScale instance by selecting **New Instance** at the top right corner (see Figure 5-90).

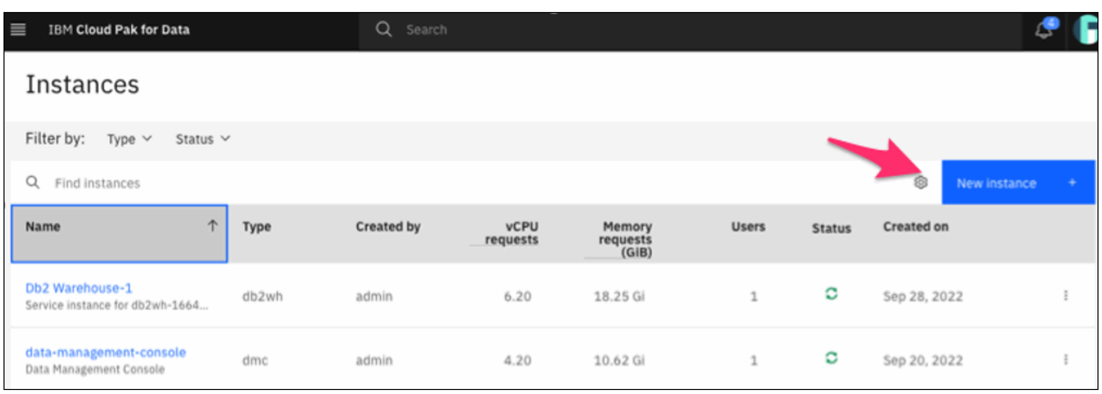

Figure 5-90 Cloud Pak for Data service instances

4. Create an OpenScale instance under the Cloud Pak for Data namespace as a CPD Administrator (see Figure 5-91).

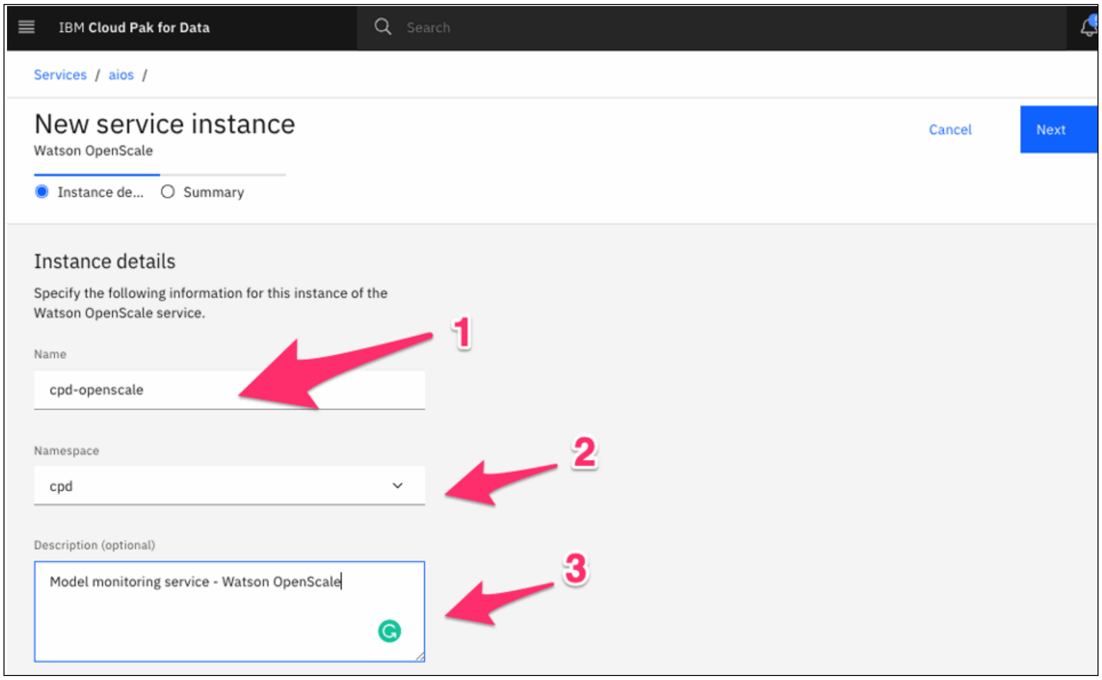

Figure 5-91 IBM Watson OpenScale new instance creation

5. On the instances page, find the OpenScale, select the **Open** drop-down menu and then, click **Open**. The order of selections is annotated, as shown in Figure 5-92.

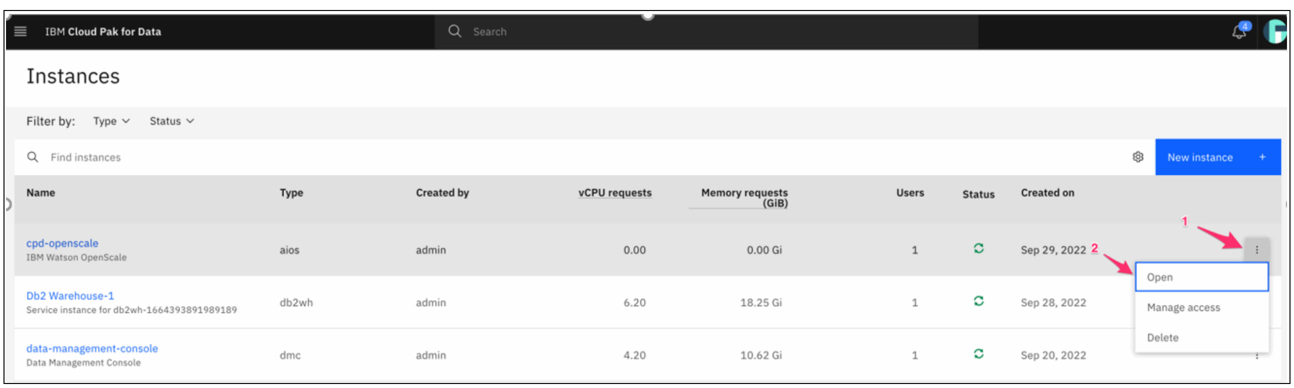

Figure 5-92 Watson OpenScale setup

6. When IBM Watson OpenScale starts, you see a landing page with a proposition to run Auto Setup. Auto setup automatically trains, deploys, and sets up monitoring of three machine learning models. Each of these models is trained on the German Credit Risk dataset to predict customer loan default. These models can be used to become familiar with the IBM Watson OpenScale and see all its capabilities without a manual monitoring setup. In general, it is useful to use Auto setup to become familiar with OpenScale and demonstrating its capabilities.

   For this example, you follow the Auto setup to configure IBM Watson OpenScale. This process requires a Db2 instance and an IBM Watson Machine Leaning instance.

In our example, we use the local instances of Db2 and WML (see Figure 5-93).

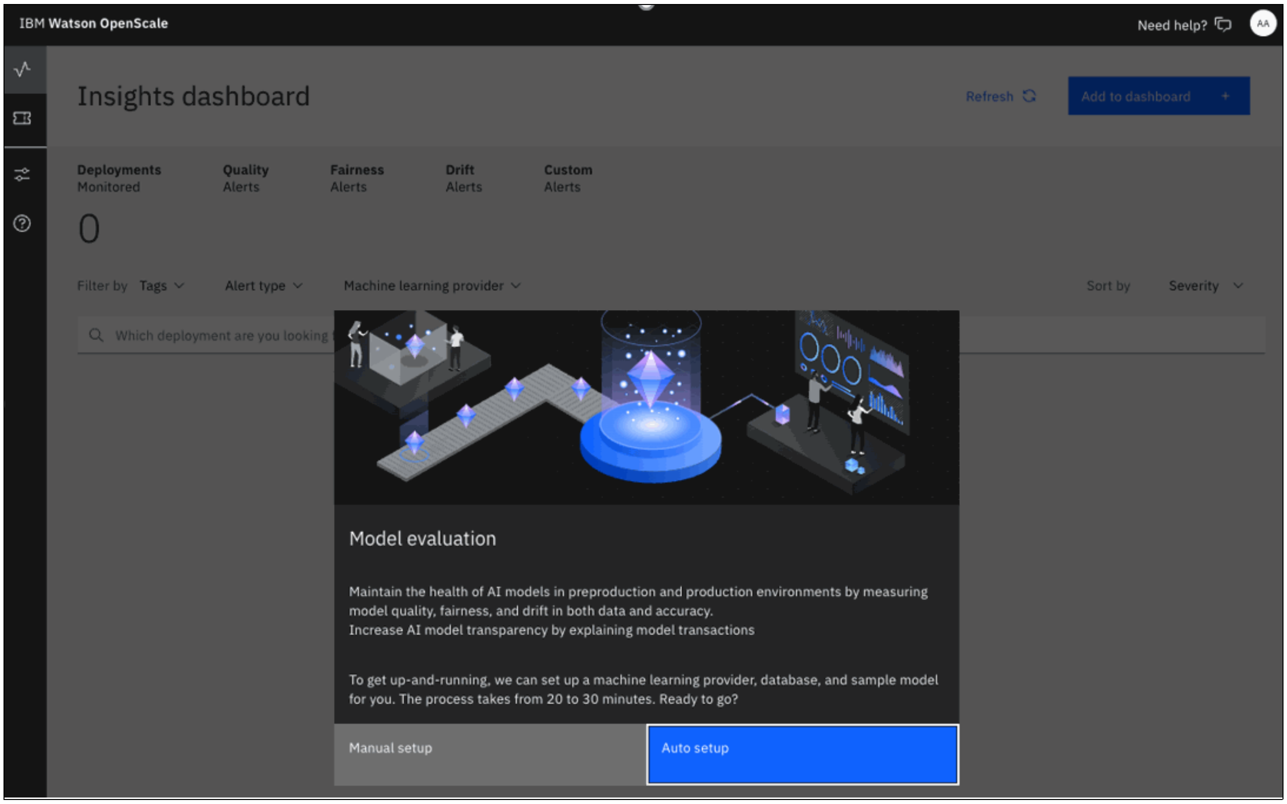

Figure 5-93 Watson OpenScale Auto setup

7. Select **Db2** as the database type and complete the Db2 connection and credentials information for the Db2 instance that is running on the same Cloud Pak for Data cluster. The following information is required to complete the setup (see Figure 5-94):
   – Hostname or IP address
   – Database port
   – Database name
   – Username and password

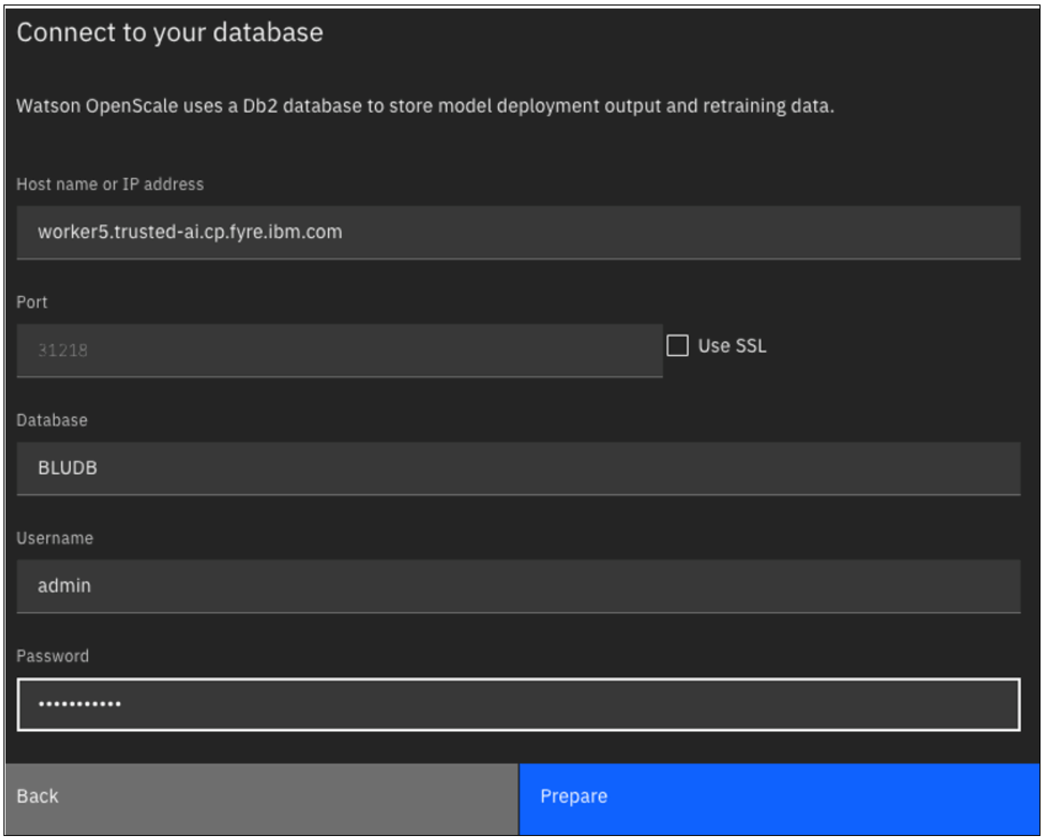

Figure 5-94 Watson OpenScale Db2 configuration

Configuring the OpenScale instance with local Db2 instance takes approximately 10 minutes to complete.

8. On successful Auto Setup, three Machine Learning models are deployed and configured with IBM Watson OpenScale monitoring, as shown in Figure 5-95.

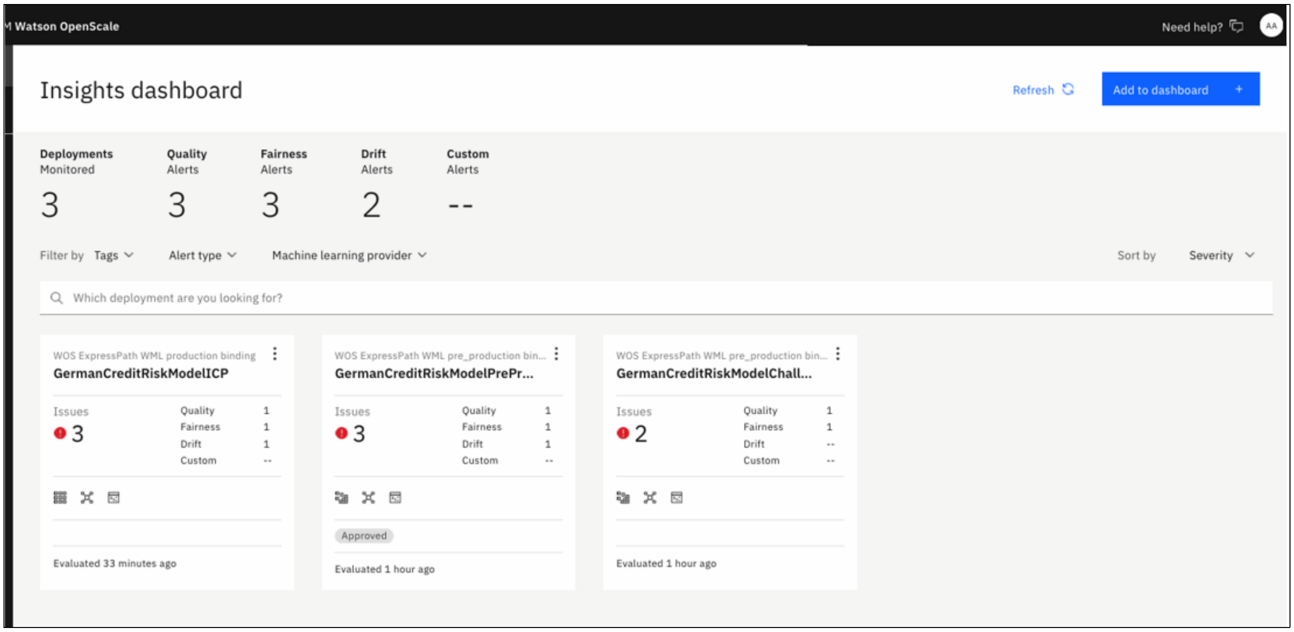

Figure 5-95 Watson OpenScale auto setup with German Credit Risk Model

This completes the initial setup of the Watson OpenScale instance.

Deploying Space Access for Admin

In this section, you add the administrator user to the `churnUATspace` deployment space. This user must be added so that IBM Watson Studio Trust (also referred to as IBM Watson OpenScale) can monitor the models that deployed in this deployment space.

If you do not complete these steps now, you are prompted to add the administrator user to your deployment space as part of register machine learning providers when you attempt to monitor models from IBM Watson OpenScale.

Complete the following steps:

1. Browse to your Deployment space by clicking the navigation menu at the top left and selecting **Deployments**.
2. In the Deployments window, select the **Spaces** tab and then, click the **churnUATspace** (see Figure 5-96).

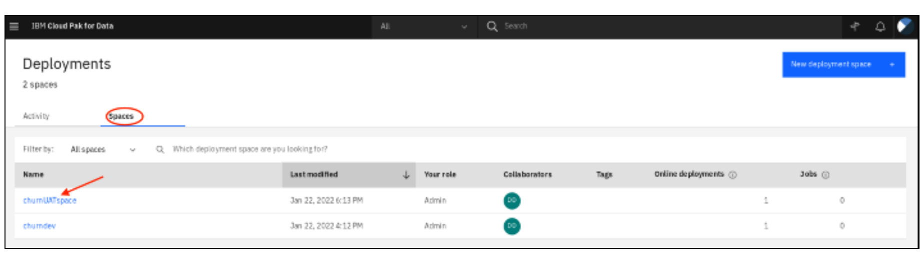

Figure 5-96 Selecting space

3. In the churnUATspace window, click the **Manage** tab and then, select **Access control**. Click the **Add collaborators** drop-down menu and select **Add users** (see Figure 5-97).

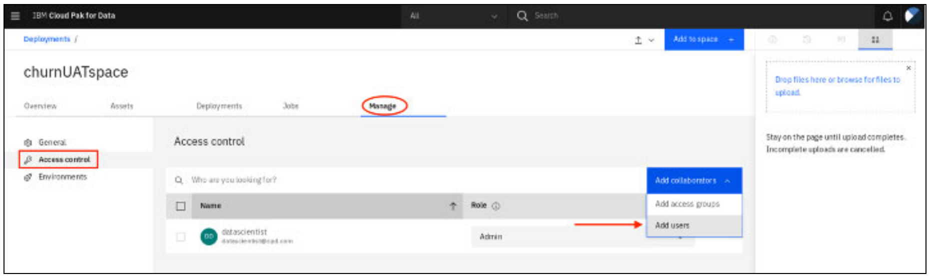

Figure 5-97 Adding users

4. Click the check box that is next to **admin** user and select the role as **Editor**. Click **Add** (see Figure 5-98).

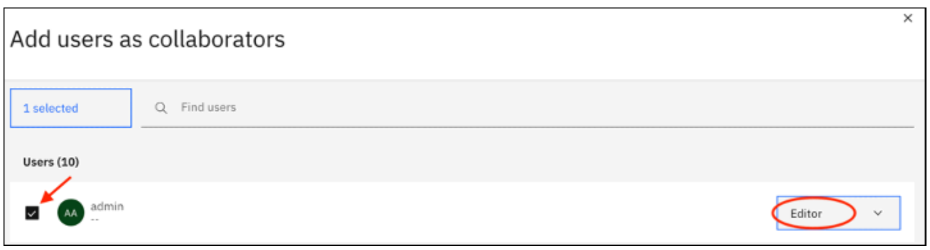

Figure 5-98 Selecting user role

Monitoring the model by using IBM Watson OpenScale

In this example, we describe the process that is used to monitor the machine learning model `customer churn` that was built and deployed under the deployment space `churn-preprod`.

IBM Watson OpenScale can support multiple machine learning environments, including third-party providers, such as Microsoft Azure ML Studio and AWS SageMaker.

In this example, IBM Watson Machine Learning is used.

Complete the following steps:

1. Select **System setup**, as shown in Figure 5-99 with arrow 1. Next, select the **Machine learning providers** to add Watson Machine Learning.

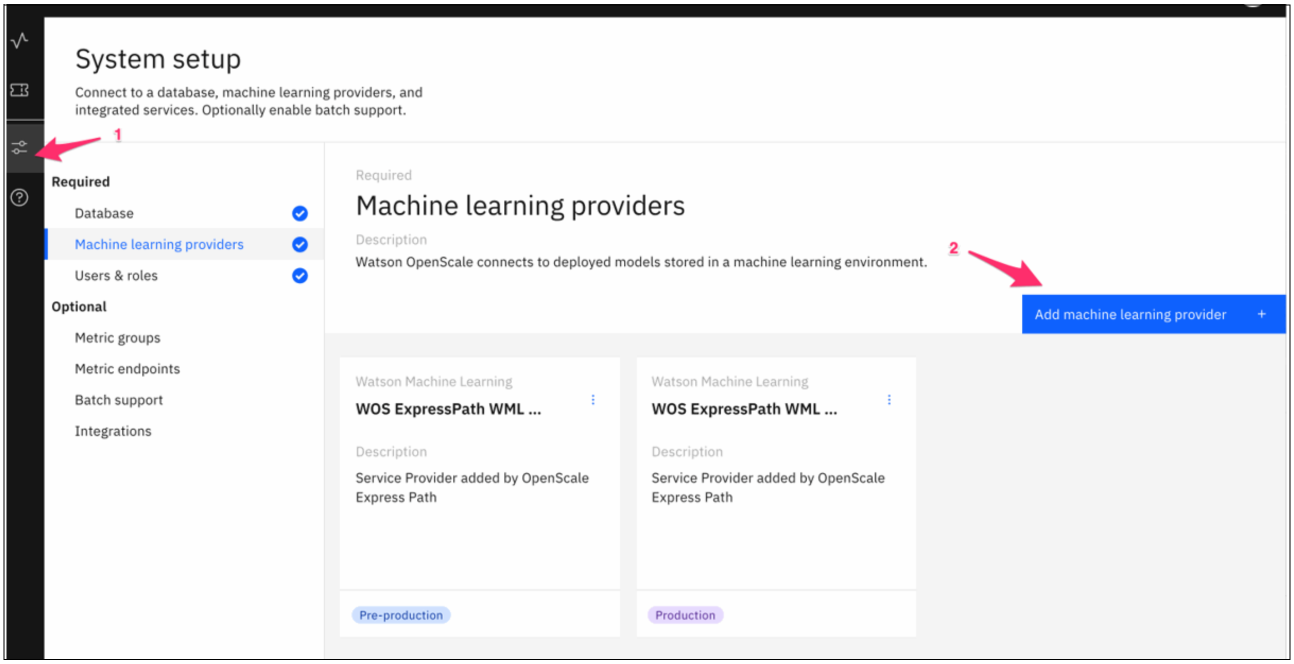

Figure 5-99 Configuring the machine learning provider

2. In this example, Watson Machine Learning is configured as the service provider by connecting to the local instance of WML with a preproduction environment type for testing (see Figure 5-100).

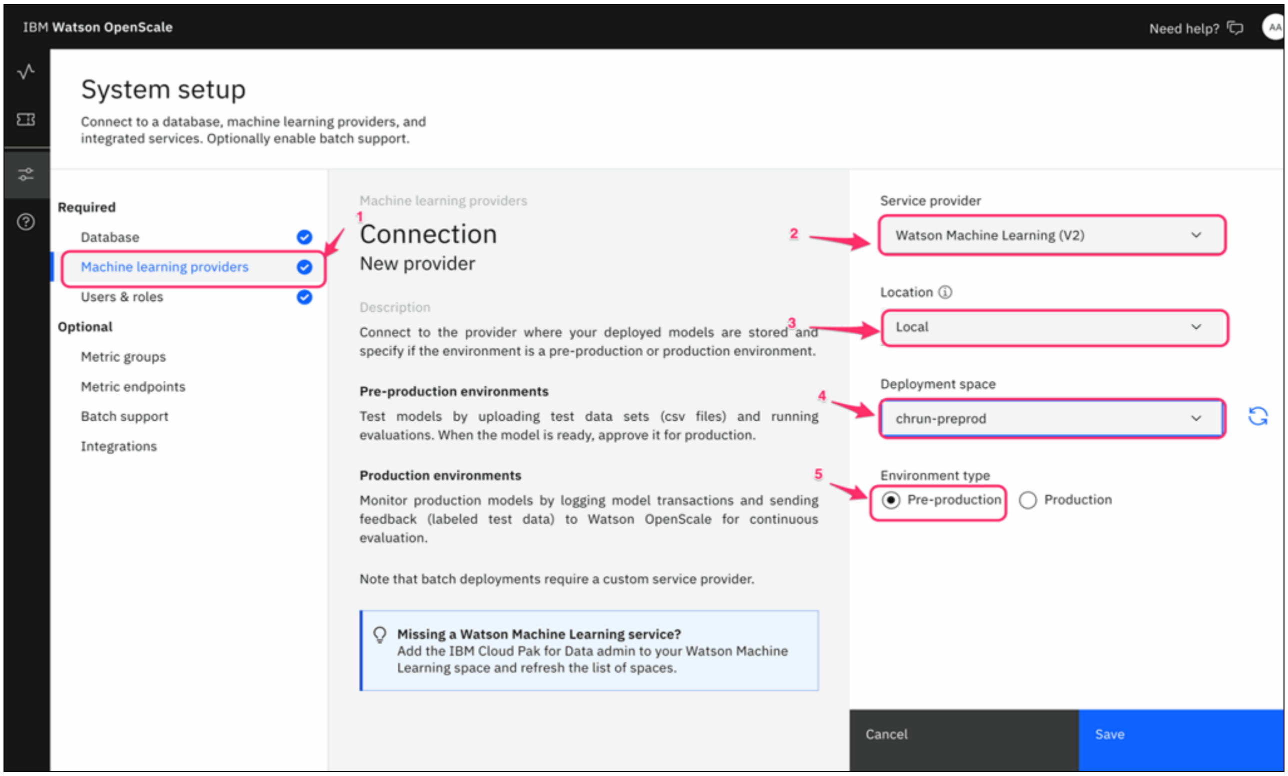

Figure 5-100 Configuring the Watson Machine Learning provider

Notice that a new tile for Watson Machine Learning is added as a machine learning provider, as shown in Figure 5-101.

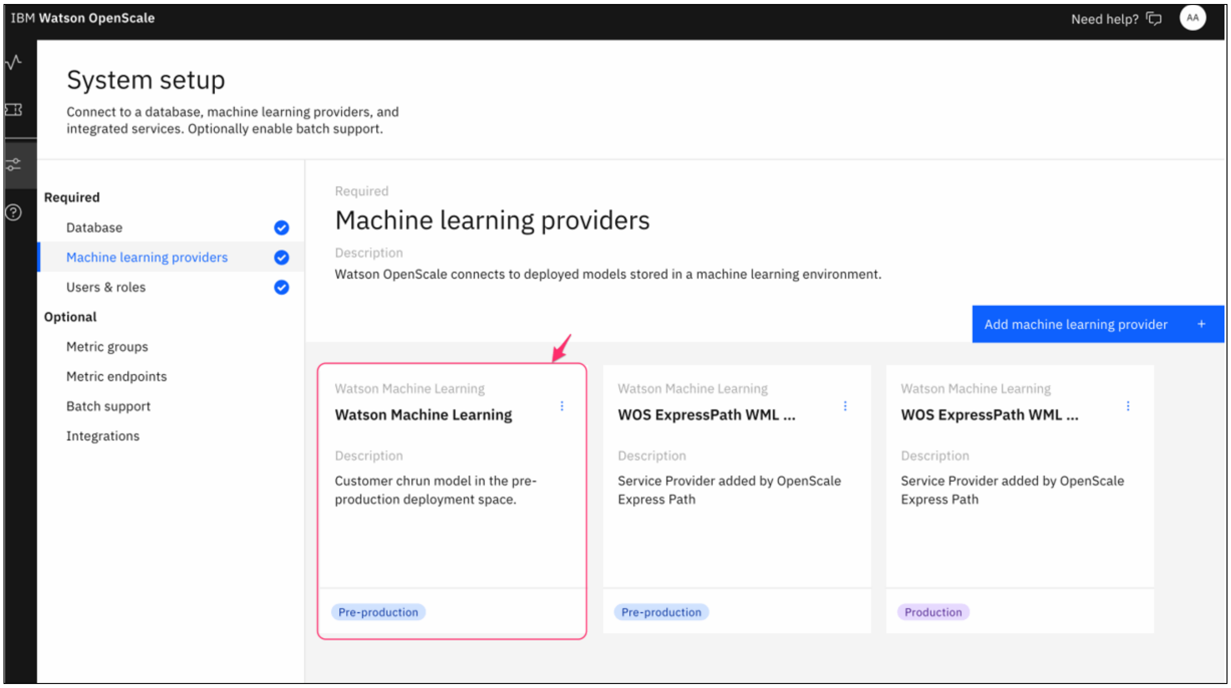

Figure 5-101 Pre-production instance of Watson Machine Learning provider

3. Click the **Action menu** and select **View & edit details**.
4. In the IBM Watson OpenScale window, select the **Insight dashboard** and select **Add to dashboard** to add the Customer churn prediction model to IBM Watson OpenScale monitoring (see Figure 5-102).

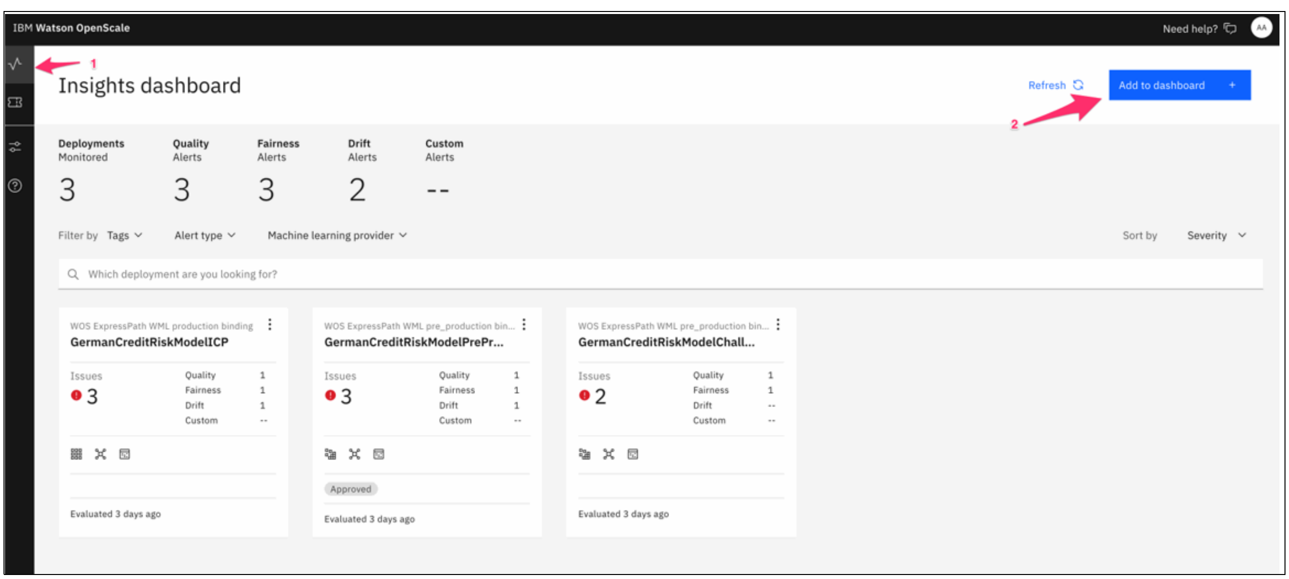

Figure 5-102 Configuring OpenScale insight dashboard

5. IBM Watson OpenScale pulls all of the deployed models of the selected machine learning providers. You can use the drop-down menu (red arrow 1 in Figure 5-103) to select a specific machine learning provider. Of the list of deployed models, select the model you want to monitor (red arrow 2 in Figure 5-103). After selecting the model, click **Configure**, which becomes active after a model is selected.

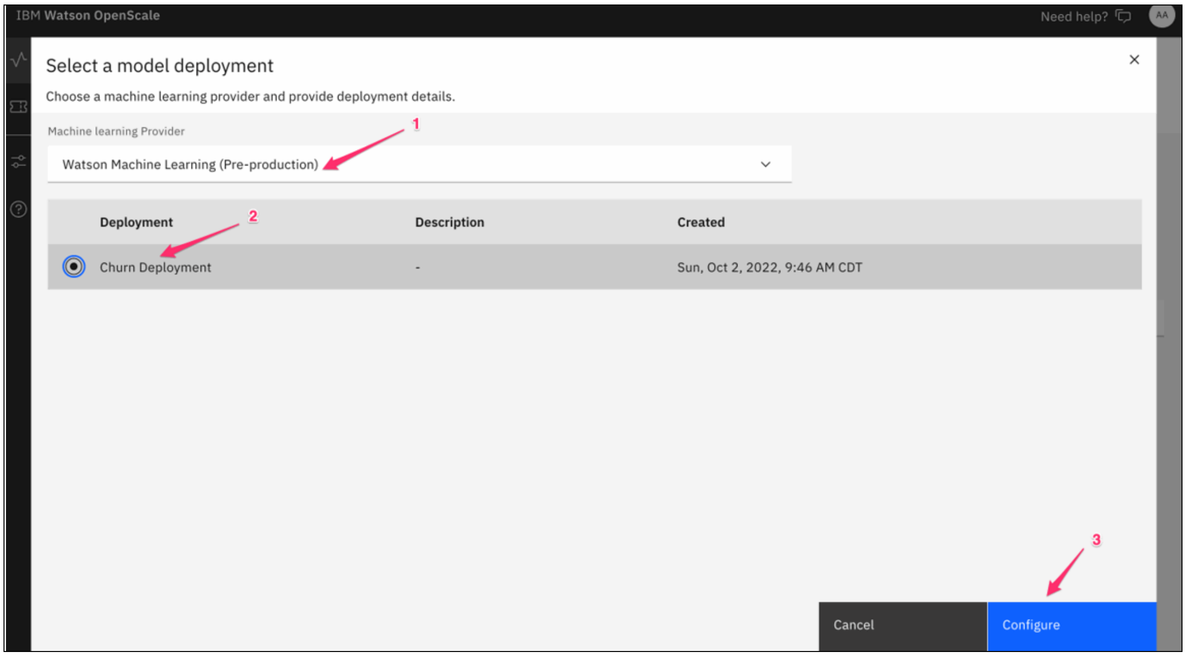

Figure 5-103 Configuring OpenScale with deployment space

6. You see the message "Selection saved", as shown in Figure 5-104. Next, select **Configure monitors**.

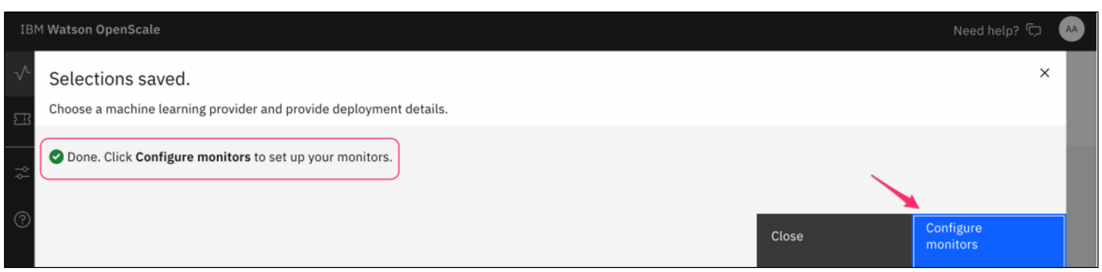

Figure 5-104 Configuring OpenScale with monitors

7. The first step in configuring monitors is to provide model details. Specifically, we must provide information about the model input and output and the training data (see Figure 5-105).

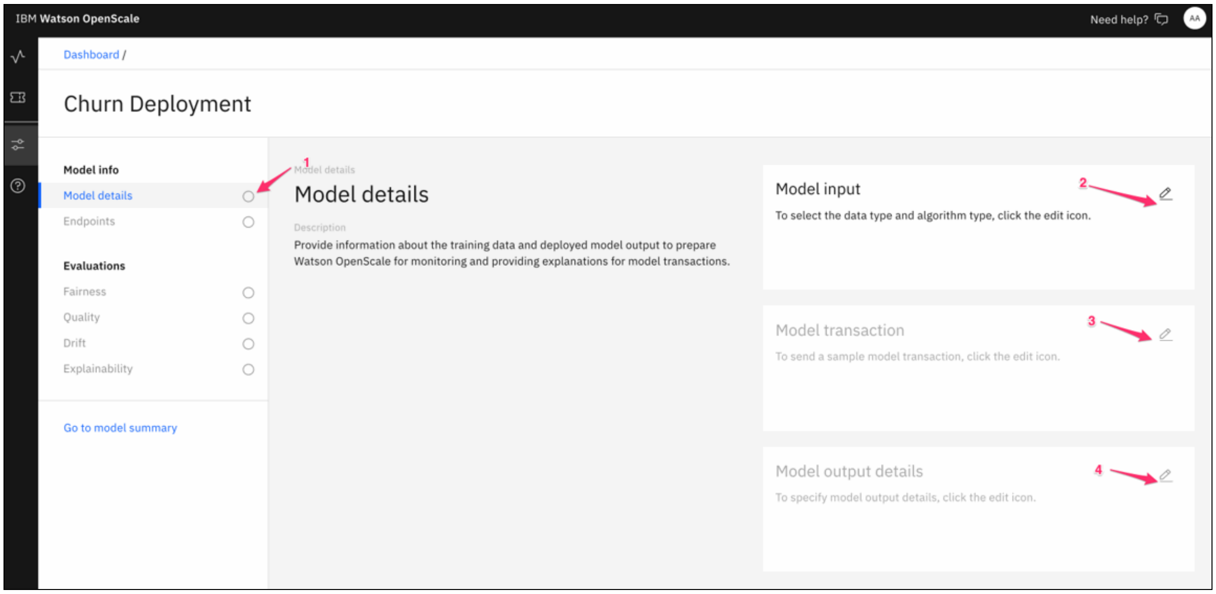

Figure 5-105 OpenScale with model details

8. Configure the model Input by clicking the pencil icon in the Model Input section and specify the data type and algorithm type. For the churn prediction model that we are monitoring, select the Data type as **Numerical/categorical** and the Algorithm type as **Binary classification** as shown in Figure 5-106. Click **Save** to continue.

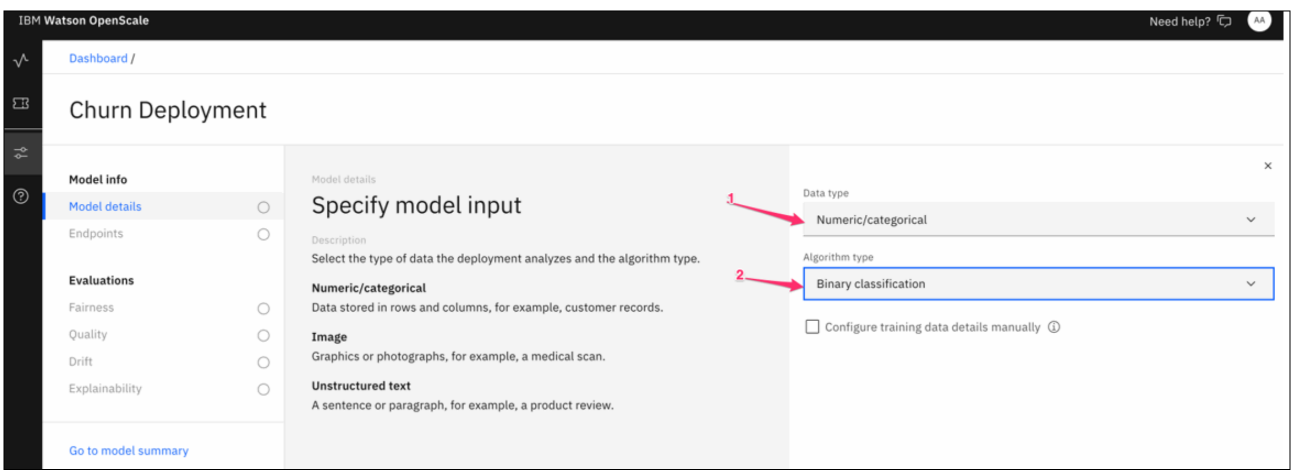

Figure 5-106 OpenScale with model type

9. Configuring the training data by providing connection information so that the IBM Watson OpenScale can connect to the training data and extract the statistics that are needed for monitoring. IBM Watson OpenScale supports reading training data from Cloud Object Storage (COS) and Db2. In this example, we use the Db2 as a source for training data:
   – Schema: `Customer`
   – Table: `Customer_training_data`

The detailed parameter inputs are shown in Figure 5-107.

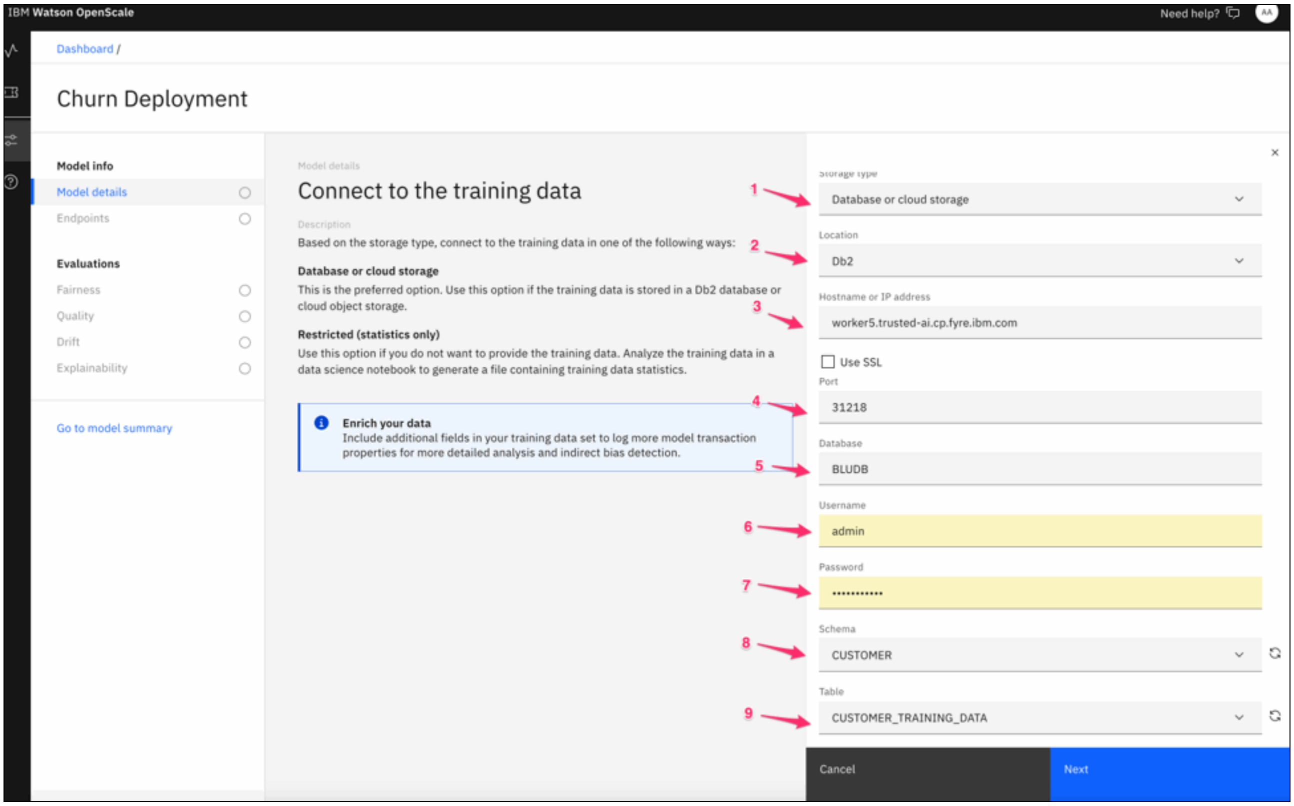

Figure 5-107 OpenScale with training data source

10. Select the label column as the **CHURN** and then, click **Next** (see Figure 5-108).

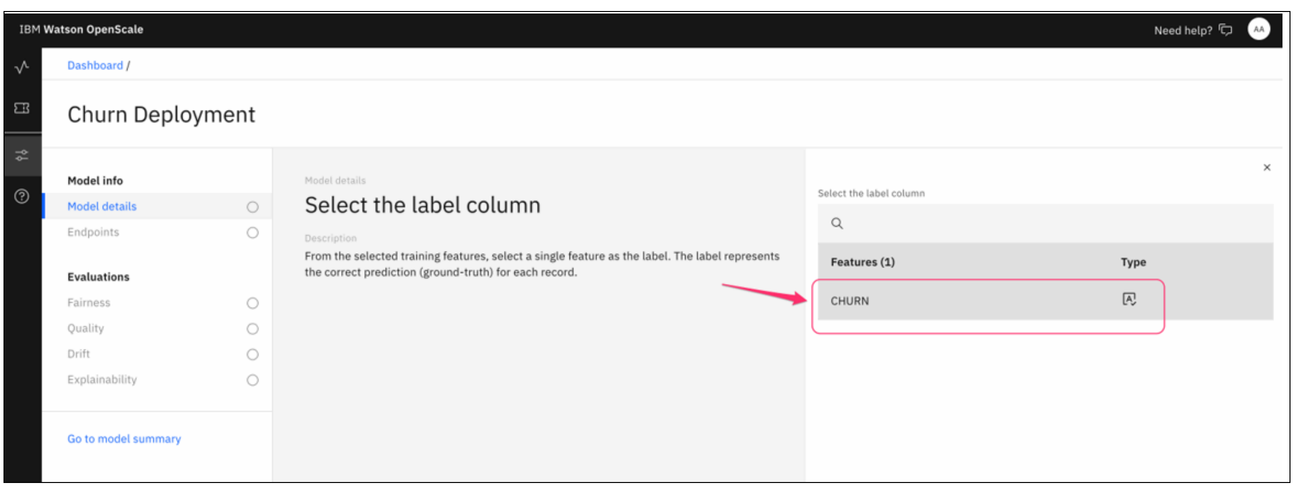

Figure 5-108 OpenScale with label column selection

11. For training features, accept the default setting, which selects all the features (see Figure 5-109). Click **Next**.

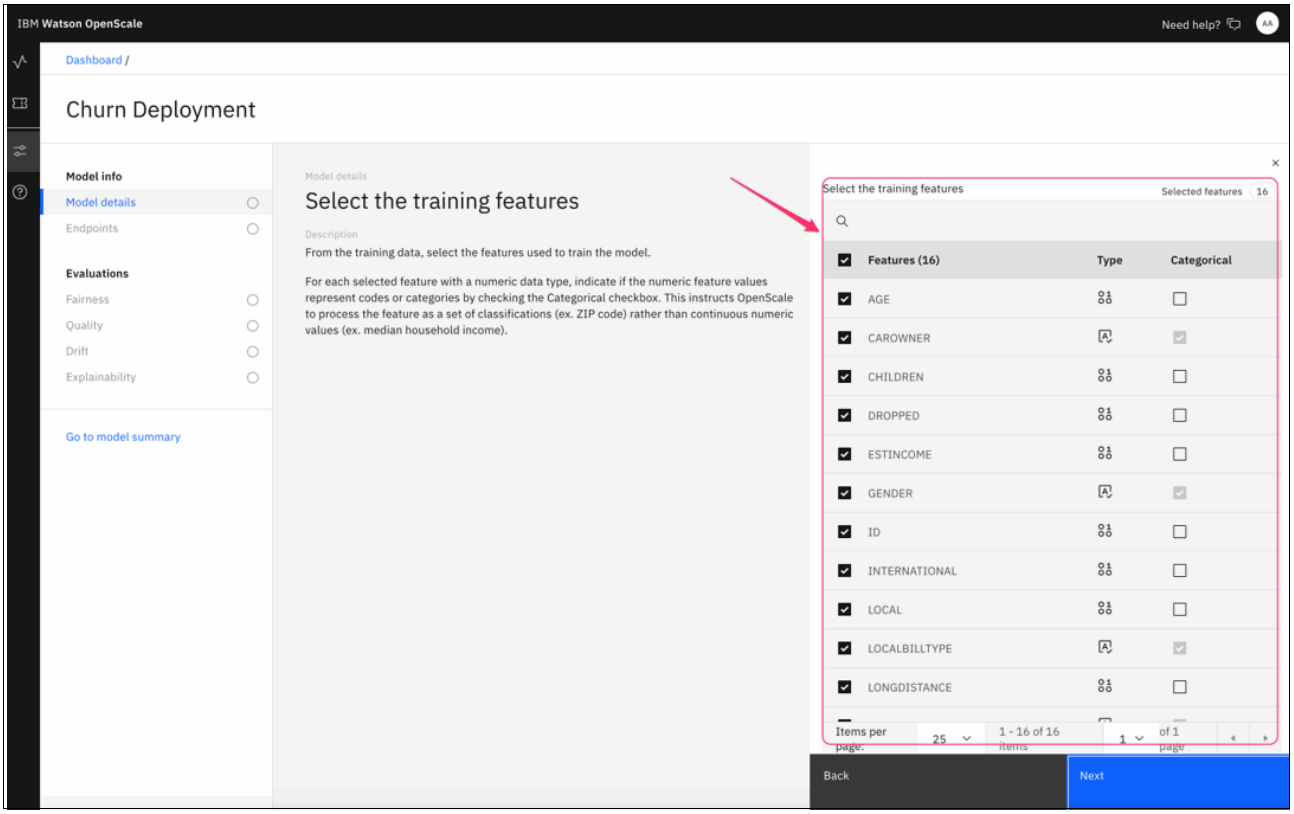

Figure 5-109 OpenScale with selecting training features

12. The Examining model output page is shown. As explained on that page, we must send a scoring request to the deployed ML model so that IBM Watson OpenScale can be prepared for tracking and storing the transactions that are processed by this ML model.

    For IBM Watson Machine Learning, the Automatic logging option is supported and is likely the easiest method to enable IBM Watson OpenScale to understand the schema of the input payload and output response. In OpenScale, select **Automatic logging**.

13. Browse to your model deployment in Cloud Pak for Data to run one transaction of inference against the model so OIBM Watson OpenScale can pick up the schema (see Figure 5-110 on page 342):

    a. Log into Cloud Pak for Data as the admin user.

    b. Browse to the deployments by clicking the navigation menu and selecting **Deployments**.

    c. On the Deployments page, select the **Spaces** tab and then, click **churn-prepod**.

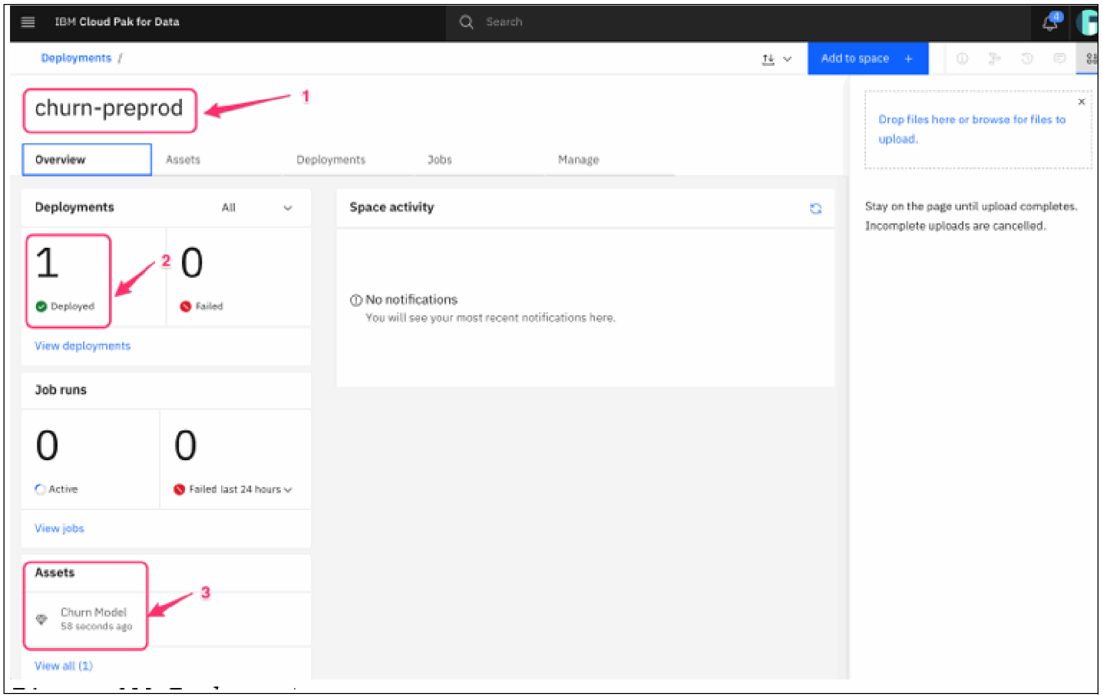

Figure 5-110 Deployment space

d. On the deployment page, click the **Test** tab. Then, provide a sample payload in JSON format by clicking **Provide input data** as JSON icon (see Figure 5-111).

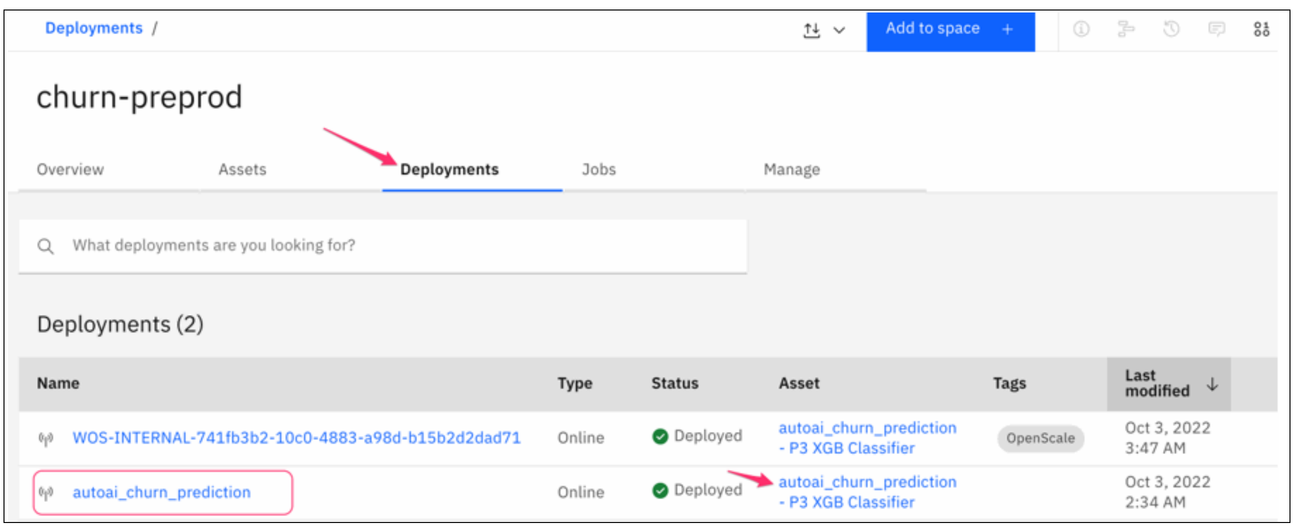

Figure 5-111 Selecting deployment space and model

Sample payload in JSON format is shown in Example 5-2.

Example 5-2 Sample payload in JSON format

```
{
"input_data": [
{
"fields":

["ID","LONGDISTANCE","INTERNATIONAL","LOCAL","DROPPED","PAYMETHOD","LOCALBILLTYPE","LONGDISTANCE
BILLTYPE","USAGE","RATEPLAN","GENDER","STATUS","CHILDREN","ESTINCOME","CAROWNER","AGE"],
"values":[[1,28,0,60,0,"Auto","FreeLocal","Standard",89,4,"F","M",1,23000,"N",45]]
}
]
}
```

e. Click **Predict** to have the trained ML model predict the likelihood of the customer to churn based on the provided feature values (see Figure 5-112).

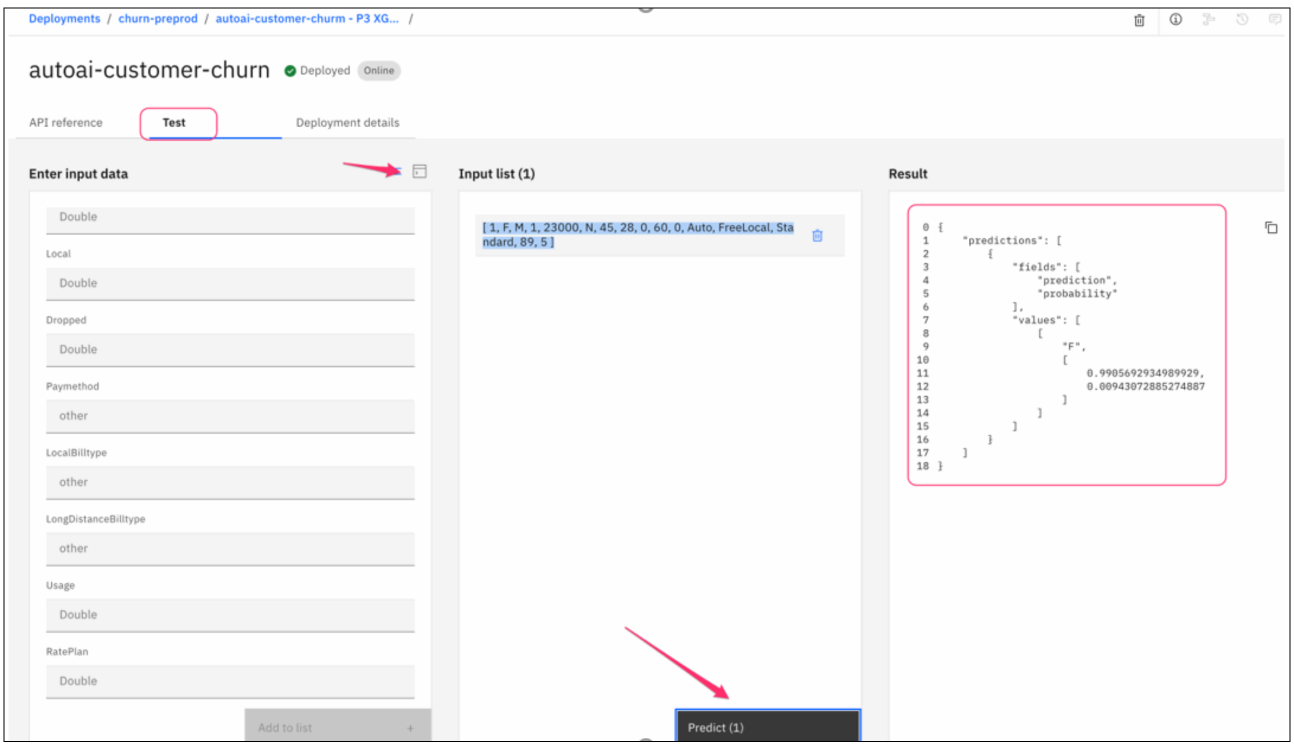

Figure 5-112 Testing the customer churn model

14. Navigate to IBM Watson OpenScale and click **Check now**. Check **Now** and you should receive a message that indicates that Logging is active. Click **Next**.

15. On the Specify model output details page, specify the prediction Label as the `Prediction`. Also, notice that the probability feature is mapped to Probability. OpenScale needs the prediction and probability information for the purposes of drift detection, debiasing, and explainability (see Figure 5-113).

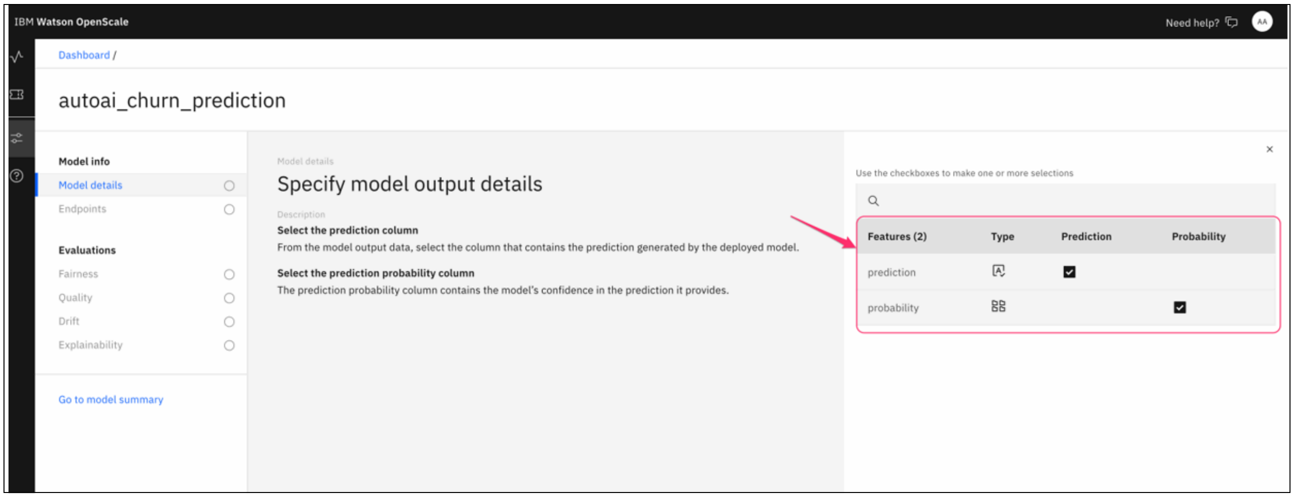

Figure 5-113 Model output details

We provided all the required information for IBM Watson OpenScale to prepare for monitoring the deployed machine learning models.

Next, we configure all three OpenScale monitoring metrics.

Quality monitoring

Complete the following steps:

1. Select **Quality** on the OpenScale evaluations and click the pencil icon (as indicated by a red arrow in Figure 5-114) to configure the quality monitor in OpenScale. IBM Watson OpenScale can monitor the quality metric that measures the model's ability to correctly predict outcomes that match labeled data.

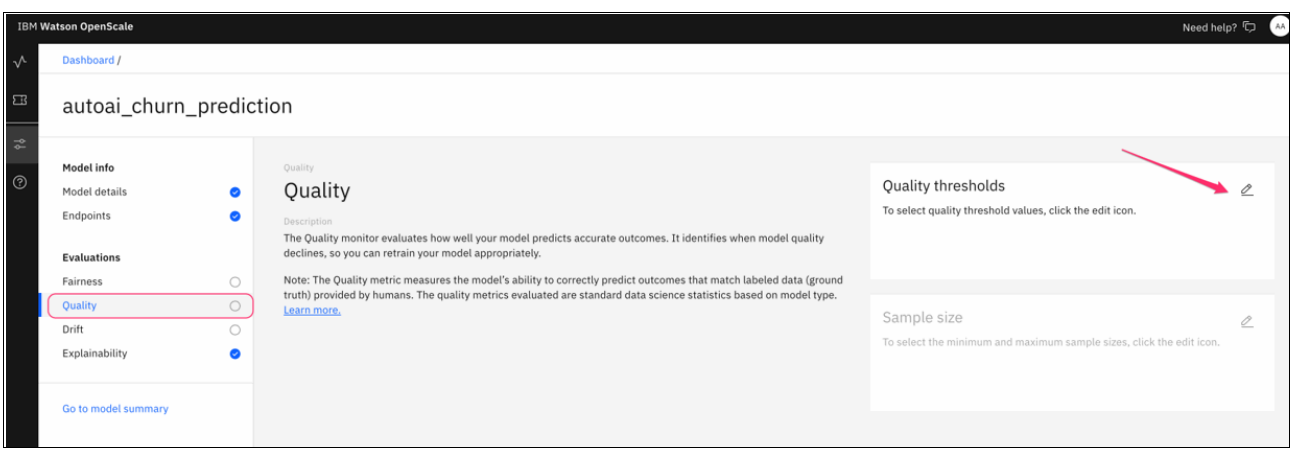

Figure 5-114 Configure OpenScale with Quality monitoring

2. Specify the Threshold value for Area under ROC to be 0.9 and click **Next**. By making this specification, the quality monitor flags an alert when the Area under ROC curve (here, denoted “Area under ROC”) is less than 0.9 (see Figure 5-115).

Figure 5-115 Configure Area under ROC

3. Change the minimum sample size to 100. In production, larger sample sizes are used to ensure that they are representative of the requests the model receives. Click **Save** (see Figure 5-116).

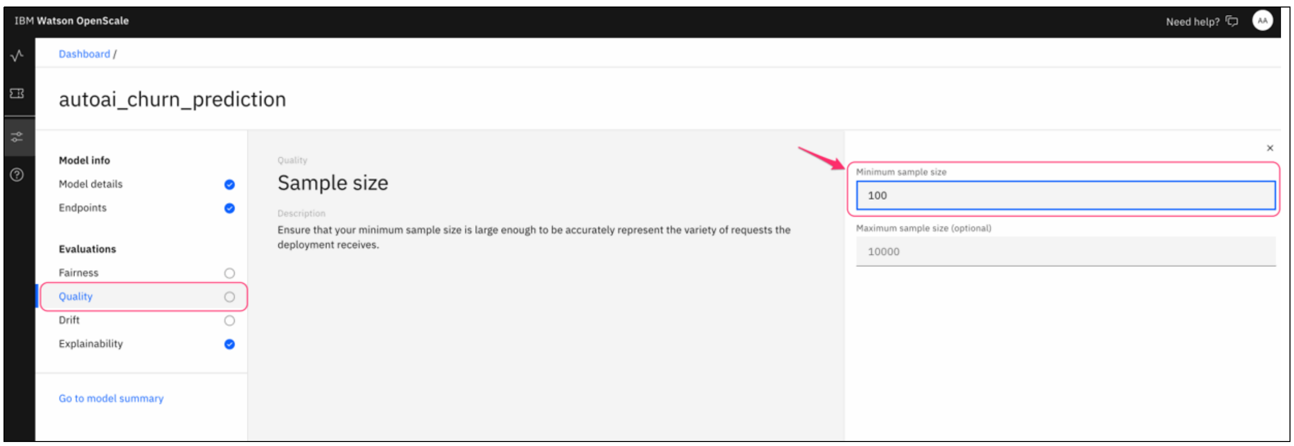

Figure 5-116 Quality monitoring sample size

4. Review the Quality monitor configuration summary page (see Figure 5-117).

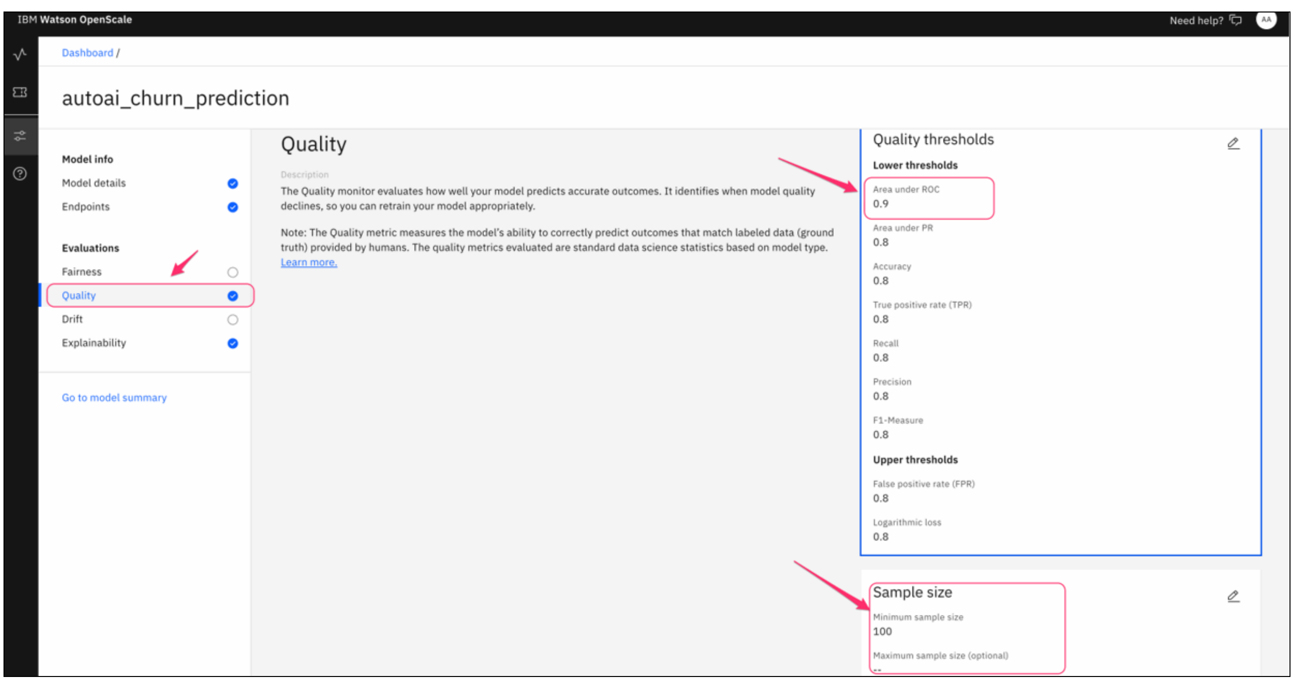

Figure 5-117 Quality threshold setup

Fairness monitoring

In the Fairness monitor, you specify to IBM Watson OpenScale which features to monitor and what are the favorable outcomes. The IBM Watson OpenScale Fairness monitor determines whether outcomes that are produced by your model are fair for the monitored group.

Complete the following steps

1. When Fairness monitoring is enabled, it generates a set of metrics every hour by default. You can generate these metrics on demand by clicking **Check fairness now** or by using the Python client.

   IBM Watson OpenScale automatically identifies whether any known protected attributes are present in a model. When IBM Watson OpenScale detects these attributes, it automatically recommends configuring bias monitors for each attribute present to ensure that bias against these potentially sensitive attributes is tracked in production.

   Currently, IBM Watson OpenScale detects and recommends monitors for protected attributes (sex, ethnicity, marital status, age, and zip code).

2. Select **Fairness** (annotated by a red rectangle in Figure 5-118) and click the **pencil** icon (annotated by a red arrow in Figure 5-118) to configure the fairness monitor.

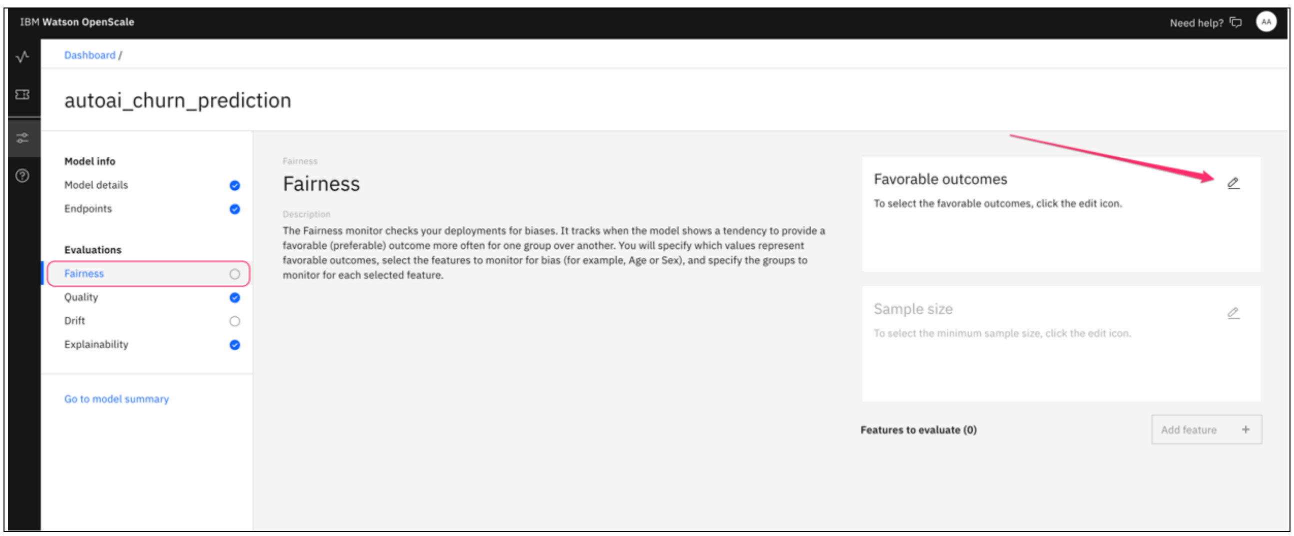

Figure 5-118 Fairness monitor setup

3. Select the favorable outcomes by specifying `F` (false) as the Favorable value and `T` (true) as the Unfavorable value. Click **Next** (see Figure 5-119).

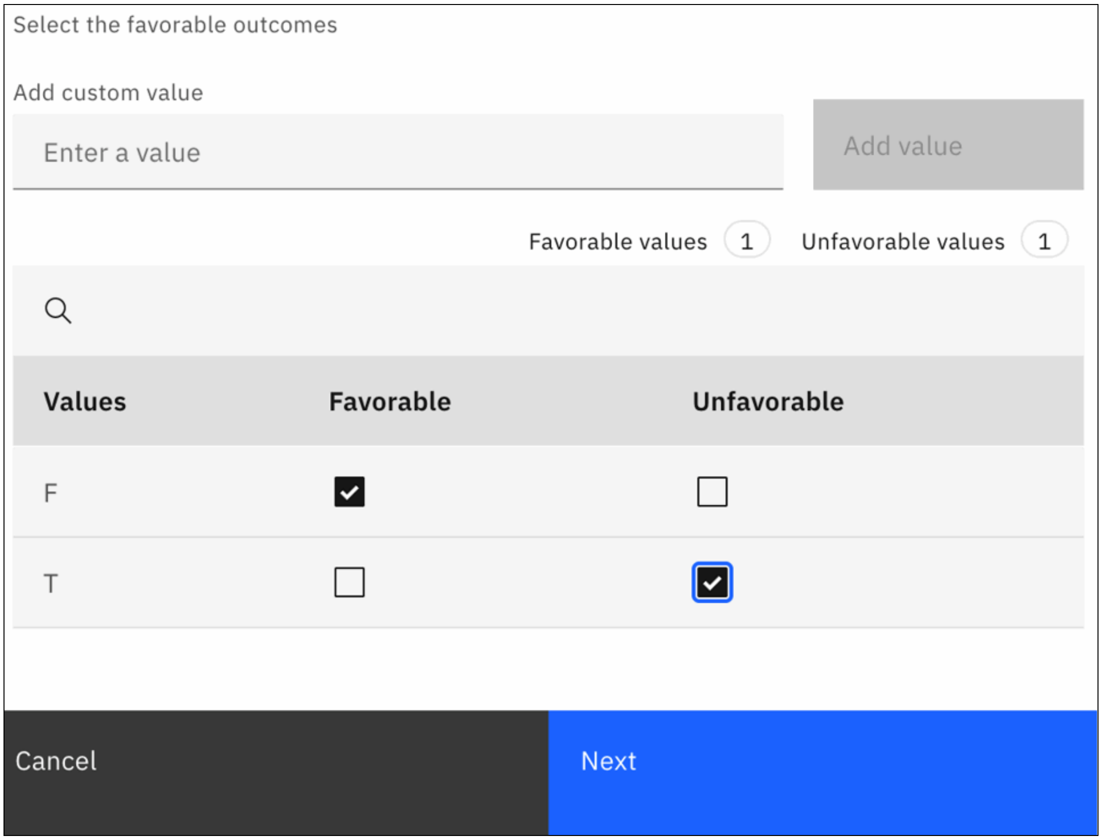

Figure 5-119 Fairness outcome setup

4. Select the minimum sample size to be **100**. In production, you might want to select a larger sample size to ensure that it is representative (see Figure 5-120).

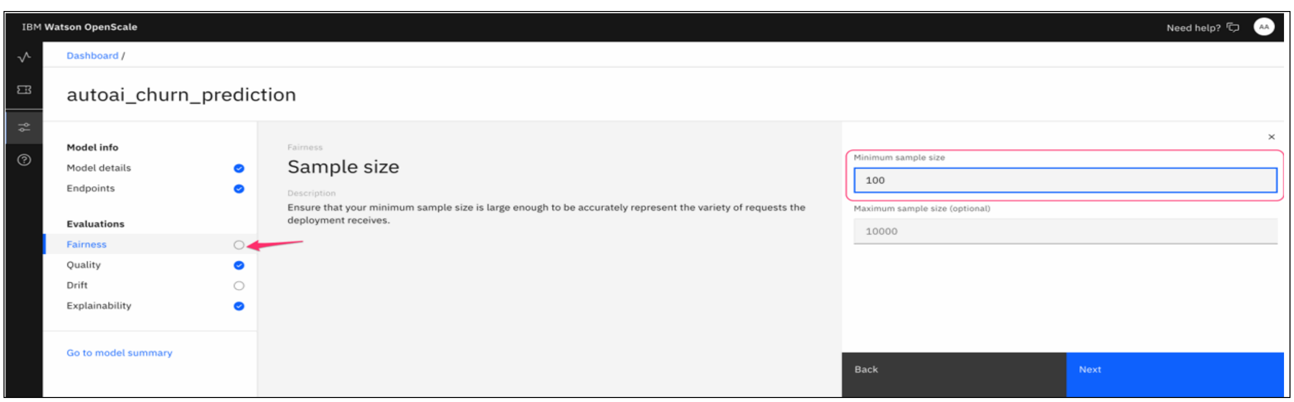

Figure 5-120 Fairness sample size setup

5. On the Select the features to monitor page, keep the default selections of GENDER and AGE features to monitor and click **Next**. These features are monitored for fairness. IBM Watson OpenScale analyzes the training data that we provided and automatically recommends which features to monitor (see Figure 5-121).

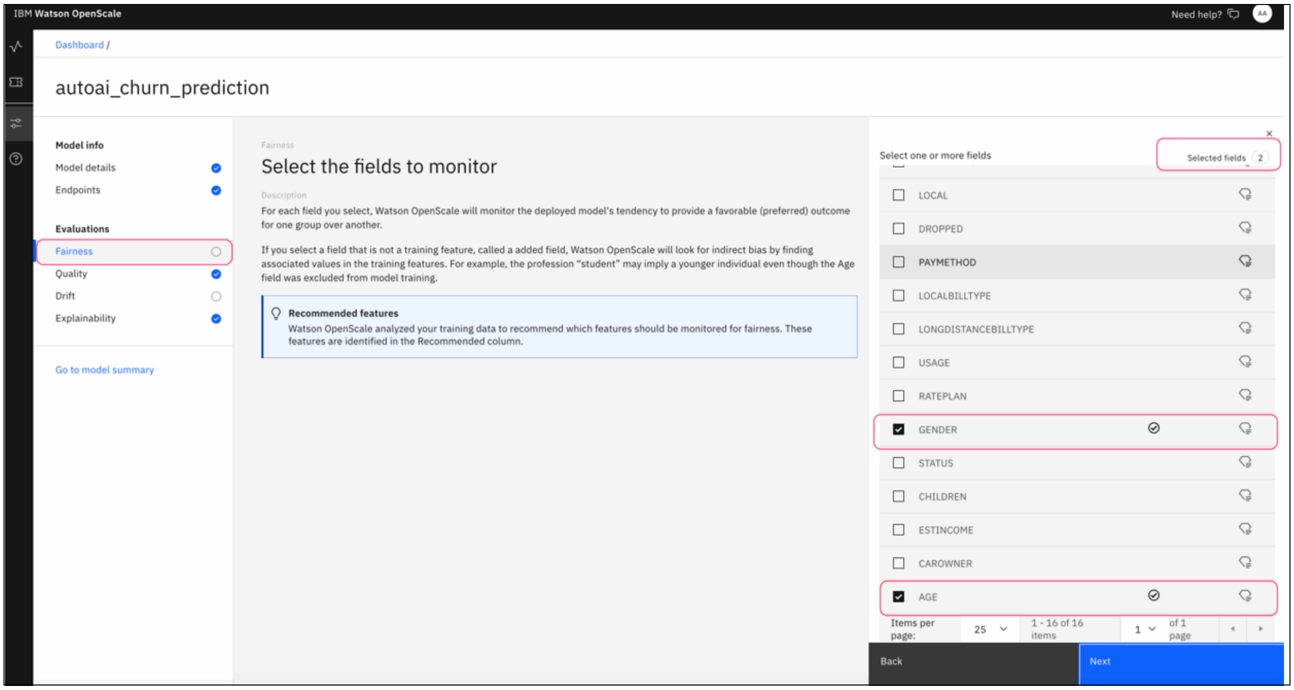

Figure 5-121 Selecting the fields to monitor for fairness

6. For the AGE feature, specify the reference and monitored groups. Again, IBM Watson OpenScale automatically recommends which group is the reference and which groups are monitored by analyzing the training data. Accept the default selections that are by IBM Watson OpenScale.

7. Set the fairness alert threshold to **95**, which effectively indicates that IBM Watson OpenScale raises an alert when the model predicts a favorable outcome for the monitored group 95% of the time less than a favorable outcome for the reference group (see Figure 5-121).

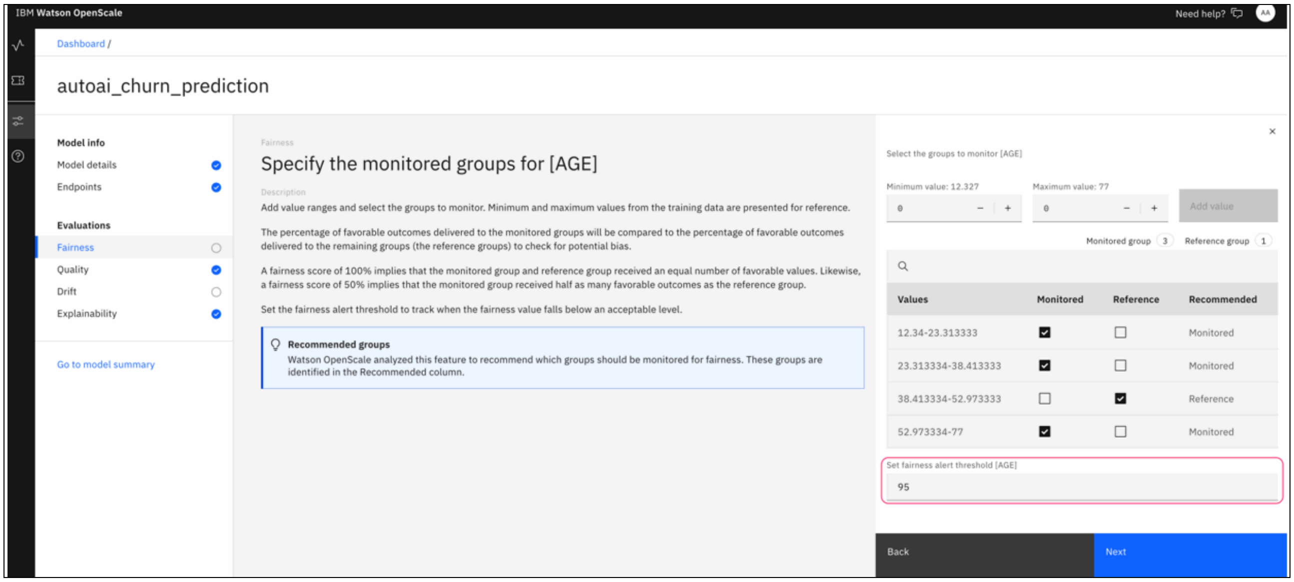

Figure 5-122 Selecting the fields to monitor for fairness

8. For the GENDER feature, specify the reference and monitored groups as `F` (Monitored) and `M` (Reference). Specify the fairness alert threshold to be **95** and click **Save** (see Figure 5-123).

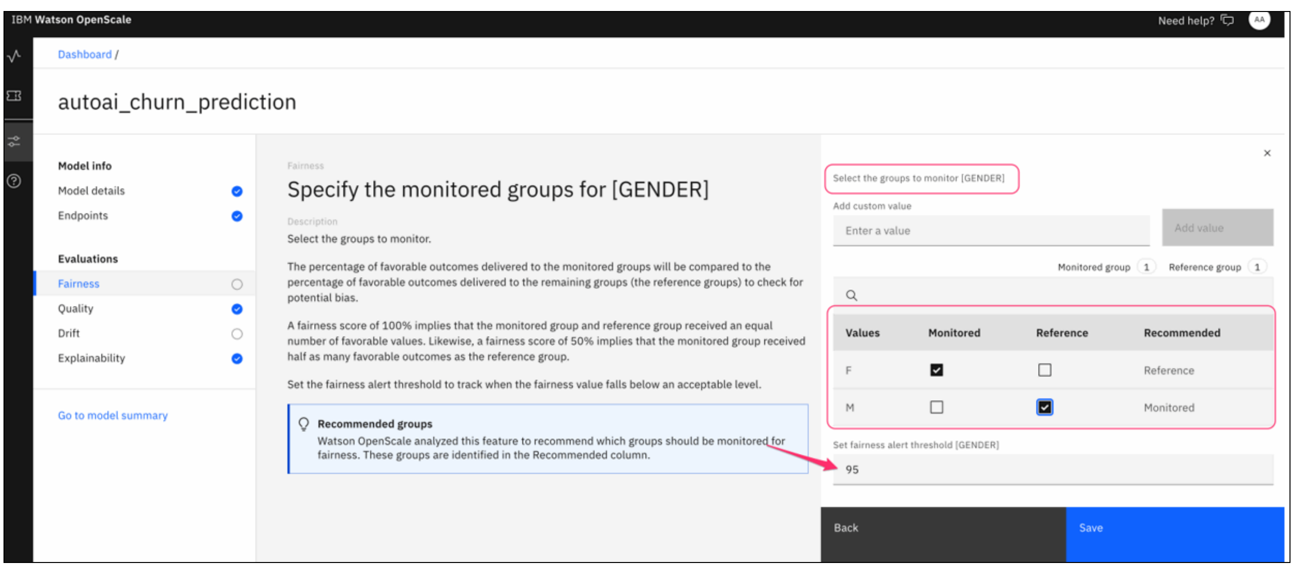

Figure 5-123 Selecting the gender group for fairness

9. Review the configuration information about the Fairness Summary (see Figure 5-124).

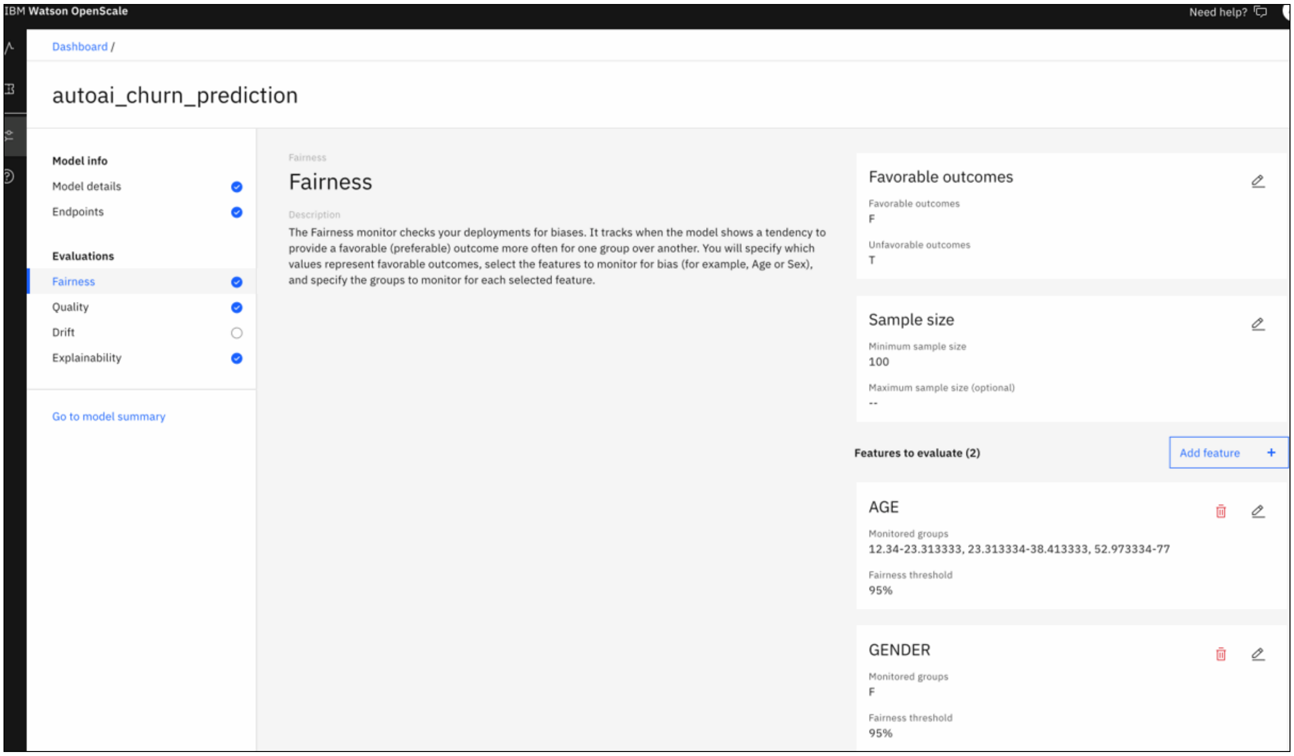

Figure 5-124 Fairness monitoring setup summary

Configuring drift detection

Complete the following steps to configure drift detection:

1. Select the **Drift mode**, as shown in Figure 5-125.

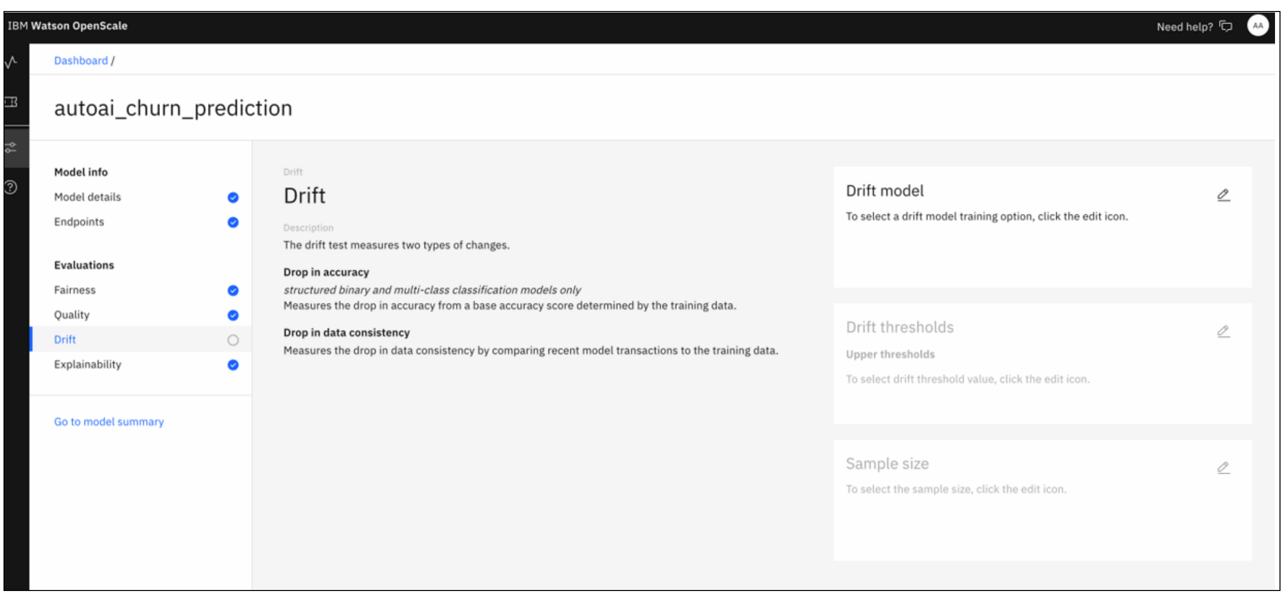

Figure 5-125 Drift monitoring setup

2. Now, the drift model must be trained. In this example, we use the Train in IBM Watson OpenScale method (see Figure 5-126).

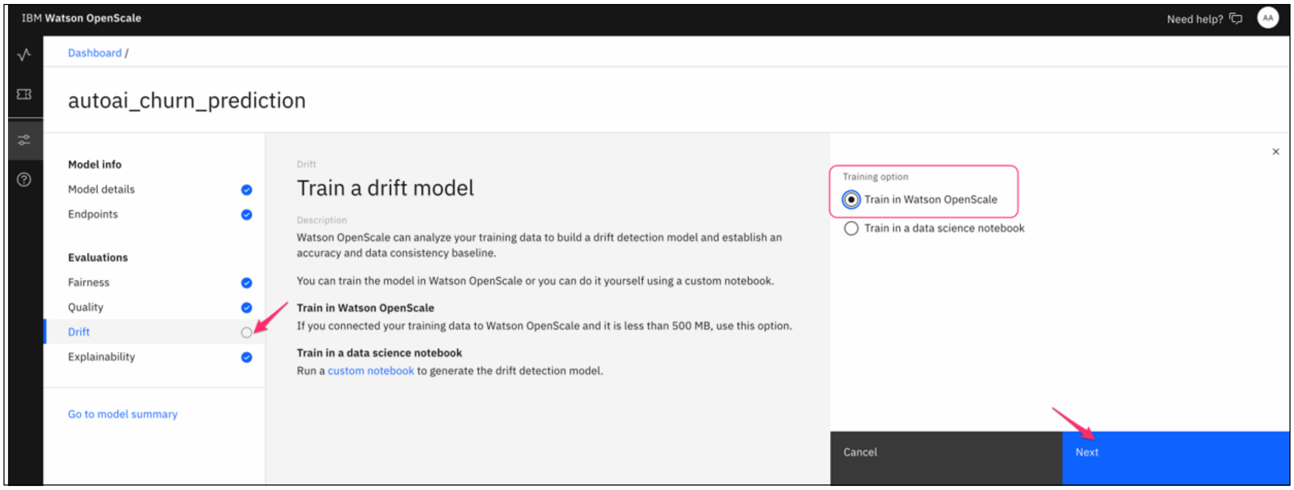

Figure 5-126 Drift monitoring setup

3. Configure the Drift threshold by selecting the threshold drop accuracy and, drop the consistency to 10% for activating the alert, as shown in Figure 5-127.

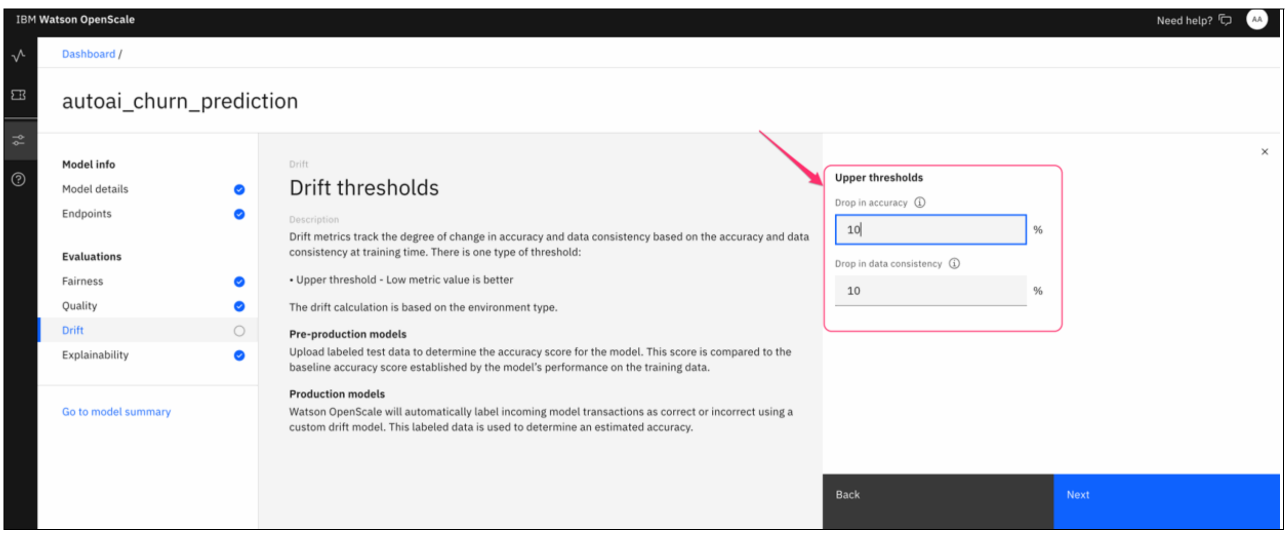

Figure 5-127 Drift monitoring upper threshold setup

4. Configure the minimum drift sample size to 100 (see Figure 5-128).

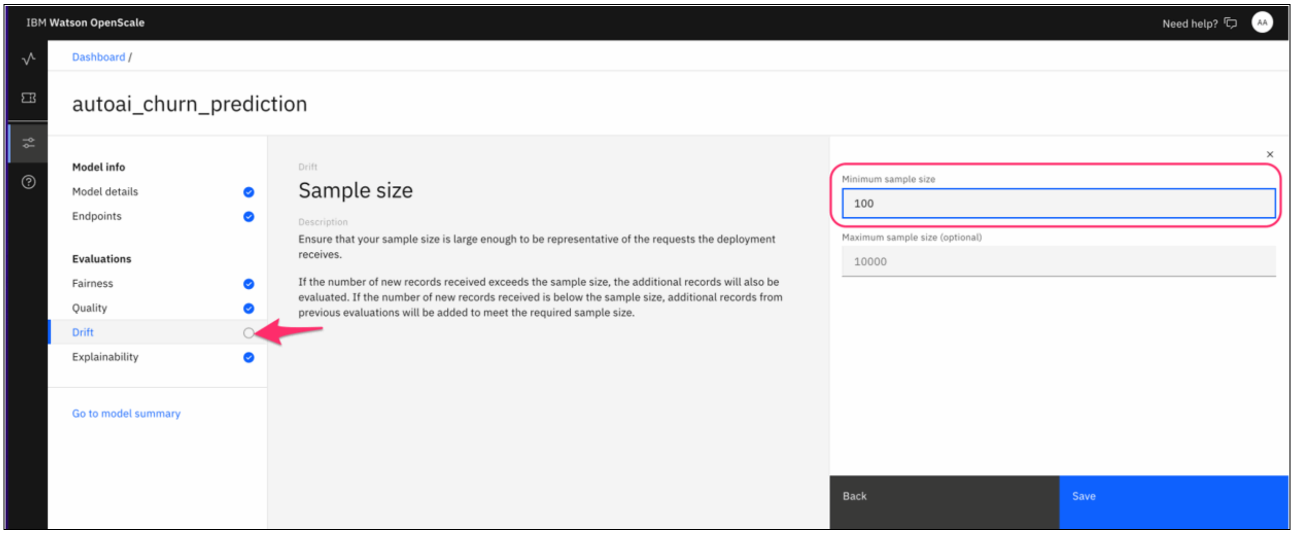

Figure 5-128 Drift monitoring sample size setup

5. Review the Drift monitor configuration summary. When saving the changes, the model training is activated. This process takes several minutes to complete (see Figure 5-129).

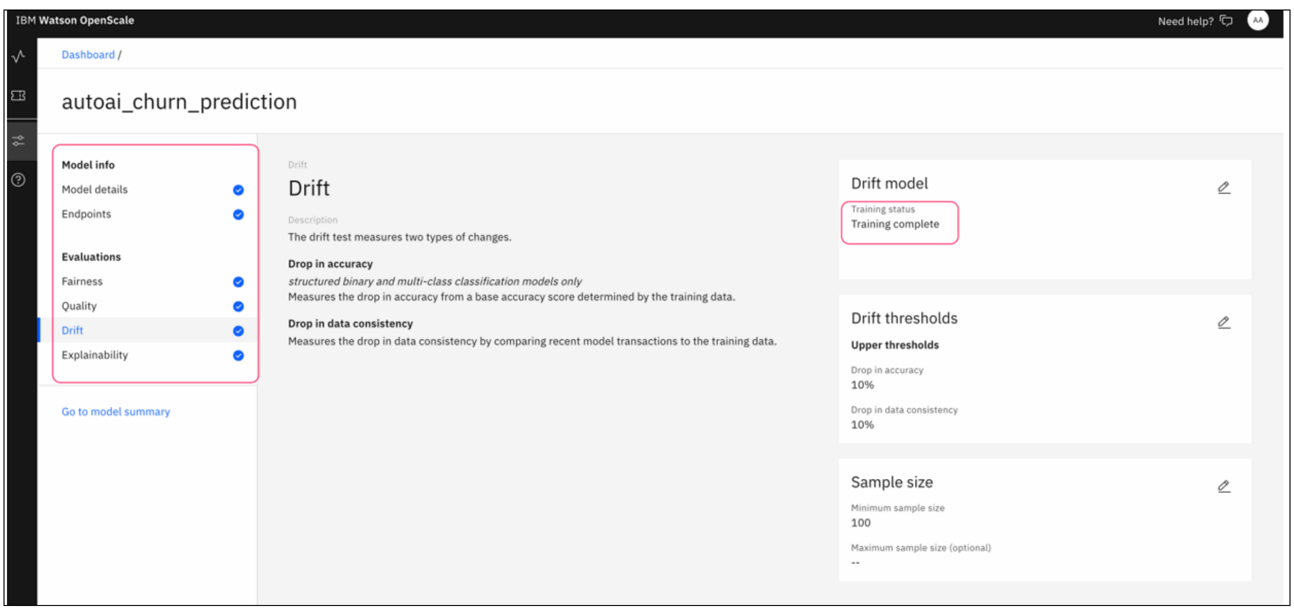

Figure 5-129 Drift monitoring summary setup

With these configurations in place, the process to configure IBM Watson OpenScale configuration is complete for fairness, accuracy, drift, and explainability to monitor the customer churn prediction machine learning model.

In production, as your machine learning model is accessed by applications, IBM Watson OpenScale monitors those scoring events and provides a dashboard (and APIs) that business or AIOps users can use to detect unwanted behavior and establish trust in the AI models.

Evaluating the model performance

Complete the following steps to run an evaluation in IBM Watson OpenScale to evaluate the performance of the model that you deployed:

1. On the churn prediction deployment dashboard, click **Action**. In drop-down menu, select **Evaluate now** (see Figure 5-130).

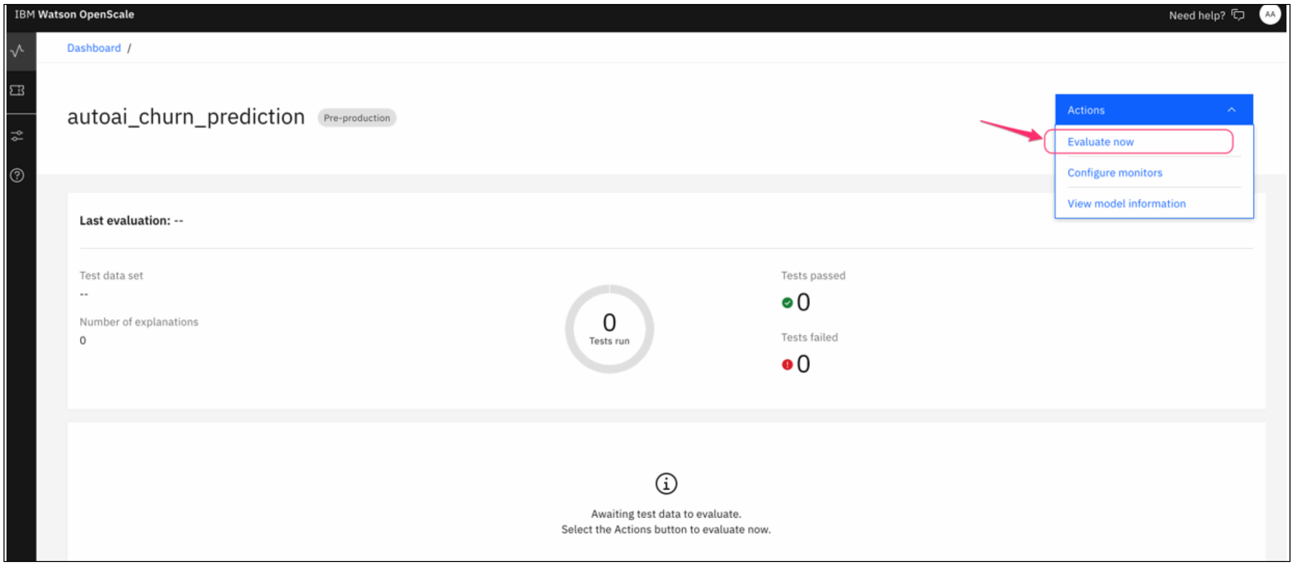

Figure 5-130 AutoAI churn prediction model evaluation

2. On the Import test data page, choose the **from CSV file** option (annotated with red arrow icon Figure 5-131) and click **Browse** to select the `customer_churn_openscale_evaluation.csv` from GitHub. Then, click **Upload and evaluate**.

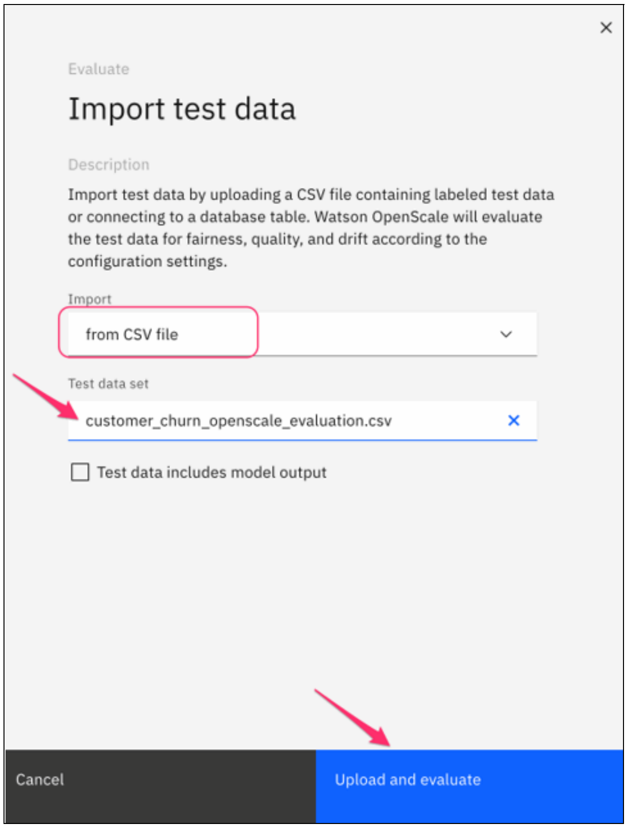

Figure 5-131 AutoAI churn prediction model evaluation: Test CSV data import

3. IBM Watson OpenScale uploads the data, runs scoring against it, and compares the model prediction to the labeled result to compute an overall quality score. It also runs the Fairness monitor to detect any fairness violations (see Figure 5-131).

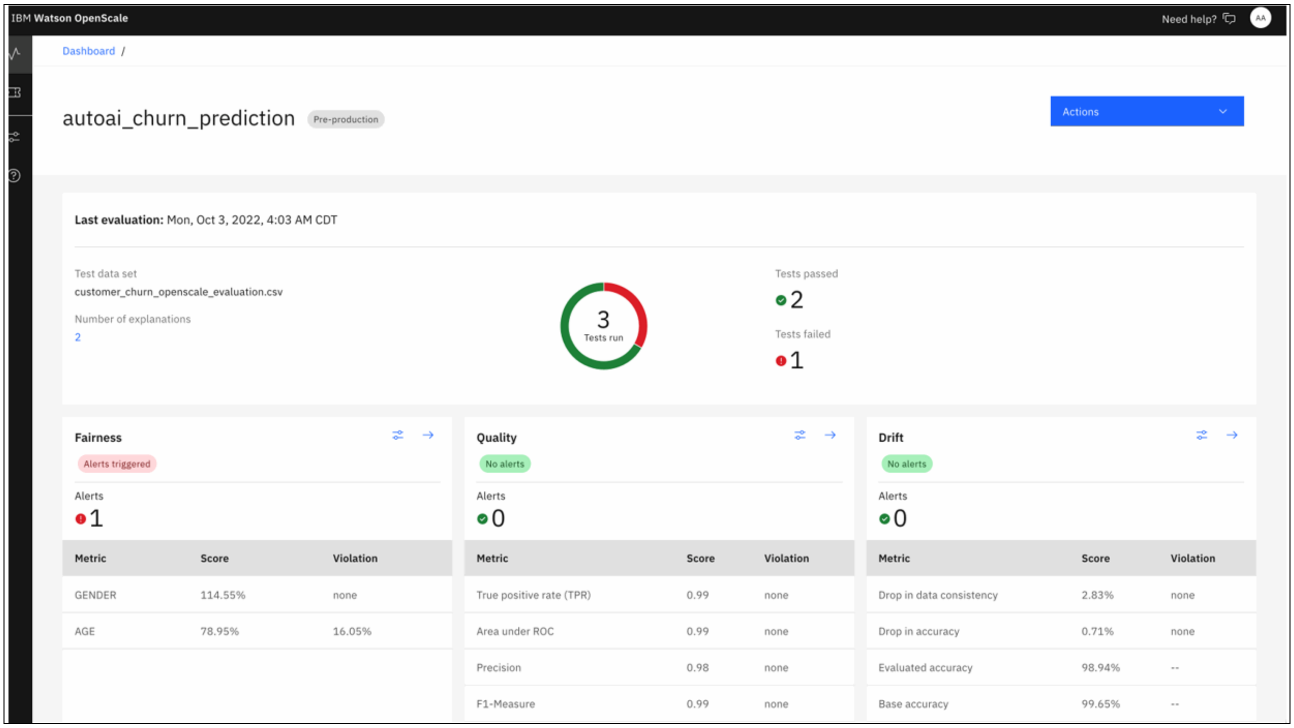

Figure 5-132 AutoAI churn prediction model evaluation summary

After the evaluation is complete, a quick view is shown in the dashboard of the fairness and quality results. In our example, no alerts for fairness or quality is shown, meaning that the model meets or exceeds the required thresholds that were set for those monitors. To investigate the fairness or quality results further, click the arrow that is next to each monitor.

> **Note:** Your results might be different and one or both of these metrics might include an alert.

4. Click the arrow that is next to the **Quality monitor** to review the quality results (see Figure 5-133).

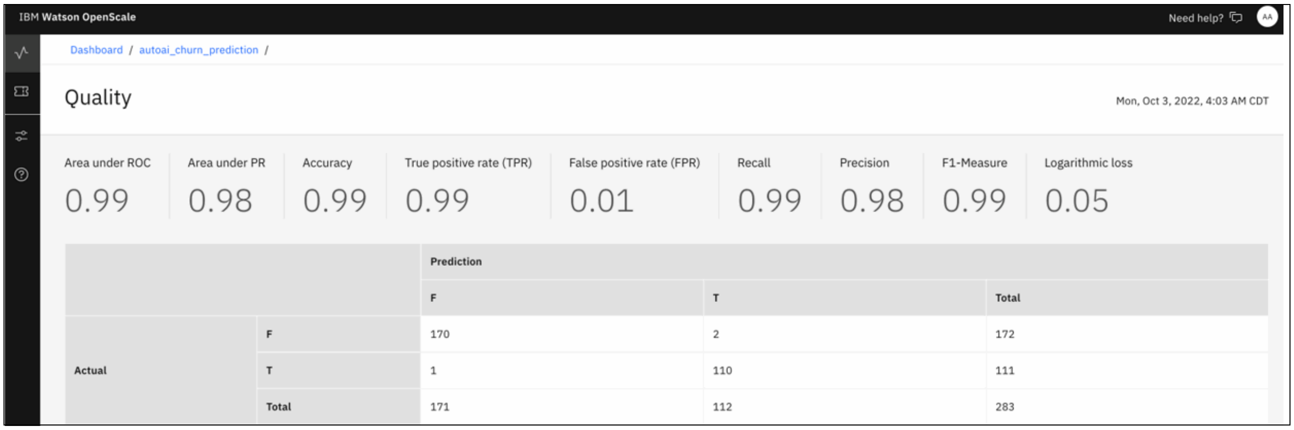

Figure 5-133 Quality monitor results

Reviewing the Fairness monitor results

Complete the following steps to review the Fairness monitor results.

1. Browse to the model dashboard and click the arrow that is next to the **Fairness monitor**. As shown in Figure 5-134, one alert was flagged from the Fairness monitor.

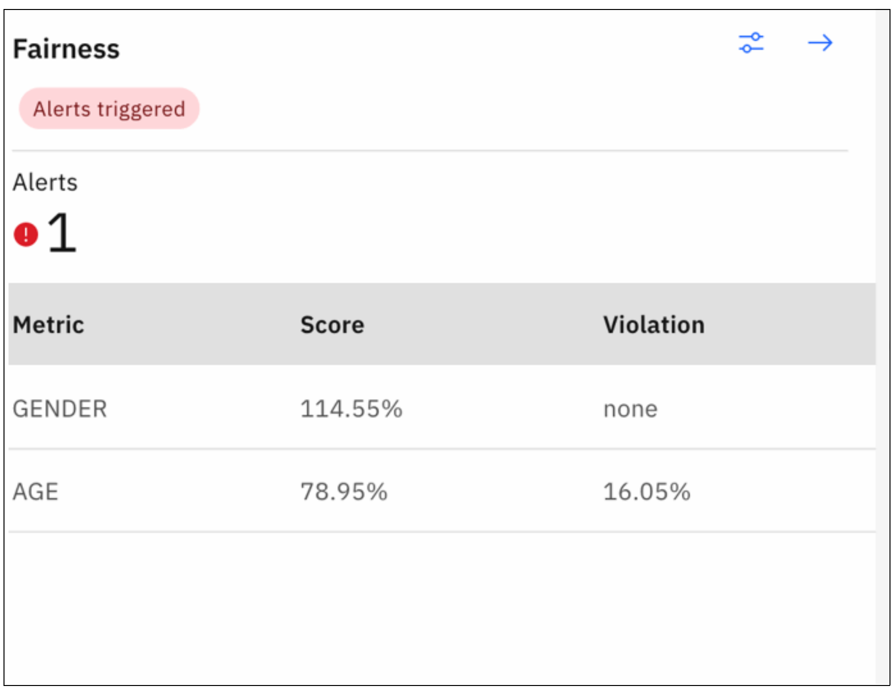

Figure 5-134 Fairness monitor alert

2. Review the fairness results and click **View payload transactions** to review the details of the transactions that were scored (see Figure 5-135).

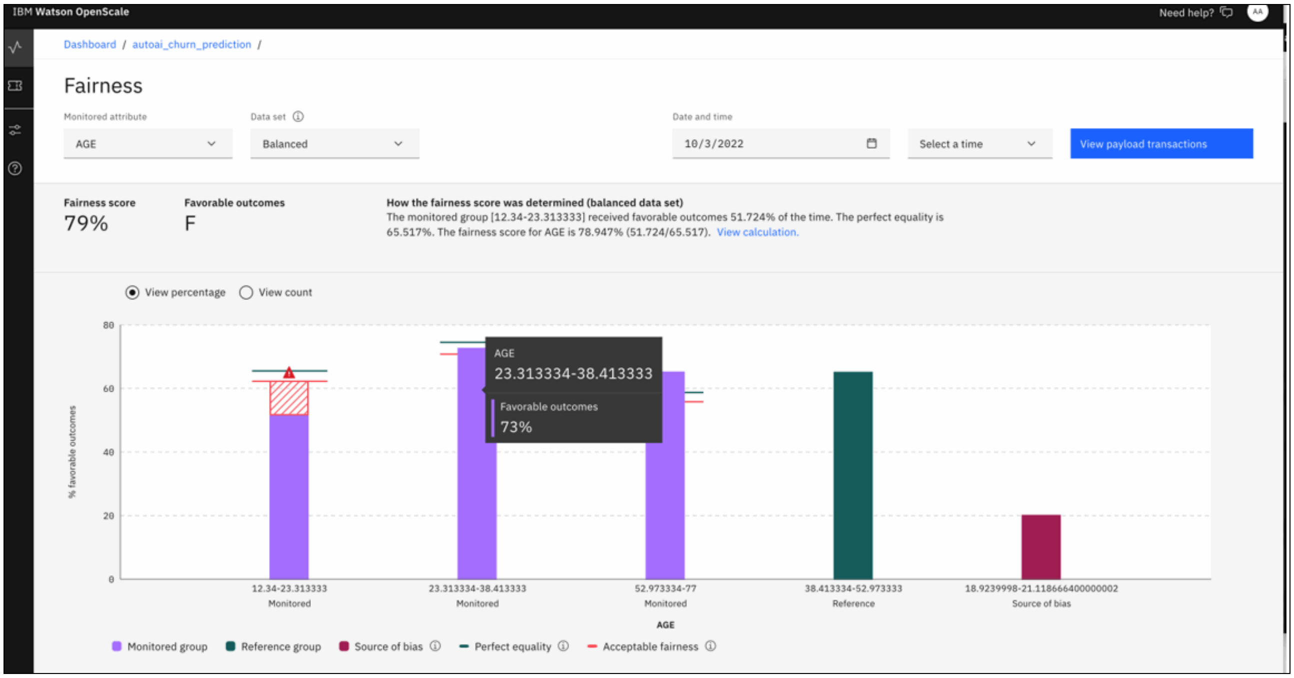

Figure 5-135 Fairness monitor alert details

3. On the Transactions page, review the results. Click **Explain prediction** (annotated with a red arrow in Figure 5-136) for one or more of these transactions to better understand how the model reached the output prediction.

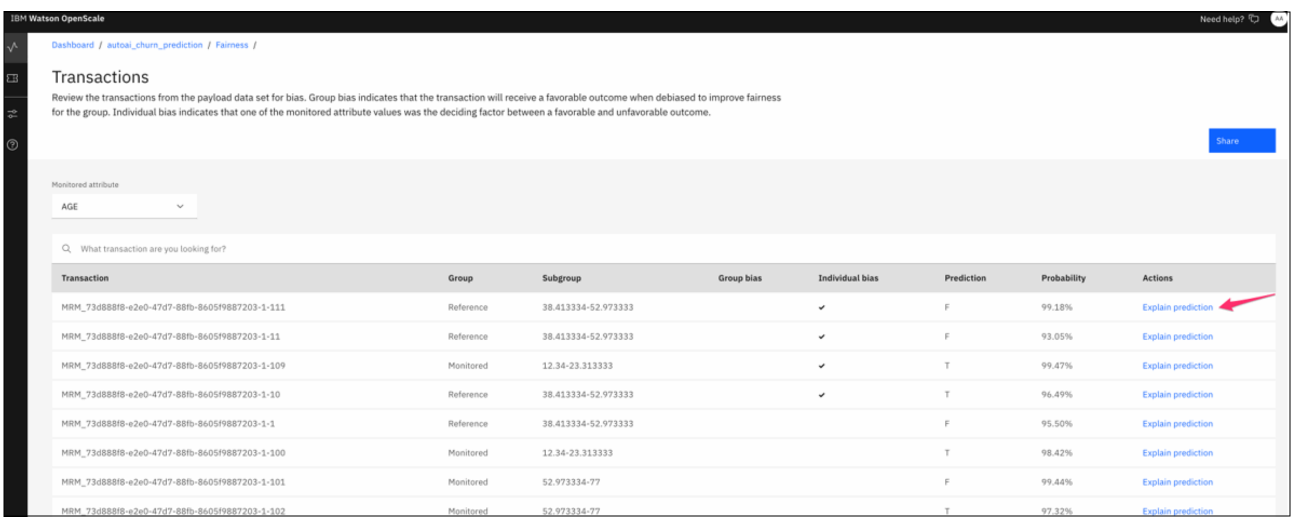

Figure 5-136 Fairness monitor transactions

In the next window (see Figure 5-137). the transaction details are explained. Explainable AI is crucial for an organization in building trust and confidence when putting AI models into production. AI explainability also helps an organization adopt a responsible approach to AI development.

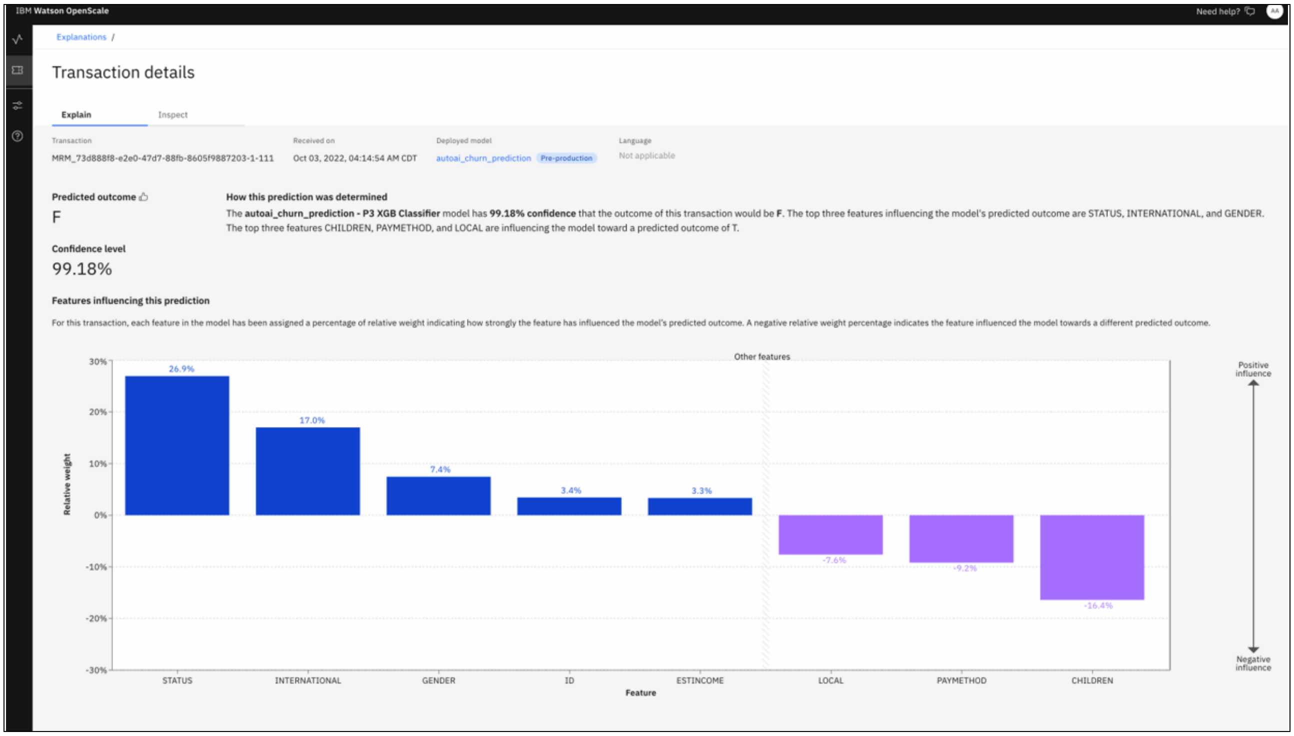

Figure 5-137 Fairness transactions: Explain

Reviewing the Drift monitor results

Complete the following steps to review the Drift monitor results:

1. Browse to the model dashboard and click the arrow that is next to the **Drift monitor**. Review the drift results and click **View payload transactions** to review the details of the transactions that were scored (see Figure 5-138).

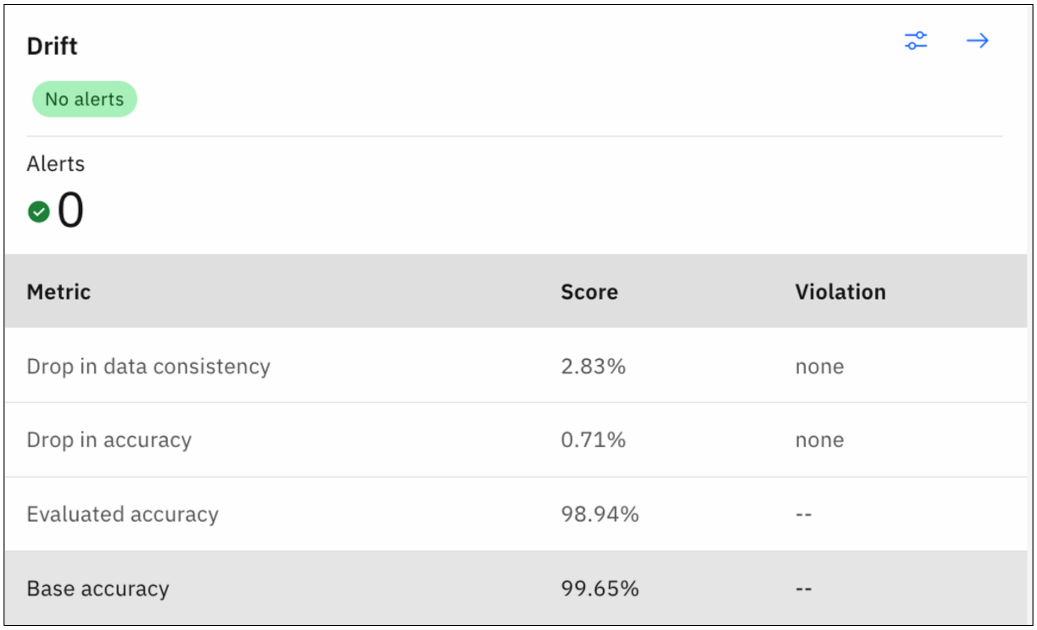

Figure 5-138 Drift monitor result

2. Review drift incorrect prediction results (see Figure 5-139).

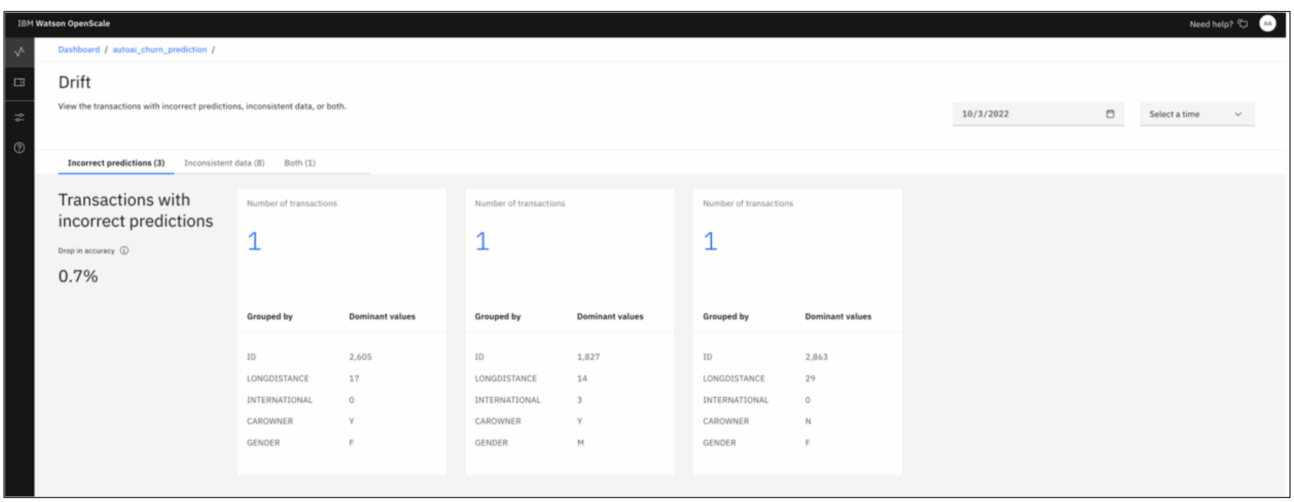

Figure 5-139 Review drift incorrect prediction results

3. Review drift monitor results of inconsistent data (see Figure 5-140).

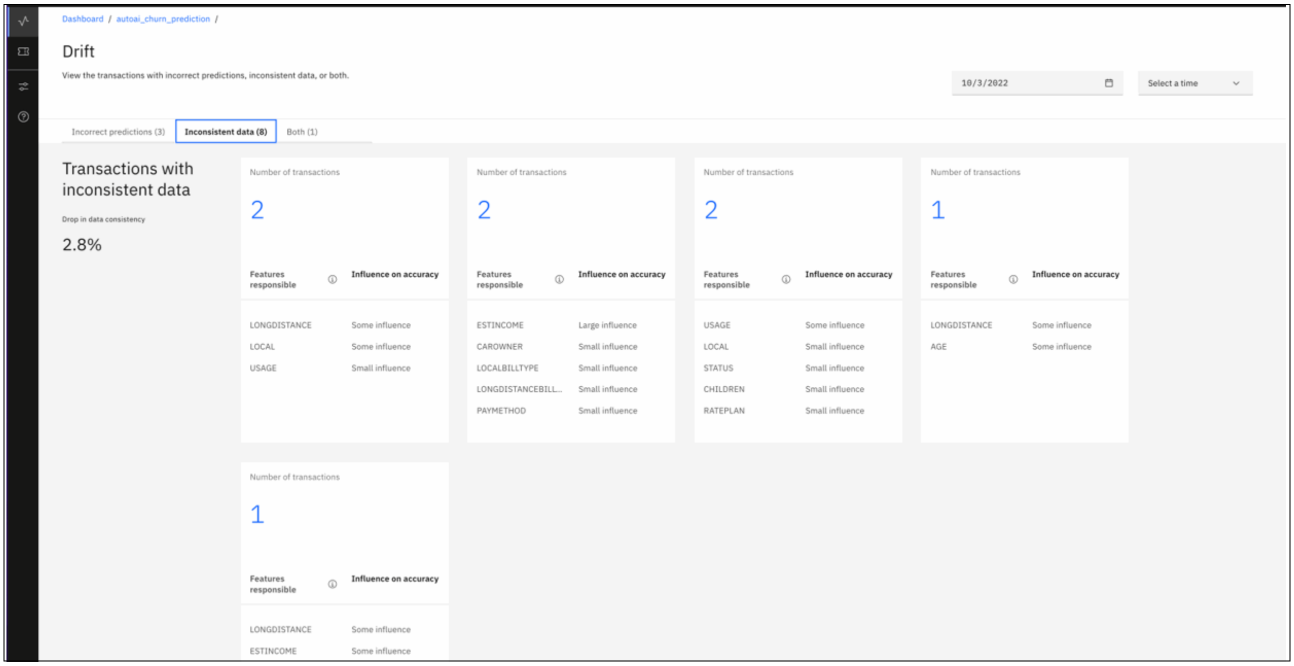

Figure 5-140 Drift monitor - incorrect prediction result

This scenario shows how to use IBM Watson OpenScale capabilities to deliver trustworthy AI by running model evaluations to validate that quality, fairness, and drift metrics are within the configured thresholds. Also, AIOps engineers, data scientists, and business users can trigger explanations of individual transactions to gain confidence in the predictions of the model.

5.4.9 Track models

IBM Watson Knowledge Catalog provides the capabilities for user to track data science models across the organization. View at-a-glance which models are in production, and which need development or validation. Use the governance features to establish processes to manage the communication flow from data scientists to ModelOps administrators.

A model inventory tracks only the models that you add to entries. You can control which models to track for an organization without tracking samples and other models that are not significant to the organization.

To monitor your models for fairness and Explainability, the customer team needs to track the production models to ensure that they are performing well.

The IBM Watson Knowledge Catalog model inventory is shown in Figure 5-141.

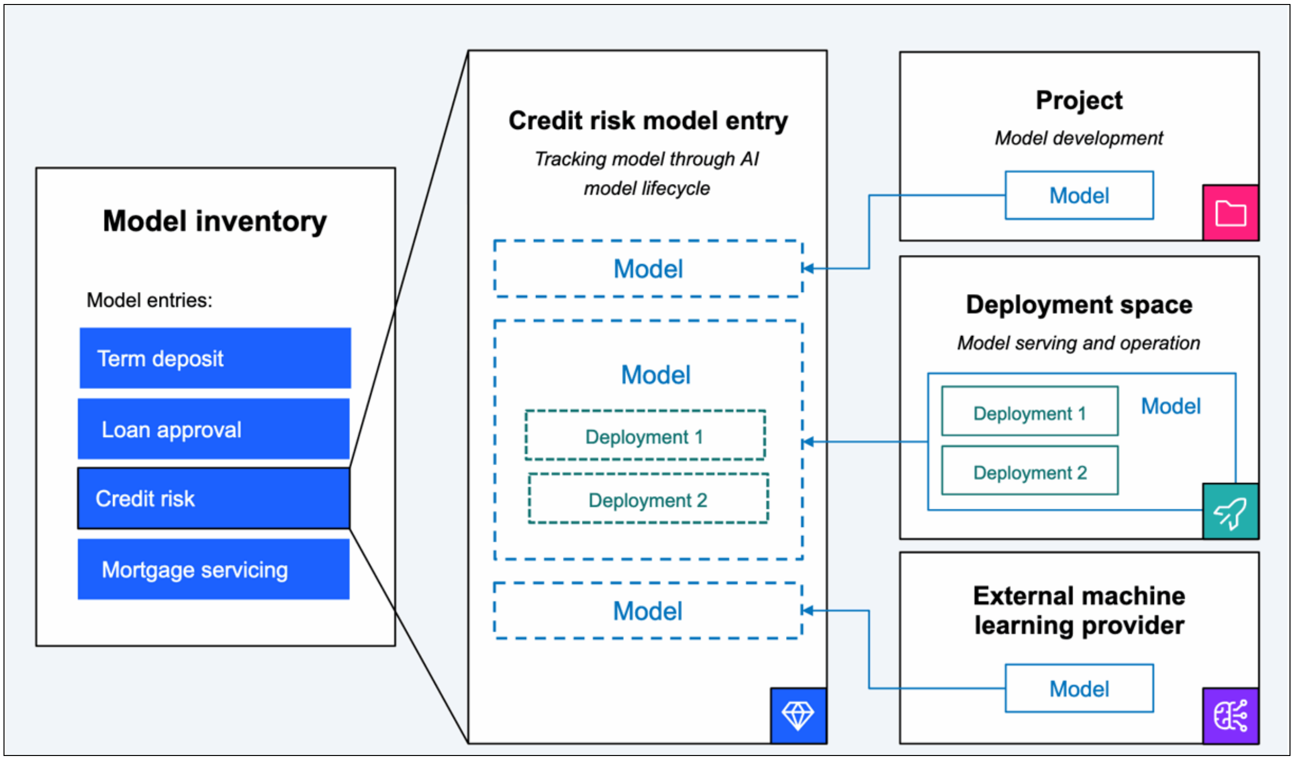

Figure 5-141 IBM Watson Knowledge Catalog model inventory

How the model inventory works

The model inventory is a view in IBM Watson Knowledge Catalog where you can request a new model and then, track it through its lifecycle. The following process shows a typical flow:

1. A business user identifies a need for a machine learning model and creates a model entry to request a new model. The business owner assigns a potential name and states the basic parameters for the requested model.
2. When the request is saved, a model entry is created in the inventory and the tracking begins. Initially, the entry is in the Awaiting development state because no assets are available to accompany the request.
3. When a data scientist creates a model for this business case, they track the model from the model details page of the project or deployment space and associate it with the model entry.
4. The model entry in the inventory can now be moved to a progress state. Stakeholders can review the assets for the entry, which now include the model.
5. As the model advances in the lifecycle, the model entry reflects all updates, including deployments and input data assets.
6. If the data scientist chooses, challenger models can be added to the entry to compare performance.
7. Validators and other stakeholders can review this and other model entries to ensure compliance with corporate protocols and to view and certify model progress from development to production.

In addition to monitoring your models for fairness and explainability, you might want to track the production models to ensure that they are performing well. In the model inventory in a catalog within IBM Watson Knowledge Catalog, you can view lifecycle status for all the registered assets and drill-down to detailed FactSheets for models or deployments that are registered to the model entry.

Creating a catalog entry for FactSheets

IBM FactSheets provide an automated, standardized way to track and store information about model development, testing, validation, and deployment. That data is stored and searchable in IBM Watson Knowledge Catalog.

Complete the following steps to create a catalog entry for FactSheets:

1. From the navigation menu in the upper-left corner, expand **Catalogs** and then, click **Model Inventory**.
2. Click **New model entry** in the top-right of the window to open the New model entry dialog form.
3. Enter a name for your model entry, such as `Customer Churn`, and an optional description. From the **Catalog** dropdown, select the catalog that you use for this exercise. Click **Save** to create the model entry (see Figure 5-142).

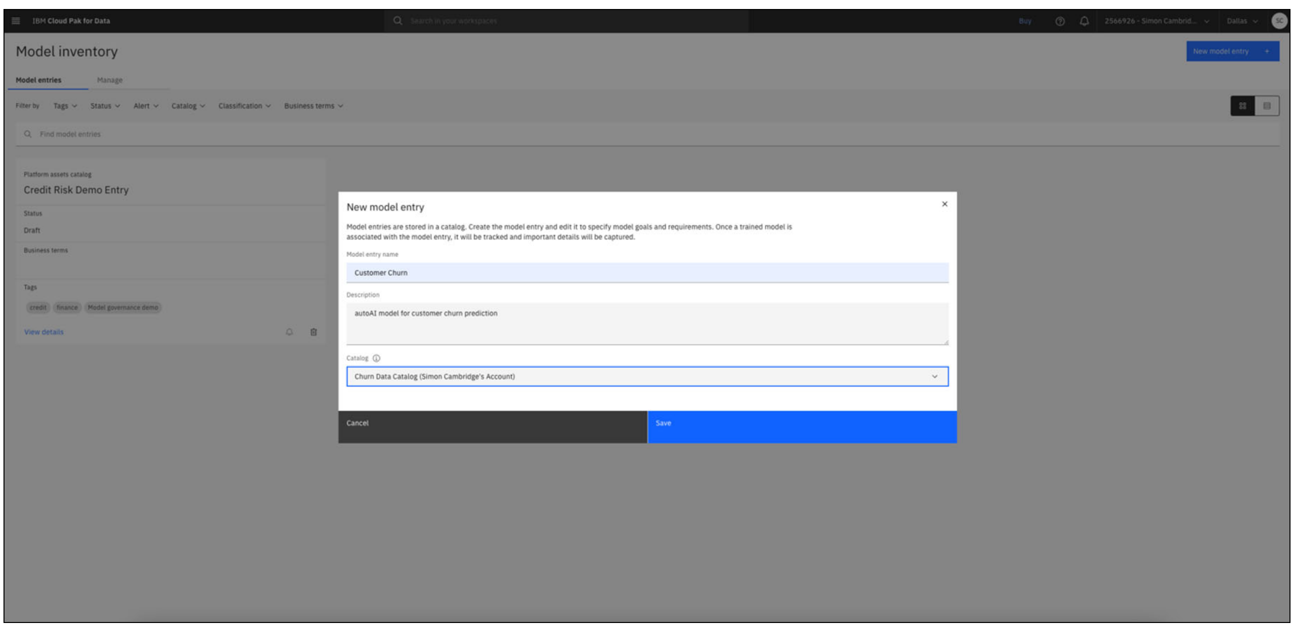

Figure 5-142 Creating a model entry

Examining the FactSheets for models

This exercise assumes that a model was created and saved as described in 5.4.5, "Building models" on page 312. If the model was not created, review that section and create a model.

Complete the following steps to examine the FactSheets for models:

1. From the navigation menu in the upper-left corner, expand **Projects** and click **View all projects**. Then, select your customer churn project.
2. Click the **Assets** tab of the project. Click **Models** from the list of the Assets types that is shown on the left side of the window.
3. The models that you created in 5.4.5, "Building models" on page 312 appear in the list of models. Click the name of one of the models to open the model details window.
4. Click **Track this model**. You now associate the model with the model entry that you created in the catalog (see Figure 5-143).

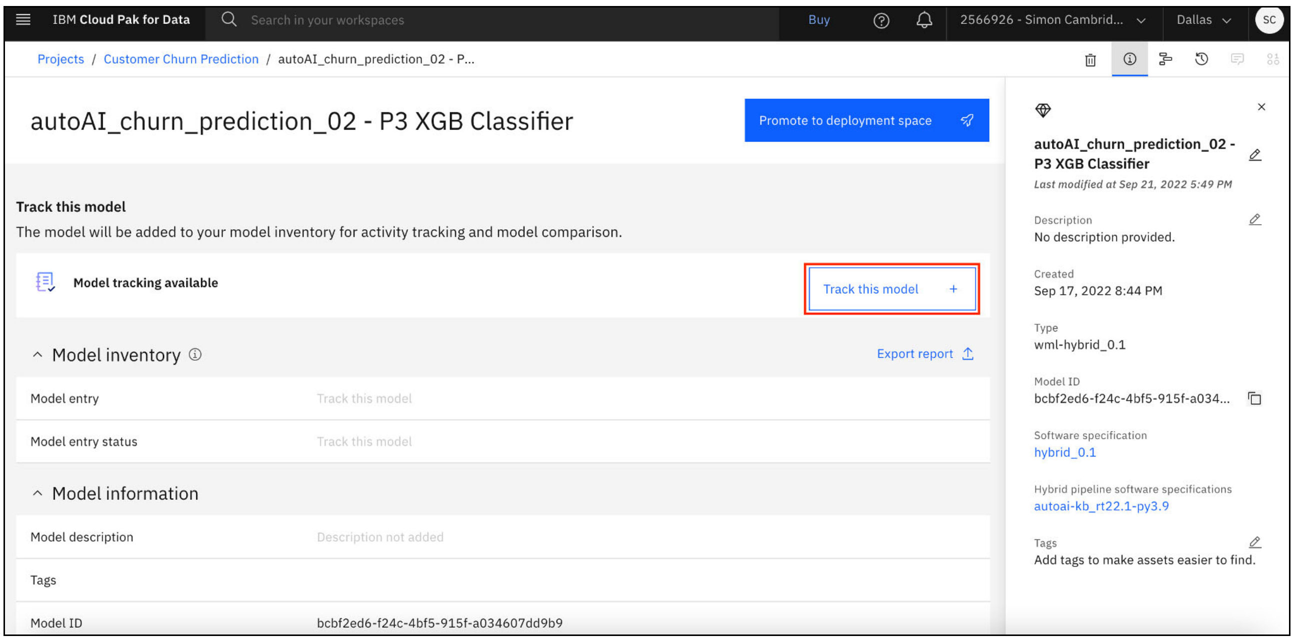

Figure 5-143 Tracking a model

5. Click **Select an existing model entry**. From the list of model entries, select the one you created. Click **Track**. You are returned to the model details page and see that model tracking is now active.
6. After you enabled tracking on the model, from the model information window, click **Open in model inventory**. The catalog entry opens. Click the **Asset tab**.

   The model inventory is divided into four buckets: Develop, Deploy, Validate, and Operate. As your models move through the lifecycle, they are automatically moved to the corresponding bucket. Because the models were just created and not deployed, they are in the Development stage.

7. Click the name of the model in the **Development** bucket. The FactSheet for the model include a great deal of automatically collected metadata (see Figure 5-144):

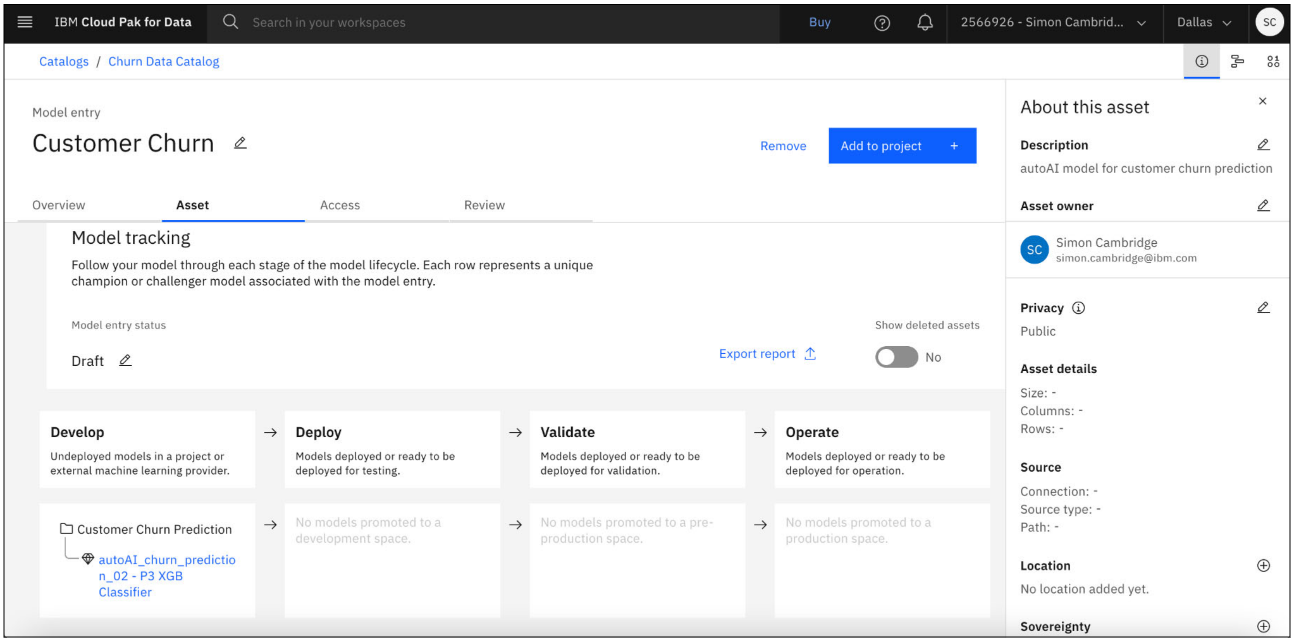

Figure 5-144 View model entry

– The Model information section includes the creation and last modified dates, and the model's type and software specification (see Figure 5-145).

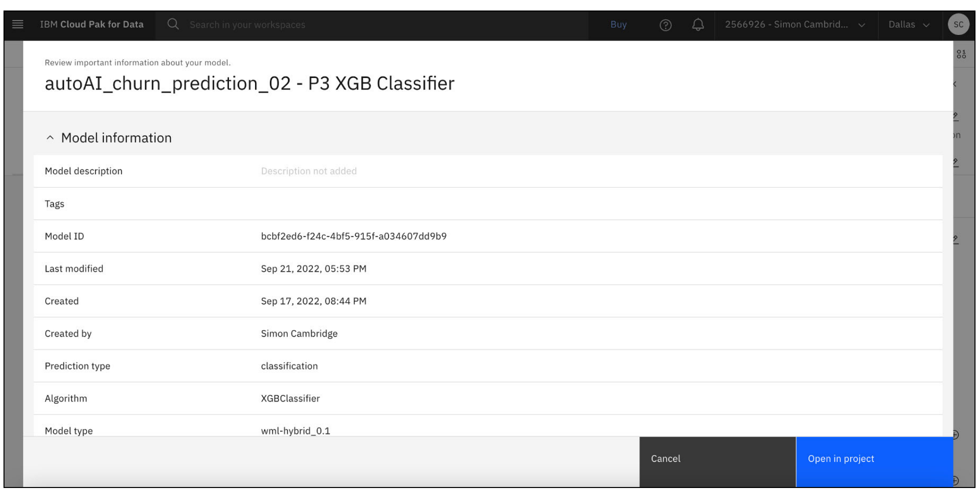

Figure 5-145 Model Information

– The Training information section includes the name of the project that was used to create the model and information about the training data (see Figure 5-146).

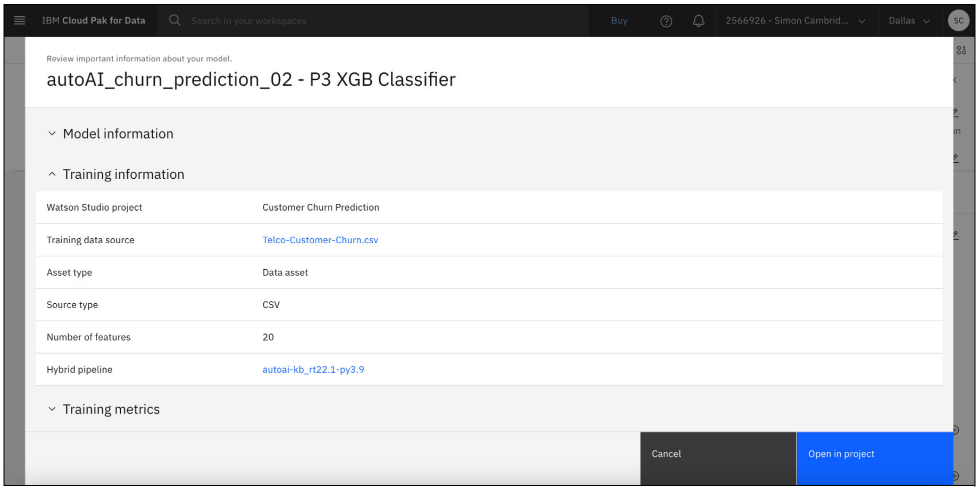

Figure 5-146 Training information

– The Training metrics section includes information that is specific to the type of model you created (see Figure 5-147).

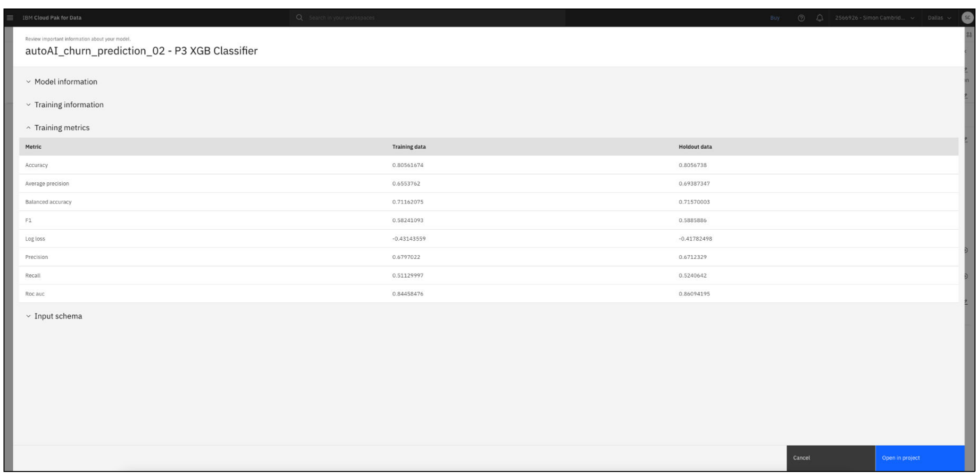

Figure 5-147 Training metrics

– The Input Schema section includes more information, such as the features that were used (see Figure 5-148).

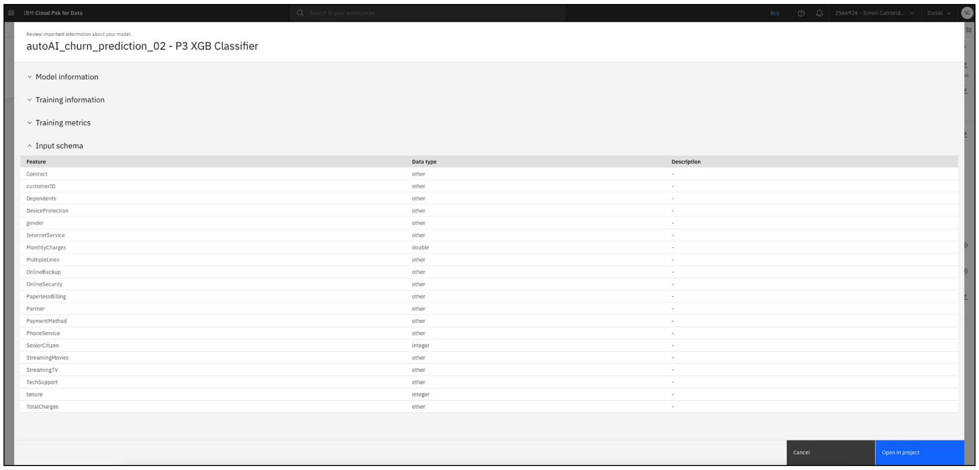

Figure 5-148 Input schema

Summary

This type of information is invaluable for model validators as they seek to understand when and how a model was built. IBM Watson Studio provides a way to standardize and automate collecting metadata; that is, data scientists can spend their time working on meaningful issues instead of collecting, maintaining, and publishing this data.

5.4.10 Automating the ML lifecycle

As organizations scale adoption of AI models in production, it becomes more important to automate the process for testing, validating, and promoting such models from development (DEV) to user acceptance testing (UAT), which also is known as *preproduction*, quality assurance or staging, to production (PRD) environments.

To enable such automation, it is important to validate and monitor the performance of AI models (fairness, quality, drift, and explainability) and automate the process of propagating the models and associated assets from one environment to another.

Previously, you learned how IBM Watson OpenScale can be used to monitor AI models. A typical scenario for developing and promoting AI models includes the following sequence:

1. Data scientists explore multiple algorithms and techniques to train best performing AI model.
2. After they are satisfied with performance results, data science leads deploy the best performing model to a PreProd deployment space.
3. The MLOps team configures IBM Watson OpenScale to monitor and run validation tests against the model deployed in Pre-Prod space.
4. After the model validation is approved, the MLOps team propagates that model from Pre-Prod to Prod.
5. The MLOps team configures IBM Watson OpenScale to monitor the production model for fairness, quality, and drift.

In this section, we describe two common approaches for implementing a governed MLOps methodology to enable the automation of propagating models from development through user acceptance testing (PreProd) to production:

- Propagating data science assets, including trained models from one environment to another.
- Git-based automation in which data assets and source code are checked into a Git repository, Git integration is used for code management, and automation is used for testing and validation in UAT (PreProd) and production environments.

In the next section, we add automation by using IBM Watson Studio Pipelines. Pipelines are critical to avoid human error when performing complex or even trivial tasks.

Choosing which of these approaches to implement is mostly dictated by the use case and the preferred method of governance that an organization chooses to adopt. In this example, we highlight both approaches and how they can be implemented by using Cloud Pak for Data.

Before discussing the details of these approaches, it is helpful to quickly review the overall data science process and various tasks or activities that are performed by the data science team.

Initially, the data science team engages with business stakeholder to discuss the business problem that is to be addressed. After understanding and scoping the business problem, data scientists search and find data assets in the enterprise catalog that might be relevant and useful for training AI models to address the identified business problem.

Data scientists experiment with various data visualizations and summarizations to get a sound understanding of available data. Real-world data is often noisy, incomplete, and might contain wrong or missing values. The use of such data as-is can lead to poor models and incorrect predictions.

After the relevant data sets are identified, data scientists commonly apply “feature engineering”, which is the task of defining and deriving new features from data features to train better-performing AI models. The feature engineering step includes aggregation and transformation of raw variables to create the features that are used in the analysis. Features in the original data might not have sufficient predictive influence and by deriving new features, data scientists train AI models that deliver better performance.

Subsequently, data scientists train machine learning models by using the cleansed data. Data scientists then train several machine learning models, evaluate them by using a holdout data set (that is, data that is not used at training time) and select the best model or multiple models (ensemble) to be deployed in the next phase.

Model building also often includes an optimization step, which aims at selecting the best set of model hyperparameters (parameters of the model). Hyperparameters are then set before training starts to improve model performance.

After data scientists build (train) an AI model that meets their performance criteria, they make that model available for other collaborators, including software engineers, other data scientists, and business analysts, to validate (or quality test) the model before it is deployed to production.

After a model goes through the iterations of development, build, and test, the MLOps team deploys the model into production. *Deployment* is the process of configuring an analytic asset for integration with other applications or access by business users to serve production workload at scale.

Two most popular types of deployment are online and batch:

- Online is a real time request/response deployment option. When this deployment option is used, models or functions are started by using a REST API. A single row or multiple rows of data can be passed in with the REST request.
- Batch is a deployment option that reads and writes from and to a static data source. A batch deployment can be started by using a REST API.

In a hybrid multi-cloud world, DEV, UAT, and PRD environments can be on-premises or in different cloud platforms. For example, the development environment can be hosted in a cloud platform, but the production environment can be on-premises.

Alternatively, the user acceptance testing environment can be on-premises while the production environment can be hosted on a public cloud platform.

For more information about model development and deployment in the AI Lifecycle Management process, see the following resources:

- Academy Publication white paper: *Artificial Intelligence Model Lifecycle Management*
- AI Model Lifecycle Management blogs:
  - Overview
  - Build Phase
  - Deploy Phase

Method 1: AI model propagation across environments

In this section, we describe the first approach that uses **`cpdctl`**, a Cloud Pak for Data command line interface (CLI) tool to automate the process of propagating trained models from one environment to another.

In practice, the environments can exist in the same Cloud Pak for Data cluster or in different Cloud Pak for Data clusters that are hosted on different cloud platforms.

For this lab, we use the same Cloud Pak for Data cluster and show how to use the **`cpdctl`** tool to propagate assets from the quality assurance (QA, which also is known as *user acceptance testing*) deployment space to the production deployment space.

The process is identical to how models are propagated from one cluster to another because the **`cpdctl`** tool is designed to handle the hybrid multi-cloud seamlessly.

Complete the following steps:

1. Log into Cloud Pak for Data as the `datascientist` user.
2. Browse to **Deployments** by clicking the navigation menu (top-left icon) and then, selecting **Deployments**.

3. On the Deployments page, click the **Spaces** tab and then, click **New deployment space** (see Figure 5-149).

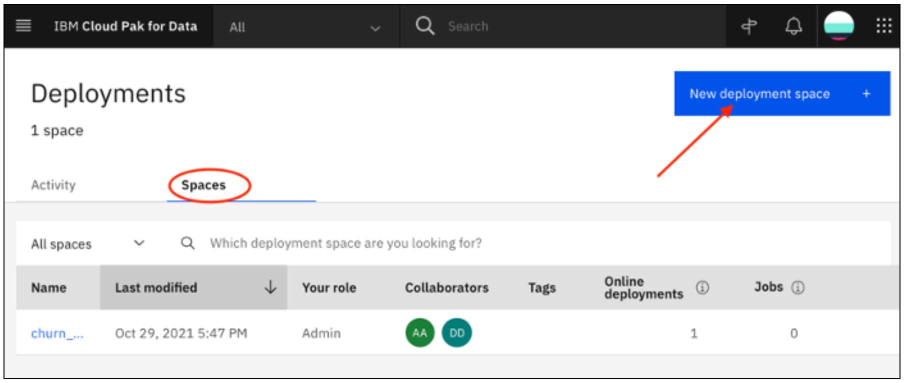

Figure 5-149 New deployment space

4. Enter a Name (for example, `echurn_prod_space`) and a Description (optional) for the deployment space and then, click **Create** (see Figure 5-150).

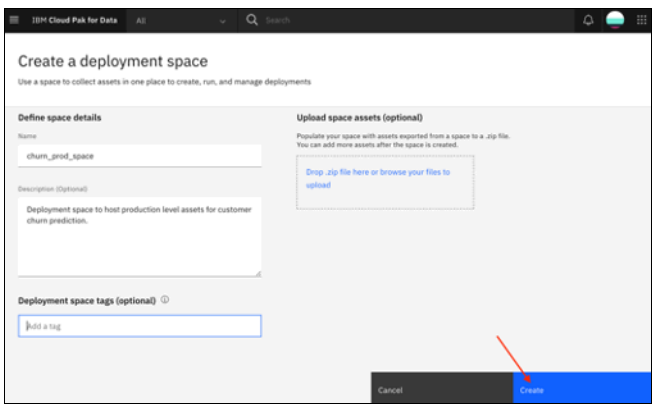

Figure 5-150 Deployment space name

5. Validate that the following deployment spaces are available (see Figure 5-151):

   – `churnUATspace`: Holds assets in quality assurance (or UAT) space.
   – `churn_prod_space`: Holds assets in production space.

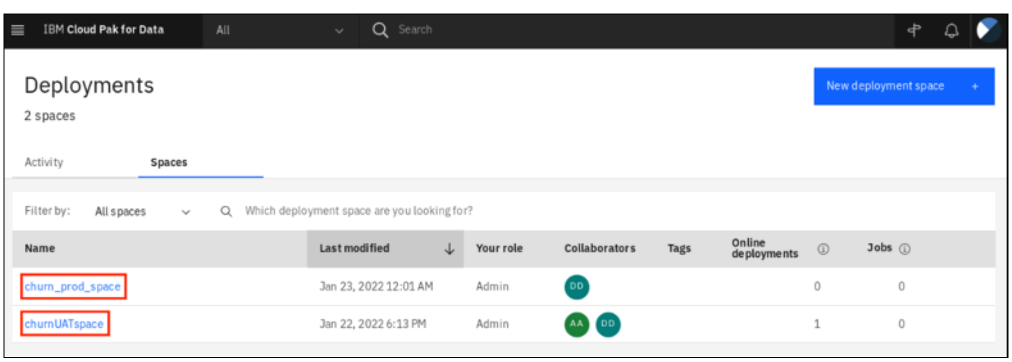

Figure 5-151 Spaces

6. Run a notebook to use the **cpdctl** tool to copy assets from `churnUATspace` to `churn_prod_space`. Return to your Customer Churn Prediction project by clicking the navigation menu in the top-left, selecting **Projects All projects** and then, clicking your Customer Churn Prediction project.
7. Click **Assets** tab and click **New asset** +. (see Figure 5-152).

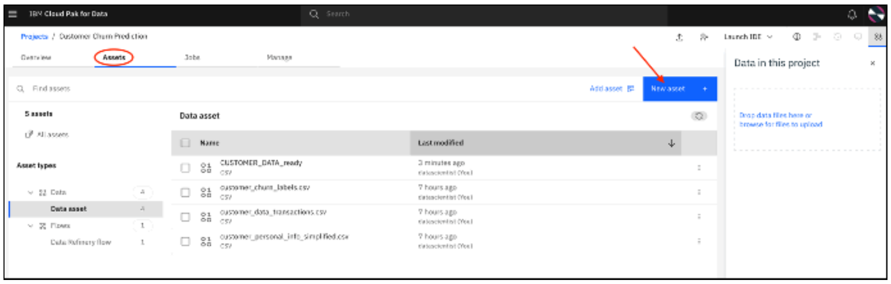

Figure 5-152 New asset

8. Scroll down and select the **Jupyter notebook editor**. You can filter asset types by selecting **Code editors** to quickly find the Jupyter notebook editor (see Figure 5-153).

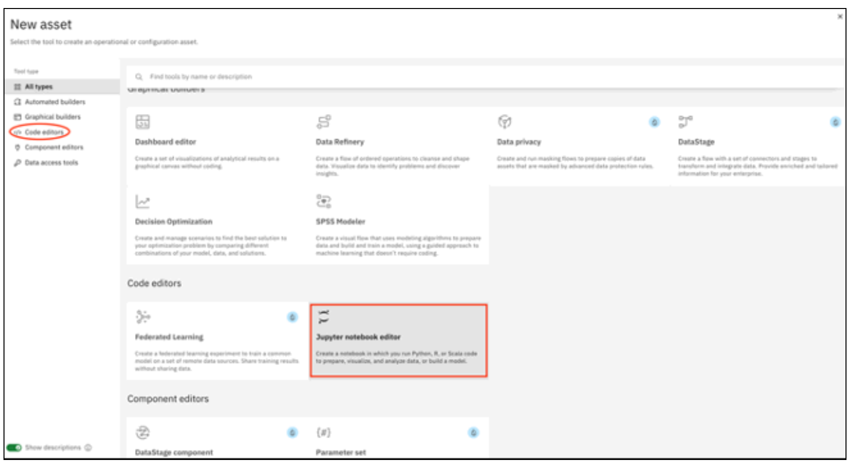

Figure 5-153 Jupyter notebook editor

9. In the New notebook window, click the **From file** tab and then, click **Drag and drop files here or upload**.

10. In the New notebook window, click the **From file** tab and then, browse to find the `CopyAssets_DeploymentSpace1_to_DeploymentSpace2.ipynb` notebook file. Add a Description (optional) and click **Create**. Verify that the selected run time is `IBM Runtime 22.1 on Python 3.9` (see Figure 5-154).

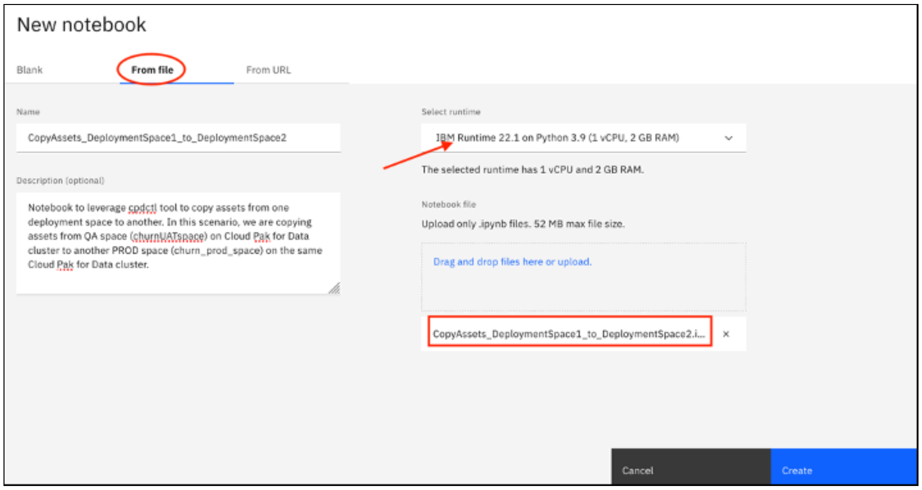

Figure 5-154 Selecting the environment

11. When the notebook loads in edit mode, run the cells of the notebook individually. Read the instructions carefully as you run these steps. You must modify the specific details about the source and target deployment spaces and model names:
    - SOURCE_DEPLOYMENT_SPACE_NAME: `churnUATspace`
    - TARGET_DEPLOYMENT_SPACE_NAME: `churn_prod_space`
    - SOURCE_MODEL_NAME: `Churn Model`
    - TARGET_DEPLOYMENT_NAME: `ChurnPredictionProd`
12. After you run all of the steps in the notebook, return to your Prod deployment space, `churn_prod_space`, to verify that all assets were copied and a new model. Then, browse to Deployments by clicking the navigation menu (top-left of the window).
13. In the Deployments window, click the **Spaces** tab and then, select `churn_prod_space` (see Figure 5-155).

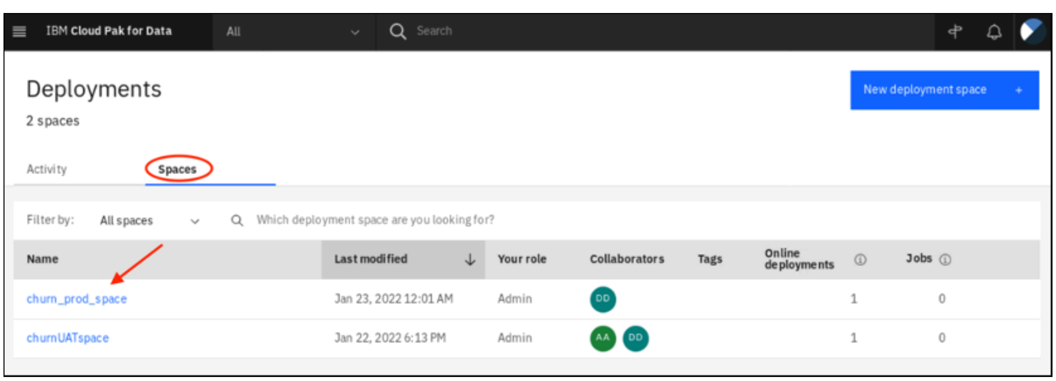

Figure 5-155 Selecting space

14.In the Prod Deployment Space window, select the **Deployments** tab and then, click the production deployment, **ChurnPredictionProd**. The name of your deployment might be different, depending on what you named it in the notebook (see Figure 5-156).

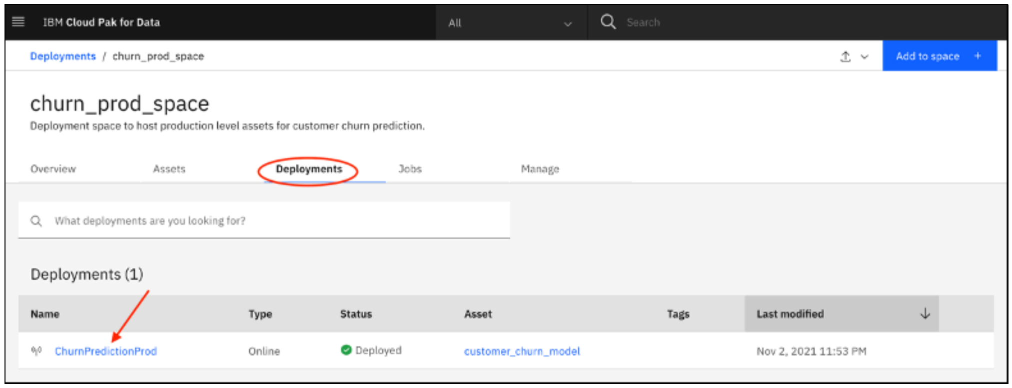

Figure 5-156 List deployments

15.In the Deployed Model window, click the **Test** tab, provide a sample dataset to score in the Body field and then, click **Predict**. Use the sample data that is shown in Example 5-3 as an example and note the prediction output that is shown in the Result section.

Example 5-3 Sample data

```
{"input_data": [
{"fields":
["ID","LONGDISTANCE","INTERNATIONAL","LOCAL","DROPPED","PAYMETHOD","LOCALBILLTYPE"
,"LONGDISTANCEBILLTYPE","USAGE","RATEPLAN","GENDER","STATUS","CHILDREN","ESTINCOME
","CAROWNER","AGE"],
"values":[[1,28,0,60,0,"Auto","FreeLocal","Standard",89,4,"F","M",1,23000,"N",45]]
}]}
```

Edit and change values and rerun the prediction to see how different features can affect the prediction (see Figure 5-157).

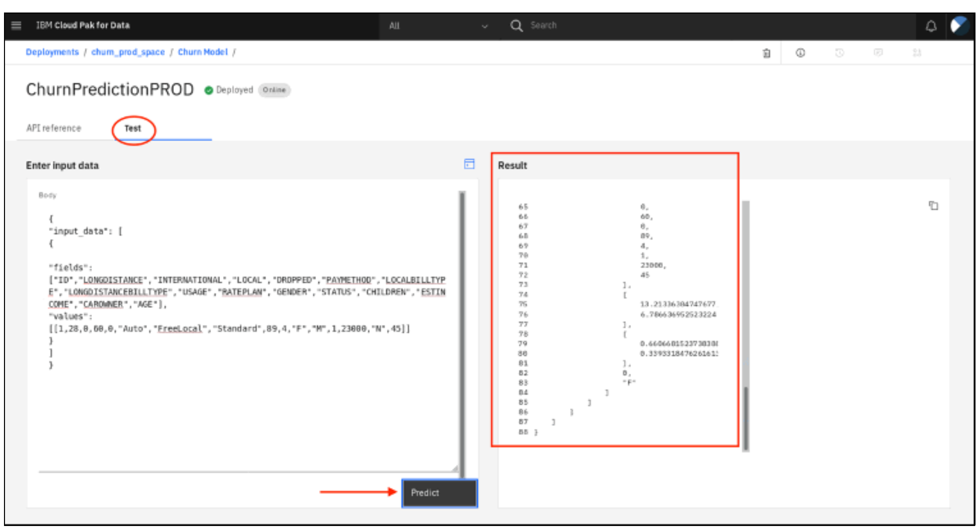

Figure 5-157 Scoring result

Thus far, we showed how to run a notebook by using the `cpdctl` tool to propagate assets from one deployment space to another. You can now create and schedule a job to run periodically and run this sample notebook.

Method 2: Git-based flow

In this part of the scenario, we use Git-based automation where data assets and source code are checked into a Git repository for source code management. Automation is used for testing and validation in UAT and production environments. Effectively, we are propagating source code only and training and deploying models in each of the environments by using notebooks and jobs (see Figure 5-158).

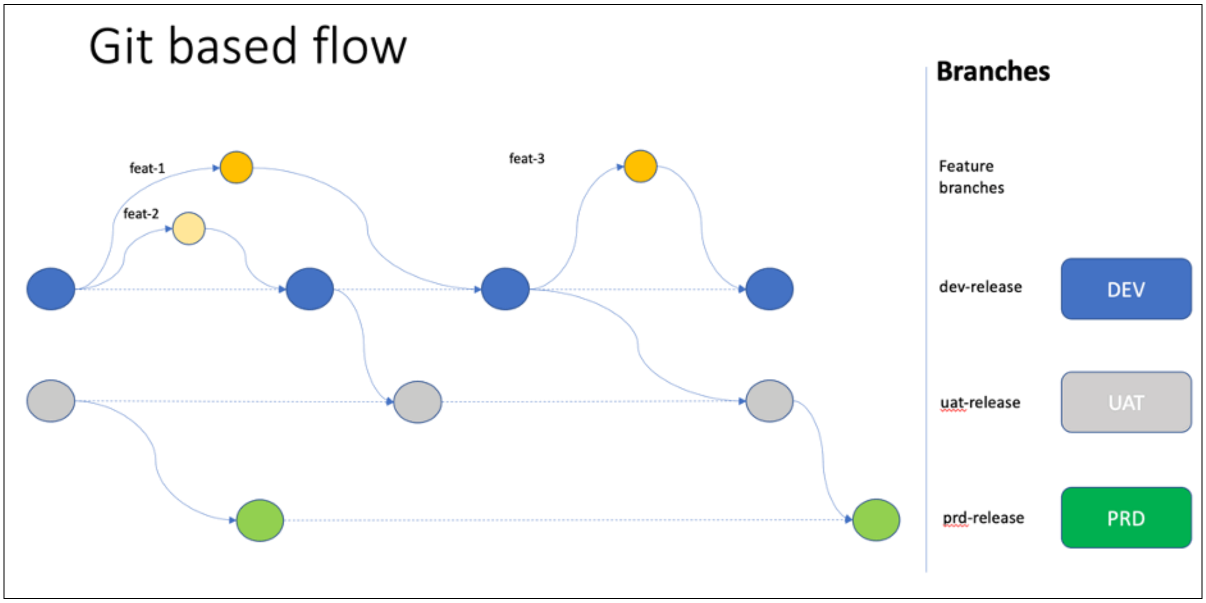

Figure 5-158 Git-based flow

Figure 5-158 shows a typical deployment process of AI models that follow a governed MLOps methodology that is applied through Git integration. Typically, the enterprise use separate clusters or namespaces (depending on isolation needs) to support the various stages of development (training and developing), validation (evaluation and preproduction) and deploying AI (production) models.

The following Git branches are shown in Figure 5-158:

- DEV (Development)
- UAT (User-acceptance testing)
- PRD (Production)

The branches correspond to the development, preproduction, and production clusters. Data scientists mainly operate in the development cluster and interact with the Git DEV branch. Data scientists use the development cluster for experimentation and exploration where they collaborate with other data scientists, data engineers, and business SMEs to identify the correct data sets and train the best performing AI models.

As shown Figure 5-158, multiple forks off of the DEV Git branch add features for improving the AI model. After data scientists are satisfied with the AI model that delivers best performance, they check the code and assets into the Git DEV branch by using pull requests.

The data science project lead, who owns the DEV Git branch, approves the submitted pull requests and tests the notebook and model.

After a review (a few back-and-forth interactions might occur), the lead data scientist creates a pull request (sometimes also referred to as *merge request*) to propagate the assets (notebooks) to the UAT Git branch for testing in the UAT environment, which typically references different data stores than the DEV environment.

Deployment of the assets in the UAT environment (from the UAT branch) typically is done by using automation, also known as *GitOps*. A general policy in many organizations mandates that deploying applications, which includes data science assets, is always fully automated without human intervention. This mandate helps to streamline the process, while also reducing the risk of failing installations because the same process is run in many stages before it reaches production.

Automation pull the assets from the UAT Git branch into the preproduction cluster to retrain the AI model and run validation testing against such a model. The data that is used for validation is different from the data that is used for training and initial testing of the AI model. Validation is run on the data assets in the UAT branch.

After UAT validation tests are completed, the final assets (code and data assets) are checked into a production Git branch by using a pull request. The MLOps team lead reviews and approves the pull request to propagate the assets into the production Git branch.

Automation picks up the assets from the production Git branch and pushes those assets into the production cluster where the code is run to retrain the AI model and validate the performance. Assuming that all performance checks meet expected targets, the model is deployed in the production cluster and is ready for inferencing at scale.

Preparing the Git environment

We already prepared a Git repository that consists of three branches: PRD (production), UAT (UAT) and DEV (development). For more information, see this GitHub web page.

Complete the following steps to create a fork of this repository to your own GitHub organization, which you also create:

1. Log in to `https://github.com`.
2. Click your user at the top-right of the window and click **Your organizations** (see Figure 5-159).

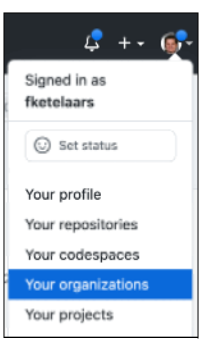

Figure 5-159 GitHub organizations

3. Create an organization and select **Free plan**.

4. Enter the Organization account name; for example, `MLOps-<your_full_name>`. The name must be unique. Also, enter your email address (see Figure 5-160).

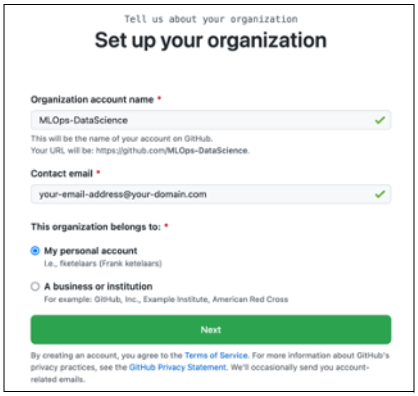

Figure 5-160 Organization name

5. In the next window, click **Complete Setup** and then, click **Submit**.
6. You now have a new github.com organization that you can use to fork the training repository. Go to `https://github.com/CP4DModelOps/mlops-churn-prediction-45` and click the **Fork** button at the right top of the page (see Figure 5-161).

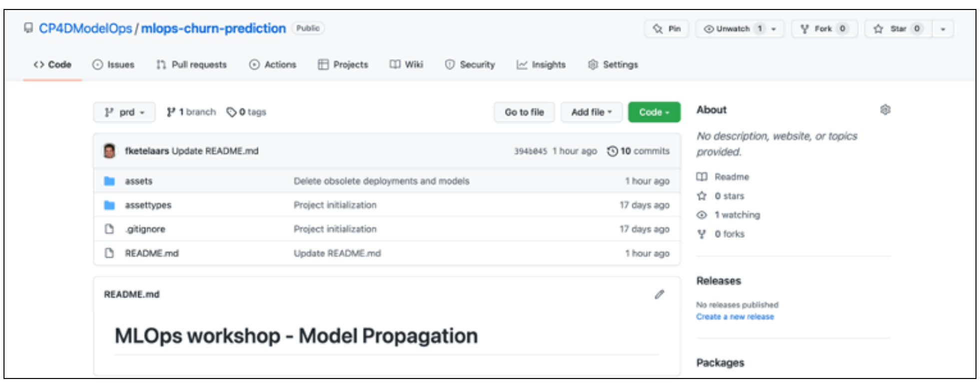

Figure 5-161 Fork repository

7. Select the organization that you just created from the list and ensure that you deselect the **Copy the prd branch only** option. This step is important because we want to fork all branches, not just the PRD branch. Click **Create fork**.
8. GitHub forks the repository and then shows it. Notice that the repository has three branches, as shown in Figure 5-162.

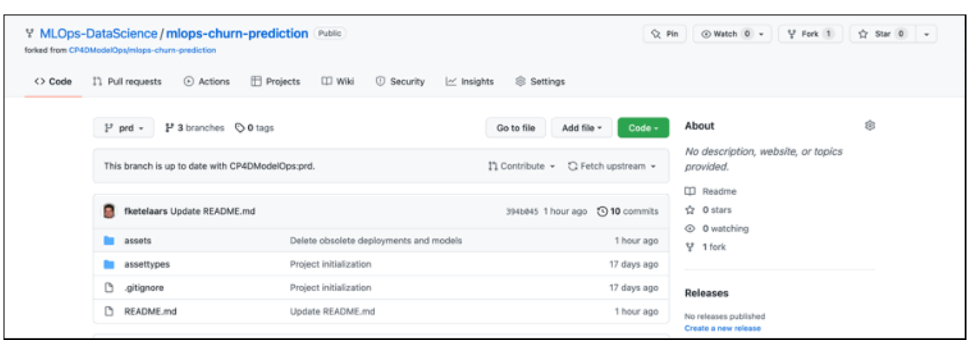

Figure 5-162 Forked repository

9. Click the area that states **3 branches** and check that you have three branches: PRD (the default), UAT, and DEV (see Figure 5-163).

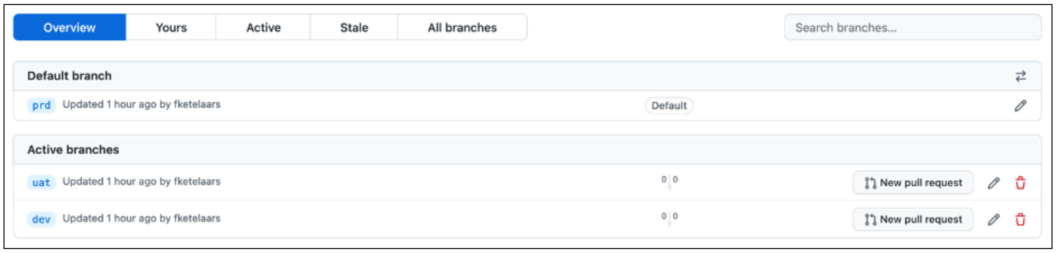

Figure 5-163   Repository branches

10. Set up a branch protection rule to simulate a real-world scenario in which pull requests must be approved before being merged. Click **Settings** and then, click **Branches** on the left side of the window (see Figure 5-164).

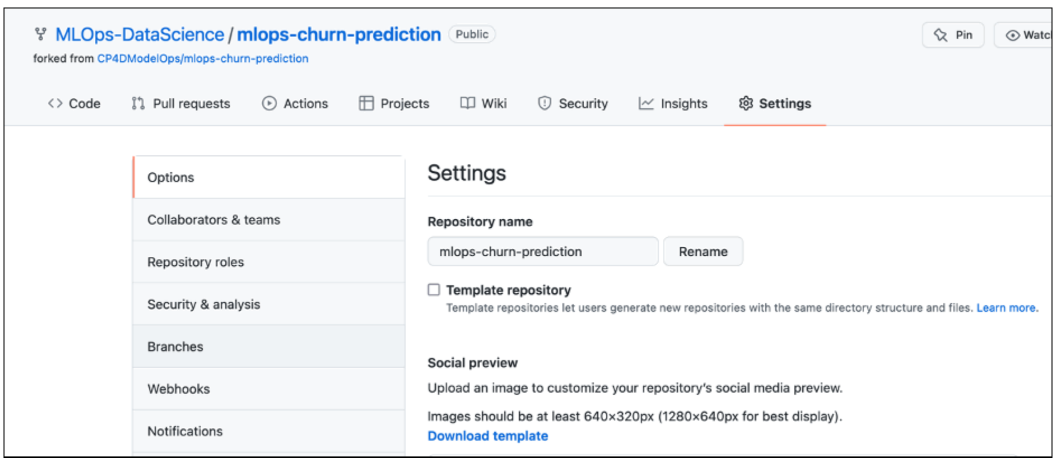

Figure 5-164   Branch settings

11. You see that no branch protection rules are set up yet. To protect the UAT and PRD branches, we create these protection rules (see Figure 5-165).

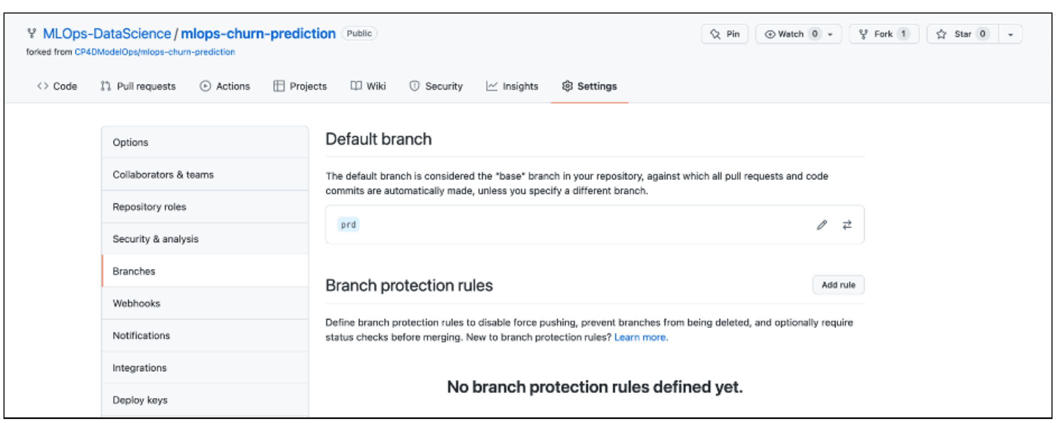

Figure 5-165   Branch protection rules

12. Click **Add rule** to create a rule for the UAT branch. Enter `uat` for the Branch name pattern and then, select **Require a pull request before merging** (see Figure 5-166.

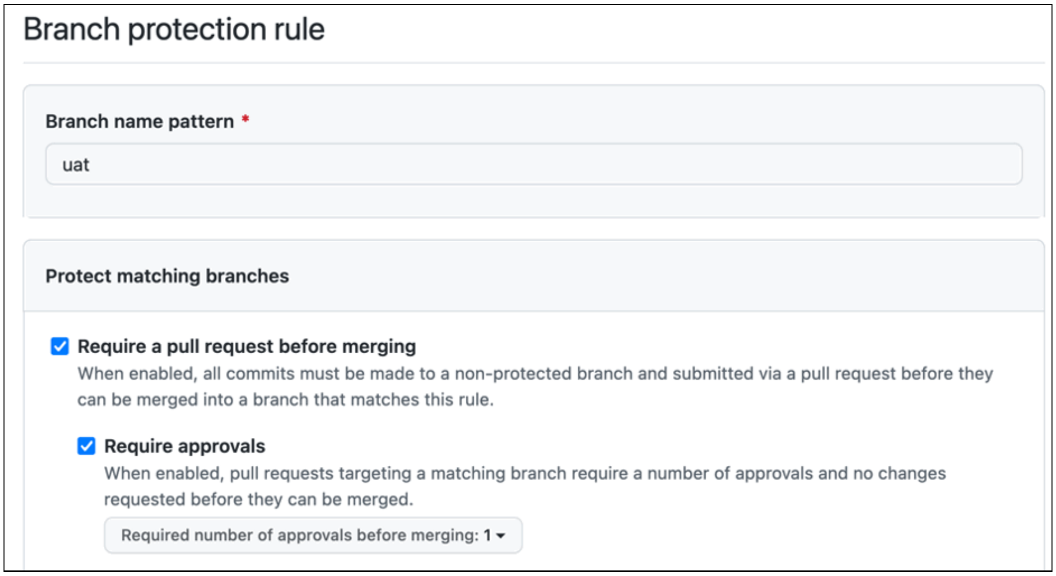

Figure 5-166 Requiring pull request

13. Scroll down and select **Restrict who can push to matching branches** (see Figure 5-167).

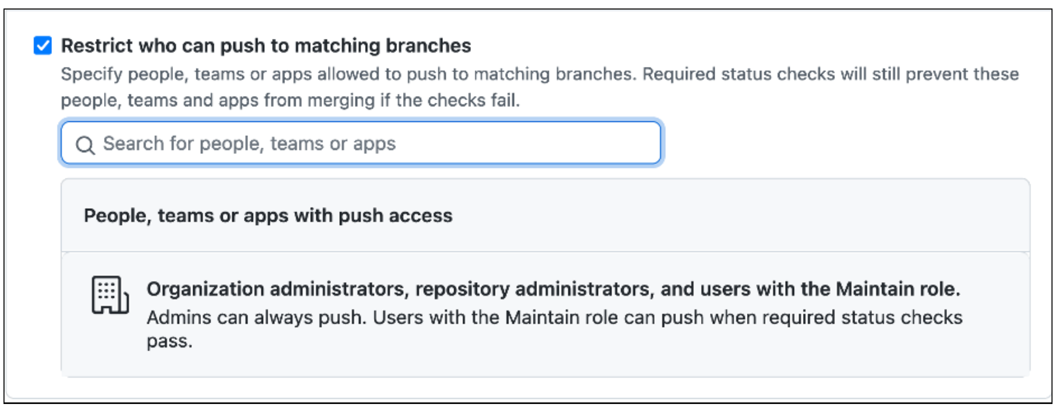

Figure 5-167 Restricting who can push

14. Scroll down and click **Create**.
15. Repeat these steps for the PRD branch. You *must* click **Branches link** in the sidebar again to create the next rule.

16. When completed, you see the branch protection rules, as shown in Figure 5-168.

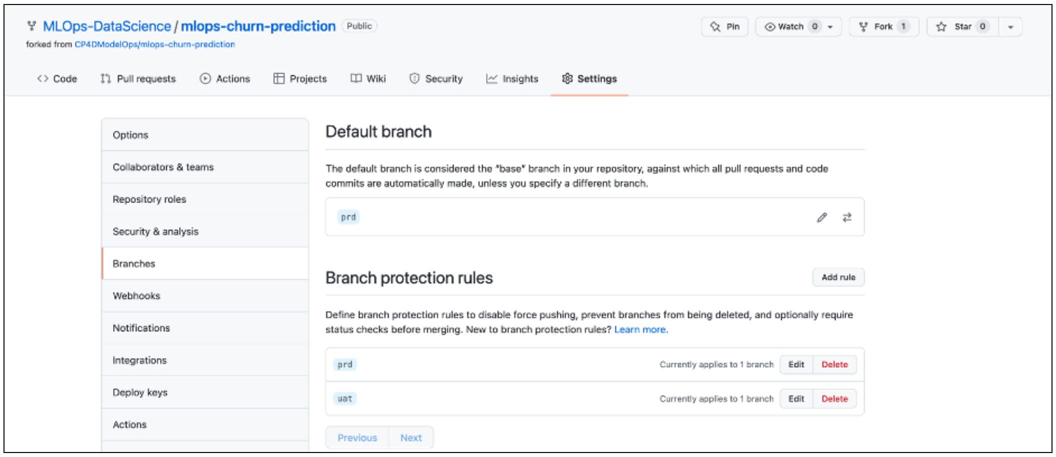

Figure 5-168 Branch protection rules

17. Because your organization has only a single user (yourself) and you are the administrator, we cannot truly set up a formal approval process in which data scientists have only write access to the repository and the MLOps staff can approve requests.

    In a real-world implementation, the production branch also can be in an upstream repository with different permissions. Setting up the GitHub topology with different repositories and teams is beyond the scope of this workshop.

18. Now that the repository is created, we can start integrating it with IBM Watson Studio in Cloud Pak for Data. IBM Watson Studio must be granted access to your repository by using a token.

    Click your user at the right-top of the window and select **Settings**, as shown in Figure 5-169.

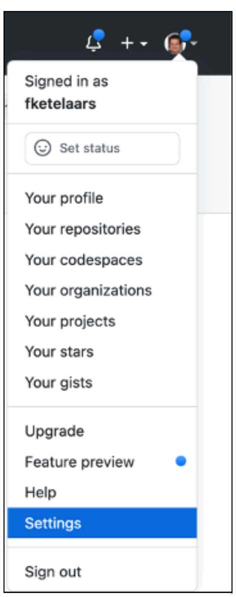

Figure 5-169 GitHub settings

19. Scroll down the left sidebar and click **Developer settings**.

20.Click **Personal access tokens** (see Figure 5-170).

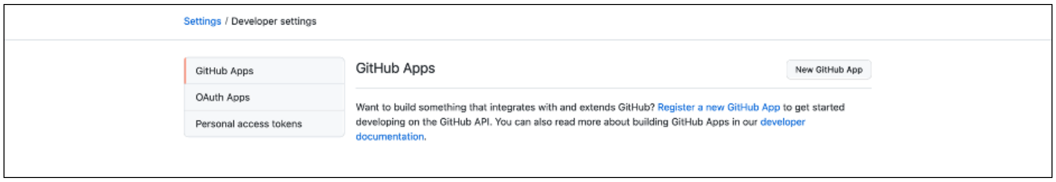

Figure 5-170 Developer settings

21.Click **Generate new token** (see Figure 5-171).

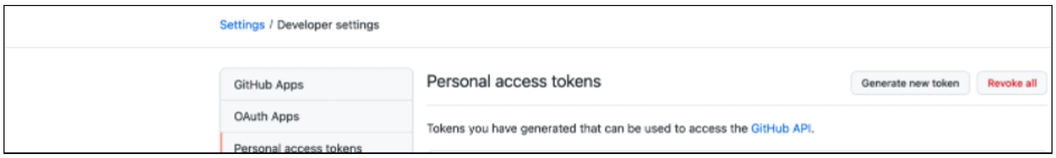

Figure 5-171 Personal access tokens

22.Enter the name of the token and ensure that the **repo** option is selected, as shown in Figure 5-172. Scroll down and create the token.

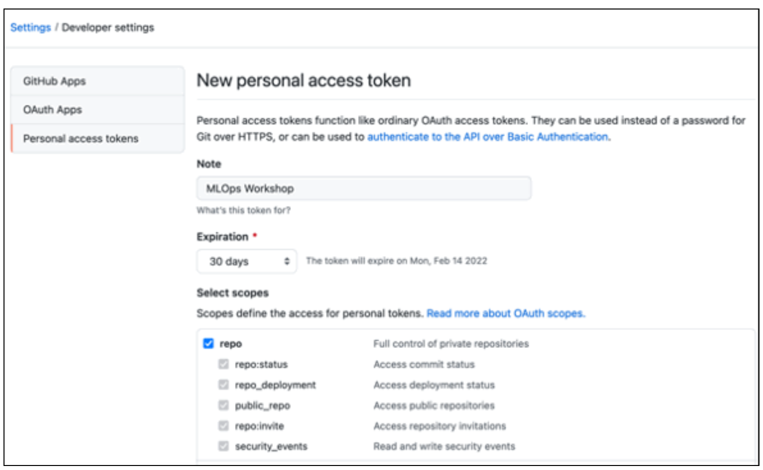

Figure 5-172 Generating the token

23.The token is displayed only once; therefore, make sure you copy it. You need it multiple times in the following steps (see Figure 5-173).

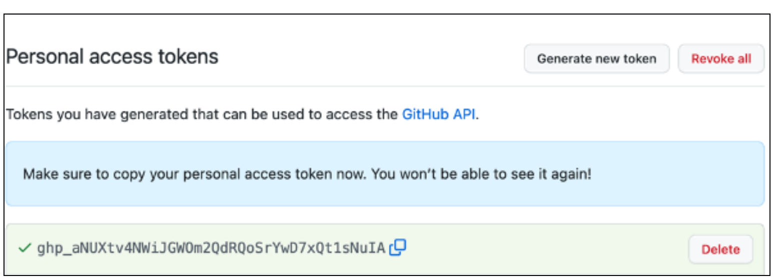

Figure 5-173 Show token

24.Log into Cloud Pak for Data as the `dslead` user.

25. Click the navigation menu (top-left navigation icon) and select **Projects** → **All projects**.

26. Click **New project** to create a project. Select the **Create a project integrated with a Git repository** option, as shown in Figure 5-174.

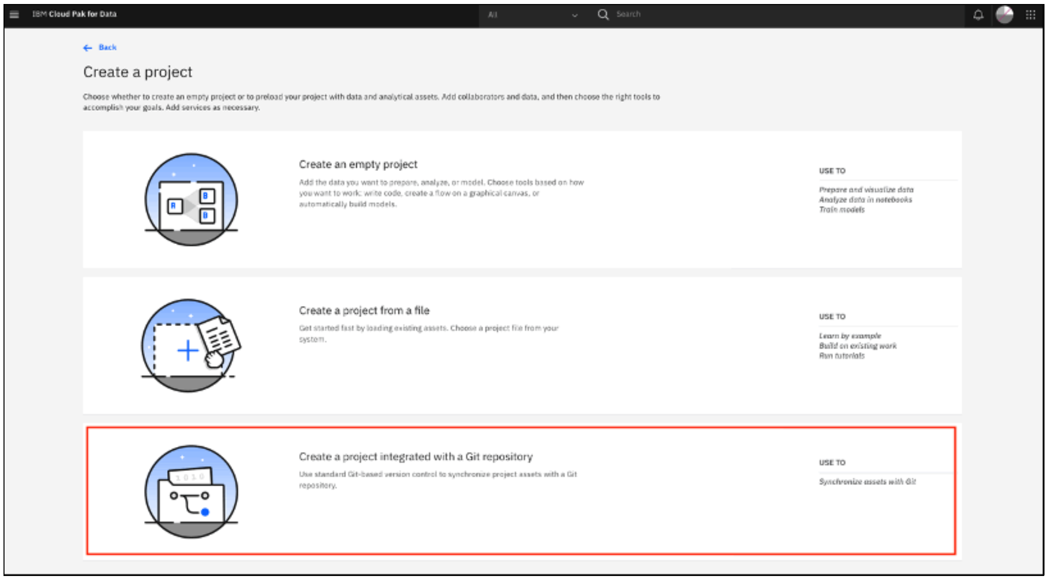

Figure 5-174 Creating a Git-integrated project

27. In the New project window, enter a Name (for example, `mlops-dev`) and Description (optional) for the project. For the integration type, select **Default Git integration**. You must upload the GitHub token that you generated to connect to your Git repository. Click **New Token** (see Figure 5-175).

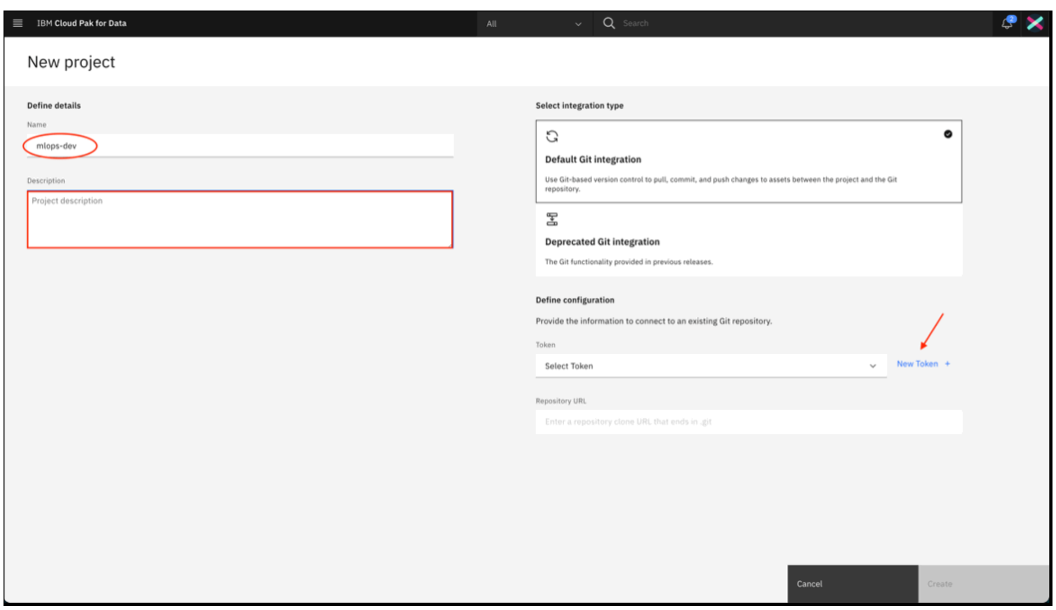

Figure 5-175 Project properties

28. In the Git integration pop-up window, provide the Git access token and enter a name for it for reference. Click **Create** (see Figure 5-176).

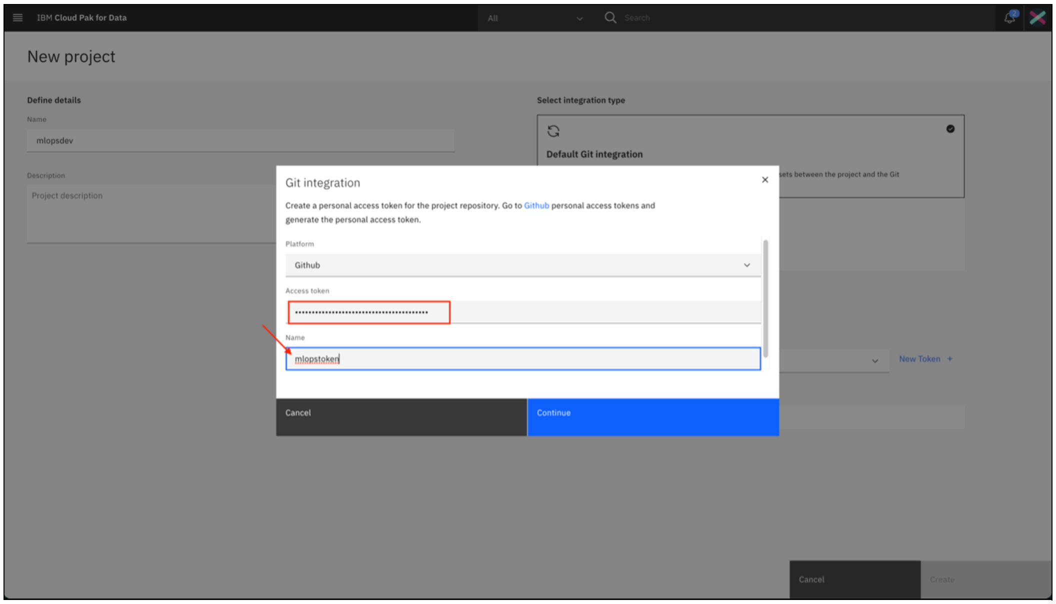

Figure 5-176 Creating a token

29. Click the **Token** drop-down menu and select the token that you created (in our example, **mlopstoken**), as shown in Figure 5-177.

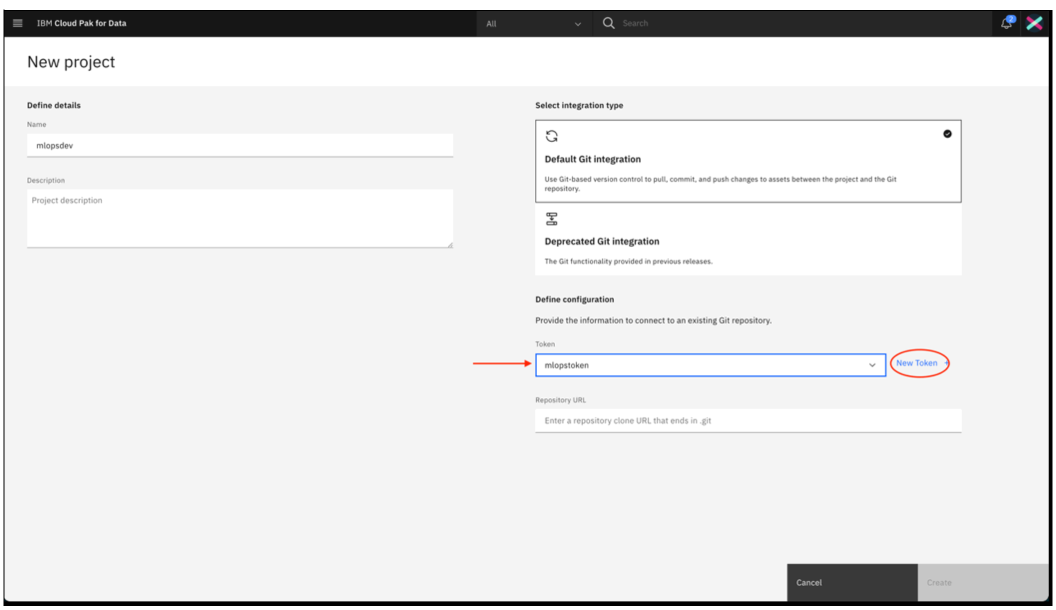

Figure 5-177 Selecting the token

30. Provide the GitHub repository and branch to associate with your project (see Figure 5-178).

> **Important:** The rest of the instructions in this module reference the repository that is found here. However, you reference your own repository that was forked from the upstream repository.
>
> For the Repository URL field, enter the HTTPS reference for your Git repository. You can copy this reference by clicking **Code** and then, clicking the **Copy** icon (highlighted by a red arrow in Figure 5-178).

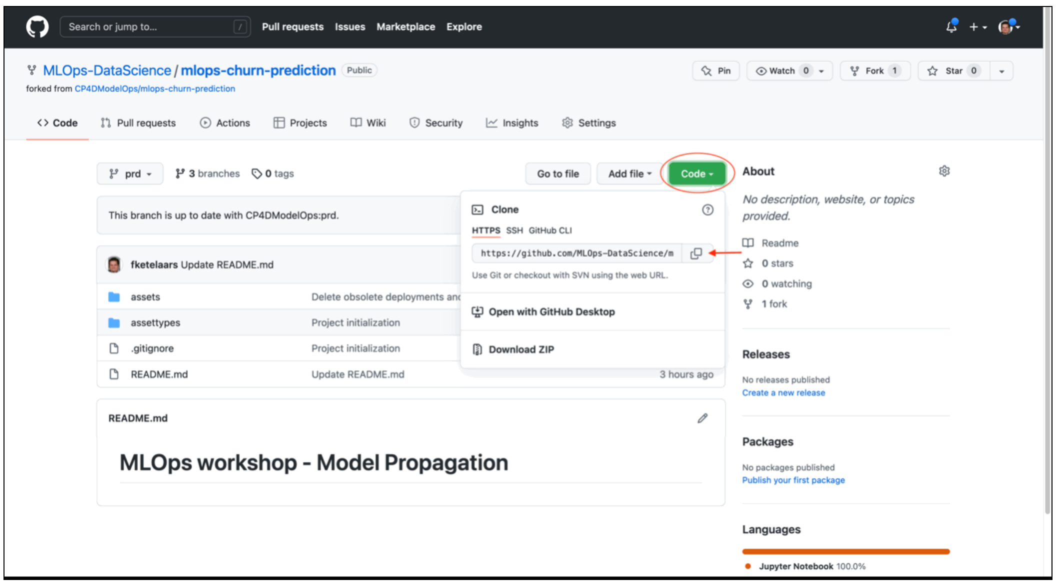

Figure 5-178 Clone URL

31. Return to your Create project window and paste the URL. Then, the Branch field becomes available.

Click the drop-down menu and select the **DEV** branch. When your new project looks like the example that is shown in Figure 5-179, click **Create** to create the project.

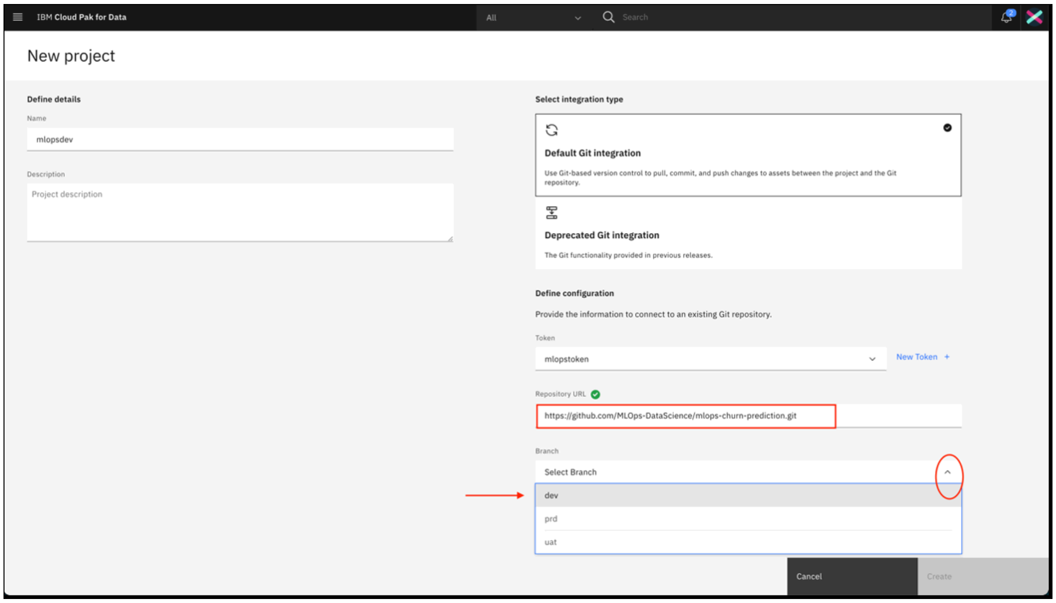

Figure 5-179 Selecting the branch

32. You should see a pop-up message that indicates that the project was created successfully. Click **View new project** to browse to the new project (see Figure 5-180).

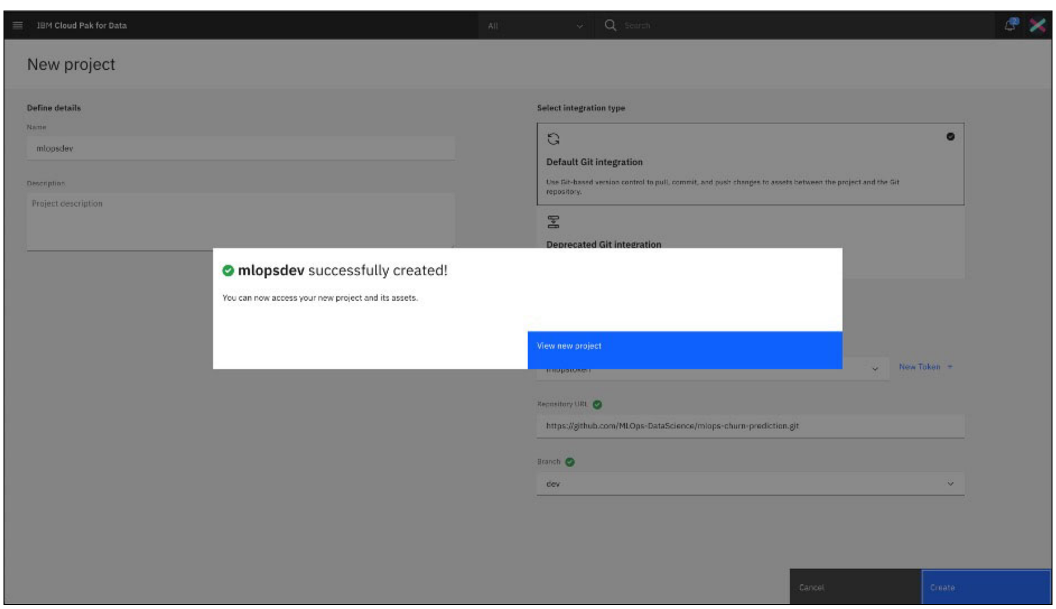

Figure 5-180 Project created

33. Take a few minutes to review the project information. The Overview tab provides general information about the project, such as Assets, Project history, and Storage used. Click the **Manage** tab and select **Access control** to review which users can access the project. In this example, the `dslead` user should be the only collaborator in this project (see Figure 5-181).

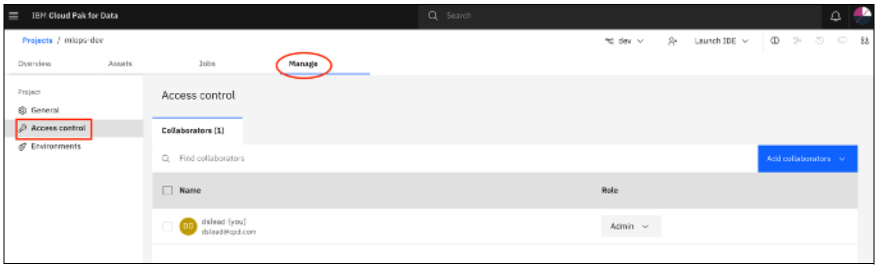

Figure 5-181 Managing access

34. You can add other users or user groups and assign them the relevant permissions to collaborate on this project. Typically, a data science project involves multiple users and roles who collaborate on developing and evaluating AI models.

    Add the `datascientist` user as an Editor to the project. Click **Add collaborators** (indicated by the arrow in Figure 5-182) and select **Add users**.

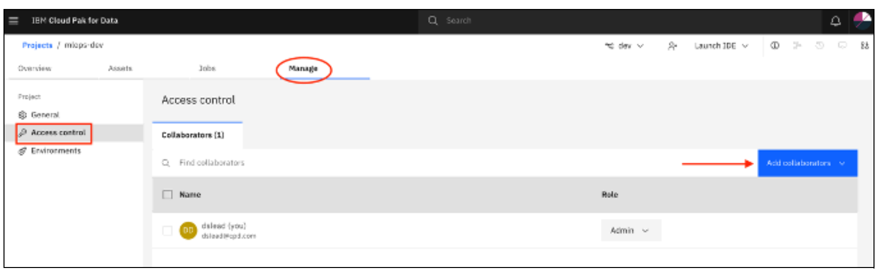

Figure 5-182 Adding users

35. Enter `datascientist` in the field that highlighted by the red rectangle in Figure 5-183 and then, select the suitable user. Specify the role as Editor and then, click **Add**.

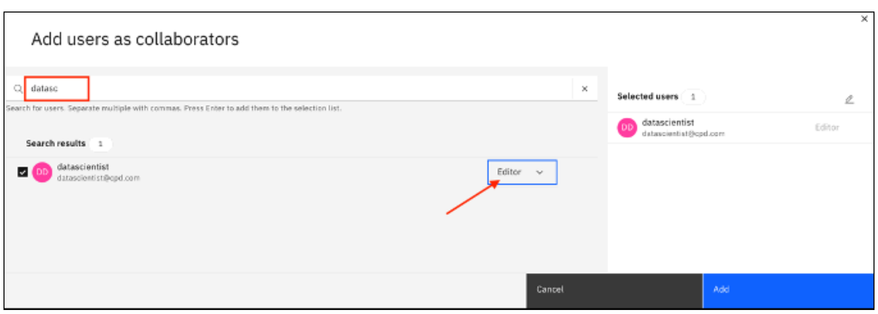

Figure 5-183 Adding user with role

36. Click the **Assets** tab to review the assets that are associated with the project. You find only the `CUSTOMER_DATA_ready.csv` file as an asset. Notebooks and other data assets can be found from the JupyterLab interface (see Figure 5-184).

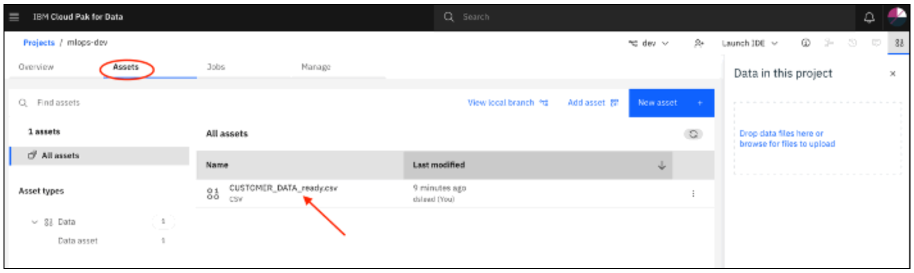

Figure 5-184 Data assets

37. Now that the project is created, the `datascientist` user starts the JupyterLab IDE to work on a new version of the churn prediction model.

Log out from Cloud Pak for Data and then, log in by using the `datascientist` user.

38. Click **All Projects** that is at the starting page.

The list of projects that the `datascientist` user can access is shown (see Figure 5-185).

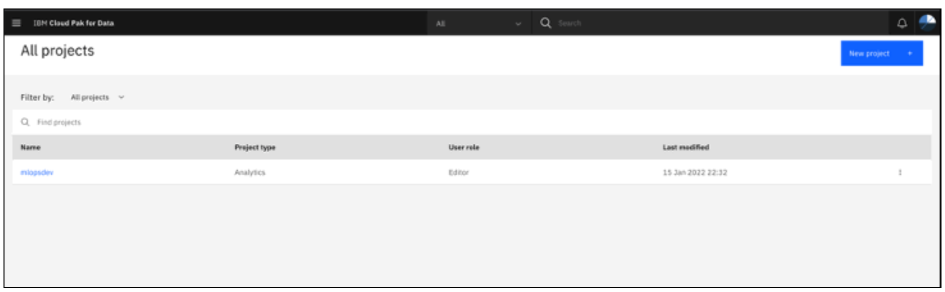

Figure 5-185 Projects for datascientist

39. Open the project by clicking its name (in our example, **mlops-dev**). Because the project is connected to a Git repository and the `datascientist` user did not open this project before, you are prompted to checkout a specific branch and specify the Git token (see Figure 5-186).

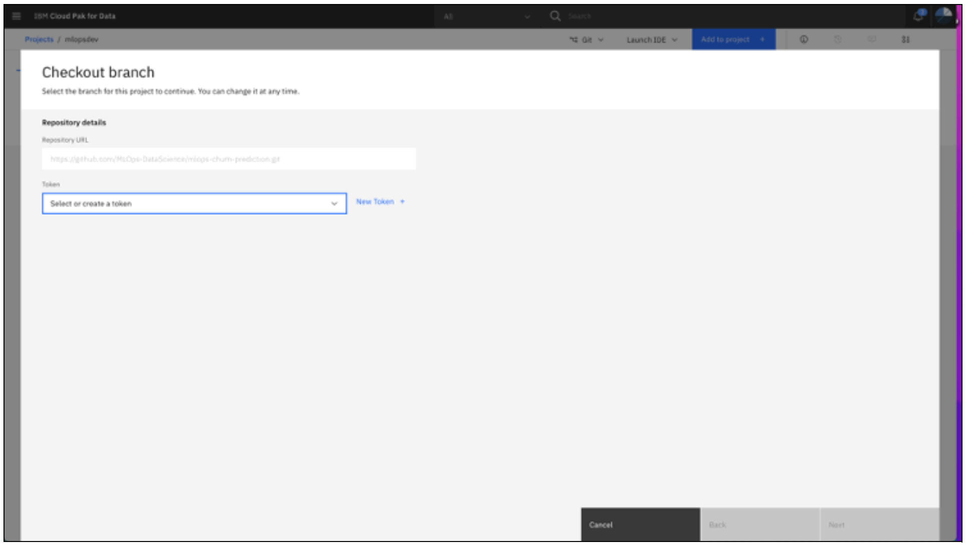

Figure 5-186 Checkout branch

40. Ignore a pop-up error (see Figure 5-187) that might be displayed that reads: `Something went wrong performing a catalogs action`.

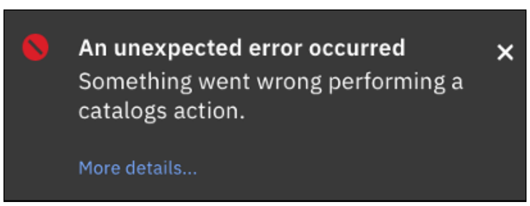

Figure 5-187 Error to be ignored

41. For convenience, we use the same Git token that was used by the `dslead` user. (Usually, every user has their own credentials to log in to the Git repository and tokens are not shared between project members.)

Click **New token link** and create a token with a new name. As shown in Figure 5-188, we use the name `mlops-datascientist`; however, you can use any name. Also, paste the token in the Access token field that you captured previously (as highlighted by the red box in Figure 5-188).

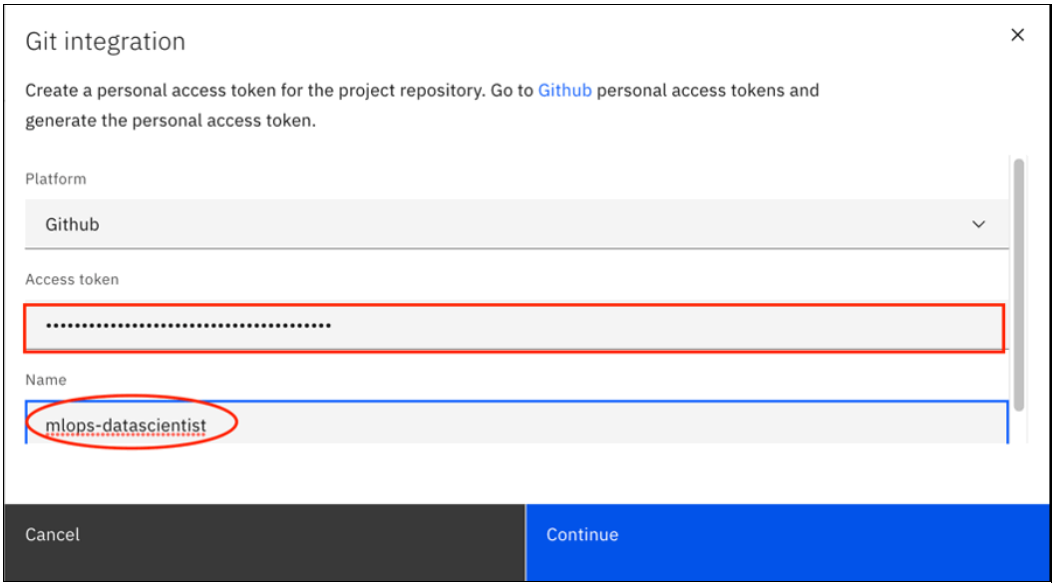

Figure 5-188 Creating a token

42. Return to the window to review the branch. Select the token that you just created and click **Next** (see Figure 5-189).

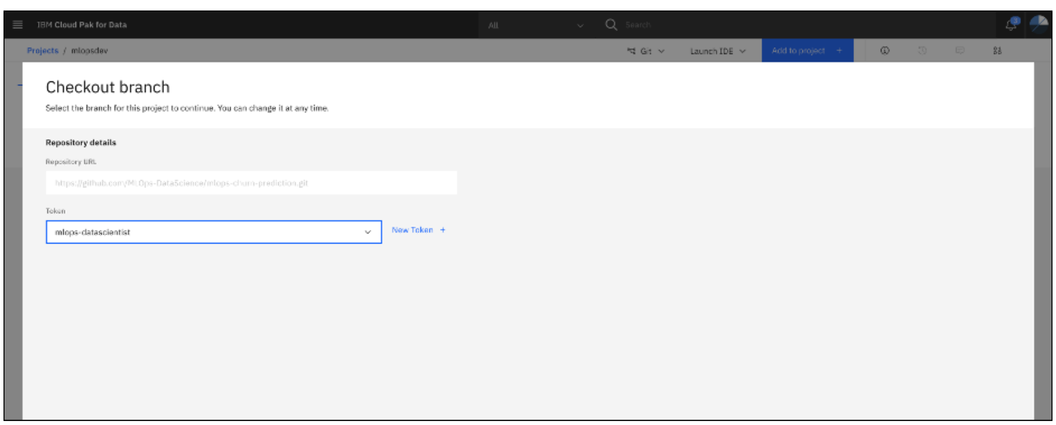

Figure 5-189 Selecting a token

43. OI the next window, select the **DEV** branch and click **Select**. A confirmation window opens (see Figure 5-190).

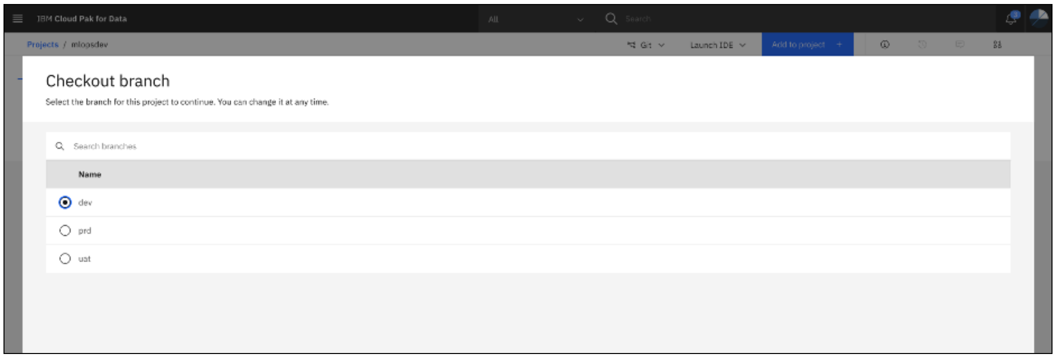

Figure 5-190 Checking out the DEV branch

44. Click **Launch IDE** at the top of the window and select **JupyterLab** (see Figure 5-191).

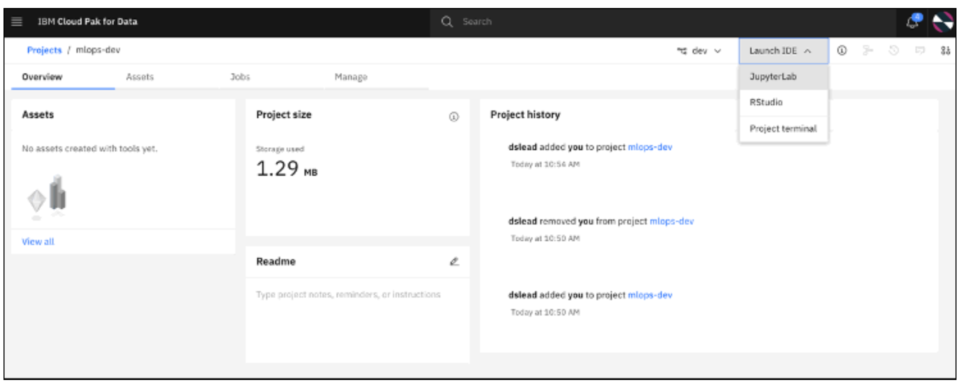

Figure 5-191 Launch JupyterLab

45. In the next window, select **JupyterLab with IBM Runtime 22.1 on Python 3.9** and then, click **Launch** (see Figure 5-192).

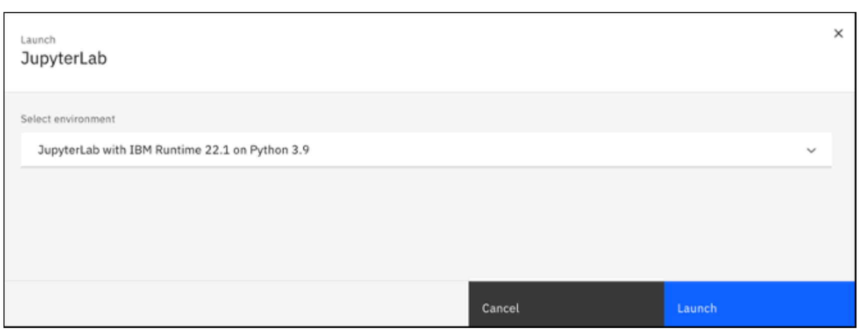

Figure 5-192 Selecting the environment

46. After JupyterLab starts, review the different areas of the window. On the far-left side (as highlighted by a red box in Figure 5-193 on page 389) you can browse the folder contents, view the active kernels and other tabs, and view the Git repository state, such as current branch and any changes that were made.

Also, on the left side, you see the “root” folder of the repository with the **assets** folder. This folder is where the input files and notebooks are stored. The middle part (as highlighted by an amber box in Figure 5-193) is used to create notebooks, open a Python console or shell terminal, and access other applications.

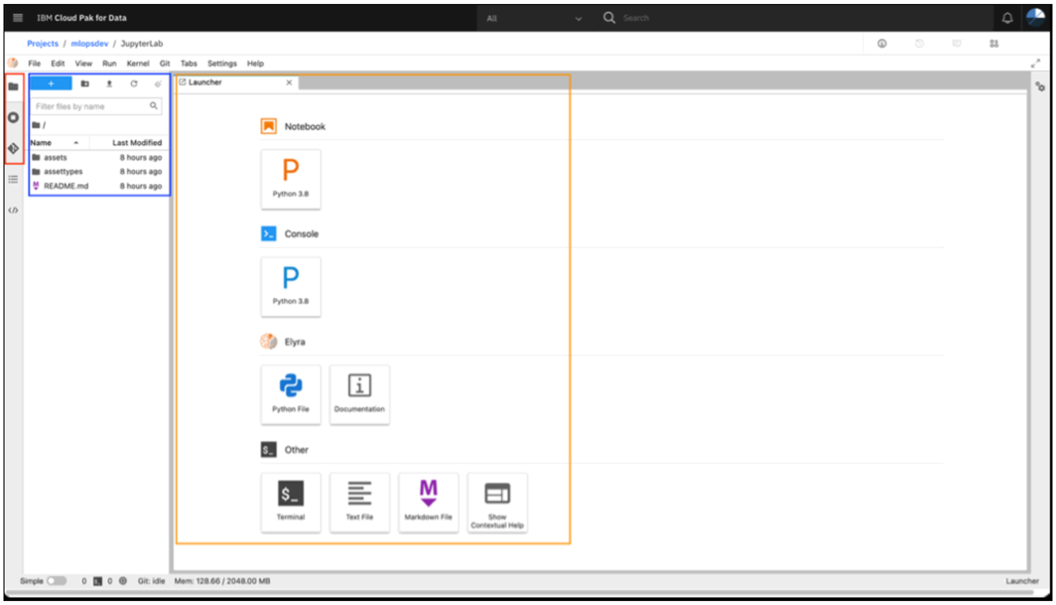

Figure 5-193 JupyterLab

47. In this exercise, we change an existing notebook. First, we make sure that we make the changes in a “feature” branch so that we can validate the changes without changing the notebook that might be used by other team members.

    Click the **Git** icon on the far left of the window, as indicated by the red arrow in Figure 5-194.

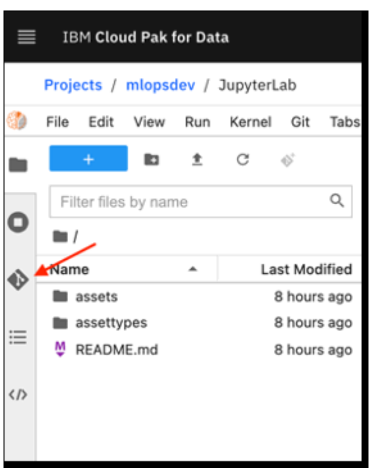

Figure 5-194 Open Git view

48. You see multiple branches in the current Git project (`dev`, `origin/HEAD`, `origin/dev`, `origin/uat` and `origin/prd`) and your current branch is DEV. Create a branch by clicking **New Branch**. A window opens in which you enter the name of the new branch. In our example, we enters the name `optimize-churn-model` name.

After you enter a name, click **Create Branch** (see Figure 5-195).

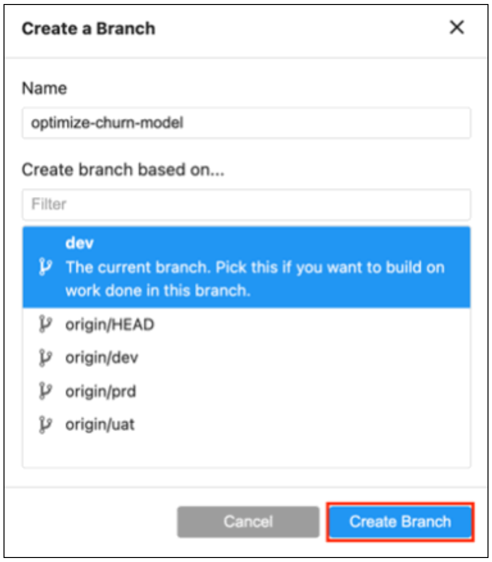

Figure 5-195 Creating a Branch window

49. Browse to the **assets/notebooks** folder by clicking the folder icon at the top left and double-click the `customer-churn-prediction.ipynb` notebook to open it (see Figure 5-196).

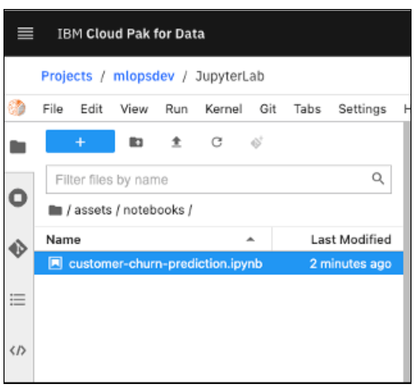

Figure 5-196 Opening a notebook

50. Scroll down to the section that is above Evaluate and notice the accuracy of the model, which is approximately 0.92, as shown in Figure 5-197.

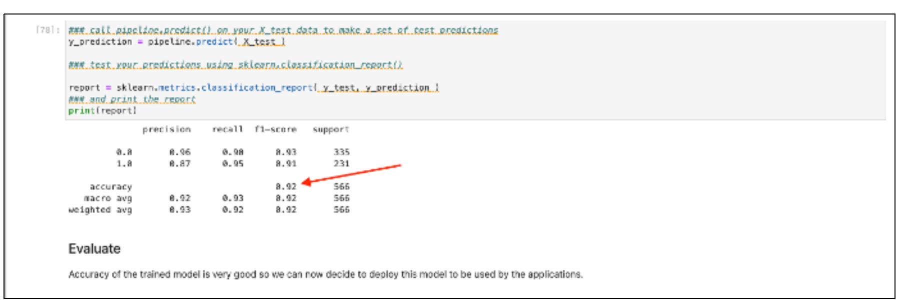

Figure 5-197 Model accuracy

51. Scroll to where the data is split and change the `test_size` from 0.4 to 0.2 (see Figure 5-198).

```
[76]: # split the data to training and testing set
      X_train, X_test, y_train, y_test = train_test_split(x, y,
          test_size=0.4,
          random_state=42, stratify=y)

[77]: # fit the model

      import sklearn.pipeline
      from sklearn.preprocessing import OneHotEncoder
```

Figure 5-198 Change test size

52. Save the notebook by clicking the disk at the top left of the notebook (as indicated by the red arrow in Figure 5-199).

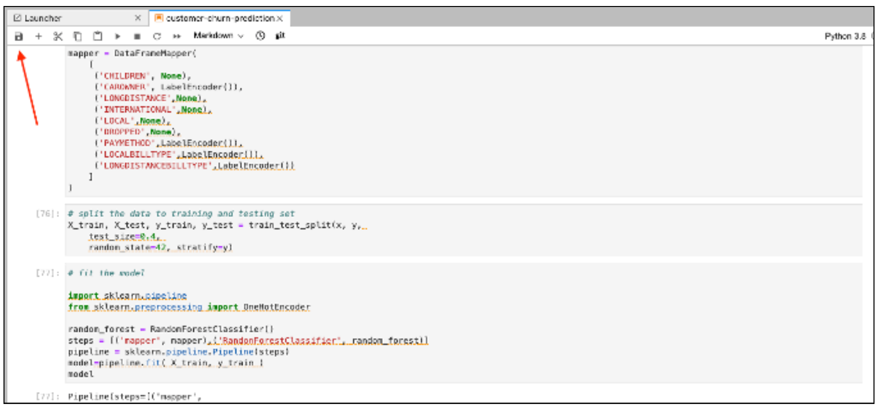

Figure 5-199 Save notebook

53. Run the notebook in its entirety by selecting **Run** → **Run all cells** from the top menu. All of the cells in the notebook are run. After they run, a sequence number appears between the square brackets on the left of each of the cells.

54. Scroll down to determine the accuracy of the model, which should be approximately 0.94 (although your results might differ slightly), as shown in Figure 5-200.

```
### call pipeline.predict() on your X_test data to make a set of test predictions
y_prediction = pipeline.predict( X_test )

### test your predictions using sklearn.classification_report()

report = sklearn.metrics.classification_report( y_test, y_prediction )
### and print the report
print(report)
              precision    recall  f1-score   support

         0.0       0.98      0.92      0.95       168
         1.0       0.89      0.97      0.93       115

    accuracy                           0.94       283
   macro avg       0.93      0.95      0.94       283
weighted avg       0.94      0.94      0.94       283
```

Figure 5-200 Model accuracy

55. Assume that these results are acceptable and want to promote the model to the DEV environment. Save the notebook again and go to the Git panel.

As shown in Figure 5-201, your changed notebook appears in the list of Changed files.

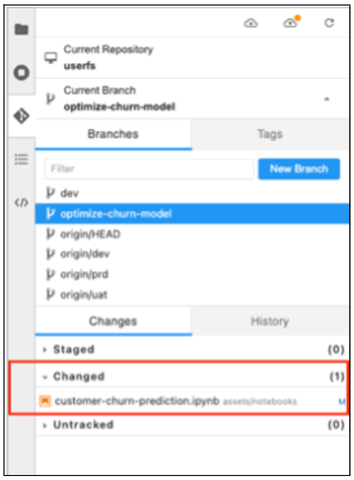

Figure 5-201 List changes

56. Click the notebook and then, click the **+** sign to stage the change for commit.The notebook now appear in the list of Staged files (see Figure 5-202).

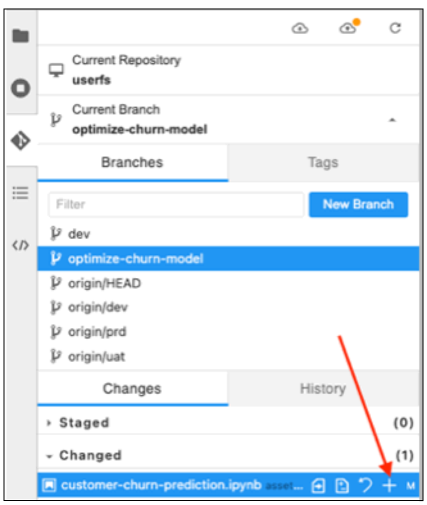

Figure 5-202 Stage changes

57. Enter a meaningful message in the Summary field and click **Commit** (see Figure 5-203).

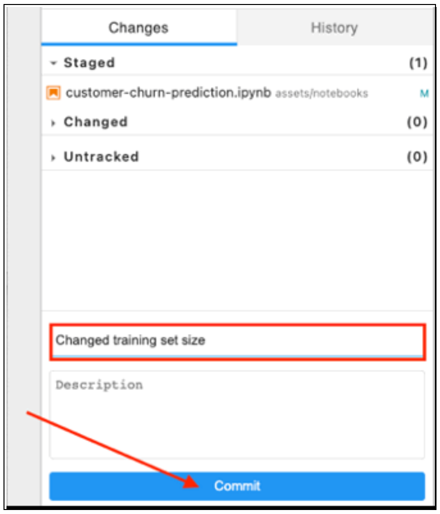

Figure 5-203 Commit

58. The change is still only local to JupyterLab. To make it available in the Git repository, we must synchronize. Click the **cloud button** to push the changes to the remote repository (see Figure 5-204).

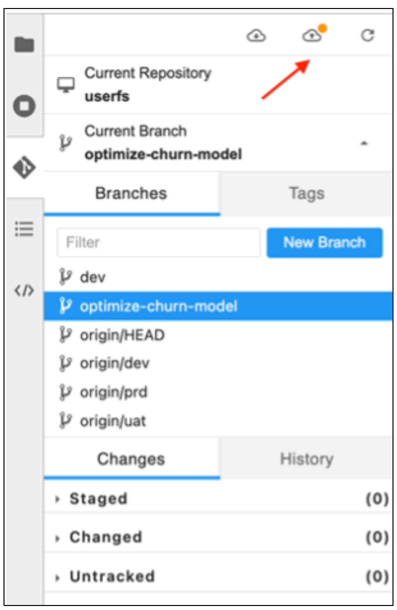

Figure 5-204 Push

59. Go to your repository on GitHub, refresh the page, and see that recent pushes were made to the branch that you created (see Figure 5-205).

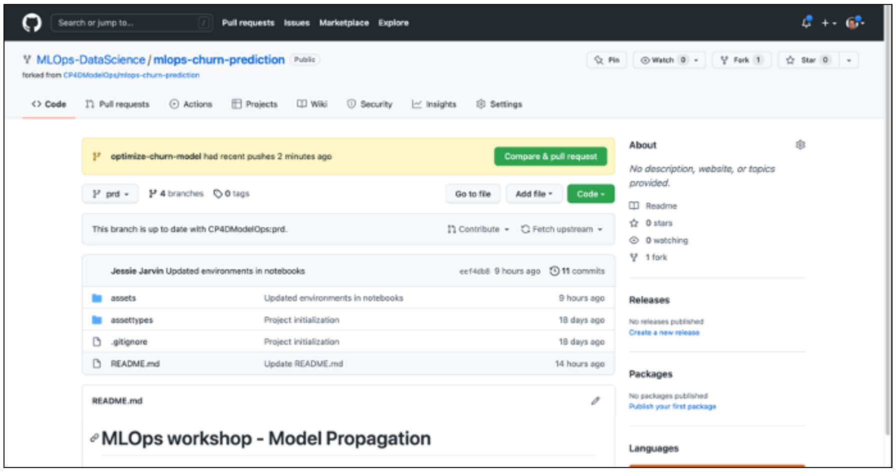

Figure 5-205 Recent pushes

60. Because you want the data science lead to review your changes, create a pull request. Click **Compare & pull request** to request your changes to be merged with the DEV branch. Ensure that you are creating a pull request for your own repository.

You find that GitHub assumes that you want to create a pull request to the upstream repository (`CP4DModelOps/mlops-churn-prediction`), as shown in Figure 5-206.

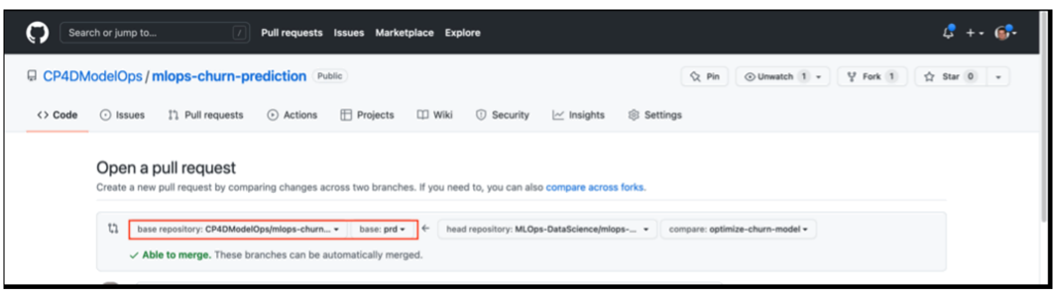

Figure 5-206 Compare and pull request

61. Because you want to merge the changes into the DEV branch of your own repository, click the **base repository** drop-down menu and select your own repository (see Figure 5-207).

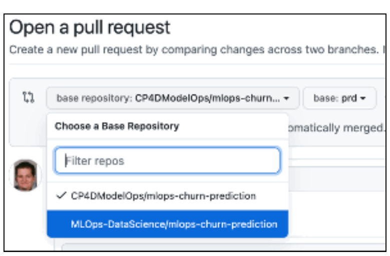

Figure 5-207 Selecting the repository

62. The upstream repository disappears and now, you can select the DEV branch by clicking the **PRD** drop-down box (see Figure 5-208).

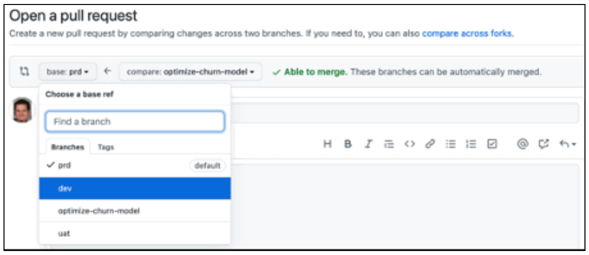

Figure 5-208 Selecting DEV branch

63. GitHub indicates that it can merge the changes automatically (Able to merge). Click **Create pull request** to request that the changes be merged into the DEV branch (see Figure 5-209).

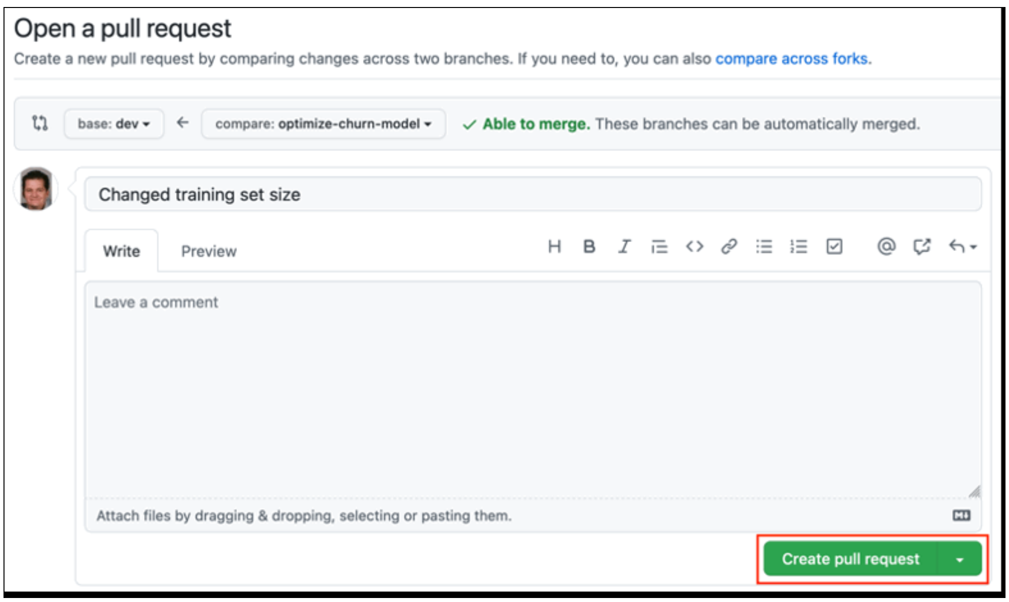

Figure 5-209 Creating a pull request

64. Assume that the Data Scientist lead received an email that a pull request is ready to be reviewed. Take the role of the lead data scientist now.

65. Click **Pull requests** at the top of the window (see Figure 5-210).

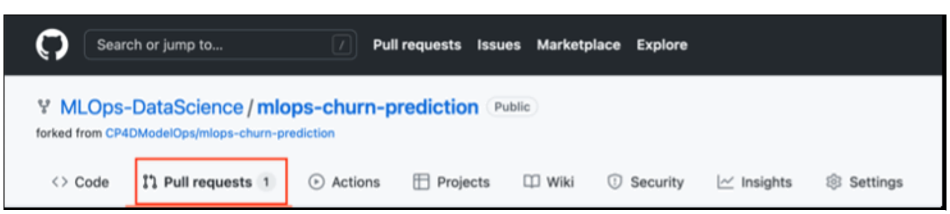

Figure 5-210 Pull requests

A list of all (in our example, one) pull requests is shown (see Figure 5-211).

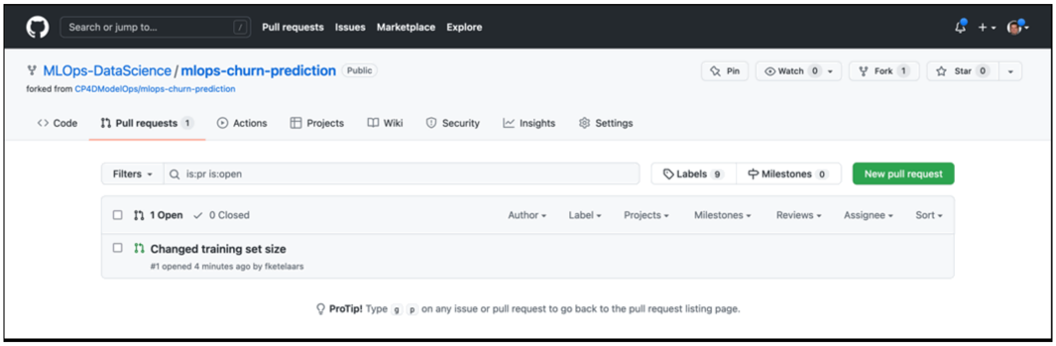

Figure 5-211 List of pull requests

66. Select the pull request that you created as the data scientist user by clicking the title. You see that the request consists of one commit and only one file was changed (as highlighted by the red boxes in Figure 5-212).

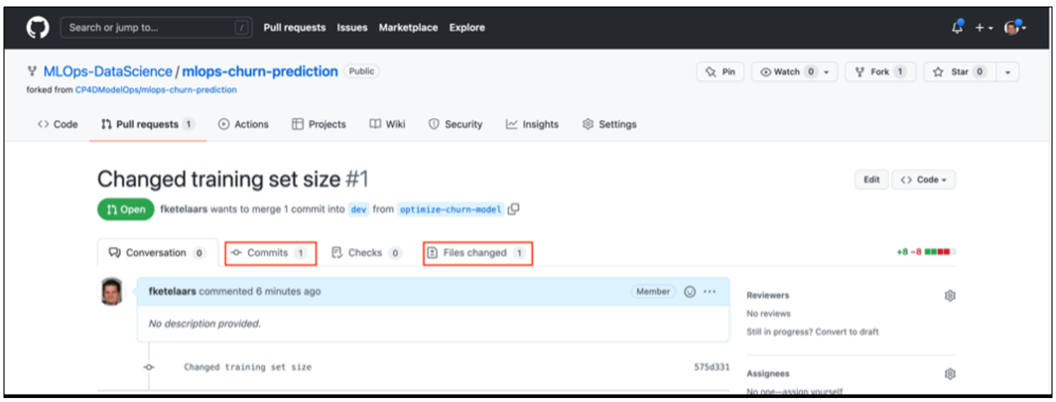

Figure 5-212 Listing pull request changes

67. Click **Files changed** to see the details of what was changed. The file is not as formatted as a notebook, but it is easy to identify the changes that were made. If wanted, you can click the ellipses (as highlighted by the red oval in Figure 5-213) and select **View file** to display the file in a notebook format.

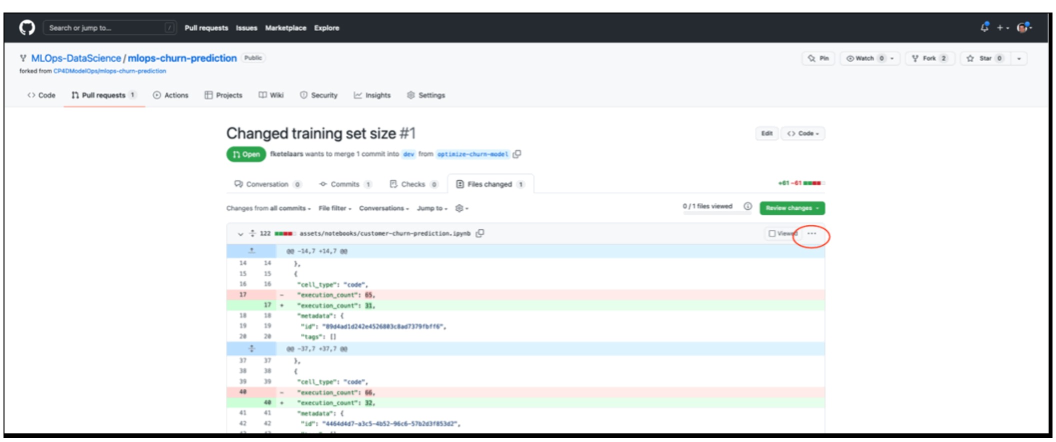

Figure 5-213 Change details

68. The changes are acceptable to the data scientist lead and approves and merges the changes to the DEV branch. (Because we have only one user for this repository, we cannot approve our own pull request.) Normally, the lead data scientist uses clicks **Review changes** to approve the request. Instead, we click the browser's **Back** button to return to the pull request.

69. Click **Merge pull request** to merge the changes with the DEV branch (see Figure 5-214).

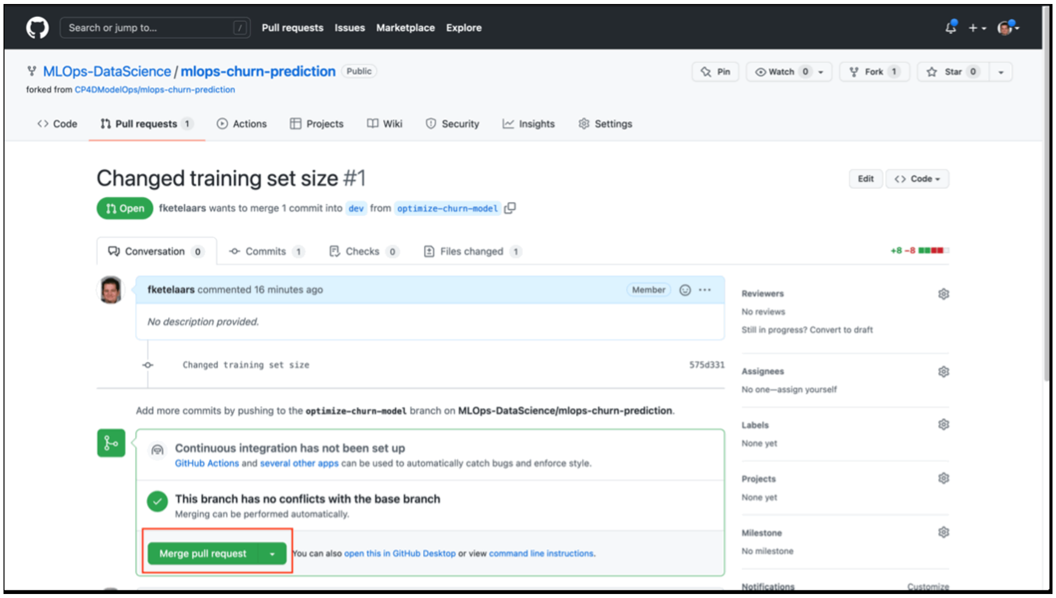

Figure 5-214 Merge pull request

70. Click **Confirm merge** (see Figure 5-215).

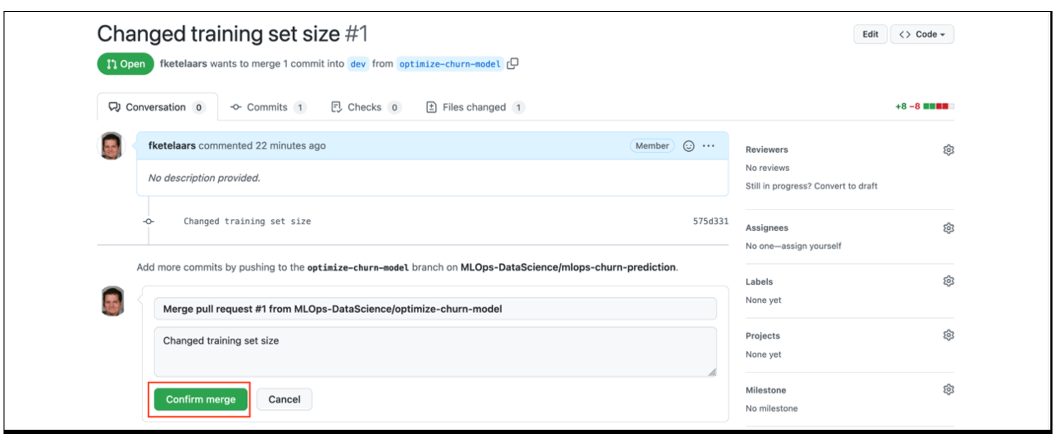

Figure 5-215 Confirm merge

71. GitHub confirms that the pull request was successfully merged and closed. As a good practice, keep the repository clean and delete the branch by clicking **Delete branch** (see Figure 5-216).

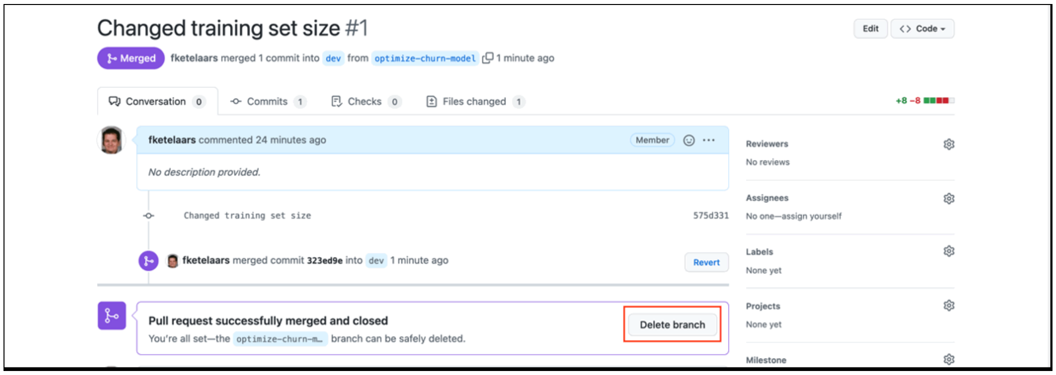

Figure 5-216 Delete branch

72. Go to your Cloud Pak for Data window. Log off as the `datascientist` user and log in as the Data Scientist Lead (`dslead`). Then, navigate to the `mlopsdev` project. Open it and start JupyterLab as you did previously (see Figure 5-217).

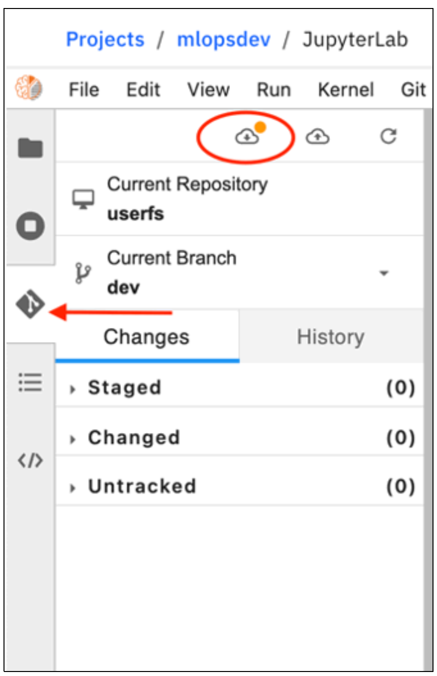

Figure 5-217 Fetch changes from remote

73. Go to the Git space in JupyterLab as indicated by the arrow in Figure 5-217. You notice amber dot next to the cloud icon with the down arrow (marked by the red oval in Figure 5-217). This dot indicates that changes exist in the remote repository that can be pulled.

74. Click the cloud icon with the arrow down inside. The changes are pulled, and the amber dot disappears. The latest commit that was done by the `datascientist` user are now downloaded.

75. Browse to the **Assets** → **notebooks** → **customer-churn-prediction.ipynb** and review the change that was made by the `datascientist` user (verify that you see the change in test-size). Run the entire notebook as the current `dslead` user by clicking **Run** → **Run All Cells**. The Data Science lead notices that the accuracy of the model improved and is that the change is acceptable.

76. Scroll down to the bottom of the notebook and observe that the notebook was successfully deployed to the `churn-dev` deployment space. You can double check this observation by clicking the navigation menu at the top left and then, open the **Deployments** page in a new tab by right-clicking **Deployments** (see Figure 5-218).

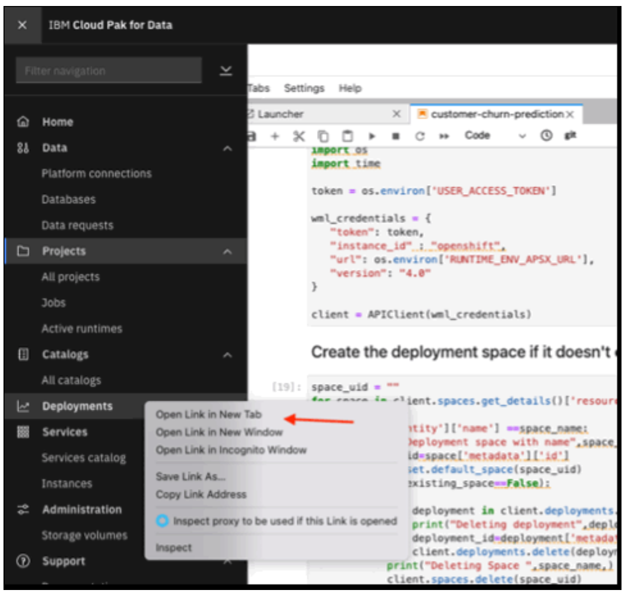

Figure 5-218 Navigating to deployment spaces

The `churn-dev` deployment space is now listed (see Figure 5-219).

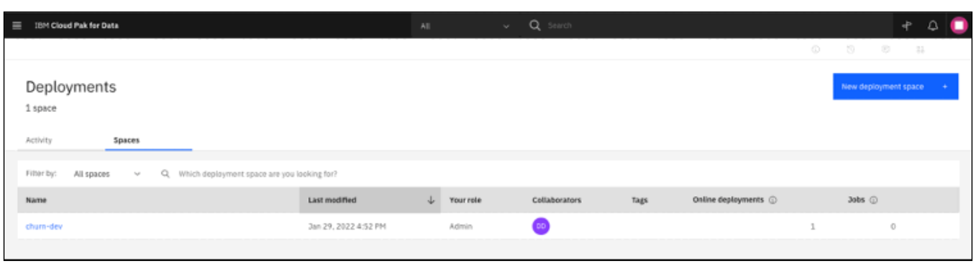

Figure 5-219 List deployments

77. Click the `churn-dev` deployment space and navigate to the **Deployments** tab. You can see that the `churn_pipeline` model was successfully deployed as `churn_pipeline_deployment` (see Figure 5-220).

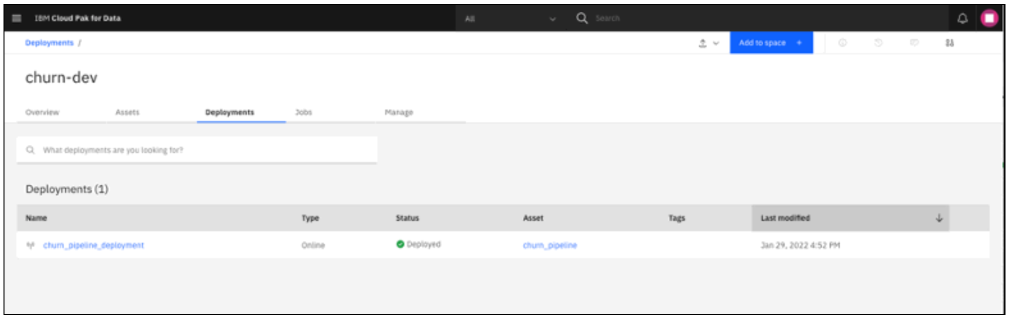

Figure 5-220 Model deployed

78. Because the data scientist only conducted the analysis and is not responsible for deployment, the lead data scientist creates a job that manages the automated deployment process in the subsequent environments, UAT and PRD. Return to the JupyterLab tab and click the **mlopsdev** breadcrumb (see Figure 5-221).

Figure 5-221 Go to mlopsdev project

79. Click the **Assets** link and then, click **View local branch** (see Figure 5-222).

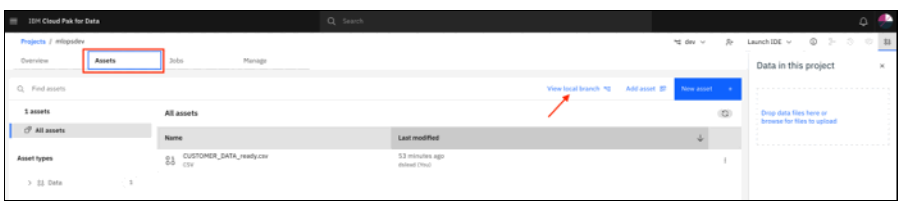

Figure 5-222 Viewing local branch

80. Create a job by clicking **New code job** in the top right of the window (see Figure 5-223).

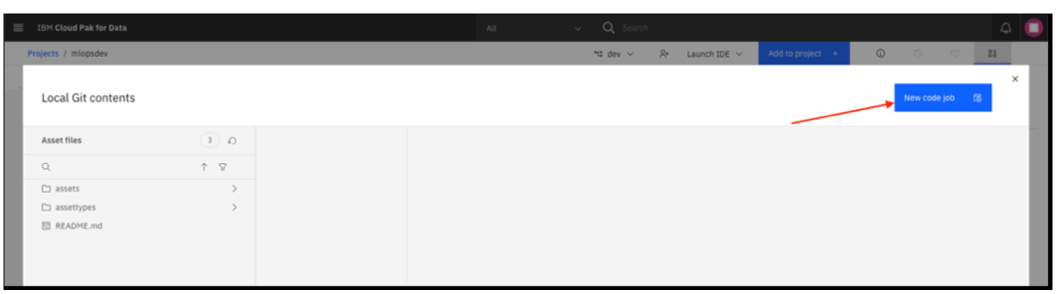

Figure 5-223 New job

81. Navigate to the notebook for which you want to create a job and click **Configure job** (see Figure 5-224).

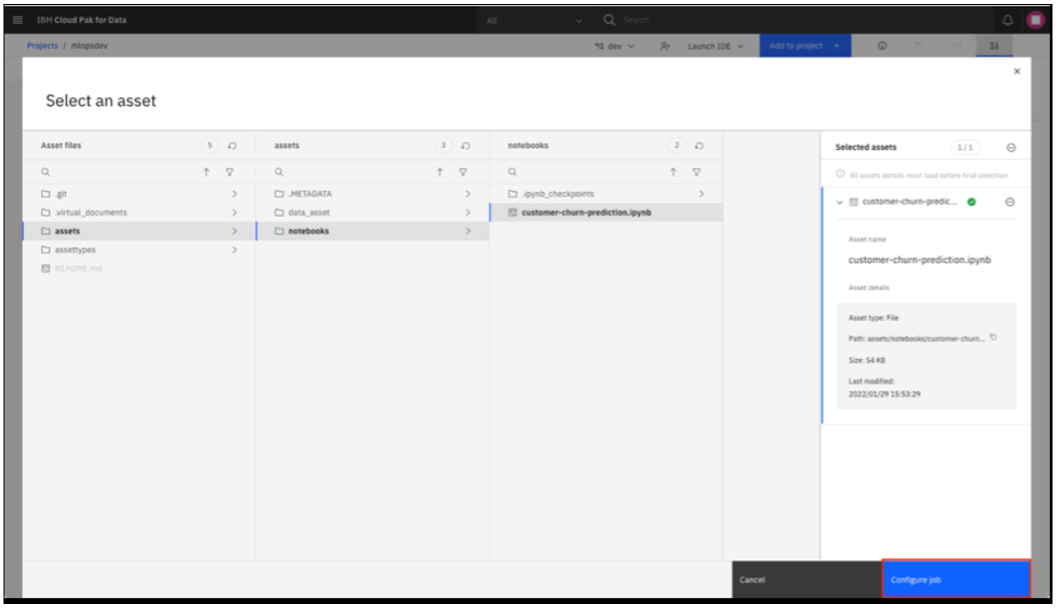

Figure 5-224 Selecting the notebook

82. Enter a meaningful name for your job (for example, `Deploy-churn-prediction-model`) and optionally enter a description of the job. Then, click **Next** (see Figure 5-225).

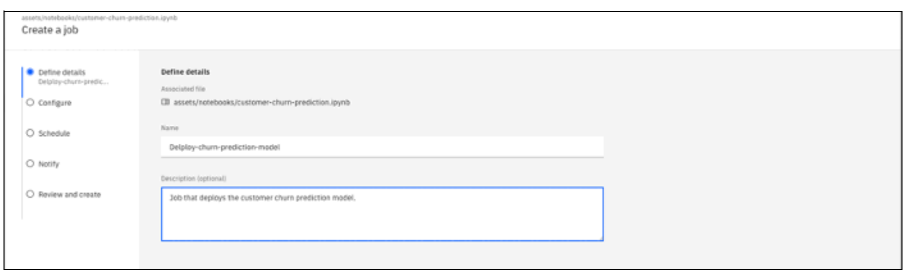

Figure 5-225 Job details

83. Choose the **IBM Runtime 22.1 on Python 3.9** environment (see Figure 5-226). Optionally, you can open the Advanced configuration to pass more environment variables to the job.

    For this exercise, we do not pass any other environment variables. Click **Next** to continue.

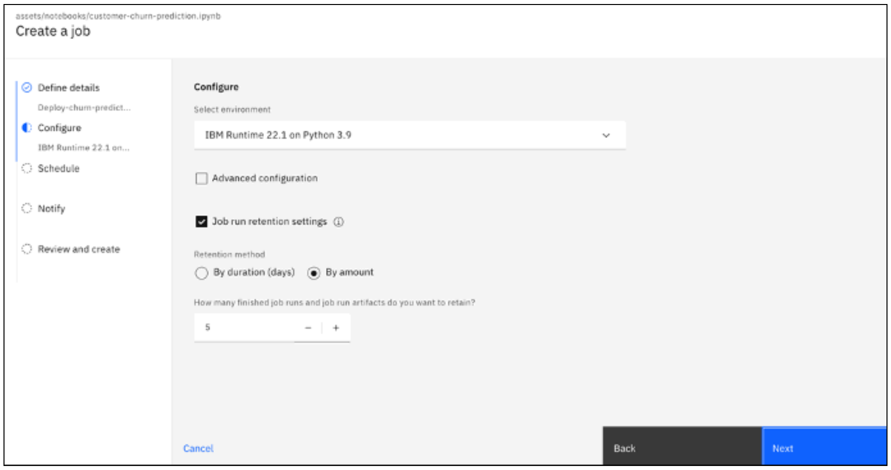

Figure 5-226 Selecting the environment

84. Because this deployment is online, we do not want to schedule a job. Instead, we want to promote it through the UAT and PRD stages and deploy it in the production environment. Click **Next** without selecting a schedule. Also, skip notifications for this job. Click **Next** again.

85. Review the parameters and create the job (see Figure 5-227).

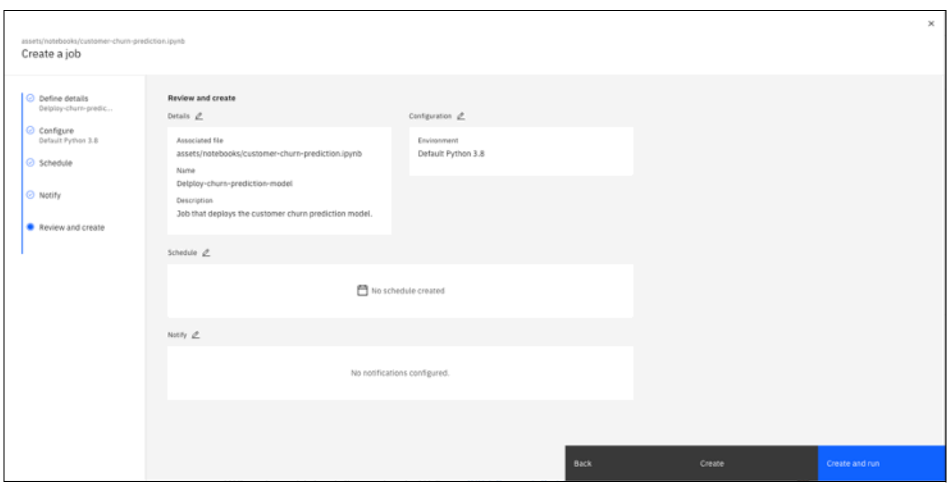

Figure 5-227 Reviewing and creating the job

86. Although the job does not need to be run now, you can choose **Create** or **Create and run**. You proceed to the Job Details page, where you can view job runs and other information about the job you just created (see Figure 5-228).

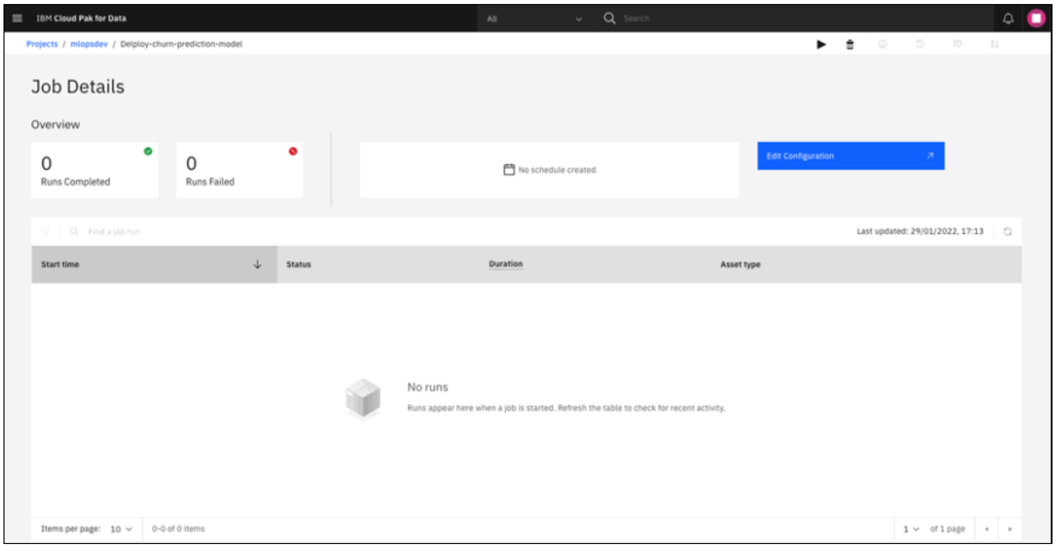

Figure 5-228 Job details

87. The lead data scientist now wants to promote the notebook and job to the next stage. First, the new job definition (which is effectively a JSON file) must be committed to the DEV branch. Click the breadcrumb of the **mlopsdev** project (see Figure 5-229).

Figure 5-229 Go to mlopsdev project

88. Commit the changes to the DEV branch. Click the **Git** icon at the top of the window and open the drop-down menu and then, click **Commit** (see Figure 5-230).

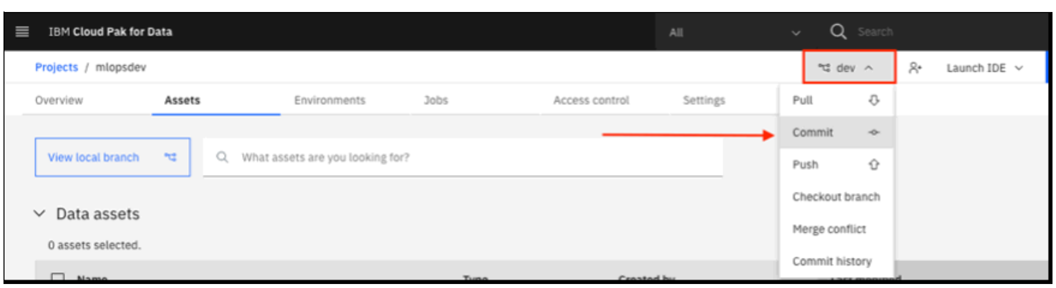

Figure 5-230 Committing to DEV

89. A new window opens (see Figure 5-231 on page 404) in which you can select the files to be included in the commit and a message that indicates any changes. Select all of the files by selecting the check box that is at the top of the list (as indicated by the arrow in Figure 5-231 on page 404) and enter a meaningful message. Then, click **Commit**.

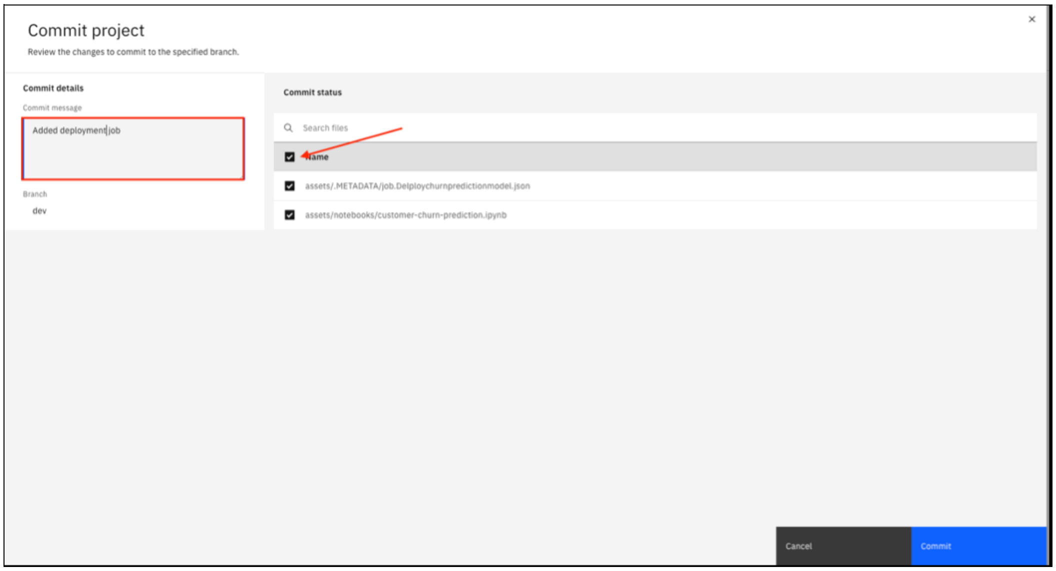

Figure 5-231 Commit

90.The changes are not yet synchronized with the Git repository. To make this synchronization, click the **Git** icon again and select **Push** (see Figure 5-232).

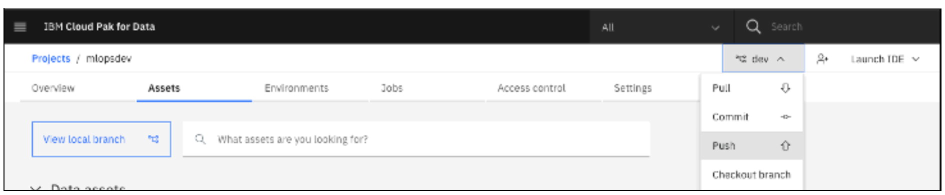

Figure 5-232 Selecting the Push option

91.All of the commits you performed are listed. In our example, you performed only one commit action, so the list is short. Confirm that your commit is listed and click **Push** (see Figure 5-233).

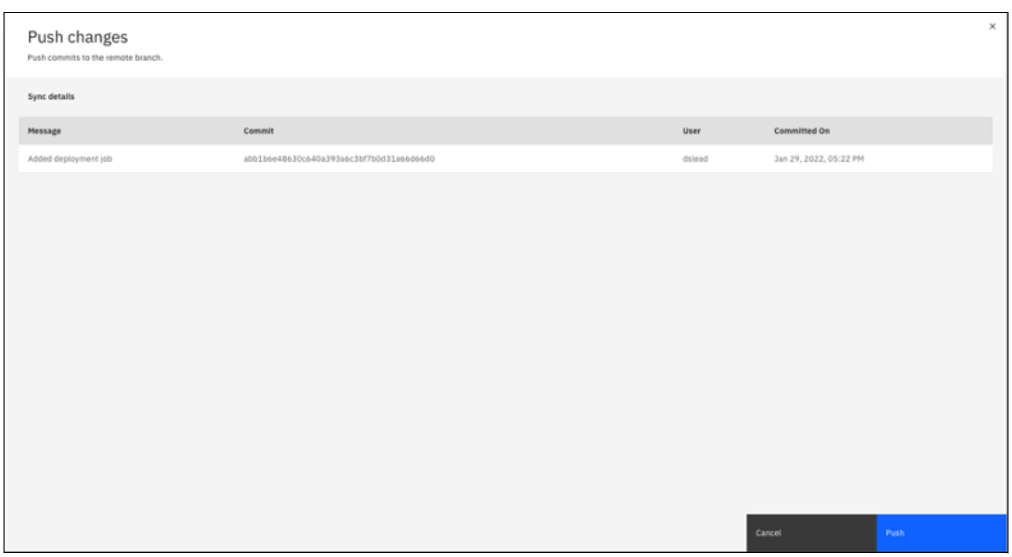

Figure 5-233 Confirming the push

92.Go to the GitHub repository in your browser. You notice a message that indicates that the DEV branch includes recent pushes (see Figure 5-234).

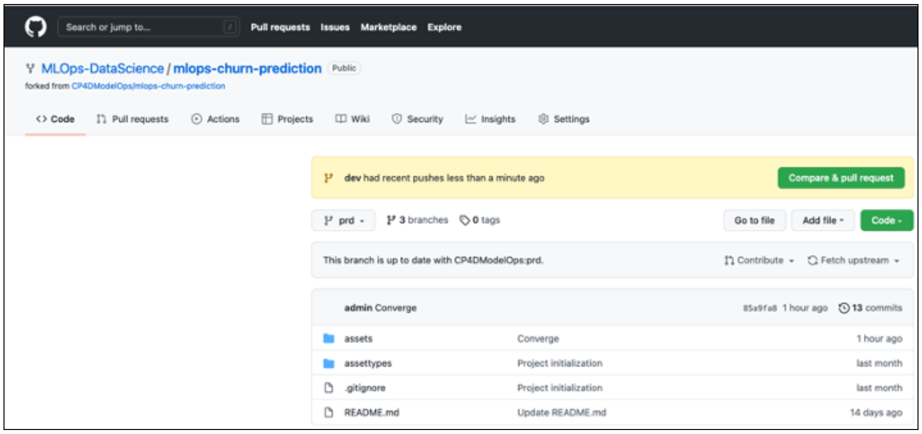

Figure 5-234 Recent pushes

93.Create a pull request to merge changes with the UAT branch. (Normally, this process is done by the lead data scientist.

Click **Compare & pull request**, change the left repository to your own repository and then, select the **UAT** branch. Your window should resemble the example that is shown in Figure 5-235.

Enter a meaningful description of the pull request and optionally, enter a longer explanation of the changes that are to be merged. Then, click **Create pull request**.

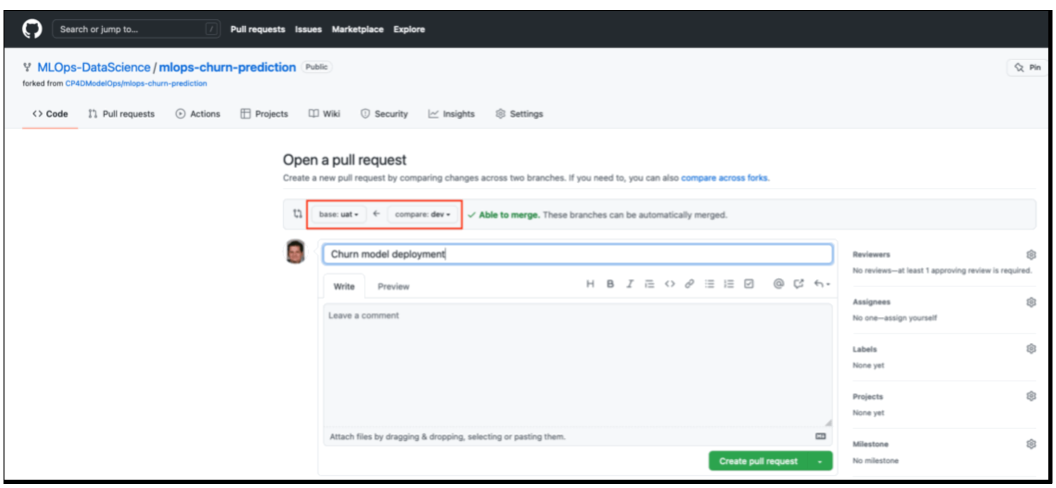

Figure 5-235 Create pull request for dev to uat

94. Observe that the merge process is blocked because it requires at least one approving review. This issue occurs because we set up the branch protection rules for the UAT branch (see Figure 5-236).

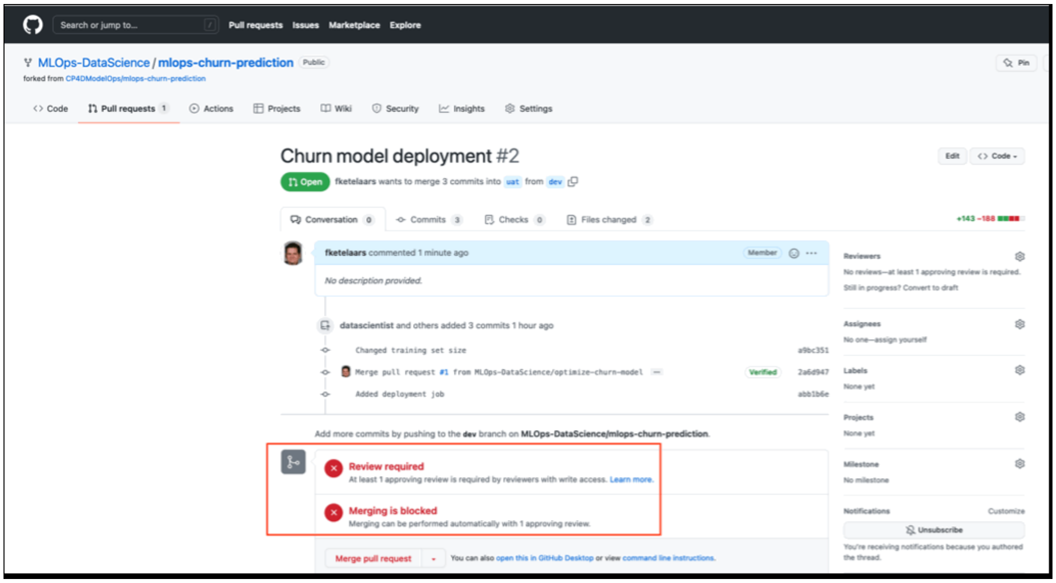

Figure 5-236 Review required

95. Because we have only a single user in the GitHub repository, a formal approval process cannot be demonstrated. For now, we force the merge by using the administrator privileges. Click **Merge pull request**. GitHub shows a check box where you can select administrator privileges. Ensure that the check box is selected and then, click **Confirm merge** (see Figure 5-237).

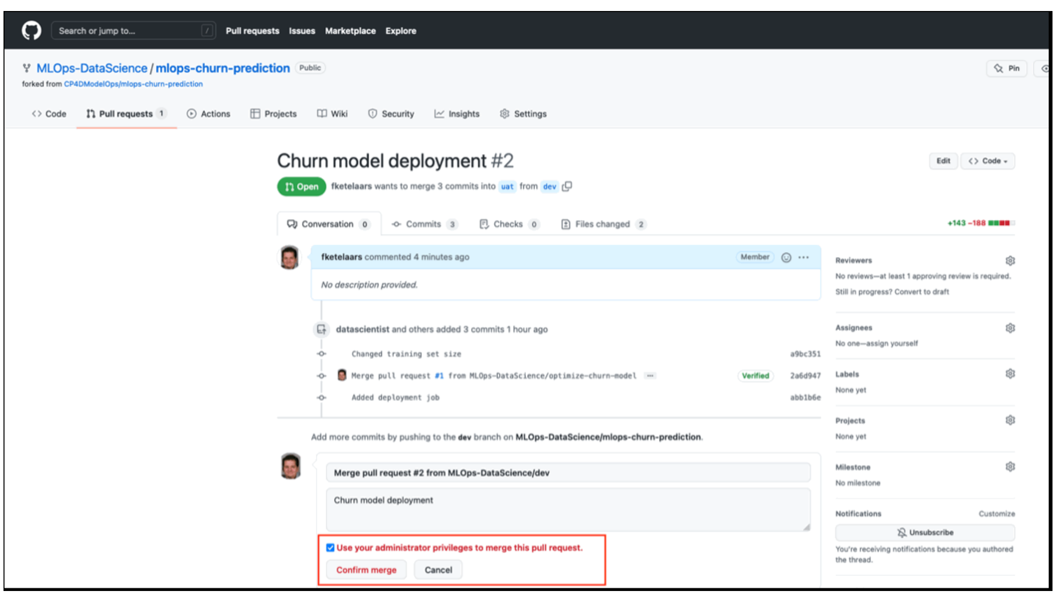

Figure 5-237 Merge as administrator

The changes are merged, and you are prompted whether you want to delete the DEV branch. In our scenario, we keep this branch in place because other team members are working in that branch.

In a real implementation, you can prevent unintentional deletion of specific branches by setting authorizations in GitHub.

96. Log out from Cloud Pak for Data and log back in again by using the `uatops` user. Navigate to the projects. You find that this user does not have any projects yet. Usually, the `uatops` user is a "service account" that runs processes as part of a continuous deployment application, such as Kubeflow, IBM Watson Studio Pipeline, Tekton, GitLab pipelines, and Jenkins. Because the environment does not have a CD application installed, we run the pipeline manually.

97. As the `uatops` user, create a project and select the option to create a project that is integrated with a Git repository (see Figure 5-238).

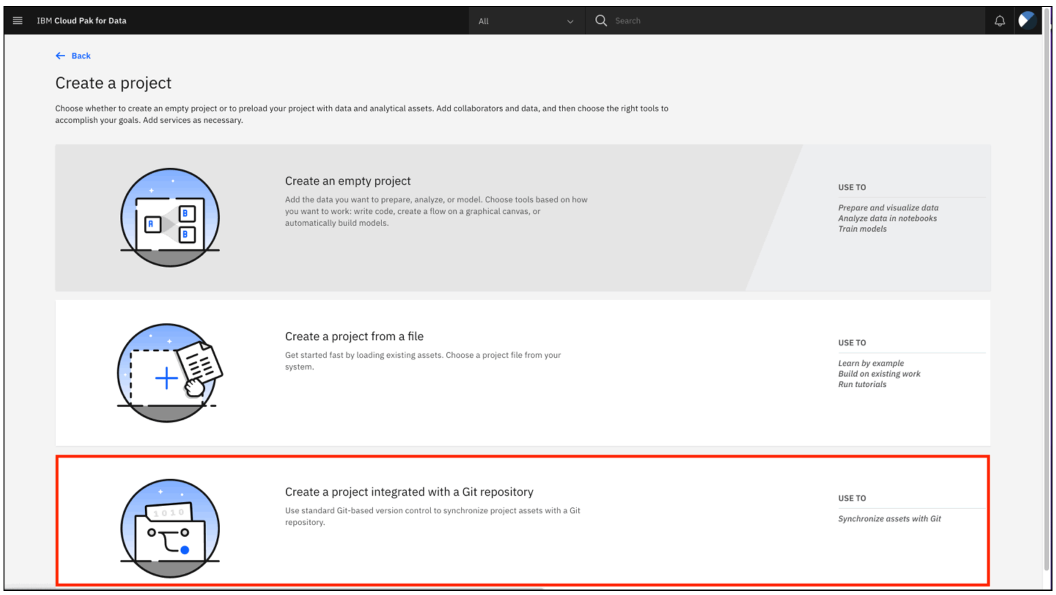

Figure 5-238 New Git-integrated project

98. Choose a representative name, such as `mlopsuat` and create a token. (You can use the same token as before.) Click **Continue** to create the token in Cloud Pak for Data (see Figure 5-239 on page 408).

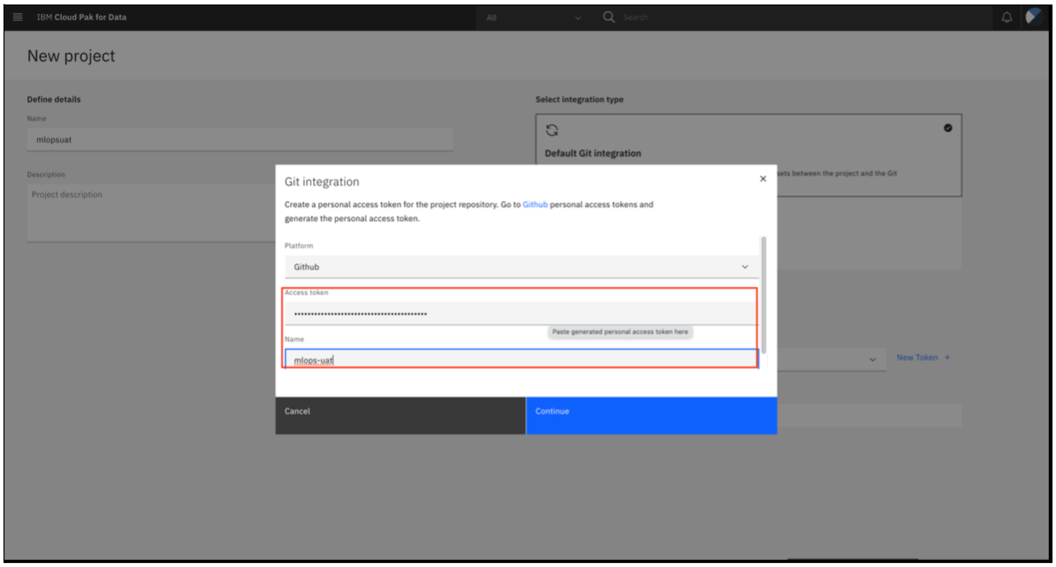

Figure 5-239 Creating a token

99.Select the token and enter the URL of your repository. You also can select the UAT branch (see Figure 5-240).

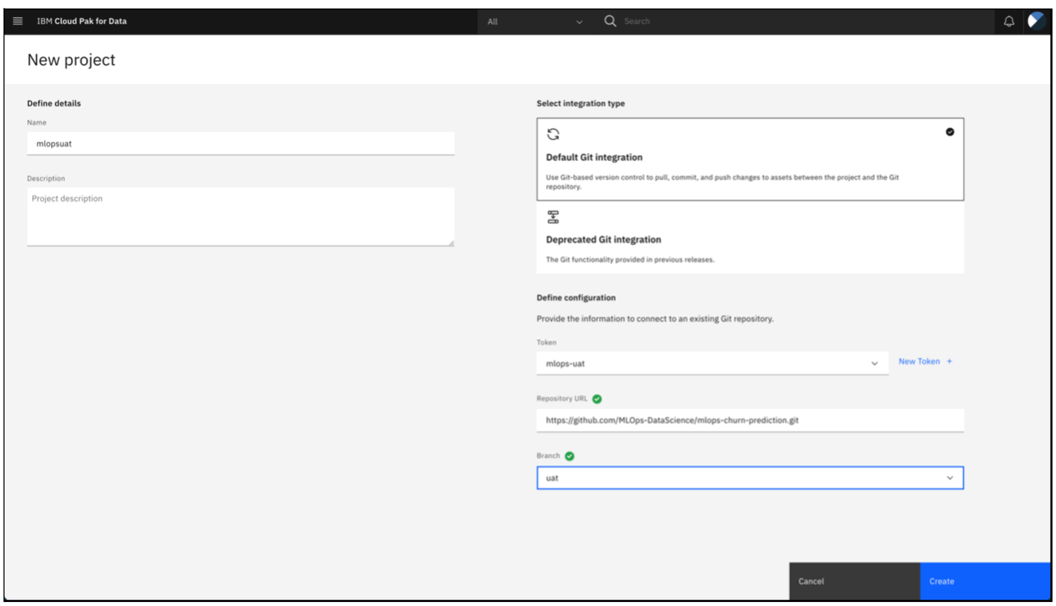

Figure 5-240 Selecting token and branch

100.After the project is imported, a completion window opens. Click **View new project** (see Figure 5-241).

Figure 5-241 Project created confirmation message

101.Navigate to the **Jobs** tab. Because you created the project from the UAT branch of the repository, the local project is already updated. If you created the Git-integrated project before, select the correct branch and pull the changes from the remote repository. You see that the `Deploy-churn-prediction-model` job already was created (see Figure 5-242).

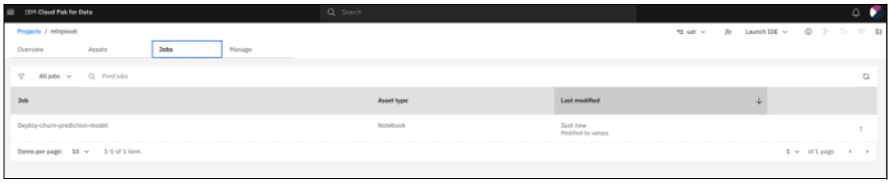

Figure 5-242 Jobs

102.Click the job name to view the details. Then, click **Edit Configuration icon** to find the Jupyter notebook that is to be run (Associated file)., as shown in Figure 5-243.

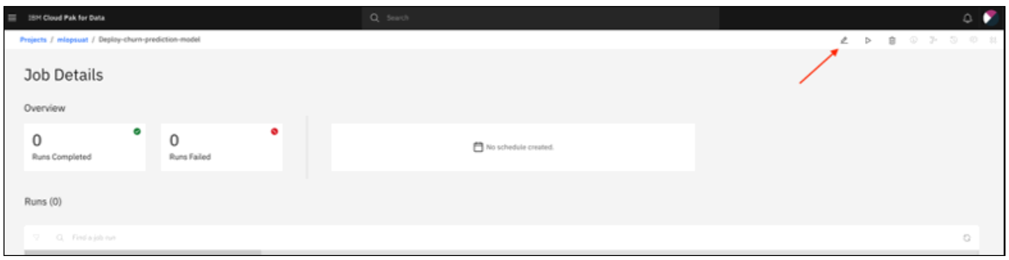

Figure 5-243 Edit job configuration

The job details are shown in Figure 5-244.

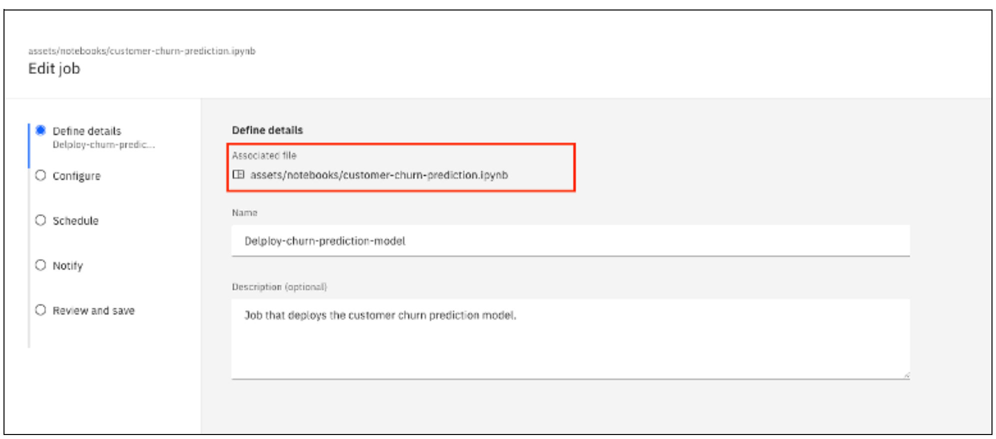

Figure 5-244 Job details

103.Run the job by clicking **Play**, as highlighted by the arrow in Figure 5-245. The job starts to run.

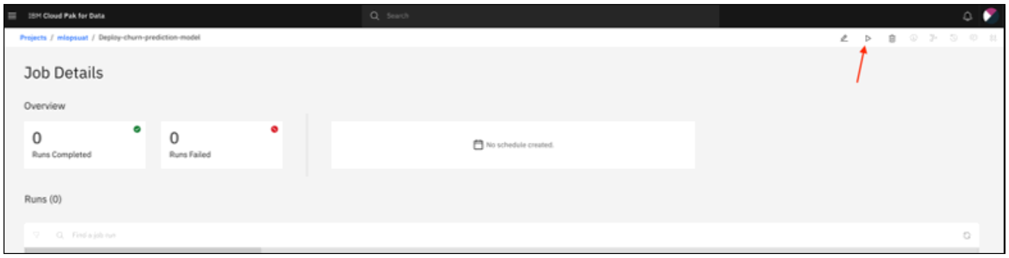

Figure 5-245 Running the job

The job details are shown in Figure 5-246.

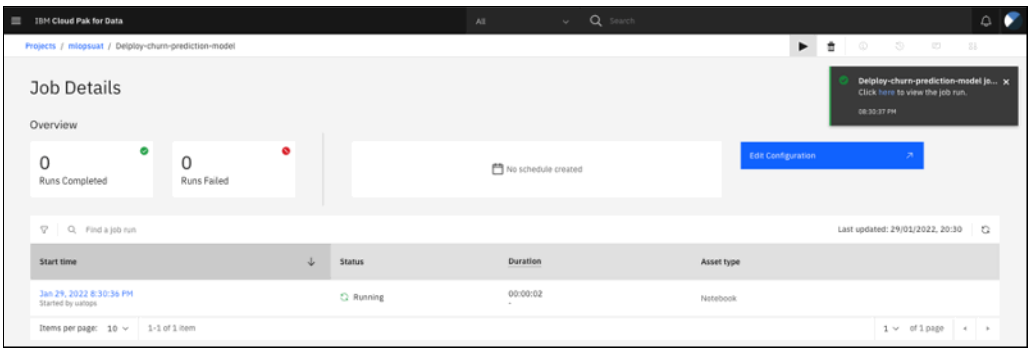

Figure 5-246 Job details

104.Click the running job in the list so that you can monitor its progress. Because the data set is small, the job takes only a few seconds to complete. When data sets are large or if more complex algorithms, such as deep learning models are run, the job can take significantly longer, such as hours, even more than a day (see Figure 5-247).

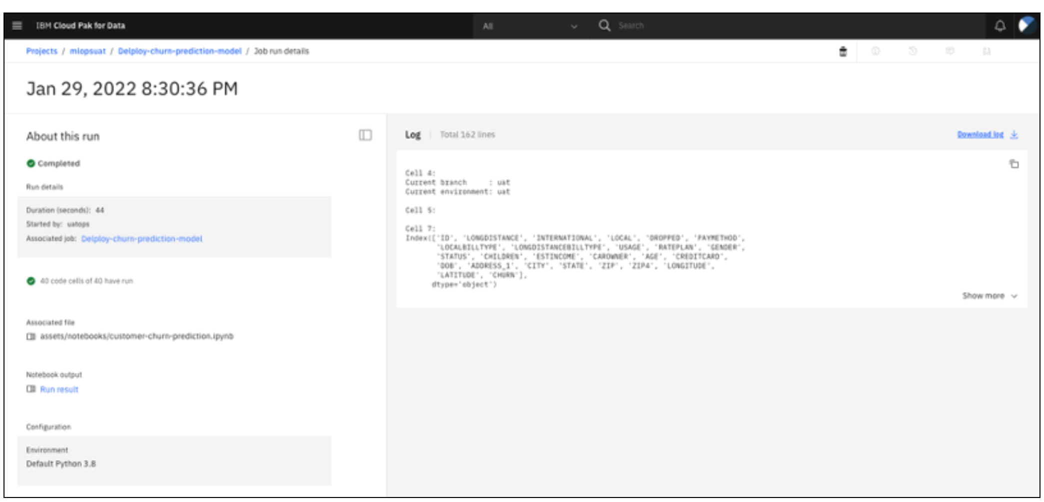

Figure 5-247 Job output

Although not directly visible in the job log, the model is being trained on the `CUSTOMER_DATA_ready-uat.csv` file, which represents the data in the UAT environment. The current branch (UAT) and environment (also UAT) is shown at the top of the job log.

105.Click the navigation menu and then, click **Deployments** to check whether a new deployment space, `churn-uat`, was created (see Figure 5-248).

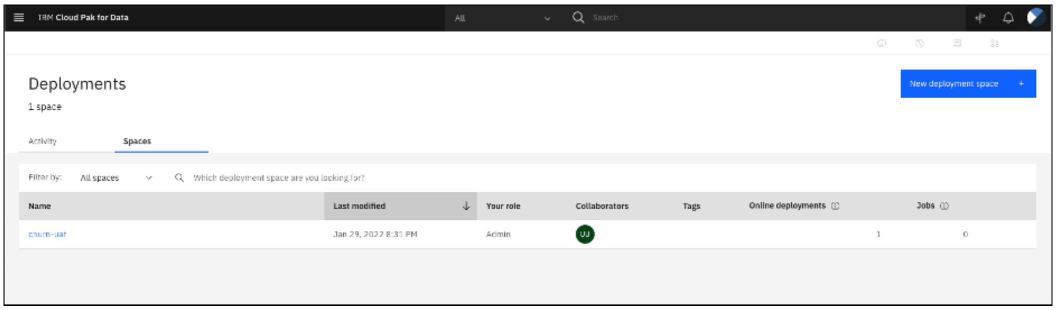

Figure 5-248 New deployment

106.Click the **churn-uat** deployment space and navigate to the **Deployments** tab. You find the `churn_pipeline` model and its deployment, `churn_pipeline_deployment` (see Figure 5-249).

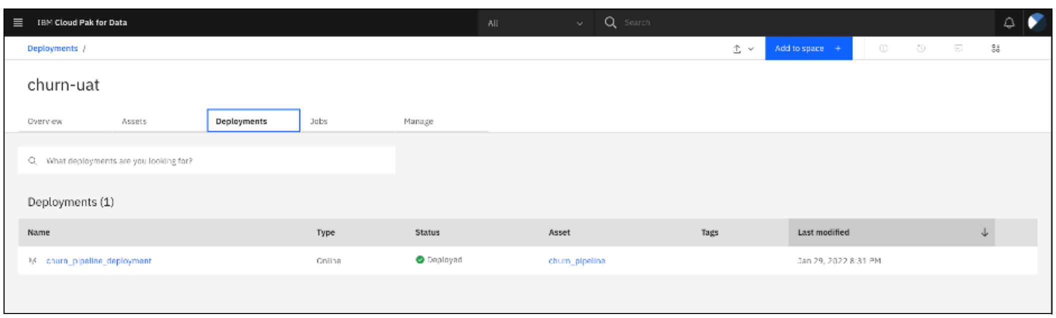

Figure 5-249 Deployments

The running of the job normally is done by a “service account” that uses a CD pipeline, such as Kubeflow, GitLab pipeline, Red Hat OpenShift pipelines (Tekton), or a similar technology. The job log is kept as part of the pipeline run. After the model is deployed, the project can be deleted.

This concludes the Git portion of the lab. Optionally, you can continue to push the changes to the PRD branch by using the same method as described here and then, deploy the model in the `churn-prd` deployment space.

5.4.11 IBM Watson Studio Pipelines

In previous modules, you created tasks that combined different data sets by using Data Refinery, and created and trained a model by using an AutoAI experiment and a notebook, and then, y deployed the trained model.

In this module, you learn how to automate all the previous steps by using IBM Watson Studio Pipelines. IBM Watson Studio Pipelines helps to manage the AI lifecycle through automation; for example, the machine learning model training, deployment, evaluation, production, and retraining.

Complete the following steps:

1. Log into Cloud Pak for Data as the `datascientist` user.
2. Go to deployment spaces by clicking the navigation menu (top left navigation icon) and clicking **Deployments**.
3. Create a deployment space that is named `pipelines`, which you use to deploy the churn model.
4. Go to projects by clicking the Navigation menu (upper left navigation icon) and clicking **Projects → All projects.**
5. Select the **Customer Churn Prediction** project that you created earlier (see Figure 5-250).

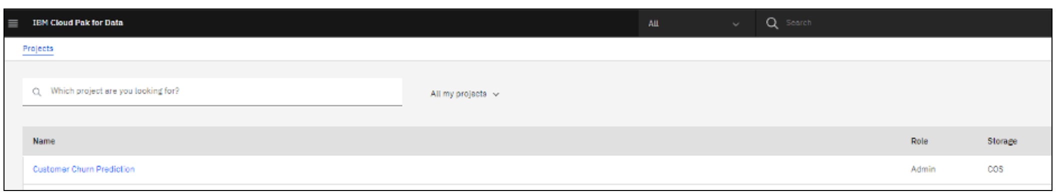

Figure 5-250 Selecting a project

6. Click **New asset** and then, select **Pipelines** (see Figure 5-251).

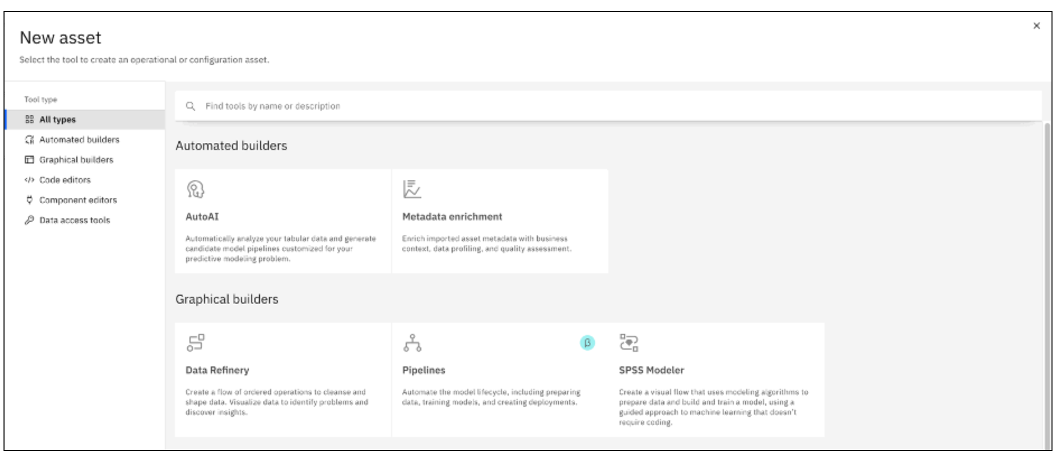

Figure 5-251 Creating a pipeline

7. Enter a Name (in our example, `Pipeline-churn`) and a Description (optional) for your pipeline and then, click **Create** (see Figure 5-252).

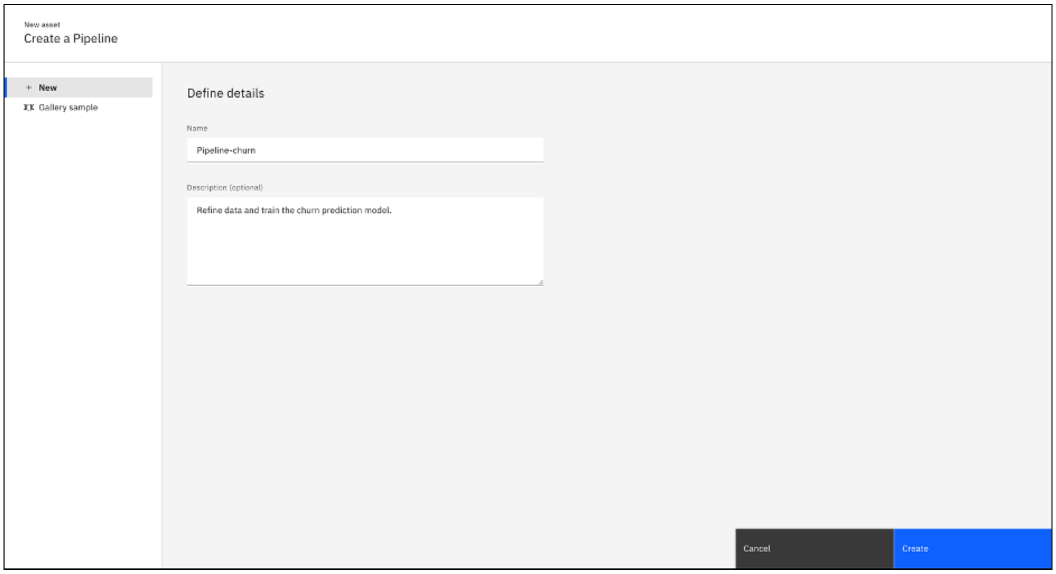

Figure 5-252 Pipeline details

8. In the left-side menu, you see the Copy option. Click the arrow to expand the menu and then, click **Copy Assets**. Drop **Copy Assets** to the canvas (see Figure 5-253).

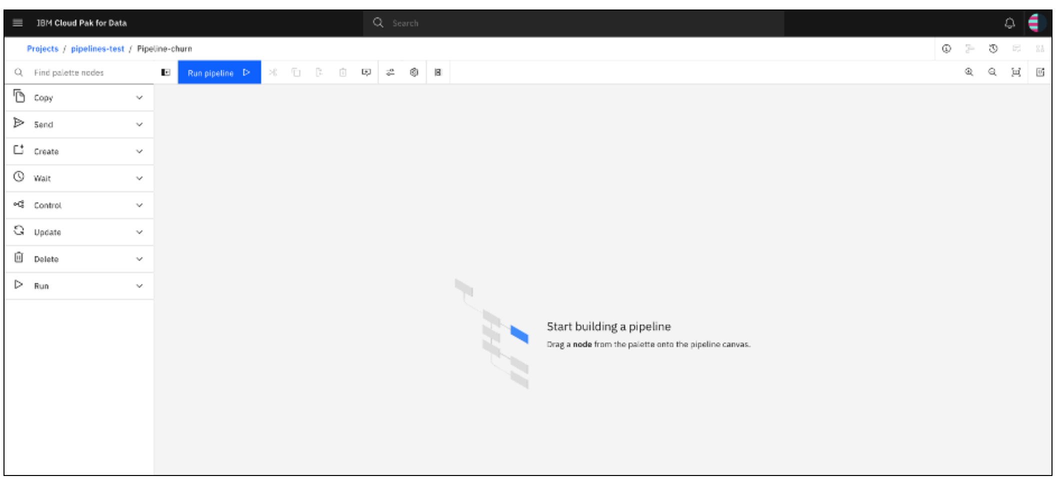

Figure 5-253 Copy assets

9. Double-click **Copy Assets** to open the properties of the node. Alternatively, when the cursor is on the top of the node, three horizontal blue points become visible on the right side of the node. Click the button and the menu pops up. Click **Open**.

10.In the properties on the right side of the window, click **Select Asset** (see Figure 5-254).

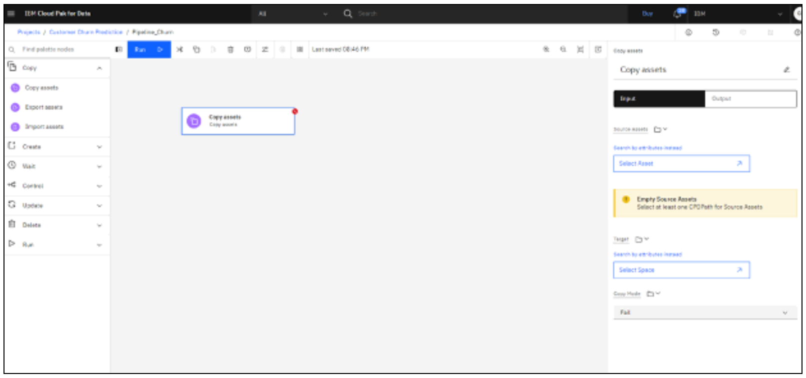

Figure 5-254 Selecting an asset

11.In the pop-up window, browse in the project that is used (`Customer Churn Prediction`). In the **Categories** tab, select **Data asset**. In the **Data assets** tab, select **customer_data_transactions.csv, customer_personal_info_simplified.csv** and **customer_churn_labels.csv**. Click **Choose** (see Figure 5-255).

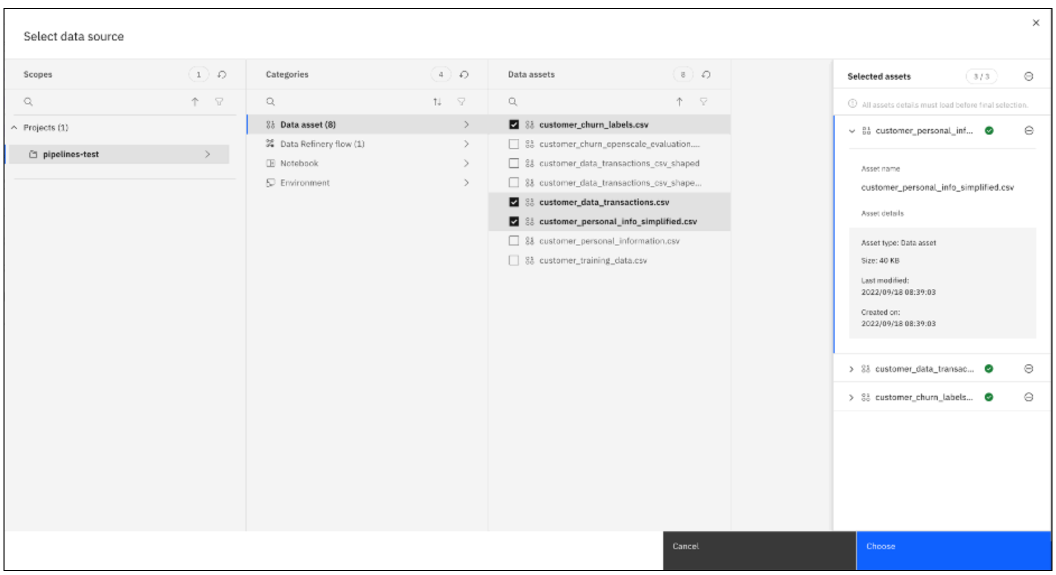

Figure 5-255 Selecting files

12. After you selected the assets, set the target to the pipeline deployment space that you created earlier and change the Copy Mode to Overwrite. This copy mode is required if you want to run the pipeline more than once and overwrite the `.csv` files that are used as input for the refinery job (see Figure 5-256).

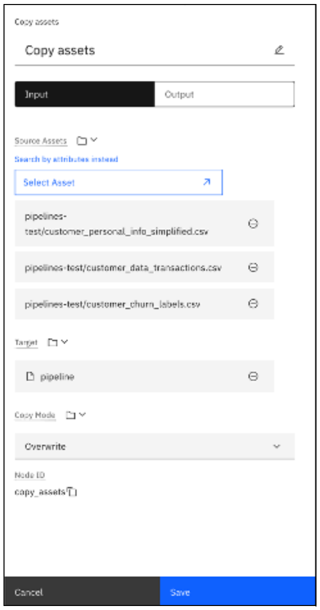

Figure 5-256 Overwrite

13. Click the **Output** tab to see the list of Output Assets. Then, click **Save** (see Figure 5-257).

Figure 5-257 Save

14. On the menu at the left side, expand the **Run** option. Drop **Run Data Refinery flow** to the gray canvas. Double-click the new **Run Data Refinery flow** node to open the node properties on the right. Click **Select Data Refinery Flow** (see Figure 5-258 on page 416).

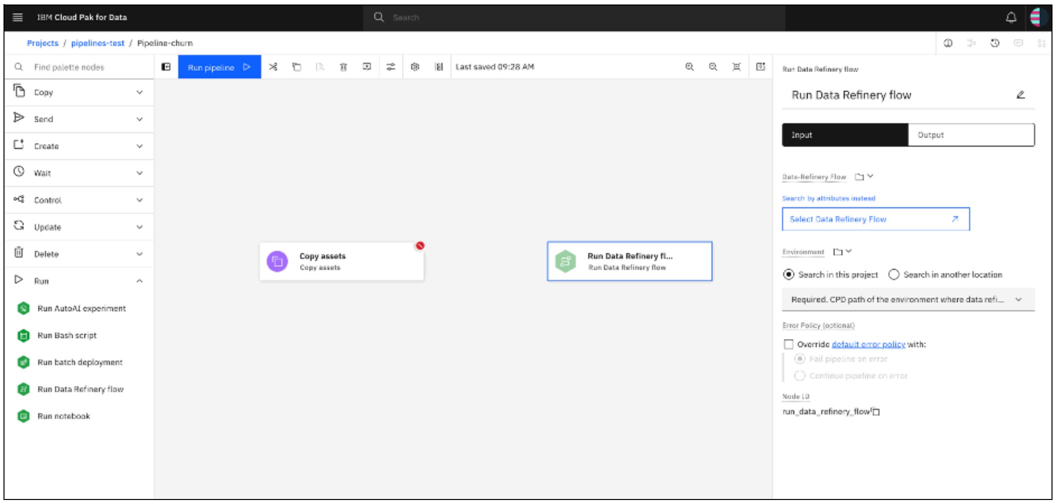

Figure 5-258 Adding Data Refinery flow

15. In the pop-up window, navigate in the project that is used (in our example, `Customer Churn Prediction`), then the **Categories** tab and then, select **Data Refinery flow**. Under the Data Refinery flow tab, select the flow that was created (**customer_data_transactions.csv_flow**). Click **Choose**.
16. In the Environment option, select **Default Data Refinery XS** and then, click **Save** (see Figure 5-259).

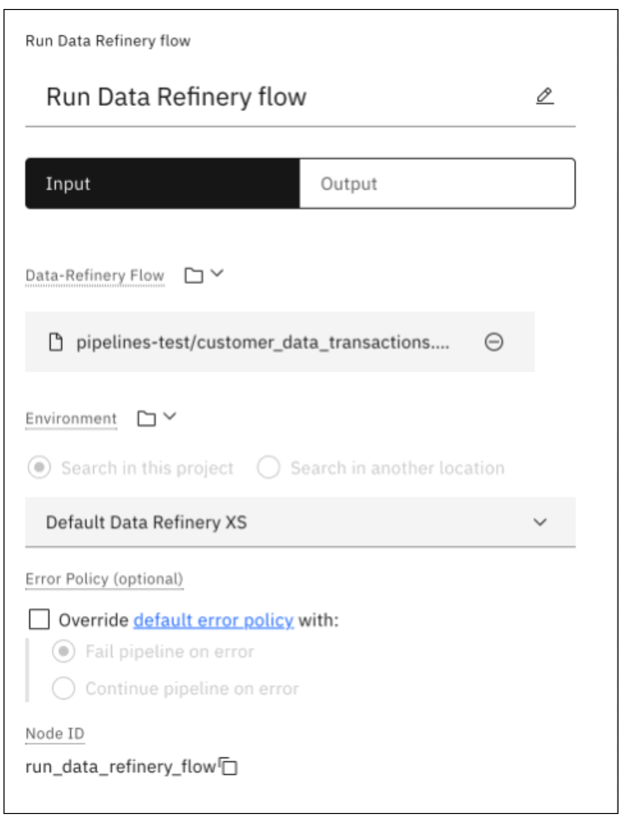

Figure 5-259 Refinery environment

17. Position the cursor at the top or bottom of the Copy Assets node. A blue arrow appears, as shown in Figure 5-260.

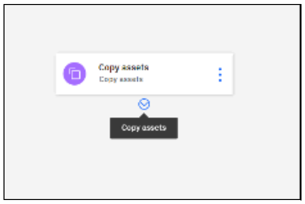

Figure 5-260 Link node

18. Click the blue arrow and drop the Run Data Refinery flow. A blue arrow is shown that connects the nodes, as shown in Figure 5-261.

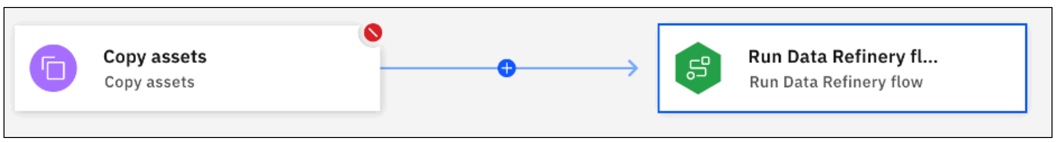

Figure 5-261 Linking two nodes

19. In the left menu, expand the Run option, locate the Run AutoAI experiment and drop the node to the gray canvas. Double-click the **Run AutoAI** experiment node. In the menu on the right side, click **Select AutoAI Experiment** (see Figure 5-262).

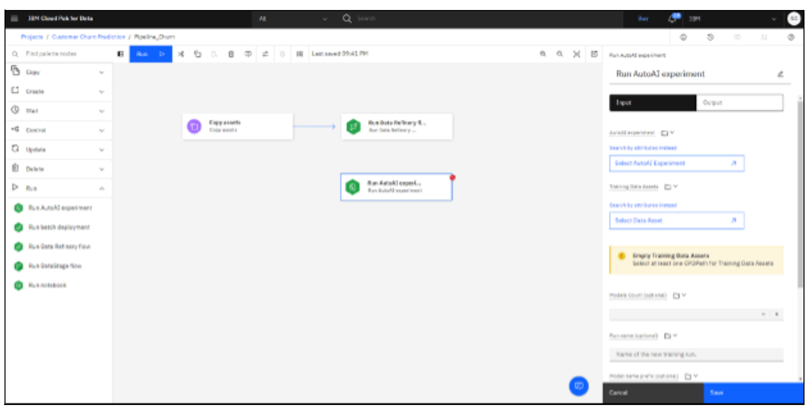

Figure 5-262 Add AutoAI experiment

20.In the pop-up window, browse in the project that is used (in our example, `Customer Churn Prediction`). Under the **Categories** tab, select **ML Pipeline**. Under the ML Pipeline tab, select the created AutoAI experiment **autoai_churn_prediction**. Click **Choose** (see Figure 5-263).

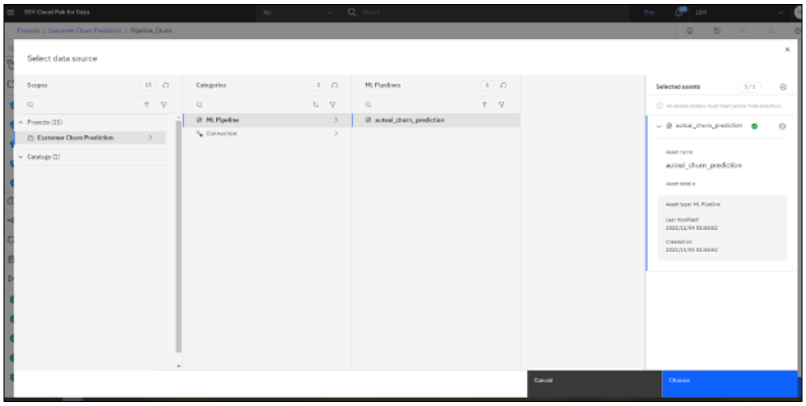

Figure 5-263 Selecting an experiment

21.In the Properties window, click **Training Data Assets** and then, select the arrow to change the source. Select the **Select from another node** option (see Figure 5-264).

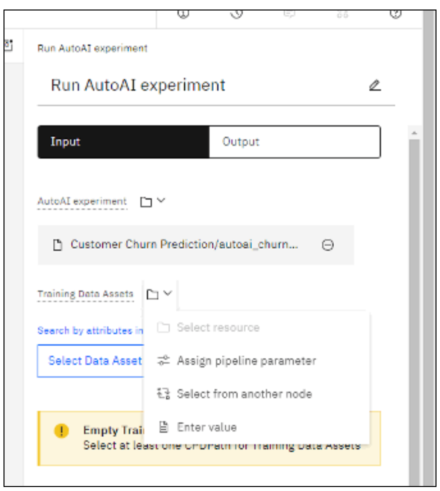

Figure 5-264 Selecting the training dataset

22. In the Select node window, select **Run Data Refinery flow** (see Figure 5-265).

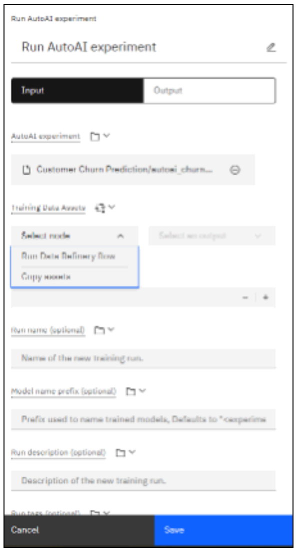

Figure 5-265 Selecting run

23. Position the cursor at the top or bottom of the **Run Data Refinery flow** node. A blue arrow appears. Click the **blue arrow** and drop the **Run AutoAI experiment** node. A blue arrow is shown that connects both nodes (see Figure 5-266).

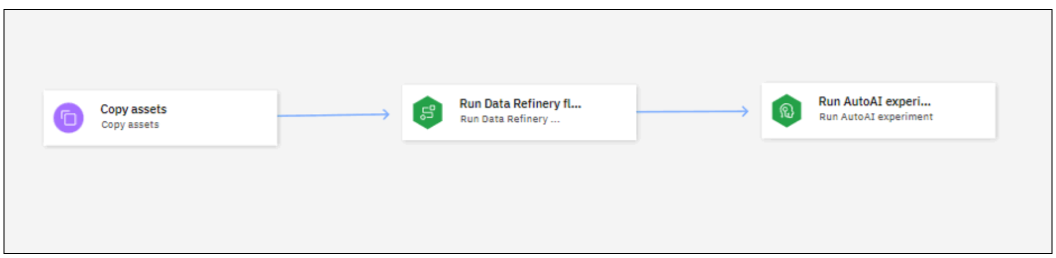

Figure 5-266 Linking two nodes

24. In the menu on the left side, expand the **Update** option, locate the **Update web service** node and drop the node to the gray canvas. Position the cursor to the right of the Run AutoAI experiment node. Click the blue arrow and drop the node to the top of the Update Web Service node. Then, double-click **Update Web Service** (see Figure 5-267).

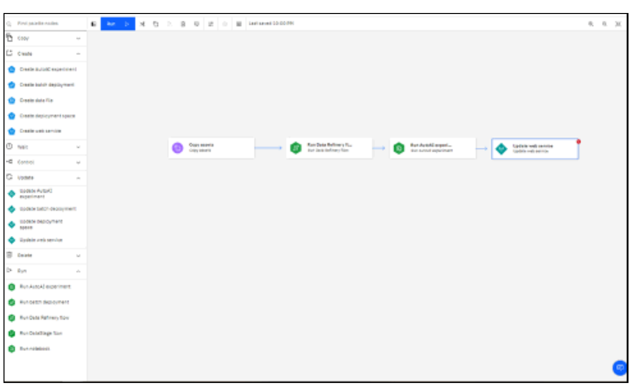

Figure 5-267 Adding the web service update

25. In the menu on the right on **Search Space**, select the name of the space that you created. In the Deployment name, enter the name of the online deployment that was created for the WML model. Click **Save**. The online deployment that is run on previous exercises is updated (see Figure 5-268).

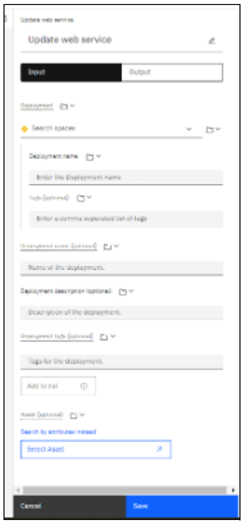

Figure 5-268 Search space

26. Click **Run** and choose **Trial Run** to test the pipeline (see Figure 5-269).

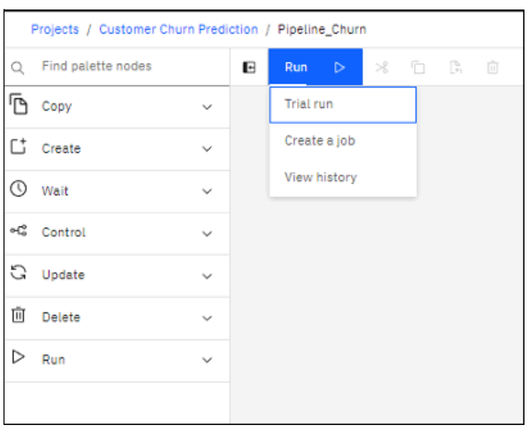

Figure 5-269 Trial run

27. A window opens in which you enter your API key. If you do not have an API Key, click **Generate new API.** A new window prompts for a name for the key. Click **Save**.

    If you have an API key, click **Use Existing API key** and enter the key. Click **Save**. In the Trial Run window, click **Run** (see Figure 5-270 on page 421).

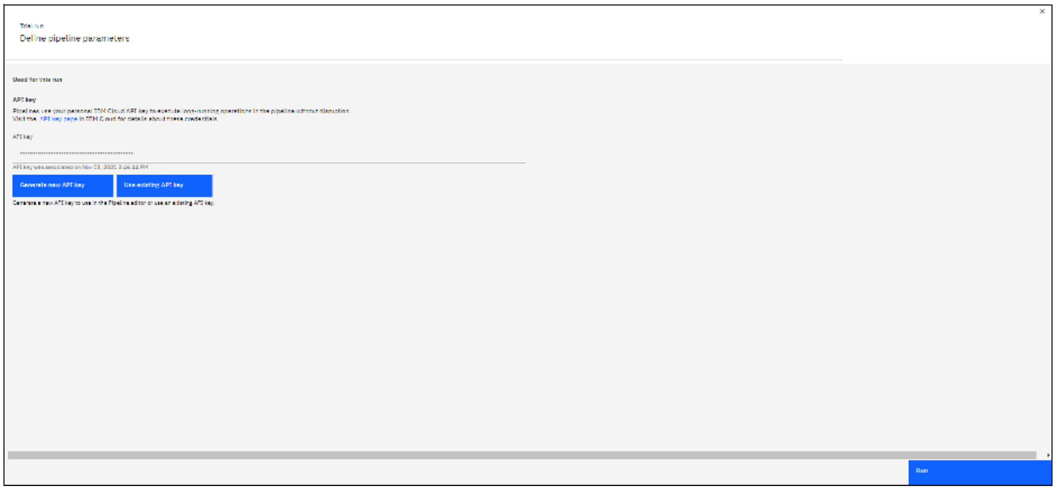

Figure 5-270 Run

The pipeline starts running (see Figure 5-271).

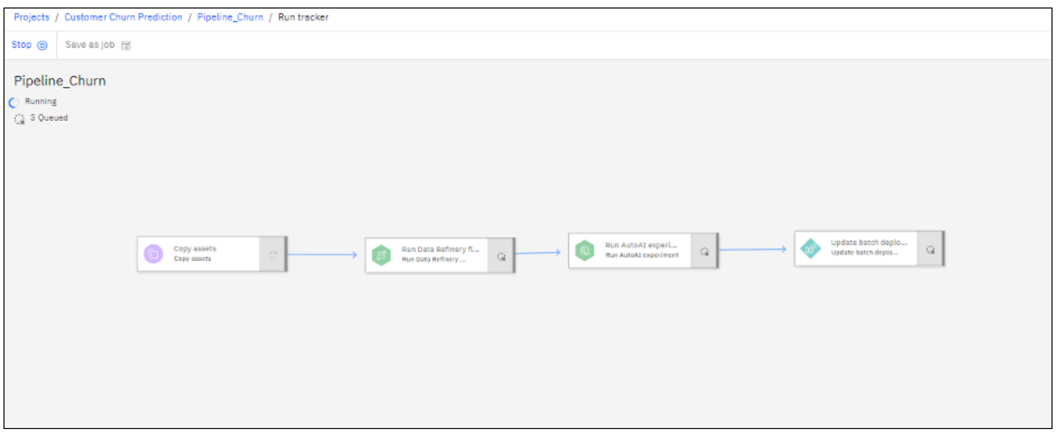

Figure 5-271 Pipeline run

After the process completes and successfully runs, the deployment that is run in previous steps is updated, including the OpenScale results.

28. Return to the pipeline creation window and click **Run** that is in the blue square and then, click **Create a job** (See Figure 5-272).

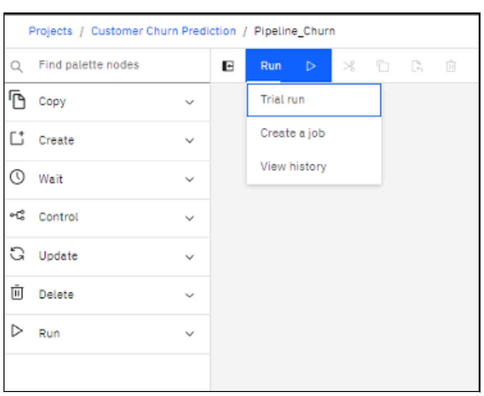

Figure 5-272 Create job

29. Enter a name for the pipeline and optionally a description. Click **Next** (see Figure 5-273).

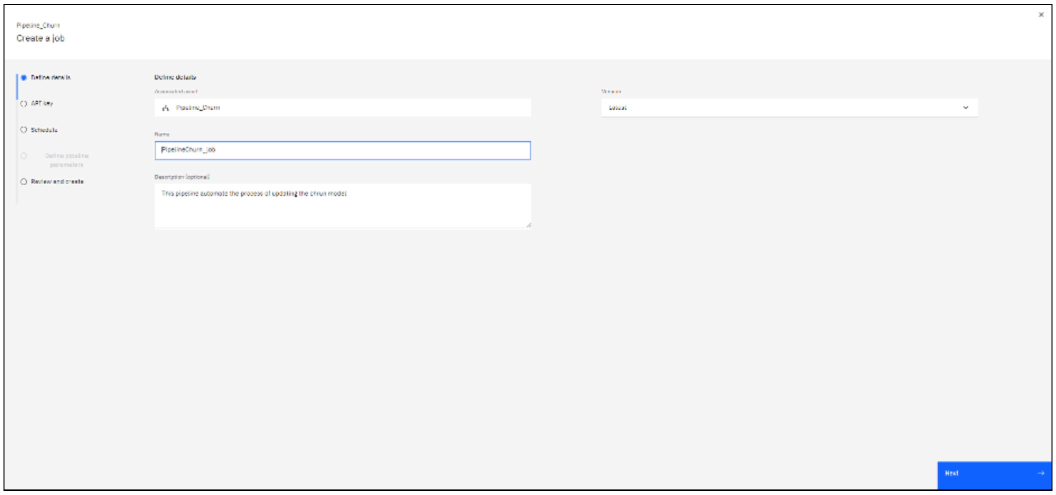

Figure 5-273 Pipeline details

30. The API key that was added previously should be associated. If not, choose between the **Generate new API key** or **Use existing API key** options. Click **Next** (see Figure 5-274).

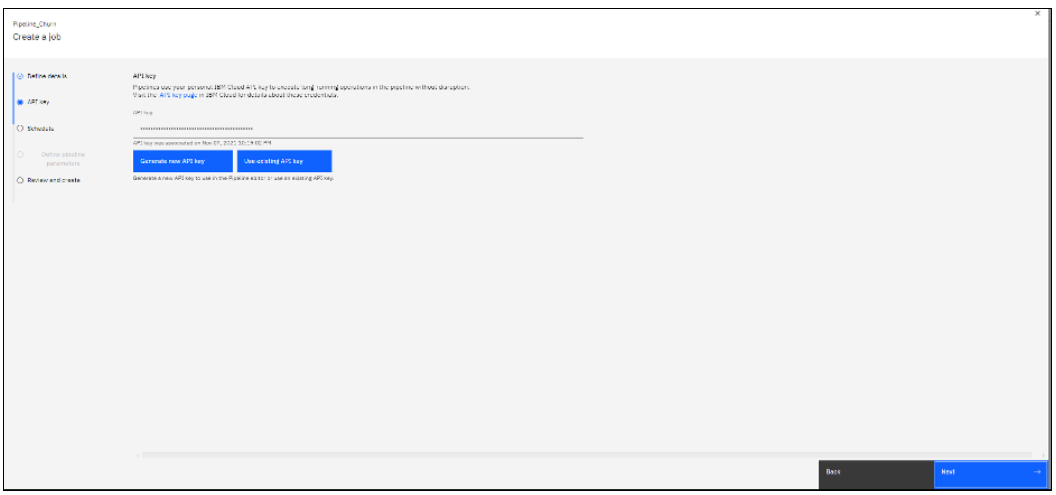

Figure 5-274 Generating API key

31. By default, the Schedule is off. You can schedule the job to run and define when to start and how often you want it to run. Click **Next** (see Figure 5-275).

Figure 5-275 Schedule

32. Review the details of the job and click **Create** (see Figure 5-276).

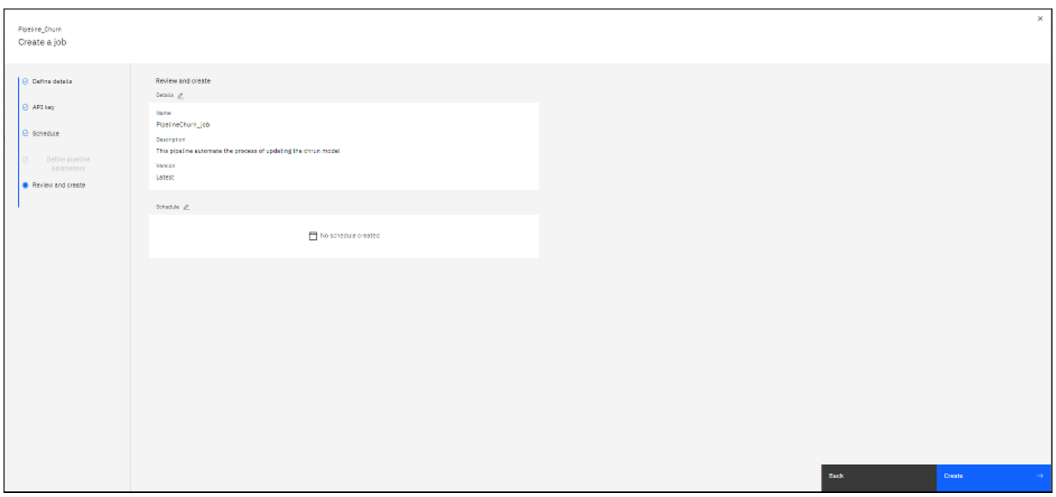

Figure 5-276 Review and create

33. To run the job, click **Start**. The job starts, and the status is provided (see Figure 5-277).

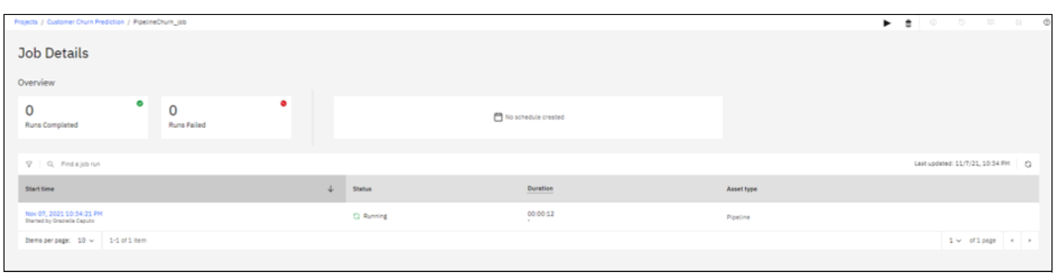

Figure 5-277 Running the job

5.4.12 Model Risk Management

IBM OpenPages for IBM Cloud Pak for Data is an integrated governance, risk, and compliance platform that companies can use as a tool to assist in managing risk and regulatory challenges across the enterprise.

It provides a set of core services and components that span risk and compliance domains, which include the following examples:

- Operational risk
- Policy management
- Financial controls management
- IT governance
- Internal audit
- Model risk governance
- Regulatory compliance management
- Third-party risk management
- Business continuity management

IBM OpenPages for IBM Cloud Pak for Data provides a powerful, highly scalable, and dynamic toolset that empowers managers with information transparency and the capability to identify, manage, monitor, and report on risk and compliance initiatives on an enterprise-wide scale (see Figure 5-278).

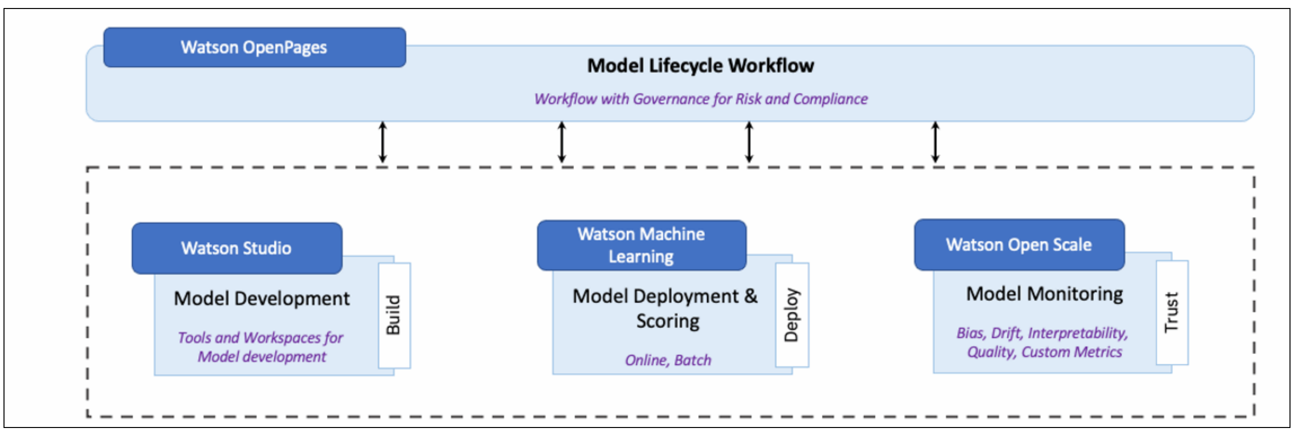

Figure 5-278 Cloud Pak for Data components for Model Governance

The core capabilities cover the functional areas that are shown in Figure 5-279.

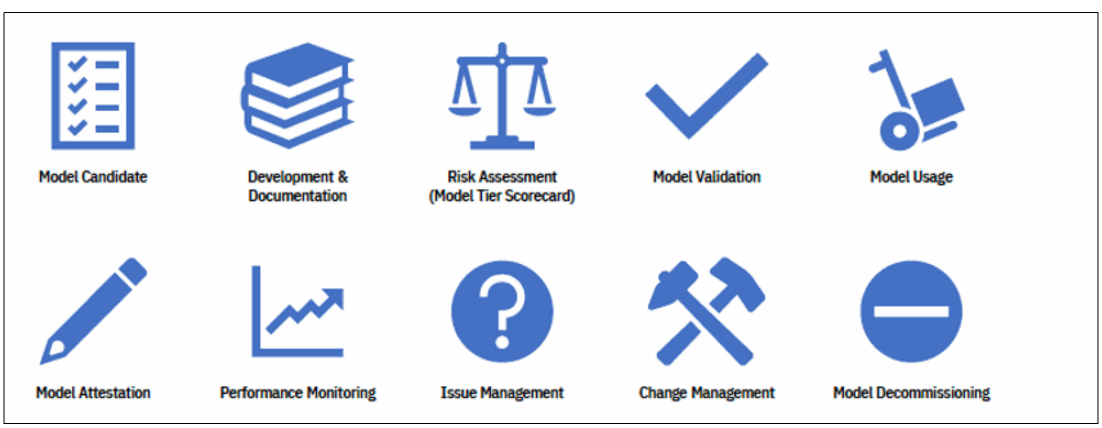

Figure 5-279 IBM OpenPages Core capabilities

Setting up model risk monitoring and governance with IBM OpenPages and IBM Watson OpenScale

The end-to-end model risk monitoring and governance solution in Cloud Pak for Data are enabled by using IBM OpenPages and IBM Watson OpenScale (see Figure 5-280).

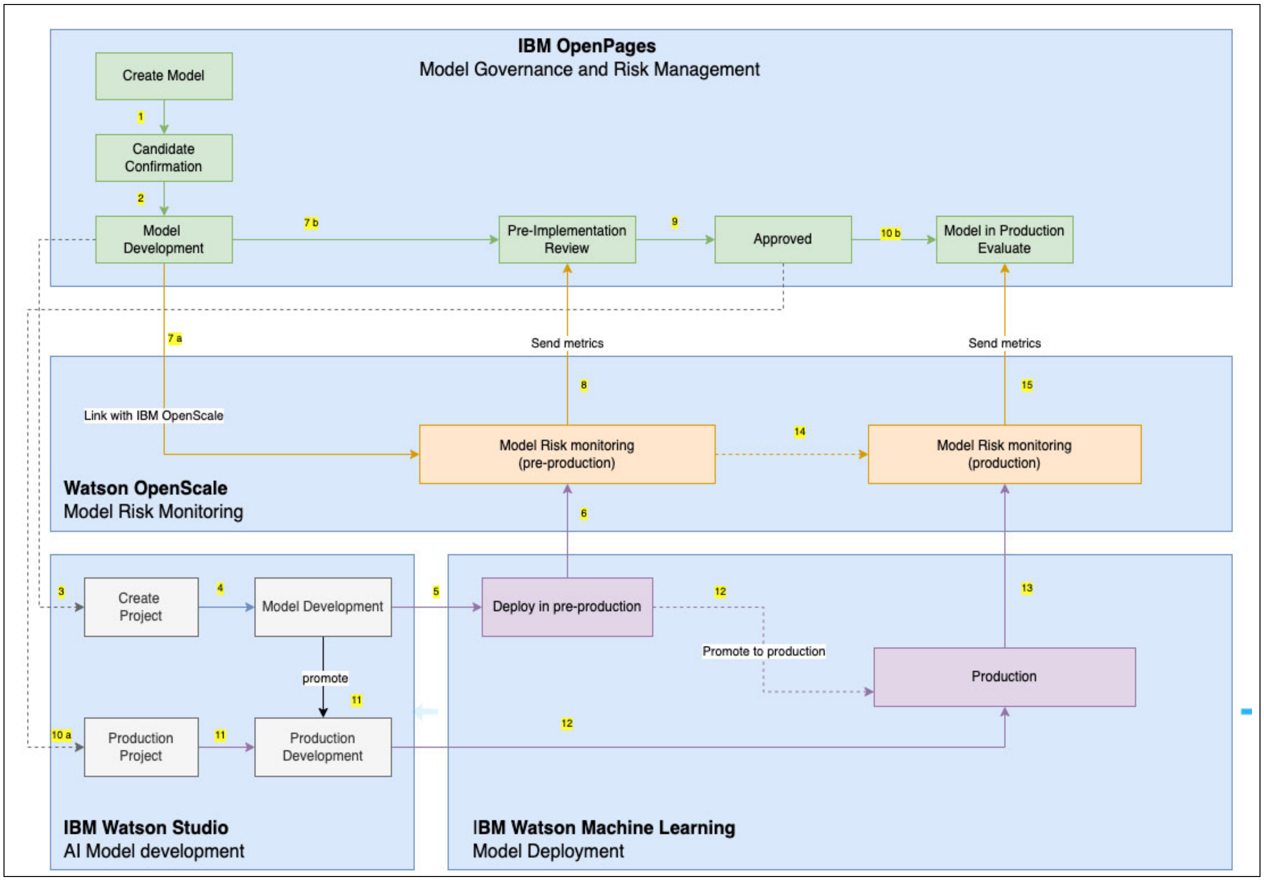

Figure 5-280 Sample Model Risk and Governance workflow using IBM OpenPages and Watson OpenScale

Setting up a new model in IBM OpenPages

Complete the following steps in IBM OpenPages to create a model to be used for Machine Learning model governance:

1. Create a Business Entity by clicking **Home** → **Organization** → **Business Entities**. Name the entity `ABC Telecom` (see Figure 5-282).

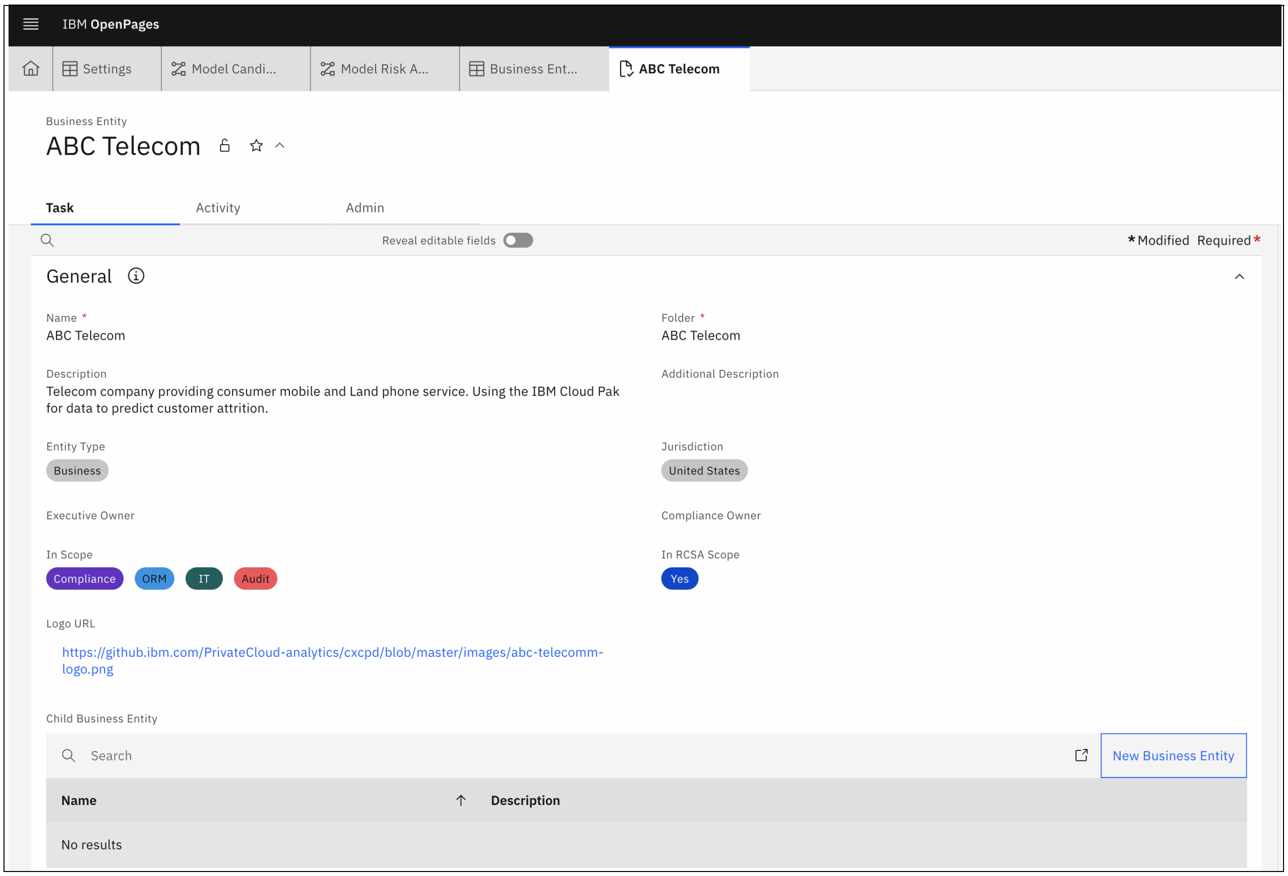

Figure 5-281 Creating OpenPages business entity

2. In the menu in the upper left of the window, click **Home** → **Inventory** → **Models**. Click **Add New** (see Figure 5-282).

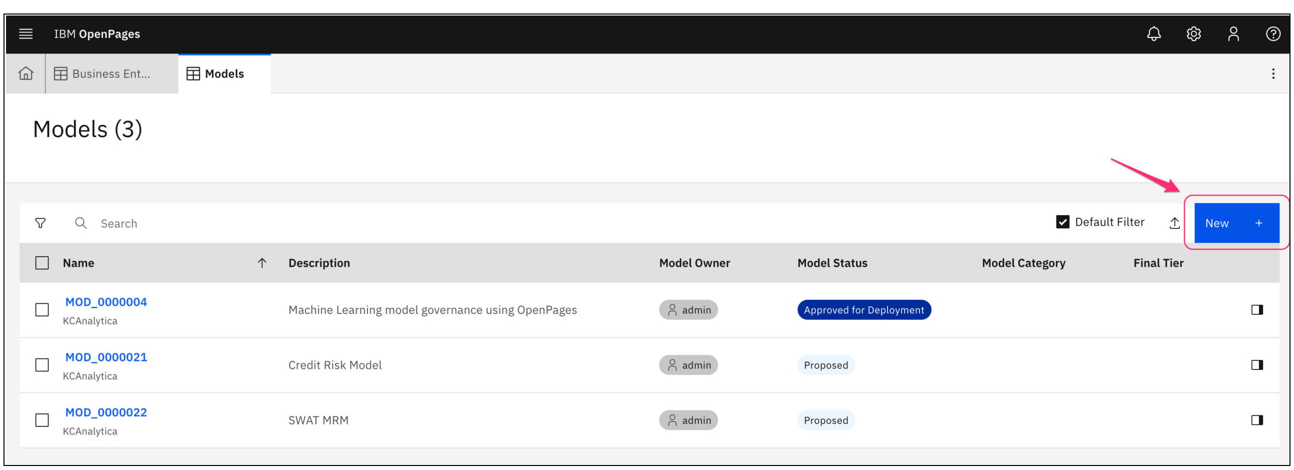

Figure 5-282 Creating a model

3. Create a model the includes the following values:
   - Description
   - Model Status: `Proposed`
   - Model Owner: *`Your account name`*
   - Model/Non Model: `Model`
   - Machine Learning Model: `Yes`
   - Monitored with OpenScale: `Yes`
   - Parent Entity, where the Business Entity is set to your organization's name; for example ABC Telecom
4. Click **Save** (see Figure 5-283).

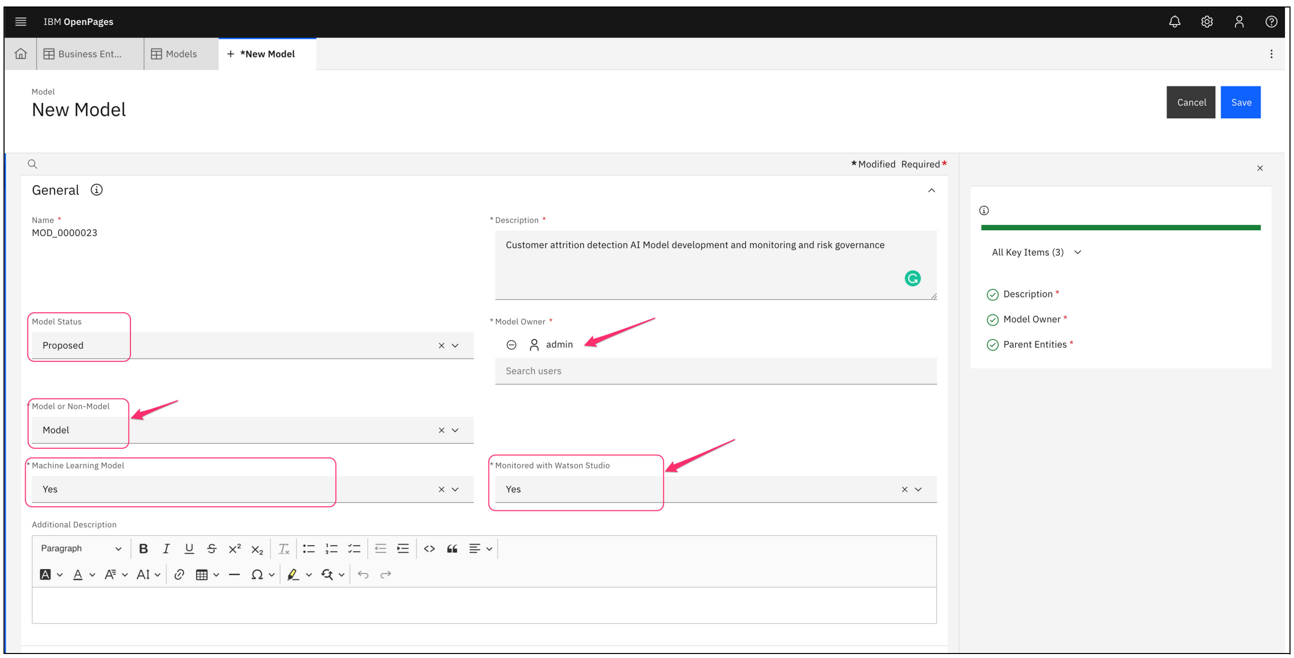

Figure 5-283 Creating a model in OpenPages

5. Move the model through the Candidate Workflow in IBM OpenPages.

   Different stages of the workflow require specific roles, such as owner, developer, and head of the model review. (In this example, we use the `admin` user for all roles), as shown in Figure 5-284.

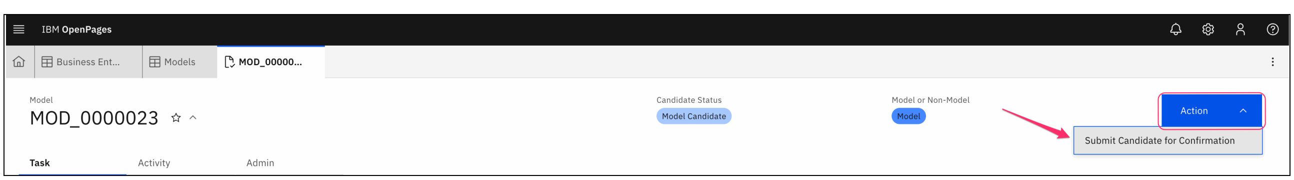

Figure 5-284 Sending the model to the Reviewer

Complete the following steps:

a. In the model that was created in Step 1, enter a Candidate Comment and then, click **Save**. In this field. the model owner can describe why the proposed model is a model and not a non-model.

b. From the Action drop-down menu, click **Submit Candidate for Confirmation**. In a live workflow, this step sends the candidate model to a reviewer for approval.

c. From the Action drop-down menu, click **Confirm Assessment**. In a live workflow, a reviewer confirms that the model candidate is a model, and the candidate workflow is complete.

6. Move the model through the Model Development workflow to the pre-implementation review stage in IBM OpenPages:

   a. From the Action drop-down menu, click **Start Model Development**.

   b. Complete the required fields in the model object that are related to the development (they are listed on the right-side of the window). Click **Save**.

   c. For the purposes of this testing, enter your user account as the developer.

   d. From the Action drop-down menu, click **Assign to Developer**.

   e. From the Action drop-down menu, click **Submit for Pre-Implementation Review** (see Figure 5-285).

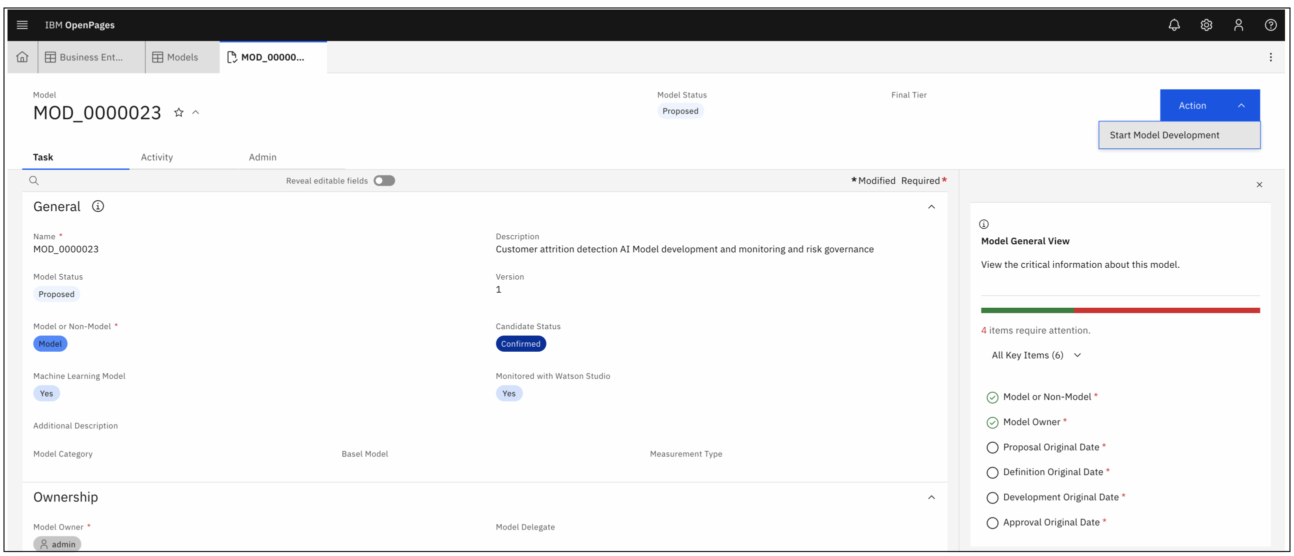

Figure 5-285 Sending the model to Start Model Development

7. Develop a Machine Learning Model in IBM Watson Studio. The following steps were completed in 5.4.5, “Building models” on page 312:

   a. In IBM Watson Studio, create a project that is named `Customer Churn Prediction`.

   b. Create a machine model by using Jupyter notebook in IBM Watson Studio.

   c. Create WML deployment space that is named `churn-preprod`.

   d. Deploy the model to the `churn-preprod` deployment space.

   e. Create the IBM Watson OpenScale instance and add the `churn-preprod` deployment space to monitoring.

8. Complete the following steps to connect IBM Watson OpenScale to IBM OpenPages:
   a. Log in to Cloud Pak for Data and open your IBM OpenPages instance.
   b. Generate the API key.
   c. Copy the URL that is displayed in the Access information section of the window (see Figure 5-286).

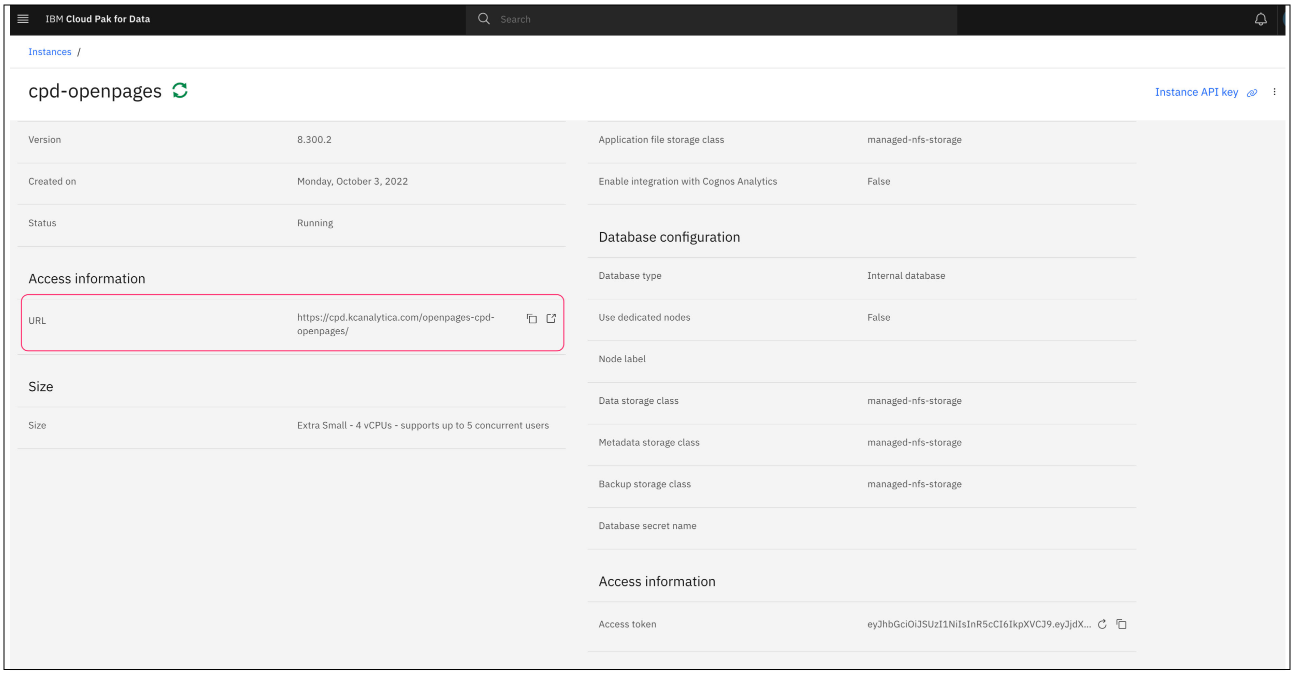

Figure 5-286 IBM OpenPages URL and credential

d. Select the **Integration** option in IBM Watson OpenScale to link IBM OpenPages (see Figure 5-288).

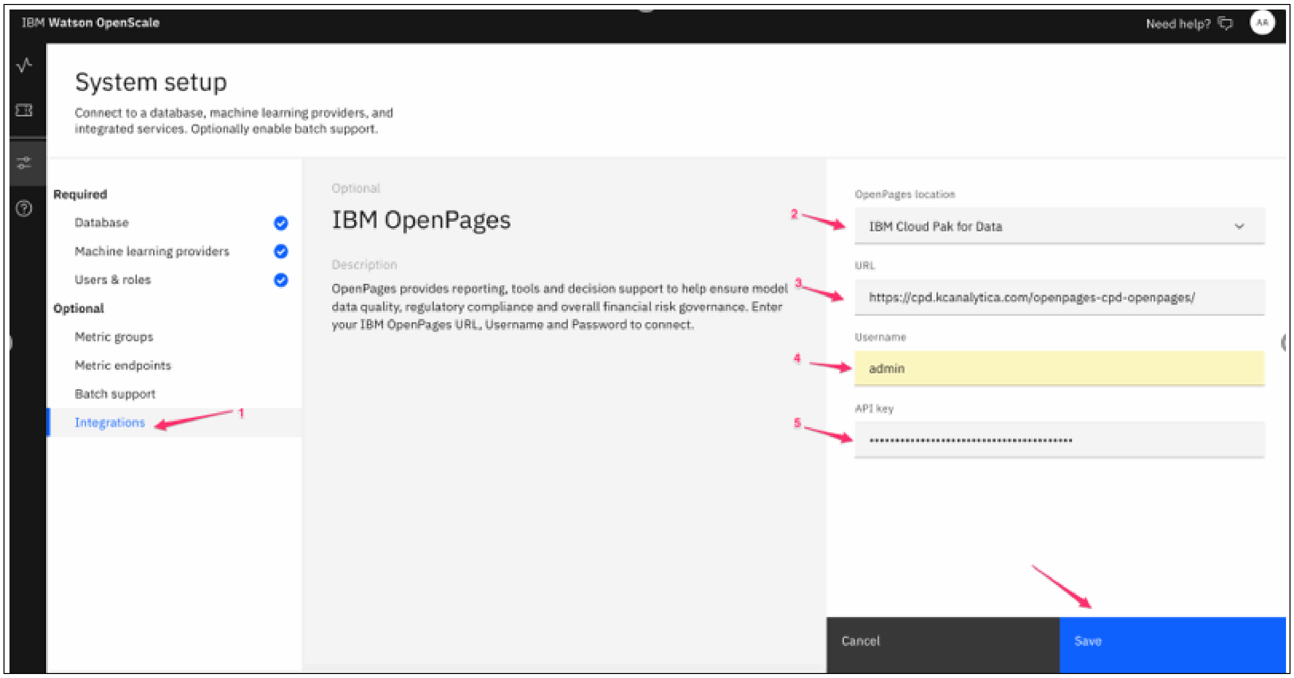

Figure 5-287 Integrate Watson OpenScale with IBM OpenPages

e. Open the AutoAI Churn Prediction model in IBM OpenScale to configure it with the IBM OpenPages model (see Figure 5-288).

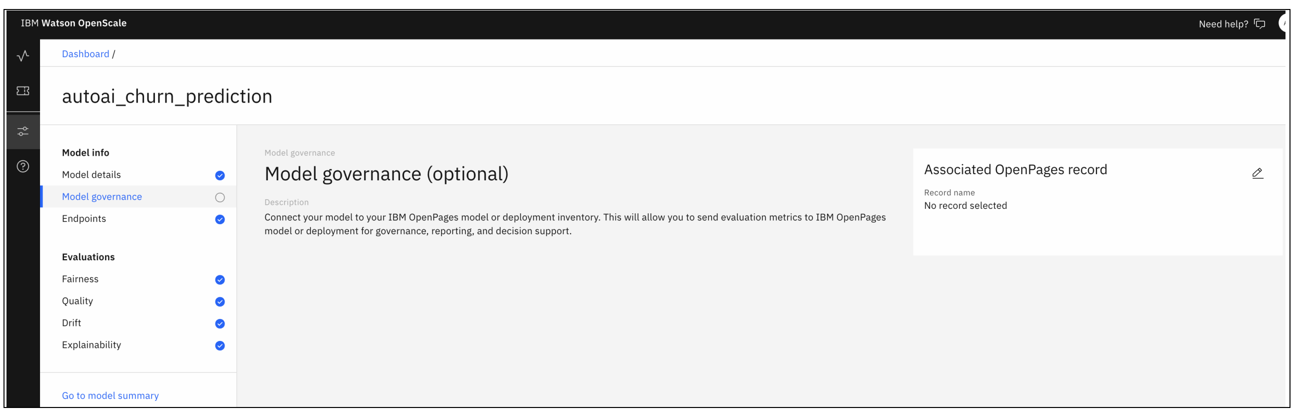

Figure 5-288 Integrate Watson OpenScale with IBM OpenPages

f. Search for the newly created IBM OpenPages model to associate with this churn prediction AI model and send the monitoring metrics to IBM OpenPages (see Figure 5-289).

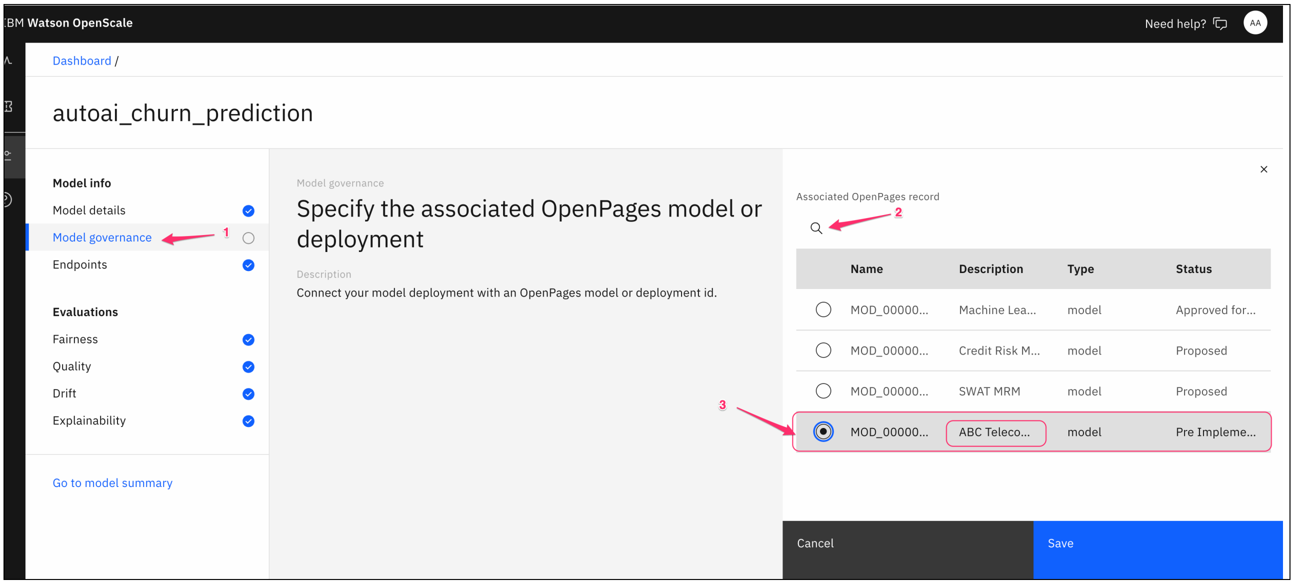

Figure 5-289 Integrate Watson OpenScale with IBM OpenPages

9. Complete the following steps to perform an analysis in IBM Watson OpenScale:

 a. Evaluate the AI model to generate new metrics to be published to IBM OpenPages.

 b. Click **Send to OpenPages**, which sends all of the metrics to the IBM OpenPages model record that you associated with the IBM OpenScale model deployment (see Figure 5-290).

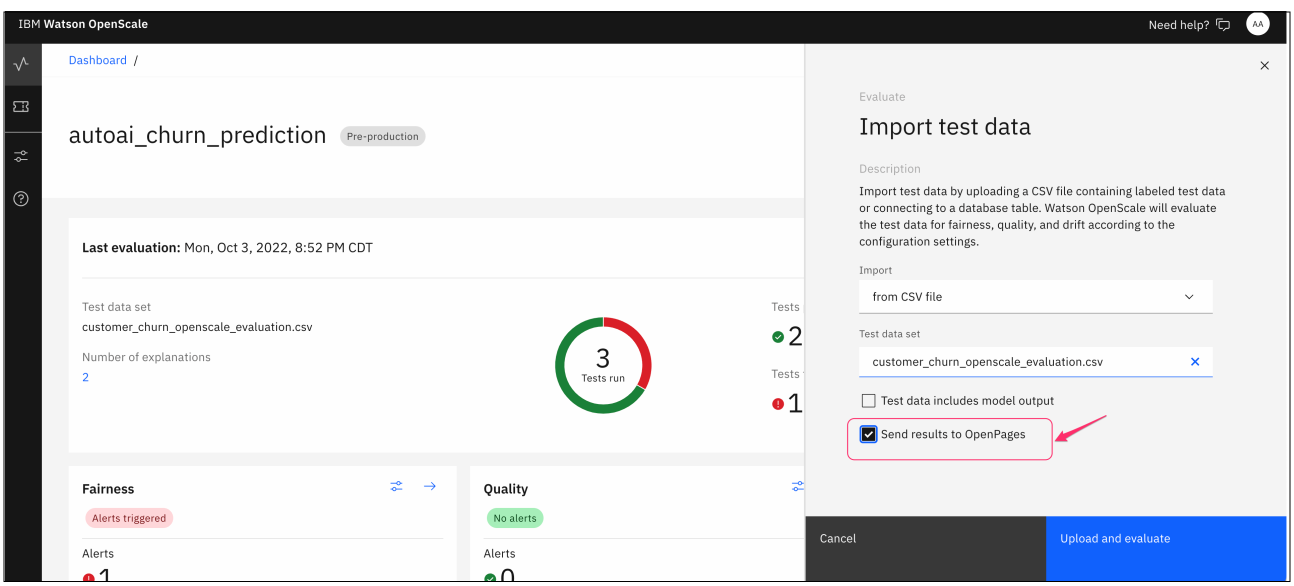

Figure 5-290 Evaluate AI model in Watson OpenScale and send to OpenPages

10. Review results in IBM OpenPages. After you send metrics to IBM OpenPages, these metrics are listed in the graphical flow diagram and can be selected so that you can drill down into each metric.

    Complete the following steps:

    a. Find the model in IBM OpenPages by using the newly created model that is named `MOD_0000023`.

    b. Review the metrics from IBM Watson OpenScale by expanding the metrics twisty.

    c. In addition to a list of metrics, you can view metrics in a graphical format by expanding the metrics summary twisty.

       The Association view provides relationships in the form of a tree.

       The Supporting files and artifacts panel provides access to all of the IBM Watson OpenScale model risk management reports that are run (see Figure 5-291).

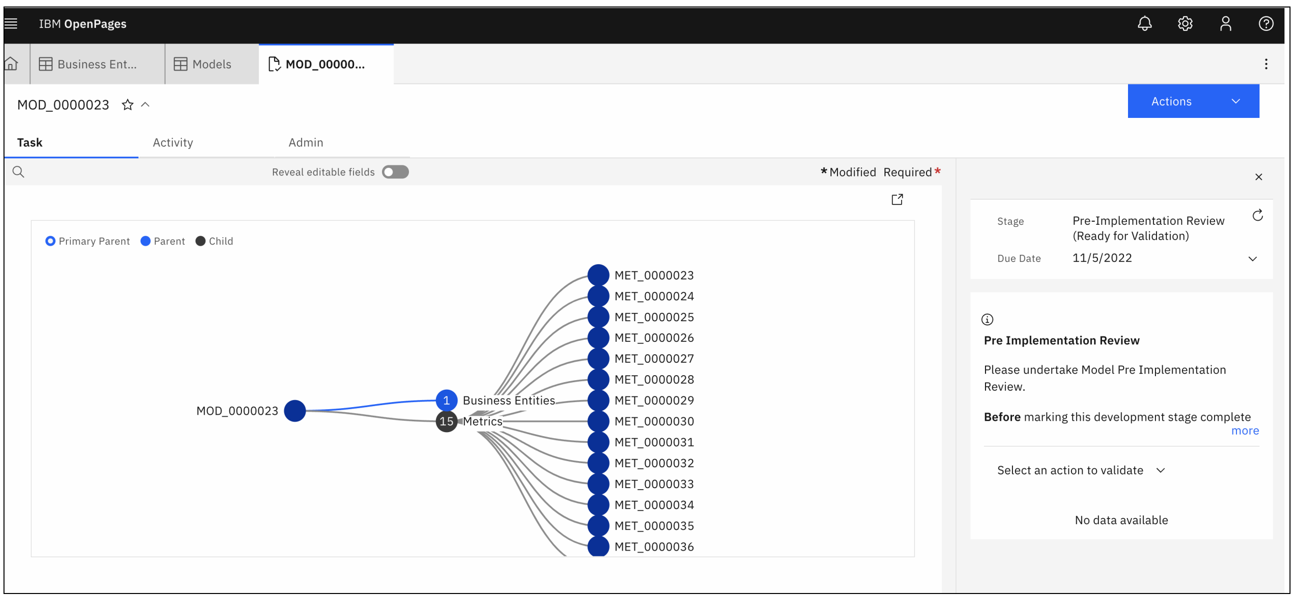

Figure 5-291 Reviewing the metrics from IBM Watson OpenScale in IBM OpenPages

d. Select one of the metrics to review the details, as shown in Figure 5-292.

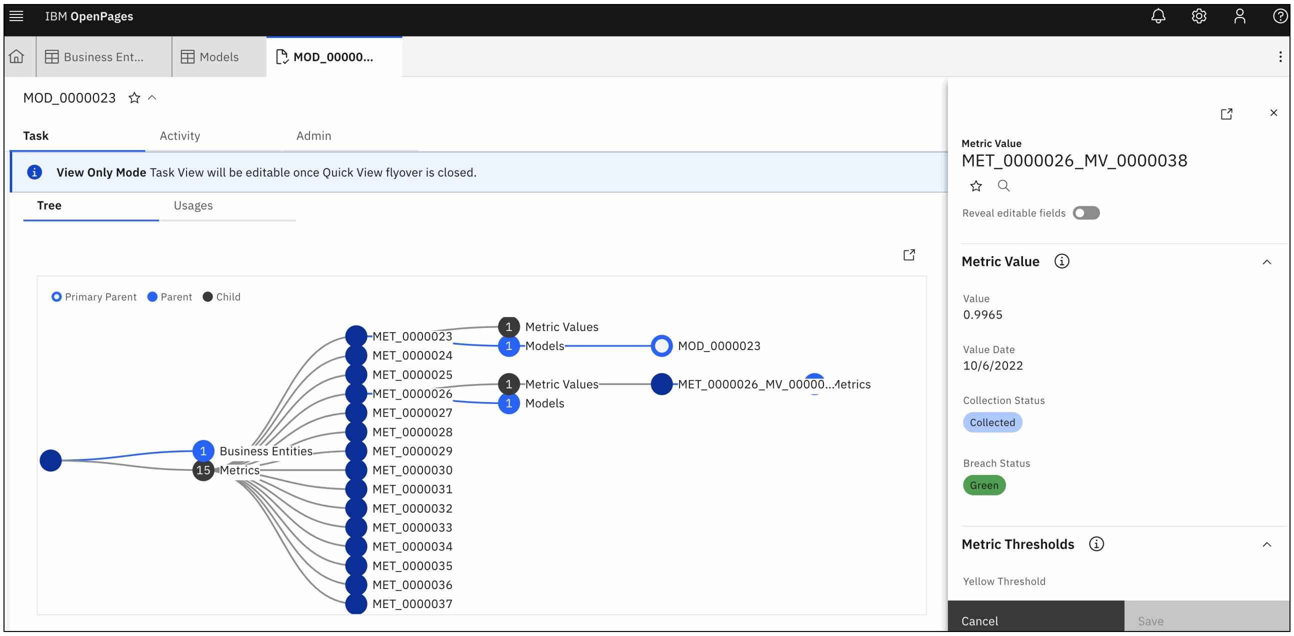

Figure 5-292 Explore the metrics from Watson OpenScale

11. Complete the following steps to approve the model in IBM OpenPages for production deployment:

   a. In IBM OpenPages, locate the `MOD_0000023` model to promote.

   b. From the Actions drop-down menu, click **Approved** (see Figure 5-293).

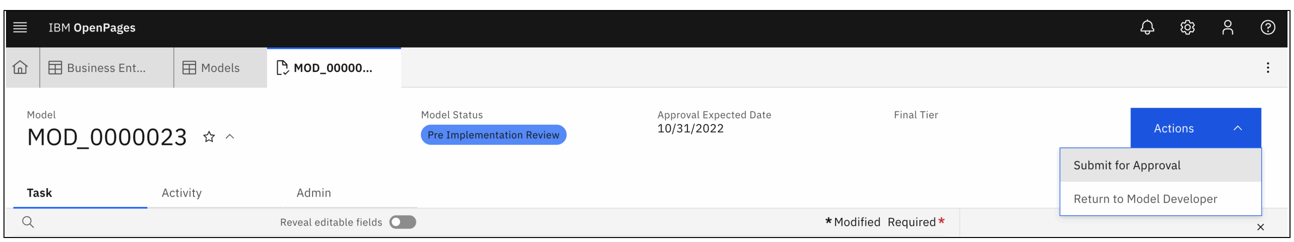

Figure 5-293 OpenPages model to submit for approval

   c. Review the status of the model in IBM OpenPages after the approval (see Figure 5-294).

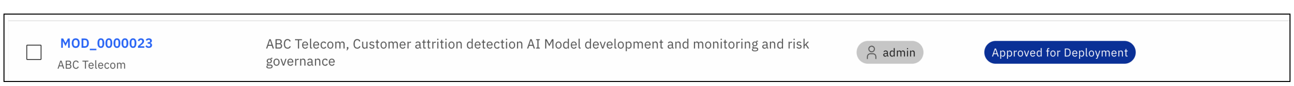

Figure 5-294 IBM OpenPages model approved for production deployment

12. Deploy the model to the production. After the model is approved for deployment in IBM OpenPages, you can send the model to production by using IBM Watson Studio and IBM Watson OpenScale. Then, you can associate the production model and metrics with the IBM OpenPages model.

Complete the following steps:

a. Create a Deployment space that is named `churn-production` if it was not yet created by using the Cloud Pak for Data web console.

b. Open the **Customer churn prediction** project and promote the model to the production deployment space (see Figure 5-295).

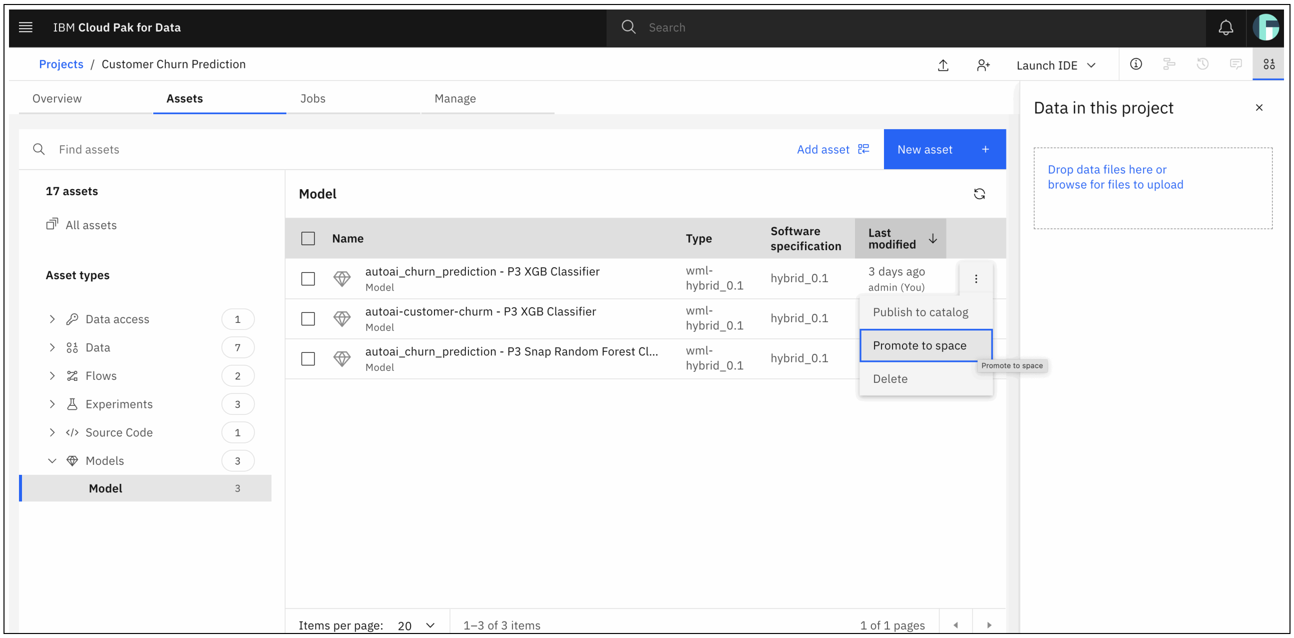

Figure 5-295 Promoting the model to production

c. Select the **churn-production** target space and click **Promote**, as shown in Figure 5-296.

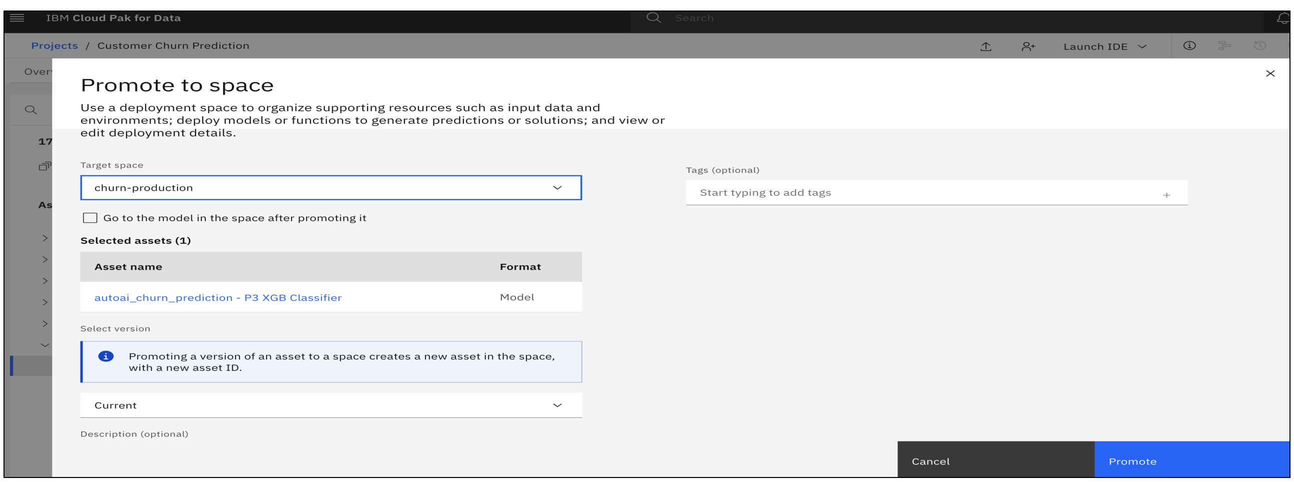

Figure 5-296 Promoting the model to production space

d. Select the **auto ai churn prediction** model and click **Deploy** (see Figure 5-297).

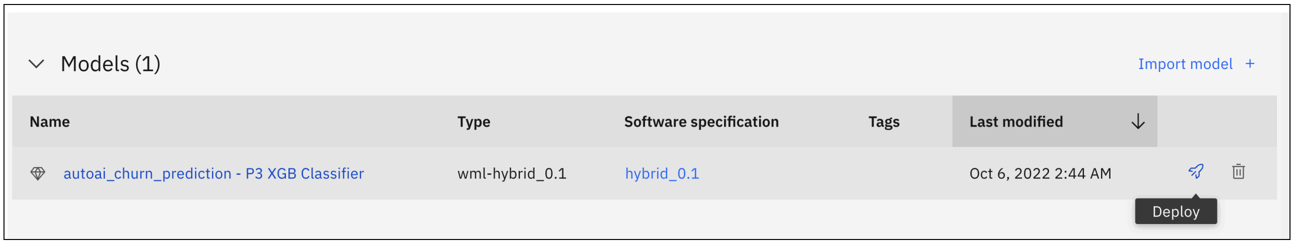

Figure 5-297 Deploy the model to production

e. Select the deployment type **Online** and save it as `customer_churn_production` (see Figure 5-298).

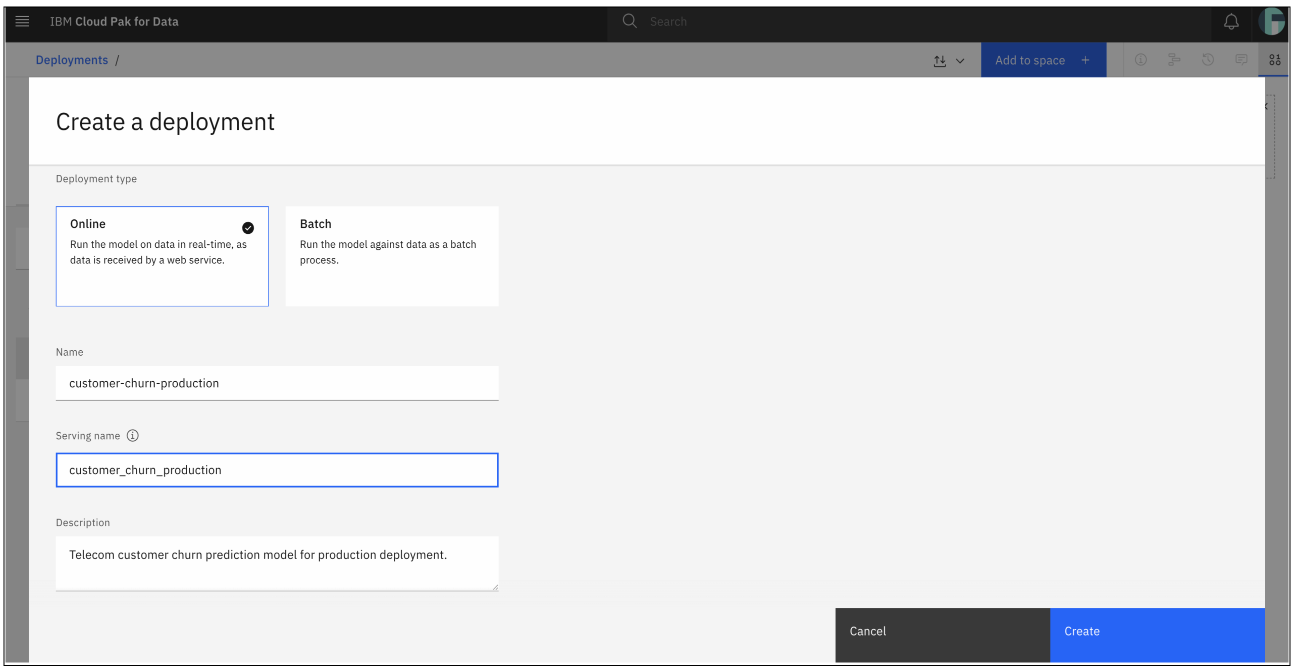

Figure 5-298 Creating deployment type "online" for this production deployment

f. Configure the IBM Watson OpenScale by using the `churn-production` deployment space (see Figure 5-299).

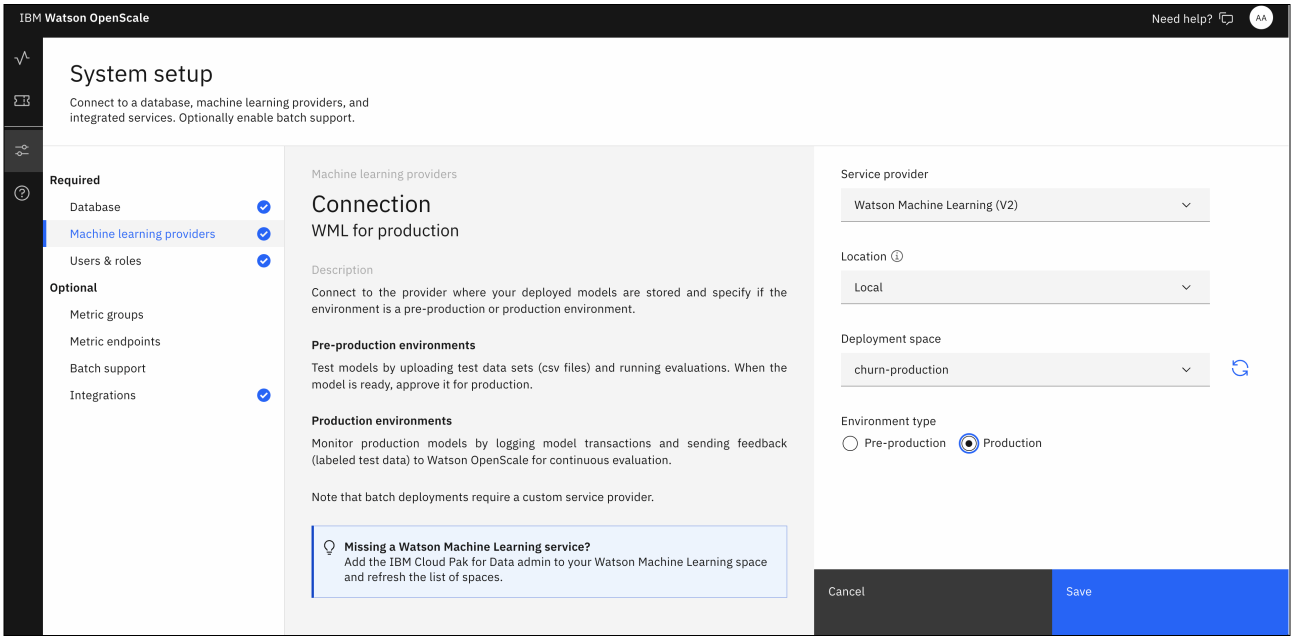

Figure 5-299 Configuring the Watson OpenScale for production deployment

g. Associate with IBM OpenPages model for production (see Figure 5-300).

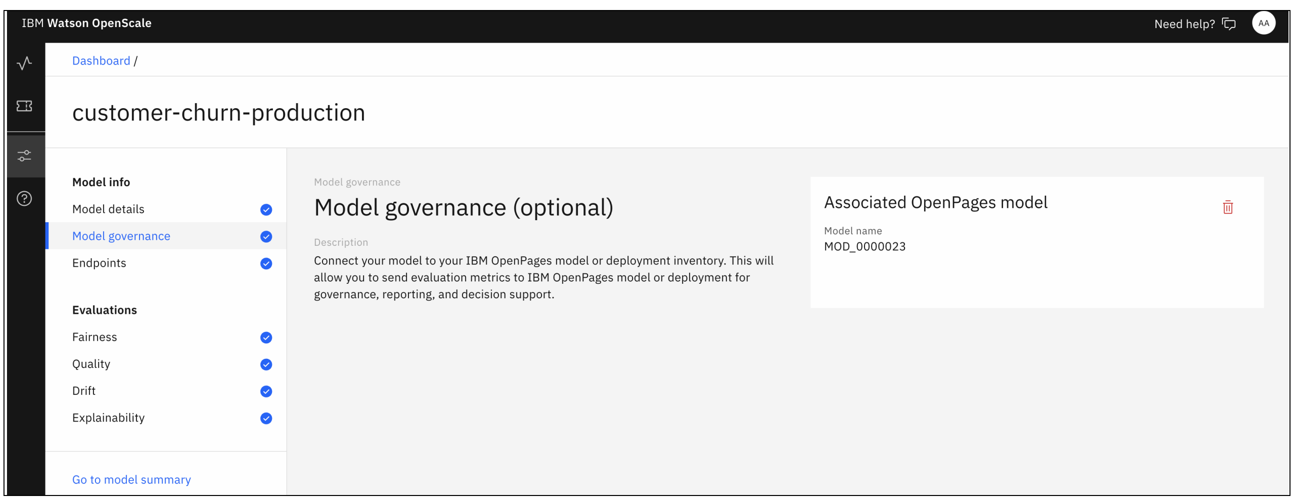

Figure 5-300 Associate the production OpenScale with the OpenPages model

h. In OpenScale, select **Action** → **Evaluate** (see Figure 5-301).

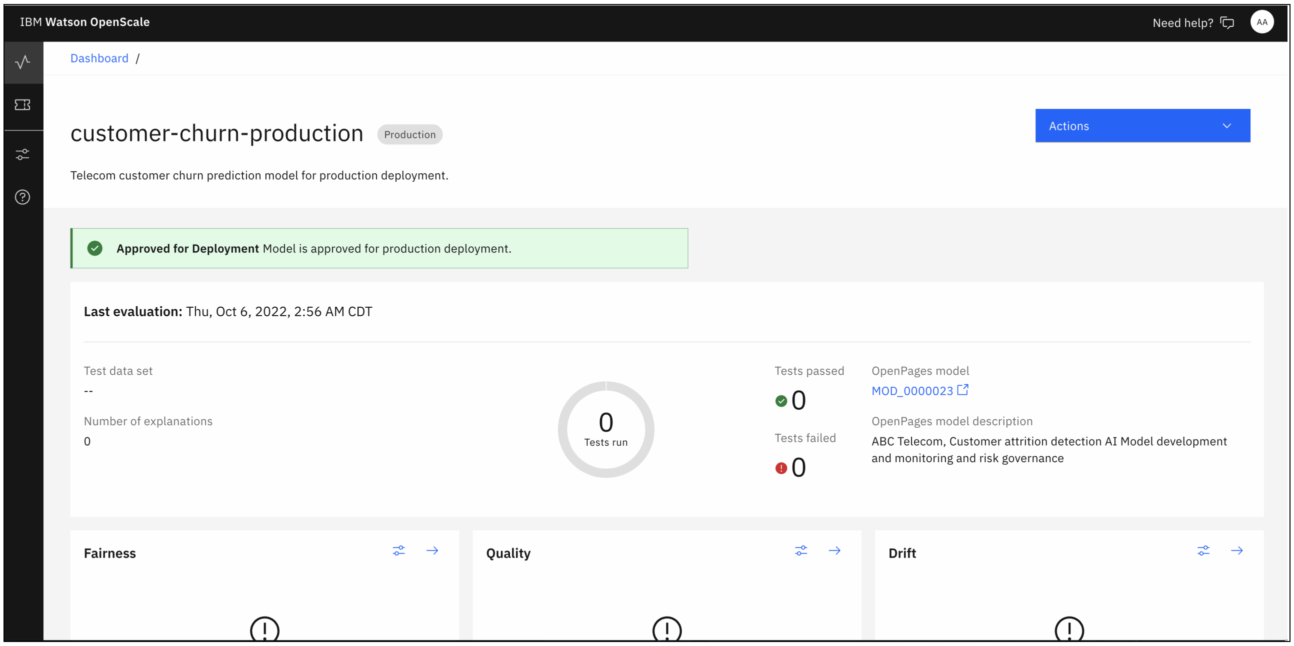

Figure 5-301 Evaluate the model with test data to start sending the metrics to OpenPages

i. Review the new metrics in OpenPages model `MOD_0000023` and under the task section (see Figure 5-302).

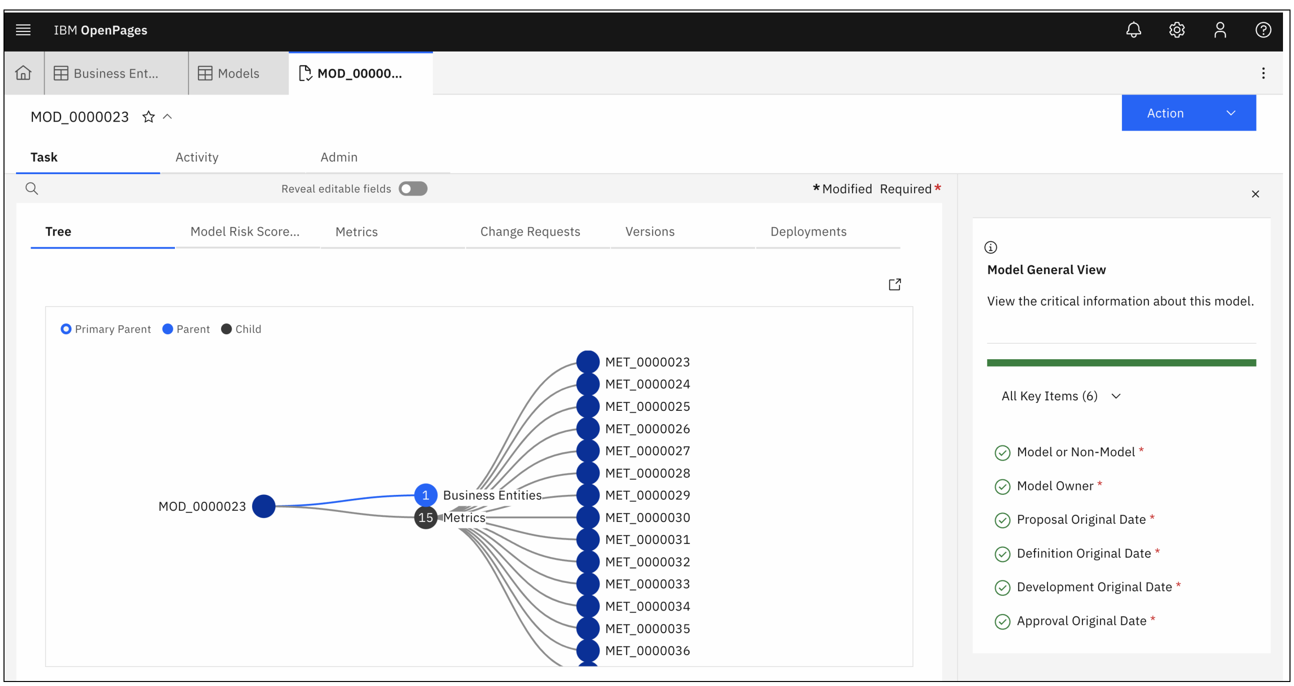

Figure 5-302 Review the metrics received in OpenPages

IBM Watson OpenScale can be configured to send alerts when fairness, quality, or drift threshold violation is raised and send an email notification, as shown in Figure 5-303.

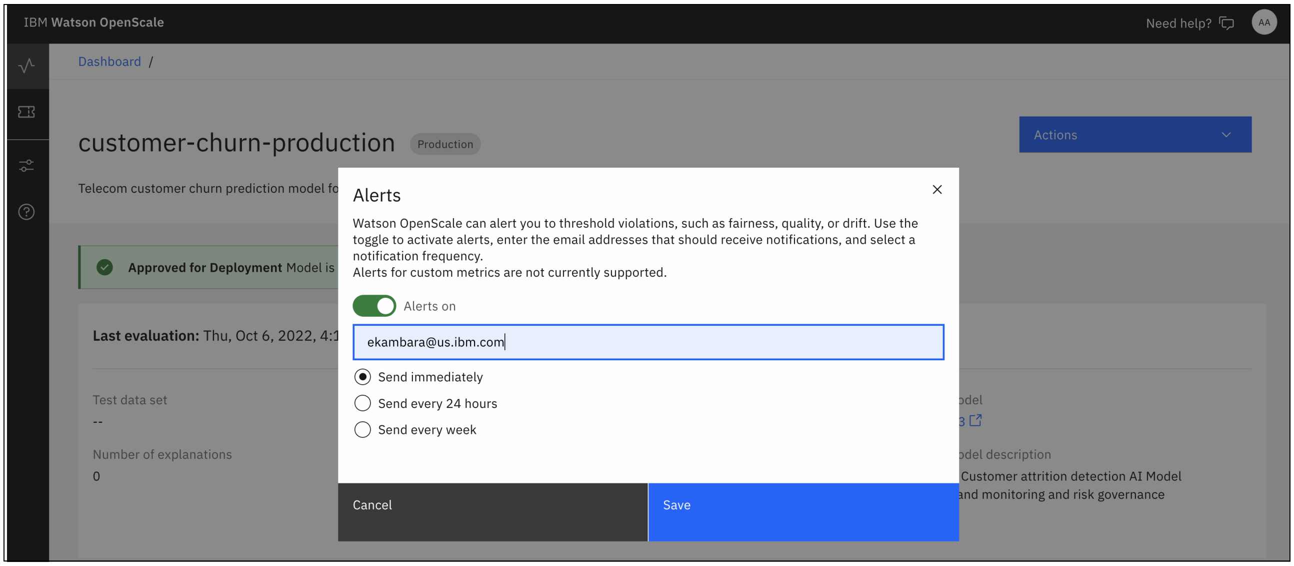

Figure 5-303 Configuring alert notifications for model threshold violation

This completes the Model Risk and Governance exercise in Cloud Pak for Data.

6

Customer care

This chapter provides an overview on the importance of the customer care solutions to be incorporated in the solutions that are developed by organizations. It also describes possible use cases for applying customer care within organizations for internal operations and external processes for clients.

The chapter provides hand-on examples and resources to create services instances and integrate Watson services with applications and includes the following topics:

6.1 Overview

Customer care can help customers accomplish what they need fast, which leaves them feeling valued, while helping them realize many benefits, such as higher customer satisfaction, reduced support costs, and the opportunity to propose offers in a nonobtrusive way.

This chapter explains what customer care is and demonstrates how the customer care portfolio at IBM can improve customer service interactions through customer care approaches that highlight key features.

6.1.1 Customer care

Customer care is a proactive approach to providing information, tools, and services for customers at each point that they interact with a brand. For a truly positive customer experience, customer care must be successfully intertwined with customer service.

Customer care and customer service together help create a positive customer experience. This experience enhances the overall impression that a person has when they interact with a company. For a good impression, building an emotional connection between the customers and the company that makes customers feel supported provides a proactive customer care experience.

Also, helping customers solve problems and answer questions before purchase in a self-service fashion, or by using the customer support team, offers reactive customer care experience.

Figure 6-1 shows that a positive customer experience both customer care *and* customer service.

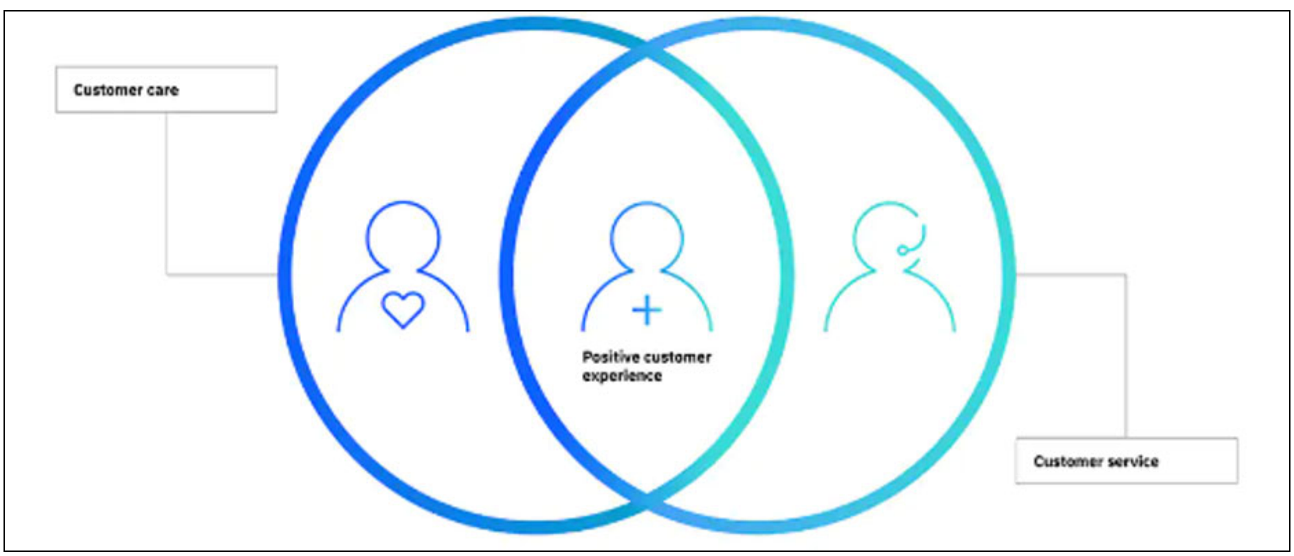

Figure 6-1 Positive customer care experience

However, finding the best way to offer a positive customer experience can be difficult. So why are positive customer service experiences so difficult to obtain?

Figure 6-2 shows three examples of what customers typically experience.

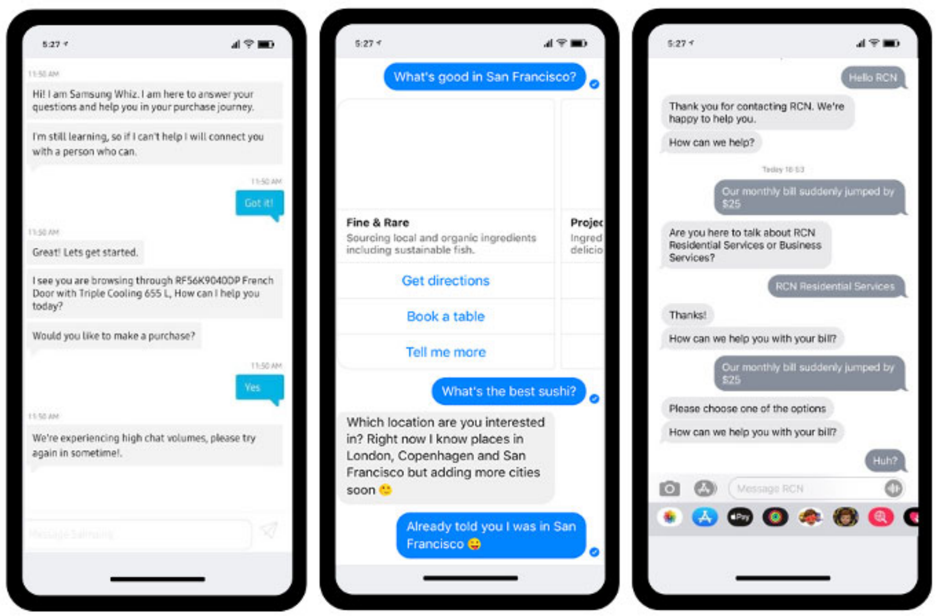

Figure 6-2 Typical customer experiences

Unfortunately, most customer experience feels more like the three conversations that are shown in Figure 6-2. These examples show how to sow customer dissatisfaction and disloyalty, while missing an opportunity to build and reinforce the relationship.

Example 6-1 shows the conversation 1 from Figure 6-2.

Example 6-1 Conversation 1

```
Bot: Hi! I am Samsung Whiz. I am here to answer your questions and help you in
your purchase journey. I'm still learning, so if I can't help I will connect you
with a person who can

Customer: Got it!

Bot: Great! Lets get started. I see you are browsing through RF56K9040DP French
Door with Triple Cooling 655 L, How can I help you today? Would you like to make a
purchase?

Customer: Yes

Bot: We're experiencing high chat volumes, please try again sometime
```

As you can see in Example 6-1, the customer starts to feel like they are getting somewhere and then, the bot suddenly cuts them off with a response that they are no longer able to help because of high chat volumes.

Example 6-2 shows the conversation 2 from Figure 6-2 on page 441.

Example 6-2 Conversation 2

```
Customer: What's good in San Francisco?

Bot: <lists some popular local businesses>

Customer: What’s the best sushi?

Bot: Which location are you interested in? Right now I know places in London,
Copenhagen and San Francisco but adding more cities?

Customer: Already told you I was in San Francisco :p
```

Example 6-2 shows a classic example of the bot not being able to take context clues from previous customer responses and apply it to a new response. This simple task is one that a human agent can do; therefore, the bot ideally should be able to comprehend this issue as well.

Example 6-3 shows the conversation 3 from Figure 6-2 on page 441.

Example 6-3 Conversation 3

```
Customer: Hello RCN

Bot: Thank you for contacting RCN. We’re happy to help you. How can we help?

Customer: Our monthly bill suddenly jumped by $25

Bot: Are you here to talk about RCN Residential Services or Business Services?

Customer: RCN Residential Services

Bot: Thanks! How can we help you with your bill?

Customer: Our monthly bill suddenly jumped by $25

Bot: Please choose one of the options. How can we help you with your bill?

Customer: Huh?
```

Example 6-3 reflects a good example of a bot that is confused about the topic of conversation with the customer. This confusion is frustrating and often makes the customer feel as though they are wasting their time.

Poor customer experience is clearly reflected in these three conversations examples. The reason companies fall short of delivering exceptional customer experience is because of the scale and complexity of customer behavior.

Figure 6-3 shows scale and complexity statistics of consumer behavior.

Figure 6-3 Scale and complexity statistics of consumer behavior

For example, a Tidio[1] study of over 200 research-based customer service statistics resulted in the following findings:

- A total of 74% of customers report the use of multiple channels to engage and complete a transaction. How consistent is their customer experiences across, for example, the company's call center and their digital chat?
- If more than half (65%) of customers prefer self-service support for simple matters, how easy is it for them to diagnose and solve common problems on their own? How difficult is it then to get help if they need it?

According to an article posted on Harvard Business Review (HBR), a 38% increase in customer hold times occurred. Even more concerning, a 68% increase in escalations to get a problem solved[2] was observed.

Figure 6-4 shows a typical customer care solution.

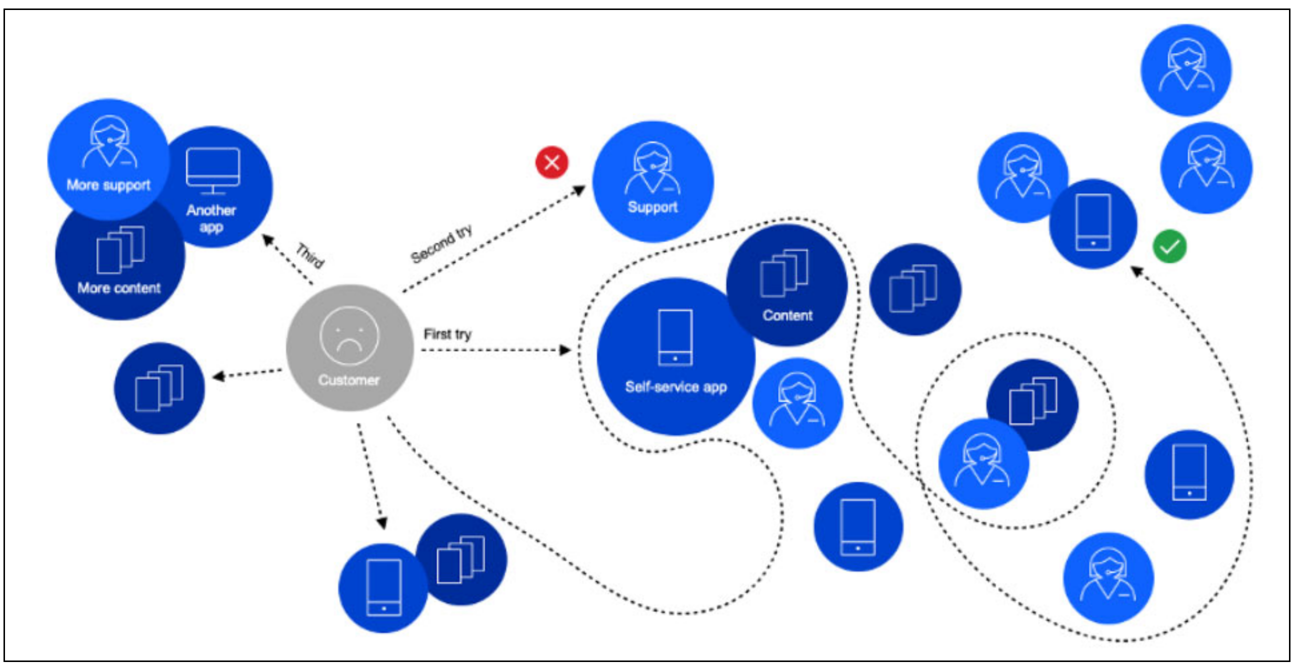

Figure 6-4 Typical customer care solutions

[1] *200+ Research-Based Customer Service Statistics (2022)*: https://www.tidio.com/blog/customer-service-statistics

[2] *Supporting Customer Service Through the Coronavirus Crisis*: https://hbr.org/2020/04/supporting-customer-service-through-the-coronavirus-crisis

For many customers, their experience looks resembles the example that is shown in Figure 6-4. This example shows a frustrating combination of multiple applications, support desks, and attempts to get help, or to even make a purchase. Not only is this situation frustrating for customers, it is inefficient and costly for the enterprise.

Figure 6-5 shows how customer care can be organized for an improved customer experience.

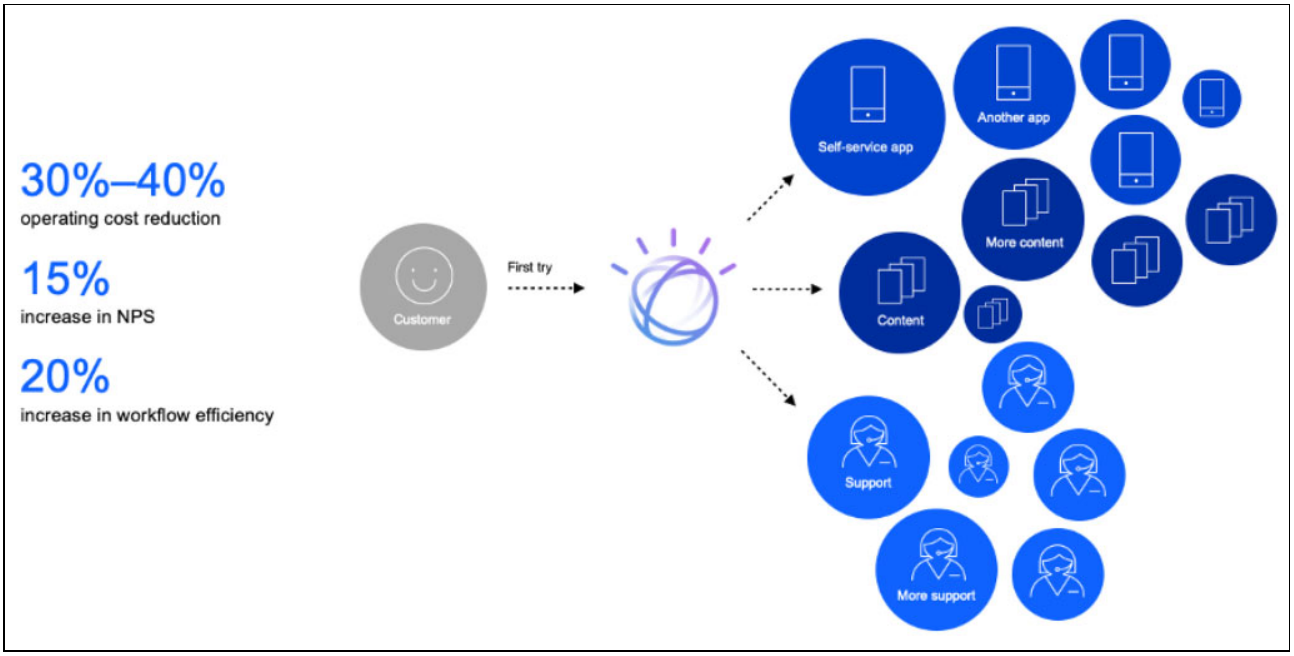

Figure 6-5 Organized and improved customer experience

Figure 6-5 shows what most enterprises want; that is, a first effort resolution that results in higher customer satisfaction and workflow and worker efficiency, and significantly lower operating costs.

Through thousands of customers engagements that were analyzed, IBM found that when clients implement artificial intelligence (AI) for customer care solutions, they realize an increase of 20% in workflow efficiency, an increase of 15 in Net Promoter Scores (NPS), while also benefiting from a 30 - 40% operating cost reduction in their contact centers.

However, if customers want to realize the benefits of improved customer satisfaction, efficiency, and lower costs, why is not every company using customer care solutions?

Figure 6-6 shows the three main virtual assistant solutions that make customer care difficult.

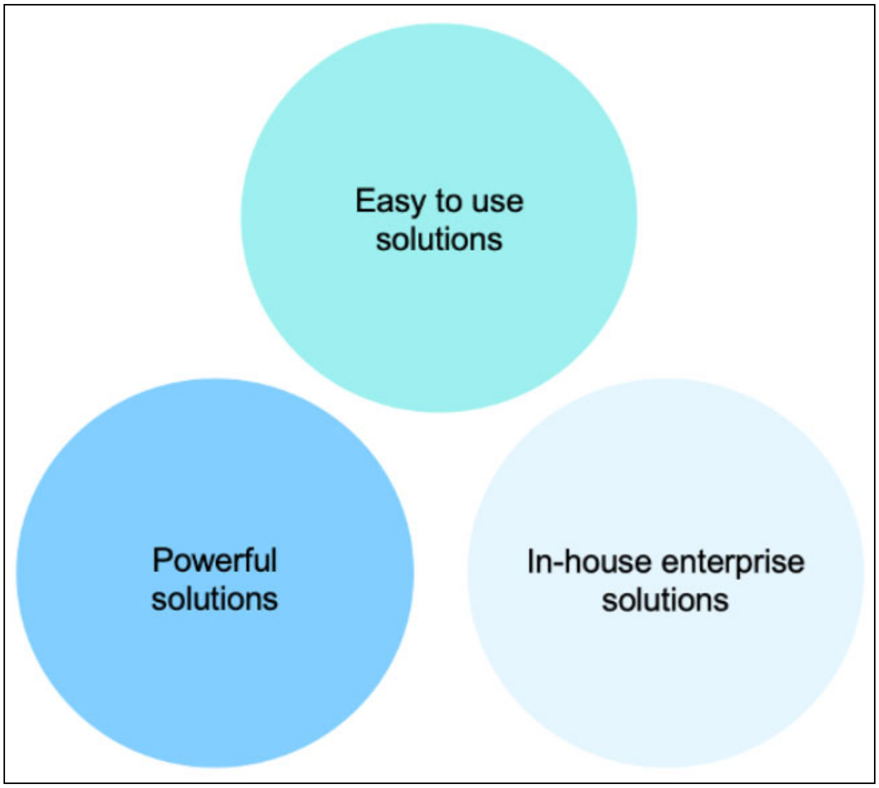

Figure 6-6 Problematic customer care virtual assistant solutions

It is not that organizations are not actively pursuing conversational AI technologies to improve their customer experiences. Historically, the issue was the tools and strategies that they choose in which to invest. Why is that? The reason is because of the following main types of issues with the available products:

- Are the products that claim to be easy to use actually easy to use or include support software? Even if they are easy to use, these products often do not scale beyond the initial use cases. They also are not powerful or flexible enough.

  The AI behind them also is not powerful, and the system is not dynamic enough to span and scale the use cases across the enterprise.

- Are the products that are powerful enough produce frictionless experiences? Even if they are powerful enough, building them often requires deep technical understanding, which often leads to expensive and time-consuming projects.

  These types of products are all targeted at AI developers, which means that they are not sustainable. When agents in the contact center encounter a new question from a customer, or the support solution for something changes, they are beholden to developers to make even the simplest of changes. This paradigm is not a realistic, economically feasible way to provide support when all refactoring must go through a development lifecycle.

- Many large enterprises are on a journey of building customer care AI solutions themselves, from the ground up. They embark on this journey because the “easy to use” solutions did not measure up to their needs, and the “powerful” solutions seemed too broad for a narrow use case.

The result often is an overly complicated solution that is designed for a specific use case. Eventually, enterprises become overwhelmed by the need for integration with other layers of customer service technology, which requires deep customization and technical development. This issue makes the in-house solutions too expensive for smaller deployments, and difficult or impossible to adapt or scale for more use cases across the enterprise.

Figure 6-7 shows groups of successful customer care solutions.

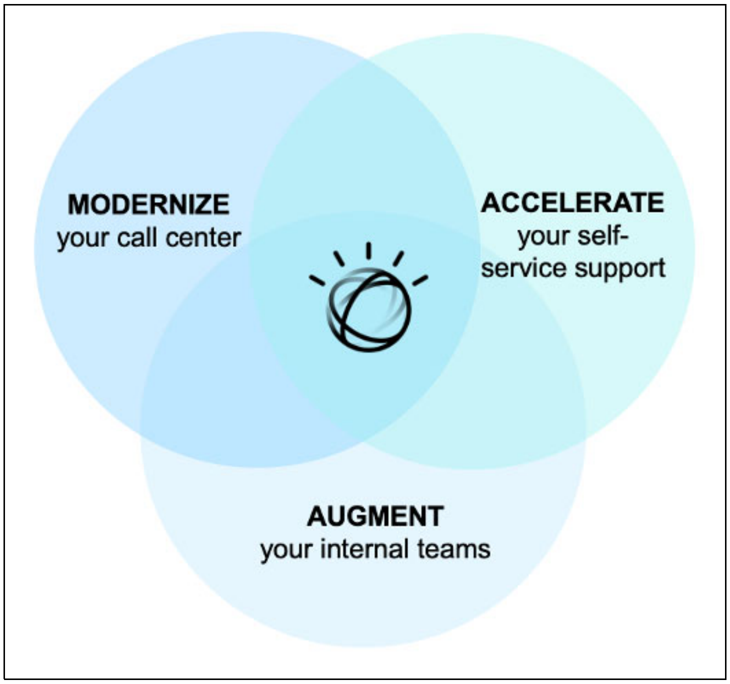

Figure 6-7 Groups of successful customer care solutions

Success requires a combination of the benefits of the three groups of customer care solutions, without any of the drawbacks. First, companies need a conversational AI platform that can deliver the robust, frictionless experiences customers demand, but that is designed for business users.

Second, customers need a solution that melds the powerful capabilities that are wielded by AI developers, yet one that is delivered through an intuitive and easy to learn interface.

Above all, customers also need a solution that can scale with their business needs that includes the flexibly to adapt to new use cases while integrating with the customer care technology stack.

In this way, customers can achieve the following goals:

- Accelerate their self-service support
- Modernize their call center
- Augment their internal teams

IBM has available a customer care portfolio that is differentiated in its ability to provide exceptional customer experiences by offering the following benefits:

- Using the latest advancements in AI to help a brand or organization's customers complete a task
- Unifying and personalizing the experience to drive positive user outcomes in a single interaction.

Applying AI to customer care is what allows companies to accelerate, modernize, and augment for a truly positive customer experience. The customer's experience is everything because it is how the most successful companies remain so successful.

Companies that want to deploy AI for customer care can realize higher customer satisfaction workflow and worker efficiency, and significantly lower operation costs.

Everything that companies do in their business, whether it is reducing risk, automating processes, securing top talent, or anything else, is ultimately for the benefit of their customers. Companies do this with the intent to better serve their customers, or to attract new customers.

Every business is centered around a customer, which means that the customer's experience is everything.

6.1.2 Importance of customer care

Customer care is involved in almost every industry. How a company cares for their customers is one of (if not the most) important aspects of their business.

The way customers engage with companies was completely revolutionized over the past few years. Customers are no longer only calling in over the phone or logging in to a web chat.

Today, customers are expecting to reach out to companies by way of social media, messaging applications, SMS, and intuitive mobile applications. In addition to this channel complexity, customers still expect a seamless and frictionless experience, regardless of their intent.

Customers might want to buy a product, run a transaction, or get support for a product or service and want immediate answers with a quick resolution to their issues in a single interaction. In these circumstances, when customers expect to be delighted, customer satisfaction is more important than ever.

Figure 6-8 shows some overall customer experience statistics[3].

Figure 6-8 Overall customer experience statistics

Also, smart customer care is not only about “delighting customers”. Rather, true loyalty is first driven by how well a company eliminates friction for their customers, and how well it delivers on its basic promises to them.

Customer care also is about how the company solves day-to-day problems for their customers. For example, according to a research study conducted by Financial, excessive customer effort is a key driver of dissatisfaction and disloyalty: a customer’s experience in dealing with a company outweighs brand, product, and service quality to the extent that 91% of unsatisfied customers part ways with a brand because of a poor experience or lack of support.

In addition, according to a paper published by Glance, 78% of customers say they will back out of a purchase, or not consider a purchase at all, because of a poor customer experience or the perception of a poor experience. That is, most customers do not settle for a mediocre experience. Instead, they want an exceptional experience. For this reason, many companies are making significant investments to satisfy their customers[4].

Industry market leaders clearly showed everyone else that customer experience is everything and the rest of the market is now working hard to catch up. To catch up, companies must truly know their customers. However, this process can be difficult when their customer data and support is siloed across different digital phone channels.

Some companies attempted to implement their own chatbots to help with their customer service. However, according to Gartner, most technologies are not flexible or scalable to evolve with the business, and it is likely that 90% of chatbots that are deployed today will be discarded by the end of 2023. As such, a great opportunity is available to use and optimize IBM Customer Care solutions.[5].

[3] Sources: 1) https://www.ibm.com/downloads/cas/QWRMQRGL
2) http://ww2.glance.net/wp-content/uploads/2015/07/Counting-the-customer_-Glance_eBook-4.pdf
3) https://www.1stfinancialtraining.com/Newsletters/trainerstoolkit1Q2009.pdf

[4] http://ww2.glance.net/wp-content/uploads/2015/07/Counting-the-customer_-Glance_eBook-4.pdf

IBM makes available a set of customer care solutions that can scale, have a lower cost of ownership, and harness the entire stack of customer data to provide a new level of the effortless experience that customers have now grown to expect. We can use IBM technology, such as IBM Watson Assistant, IBM Watson Discovery, and IBM Watson Speech to understand and infuse the knowledge into each customer interaction to predict and resolve the changing customer needs.

Customer care is now more important than ever, and IBM provides AI solutions that can be easily harnessed.

6.1.3 Approaches to apply customer care

Many approaches are available to apply customer care. IBM features an AI for Customer Care portfolio that uses the latest advancements in AI to help a brand or organization's users complete a task on their first attempt. With Cloud Pak for Data add-on services, such as IBM Watson Assistant, IBM Watson Discovery, and IBM Watson Speech services, it is uniquely designed to achieve the following results:

- Unify the customer experiences across all channels and departments
- Create a personalized and delightful experience
- Drive the outcomes customers seek, including first-contact resolution

Figure 6-9 shows what the IBM AI for Customer Care portfolio is designed to do.

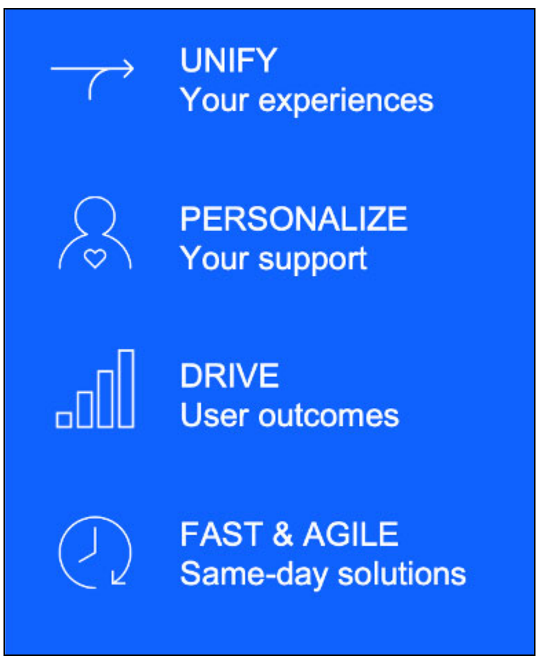

Figure 6-9 IBM AI for Customer Care portfolio intentions

The IBM Watson offerings that are a part of the IBM AI for Customer Care portfolio can be deployed quickly and easily on Cloud Pak for Data to provide for the maximum flexibility of deployment options.

[5] https://build-ecosystem.com/event/on-demand-building-best-in-class-conversational-ai-apps-that-delight-customers/

IBM Watson Assistant is at the core of the IBM AI for Customer Care solution and is a breakthrough in optimizing customer care. This AI-powered virtual agent delivers everything that an organization needs to build, train, integrate, and support a virtual agent in digital formats (web, tablet, or smartphone) or voice channels.

It also provides customers with fast, consistent, and accurate answers across any messaging platform, application, device, or channel. Its natural learning processes analyze customer conversations, which improves its ability to resolve issues the first time.

Figure 6-10 shows the key functions of IBM Watson Assistant.

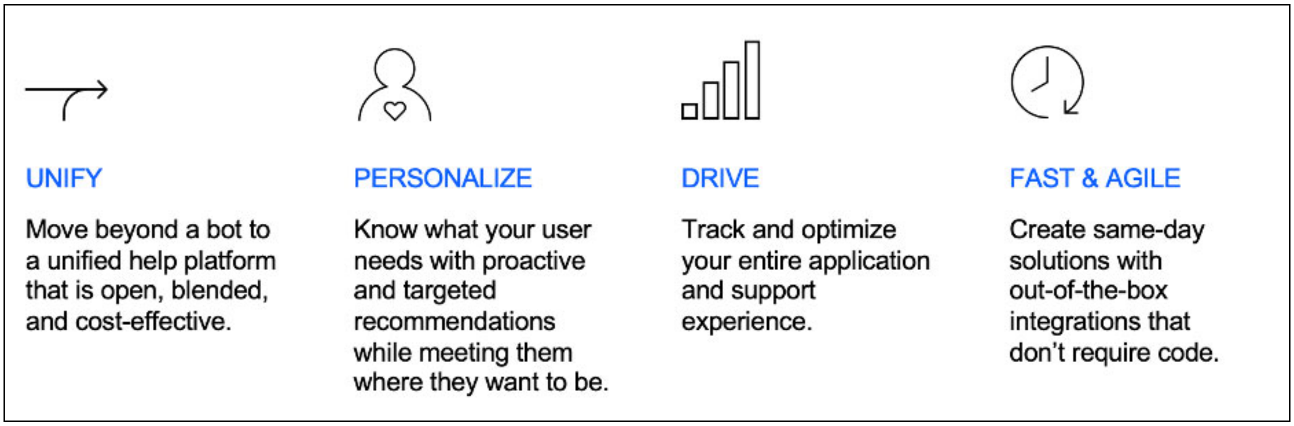

Figure 6-10 Key functions of IBM Watson Assistant

IBM Watson Assistant enables companies to achieve the following goals:

- Move beyond a bot to a unified help platform that is open, blended, and cost-effective.
- Know what their customers need, with proactive and targeted recommendations, while meeting them where the customers are: in the application or channel of their choice.
- Track and optimize their entire application and support experience.
- Create same-day solutions with immediately available integrations that do not require code.

An AI technology-first approach to customer care improves containment and customer satisfaction. IBM Watson Assistant takes over the basic and repetitive tasks that customer support agents perform, which frees them up to ficus more meaningful tasks, such as closing the customer feedback loop.

Companies can realize a two-fold benefit:

- More efficiently use the workforce.
- Improve customer experience metrics, such as IBM Net Promoter Score (NPS®) and customer satisfaction (CSAT).

Businesses saw significant customer care improvement, delivering a 383% ROI, according to a83 Forrester TEI report.[6]

[6] https://www.ibm.com/blogs/watson/2021/03/forrester-study-ibm-watson-discovery-6m-benefits/

IBM Watson Discovery

Building a comprehensive knowledge base is another means for providing a higher level of customer care. It makes call center response times faster and empowers customers to find the answers on their own.

IBM Watson Discovery is an award-winning, AI-powered intelligent search and text-analytics platform that uses the data that the company uses to improve customer experience efficiently. With innovative, market-leading natural language processing, it can quickly uncover meaningful business insights from documents, webpages, and big data, which reduces research time by 50%.

Figure 6-11 shows the key functions of IBM Watson Discovery.

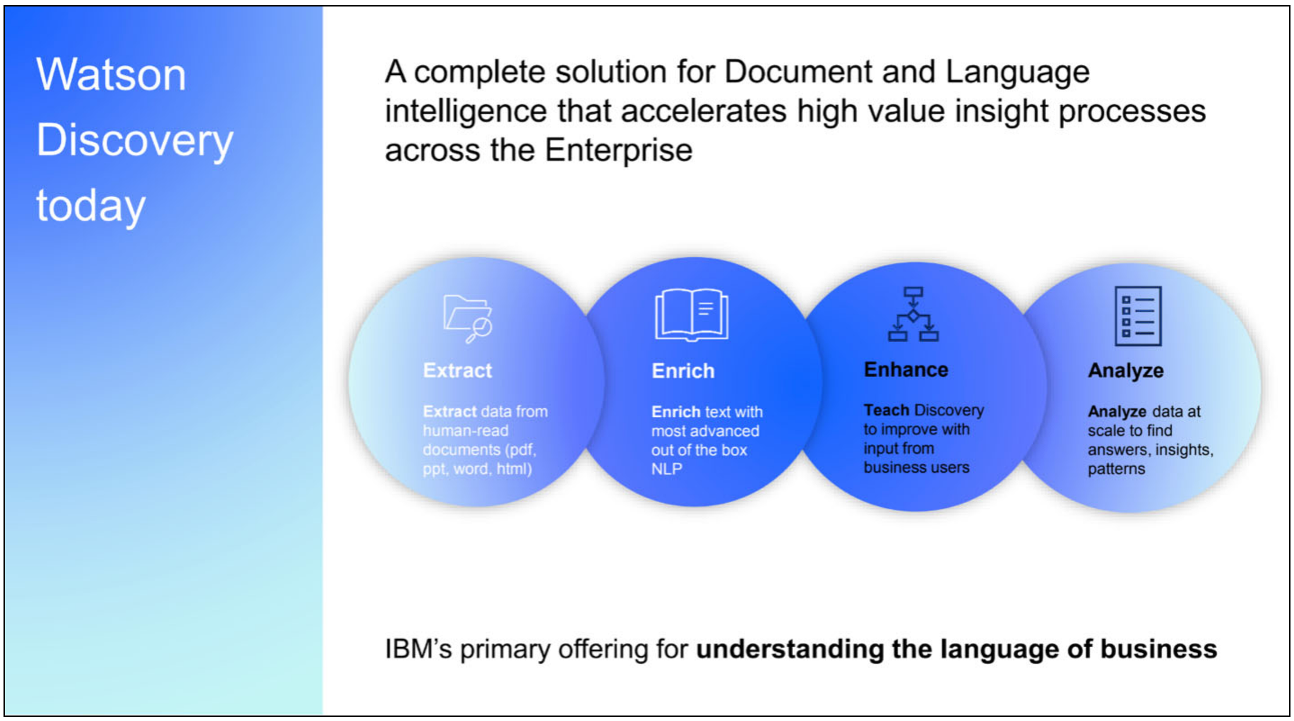

Figure 6-11 Key functions of IBM Watson Discovery

IBM Watson Discovery empowers an organization's experts and knowledge workers with the right information at the right time, and in the right context.

IBM Watson Discovery accelerates high value processes across the enterprise and allows companies to understand the language of their business to better serve their customers. It provides this acceleration through the four-step process that is shown in Figure 6-11: Extract, Enrich, Enhance, and Analyze.

IBM Watson Discovery enables companies to unlock hidden value in unstructured data to find answers, monitor trends, and surface patters. This feature augments an expert's ability to understand complex, high-value documents by providing immediate access to concise, trusted, and personalized information.

When customers interact with an enterprise to get something done, they want the experience to be fast and pain-free. But, because of the size and complexity of an enterprise, the customer service experience that is delivered is often disjointed and frustrating.

This issue occurs because often the business is unorganized and hard-to-reach informational content is connected to their customer service experience. IBM Watson Discovery solves this issue and enhance the customers services solution for a positive interaction.

In addition to the IBM conversational AI and content intelligence IBM Watson offers to improve customer care experiences, IBM also offers IBM Watson Speech services. IBM Watson Speech includes IBM Watson Speech to Text and IBM Watson Text to Speech.

IBM Watson Speech to Text enables fast and accurate speech transcription in multiple languages for various use cases, including customer self-service, agent assistance, and speech analytics. IBM Watson Text to Speech enables the conversation of written text into natural-sounding voices within any application.

Figure 6-12 shows how IBM Watson Speech to Text works.

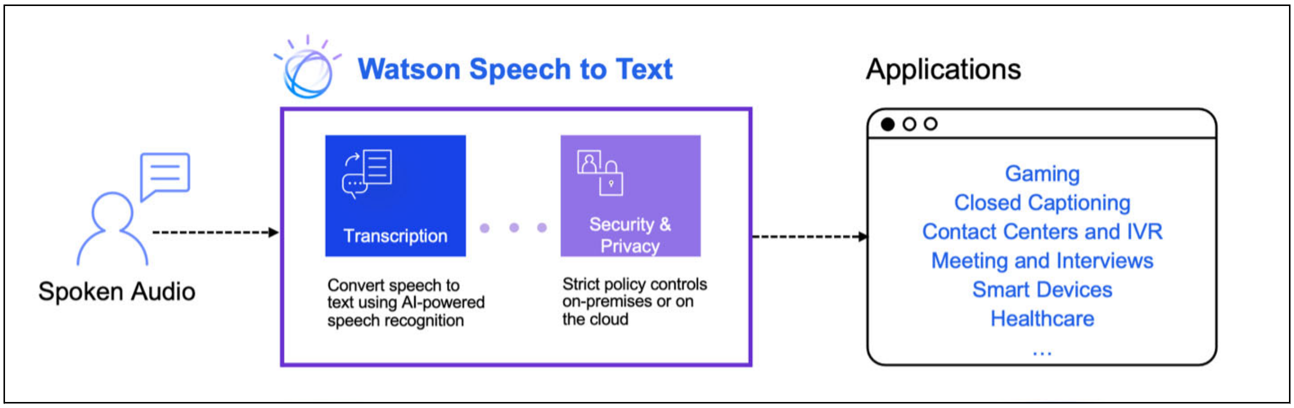

Figure 6-12 How IBM Watson Speech to Text works

Figure 6-13 shows how IBM Watson Text to Speech works.

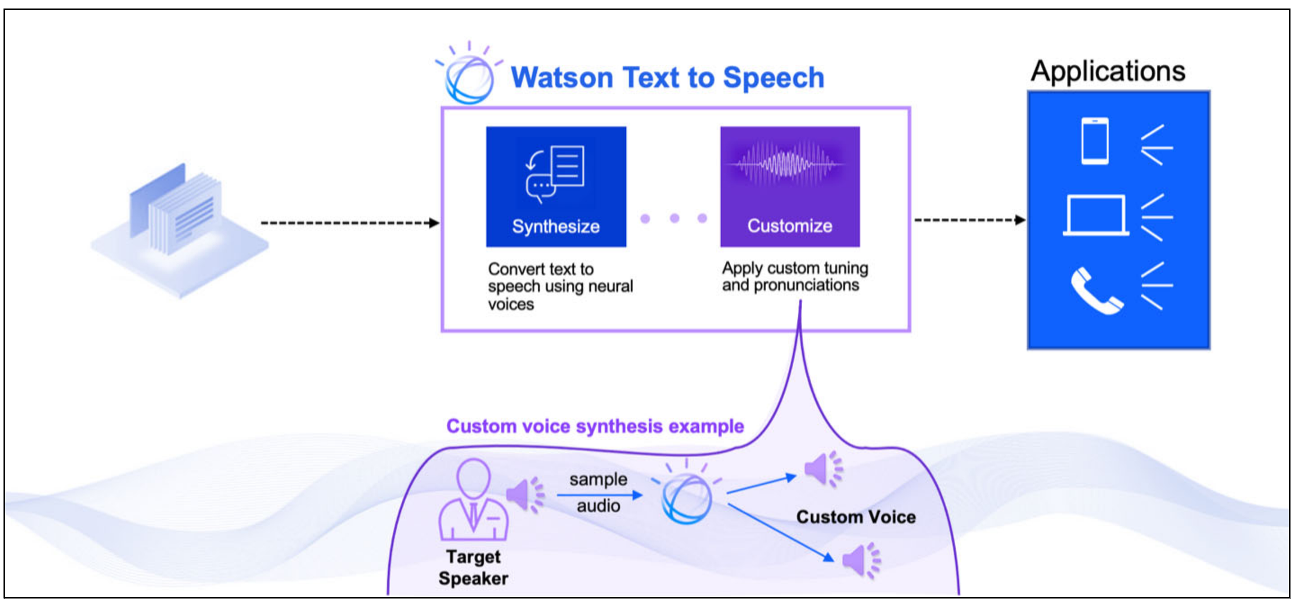

Figure 6-13 How IBM Watson Text to Speech works

The Watson Speech services help make the customer service experience feel more human-like and ultimately provide for a more satisfied customer.

Ultimately, all of these AI and ML-powered solutions not only answer customer questions, but do so in a more human way, which creates the emotional connections that are so important to customer care and customer loyalty.

Conversational AI with IBM Watson Assistant easily integrates with the IBM Watson Discovery for content intelligence and the IBM Watson Speech to Text and IBM Watson Text to Speech services to provide a friendlier experience for companies' customers.

In the next few sections, we consider some applications of the IBM AI for Customer Care portfolio with several Cloud Pak for Data services.

6.2 Use case description

Customer care is becoming a differentiator across all industries. For example, the energy and utility industry is being challenged to personalize their customer interactions while controlling costs.

In the telecommunications industry, the challenge is to improve the customer support experience, which can be considerably enhanced by using an AI assistant that understands a large percentage of the help requests.

6.2.1 Business use case definition

Customer centricity is needed throughout organizations. That is, knowing customers as individuals, and providing personalized and proactive care across many channels, when and where customers want.

Although providing this level of customer care is challenging, customer satisfaction scores are higher than ever before by applying practices to achieve such scores. Also, costs are reduced in comparison to before AI-enabled solutions were used.

Customer engagement analytics and AI-infused virtual assistants (chatbots) can help provide a superior level of personalized customer service when and where your customers want to communicate. AI helps cut costs, automate queries, and free employees' time by working on customer care to focus on higher value interactions, which makes processes less complex for customers, employees, and businesses.

6.2.2 Understanding business requirements for customer care

Business requirements differ from one organization to another regarding customer care. Requirements can be centered around the following concepts:

- Smarter AI for customer service to improve the clients' satisfaction.
- Detection of emerging trends, performing predictive analytics, and gaining operational insights can be another requirement to enhance the operations of organizations.
- Understanding customer's voice and turn it into a searchable text in real time to provide immediate support.
- Engaging with customers through voice services by using the customer's own language to create an automated conversation.

Improving customer care with IBM Watson

Each of the business requirements has a different solution; therefore, they are addressed differently through different IBM Watson tools:

- IBM Watson Assistant addresses business needs when it comes to having an AI conversion.
- IBM Watson Discovery addresses the need to enhance internal operations by the predictive analytics and operational insights.
- IBM Watson Speech to Text is used to turn the customer's voice into searchable text in real time to enable a seamless experience for customers that use a voice agent. Also, call recordings can be transcribed for use in spotting trends or training issues, and speed live agent call resolution by supplying agents with relevant information as customers say it.
- IBM Watson Text to Speech enables the customer experience improvement and engagement by interacting with users in their own languages by using any written text.

6.2.3 Use case examples

In this section, we present some use case examples for customer care.

Customer self-service

With this pattern, AI powers conversations between a business and its customers in natural language (written and spoken). For this process to occur, our customers cannot rely on chatbots that are powered by a simple rules engine.

For this reason, IBM developed IBM Watson Assistant, which ensures that customers get help in an efficient way without encountering dead ends. By using advanced natural language understanding (NLU) and AI-powered search, IBM Watson Assistant dynamically uses a broad and constantly growing set of relevant information to find the most accurate answer.

If this approach does not resolve a customer's need, IBM Watson Assistant can pass the conversation over to a human agent, including the conversation history and context (for example, if AI models were run for a sentiment score or churn risk) so that the customer does not have to restart the contact cycle from the beginning.

Employee AI self-service

This use case creates a single path for your employees to solve even their most complex issues, across any digital or voice channel. Applying customer care solutions for employees reduces costs while accelerating resolutions by using natural language processing, which leads to higher containment rates and simpler deployments.

It also helps in getting insights from your data and connects with existing tools and uses AI search to surface the most relevant answers within existing content across your organization.

These solutions improve efficiency and streamline how organizations empower employees to resolve inquiries across channels, including web, messaging, and voice. The same approach is applied to a AI Telecommunications Virtual Assistant use case.

6.3 Conversational AI

Conversational AI is a capability that is provided by the IBM Watson Assistant service on Cloud Pak for Data. This capability is used to create chatbots that use AI to understand the customers in context. With this context, customers are provided with fast, consistent, and accurate answers in an omnichannel environment.

The other use cases that are presented in this chapter can be used to enhance the chatbot with natural language capabilities and document understanding.

In this section, we focus on the base chatbot capabilities.

6.3.1 Reference architecture

Figure 6-14 shows the reference architecture for the conversational AI.

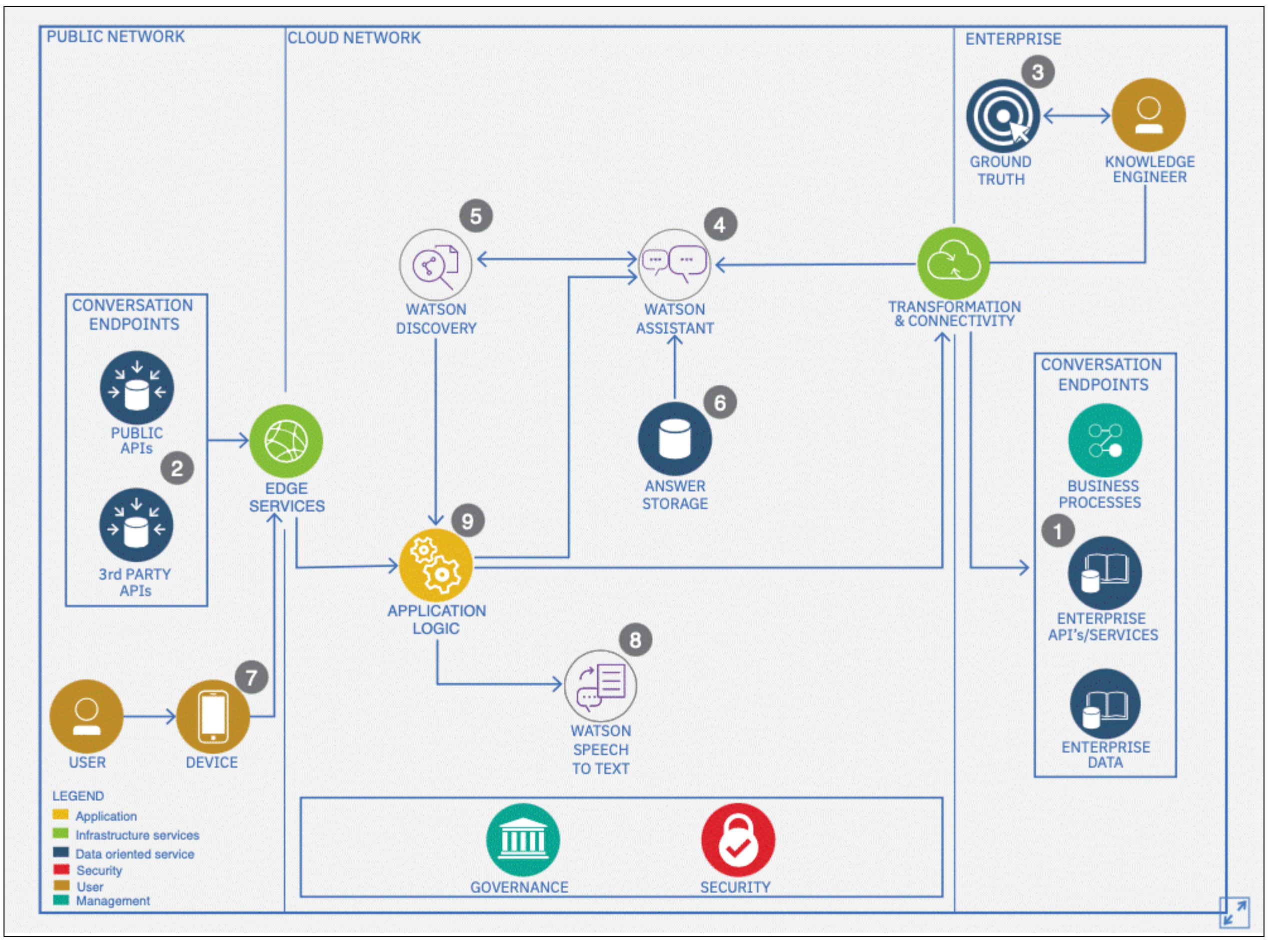

Figure 6-14 Reference architecture for the conversational AI

For more information, see this IBM Cloud web page.

6.3.2 Prerequisites

Before we can start building an assistant a Cloud Pak for Data Administrator, we must install the IBM Watson Assistant service. For more information about this installation process, see this IBM Documentation web page.

As part of the installation process, several configuration options are available that are provided to the installer as an installation options yaml file.

In our example, we use a minimally sized installation and a single language: English. For more information about a minimum cluster size for IBM Watson Assistant, see this IBM Documentation web page.

This sizing is adequate for the hands-on lab that is described later in this chapter. At least four worker nodes are required on the cluster for the service to install. For a production ready-sized instance of IBM Watson Assistant, several key criteria dictate the sizing, including the following examples:

- Number of users per month.
- The languages that are installed.
- Whether high availability (HA) is required; HA is recommended for production systems.

For a thorough sizing, contact your IBM sales representative.

6.3.3 Assistant artifacts

Before we begin creating an IBM Watson Assistant implementation, we review the chatbot components.

Instance

We start with an instance of IBM Watson Assistant. At the time of writing, each deployment of Cloud Pak for Data with IBM Watson Assistant installed can support up to 30 instances of the service. The instance supports the second of the tenancy models that are defined at this IBM Documentation web page.

An instance of Watson Assistant is created by using the navigation menu to browse to the service section and clicking **Service Instances**, as shown in Figure 6-15.

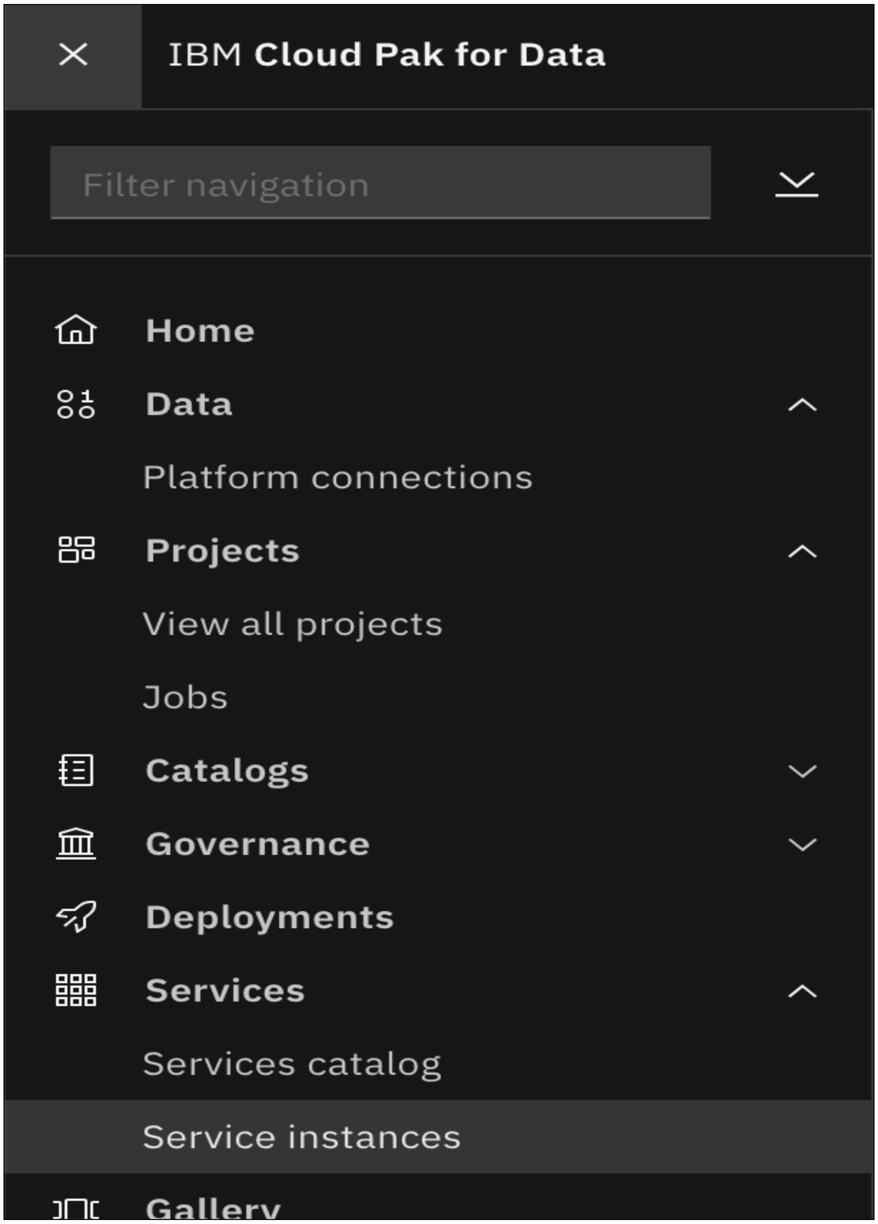

Figure 6-15 Selecting Service instances

Click **New Instance** in the top upper right of the window and then, select the **Watson Assistant** tile to create an instance of the assistant.

You also can search for the Watson Assistant service from the Services catalog to create an instance for the IBM Watson Assistant service, as shown in Figure 6-16.

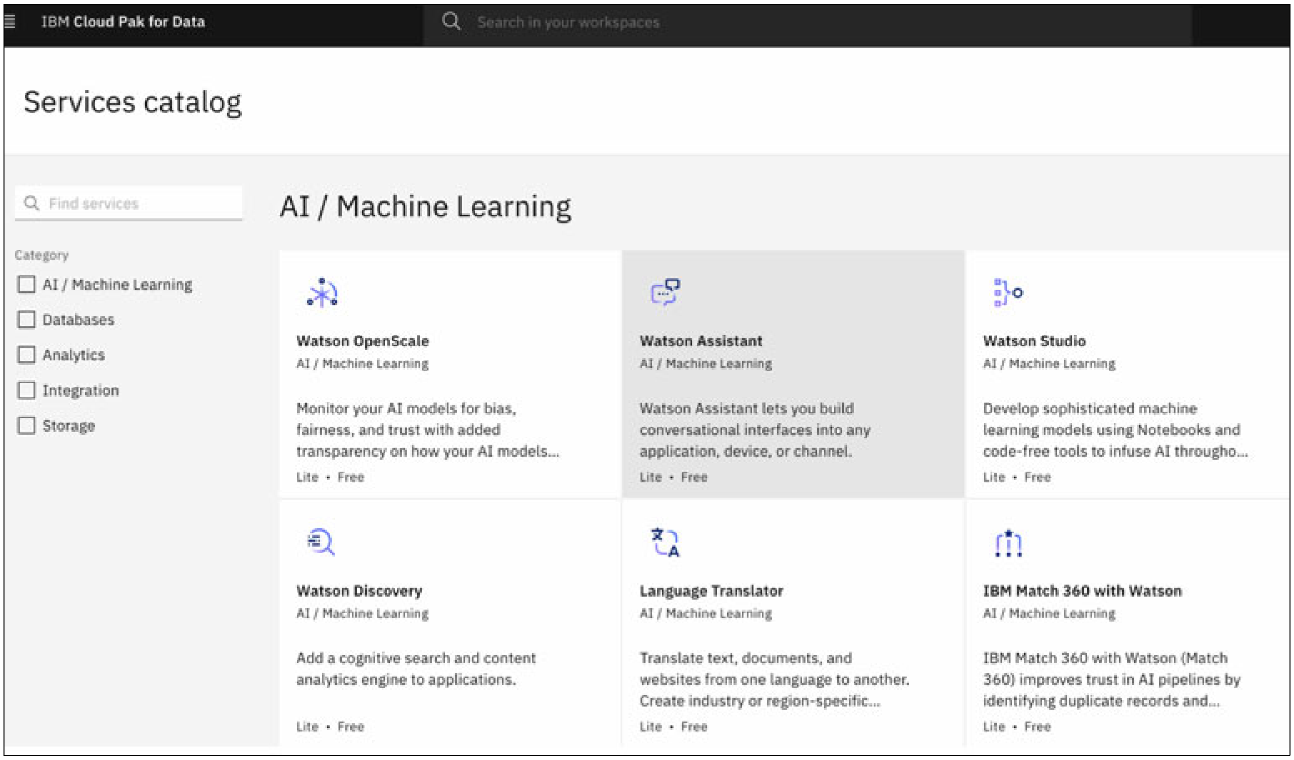

Figure 6-16 Selecting Watson Assistant service from the services catalog

Access to instances can be controlled by using the Manage Access tab of the instance. Cloud Pak for Data users and groups can be added to the instance as administrators or users. For more information about configuring access control, see this IBM Documentation web page.

Skills

Skills are added to your instance of Watson Assistant to realize your business requirements. These skills are available in the following versions:

- Dialog

  This skill uses IBM Watson natural language processing and machine learning technology to understand your questions, and respond to them with answers that are authored by you.

  Two key methods are available for creating an instance of IBM Watson Assistant: by creating actions or dialogs. The use of actions is a new simplified development process for creating IBM Watson Assistant implementations.

  A user can switch between these two approaches at any time without losing any data. For more information about use the actions approach, see this IBM Documentation web page.

  For more information about the pros and cons of each approach to provide the conversational skills, see this IBM Cloud Docs web page.

  The focus in this book is on the use of the dialog approach.

- Search

  An assistant uses a search skill to route complex customer inquiries to IBM Watson Discovery for IBM Cloud Pak.

  For more information about the use of IBM Watson Discovery, see 6.5, "Content intelligence" on page 474.

The dialog skill is the container for the other elements of your chatbot. The dialog skill can be exported from one instance of IBM Watson Assistance and imported into another to provide the key capabilities for the software development lifecycle and movement of the IBM Watson Assistant code through the development, testing, and production process.

Intent

An *intent* is a collection of user statements that have the same meaning. By creating intents, you train your assistant to understand the various ways users express a goal. An intent features a name and a selection of user examples that enable training IBM Watson Assistant. An example of a hotel booking intent is shown in Example 6-4.

Example 6-4 Creating an intent

```
Name: #BOOK_HOTEL
Description: An optional description of the intent
User examples:
I want to book a room
I want to book a hotel
I want to find a hotel
I need a place to stay
Could you create a hotel reservation for me
```

Intents are always prefixed with a # sign. In Example 6-4, five user examples are provided and they are used to train IBM Watson Assistant. After the assistant is trained, it recognizes various other user inputs, such as the following examples:

- Could you find me a hotel.
- I need a hotel.
- Can you reserve a room?

As intents become more complex than the simple example that is shown in Example 6-4, a computational linguist is used to garner the best results.

> **Note:** A standard set of intents are included with the IBM Watson Assistant platform that can be imported into your dialog skill.

Entities

If intents can be seen as the verbs of the system, entities are the nouns. By building out your business terms in entities, your assistant can provide targeted responses to queries.

An *entity* represents a term or object that is relevant to your intents and it provides a specific context for an intent. The name of an entity is always prefixed with the @ character.

IBM Watson Assistant detects entities through a dictionary-based mechanism or an annotation-based mechanism.

Dictionary based approach

You can train the skill to recognize your entities by providing entity term values and synonyms, entity patterns, or by identifying the context in which an entity is typically used in a sentence.

Entities can be defined by using one of the following mechanisms:

- Synonym: You define a category of terms as an entity, values are added to the entity, and a collection of synonyms can be added for each value. IBM Watson Assistant also can make recommendations for synonyms.
- Pattern: You define a category of terms as an entity, values are added to the entity, and for each value that is added a regular expression can be defined that defines the textual pattern of the value.
- System entities: These entities are included with the platform. They support currency, dates, time, numbers, and percentages. For more information about system entities, see this IBM Documentation web page.

To fine-tune your dialog, add nodes that check for entity mentions in user input in addition to intents. As shown in Figure 6-17, the dialog recognizes if a specific Disney character as an Entity is selected and then, adds a specific response when a specific entity is recognized by the assistant.

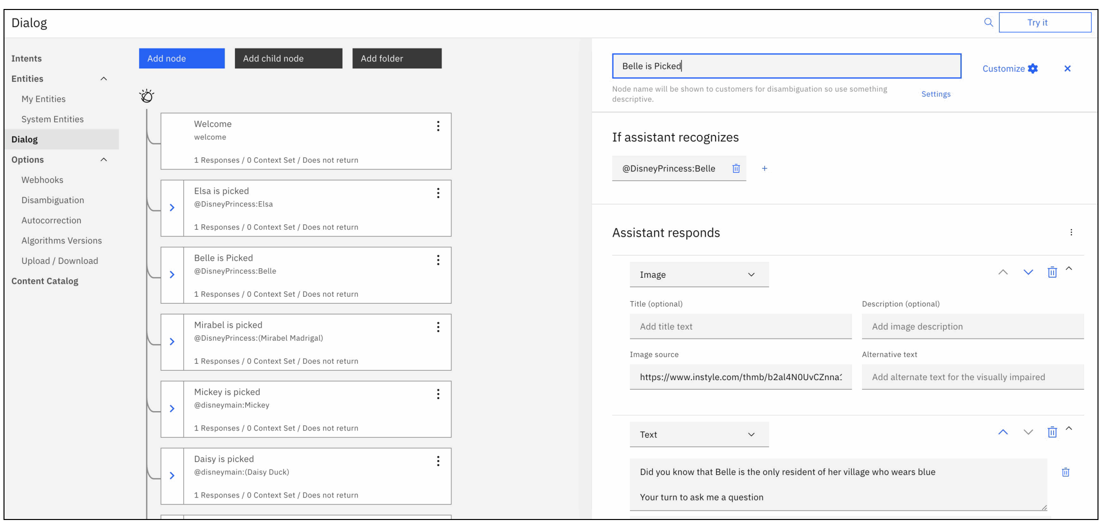

Figure 6-17 Dialog nodes to detect entities

Annotation-based approach

An annotation entity also is known as a *contextual entity* and is trained on the term and the context in which the term is used. A category of terms is first created as an entity and then, the user examples of the intents you use are mined to find references to these entities. These references are then labeled with the entity.

As shown in Figure 6-18 on page 461, the entities are defined and preceded by '@' annotation. For each entity, some terms are identified so that IBM Watson Assistant can recognize user input easier through these terms and map it to the corresponding entity.

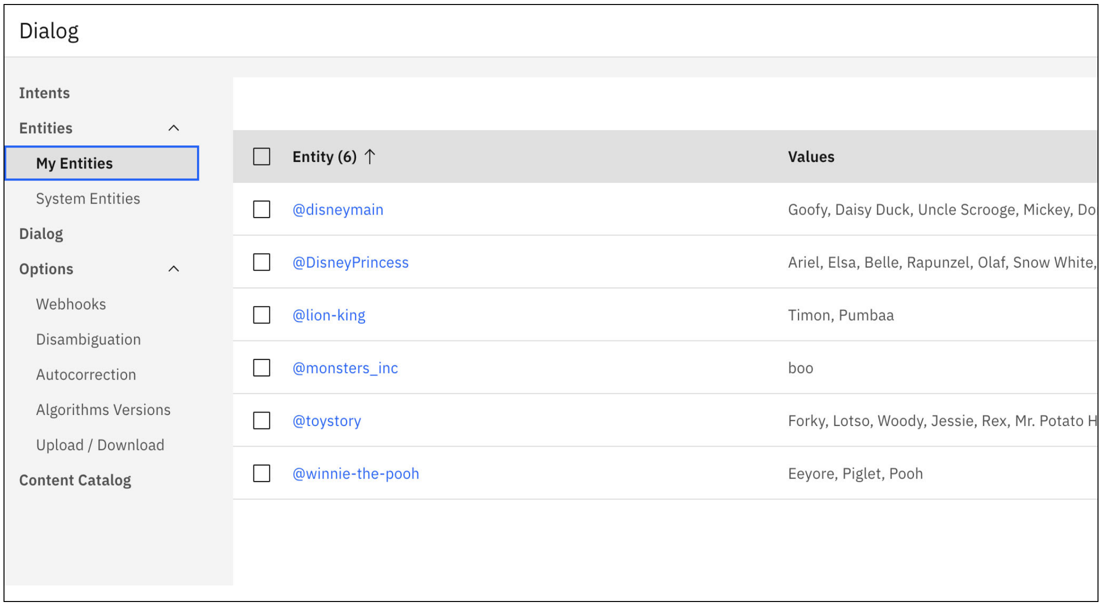

Figure 6-18 Annotation-based or Entity definition in Watson Assistant

Node

A dialog skill is defined as a tree. A branch is created for each intent that must be processed, and each branch consists of one or more nodes. The tree is traversed from top to bottom until a match is found.

An example for the processing flow is shown in Figure 6-19. Each node contains at a minimum one condition and one response. The condition for the primary node of a branch often is an intent. The conditions for the other nodes in a branch can be an entity type an entity value or a variable from the users context.

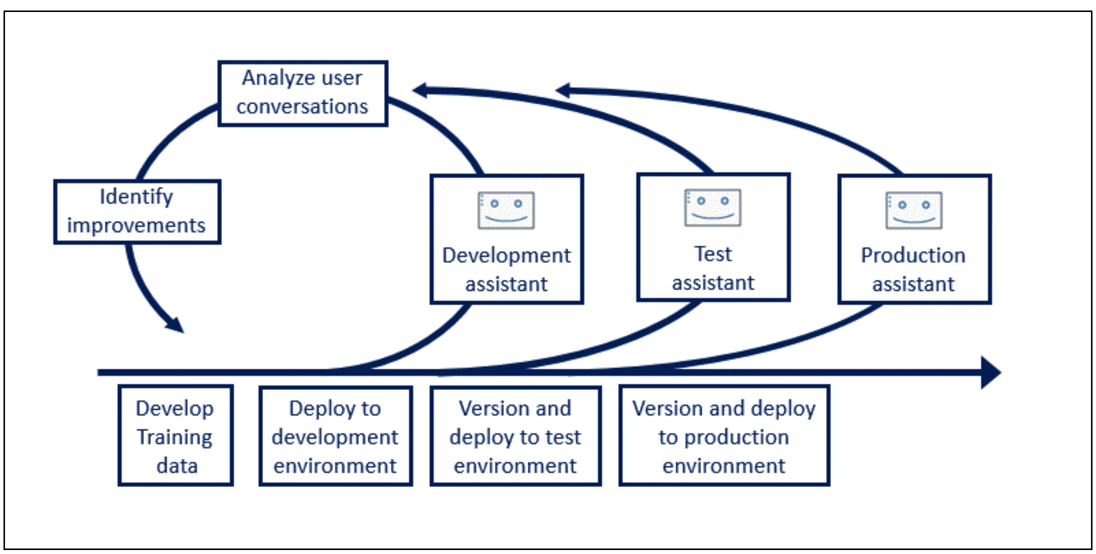

Figure 6-19 Dialog skill showing tree processing

A *condition* specifies the information that must be present in the user input for this node in the dialog to be triggered. The simplest condition is a single intent, but conditions can be any combination of intent, entity type, entity value, context variable, or special condition.

The following special conditions are part of the Watson Assistant platform:

- `anything_else`: You can use this condition at the end of a dialog to be processed when the user input does not match any other dialog nodes. The Anything else node is triggered by this condition.
- `welcome`: This condition is evaluated as true during the first dialog turn (when the conversation starts), only if the initial request from the application does not contain any user input. It is evaluated as false in all subsequent dialog turns. The Welcome node is triggered by this condition.
- `converstaion_start`: This condition is evaluated as true during the first dialog turn. Unlike `welcome`, it is true whether the initial request from the application contains user input. A node with the `conversation_start` condition can be used to initialize context variables or perform other tasks at the beginning of the dialog.
- `true`: This condition is always evaluated to true. You can use it at the end of a list of nodes or responses to catch any responses that did not match any of the previous conditions.
- `false`: This condition is always evaluated to false. You might use this condition at the start of a branch that is under development to prevent it from being used, or as the condition for a node that provides a common function and is used as only the target of a Jump to action.
- `irrelevant`: This condition evaluates to true if the user's input is determined to be irrelevant by the IBM Watson Assistant service.

A *response* is the way that the assistant responds to the user. The response can be an actual answer, image, list, or programmed action.

Slots are added to nodes to gather multiple pieces of information from the user. These slots then allow the node to provide a more specific response. For example, a hotel booking requires collecting information about the length of stay, number of guests, room type, and so on, *before* making a room offer.

A *slot* consists of a variable and a prompt. That prompt is passed back to the user to trigger the user to input the new required data. If the user responds with information for multiple slots, IBM Watson Assistant picks up information and populates the relevant slots.

As shown in Figure 6-17 on page 460, the node is added to the dialog. In the node's configuration, you can identify the intent or entity to be detected by IBM Watson Assistant and add a sample response that the assistant sends to the user when this entity or intent is recognized. An option is available to respond with text, image, video, audio, and other options.

Another concepts that are relevant for nodes are digressions. *Digressions* enable a user to interrupt the dialog for a specific intent and jump to a node that supports a new intent. Depending on how the nodes are configured for digression, the user can be brought back into the dialog branch for the initial intent.

For more information, see this IBM Documentation web page.

In our example, two dialog branches are used to support the `#Restautant` booking and the `#Restaurant` opening intents. While trying to book a restaurant, the user of the bot digresses to the restaurant opening hours.

Context

To personalize the conversation, the assistant can manage contextual information. Information can be collected from the user and stored so it can be used later in the conversation. Contextual information is managed by using context variables.

The dialog approach to creating an assistant is stateless. If the use of the newer interface than the interaction is stateful, context variables are still possible, but the context is managed server side. (We focus on the dialog mechanism for building an assistant.) In this example, contextual information is saved from message interaction and resubmits it on the next interaction.

The contextual information can be passed to the dialog nodes and from node to node through a context variable that is defined in a node. A default value can be specified for it. Then, other nodes, application logic, or user input can set or change the value of the context variable.

Conditions that include the context variable values are used by referencing a context variable from a dialog node condition to determine whether to run a node. A context variable can be referenced from dialog node response conditions to show different responses, depending on a value that is provided by an external service or by the user.

For example, as shown Figure 6-20, your application can set a `$time_of_day` context variable, and pass it to the dialog, which can use the information to tailor the greeting it displays to the user.

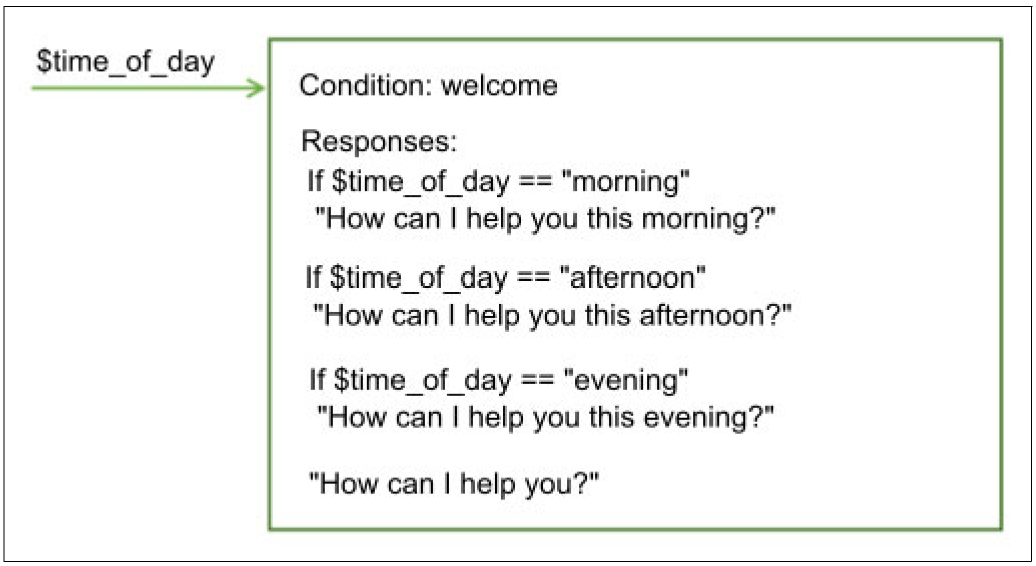

Figure 6-20 passing context to Watson Assistant node

Webhooks

A *webhook* is a mechanism with which you call out to an external program that is based on events in your dialog. Nodes can be configured to use webhooks by clicking customize for the node and then, enabling the webhook capability.

Webhooks are used in the answer store programming pattern that is defined next. For more information about an example demonstration that uses a webhook to interact with an answer store, see 6.3.5, “Creating an assistant” on page 465.

Answer store

The answer store is shown in the reference architecture (see Figure 6-14 on page 455). Although it is not a true component of IBM Watson Assistant, it is a common design pattern in which the flow is decoupled from the answers, which allows IBM Watson Assistant experts to create the flow and business SMEs can create the answers.

The answer store is a means of maintaining the answers in external answer storage. Nodes are defined in such a way that the answer is an ID that can be used by post response webhooks to get the correct answer from the answer store to be provided back to the user after the intent and entities are understood.

6.3.4 Development process

The development process is an iterative process. Generally, an IBM Watson Assistant initiative starts small with a set of customer needs that the assistant must address and any integrations that are needed to support those needs. The development lifecycle is shown in Figure 6-21.

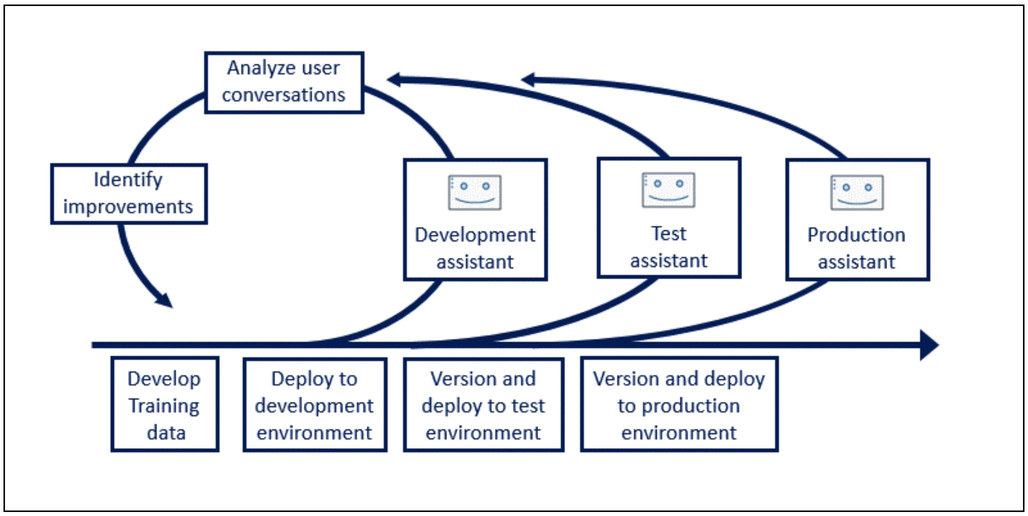

Figure 6-21 IBM Watson Assistant development lifecycle

The needs are implemented as intents, and the intents are enriched with entities where needed. Each intent is tested by using the user interface, when the user conversation is analyzed to be “happy”. The dialog can be versioned and processed by using the Software Development Lifecycle.

As the assistant progresses, the chat logs of the assistant are monitored to see whether any identifiable improvements to the assistant can be made. These improvements are fed back into the development process.

6.3.5 Creating an assistant

Many publicly available materials provide hands-on experiences with IBM Watson Assistant. Some are available as demonstrations within the IBM cloud, others are GitHub repositories that can be used as the basis for the creating an assistant in an on-premises deployment of Cloud Pak for Data. Some of the hands-on material that is available is listed in Table 6-1.

Table 6-1 Available material

Description	Location
Virtual insurance assistant	https://github.com/IBM/virtual-insurance-assistant
Watson Assistant Slots Intro	https://github.com/IBM/watson-assistant-slots-intro
Watson Voice Bot	https://github.com/IBM/watson-voice-bot
Retail Customer Service chatbot.	https://developer.ibm.com/tutorials/create-your-first-assistant-powered-chatbot/?mhsrc=ibmsearch_a&mhq=Watson%20assistant
Integrate a custom live agent with IBM Watson Assistant	https://developer.ibm.com/tutorials/integrate-a-custom-live-agent-with-ibm-watson-assistant/?mhsrc=ibmsearch_a&mhq=Watson%20assistant
Create an assistant with an answer store	https://github.com/foward/answer-store-watson
Using digressions	https://cloud.ibm.com/docs/assistant?topic=assistant-tutorial-digressions

6.4 Speech services

In this section, we discuss the speech services that are available in Cloud Pak for Data.

6.4.1 IBM Watson Text to Speech

IBM Watson Text to Speech is an API cloud service that enables the conversion of written text into natural-sounding audio in various languages and voices within an application or IBM Watson Assistant.

The voices can be customized to brands and leads to improving customer experience and engagement by interacting with users in their native language. By using IBM Watson Text to Speech, you can increase accessibility for users with different abilities, provide audio options to avoid distracted driving, or automate customer service interactions to eliminate hold times.

IBM Watson Text to Speech Service helps users realize the following benefits:

- Improve user experience

  Helps all customers comprehend the message that you want to send or receive by translating written text to audio.

- Boost contact resolution

  Solves customer issues faster by providing key information in their native language.

- Protect your data

  Enjoy the security of IBM’s world-class data governance practices.

- Run it anywhere

  Supports global languages and deploys on-premises or on any cloud.

IBM Watson Text to Speech services include the following features:

- Natural-sounding neural voices

  Benefit from IBM's deep neural networks that are trained on human speech to automatically produce smooth and natural sounding voice quality.

- Custom voices

  Users can design their own unique branded neural voice that is modeled after choosing the speaker by using as little as one hour of recordings.

- Controllable speech attributes

  Easily adjust pronunciation, volume, pitch, speed, and other attributes by using Speech Synthesis Markup Language.

- Customized word pronunciations

  Clarify the pronunciation of unusual words with the help of IPA or the IBM SPR.

- Expressiveness

  Control the tone of voice by choosing a specific speaking style: GoodNews, Apology, and Uncertainty.

- Voice transformation

  Personalize voice quality by specifying attributes, such as strength, pitch, breathiness, rate, timbre, and more.

From the services catalog, search for Text to Speech service to create a Text to speech service instance, as shown in Figure 6-22.

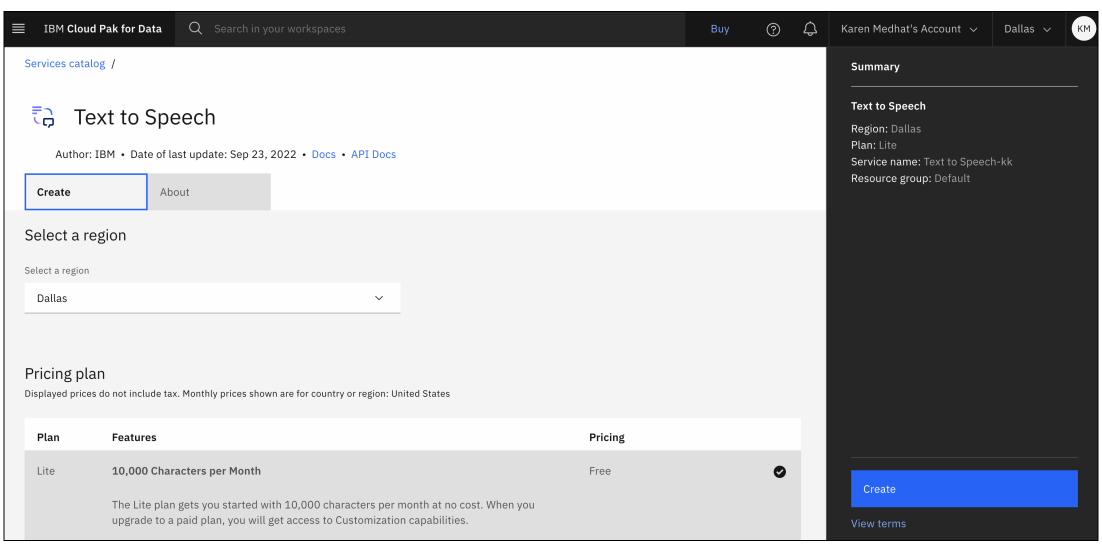

Figure 6-22 Text to Speech service creation page

After creating the service instance, as shown in Figure 6-23, you find a URL for the API reference page with all of the API details that are needed to integrate with the Text to Speech instance in the application. The service credentials on the left side menu of the landing page are the needed credentials for accessing the APIs of the service instance.

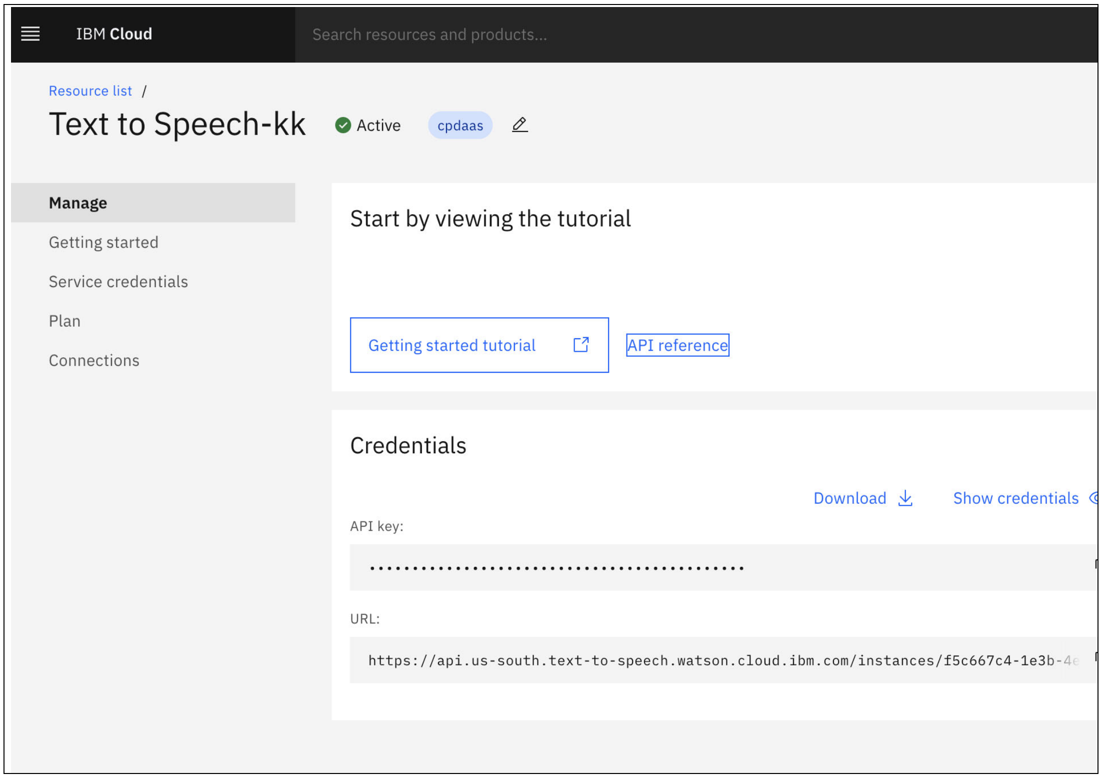

Figure 6-23 Landing page of Text to Speech service instance

6.4.2 IBM Watson Speech to Text

IBM Watson Speech to Text technology enables fast and accurate speech transcription in multiple languages for various use cases, including but customer self-service, agent assistance, and speech analytics. It features advanced machine learning model components that are immediately available and can be customized according to the business use case.

The use of Watson Speech to Text Service includes the following benefits:

- More accurate AI

  Embedded within IBM Watson Speech to Text, the AI understands the customers.

- Customizable for your business

  Train IBM Watson Speech to Text on unique domain language and specific audio characteristics.

- Protects the data

  IBM Watson Speech to Text is used with confidence under the world-class data governance practices of IBM.

- Truly runs anywhere

  Built to support global languages and deployable on any cloud, public, private, hybrid, multicloud, or on-premises.

IBM Watson Speech to Text Service includes the following features:

- Pre-trained speech models

  Activate voice applications with speech models that are tuned for the customer care domain.

- Model training options

  Improve speech recognition accuracy for the business use case with language and acoustic training options.

- Fine-tuning features

  Improve speech recognition accuracy for extracting phrases, words, letters, numbers, or lists.

- Low latency transcription

  The models that are used in the service are optimized for low latency in real-time speech applications.

- Audio diagnostics before transcription

  Analyze and correct weak audio signals before transcription begins.

- Interim transcription before final results

  Improve application response times by using speech transcription as it is generated and throughout the finalization process.

- Smart formatting

  Transcribe dates, times, numbers, currency values, email, and website addresses in your final transcripts by converting them into conventional forms.

- Speaker diarization

  Recognize who said what in a multi-participant voice exchange. Currently, optimized for two-way call center conversations but can detect up to six different speakers.

- Word spotting and filtering

  Filter for specific words or inappropriate content by using keyword spotting and profanity filtering features (US English only).

From the services catalog, search for Speech to Text service to create a Speech to Text service instance as what was done for the Text to Speech service that is shown in Figure 6-22 on page 466.

After creating the service instance, you find a URL for an API reference page with all of the API details that are needed to integrate with the Speech to Text instance in the application. The service credentials on the left side menu of the landing page are the needed credentials for accessing the APIs of the service instance, which is the same as the Text to Speech service instance shown in Figure 6-23 on page 467.

6.4.3 Speech services with IBM Watson Assistant

IBM Watson Speech to Text and IBM Watson Text to Speech services can be used with IBM Watson Assistant to enhance the automated self-service experience. IBM Watson Assistant is at the core of IBM's AI for the customer care solution because it delivers what is needed to build, train, integrate, and support a virtual agent in digital channels as web, or smartphone or video channels.

Figure 6-24 shows the capabilities that are enabled by Watson Assistant.

Figure 6-24 IBM Watson Assistant at the core of IBM's AI for customer care

6.4.4 Architectural patterns

The following common architectural patterns are available for the use of IBM Watson Assistant with IBM Watson Speech services.

- Accelerate the digital business
- Augment employee teams
- Modernize the contact center with AI

These common patterns are described next.

Accelerate digital business

One example of the accelerate digital business pattern is digital self-service, as shown in Figure 6-25. In this pattern, the user sends a message to the web UI of IBM Watson Assistant, which then analyzes the user input by using IBM Watson Discovery service to search for the skill and then, respond to user.

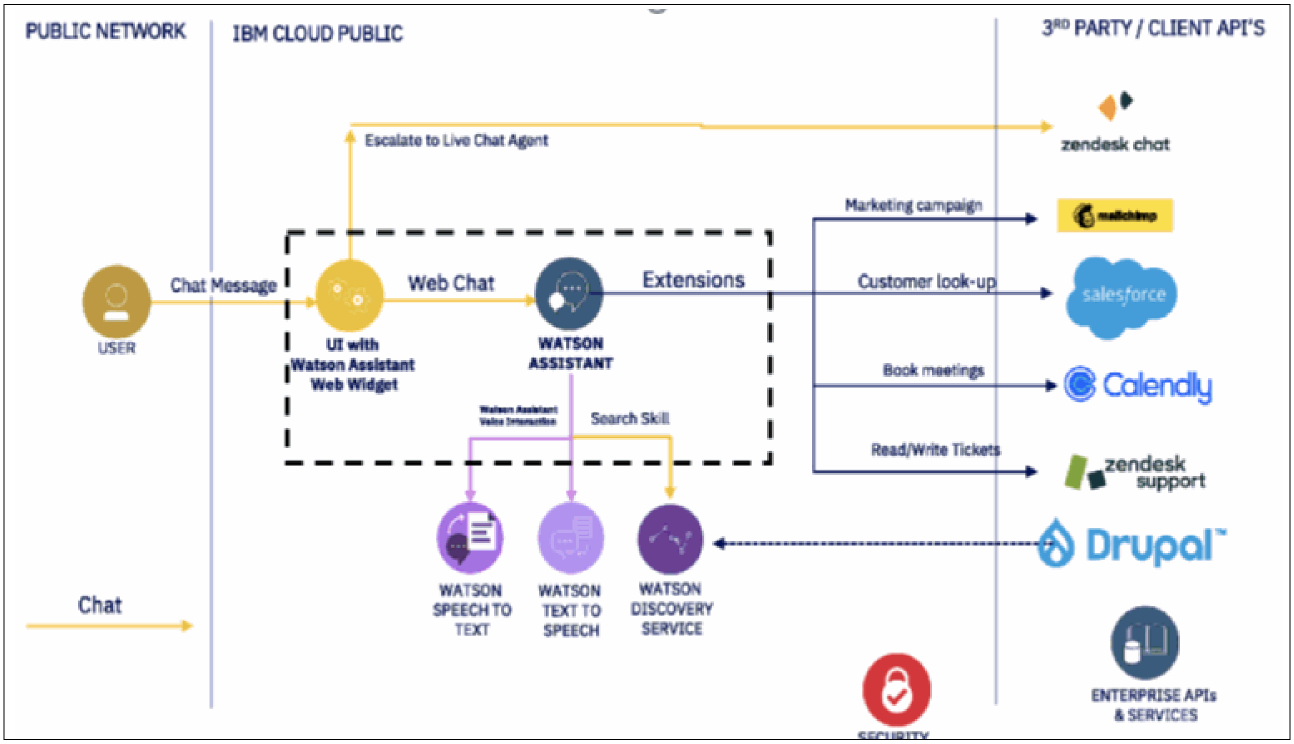

Figure 6-25 Digital self-service

If the user input has voice or if the response must be voice, the IBM Watson Assistant interacts with the IBM Watson Speech to Text and Text to Speech services to perform the needed action.

The IBM Watson Assistant also is integrated with other systems to trigger some tasks, such as booking a meeting, raising tickets, looking up a customer, or being connected to a live agent.

The other example for accelerating digital business is having omnichannel customer experience.

Omnichannel customer service is assistance and advice for customers across a seamless and integrated network of devices and touchpoints. Businesses with robust omnichannel customer service can maintain consistently great experiences for their customers, regardless of the communication channel.

The growth of digital channels and new communication technologies enabled businesses to adopt an omnichannel approach to customer support. In doing so, they can manage interactions across multiple channels, such as call centers, web chats, SMS, messaging, email, and social media.

For example, a customer support conversation might begin on Twitter and then, continue with text messages and end with a phone call, all in a seamless, connected experience.

Customers do not have to stop and explain their problem at each channel interaction. As shown in Figure 6-26, the user can start the interaction by a phone call or web chat and receive the response back by SMS or phone call. Figure 6-26 shows one layer for CCaaS (live voice agent transfer) and one layer for service desk providers (live chat).

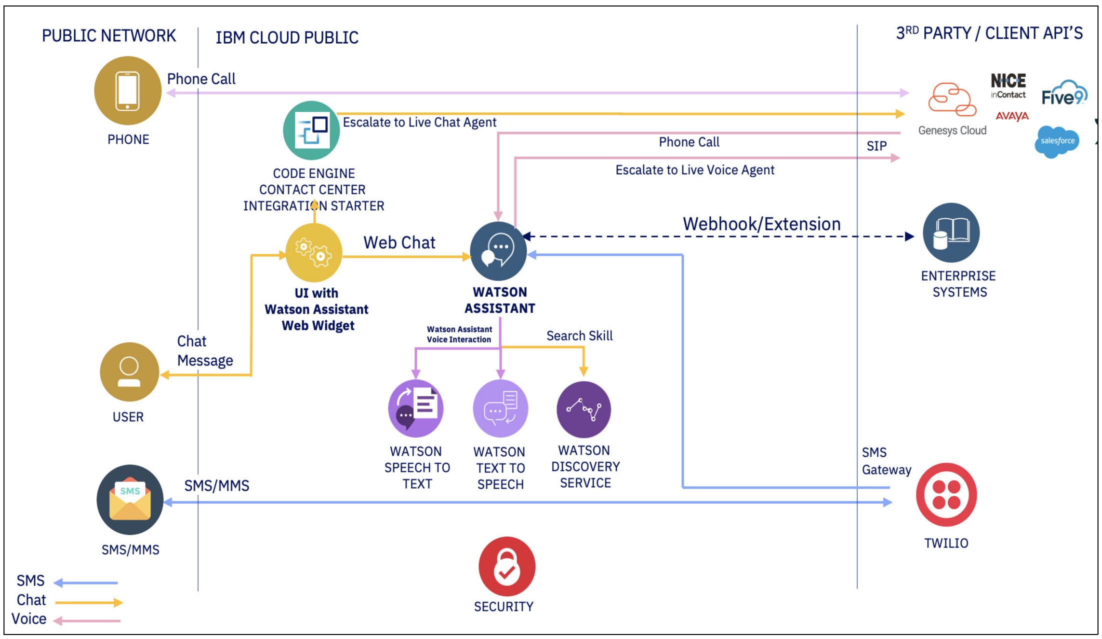

Figure 6-26 Omnichannel concierge

Augment Employee Teams

One example of augment employee teams is the HR support, as shown in Figure 6-27.

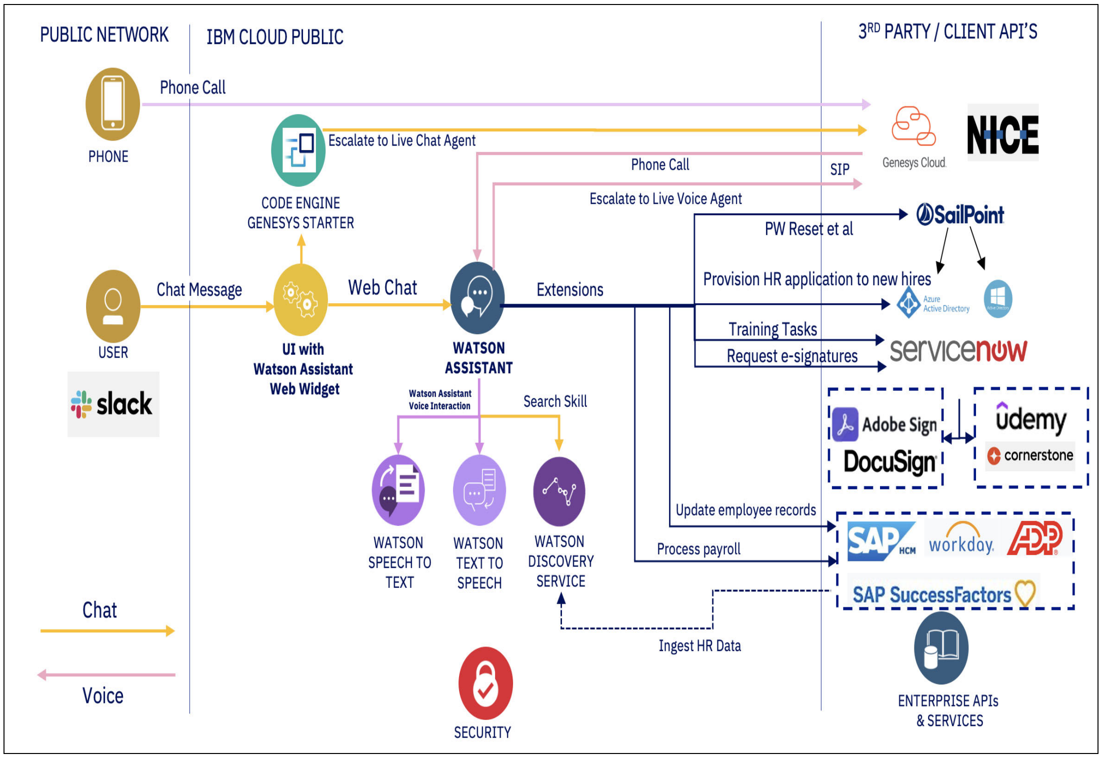

Figure 6-27 Augment Employee Teams HR

The user sends a message to the web UI of IBM Watson Assistant or though a phone call, which then analyzes the employee input and whether escalate to a live agent for trigger a task within one of the integrated tools or extensions for handling the HR processes.

Modernize the contact center with AI

One example of contact center modernization is voice automation with phone integration, as shown in Figure 6-28 where the interaction with the system is through phone calls and the user receives automated voice responses through the speech services.

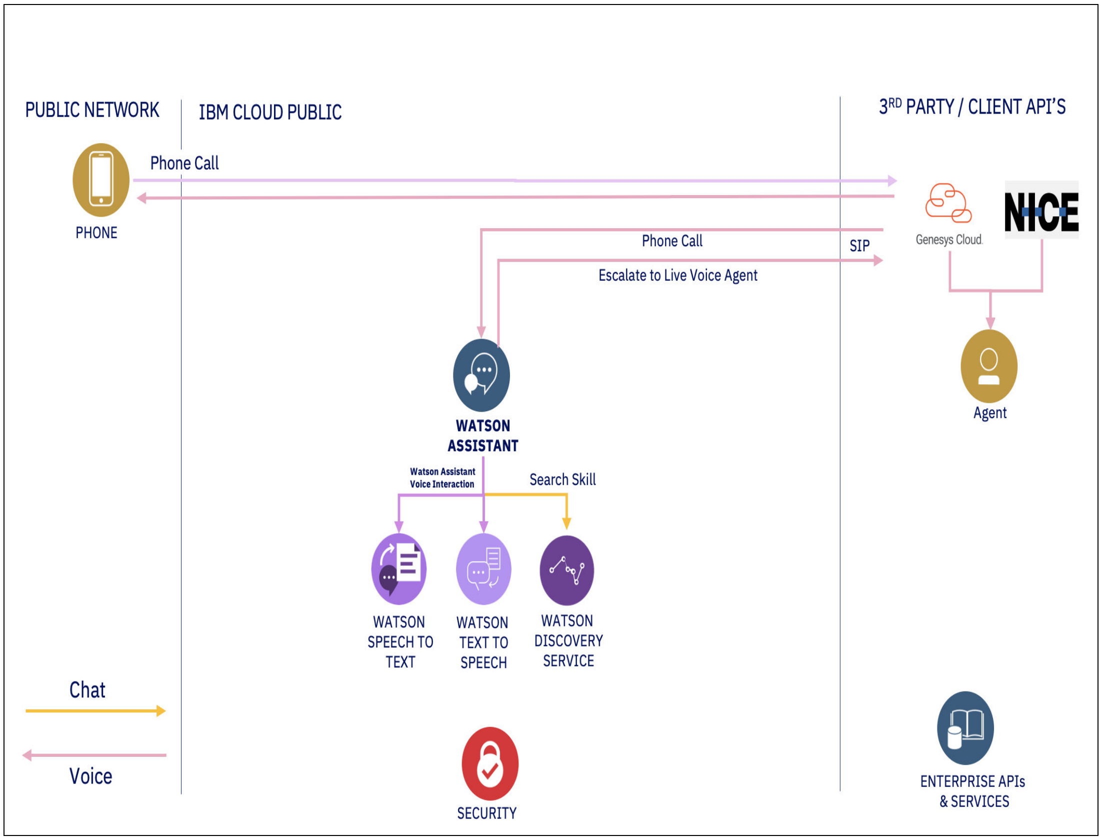

Figure 6-28 Modernize the contact center-voice automation with phone integration

6.4.5 Speech services

Several published publicly available materials provide hands-on experiences with IBM Watson Text to Speech and Speech to Text services. Some are available as demonstrations within the IBM Cloud and others are GitHub repositories that can be used as examples of the use of an IBM Watson Speech services deployment on Cloud Pak for Data.

Publicly available resources with examples are listed in Table 6-2.

Table 6-2 Watson Speech services resources

Description	Location
IBM Watson Speech to text online demonstration	https://www.ibm.com/demos/live/speech-to-text/self-service/home
IBM Demos Watson Hands-On Labs	https://www.ibm.com/demos/live/content/watson/stt/lab/hands-on-lab-customization.pdf
IBM Watson Text to Speech live demo	https://www.ibm.com/demos/live/tts-demo/self-service/home
API documents for speech to text	https://cloud.ibm.com/apidocs/speech-to-text
API documents for text to speech	https://cloud.ibm.com/apidocs/text-to-speech
Integrate voice services into your Watson Assistant	https://developer.ibm.com/tutorials/integrate-voice-services-to-your-watson-assistant/

6.5 Content intelligence

Enterprises around the world are evolving to become more digitally focused as they look to use structured and unstructured data for their most complex problems and processes that is often distributed across multiple systems and departments.

As this evolution occurs, a clear need emerges for technology that can surface facts, insights, and hidden patterns to different knowledge workers through the business. During customer service experiences, failure to find the right answers fuels agent turnover and leaves customer with a bad experience.

Content intelligence is a key part of the IBM customer care portfolio that can help provide a positive customer experience by providing AI solutions with IBM Watson Discovery as a complete solution for document and language intelligence.

The IBM Watson Discovery service on Cloud Pak for Data adds a cognitive engine to applications to search content and apply analytics to it. With the information, patterns and trends can easily be identified to gain insights that drive better decision making.

Imagine having all of the data that is needed to answer questions. The IBM Watson Discovery service can process data from various sources where the processing occurs in several ways, including the following examples:

- With APIs, content can be uploaded by using an application, or a custom mechanism can be created.
- With discovery tools, locally accessible files for configuration and testing can be uploaded.
- With Data Crawler, many files from a supported repository can be uploaded by using the command line.

Figure 6-29 shows a high-level overview of how the IBM Watson Discovery processes can be interpreted.

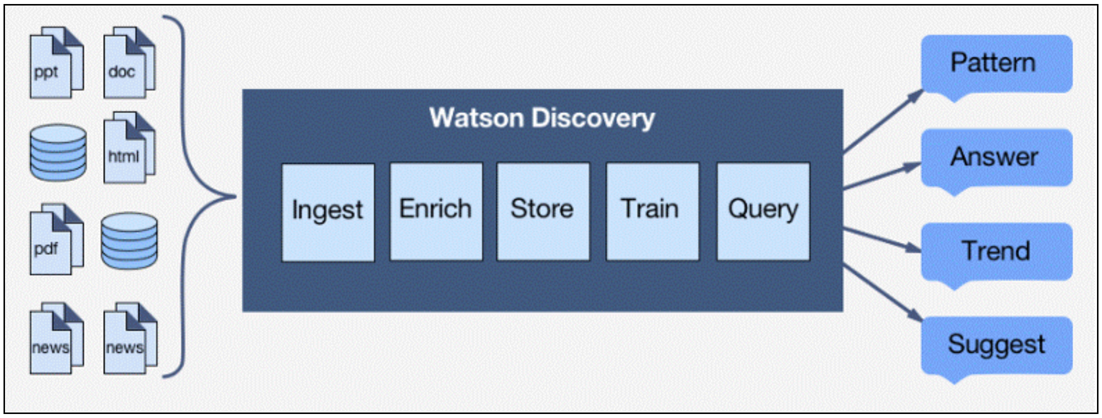

Figure 6-29 Watson Discovery process

IBM Watson Discovery provides an intelligent document understanding platform for the enterprise that is built to understand the language of business and accelerate the knowledge process for customer experience.

Publicly available data and enterprise-specific data are processed and then enriched to be used by the application. Data enrichment provides a collection of text analysis functions that derive semantic information from the enterprise's context.

Text, HTML, or a public URL can be provided, and Natural Language Processing (NLP) techniques are used to get a high-level understanding of the enterprise's context to obtain detailed insights, such as directional sentiment from entity to object.

Specific benefits of the intelligent document processing capabilities of Watson Discovery include the following examples:

- Extract data from human-read documents.
- Enrich text with the most advanced immediately available NLP.
- Enhance understanding by teaching IBM Watson Discovery to improve with input from business users.
- Analyze data at scale to find answers, insights, and patterns.

IBM Watson Discovery connects to, and extracts data from documents that are intended to be read by humans, including PDFs, Microsoft Word documents, HTML, contacts, and even Microsoft PowerPoint presentations.

After IBM Watson Discovery extracts the information, it needs to Enrich the document and understand the material. With advanced immediately available NLP, IBM Watson Discovery detects entities, keywords, concepts, and sentiment. This process mimics how humans read and process human-read documents.

The IBM Watson Discovery NLP document enrichment is great at understanding most content and context. However, at times it must be trained on specialized or domain-specific terminology and jargon. Industry and enterprises have their own specific language, with nuances that, if not understood in the correct context, can fundamentally change the meaning of the content.

The IBM Watson Discovery Enhance step is where the training occurs to improve its understanding of the industry and company-specific terms and jargon. After IBM Watson Discovery extracts, enriches, and enhances the documents, this information can now be Analyzed at scale to find answers, insights, and patterns to accelerate complex business processes.

After the data is processed and enriched, it is securely stored and available to only the application. This information can then be gathered by using search queries.

The IBM Watson Discovery services provide search capabilities through queries. The search engine finds matching documents from the processed data. Then, the engine applies a formula that provides relevance scoring to return the best answer to the query.

When an application uses the IBM Watson Discovery service, the people who use that application can gain insight from textural data. Customer service is a typical example of how IBM Watson Discovery is used for an improved customer care experience by using its powerful cognitive search engine as part of the IBM customer care portfolio on Cloud Pak for Data.

6.5.1 IBM Watson Discovery architecture

The IBM Watson Discovery architecture enables the ability to rapidly build AI, cloud-based exploration applications that unlock actionable insights in unstructured data. That data includes the enterprise unique proprietary data, and public and third-party data.

The AI Discovery design time architecture and process flow that is shown in Figure 6-30 on page 477 includes steps to define requirements, prepare and configure the environment, and integrate into applications.

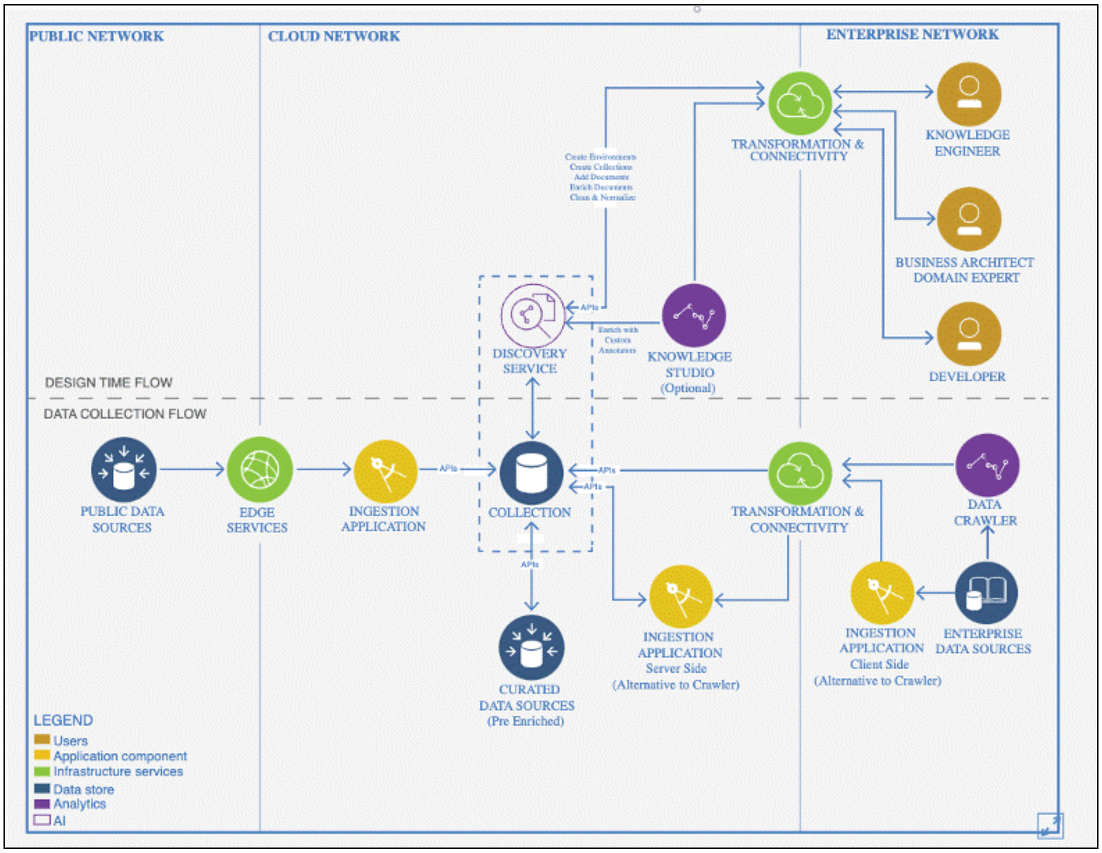

Figure 6-30 Watson Discovery reference architecture

Table 6-3 lists the architectural concepts that are referenced in Figure 6-30.

Table 6-3 Watson Discovery architectural concepts

Concept	Definition
Edge services	Enables data to flow safely from the internet.
Public data sources	Structured and unstructured data sources from the public internet domain.
Discovery service	Ingests, parses, indexes, and annotates content by using cognitive functions.
Collection	A grouping of content with the environment. At least one collection must be created to upload the content.
Curated data sources	Includes vendor provided data feeds.
Ingestion application	Ingests data by using discovery service APIs.
Transformation and connectivity	Includes scalable messaging and transformation and secure connectivity. Application logic also can strengthen the response by supplementing structured data (such as user profile, past orders, and policy in-formation) from the enterprise network. The connection to the enterprise network is established through the transformation and connectivity component.
Data crawler	Crawls data sources and ingests into discovery service collection.
Knowledge studio	Domain-specific text and content analytics that uses machine learning and rules-based annotators.

Concept	Definition
Knowledge engineer	Product specialist who knows technology in depth and designs the solution based on the specifications of the business architect.
Business architect	Subject matter expert and domain expert who understands the data in depth and helps define the requirements and specifications of the overall solution.
Developer	Programmer who can develop the custom components of the solution according to the specifications of the business architect and the design of the knowledge engineer.
Enterprise data source	Data sources from the enterprise domain; can include ECM repositories, file shares, databases, and others.

The Watson Discovery architecture that is referenced in Figure 6-30 on page 477 allows for seamless integrations with other products in the IBM customer care portfolio.

For more information about how to integrate IBM Watson Discovery for content intelligence by using an AI-power search skill, see “Hands-on experience with Watson Discovery” on page 480.

6.5.2 Using IBM Watson Discovery

The AI powered search capabilities with IBM Watson Discovery help agents respond to complex queries. Complex queries can require an agent, yet IBM Watson Discovery that is integrated with IBM Watson Assistant or other speech services in the IBM customer care portfolio on Cloud Pak for Data still plays a role.

As described in 6.3, “Conversational AI” on page 455, IBM Watson Assistant can help agents answer questions faster and with more confidence. When paired with IBM Watson Discovery by using a search query, IBM Watson Assistant delivers accurate information that is drawn from enterprise data.

By using Natural Language Processing (NLP) to understand the industry’s unique language, IBM Watson Discovery finds answers fast. IBM Watson Discovery can be easily paired with IBM Watson Assistant and other speech services. For more information about IBM Watson Discovery concepts and capabilities, how to set up the service, and publicly accessible examples, see 6.5.2, “Using IBM Watson Discovery” on page 478

IBM Watson Discovery concepts

Before getting started with IBM Watson Discovery for customer care or any other use case, it is important to understand the concepts and vocabulary that is used.

Figure 6-31 on page 479 shows the eight types of Natural Language Understanding (NLU) information that IBM Watson Discovery uses to enrich business documents and other content.

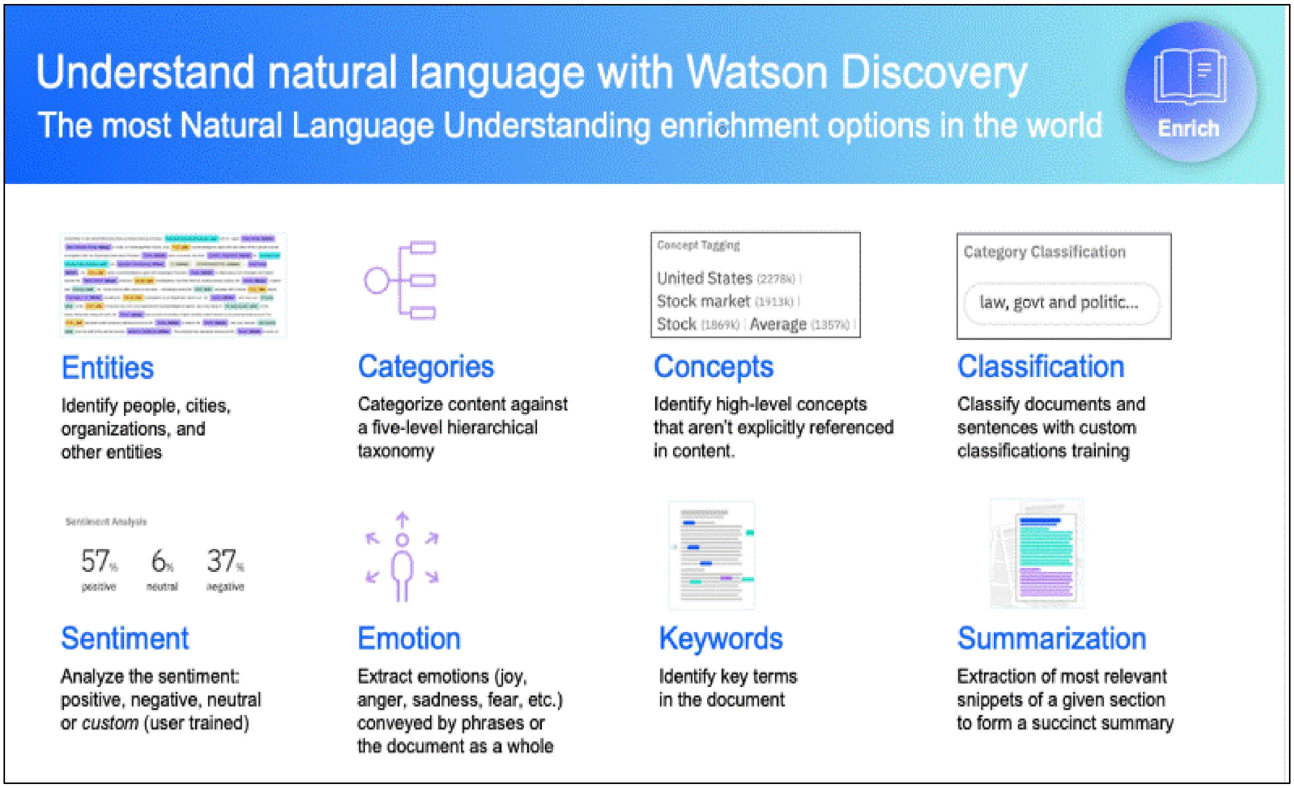

Figure 6-31 Watson Discovery NLU enrichments

IBM Watson Discovery is built on IBM Watson NLU, which provides more content enrichment options than other vendors. Immediately available, IBM Watson Discovery can automatically enrich business documents and other content with the following types of information:

Entities Identifies people, cities, organizations, and more.

Categories Categorizes content hierarchically, whether the hierarchy is accounts, organizations, products, geographical, and so on.

Concepts By understanding how concepts relate, IBM Watson Discovery can identify high-level concepts that are not necessarily referenced specifically in content. For example, if a page references stocks and stockbrokers, IBM Watson Discovery can identify the stock market as a concept, even if that term is not mentioned specifically in the page. Concept tagging enables higher-level analysis of input content than only basic keyword identification.

Classification Classifies documents and sentences with custom classifications that are trained by the user.

Sentiment Analyzes the sentiment (positive, negative, or neutral), and supports custom sentiments that are based on a user-trained model.

Emotion Extracts emotions (joy, anger, sadness, fear, and so on) that are conveyed by specific target phrases, or based on the document as a whole.

Keywords Identifies key terms in the document, which has various uses, such as indexing data, generating word clouds, or to improve the search function.

Summarization Extracts the most relevant snippets of a specific section to form a succinct summary, which enables IBM Watson Discovery to provide a short answer out of a long passage from a document.

IBM Watson Discovery supports 27 languages and can be extended with customization to enrich documents in other languages. After IBM Watson Discovery is finished enriching a customer's data and documents, knowledge workers can rapidly use answers from complex business documents, regardless of form (including tables, factoids, narrative generation, and charts), which enables them to solve their business problems.

For more information about the Watson IBM Discovery service on Cloud Pak for Data, see this IBM Documentation web page.

IBM Watson Discovery prerequisites

Before getting started with the use of IBM Watson Discovery, a Cloud Pak for Data administrator must install the IBM Watson Discovery service. For more information about how to install the service, see this IBM Documentation web page.

For more information about the recommended cluster size for IBM Watson Discovery, including system requirements and dependencies, see this IBM Documentation web page.

The minimum sizing and requirements are adequate for the hands-on scenario in this chapter. For a production-ready sized instance of IBM Watson Discovery, the following key criteria must be met:

- 23 vCPU and 150 GB RAM.
- CPUs support the AVX2 instruction set.
- Production instances have multiple replicas per service.
- Only one installation in a single namespace and only one instance per installation is allowed. However, multiple installations in separate namespaces are allowed.

Work with IBM Sales for more accurate sizing that is based on your expected workload.

IBM Watson Discovery Instance

After you install IBM Watson Discovery, you must provision an instance of the service to begin using it. In a multitenant environment of Cloud Pak for Data, you install the IBM Watson Discovery service once and then, you can deploy up to 10 separate instances of the service.

An instance of IBM Watson Discovery can be created by using the navigation menu to browse to the Services section and clicking **Instances**. Click **New instance** in the upper right of the window and select the **IBM Watson Discovery service tile** to create an instance of the service. Wait 1 - 2 minutes for the new instance to be provisioned.

For more information, see this IBM Documentation web page.

Hands-on experience with Watson Discovery

Several publicly available materials provide hands-on experiences with IBM Watson Discovery. Some are available as demonstrations within the IBM Cloud and others are GitHub repositories that can be used as examples of the use of a Watson Discovery deployment on Cloud Pak for Data.

Publicly available resources with examples are listed in Table 6-4.

Table 6-4 IBM Watson Discovery examples

Description	Location
Process complex insurance documents with IBM Watson Discovery	https://github.com/IBM/virtual-insurance-assistant
Add Watson Discovery skills to your IBM Watson Assistant chatbot	https://developer.ibm.com/tutorials/add-watson-discovery-skills-to-your-watson-assistant-chatbot/
Create an intelligent search application by using IBM Watson Discovery UI components	https://developer.ibm.com/tutorials/create-an-intelligent-search-app-using-watson-discovery-ui-components/
Process, understand, and answer policy questions with smart documents understanding	https://developer.ibm.com/tutorials/analyze-and-answer-policy-questions-with-smart-document-understanding/
Deploy an IBM Watson Discovery application with New Relic observability	https://developer.ibm.com/tutorials/deploy-a-simple-app-on-kubernetes-that-serves-a-web-app-and-communicates-with-watson-discovery/
Use of discovery components	https://github.com/watson-developer-cloud/discovery-components

For more information about how to use the IBM Watson Discovery service, see this IBM Documentation web page.

7

Business analytics

This chapter provides an overview of business analytics on Cloud Pak for Data and includes example use cases.

This chapter includes the following sections:

7.1 Overview

Business analytics is a set of data analytics practices and tools that organizations can use to gain insight from their data, enable data-driven decision making, and plan for the future. Business intelligence is a subset of business analytics, and involves developing reports and dashboards by using historical data.

Business analytics encompasses for key areas: descriptive analytics, diagnostic analytics, predictive analytics, and prescriptive analytics. These four areas often overlap in implementation. Figure 7-1 highlights the key questions that are answered by each of these four areas.

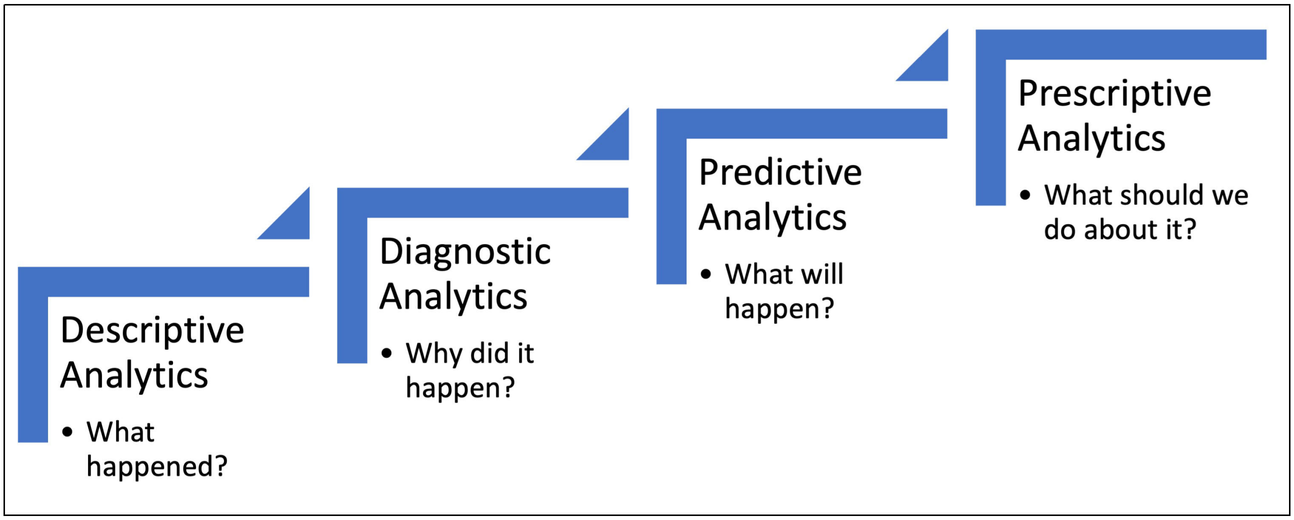

Figure 7-1 Four areas of business analytics

7.1.1 Descriptive analytics

Descriptive analytics involves summarizing historical data, highlighting key findings, and displaying interesting patterns and trends. These use cases answer several questions, such as “What happened?” and “What have we done in the past?”

Reports and dashboards are common descriptive analytics products. For example, an insurance company can create a dashboard that displays the total number of claims over a period, or a financial services company can develop a dashboard to visualize trades processed during a period. These dashboards highlight key features of these claims or trades that are based on historical data.

7.1.2 Diagnostic analytics

Diagnostic analytics builds on descriptive analytics. Key features of diagnostic analytics include the ability to drill down into a data set and data discovery. These features are commonly found in business intelligence tools and are incorporated into dashboards.

Diagnostic analytics use cases answer questions such as: “Why did this happen?” and “Why have we seen past results?”

Returning to the claims dashboard example from the previous section, the claims team can use a business intelligence tool to further drill down into the data. For example, although the dashboard might show the total number of claims for a specific year, a user can drill down into the data by using filters and the suitable BI tool to review claims that were created based on a specific event, such as a natural disaster.

7.1.3 Predictive analytics

Predictive analytics involves predicting outcomes about what is likely to happen. Use cases in this space answer questions, that are similar to “What will happen?” or “What is likely to happen?”

Predicting a numerical or categorical value is a common goal among predictive analytics use cases. In predictive analytics, regression problems are typically where the target you are predicting is a numerical value, while classification problems are where the target you are predicting is a categorical value. A common regression problem in predictive analytics is predicting the price of a house, while a common classification problem is predicting if a loan will default.

7.1.4 Prescriptive analytics

Prescriptive analytics builds on predictive analytics. The goal here is to provide recommendations on a next course of action. “What should we do about it?” is a key question that is answered in prescriptive analytics use cases.

Prescriptive analytics is common in the retail industry. Many retailers, such as Amazon, create tailored product recommendations for customers that are based on previous purchase history.

7.1.5 Roles and processes

A few standard processes are available that can be applied when developing business analytics use cases. The Cross Industry Standard Process for Data Mining (CRISP-DM) and the IBM Analytics Solutions Unified Method (ASUM) are two common processes in the analytics space that are specific to predictive analytics use cases.

Business analytics is an iterative process, and generally consists of the following steps:

1. Understand the business.

   Develop an understanding of the business objective. Determine the goals of the user, the overall business value of the analysis, and how stakeholders want to access the results.

2. Collect and understand the data.

   Identify what data is needed, where it is, and develop a business understanding of the data that is available for analysis.

3. Prepare the data.

   Clean and combine the data that is needed for analysis.

4. Analyze the data.

   Analyze the data to answer the business question. The output of this step can result in an ad hoc analysis, report, dashboard, or model, depending on the business question and need.

5. Interpret and share the results.

   Evaluate and interpret findings with stakeholders to enable data-driven decision making. Obtain feedback on the analysis and determine whether more steps are required.

6. Repeat.

   Make updates to the data that are based on feedback that is received from stakeholders.

Again, because business analytics is an iterative process, the amount of time that is spent on each step, and the number of repetitions through this process, varies by project. Each step in this process involves different technical and nontechnical roles.

The roles that are involved in any business analytics project can vary based on the use case, organization, and overall data maturity of the team. Common roles include business users, stakeholders, data analysts, data scientists, data engineers, and business analysts.

Table 7-1 lists the common roles in business analytics projects and the responsibilities of each role. Responsibilities that are performed by each role vary by team and project.

Table 7-1 Common roles in business analytics projects

Role	Responsibilities
Business User and Stakeholders	Define business goal and need
Business Analyst	Collaborate with stakeholders to define requirements
Data Analyst	Develop reports and dashboards
Data Scientist	Develop models by using statistical techniques and machine learning algorithms
Data Engineer	► Prepare data for operational use ► Build data pipelines

7.1.6 Common challenges

Teams today can encounter various challenges when they are seeking to develop and implement business analytics solutions within their organization. These common challenges include the following examples:

- Analyzing and visualizing disparate data sources

  Organizations have a wide range of data that can be used for reporting and analysis; however, it is often spread across many systems and applications. Identifying and aggregating the correct data sets for analysis takes time and effort.

- Poor user adoption

  The outcome of a business analytics use case can be presented in various formats, including reports, dashboards, stories, or ad hoc analysis. At times, the products that are developed and insights that are gained from these use cases are not fully adopted by the user. Common reasons for not adopting these products and insights include poor design and lack of integration into workflows and processes.

- Platform and tools

  Before any business analytics initiatives can be taken, teams must decide what platform and tools they are to use to enable their analysis. Several factors must be considered when selecting a platform and tools for business analytics, including cost, ease-of-use, and available features and capabilities.

For more information about how these challenges can be addressed with the Cloud Pak for Data platform, see 7.2, “Business analytics on Cloud Pak for Data” on page 487.

7.2 Business analytics on Cloud Pak for Data

In this section, we discuss business analytics on Cloud Pak for Data.

7.2.1 Cloud Pak for Data business analytics advantages

When presented with the challenges that are common to business analytics objectives, decision makers typically have many product choices. Many organizations find themselves choosing multiple products and services to meet their needs, instead of using a single integrated toolset.

This decision often leads to data silos, a lack of integration between products, complications across the business analytics stack, poor user adoption, and eventually inferior outcomes.

Cloud Pak for Data resolves these central challenges by integrating the best-in-class IBM Business Analytics tools within a powerful, easy-to-use environment. The use of the trio of Cognos Dashboards, Cognos Analytics, and Planning Analytics services enables users to take advantage of their respective features while also benefiting from the complete Cloud Pak for Data toolkit, from management of disparate data sources to deployment of analytical results.

This integration between business analytics tools is complemented by the project management and data preparation, management, and governance capabilities of the Cloud Pak for Data interface. The resulting environment drives user adoption, increases productivity and product coherence, eliminates data silos, and produces great time to value.

7.2.2 Business Analytics Services overview

The following is an overview of Business Analytics Services.

Cognos Dashboards

The Cognos Dashboards service provides access to sophisticated visualization capabilities directly within the CP4D interface, which enables simple drag and drop to construct meaningful, communicative dashboards for decision makers.

By using this service, users can quickly build the dashboards that are necessary to answer important questions or provide the foundation for more in-depth analysis (see Figure 7-2).

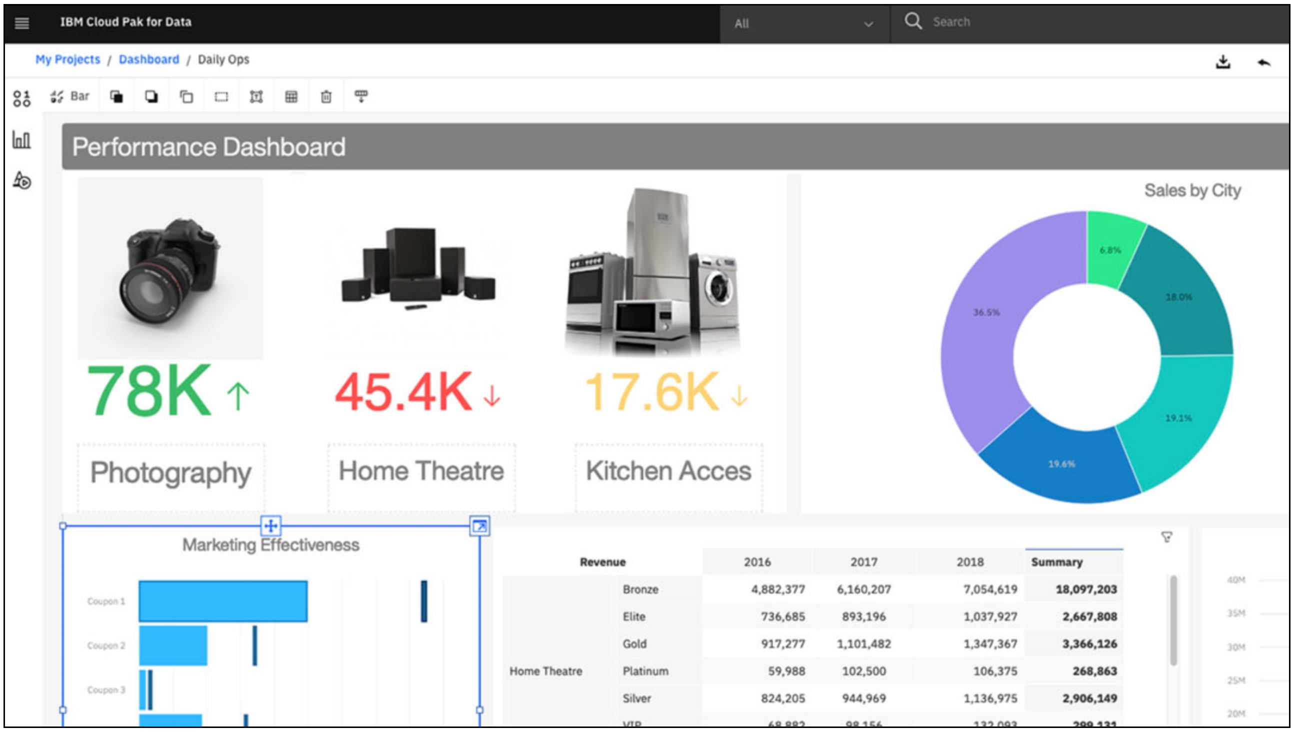

Figure 7-2 Sample Cognos Dashboards example

Various visualization layouts are available as templates on the dashboard canvas. These layouts enable assembling nearly endless combinations of interactive graphics for communication of comparisons, relationships, and trends in data.

Also, no coding or SQL is required, with all fields presented as options for use in individual components of the chosen template.

Getting started is as easy as adding a data asset in the project from one of these sources to a dashboard and then, immediately starting on visualizations.

A sample template for the dashboard canvas is shown in Figure 7-3 on page 489, which is ready for visualizations with just a few clicks.

For more information about how to use Cognos Dashboards for student performance analysis, representing an example of the work from data preparation to dashboard build of multiple roles within the project team, see 7.3.1, “Use case #1: Visualizing disparate data sources” on page 496.

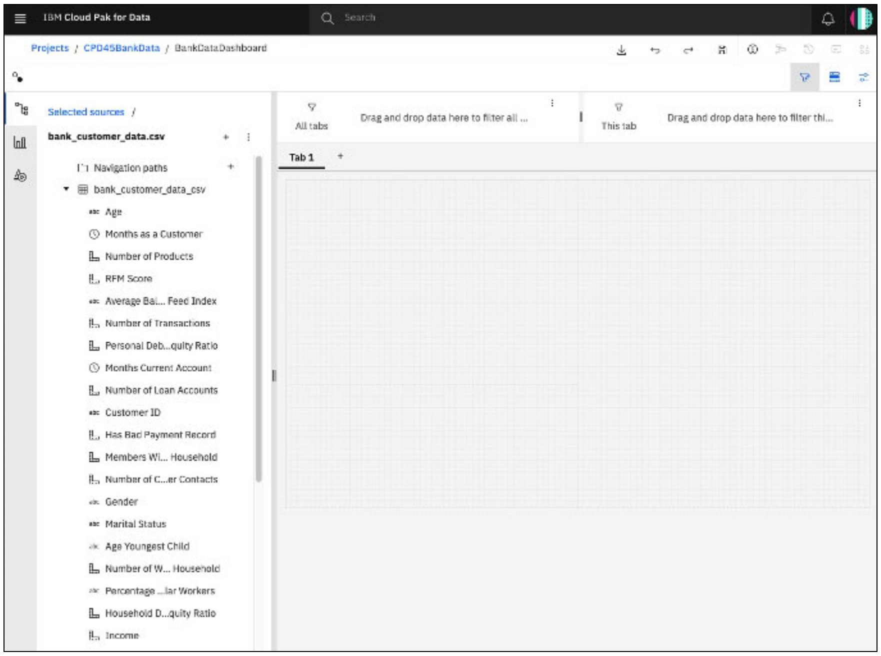

Figure 7-3 Sample blank dashboard, illustrating simple drag visualizations from your data

After the dashboards are created, they can be managed, edited, and shared as part of the overall project structure within IBM Watson Studio. Observed trends or interesting factors can be instantly shared with project collaborators for deeper analysis, or the dashboard can be deployed as a final product for distribution.

Cognos Dashboards can be even more effective when they are used with the complementary IBM Watson Studio, IBM Watson Knowledge Catalog, Data Refinery, and Data Virtualization features to manage the project and prepare the data.

Cognos Dashboards are compatible with various data formats in Cloud Pak for Data, including:

- CSV files
- IBM Data Virtualization tables
- IBM Db2
- Db2 Warehouse
- DB2 on Cloud
- IBM Cloud Databases for PostgreSQL
- PostgreSQL
- Microsoft SQL Server
- IBM Netezza® Performance Server

Because of the compatibility with Data Virtualization, other sources that are supported by Data Virtualization also are supported by proxy.

Note: The Cognos Dashboards service is not available by default within Cloud Pak for Data; it must be provisioned by an administrator.

For more information about the Cognos Dashboards service on Cloud Pak for Data 4.5, see this IBM Documentation web page.

Cognos Analytics

The Cognos Analytics service in Cloud Pak for Data provides a complete AI-infused, self-service solution for dashboards and reporting. With Cognos Analytics, stunning visualizations are created with help from the AI Assistant and a long lineage of proven, dependable analytics that enable decision makers to have the most important and compelling information at their fingertips.

Integrated into the Cognos Analytics toolset are reporting, dashboards, stories, modeling, analysis, and event management. These features provide an end-to-end personalized analytics experience that is driven by decades of leadership in the Business Analytics space.

Because of the long lineage of Cognos, an active IBM Cognos Analytics community exists with many resources for users to use their results, particularly with the modern AI components. Functions even include Natural Language Processing technology to help explore textual data and suggest directions for analysis.

Cognos Analytics is started from the Instances list in the Cloud Pak for Data interface by opening the Cognos Analytics instance, as shown in Figure 7-4.

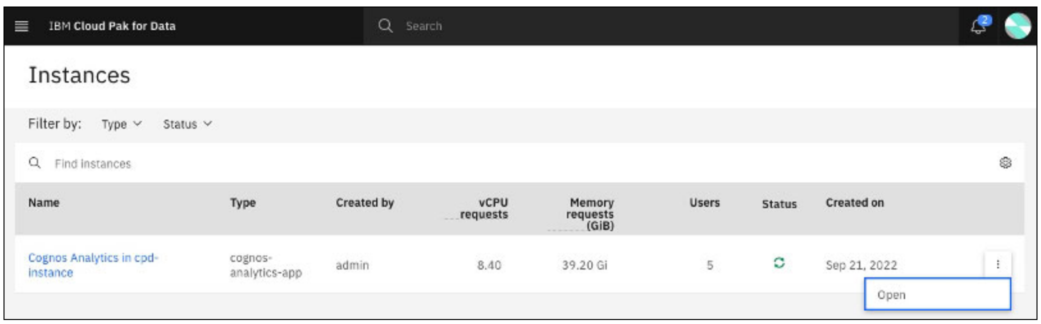

Figure 7-4 Instances list in Cloud Pak for Data

The Cognos Analytics service in Cloud Pak for Data opens in a separate browser window, as shown in Figure 7-5. Users can immediately use the Quick Launch functions by uploading data, such as spreadsheets, flat files, and various other data sources, including connections to Cloud Pak for Data data sources.

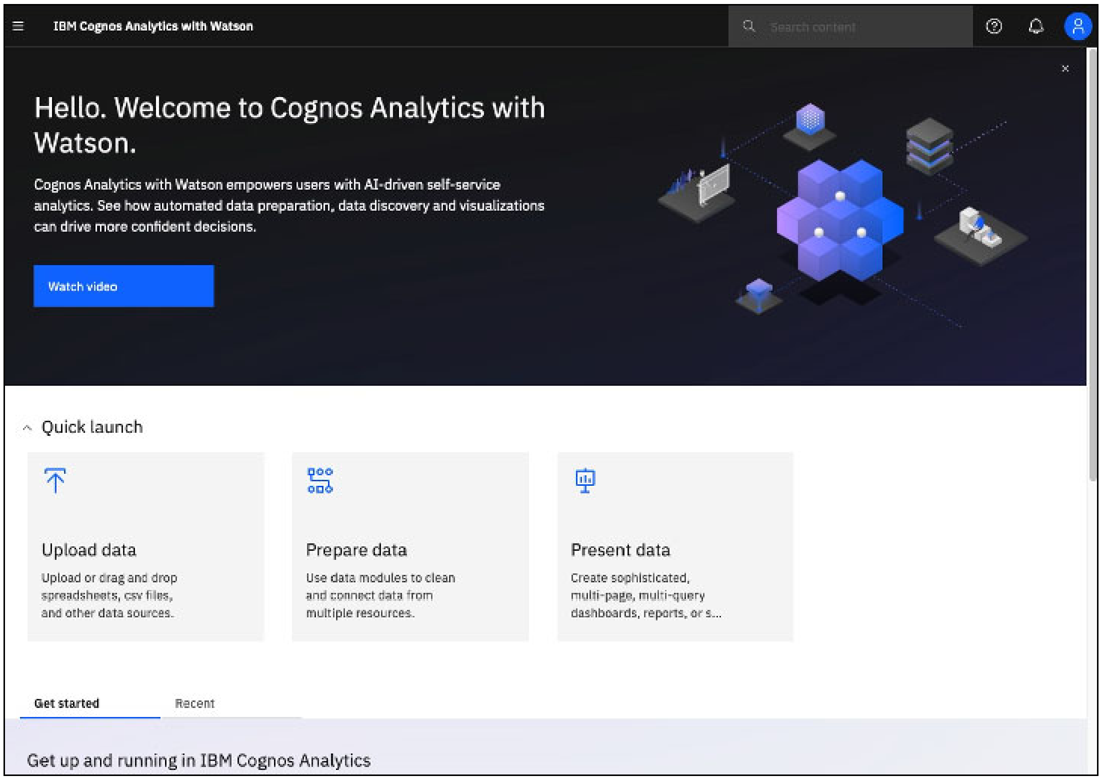

Figure 7-5 Cognos Analytics service home page

Users can get started with a sample bank data CSV data file by clicking **Upload data** and then, clicking **New Dashboard**. A blank canvas opens onto which visualizations can be added in just a few clicks within the interface. (This process is similar to the Cognos Dashboards that were introduced in “Cognos Dashboards” on page 487).

The result is shown in Figure 7-6 on page 492.

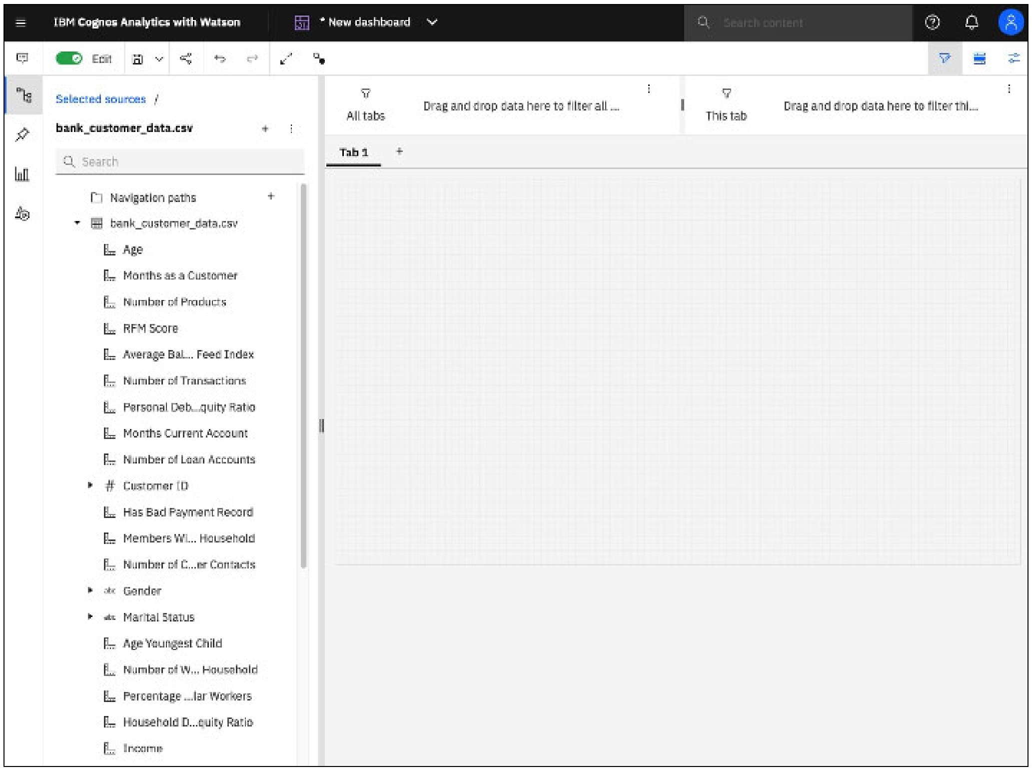

Figure 7-6 Cognos Analytics Dashboard

The results of the use of the Cognos Analytics service include stunning, functional dashboards, stories, and reports that are ready for sharing across organizations. An example of a result that uses one of the myriad visualization methods along with the Explore function in the service is shown in Figure 7-7. This example incorporates the AI Assistant and its natural language exploration capabilities, as shown in the left side panel of the window, and annotations to help the user.

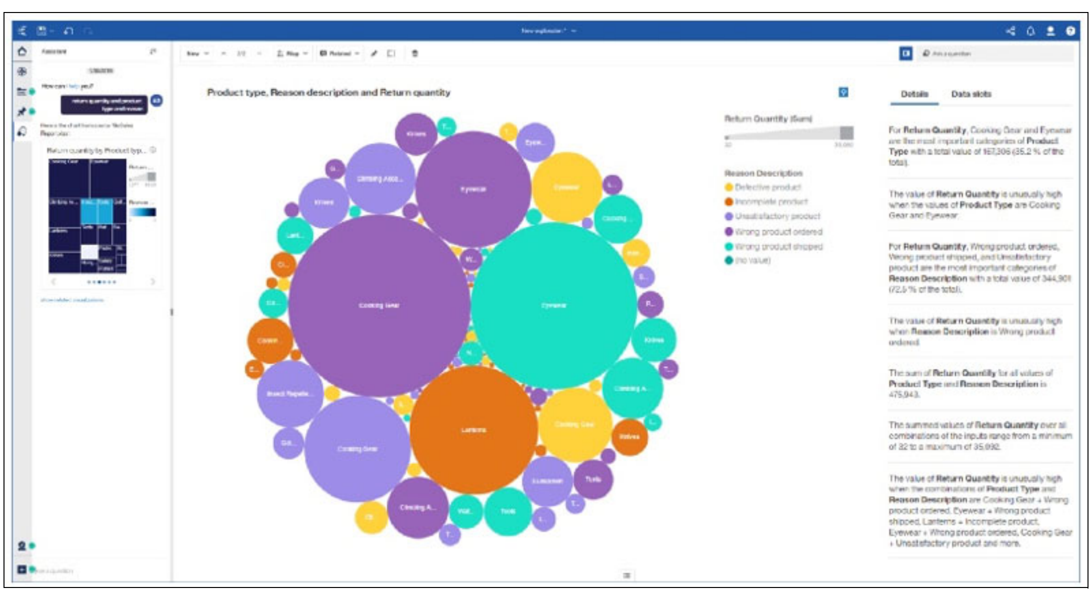

Figure 7-7 Cognos Analytics full-featured sample visualization

Cognos Analytics supports a broad range of data sources, including the following examples:

- Data modules: Data modules contain data from data servers, uploaded files, data sets, other data modules, and from relational, dynamic query mode packages.
- Packages: A package is a subset of a model (up to the whole model) that is made available to the Cognos Analytics application.
- Data sets: Data sets are frequently used collections of data items. As updates are made to the data set, the dashboards, stories, or explorations that use that data set also are updated the next time that they are run.
- Uploaded files: For some quick analysis and visualizations with data files, users can manually upload the files to IBM Cognos Analytics. Data files must meet size and structure requirements.

Although every role, from decision maker to data scientist, can take advantage of the functions within Cognos Analytics (particularly with the AI Assistant and complementary Cloud Pak for Data services to help), the depth and breadth of the tool encourages collaboration between these roles to take full advantage of the service.

For more information about examples of the use of Cognos Analytics and complementary Cloud Pak for Data services to develop reports and answer questions for decision makers, see 7.3.1, “Use case #1: Visualizing disparate data sources” on page 496, 7.3.2, “Use case #2: Visualizing model results” on page 522, and 7.3.3, “Use case #3: Creating a dashboard in Cognos Analytics” on page 534.

These use cases explore how multiple team roles, from data engineer and data scientist on to the analysts and decision makers, can be used to deliver the results.

> **Note:** The Cognos Analytics service is not available by default within Cloud Pak for Data; it must be provisioned by an administrator.

For more information about the Cognos Analytics service on Cloud Pak for Data 4.5, see this IBM Documentation web page.

Planning Analytics

The Planning Analytics service in Cloud Pak for Data provides continuous, integrated planning and reporting with organizational data. AI-infused, Planning Analytics goes beyond manual planning and enables users to break down organization silos and efficiently create more accurate plans and forecasts.

As a result, users can pivot in real time with current data, optimize forecasts, plan continuously, scale when needed, and deploy where needed.

Planning Analytics users can take advantage of several best-in-class modern features, including a Microsoft Excel interface for embracing your spreadsheets, and conducting “what-if” scenario testing, in addition to the robust analysis, dash boarding, and reporting tools.

To start the Planning Analytics service, users start in the Cloud Pak for Data interface instances list and open the Planning Analytics Instance, as shown in Figure 7-8.

Figure 7-8 Cloud Pak for Data Instances list

The interface for Planning Analytics on Cloud Pak for Data opens in a new browser tab, and is similar in structure to Cognos Analytics, as shown in Figure 7-9. Users can immediately create applications and plans, reports, dashboards, and analyses after they are familiar with the data structure that is present in TM1. TM1 is the complementary database and calculation engine that is required for the Planning Analytics installed service to be provisioned as an instance within Cloud Pak for Data.

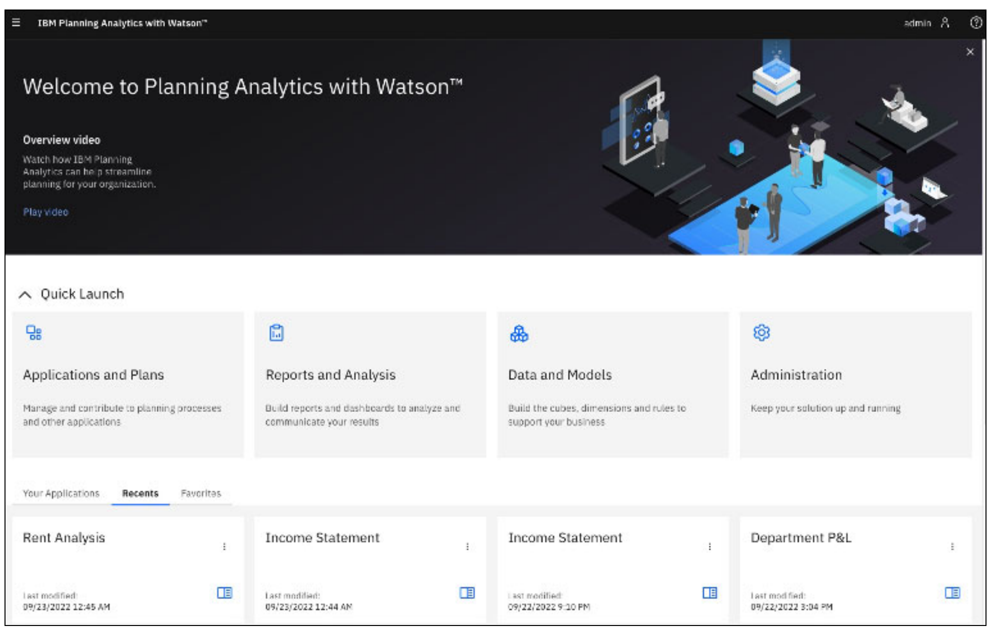

Figure 7-9 Planning Analytics service home page

As a powerful, full-featured tool, Planning Analytics typically requires multiple team members in multiple roles to take full advantage of its capabilities. For more information about an example that starts with pre-built database objects for this reason, see 7.3.4, "Use case #4: Planning Analytics" on page 552.

Typically, the task of developing the underlying TM1 data model is placed with a dedicated TM1 database developer who works with the Planning Analytics analysts and decision makers to help understand the cubes, dimensions, and other data elements in the database.

After this data understanding is achieved, essential visualizations are readily constructed within the service, such as the example that is shown in Figure 7-10. In this example, an interactive dashboard helps decision makers understand budgetary summaries for their organization and enables them to perform “what-if” analyses.

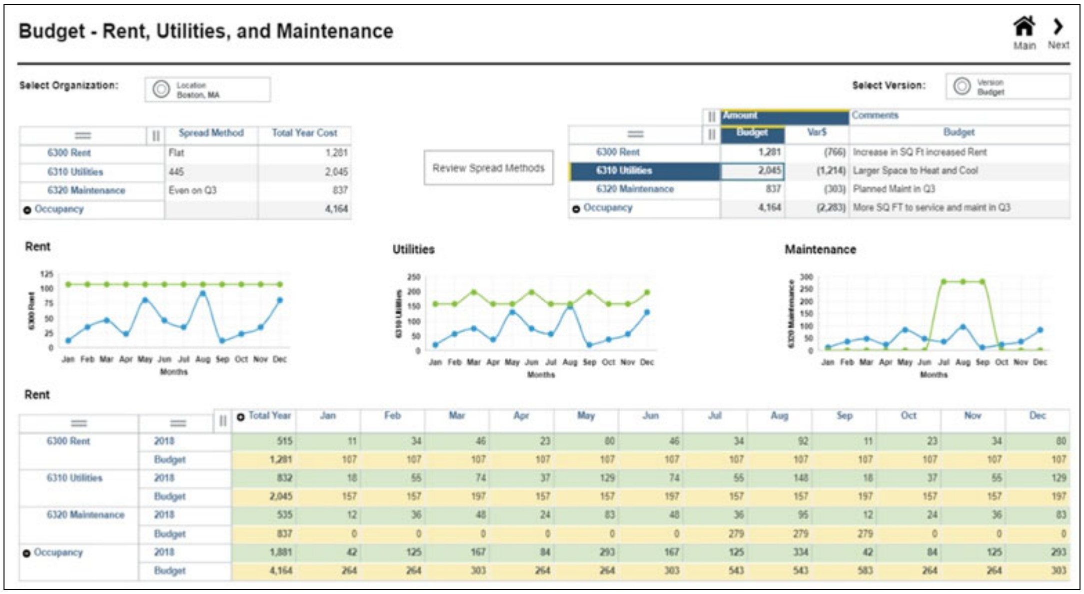

Figure 7-10 Sample Planning Analytics interactive dashboard for forecasting and what-if analysis

A new version (which is under technical preview as of this writing) of the Planning Analytics service, the Planning Analytics Engine, is the next generation Planning Analytics database that removes the need for a separate TM1 server.

Planning Analytics Engine is enterprise class, cloud-ready, and available exclusively on the Cloud Pak for Data platform. By using the Docker and Kubernetes container infrastructure, Planning Analytics Engine runs as a service on public and private clouds.

Planning Analytics Engine includes several new features, including the following examples:

- Database as a service: Planning Analytics Engine runs as a service and enables you to manage all of your Planning Analytics Engine databases through a single service endpoint.
- High availability: Planning Analytics Engine can run individual databases in High Availability mode. When running in High Availability mode, the service manages multiple replicas of the database in parallel, which ensures that all changes are propagated to all replicas while dispatching requests in such a way as to spread the load on the overall system.
- Horizontal scalability: Planning Analytics Engine allows the number of replicas of any database to be increased or decreased without any downtime. This ability allows customers to scale up during peak periods and scale down during quiet times without any interruption to users.

> **Note:** The Planning Analytics service is not available by default within Cloud Pak for Data; it must be provisioned by an administrator.

For more information about the Cognos Analytics service on Cloud Pak for Data 4.5, see this IBM Documentation web page.

7.3 Use cases

In this section we, cover four business analytics use cases.

7.3.1 Use case #1: Visualizing disparate data sources

Analyzing and visualizing disparate data sources is a common challenge that teams encounter when developing business analytics solutions. This use case shows how users can create dashboards on Cloud Pak for Data by using multiple data sources.

Specifically, this use case shows two ways a dashboard can be created on Cloud Pak for Data by using IBM Db2, Data Virtualization, IBM Watson Studio, Cognos Dashboards, and Cognos Analytics.

The following example uses the Student Performance data sets from the UCI Machine Learning Repository. This data can be downloaded from the UCI Machine Learning Repository. [1]

These data sets show student grades in Math and Portuguese for two high schools that are in Portugal. Each data set contains 33 attributes for each record, and each record represents one student. Student Math grades and student Portuguese grades are in two separate Db2 databases on Cloud Pak for Data.

Table 7-2 lists the attributes in the data set.

Table 7-2 Attributes in Math and Portuguese student performance data sets

Column name	Column description	Sample values
school	Student’s School	GP (Gabriel Pereira), MS (Mousinho da Silveira)
sex	Student’s Gender	F, M
age	Student’s Age	15, 16, and so on
address	Home Address Type	U (Urban), R (Rural)
famsize	Family Size	LE3, GT3
Pstatus	Parent’s Cohabitation Status	T (living together), A (living apart)
Medu	Mother’s Education	0 - none 1 - primary education (4th grade) 2 - 5th - 9th grade 3 - secondary education 4 - higher education
Fedu	Father’s Education	0 - none 1 - primary education (4th grade) 2 - 5th - 9th grade 3 - secondary education 4 - higher education

[1] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Column name	Column description	Sample values
Mjob	Mother's Occupation	teacher, health care related, other, and so on
Fjob	Father's Occupation	teacher, health care related, other, and so on
reason	Reason to Choose this School	close to home, school reputation, other, and so on
guardian	Student's Guardian	Mother, Father, and so on
traveltime	Home to School Travel Time	1 - <15 min 2 - 15 - 30 min 3 - 30 min - 1 hr 4 - >1 hr
studytime	Weekly Study Time	1 - < 2 hours 2 - 2 - 5 hours 3 - 5 - 10 hours 4 - > 10 hours
failures	Number of Past Class Failures	1, 2, 3, 4
schoolsup	Extra Educational Support	Yes, No
famsup	Family Educational Support	Yes, No
paid	Extra Paid Classes within the Course Subject	Yes, No
activities	Extra-curricular Activities	Yes, No
nursery	Attended Nursery School	Yes, No
higher	Wants to Take Higher Education	Yes, No
internet	Internet Access at Home	Yes, No
romantic	With a Romantic Relationship	Yes, No
famrel	Quality of Family Relationships	1 (very bad) - 5 (excellent)
freetime	Free time after school	1 (very low) - 5 (very high)
goout	Going out with friends	1 (very low) - 5 (very high)
Dalc	Weekday Alcohol Consumption	1 (very low) - 5 (very high)
Walc	Weekend Alcohol Consumption	1 (very low) - 5 (very high)
health	Current Health Status	1 (very bad) - 5 (very good)
absences	Number of School Absences	0, 1, 2, ..93
G1	First Period Grade	0, 1, 2,...20
G2	Second Period Grade	0, 1, 2,...20
G3	Final Grade	0, 1, 2,...20

Step 1: Creating a Db2 connection in Data Virtualization

Complete the following steps to create a Db2 connection in Data Virtualization:

1. In Cloud Pak for Data, browse to **Data Virtualization** to create a connection to Db2. Figure 7-11 shows how to access Data Virtualization from the navigation menu.

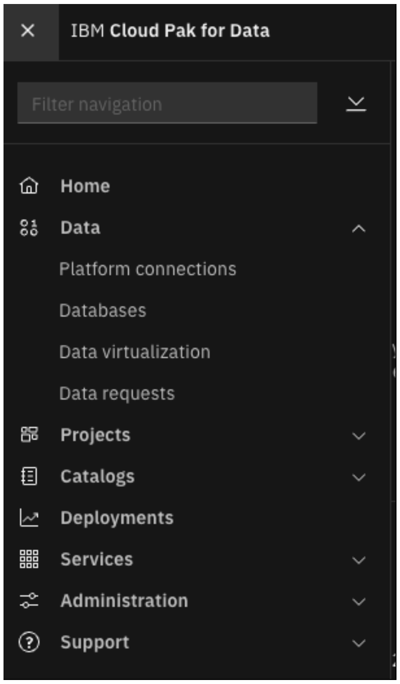

Figure 7-11 Access Data Virtualization from Navigation Menu

2. In Data Virtualization, add a connection by selecting **Add connection** on the right side of the window. Then, click **New connection** to add a data source connection (see Figure 7-12).

Figure 7-12 Adding a connection in Data Virtualization

3. In the next window, search for the `Db2 connection type` by using the search bar, as shown in Figure 7-13. From the connection options, select **Db2** and then, click **Select**.

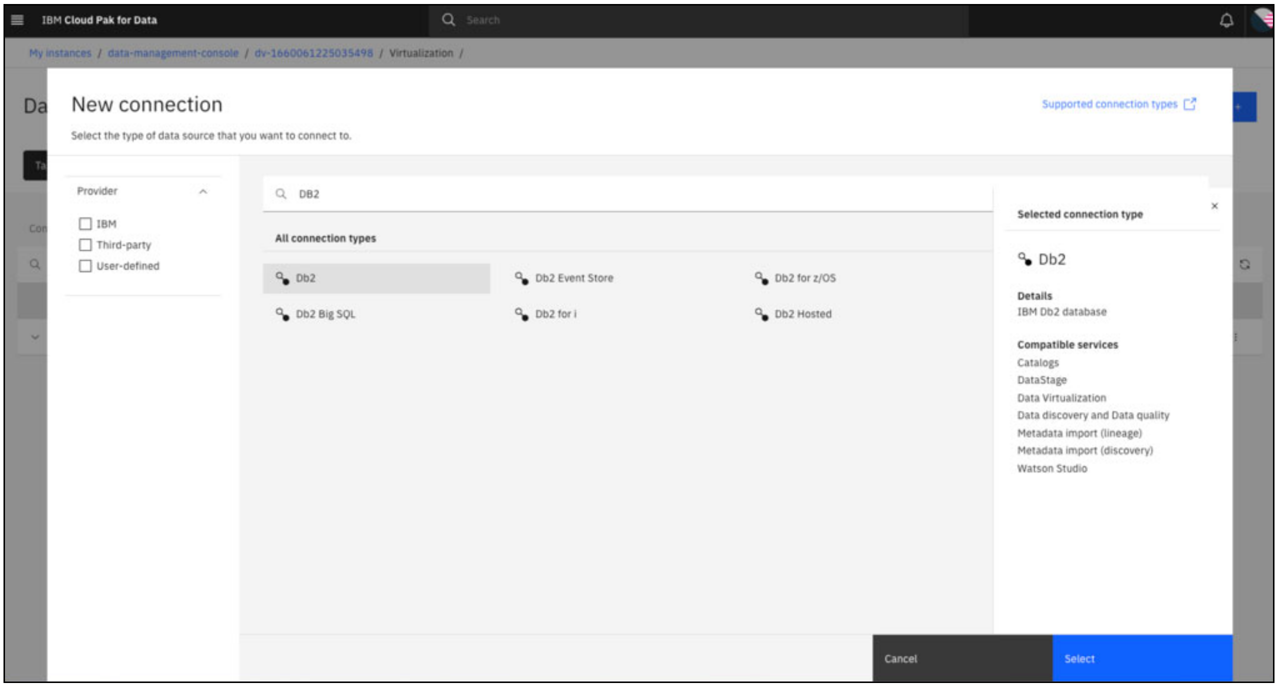

Figure 7-13 Selecting Db2 Connection Type

4. A form that includes the following sections must be completed: Connection overview, Connection Details, Credentials, and Certificates. Figure 7-14 shows the Connection overview section, with the Name and Description fields completed.

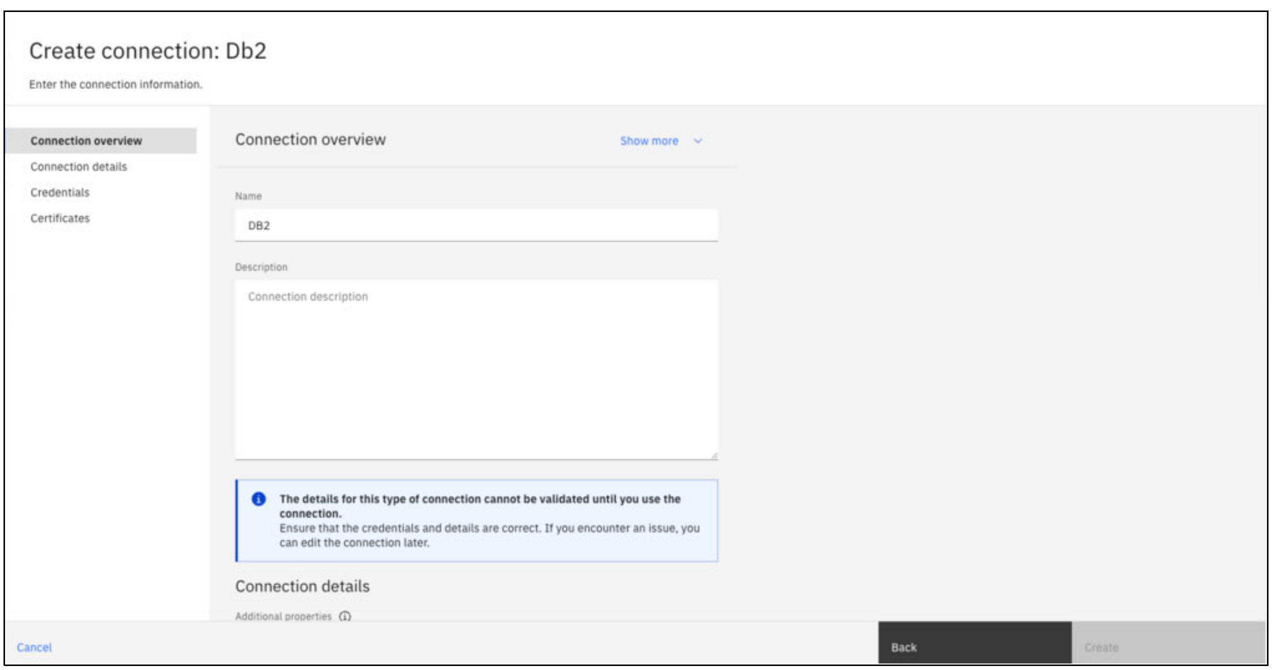

Figure 7-14 Connection overview details

5. In the Connection details section, enter the connection details for the Db2 database, including database name, hostname, port, and login credentials (see Figure 7-15).

Note: If Db2 on Cloud Pak for Data is used, enter the login credentials for Cloud Pak for Data that are used for the Db2 database.

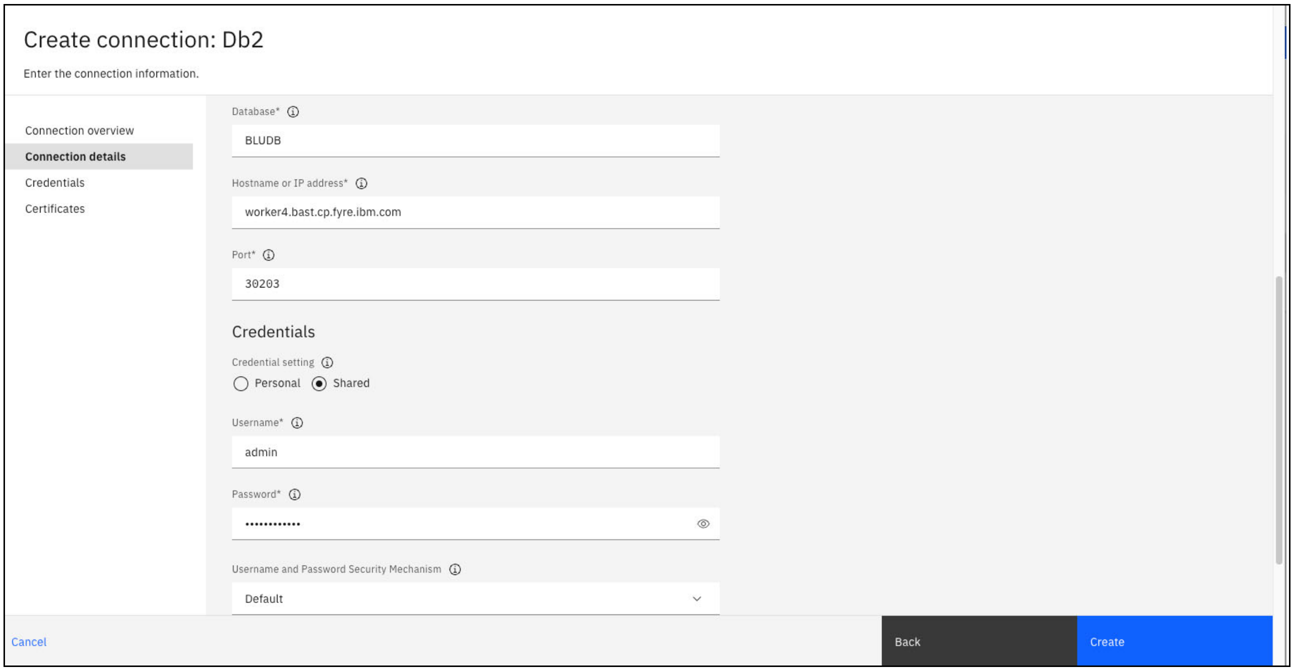

Figure 7-15 Connection and credentials details

6. When adding a connection, an option is available to specify whether the port is SSL-enabled, as shown in Figure 7-16. After all of the fields are completed, click **Create** to create the Db2 connection and continue.

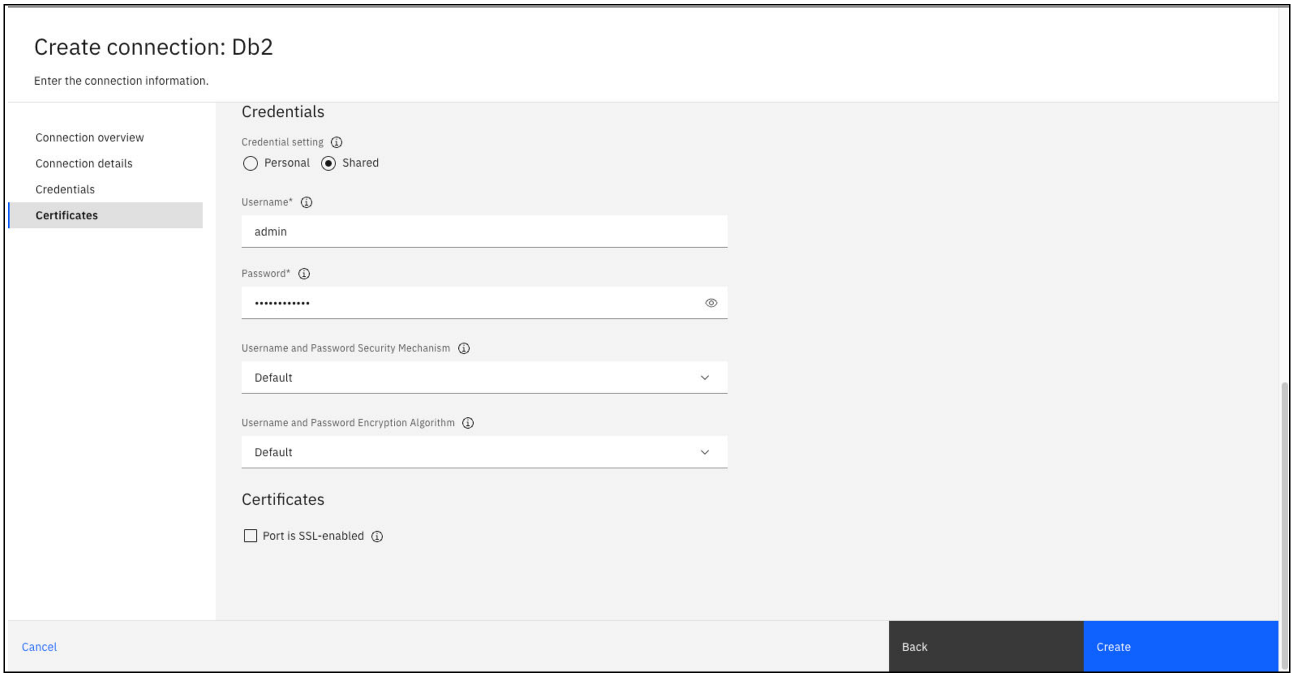

Figure 7-16 Certification details

7. In the next window (see Figure 7-17) you can add a remote connector, if wanted. Select **Skip** to continue without adding an optional remote connector.

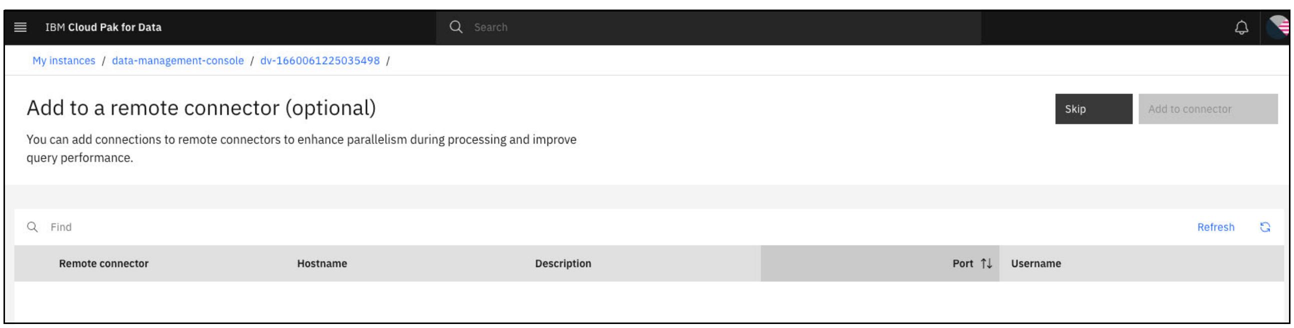

Figure 7-17 Adding a remote connector (optional)

After the connection is created, it appears under Data Sources in Data Virtualization. The new Db2 connection is shown in Figure 7-18.

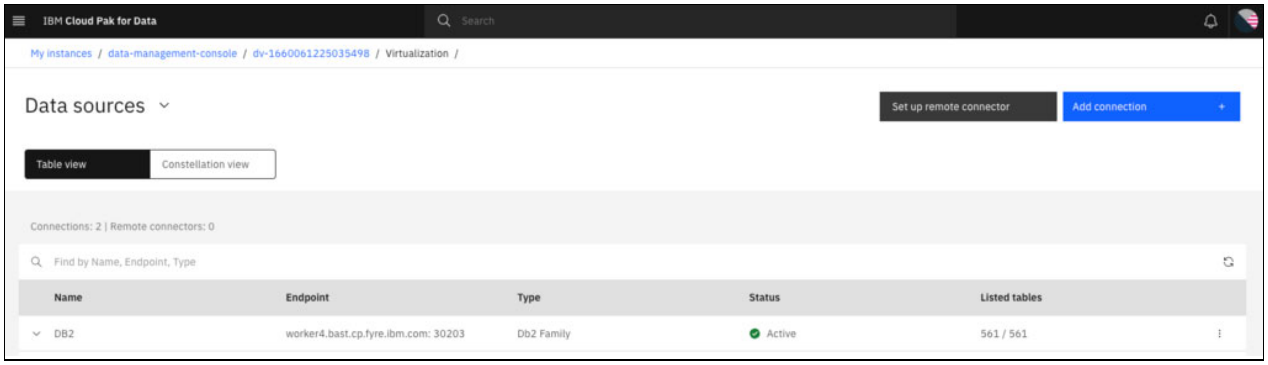

Figure 7-18 Db2 connection in Data Virtualization

Step 2: Virtualizing tables from Db2 database

After the connection to Db2 is created, complete the following steps to virtualize the tables that are needed for analysis:

1. Select **Virtualize** from the drop-down menu, as shown in Figure 7-19.

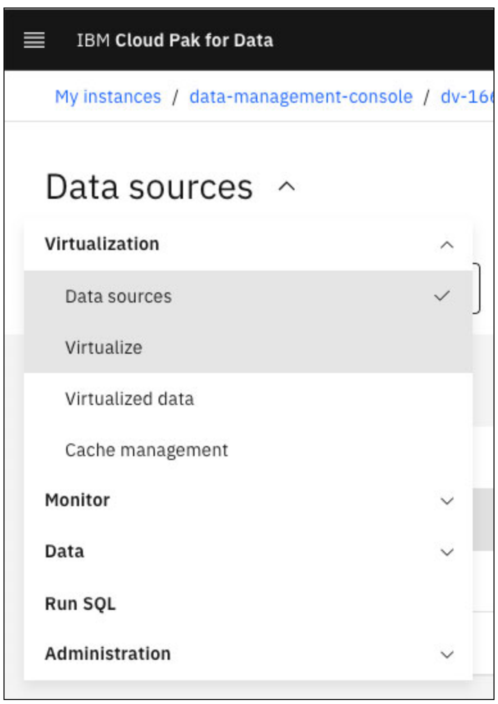

Figure 7-19 Data Virtualization menu options

2. On the Virtualize page, select the tables to virtualize by using the filters at the top of the GUI to help find your table quickly. Figure 7-20 shows that the Math table appears after filtering for the IBM DB2® database and the STUDENTS schema.

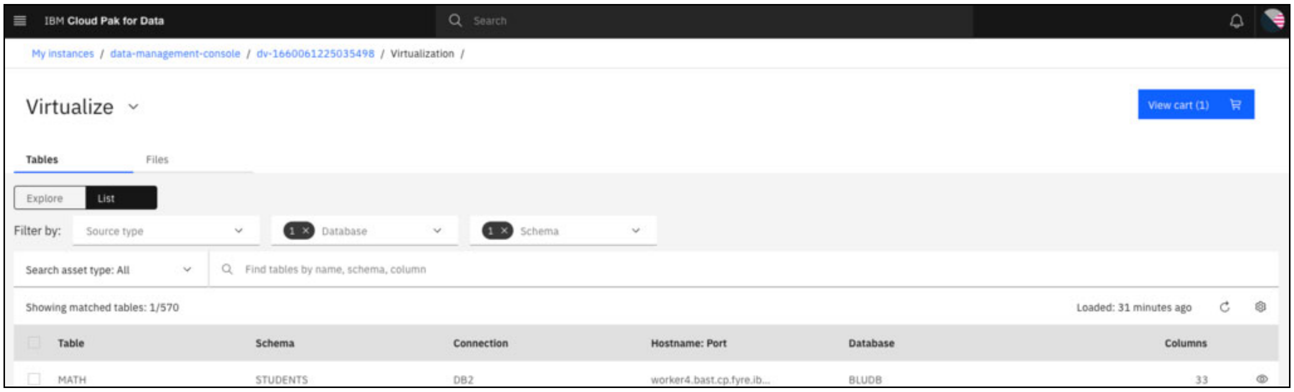

Figure 7-20 Virtualize page in Data Virtualization

3. After the tables are added to the cart, they can be virtualized. As shown in Figure 7-21, the table `MATH` was selected for virtualization and is assigned to `Virtualized data`. To virtualize the tables, select **Virtualize** in the upper-right corner.

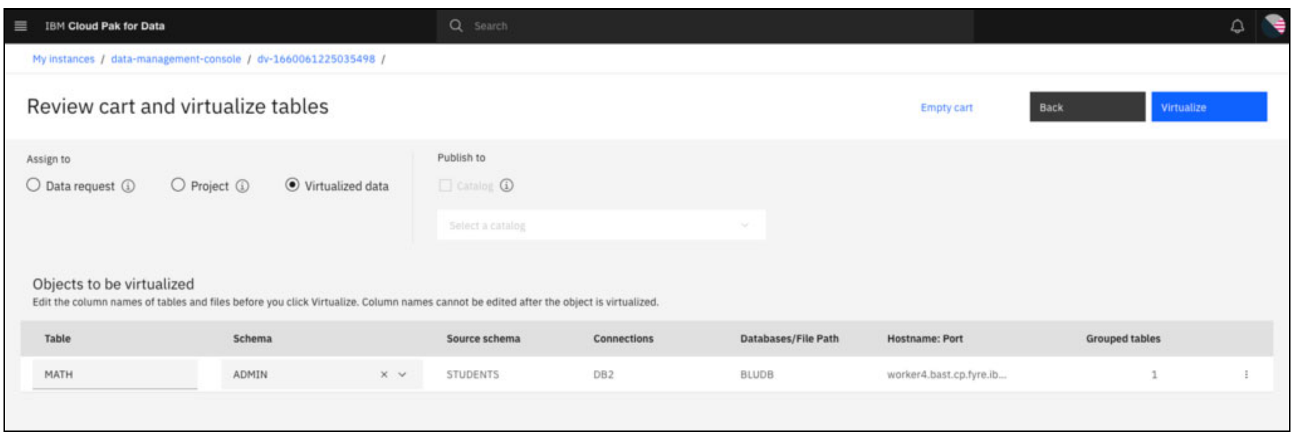

Figure 7-21 Review cart and virtualize tables window

4. After the tables are virtualized, they appear under **Virtualized data** in Data Virtualization. Figure 7-22 shows the `MATH` and `PORTUGUESE` virtualized tables.

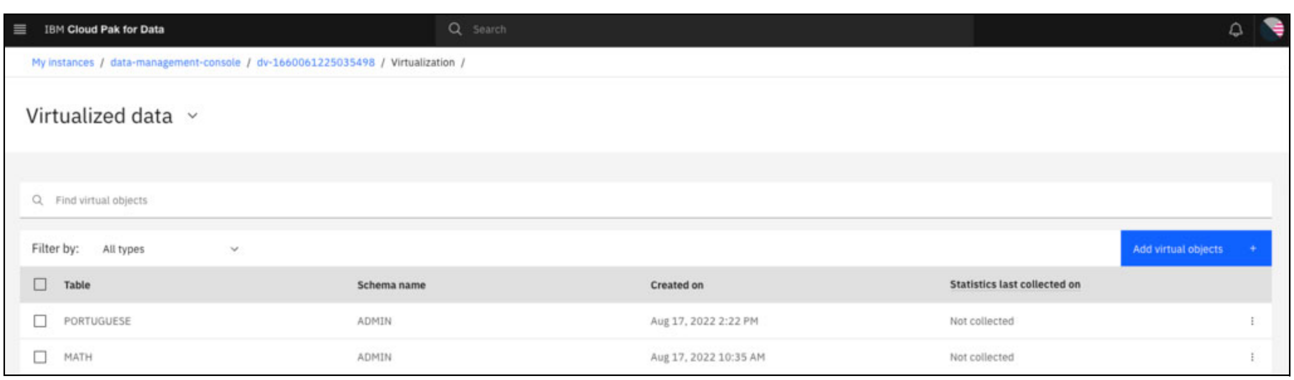

Figure 7-22 Virtualized data with MATH and PORTUGUESE tables

Step 3: Creating a virtualized view

Virtualized views can be created by using one or more tables. The Run SQL editor in Data Virtualization can be used to query virtualized tables and create views. Figure 7-23shows how Run SQL can be accessed from the Data Virtualization drop-down menu.

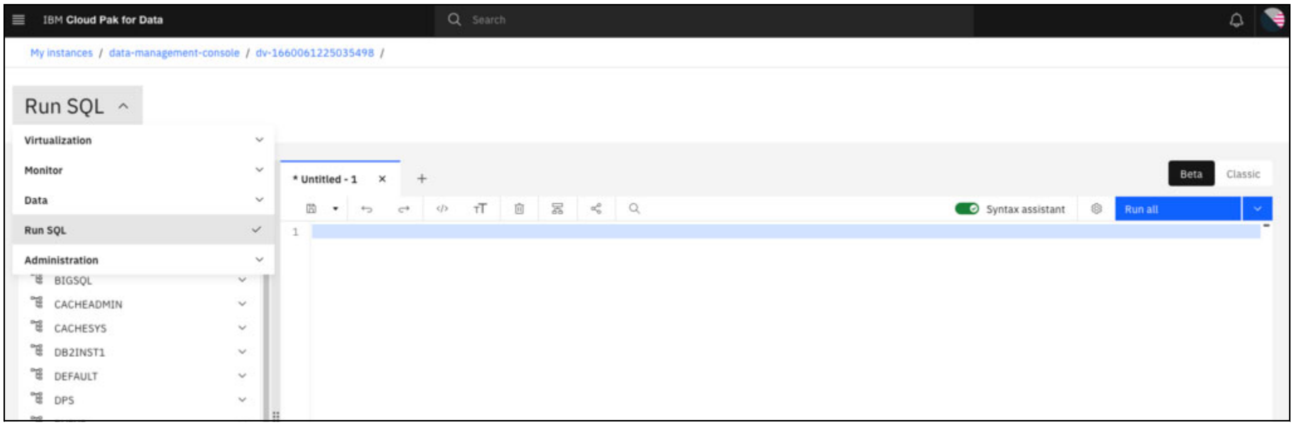

Figure 7-23 Running SQL Editor

Complete the following steps to create virtualized view:

1. Use the search bar in the Run SQL window to quickly find tables. As shown in Figure 7-24, SQL queries can be automatically generated for a table by right-clicking the table name and selecting from the available options.

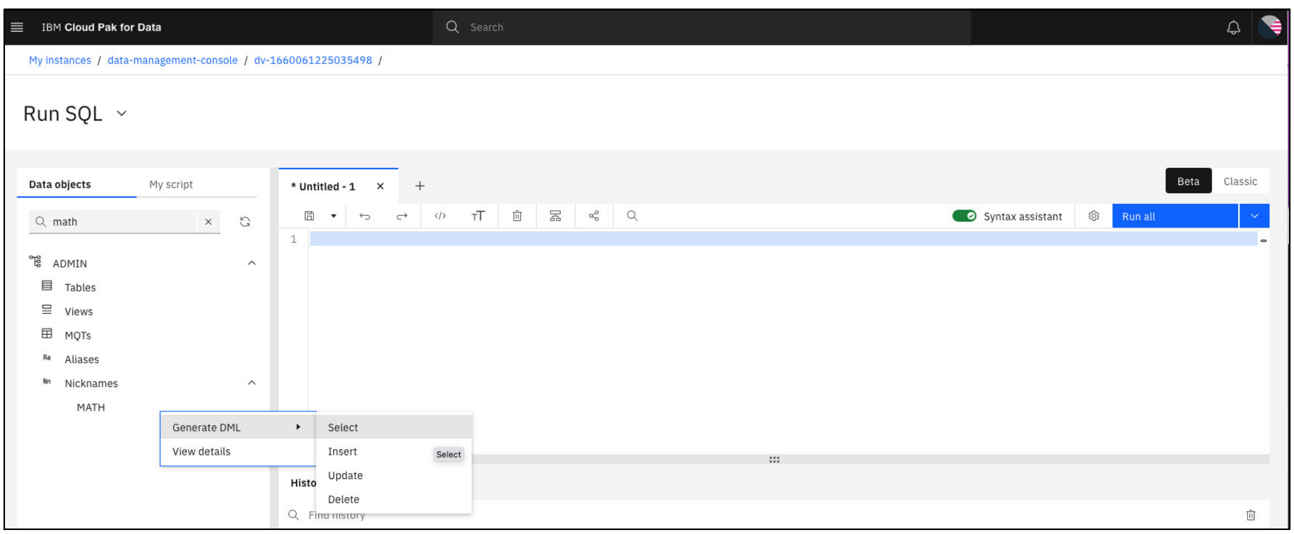

Figure 7-24 Auto-generated DML options that are available in Run SQL

2. In this example, the `MATH` table and `PORTUGUESE` tables are combined into one view for analysis. Example 7-1 shows the SQL code that is used in Run SQL to combine the tables into one view that is named `STUDENT_GRADES`.

Example 7-1 Creating view SQL statement that is used to create STUDENT_GRADES view

```
CREATE VIEW STUDENT_GRADES AS
SELECT "SCHOOL", "SEX", "AGE", "ADDRESS", "FAMSIZE", "PSTATUS", "MEDU", "FEDU",
"MJOB", "FJOB", "REASON", "GUARDIAN", "TRAVELTIME", "STUDYTIME", "FAILURES",
"SCHOOLSUP", "FAMSUP", "PAID", "ACTIVITIES", "NURSERY", "HIGHER", "INTERNET",
"ROMANTIC", "FAMREL", "FREETIME", "GOOUT", "DALC", "WALC", "HEALTH", "ABSENCES",
"G1", "G2", "G3"
FROM "ADMIN"."MATH"
UNION
SELECT "SCHOOL", "SEX", "AGE", "ADDRESS", "FAMSIZE", "PSTATUS", "MEDU", "FEDU",
"MJOB", "FJOB", "REASON", "GUARDIAN", "TRAVELTIME", "STUDYTIME", "FAILURES",
"SCHOOLSUP", "FAMSUP", "PAID", "ACTIVITIES", "NURSERY", "HIGHER", "INTERNET",
"ROMANTIC", "FAMREL", "FREETIME", "GOOUT", "DALC", "WALC", "HEALTH", "ABSENCES",
"G1", "G2", "G3"
FROM "ADMIN"."PORTUGUESE";
```

After a view is created, it can be accessed from the **Views** page in Data Virtualization, as shown in Figure 7-25.

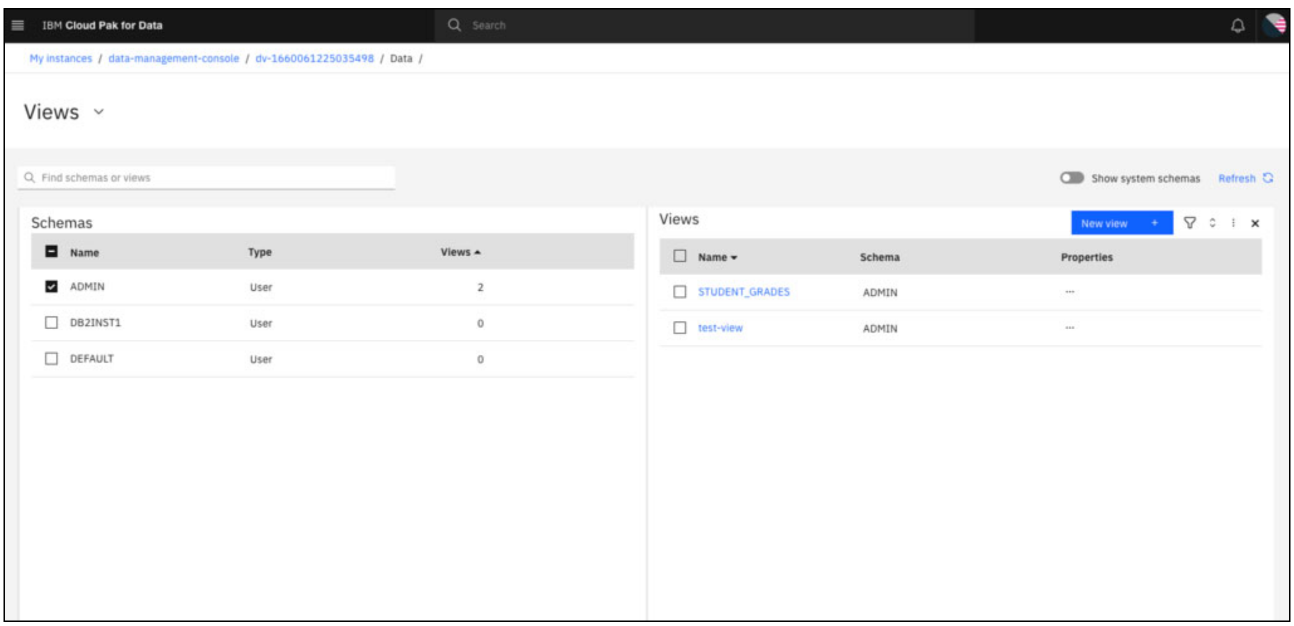

Figure 7-25 Views window

Step 4: Visualizing Virtualized view

In Cloud Pak for Data, a dashboard can be generated by using:

- ► Cognos Dashboards within IBM Watson Studio
- ► Cognos Analytics

If the Cognos Analytics service is available, users also can generate reports and stories. The following steps outline two different approaches to developing a dashboard by using a Data Virtualization view on Cloud Pak for Data.

Option 1: Visualize by using Cognos Dashboards

To create a dashboard by using a Data Virtualization view, the virtual object must be assigned to an IBM Watson Studio project. Complete the following steps:

1. To create a dashboard with the `STUDENT_GRADES` view, select this view from the **Virtualized data** page and then, select **Assign**, as shown in Figure 7-26.

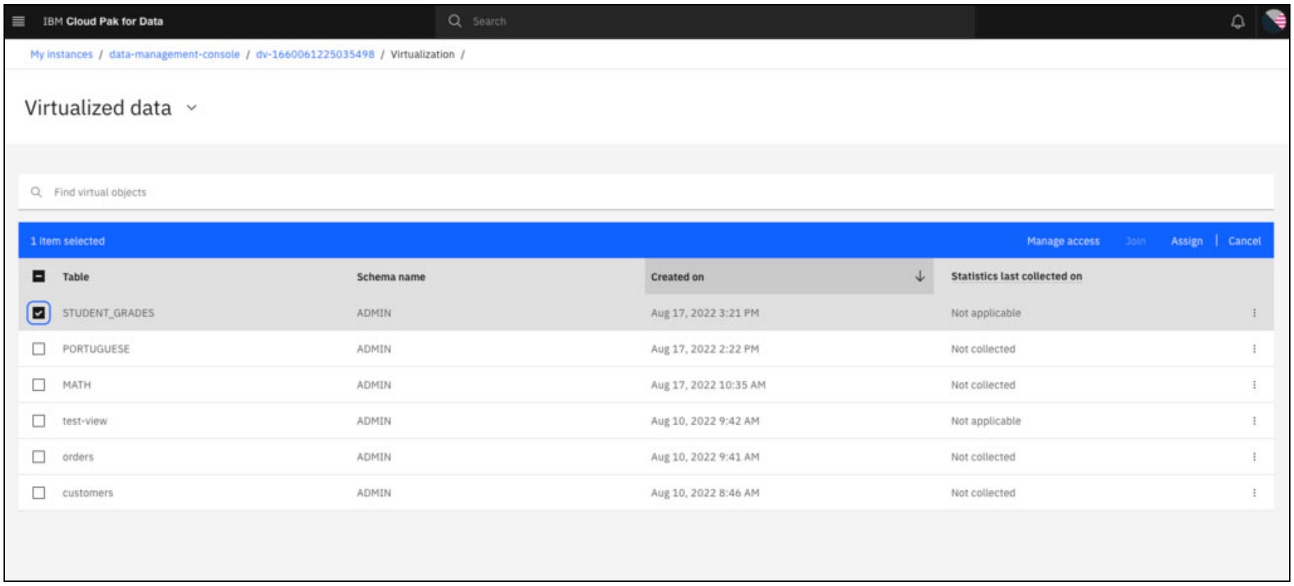

Figure 7-26 Virtualized data

2. In the next window, select **Project** and the **Project Name** to which the view must be assigned. In the student grades example, the **Student Performance Analysis** project is selected, as shown in Figure 7-27. After the suitable project is selected, select **Assign** to assign the view to the project.

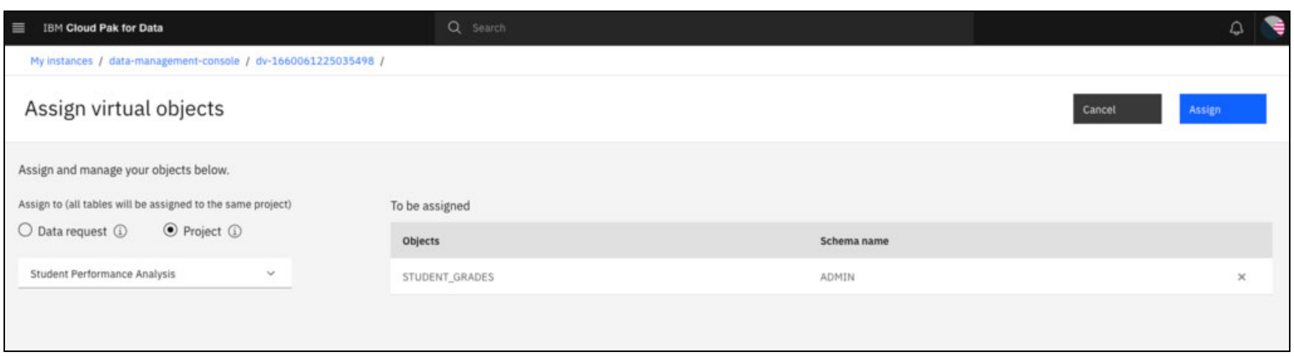

Figure 7-27 Assigning virtual objects to IBM Watson Studio project

3. After the view is assigned to a project, browse to that project in IBM Watson Studio. Select the **Assets** tab within the IBM Watson Studio project to see the view within the Data assets section. Figure 7-28 shows the `STUDENT_GRADES` view under Data assets in the Student Performance Analysis project.

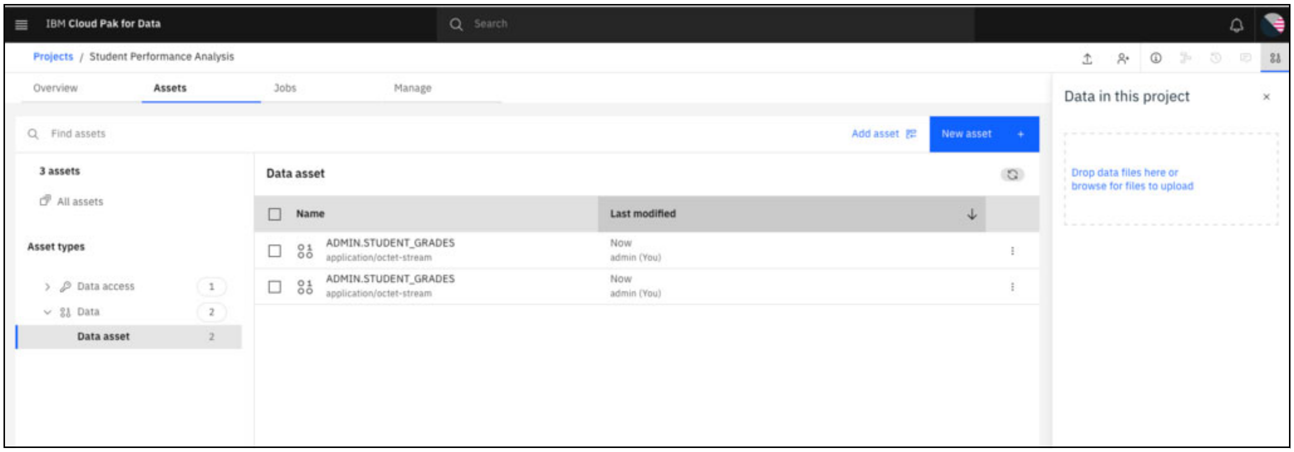

Figure 7-28 Data Assets within Student Performance Analysis Project

4. Create a dashboard from the Assets page by selecting **New asset** in the upper-right corner of the GUI, as seen in Figure 7-28. In the next window, you are prompted for information about the dashboard, including a name and description. In the Student Performance Analysis project, a new dashboard that is named `Student Performance Dashboard` was added to the project, as shown in Figure 7-29.

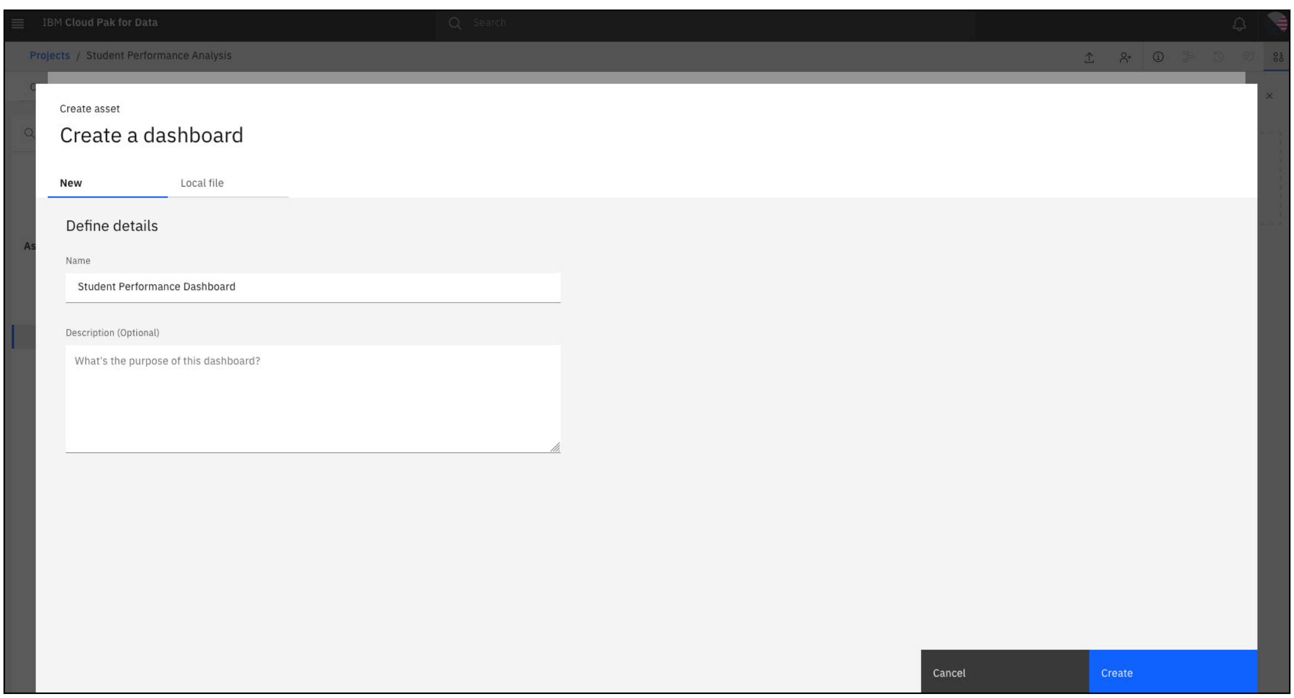

Figure 7-29 Defining details for new dashboard

5. After the dashboard details are entered, a dashboard is created. Before any visualizations can be created, a source must be selected for the dashboard. As shown in Figure 7-30, the Data panel on the left is empty; new data sources can be added by selecting **Select a source**.

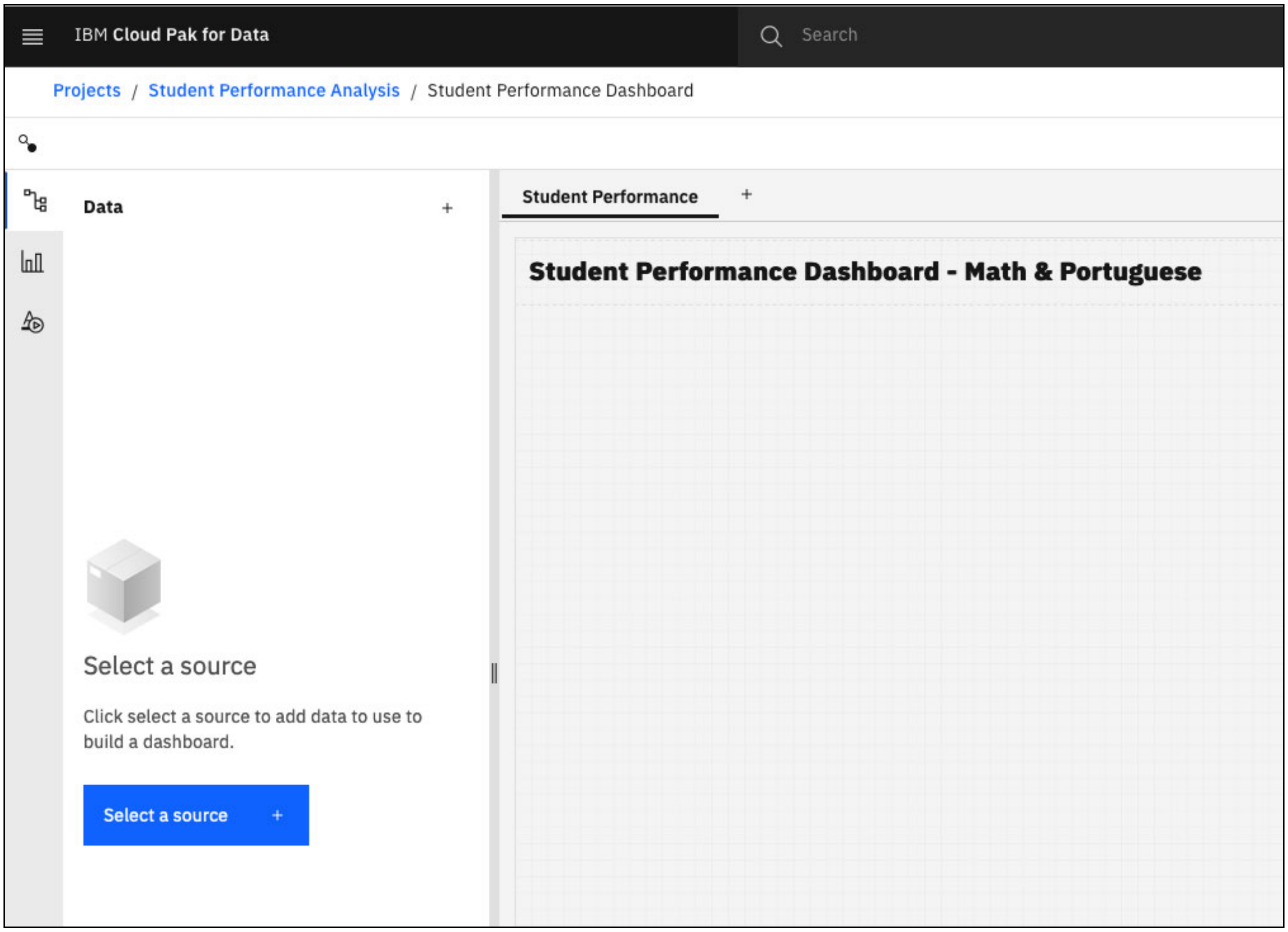

Figure 7-30 New Cognos Dashboard in IBM Watson Studio project

6. In the next window, the data assets that are available in the project are shown. Figure 7-31 on page 509 shows an error message that states: `Missing personal credentials`, when the `STUDENT_GRADES` view is selected.

   This is because before a virtual object from Data Virtualization can be used, the credentials that are used to access this data source must be provided. Select the link that is provided in the error message to enter the credential details.

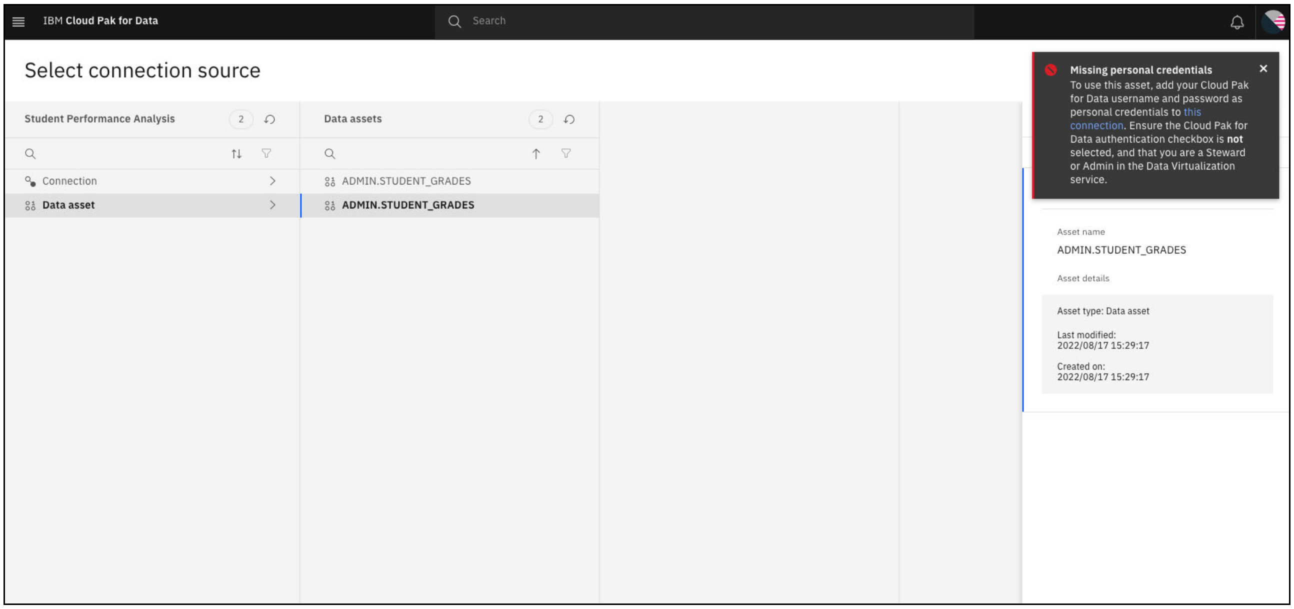

Figure 7-31 Selecting a connection source for Cognos Dashboard

7. Figure 7-32 shows the connection details for the Data Virtualization virtual object. Enter the credentials for the data source, which in this example are the admin credentials, and select **Save** to update the data connection.

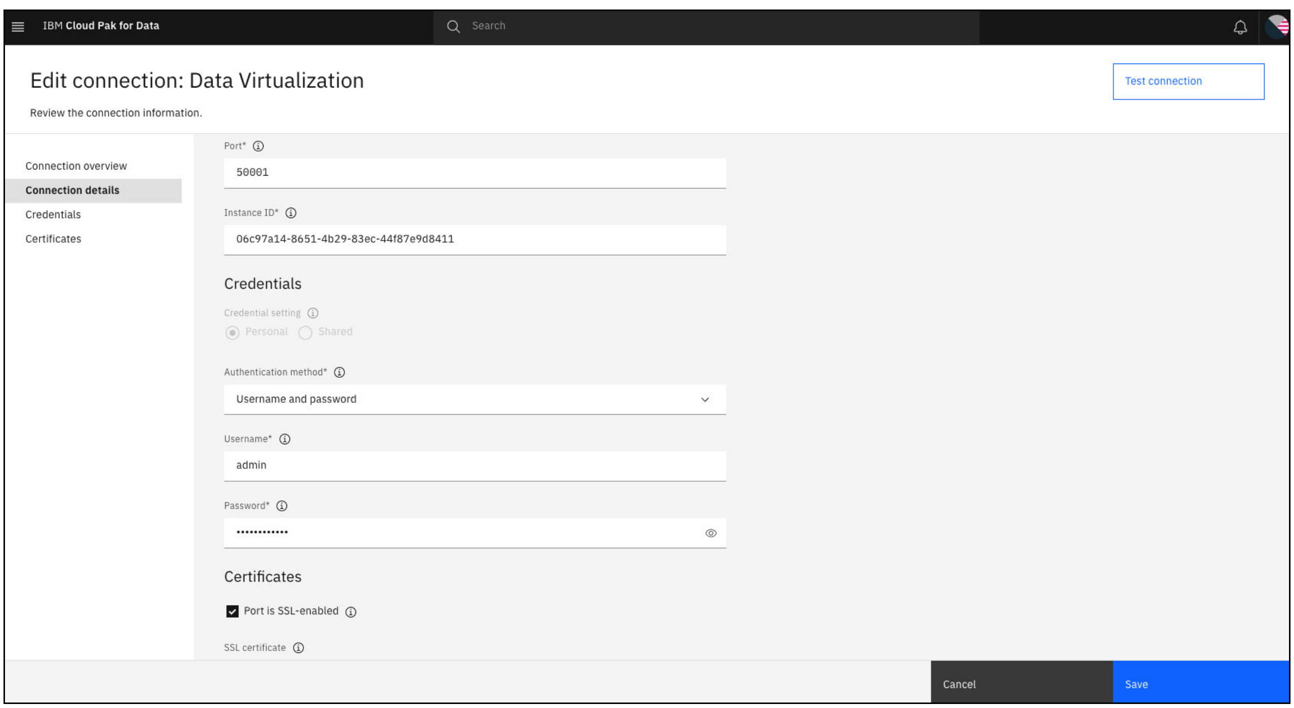

Figure 7-32 Updating data source connection credentials

8. After the data source is configured, the Data panel displays the available fields in the `STUDENT_GRADES` view, as shown in Figure 7-33.

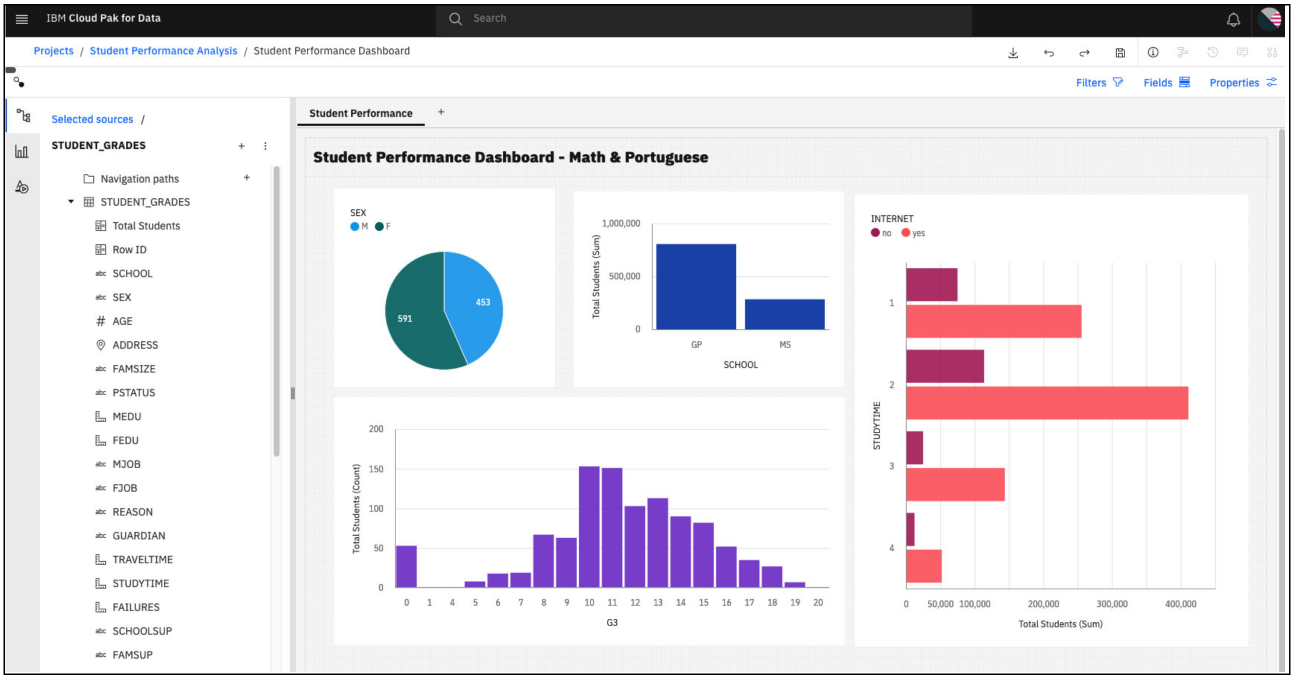

Figure 7-33 Student Performance dashboard

Option 2: Visualize by using Cognos Analytics

Cognos Analytics on Cloud Pak for Data also can be used to visualize a Data Virtualization view. To create a dashboard by using a Data Virtualization view, such as `STUDENT_GRADES`, a connection must be configured in Cognos Analytics to Data Virtualization.

Before a connection in Cognos Analytics can be created, the load balancer must be updated to allow external traffic to be routed to Data Virtualization.

Example 7-2 shows a sample of what must be added to the load balancer configuration if HAProxy is used. For more information, see this IBM Documentation web page.

Example 7-2 Updating HAProxy configuration file

```
frontend dv-nonssl
       bind *:NodePort for 50000
       default_backend dv-nonssl
       mode tcp
       option tcplog
backend dv-nonssl
       balance source
       mode tcp
       server master0 Master0-PrivateIP:NodePort for 50000
       server master1 Master1-PrivateIP:NodePort for 50000
      (repeat for each master node in the cluster)
```

Next, gather the connection details from Data Virtualization. This information is used to configure the BigSQL connection in Cognos Analytics. Figure 7-34 shows how the connection details for Data Virtualization can be retrieved from the Data Virtualization drop-down menu, under Configure connection.

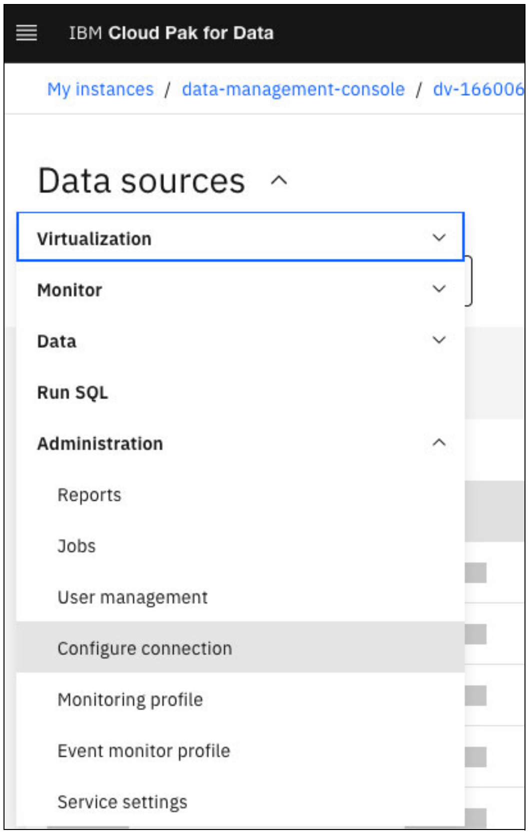

Figure 7-34 Configuring connection drop-down option

Next, complete the following steps to create a BigSQL connection in Cognos Analytics:

1. Open the Cognos Analytics instance to configure a connection to Data Virtualization. Figure 7-35 shows how to open Cognos Analytics on Cloud Pak for Data from the **Instances** page.

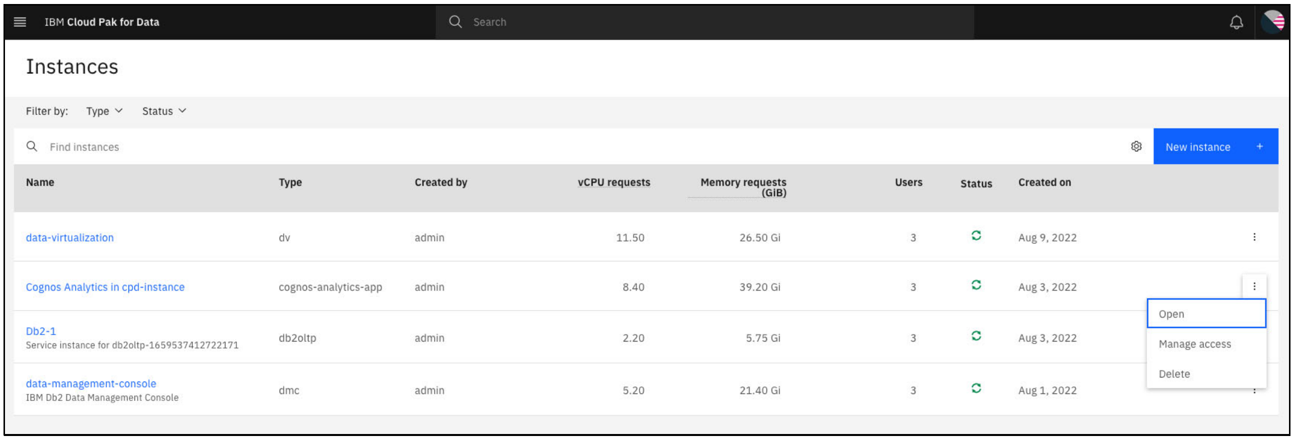

Figure 7-35 Accessing Cognos Analytics from the Instances page

2. On the Cognos Analytics home page, select **Manage** from the navigation menu. Figure 7-36 shows how the navigation menu on the left can be used to access the **Manage** page.

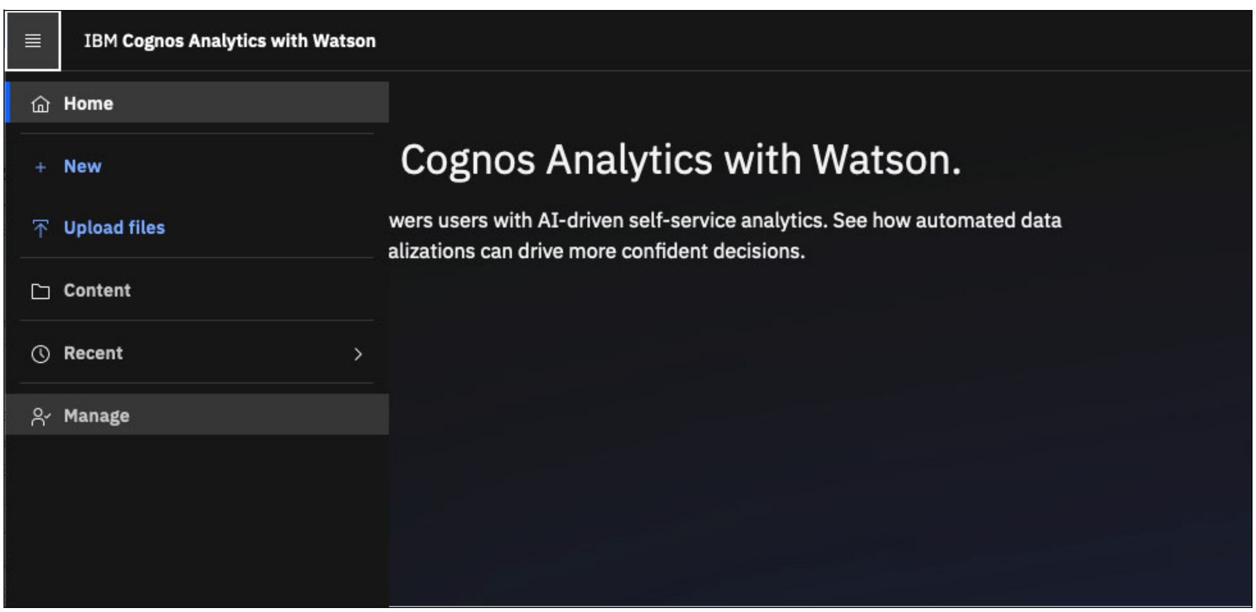

Figure 7-36 Cognos Analytics home page

3. Select **Data server connections** from the available options, as shown in Figure 7-37.

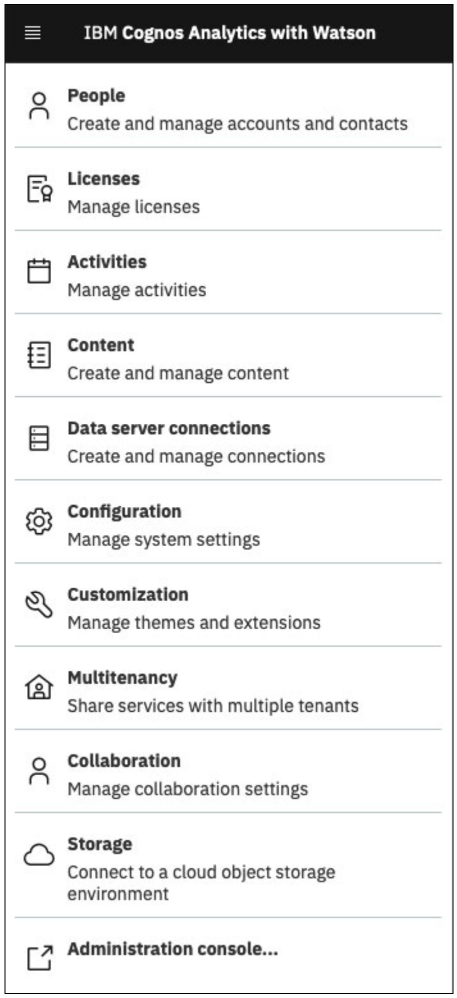

Figure 7-37 Managing available options in Cognos Analytics

4. Search for `IBM Big SQL` under **Select a type**, as shown in Figure 7-38. IBM Big SQL is selected as the data server connection type because IBM Big SQL is the underlying SQL engine for Data Virtualization.

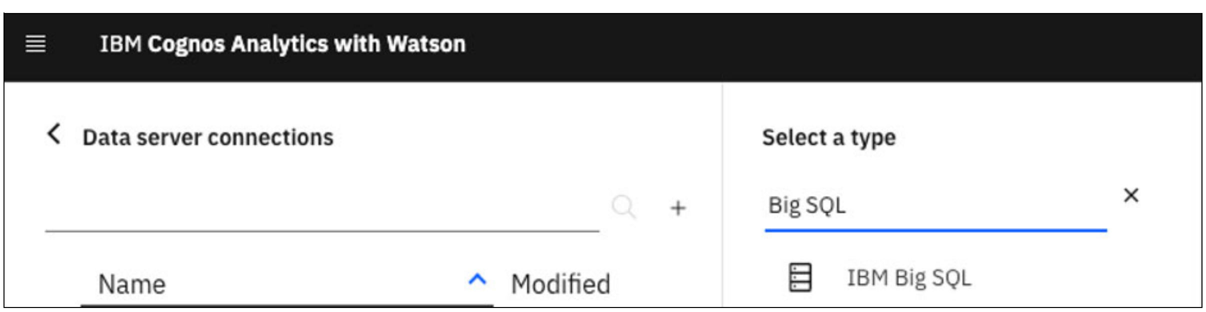

Figure 7-38 Searching for IBM Big SQL under the Select a type option

5. After **IBM Big SQL** is selected as the data server connection type, the window that is shown in Figure 7-39 opens. In this window, you are prompted for information about the new connection. Select **Settings** and update the `JDBC URL` with the information that is retrieved from Data Virtualization.

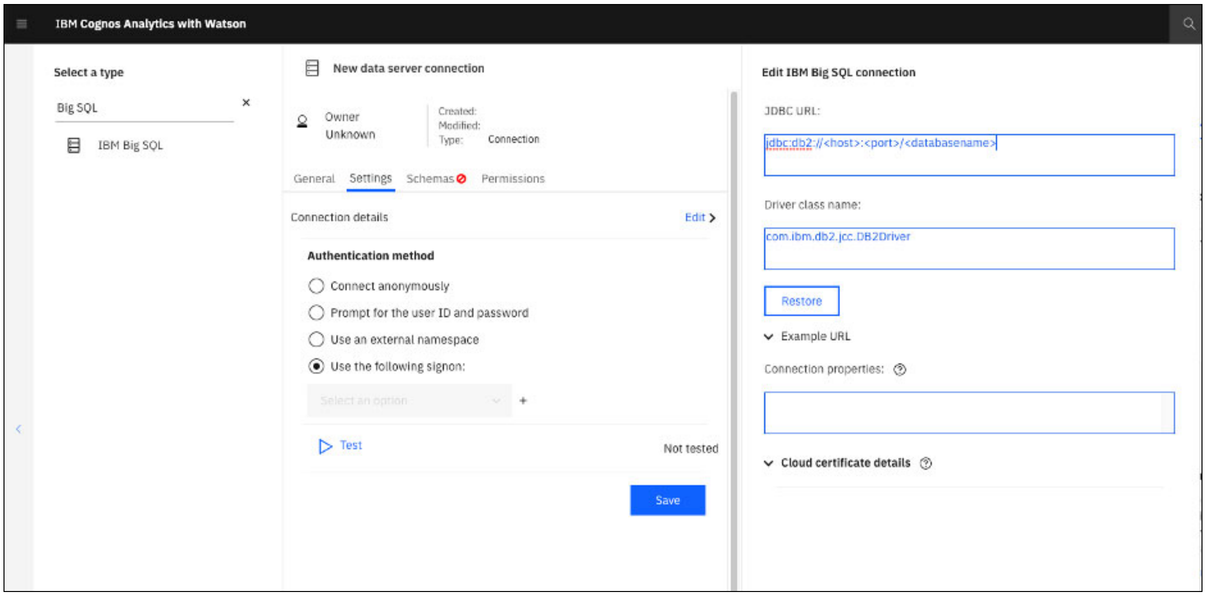

Figure 7-39 IBM Big SQL connection requirements

6. In the Authentication method window, select **Use the following signon:** and then, the **+** button to add a sign on (see Figure 7-40).

Figure 7-40 Adding sign-on for IBM Big SQL connection

7. In the next window, select the **Credentials** tab and then, enter the User ID and Password for the credentials that are used to access IBM Big SQL, as shown in Figure 7-41.

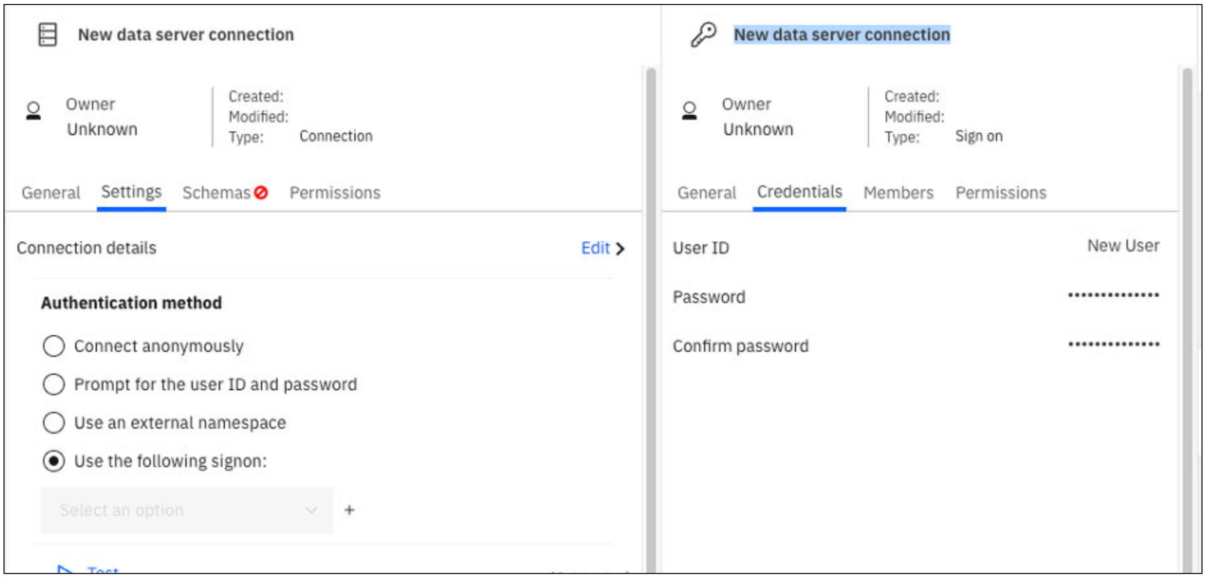

Figure 7-41 Data server connection credentials

8. After all of the required information is entered, test the connection. Figure 7-42 shows the expected output after testing a connection. Select **Save** after the data server connection test passes.

Figure 7-42 Testing data server connection

Load metadata for schemas with views needed

After the data server connection is created, the metadata must be loaded for the schemas that contain the tables and views that are needed for analysis.

As shown in Figure 7-43 on page 516, the metadata was loaded for the ADMIN schema, but not for the BIGSQL schema.

Complete the following steps:

1. The schemas for a data server connection can be accessed by selecting the data server connection, and then selecting the **Schemas** tab (see Figure 7-43).

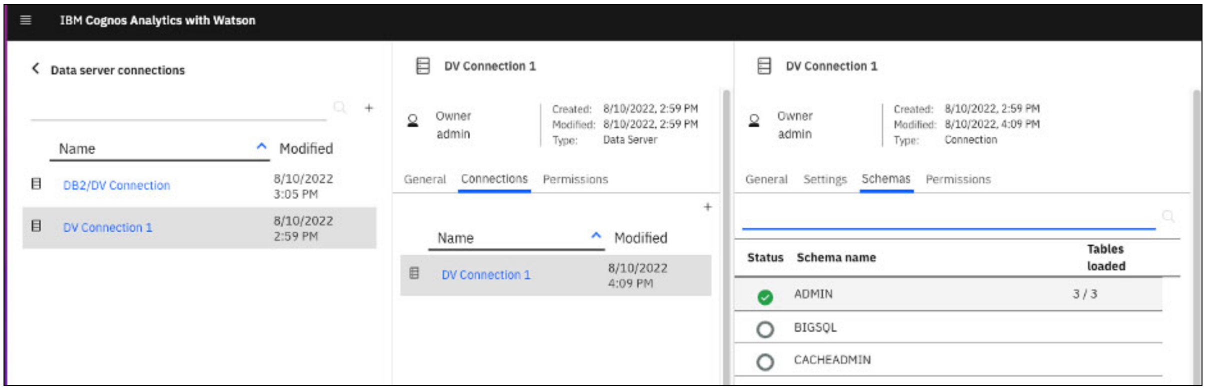

Figure 7-43 Data server connection schemas

2. Right-click the schema name and select **Load metadata**. Figure 7-44 shows how to load the metadata for the BIGSQL schema.

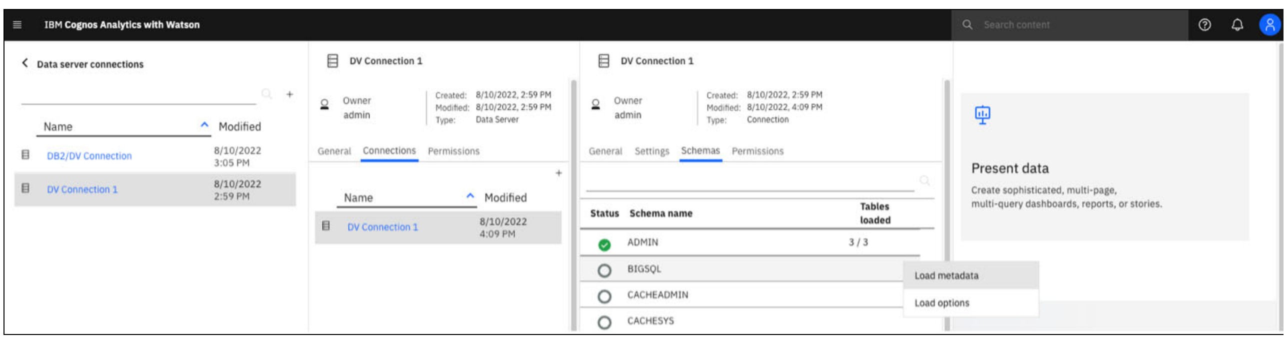

Figure 7-44 Loading metadata for BIGSQL schema

3. After the metadata for the schema is loaded, a green check mark appears next to the schema name, as shown in Figure 7-45.

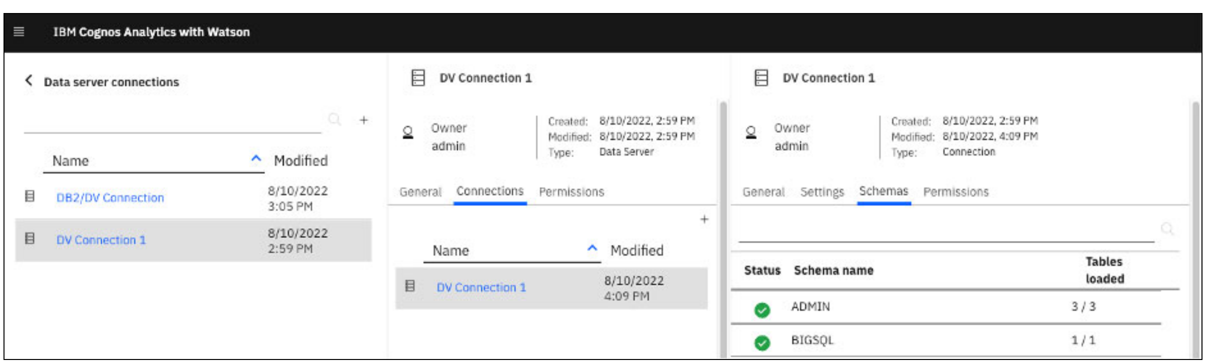

Figure 7-45 Completed metadata load for BIGSQL schema

Creating a data module

Complete the following steps to create a data module in Cognos Analytics by using the tables in the ADMIN schema in Data Virtualization:

1. From the navigation menu, select **+New** and then, **Data module**, as shown in Figure 7-46.

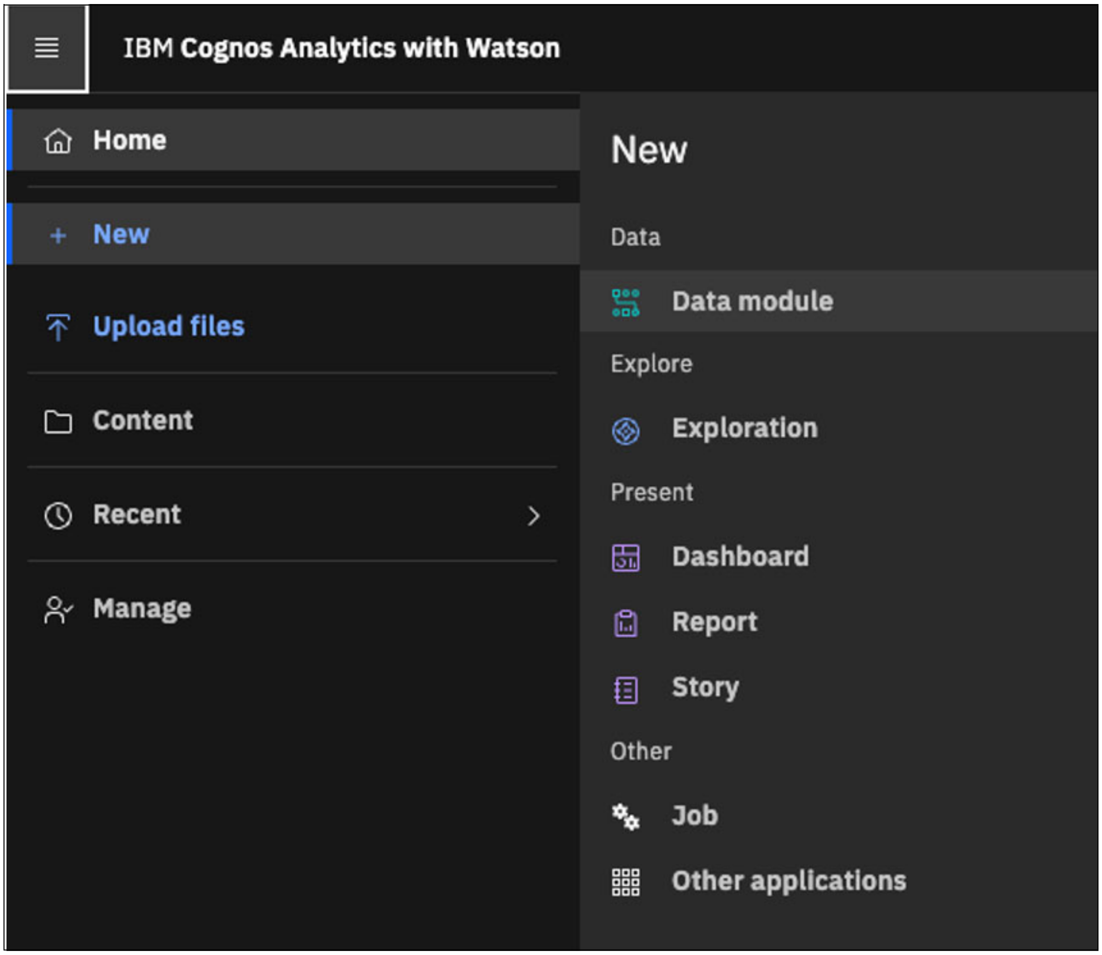

Figure 7-46 Creating a data module

2. In the next window, select the **Data Source** icon on the left, and the name of the data server connection that includes the data that is needed for analysis. In our example, `DV Connection 1` contains the tables that are needed for analysis, as shown in Figure 7-47.

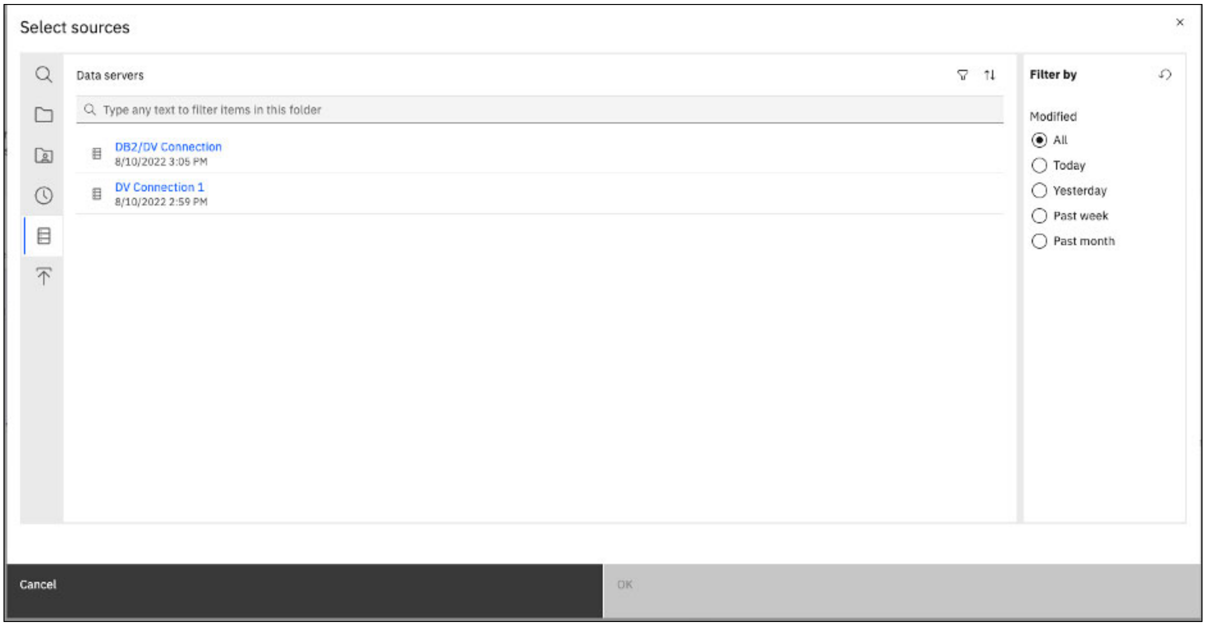

Figure 7-47 Selecting a data source

3. Select the schema that includes the tables that are needed for analysis. In our example, two schemas are available: `ADMIN` and `BIGSQL`, as shown in Figure 7-48. Select the **ADMIN** schema.

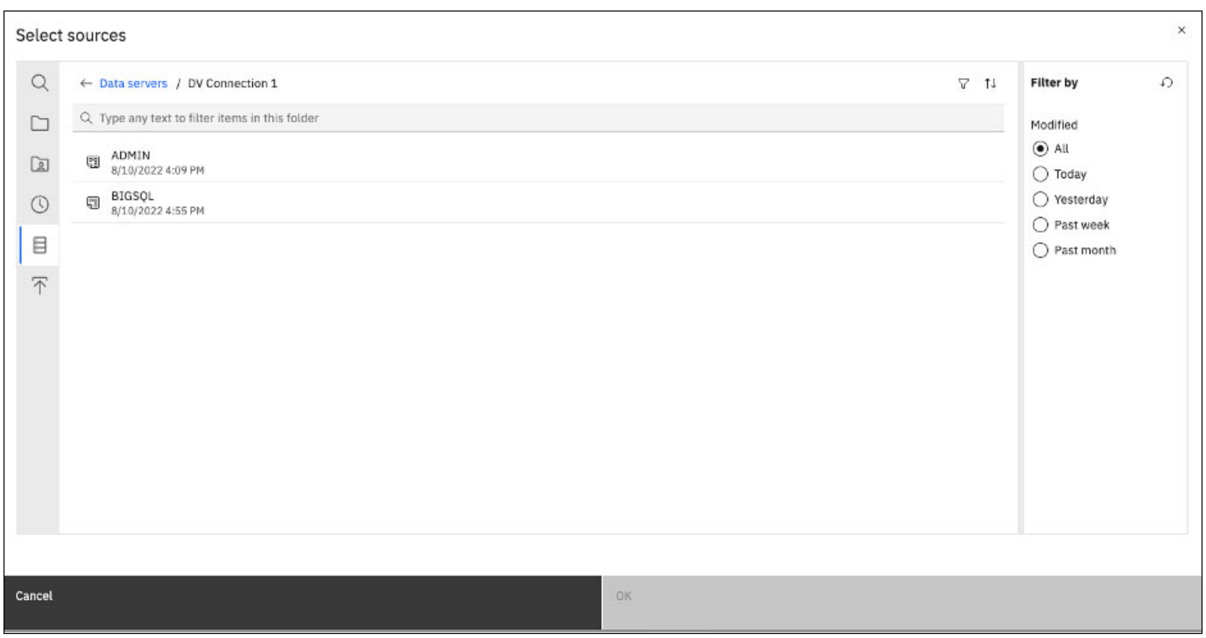

Figure 7-48 Selecting schema

4. To add tables, choose the **Select tables** options and then. select **Next**, as shown in Figure 7-49.

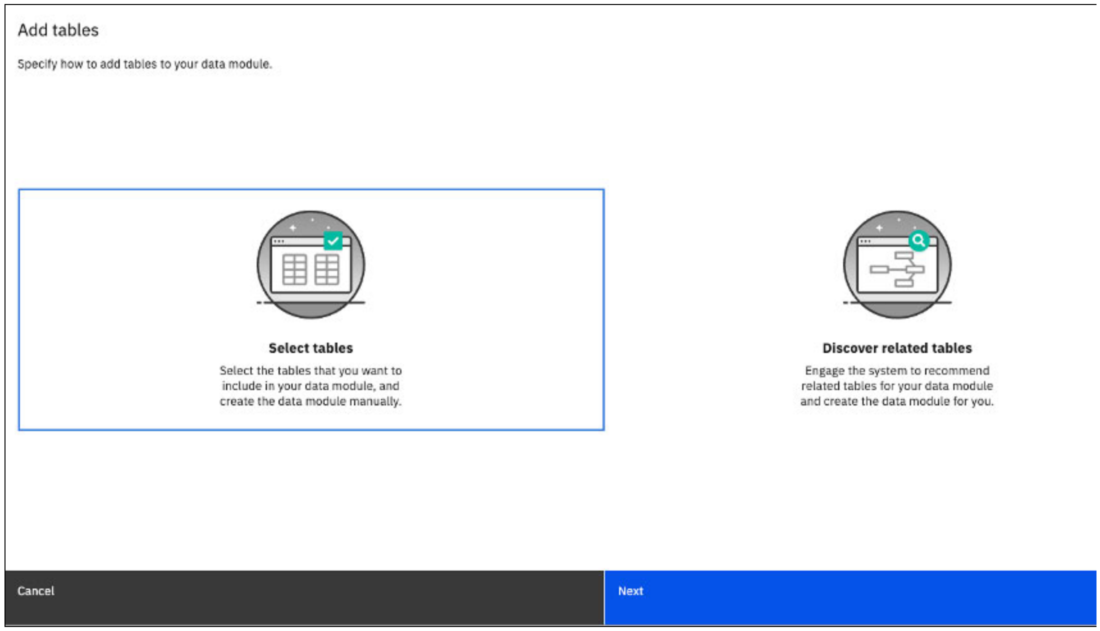

Figure 7-49 Adding tables to data module

5. The tables and views that are available for analysis are listed on the left side of the window, as shown in Figure 7-50. Select the **STUDENT_GRADES** view and then, click **OK**.

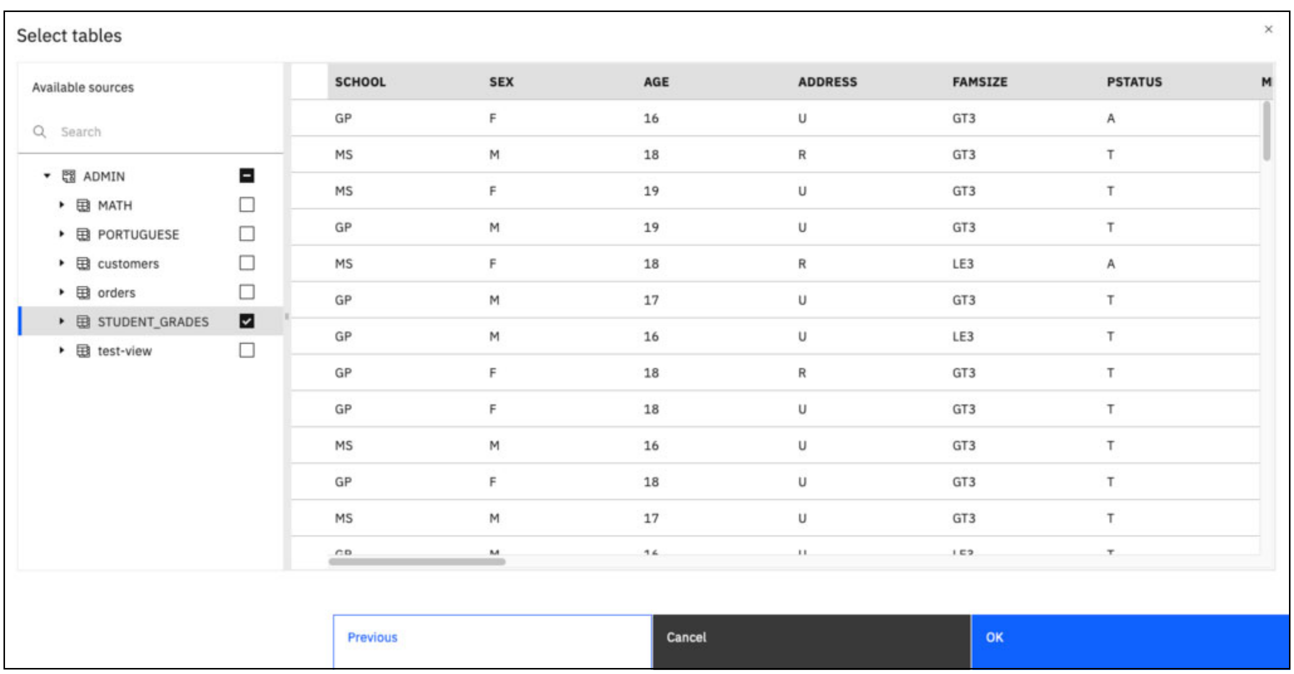

Figure 7-50 Selecting tables or views for data module

6. Save the data module in a folder. Figure 7-51 shows the data module that is saved as `Student Grades Data Module` within the Team content folder. After the data module is saved, more changes can be made within the data module.

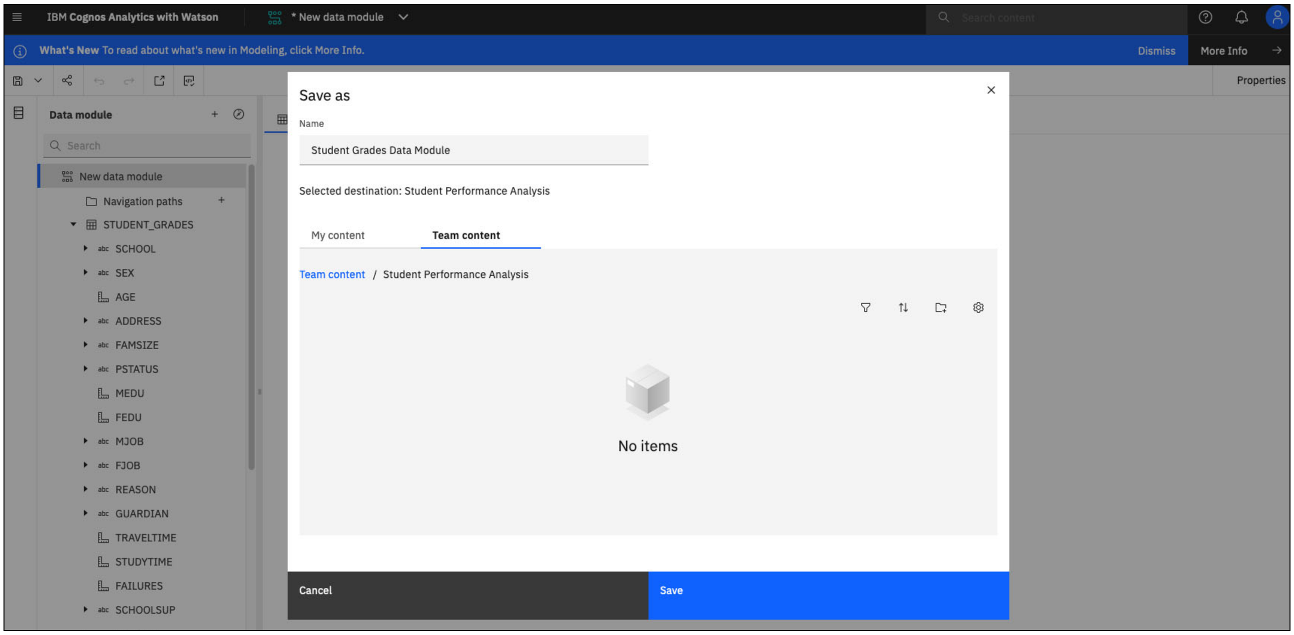

Figure 7-51 Saving the data module

Next, complete the following steps to create a dashboard:

1. Select **+New** and then **Dashboard** from the navigation menu, which is similar to the process for creating a data module.
2. After a new dashboard is created, select **Select a source** from the left side to add a data source, as shown in Figure 7-52.

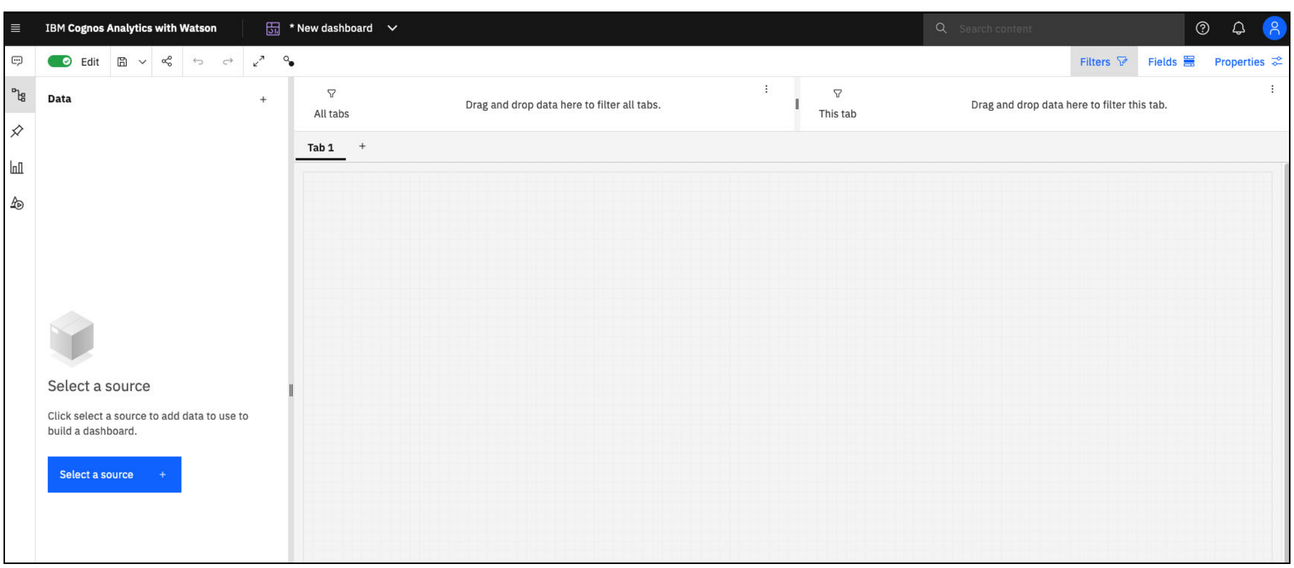

Figure 7-52 New dashboard in Cognos Analytics

3. Figure 7-53 shows the Student Grades Data Module that was created. Select this data module and then, click **Add**.

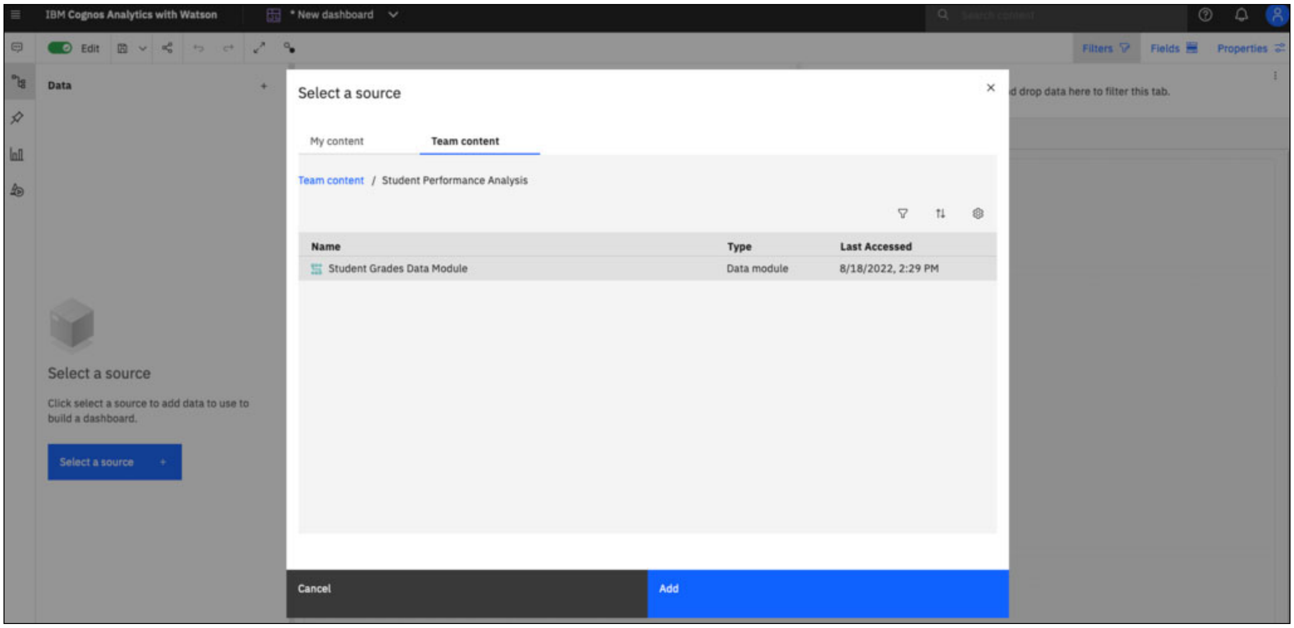

Figure 7-53 Adding data module as dashboard data source

4. After the data source is configured, the Data panel displays the fields that are available in the `STUDENT_GRADES` view, as seen in Figure 7-54.

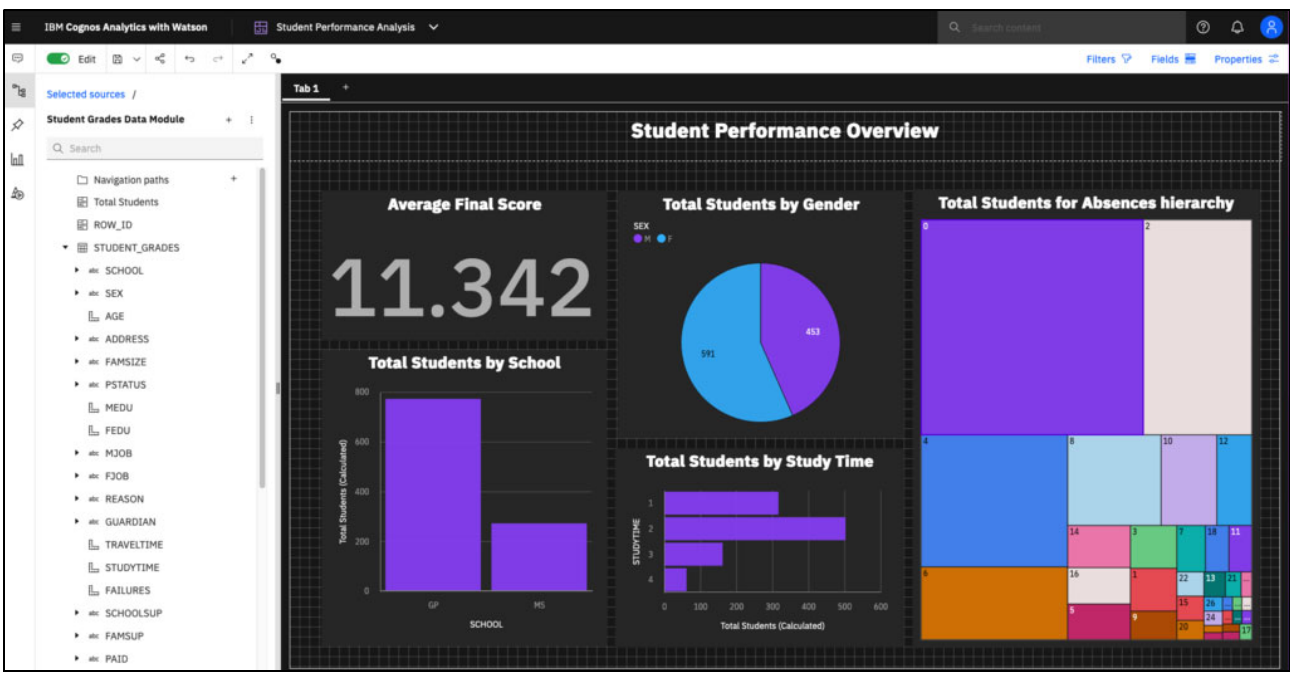

Figure 7-54 Student Performance sample dashboard

7.3.2 Use case #2: Visualizing model results

Predictive analytics, which is the ability to make predictions about future outcomes based on historical data, is a common use case. Cognos Analytics provides built in-capabilities for forecasting and advanced analytics; however, a model can be developed by using another tool and the results must be shared with users in an easily understandable format.

This use case describes how to develop a model in IBM Watson Studio, and share the results in Cognos Analytics on Cloud Pak for Data.

The following use case uses the Seoul Bike Sharing Demand data set from the UCI Machine Learning Repository.[2] This data set shows the bike count demand per hour in Seoul, South Korea, December 2017 - November 2018. Each record in this data set represents a one-hour timeframe. The data set contains 14 attributes.

Table 7-3 lists the attributes in the data set.

Table 7-3 Attributes in Seoul Bike Sharing Demand data set

Column name	Column description	Sample values
Date	Date bike was rented (year-month-day)	1/12/2017
Rented Bike Count	Count of bikes rented each hour	254
Hour	Hour of the day bike was rented	0, 1, 2, and so on
Temperature	Temperature in Celsius	-5.2, -7.4, 1.9, and so on
Humidity	% Humidity	21, 23, 37, and so on
Windspeed	Windspeed in m/s	0.8, 1, 2.2, and so on
Visibility	in m	793,1256, 2000, and so on
Dew point temperature	Dew point temperature in Celsius	-17.6, -7, -5, and so on
Solar radiation	MJ/m2	0, 0.3, 1, and so on
Rainfall	Rainfall in mm	0, 0.1, 2.5, and so on
Snowfall	Snowfall in cm	0, 0.3, 0.9, and so on
Seasons	Season bike was rented	Winter, Spring, Summer, Autumn
Holiday	Indicate whether day bike is rented on holiday	Holiday, No Holiday
Functional Day	Indicates whether day bike is rented on functional day	Fun, NoFunc

[2] [1] Sathishkumar V E, Jangwoo Park, and Yongyun Cho. 'Using data mining techniques for bike sharing demand prediction in metropolitan city.' Computer Communications, Vol.153, pp.353-366, March, 2020
[2] Sathishkumar V E and Yongyun Cho. 'A rule-based model for Seoul Bike sharing demand prediction using weather data' European Journal of Remote Sensing, pp. 1-18, Feb, 2020

Step 1: Creating an IBM Watson Studio project

Complete the following steps to create an IBM Watson Studio project on Cloud Pak for Data:

1. Select **All projects** under the Projects heading (see Figure 7-55).

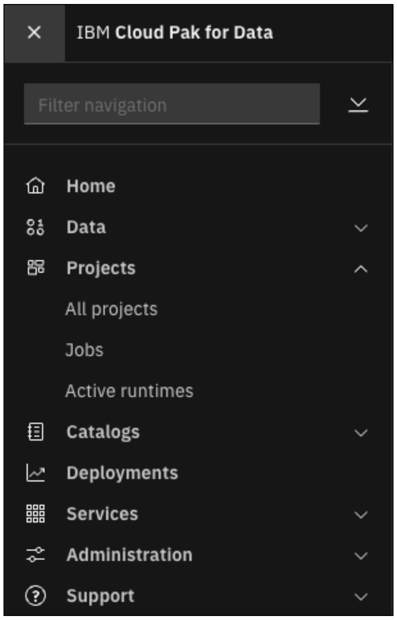

Figure 7-55 Accessing All projects on Cloud Pak for Data

2. Select **New Project** on the Projects page, as shown in Figure 7-56.

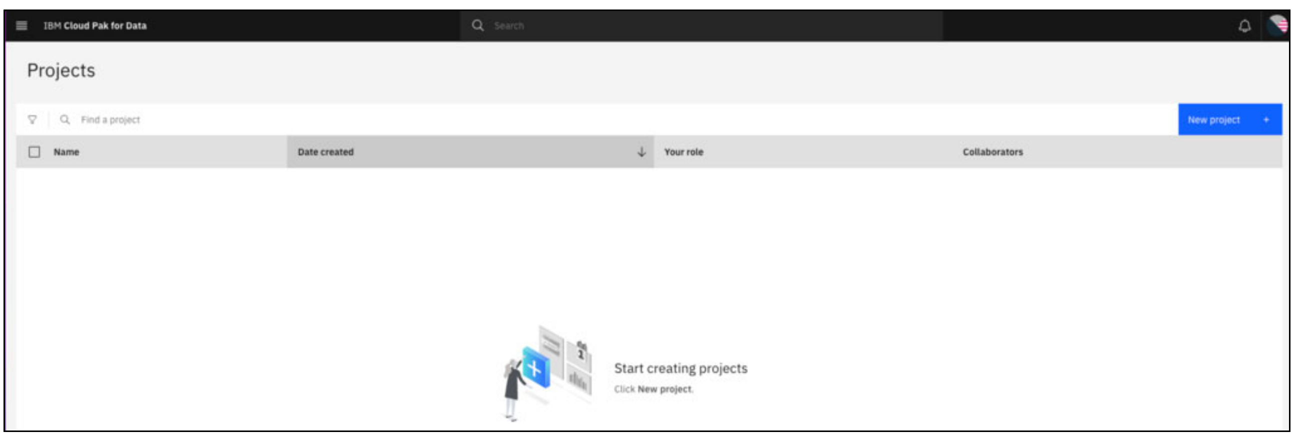

Figure 7-56 Projects page on Cloud Pak for Data

3. The following options are available when an IBM Watson Studio project is created, as shown in Figure 7-57 on page 524:
   - Create an empty project.
   - Create a project from a file.
   - Create project integrated with a Git repository.

   Select **Create an empty project**.

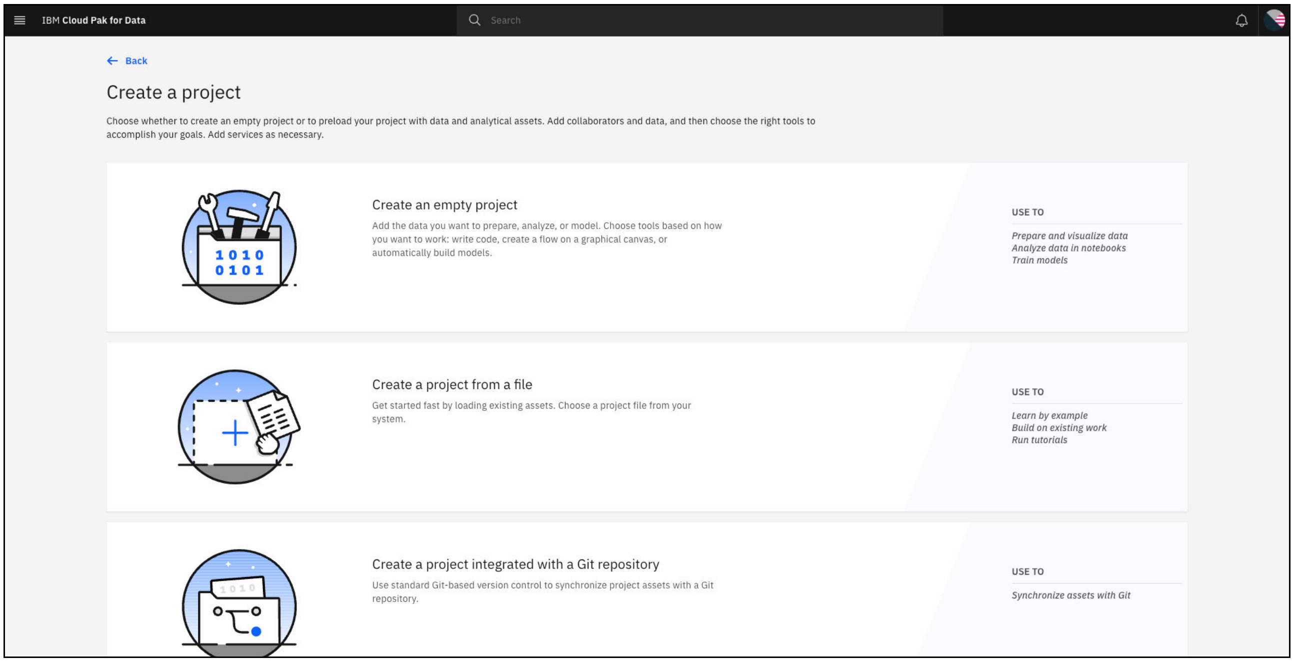

Figure 7-57 Creating project option

4. In the next window, enter a name for the project and select **Create**. In Figure 7-58, the name of the project that is created is shown as `Seoul Bike Sharing Demand Analysis`.

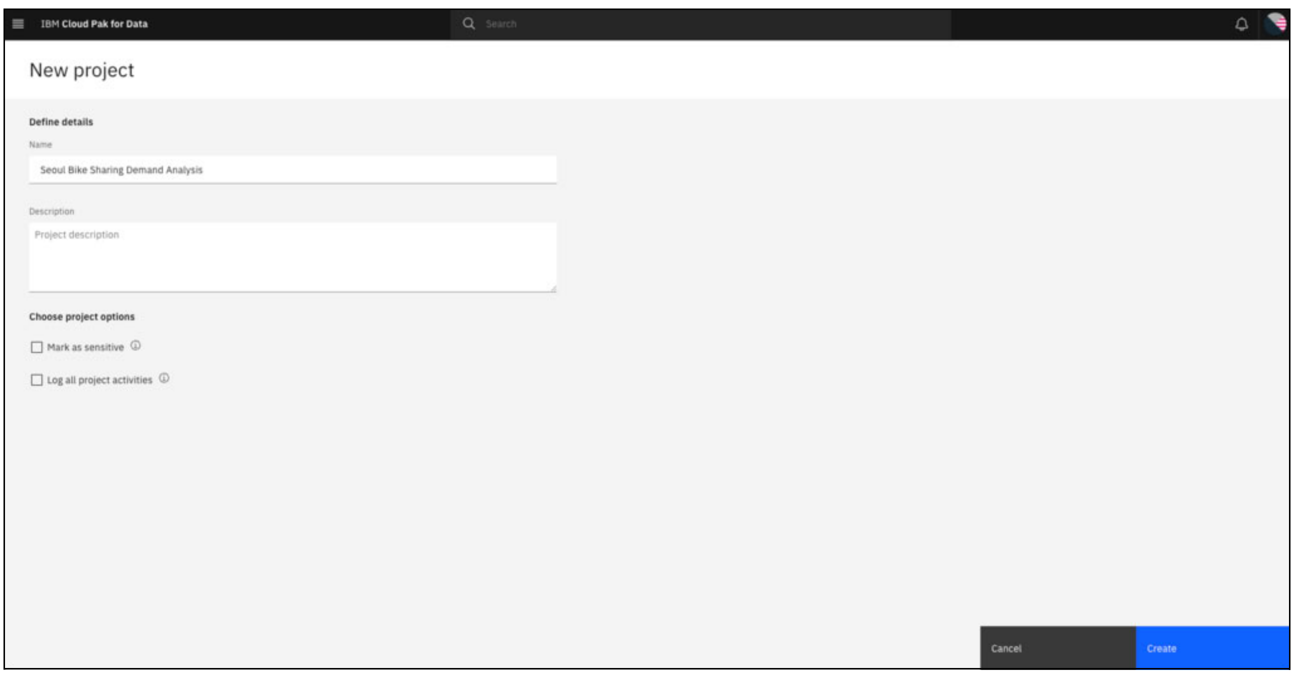

Figure 7-58 Entering details for a new project

After the project is created, it is accessible from the Projects page on Cloud Pak for Data, as shown in Figure 7-59.

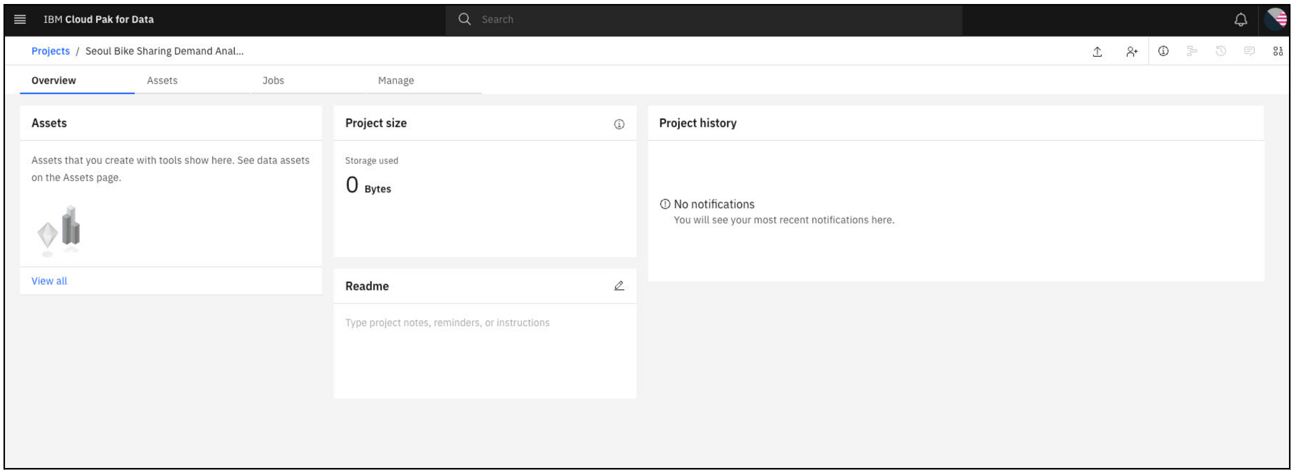

Figure 7-59 Project overview

Step 2: Uploading the data set

After a project is created, complete the following steps to upload files to the project:

1. Figure 7-60 shows the Assets tab within the `Seoul Bike Sharing Demand Analysis` project. To upload data to this project, select the icon on the right, **Upload asset to project**. Then, select **Drop data files here or browse for files to upload** on the right side panel. A new window opens in which you can search for local files to upload to the project. This example uses the `Seoul Bike Sharing Demand` data set and can be downloaded from the UCI Machine Learning Repository.

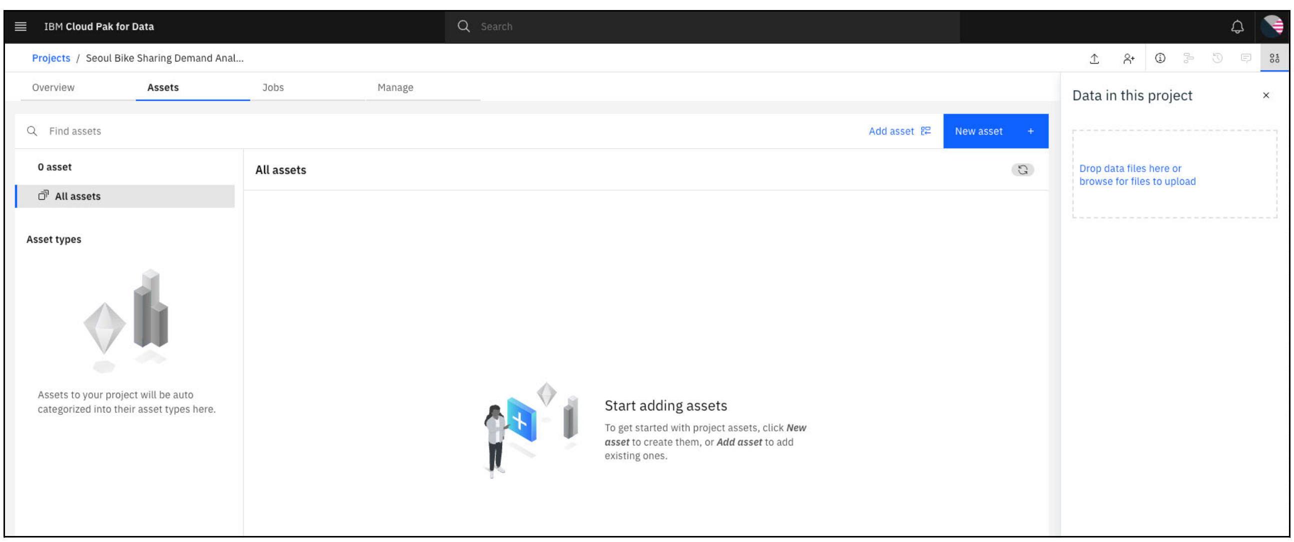

Figure 7-60 Uploading a data file to project

After the file is uploaded to the project, it is listed under Data asset on the Assets page, as shown in Figure 7-61.

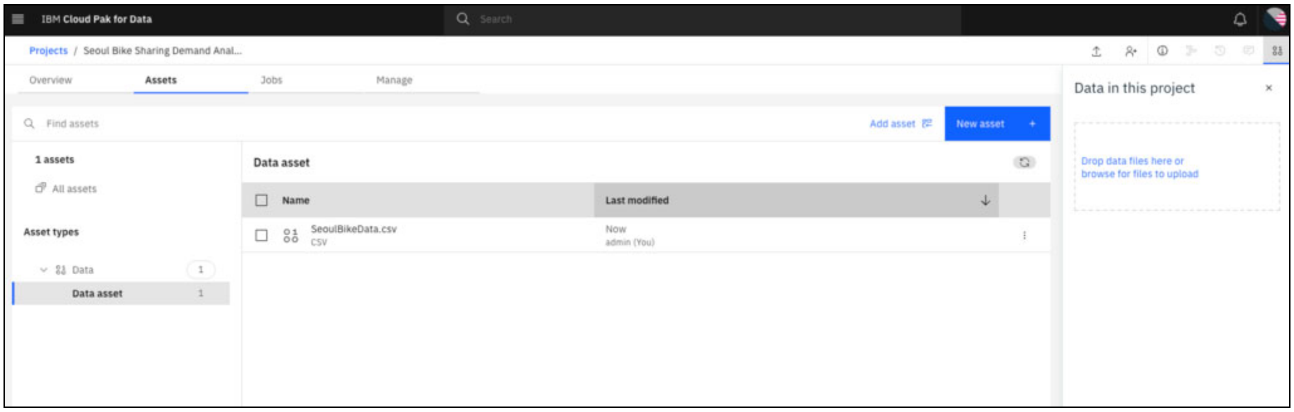

Figure 7-61 Project after data file is uploaded

Step 3: Creating a notebook

Complete the following steps to create a model in IBM Watson Studio by using a Jupyter notebook:

1. Add a Jupyter notebook asset by selecting **New asset**, as shown in Figure 7-62.

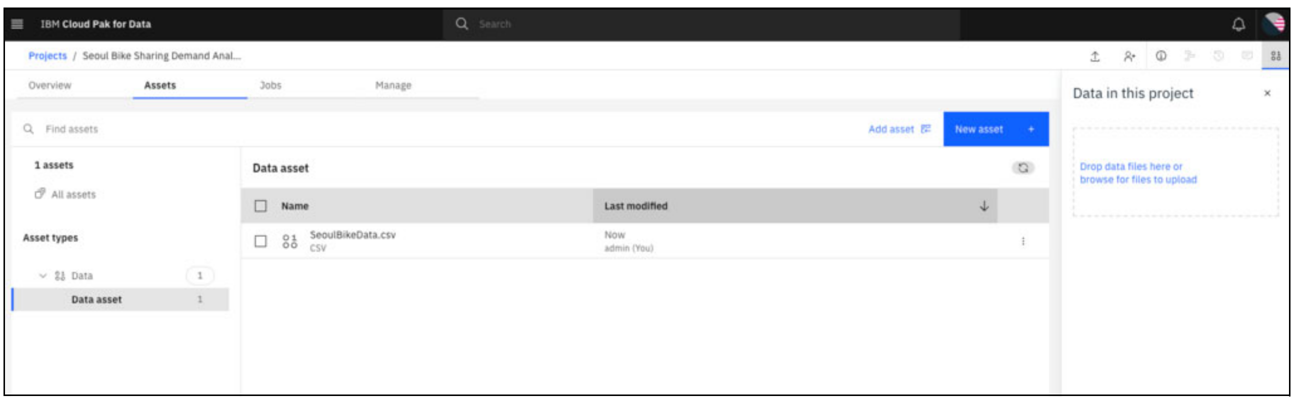

Figure 7-62 Adding an asset to a project

2. Figure 7-63 shows the available asset types to add to the project. Select **Jupyter notebook editor** from the options available.

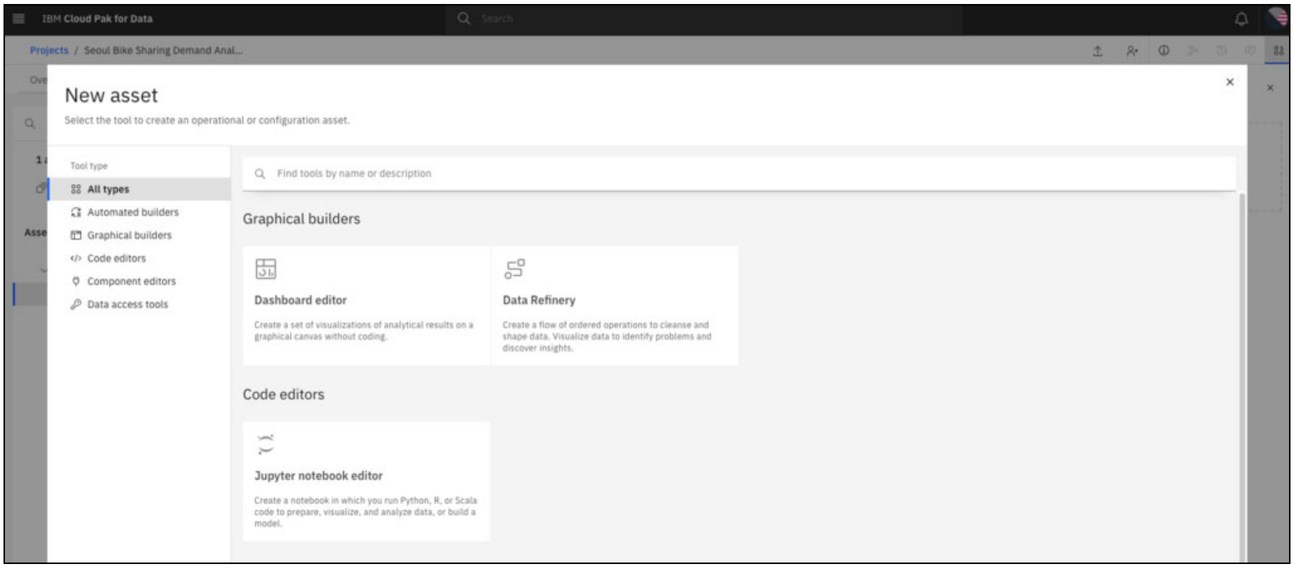

Figure 7-63 Available asset types to add to project

3. Enter a name for the new notebook and then, select **Create**, as shown in Figure 7-64.

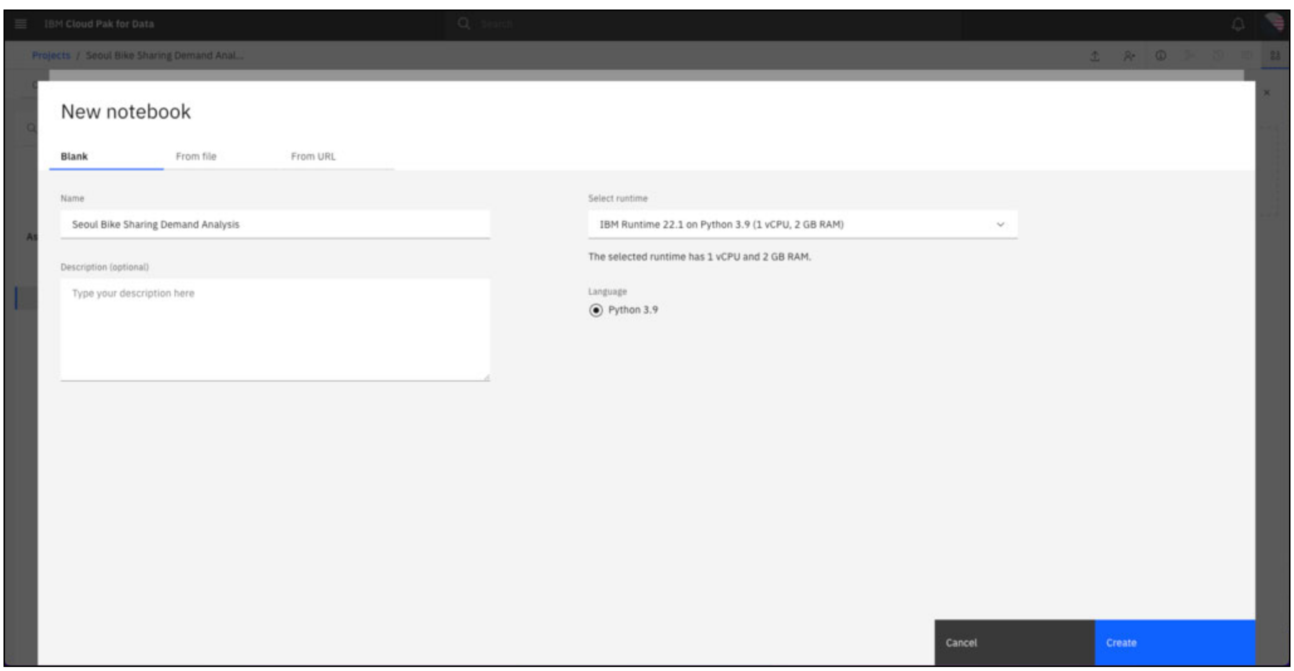

Figure 7-64 Entering notebook details

The notebook takes some time to create. Figure 7-65 shows the instantiating runtime window that is displayed after a notebook is created.

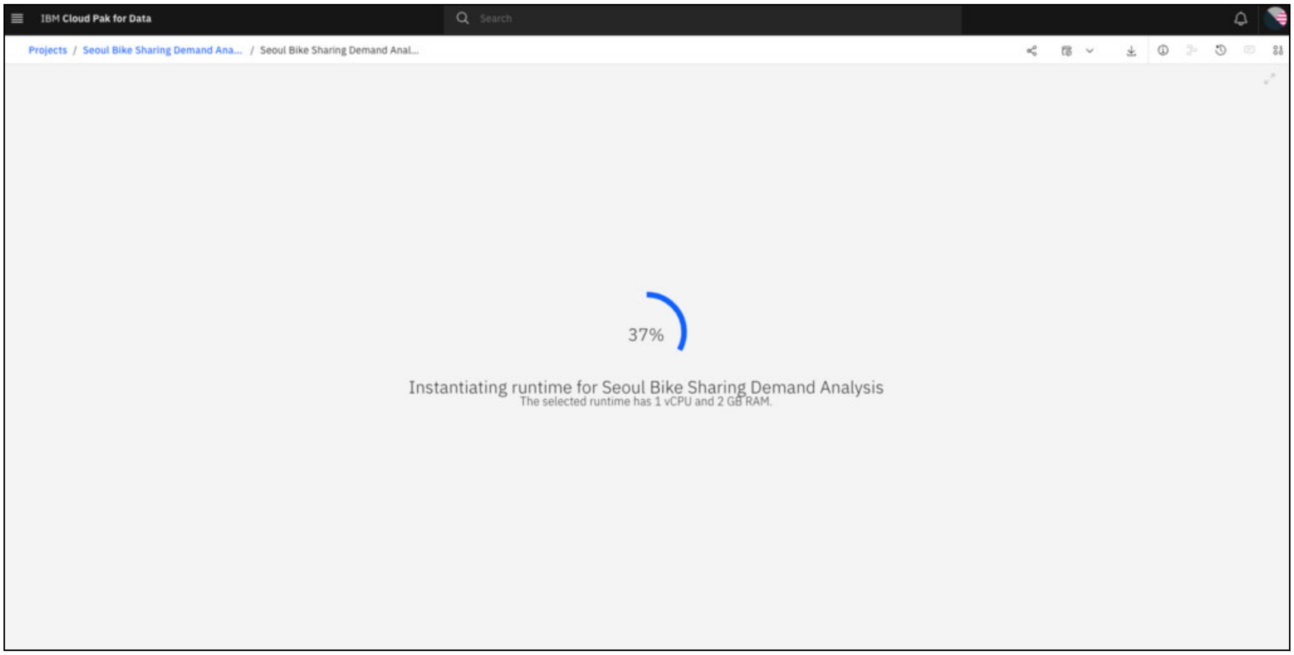

Figure 7-65 Instantiating runtime for new notebook

Step 4: Connecting to data set within notebook

Complete the following steps to connect to a data set within a notebook:

1. Select the **Find and add data** icon on the right. Then, select **pandas.DataFrame**, which is under the `SeoulBikeData.csv` data source, as shown in Figure 7-66.

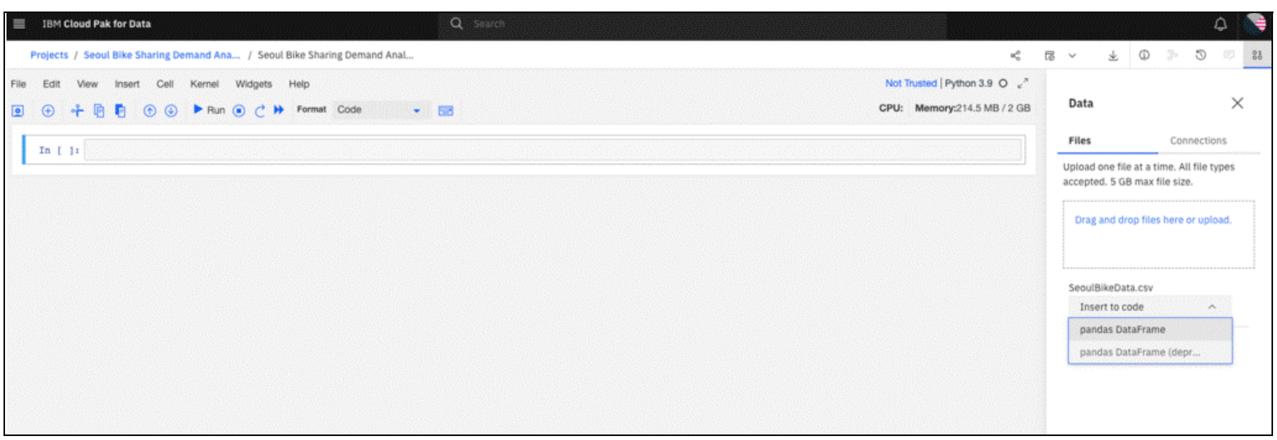

Figure 7-66 Inserting SeoulBikeData.csv as pandas.DataFrame

2. Code is inserted into the notebook to read the `.csv` in as a pandas data frame. Figure 7-67 shows the code to import the `SeoulBikeData.csv` file as a pandas data frame.

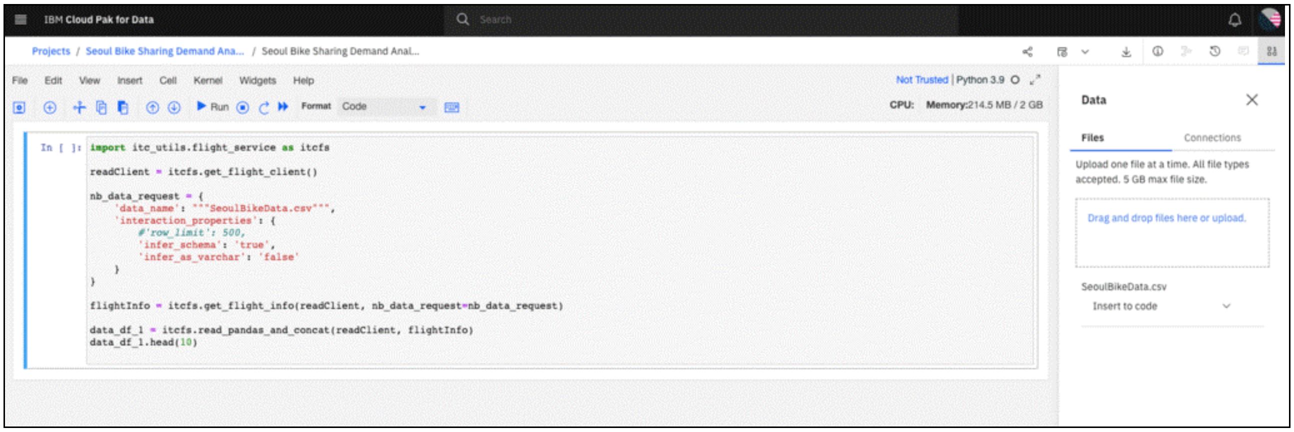

Figure 7-67 Python code to insert SeoulBikeData.csv as pandas data frame

3. Run the code to create the data frame and then to view the contents, as shown in Figure 7-68.

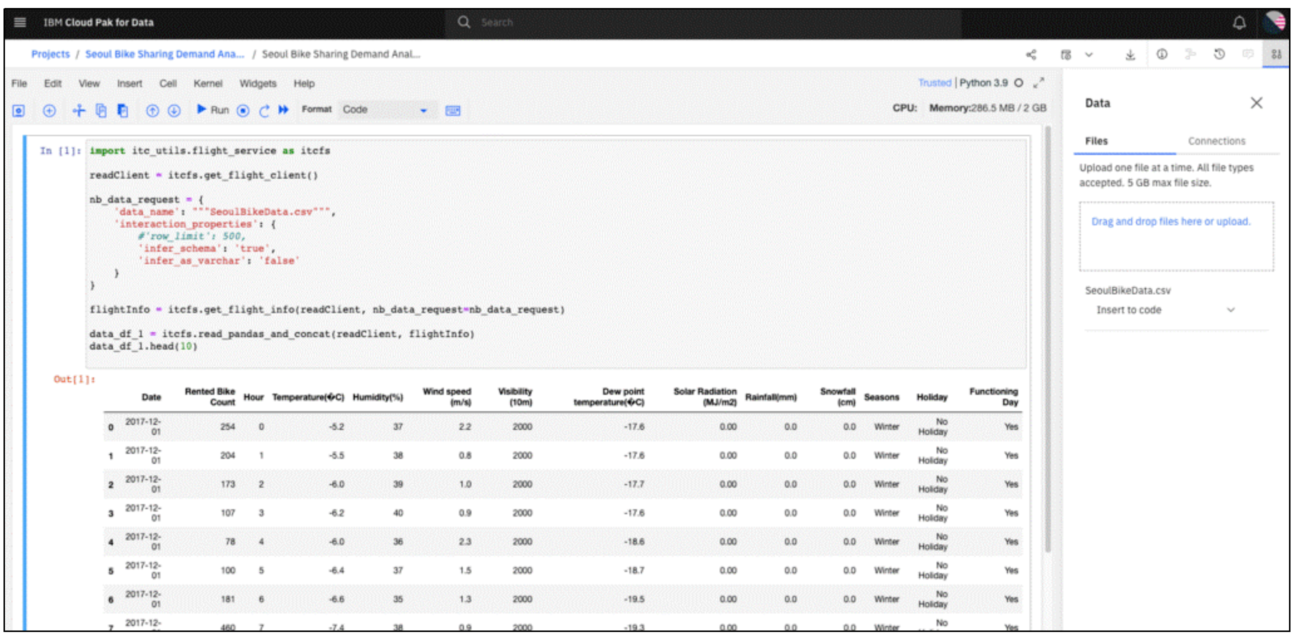

Figure 7-68 Running code and view contents of SeoulBikeData data frame

Step 5: Developing a model

After the data is loaded into a data frame, a model can be developed. The following example shows a regression tree with `Rented Bike Counts` as the target variable.

This section provides a high-level overview of some of the steps that are taken during the model building process. We show how some of the output from this section can be moved to Cognos Analytics for visualization and presentation.

Figure 7-69 shows a few of the steps that are taken during exploratory data analysis.

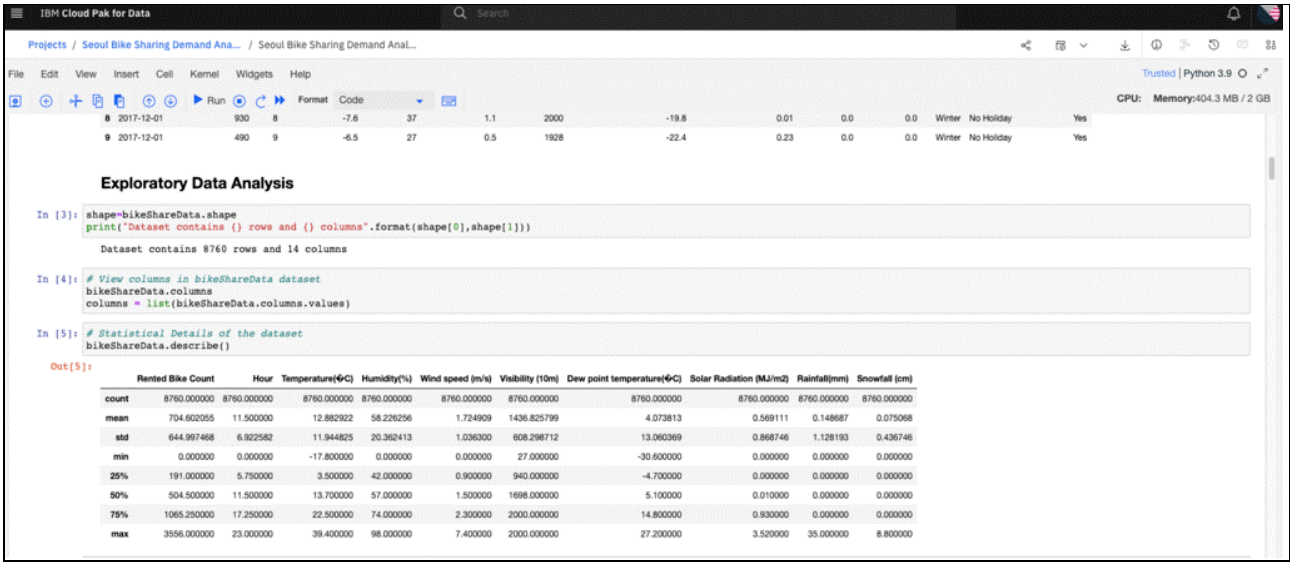

Figure 7-69 Exploratory Data Analysis

Figure 7-70 shows the Python code that is used to generate regression tree.

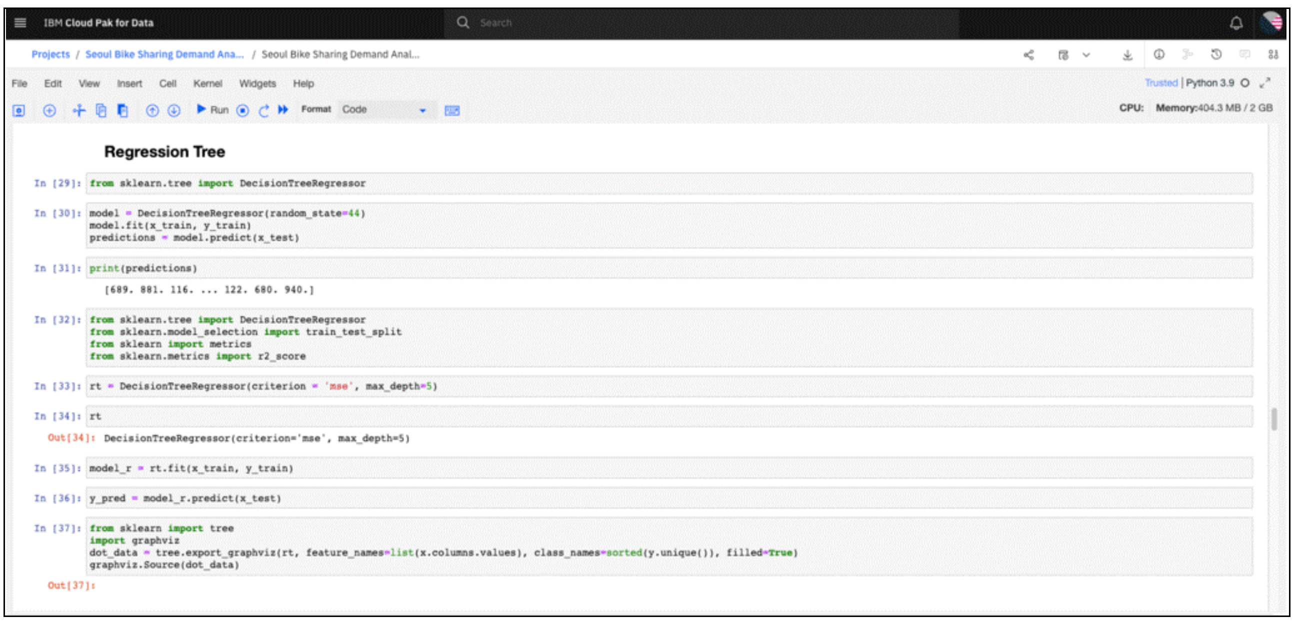

Figure 7-70 Creating regression tree

Figure 7-71 shows the creation of a data frame that contains the feature importance that is based on the generated model.

```
# create dataframe for feature importance
FeatureImportance = pd.DataFrame(columns=['feature-name', 'importance'])

for importance, name in sorted(zip(rt.feature_importances_, x_train.columns),reverse=True):
    # print (name, importance)
    FeatureImportance = FeatureImportance.append({'feature-name' : name,
                    'importance' : importance} ,
                    ignore_index=True)
```

```
FeatureImportance
```

0]:

	feature-name	importance
0	Seasons_Winter	0.367353
1	Humidity(%)	0.221828

Figure 7-71 Feature importance data frame

These steps show a few examples of the output from the model building process if a model is developed by using Python. For more information about how some of this output can be written to a Cognos Analytics folder, see "Step 6: Writing output results to Cognos Analytics Folder" on page 531.

Step 6: Writing output results to Cognos Analytics Folder

Example 7-3 shows an example of how to import the Cognos Analytics Data Connector, and write the `bikeShareData` data frame to a folder within a user's personal folder, and to a folder within the Team Content folder. Documents in a user's personal folder are accessible only to that user by default; documents in the Team Content folder can be accessed by other users based on permissions set on the folder.

Example 7-3 Cognos Analytics Data Connector implementation

```
# Import Cognos Analytics Data Connector
from ca_data_connector import CADataConnector

# Enter URL to CA Instance
CADataConnector.connect({'url':'https://cpd-cpd-instance.apps.payalpn.cp.fyre.ibm.
com/cognosanalytics/bi/?perspective=home'});

CADataConnector.connect({'url':'https://cpd-cpd-instance.apps.bast.cp.fyre.ibm.com
/cognosanalytics/cpd-instance/bi/?perspective=home'});


# Write data set to folder in Cognos Analytics - Specify name of data set, folder
in Cognos AnalyticsInstance, and mode to write
## Example 1 - Write data set to folder in personal folder
data = CADataConnector.write_data(bikeShareData,
path=".my_folders/BikeShareAnalysis/bikeShareData", mode="w") #example using
user's personal folder on Cognos Analytics

## Example 2 - Write data set to folder in public folder (team content)
data = CADataConnector.write_data(bikeShareData,
path=".public_folders/BikeShareAnalysis/bikeShareData", mode="w") #example using
folder within Team Content on Cognos Analytics
```

This process can be repeated for other data frames in the notebook, such as the `FeatureImportance` data frame.

Complete the following steps to write different data frames in Jupyter notebook to a folder in Cognos Analytics for analysis by using the Cognos Analytics Data Connector:

1. After the data sets are written to a folder in Cognos Analytics, open the Cognos Analytics instance and browse to the folder.
2. Figure 7-72 shows how folders can be accessed within Cognos Analytics from the navigation menu. Select **Content** from the menu options.

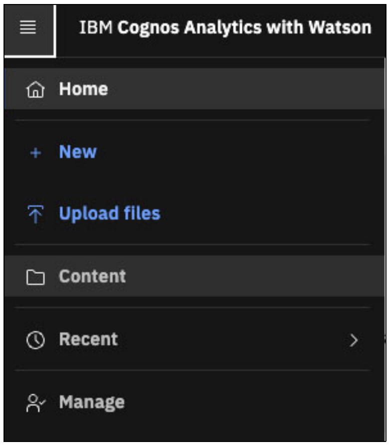

Figure 7-72 Accessing content folders within Cognos Analytics

3. Select **Team content** and the **BikeShareAnalysis** folder. Figure 7-73 shows the `FeatureImportance` and `bikeShareData` in the BikeShareAnalysis folder.

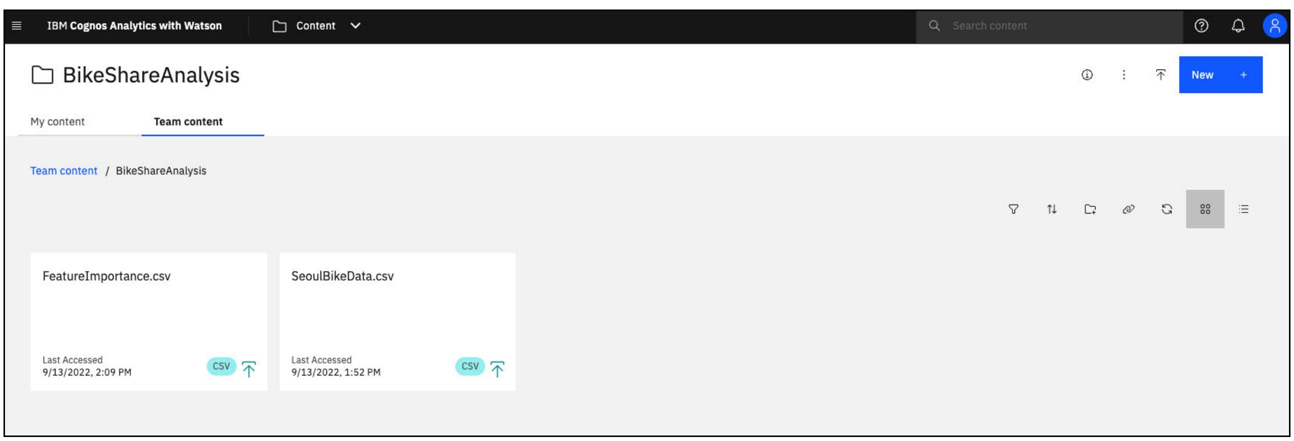

Figure 7-73 Datasets in BikeShareAnalysis folder on Cognos Analytics

Step 7: Creating a dashboard in Cognos Analytics

To create a dashboard by using these data sets, select **New** on the right in the GUI and then, select **Dashboard**, as shown in Figure 7-74.

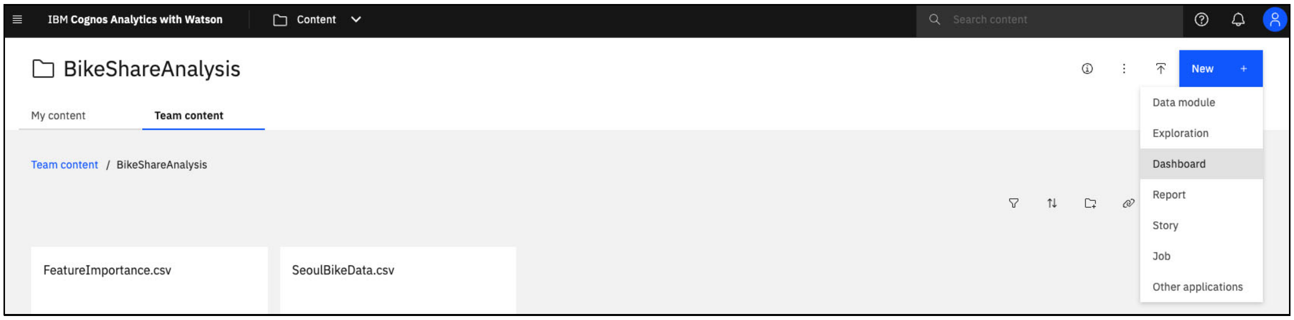

Figure 7-74   Creating a dashboard in BikeShareAnalysis folder

Figure 7-75 shows a sample dashboard that uses the two data sets in the BikeShareAnalysis folder. The left panel shows the two data sources that are used to create the dashboard. This dashboard contains visualizations that show the rented bike county by holiday, hour, and seasons.

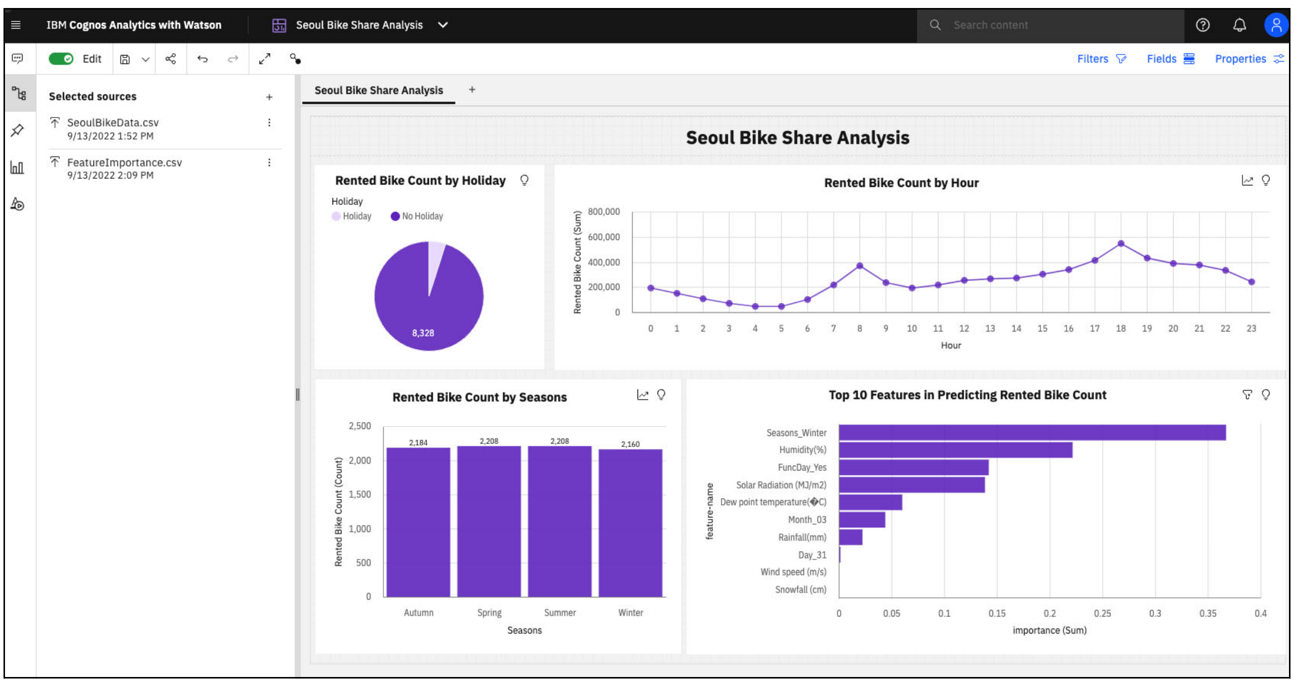

Figure 7-75   Seoul Bike Share Analysis Sample dashboard

Another graph shows the top 10 features in predicting rented bike count. For more information about dashboard design best practices and recommendations, see 7.3.3, “Use case #3: Creating a dashboard in Cognos Analytics” on page 534.

7.3.3 Use case #3: Creating a dashboard in Cognos Analytics

Cognos Analytics enables users to develop reports, dashboards, and stories. Dashboards are a common data product that is developed by teams to visually display and share information with members in an organization.

This use case provides an example of developing a dashboard in Cognos Analytics. It also highlights key features and helpful tips.

The following example uses the Student Performance data set again from the UCI Machine Learning Repository., specifically the Math students data set. This data set shows student grades in Math for two high schools that are in Portugal. Each data set contains 33 attributes for each record, and each record represents one student.

Table 7-2, in 7.3.1, "Use case #1: Visualizing disparate data sources" on page 496 highlights the attributes in the data set, including column name, column description, and sample values.

Accessing Cognos Analytics

To create a dashboard in Cognos Analytics, access the Cognos Analytics instance on Cloud Pak for Data from the **Instances** page, as shown in Figure 7-76.

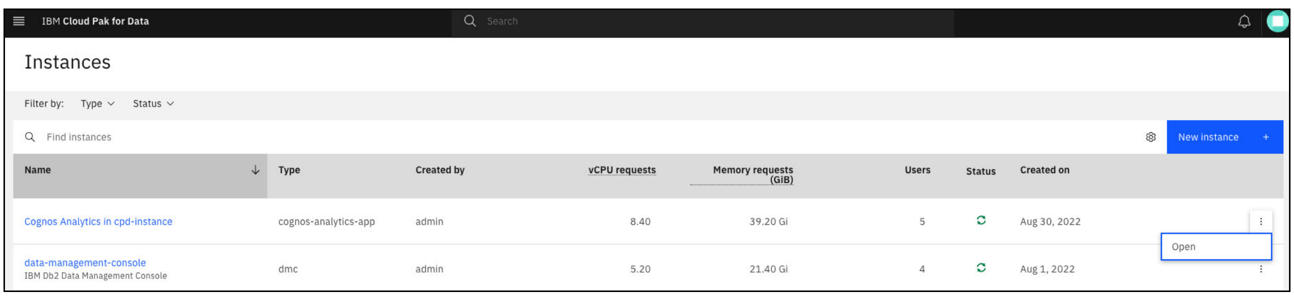

Figure 7-76 Instances on Cloud Pak for Data

Figure 7-77 shows the Cognos Analytics main page.

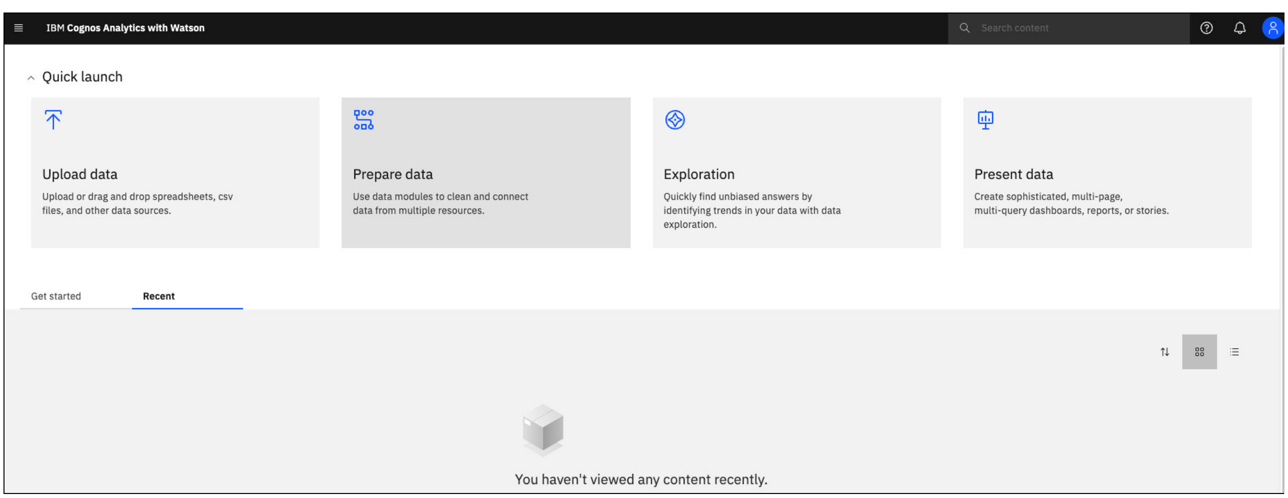

Figure 7-77 Cognos Analytics home page

Creating a folder

In Cognos Analytics, content is organized into folders. Users can add objects, such as data modules, reports, and dashboards to a folder. Folders can be created within a user's personal directory, called My content, or within the Team content folder.

Complete the following steps to create a folder:

1. To access the folders within My content and Team content, select **Content** from the navigation menu, as shown in Figure 7-78.

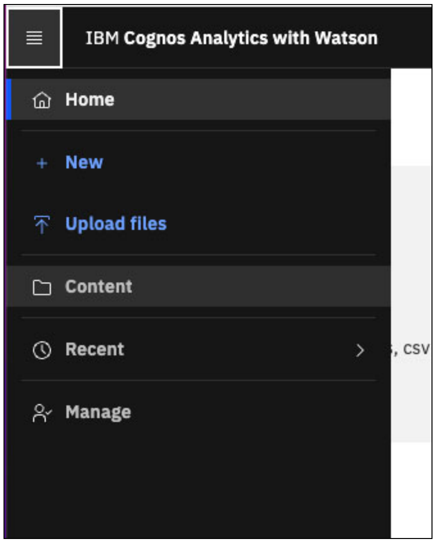

Figure 7-78 Content page in Cognos Analytics Navigation Menu

2. Figure 7-79 shows the Content page in Cognos Analytics. Folders can be created in My content or Team content. This example shows how to create a folder within My content, as shown in Figure 7-79. To create a folder, select the **Add folder** icon the left side of the window.

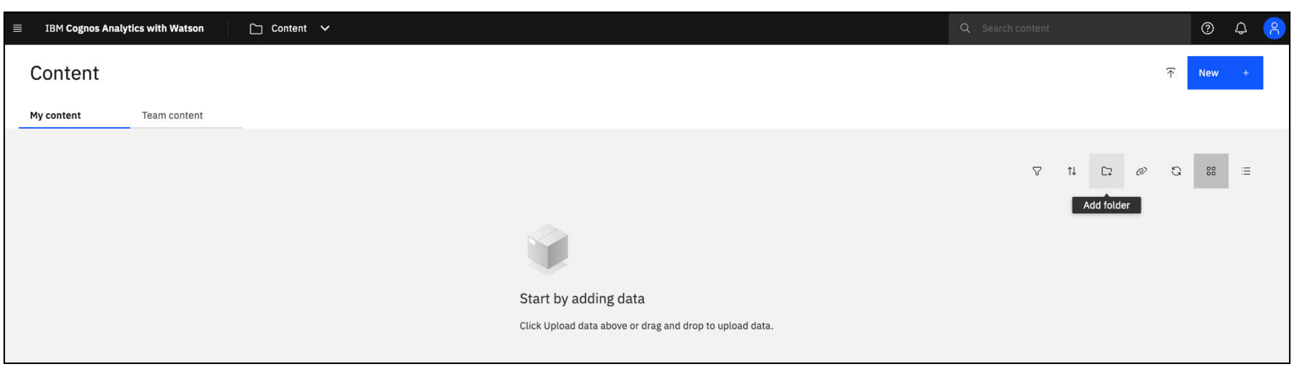

Figure 7-79 Adding a folder to My content

3. Enter a name for the new folder and select **Add**, as shown in Figure 7-80.

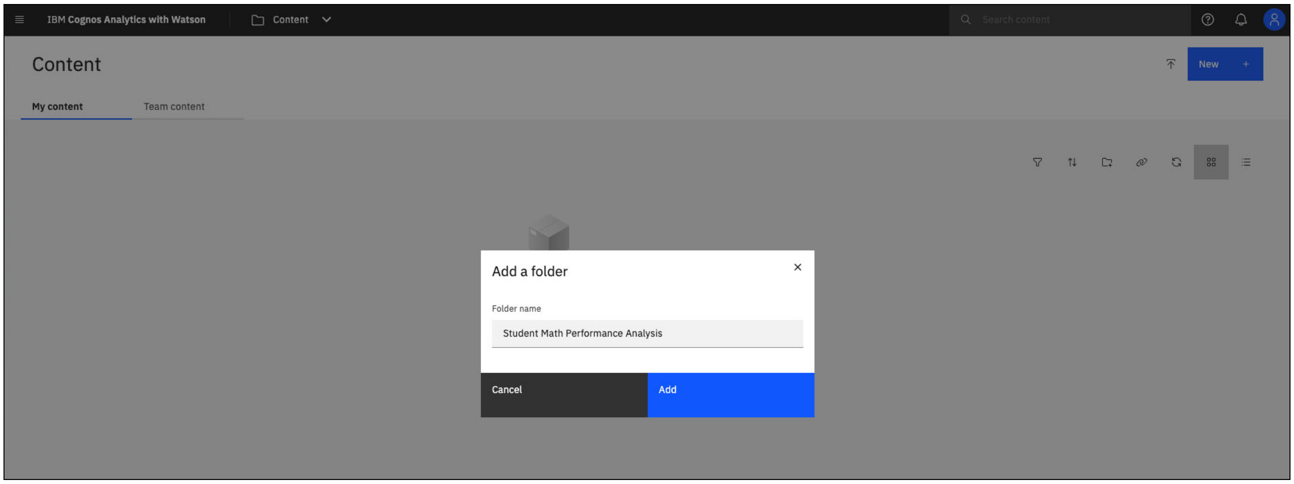

Figure 7-80 Creating a folder named Student Math Performance Analysis

Uploading a data set

After a folder is created, objects can be created and uploaded to the folder. Figure 7-81 shows the Student Math Performance Analysis folder that was created in the previous step. Select the folder to view its contents.

Figure 7-81 Folders in My content

As shown in Figure 7-82, no assets are in the Student Math Performance Analysis folder. To add a data file from a local machine, select the **Upload data** icon on the right and then, select the file to upload from your local machine.

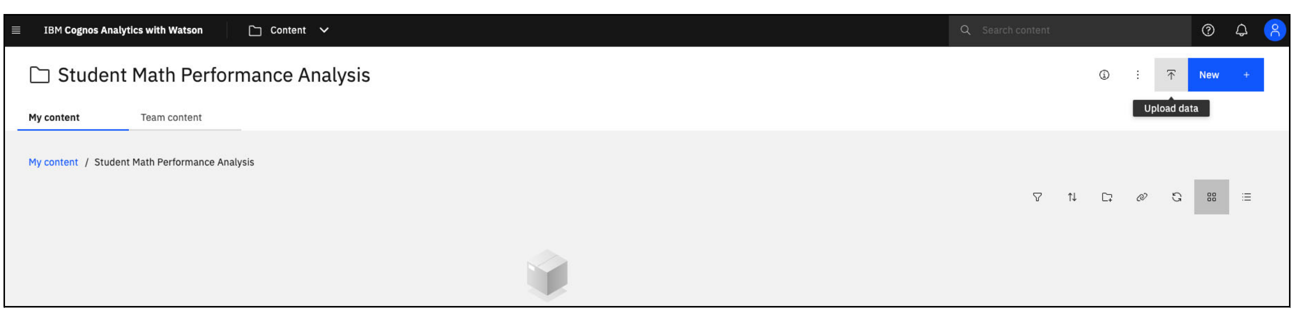

Figure 7-82 Student Math Performance Analysis folder

Creating a dashboard

To create a dashboard, select **New** and then, select **Dashboard**, as shown in Figure 7-83.

Figure 7-83 Adding dashboard object to Student Math Performance Analysis folder

Figure 7-84 shows the available dashboard templates. Select the first template and then, select **Create**.

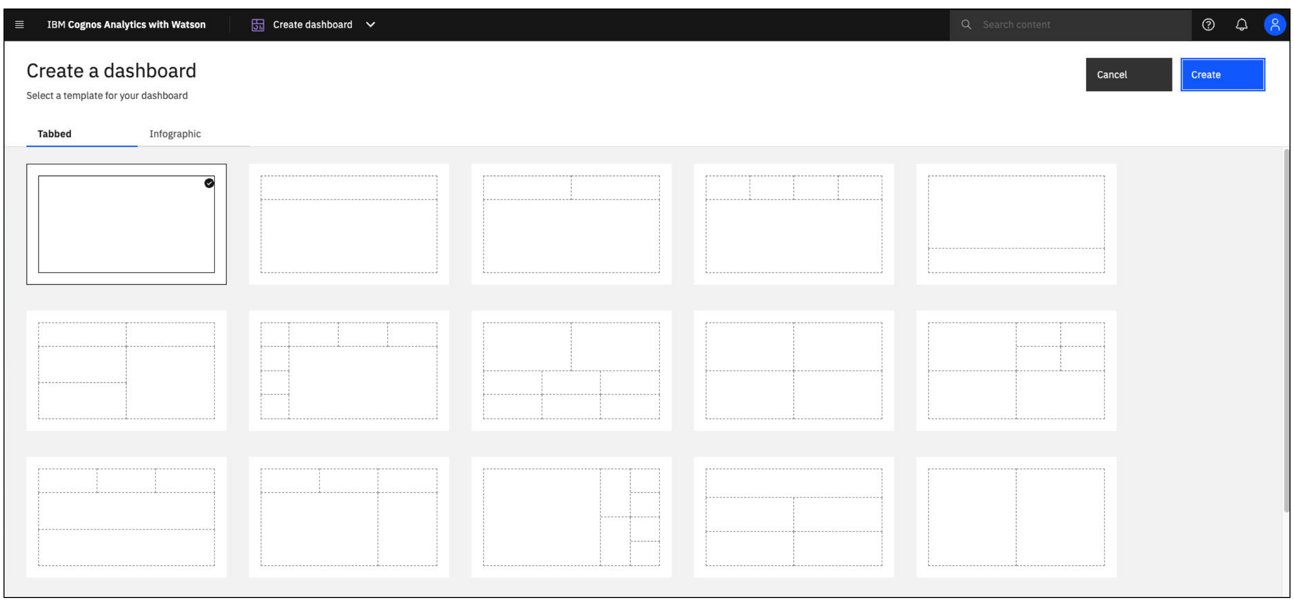

Figure 7-84 Dashboard templates in Cognos Analytics

Figure 7-85 shows the new dashboard. For more information about how to connect a data source to the dashboard, see “Adding a data source” on page 538.

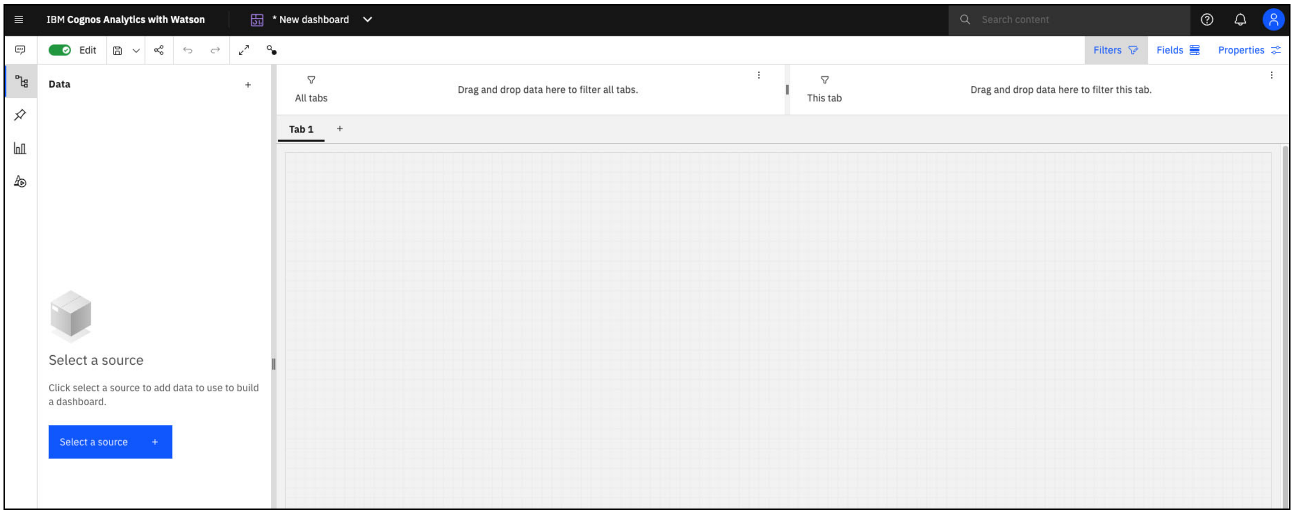

Figure 7-85 New dashboard in Cognos Analytics

Adding a data source

Before any visualizations can be created in a dashboard, a data source must be configured. As shown in Figure 7-85, the Data panel is on the left of the dashboard. Select the **Select a source** option to add a data source to the dashboard.

Then, select the source from My content or Team content, and select **Add**, as shown in Figure 7-86.

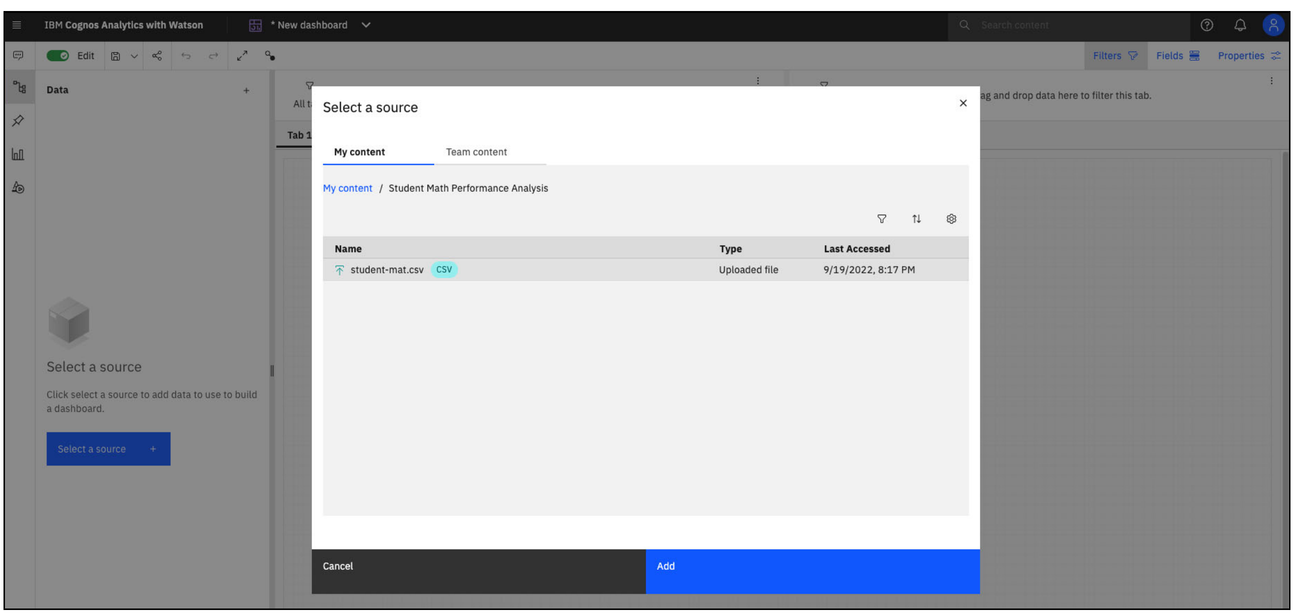

Figure 7-86 Selecting a data source from Student Math Performance Analysis folder

Dashboard properties

To modify properties of the dashboard that are related to areas, such as the canvas, color, theme, and Tabs, select the **Properties** tab on the right of the dashboard, as shown in Figure 7-87. To change the theme of the dashboard, select **Carbon X Dark**, under the Color and theme subheading, as shown in Figure 7-87.

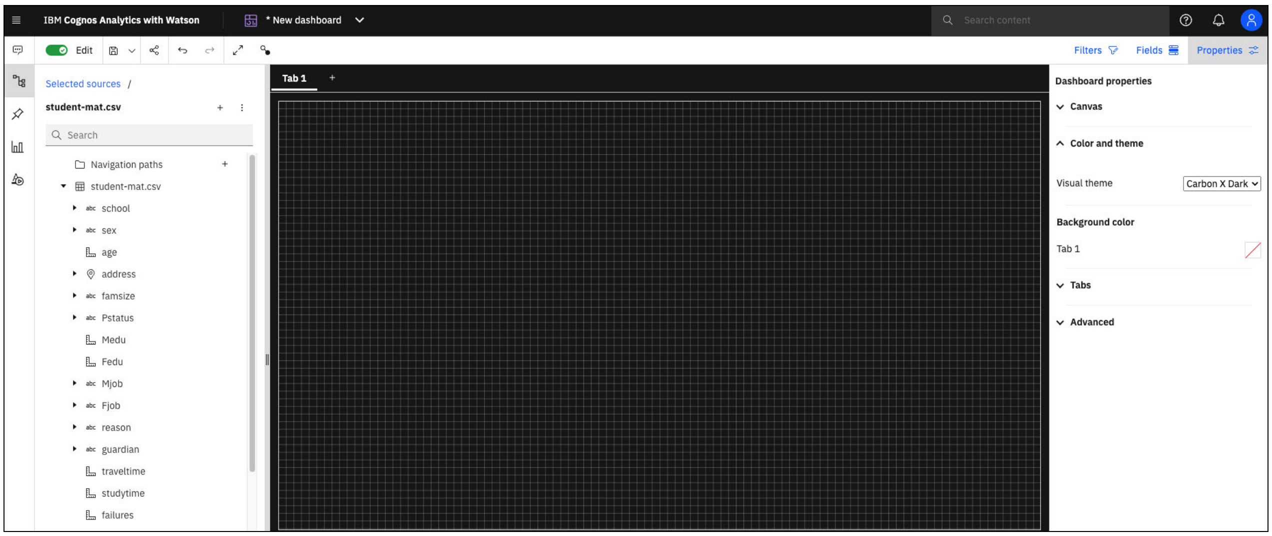

Figure 7-87 Modifying dashboard properties

Adding a calculation

In Cognos Analytics, calculations can be created by using fields in the original data set. A wide range of functions is available to create a calculation.

A common calculation is counting the total rows in a data set, or in the case of this data set, counting the total number of students.

To create a calculation, right-click the `student-mat.csv` data source, and select **Calculation**, as shown in Figure 7-88.

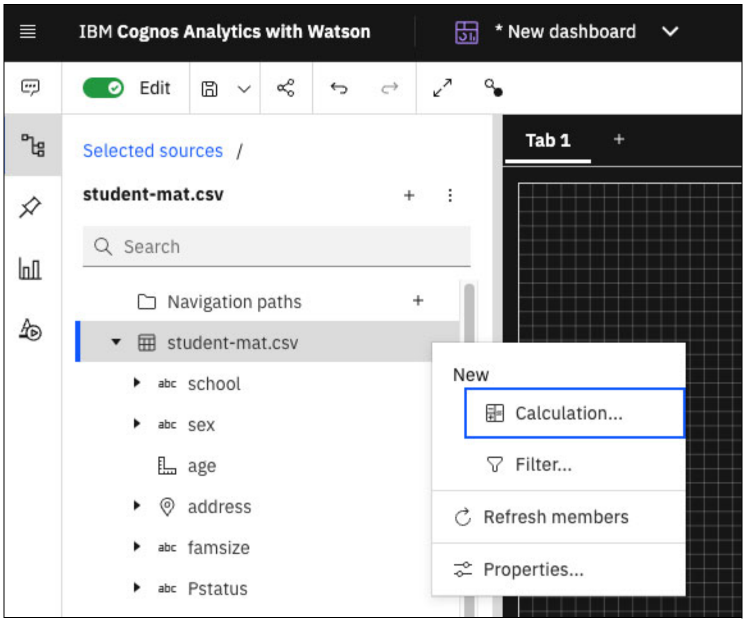

Figure 7-88   Creating a calculation

In the next window, enter the name of the calculation and the calculation expression and then, select **OK**. Figure 7-89 shows an example calculation definition for *Total Students*.

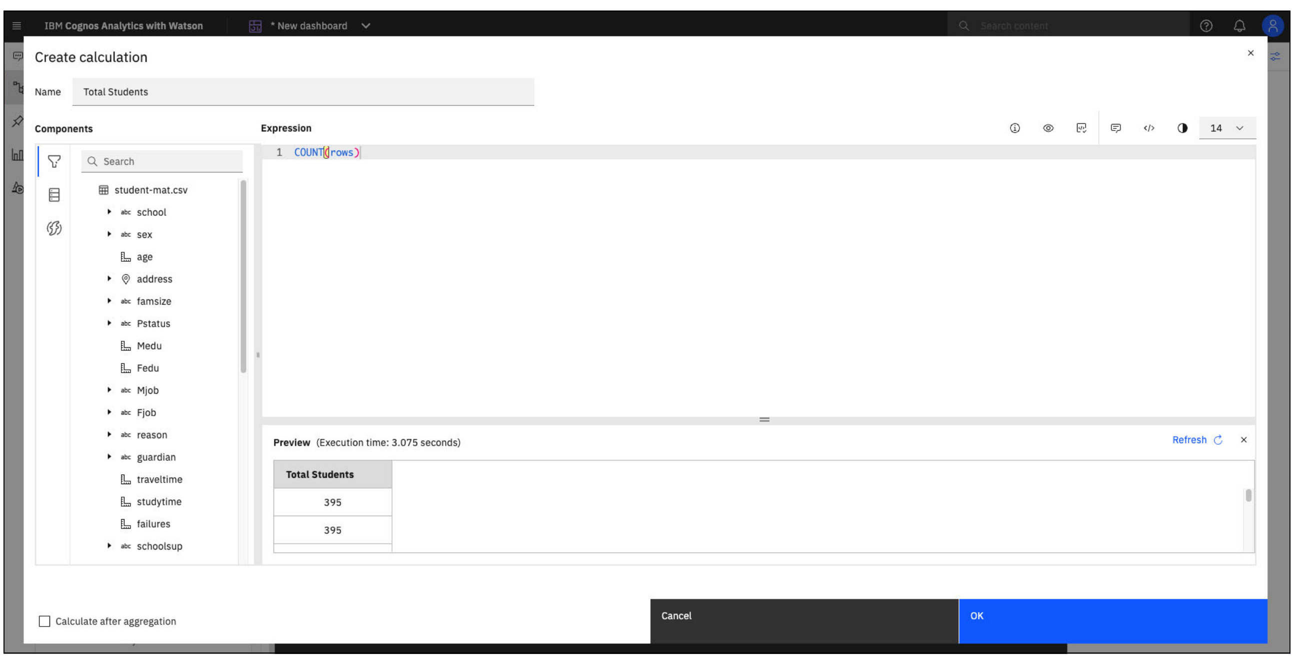

Figure 7-89   Total Students calculation definition

Visualizations tab

Several visualization templates are available, including bar charts, line charts, bubble charts, and network graphs. Each visualization template features specific requirements, depending on the visualization type. To add visualizations to a dashboard, select the **Visualizations** tab on the left, as shown in Figure 7-90.

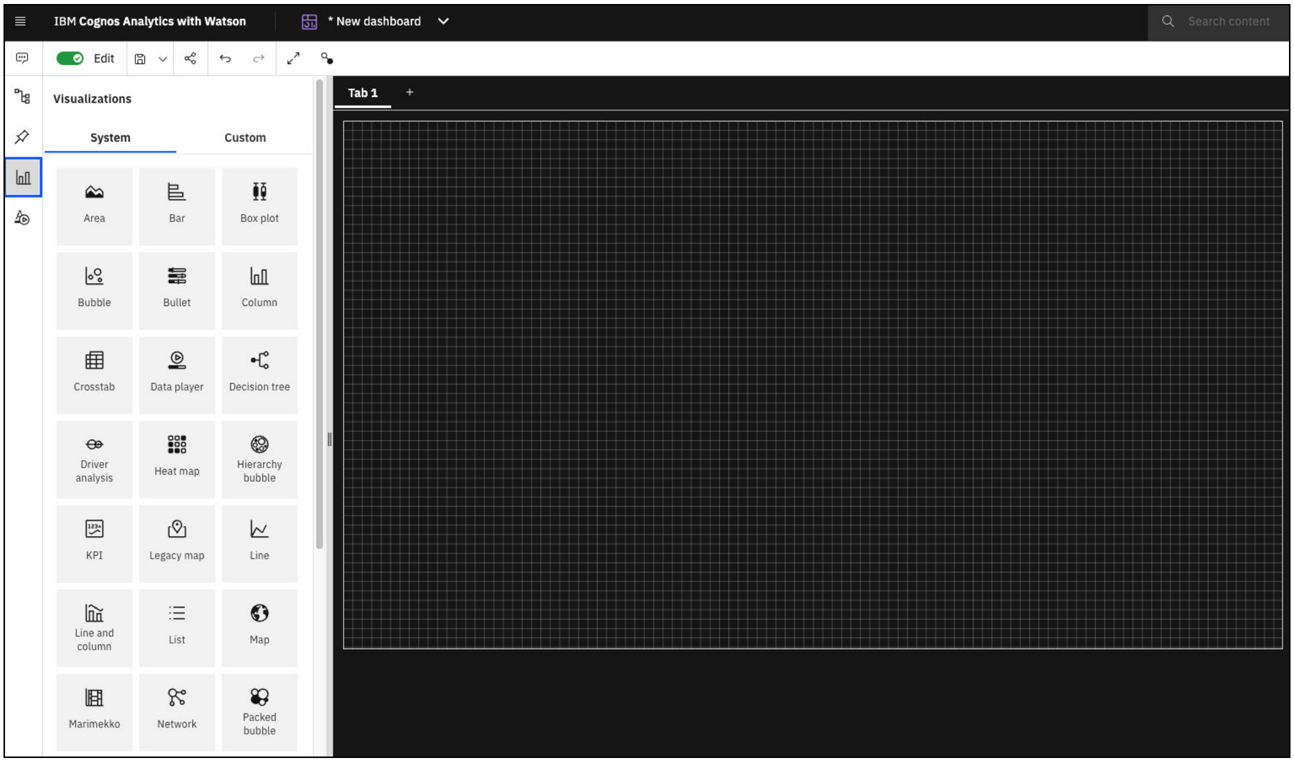

Figure 7-90 Visualizations tab

Custom visualization also can be created and added to a dashboard. For more information about getting started with custom visualizations in Cognos Analytics, see this IBM Documentation web page.

Pie chart example

Each visualization type features properties that can be modified, similar to the properties of the overall dashboard.

Complete the following steps to create and format a pie chart:

1. To add a pie chart, drag the pie chart template from the **Visualizations** tab onto the dashboard. Pie charts require two fields: Segments and Size, as shown in Figure 7-91. From the data panel on the left, drag the `school` field that is under Segments, and the `Total Students` field that is under Size, to create a pie chart that shows the Total Students by School.

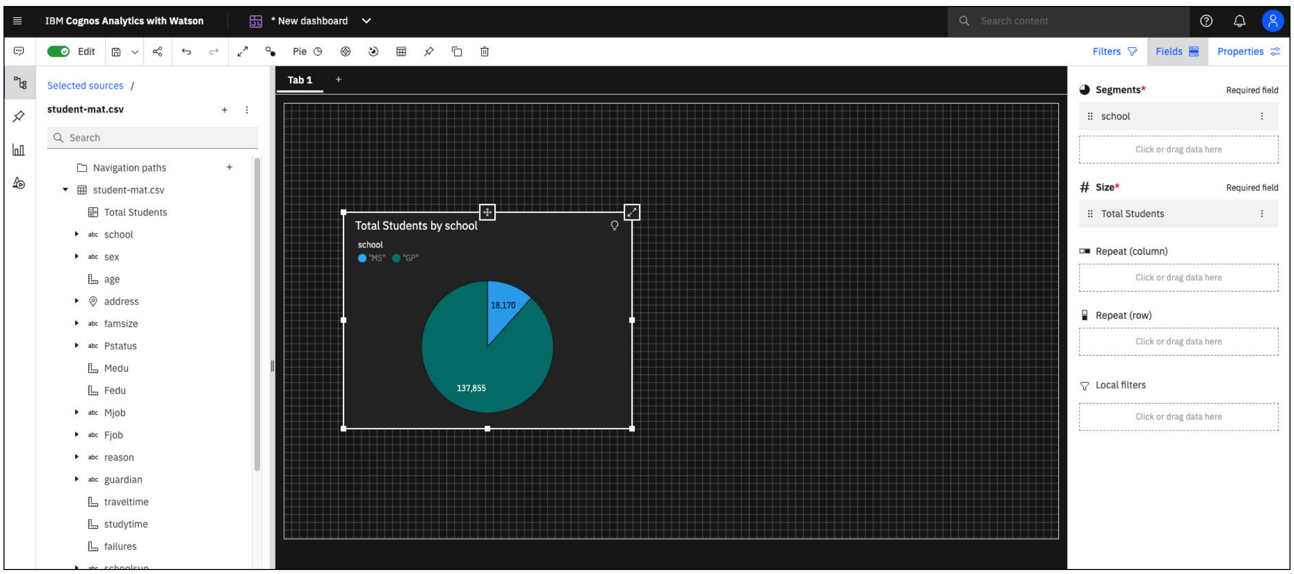

Figure 7-91 Pie chart template

2. To format the pie chart, select the **Properties** tab on the left, as shown in Figure 7-92. This tab provides tools to modify the color, legend, and chart.

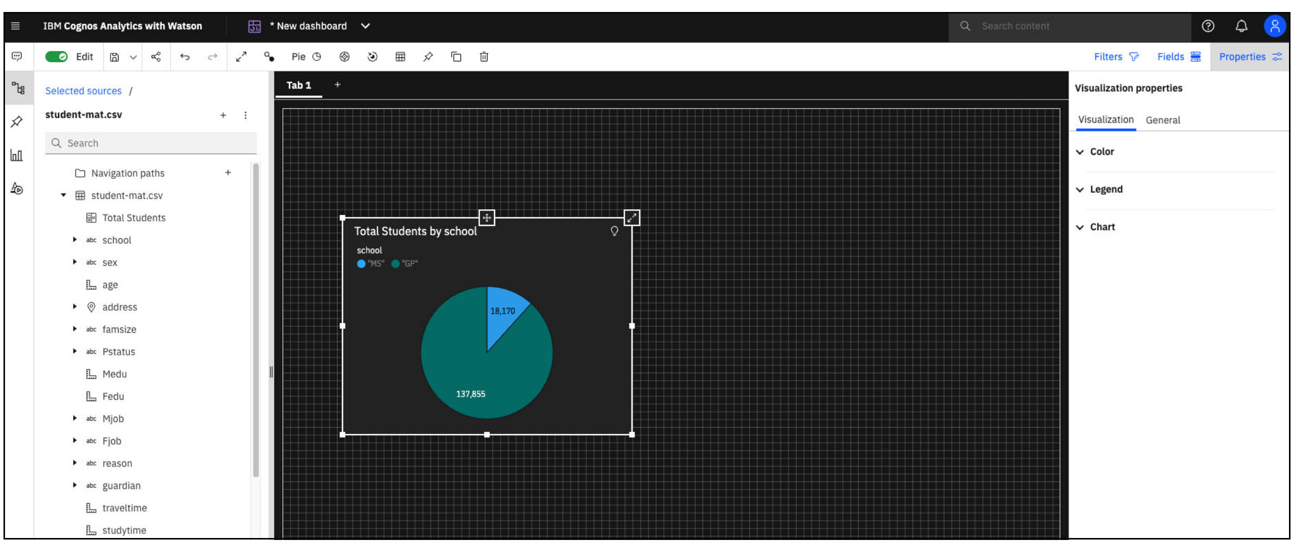

Figure 7-92 Pie chart Properties tab

3. The legend can be moved to the right by selecting **Right** as the Legend Position under the **Legend** subheading, as shown in Figure 7-93.

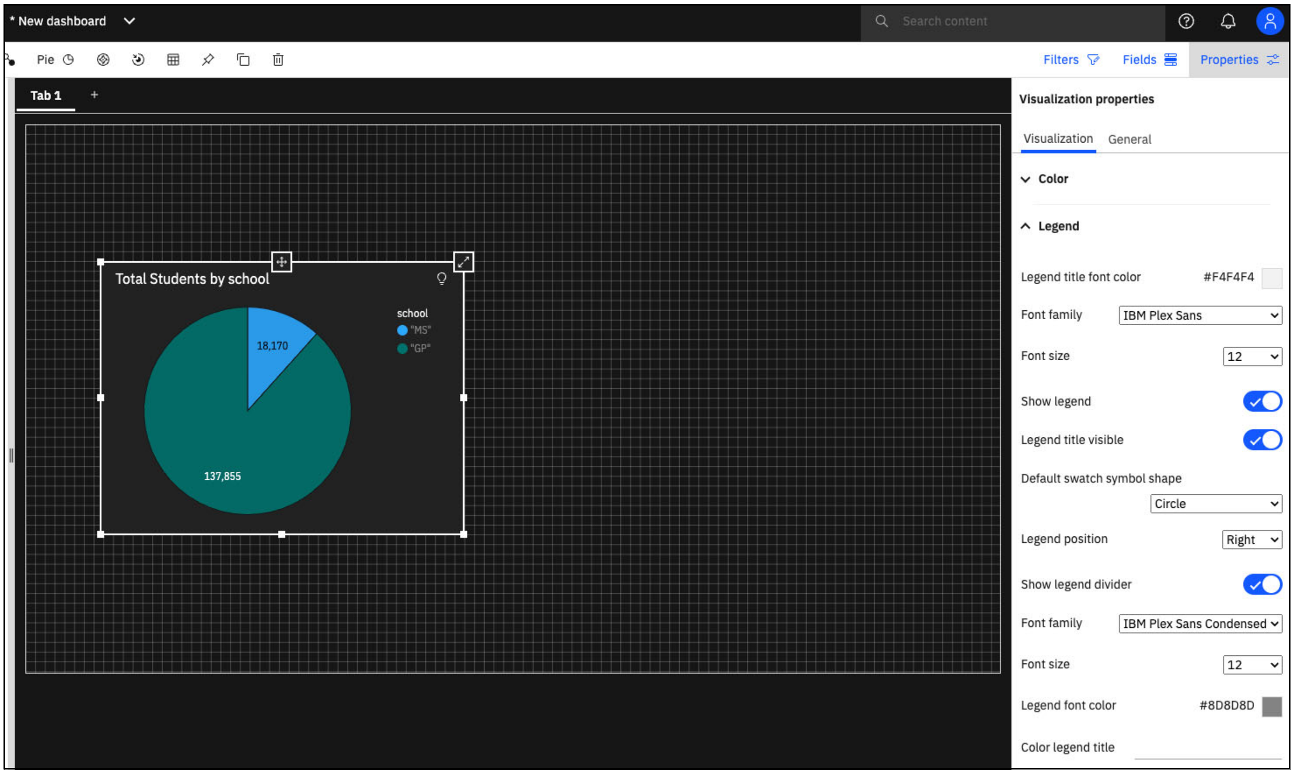

Figure 7-93 Changing legend location on pie chart

4. To change the values in the pie chart from numbers to percentages, select **Display %** under the **Chart** subheading, as shown in Figure 7-94 on page 544. Displaying percentages in a pie chart allows for quick comparison of the individual segments, or parts, to the whole.

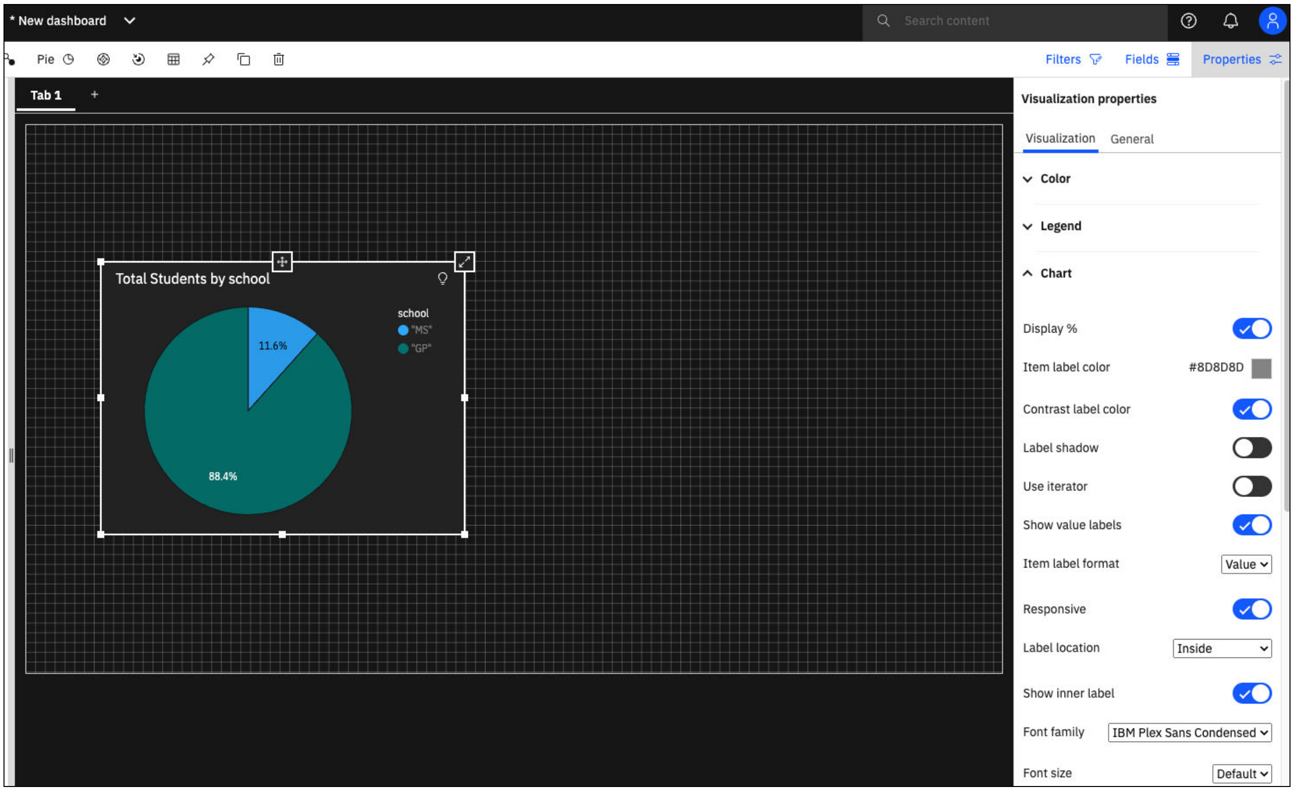

Figure 7-94 Displaying pie chart percentages

5. Figure 7-95 shows how to format the title of the pie chart by highlighting the title text and then, modifying the parameters under **Text details** in the Properties tab. The text also can be formatted by using the options that are directly below the dashboard title. Centering and using bold face type in the chart text can help users quickly identify the purpose of the chart. This feature is especially helpful when visualizations are added to the dashboard.

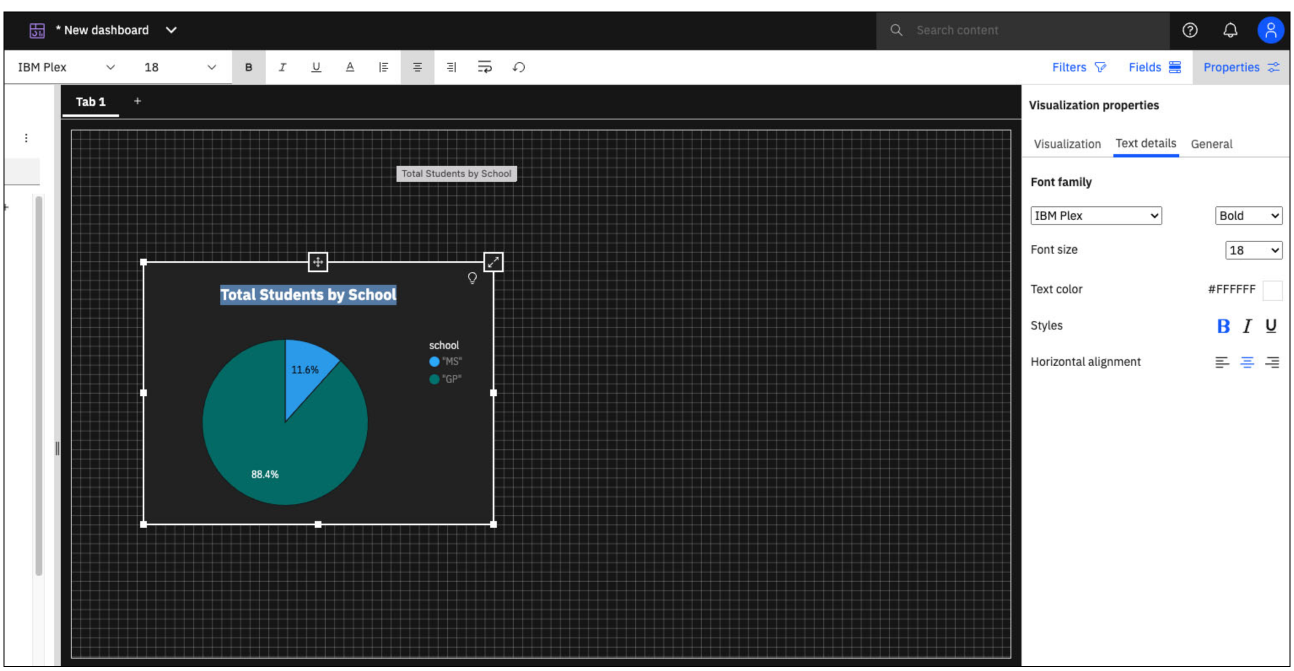

Figure 7-95 Formatting pie chart title

Bar chart example

Each visualization type includes different formatting available options. Complete the following steps to create and design a bar chart:

1. To add a bar chart, drag the bar chart template from the **Visualizations** tab onto the dashboard. Bar charts require two fields, in this case Bars and Length, as shown in Figure 7-96. From the data panel on the left, drag the `reason` field under Bars, and the `Total Students` field under Length, to create a bar chart that shows the Total Students by Reason.

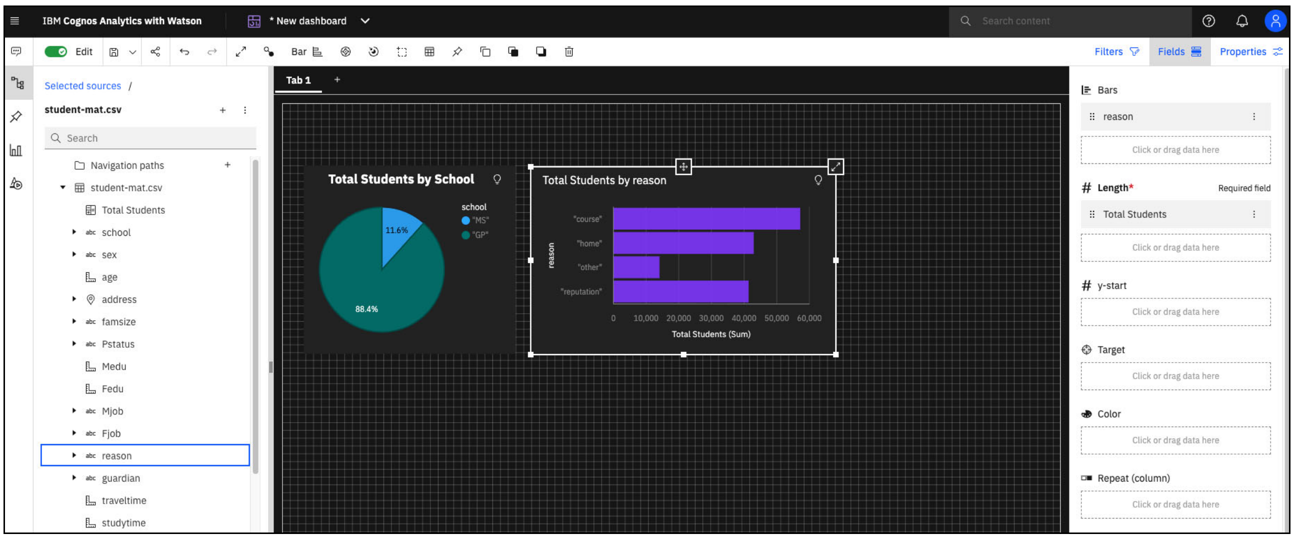

Figure 7-96 Creating bar chart in Cognos Analytics

2. Bar charts can be sorted alphabetically or by value of each category. Sorting by the value in each category enables users to quickly identify which category has the most value and which category has the least value. Figure 7-97 shows how to sort the bar chart by Total Students in descending order.

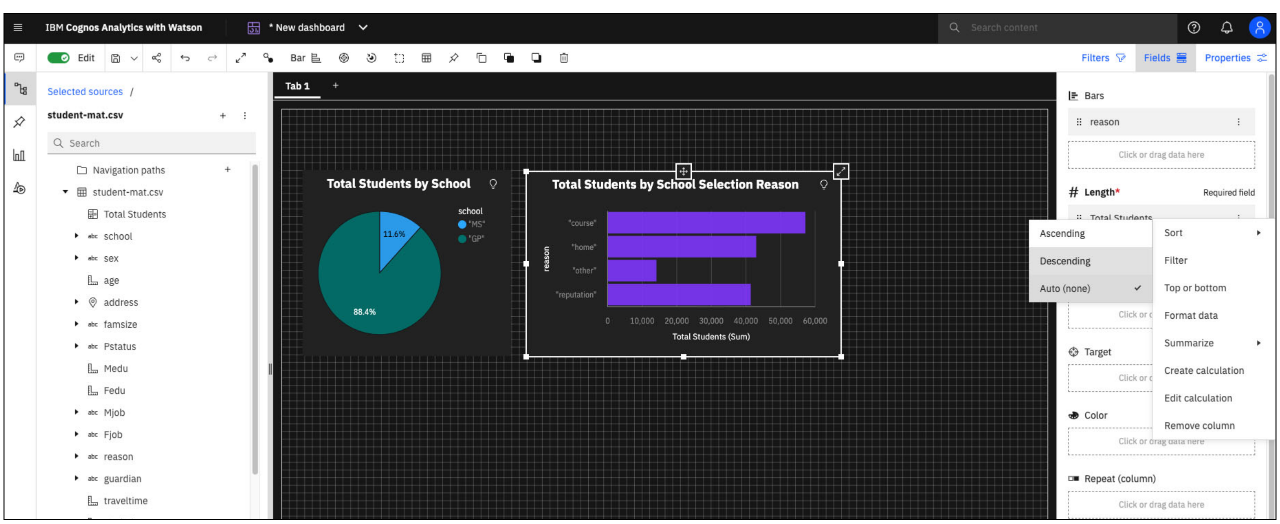

Figure 7-97 Sorting bar chart in descending order

3. Hovering over the different segments in the bar chart shows the value for each category. The values are summed by default. To change this setting to reflect the count, select the field name that is under **Length** and then, select **Summarize** and then, select **Count,** as shown in Figure 7-98.

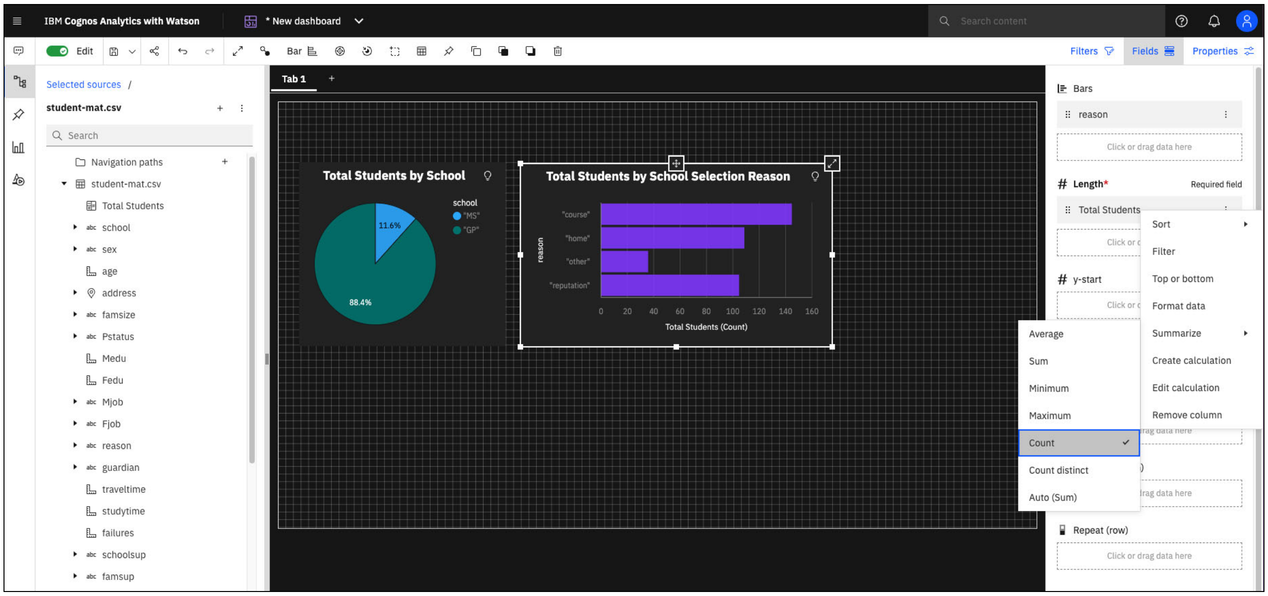

Figure 7-98 Changing Total Fields Summarization Method

4. Adding value labels to a bar chart makes it easier for users to readily identify the values for each category without having to hover over or select a specific bar.

   Figure 7-99 shows how to add values labels to a bar chart by selecting **Show value labels** under the Properties tab.

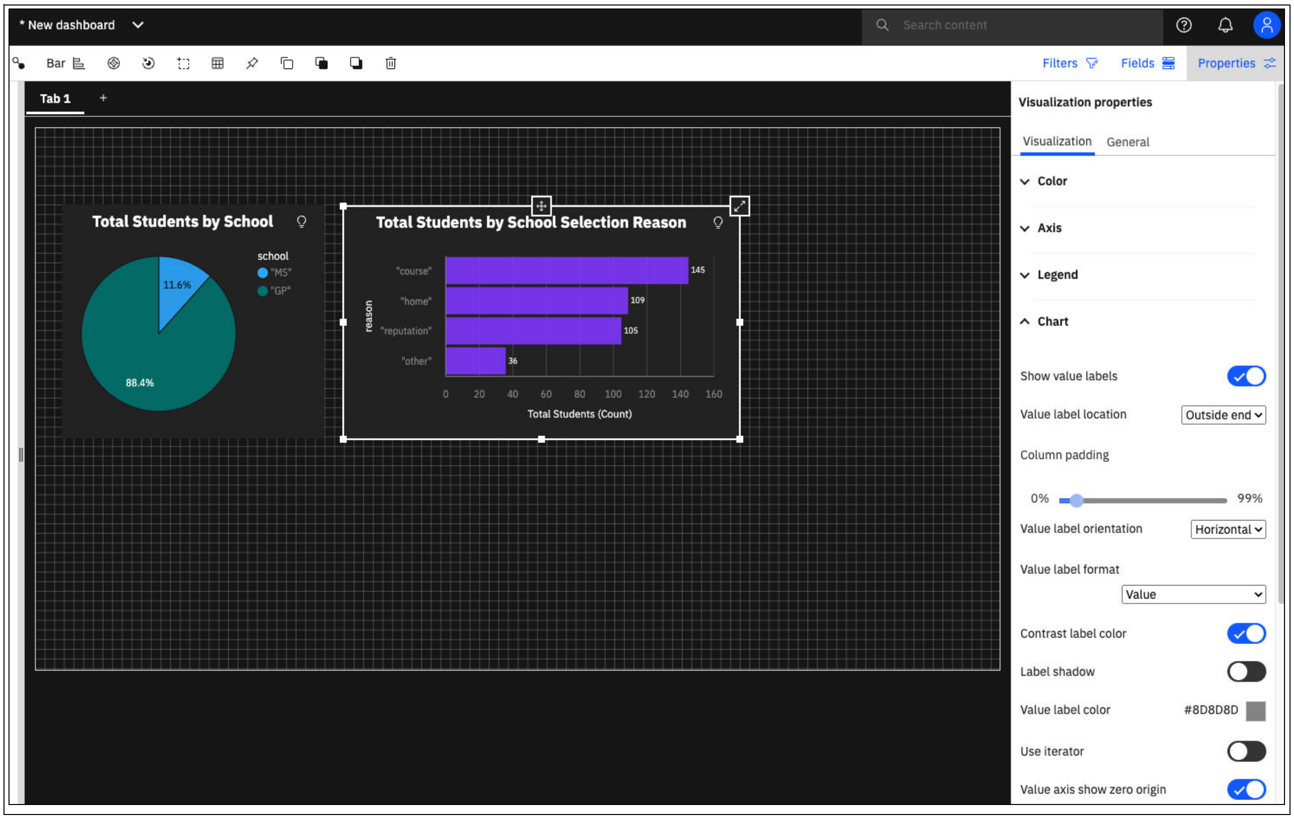

Figure 7-99 Adding value labels to bar chart

5. Similar to chart titles, the text for chart axes also can be modified. Modifying the default settings can increase readability of the chart.

   To modify the text of an axis title, right-click the axis title and then, select **Style text**, as shown in Figure 7-100.

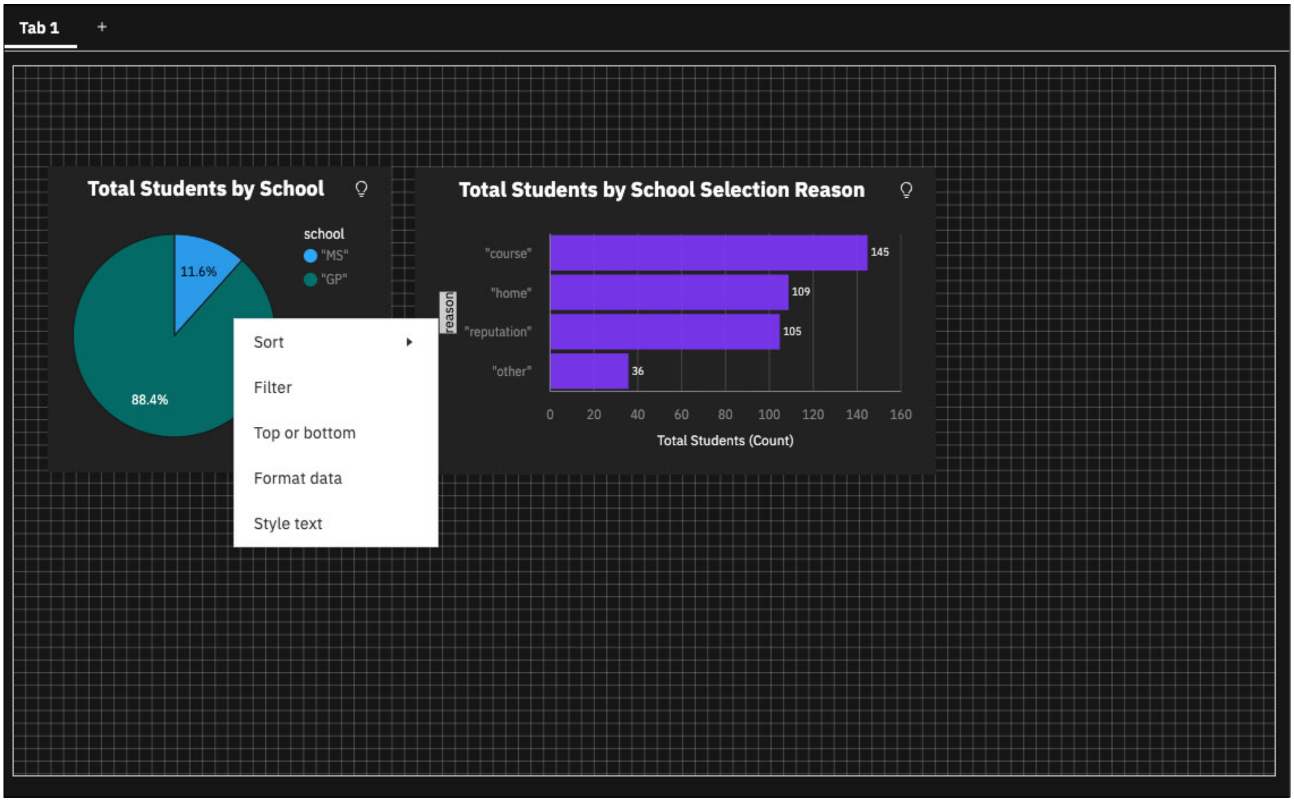

Figure 7-100 Bar chart axis title options

6. Select the font style and size. Figure 1-93 shows how the title of the reason axis is changed to School Selection Reason, and is displayed in bold font.

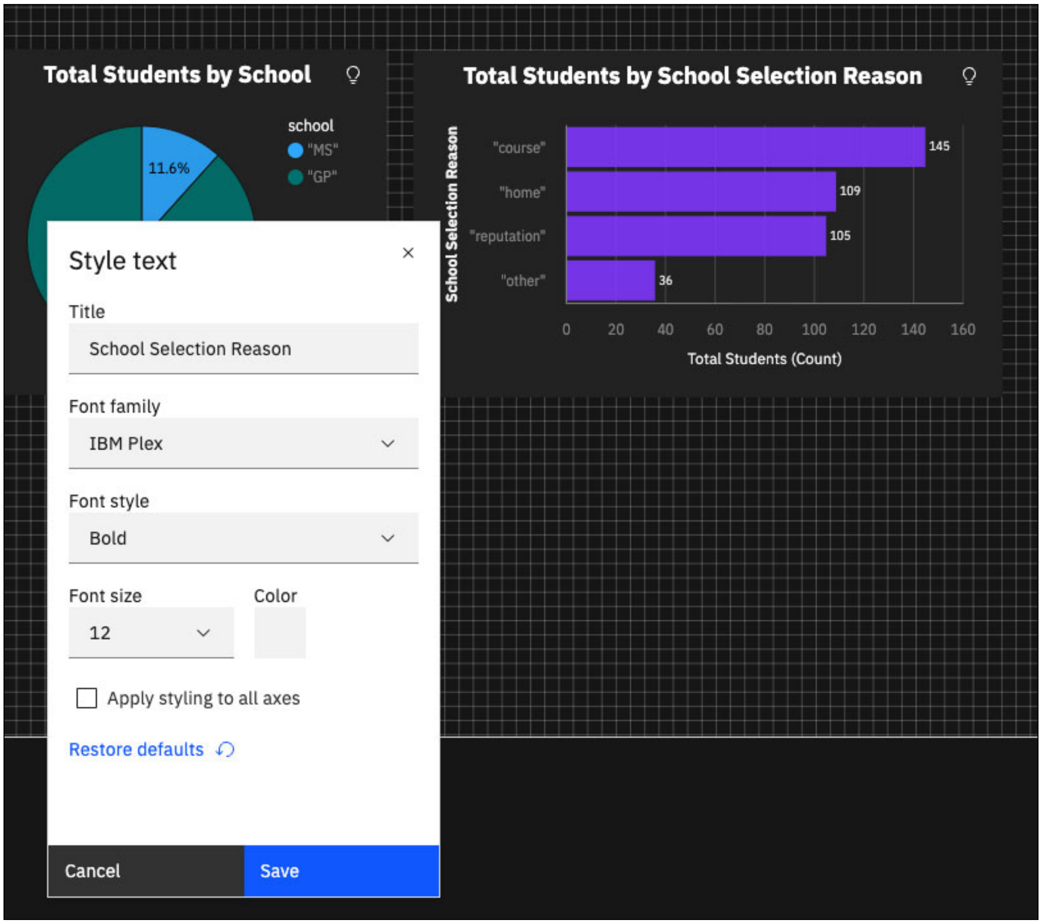

Figure 7-101 Changing text format of reason axis

Widgets

In addition to visualizations, Cognos Analytics provides various widgets that can be added to dashboards. Widgets include text boxes, images, webpages, and shapes. Figure 7-102 shows the Widgets tab on the left of the dashboard.

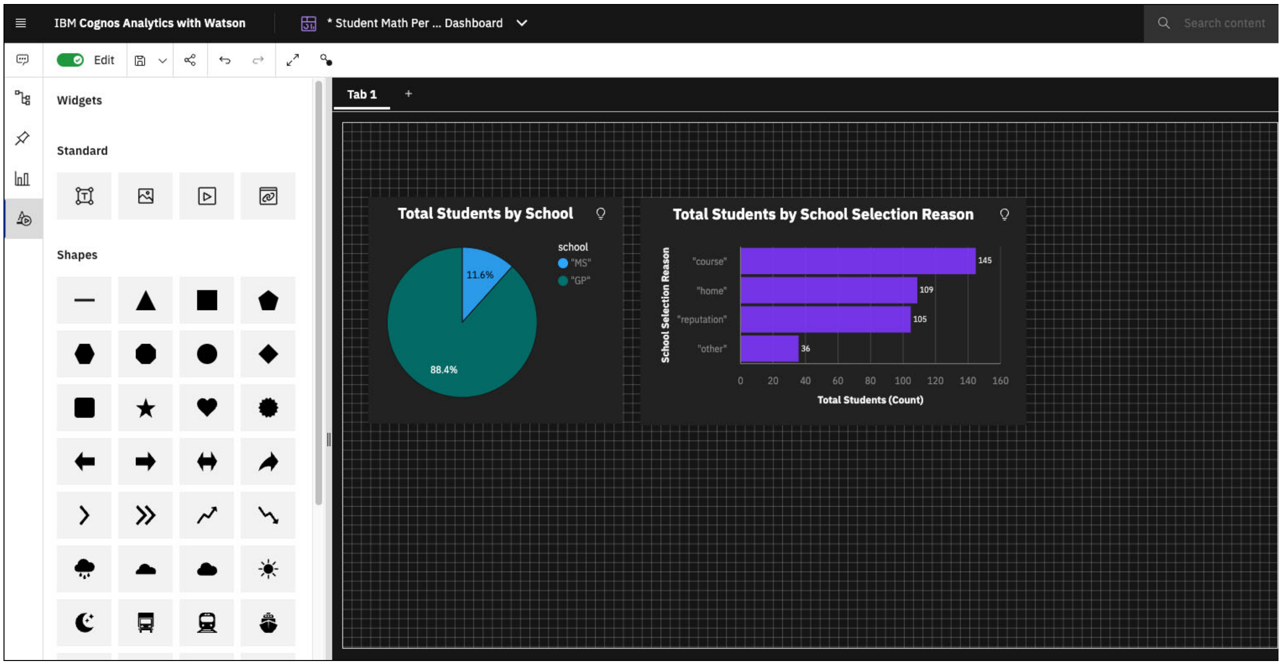

Figure 7-102 Widgets tab in Cognos Analytics

Text box widgets can be used to add titles and other context to a dashboard. Figure 7-103 shows how the text box widget is used to add a dashboard title, Student Math Performance.

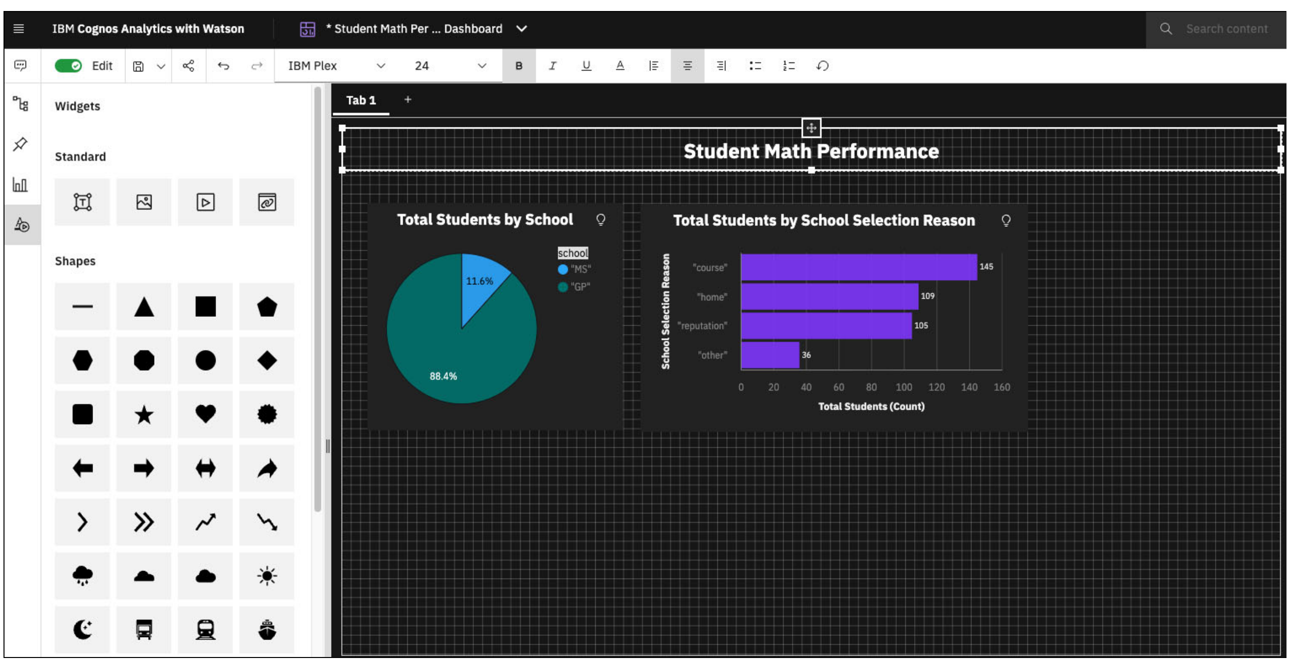

Figure 7-103 Adding dashboard title with text box widget

Tabs

Tabs can be used to organize visualizations in a dashboard. To add a tab to a dashboard, select the **+** icon that is next to the tabs. To edit the title of a tab, double-click the title name and select **Edit the title**, as shown in Figure 7-104.

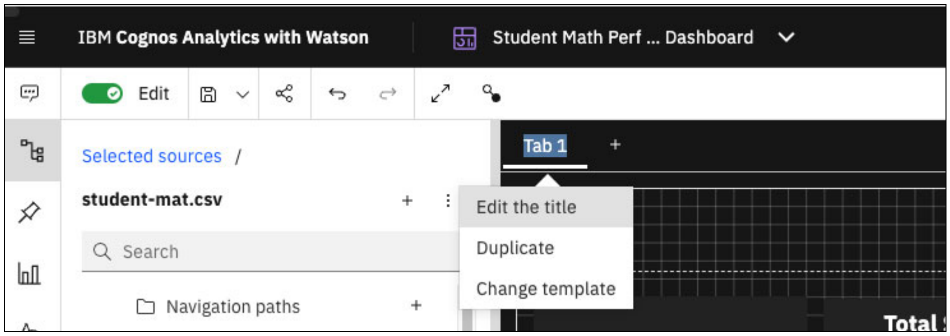

Figure 7-104 Selecting the Edit the title option

Filters

Visualizations on a dashboard can be used to filter the data. In addition to filtering by using the charts, fields can be placed in the filters panel at the top, as shown in Figure 7-105. In this example, the field `Internet` is applied as a filter to one specific tab. Filters can be applied to one tab or all tabs within a dashboard.

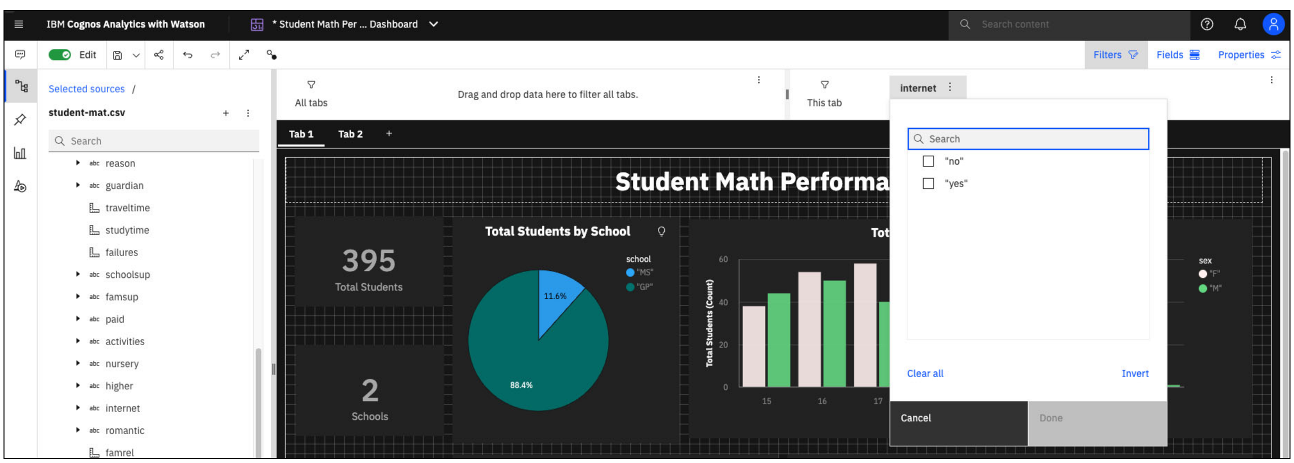

Figure 7-105 Adding Internet Field as Dashboard Filter

Recap

After the dashboard is created, it can be saved and shared with users.

This use case highlighted a few features in Cognos Analytics. For more information about features and capabilities in Cognos Analytics, view this IBM Documentation web page.

7.3.4 Use case #4: Planning Analytics

This use case shows Planning Analytics (PA) functions, and uses PA visualizations to deliver on enterprise requirements.

In this example, a Cognos TM1 sample database with the fictitious 24Retail organization is used. This data is included with the TM1 Server product, and features cubes and dimensions that are configured by using Planning Analytics Workspace and TM1.

For more information about the database samples in TM1, see this IBM Documentation web page.

Scenario overview

The scenario is that you are a financial analyst at 24Retail, a consumer products company. Your manager asked you to analyze trends and provide the details for Q4 Rent by Month, with Actuals and Budget.

You know your data well, and thus know that this process can be done readily in the Planning Analytics interface. The cubes, dimensions, and other facets of the TM1 database you often use include all of the information that is required for this process, and Planning Analytics makes it easy without having to code.

A workbook was created that contains the Income Statement data that is required, which is used to get a fast start on the effort. Complete the following steps:

1. To start Planning Analytics, begin in the Cloud Pak for Data interface Instances list and open the **Planning Analytics Instance**, as shown in Figure 7-106.

Figure 7-106 Planning Analytics Instance in Instances interface in Cloud Pak for Data 4.5

2. Opening the instance opens the new browser window for Planning Analytics, as shown in Figure 7-107.

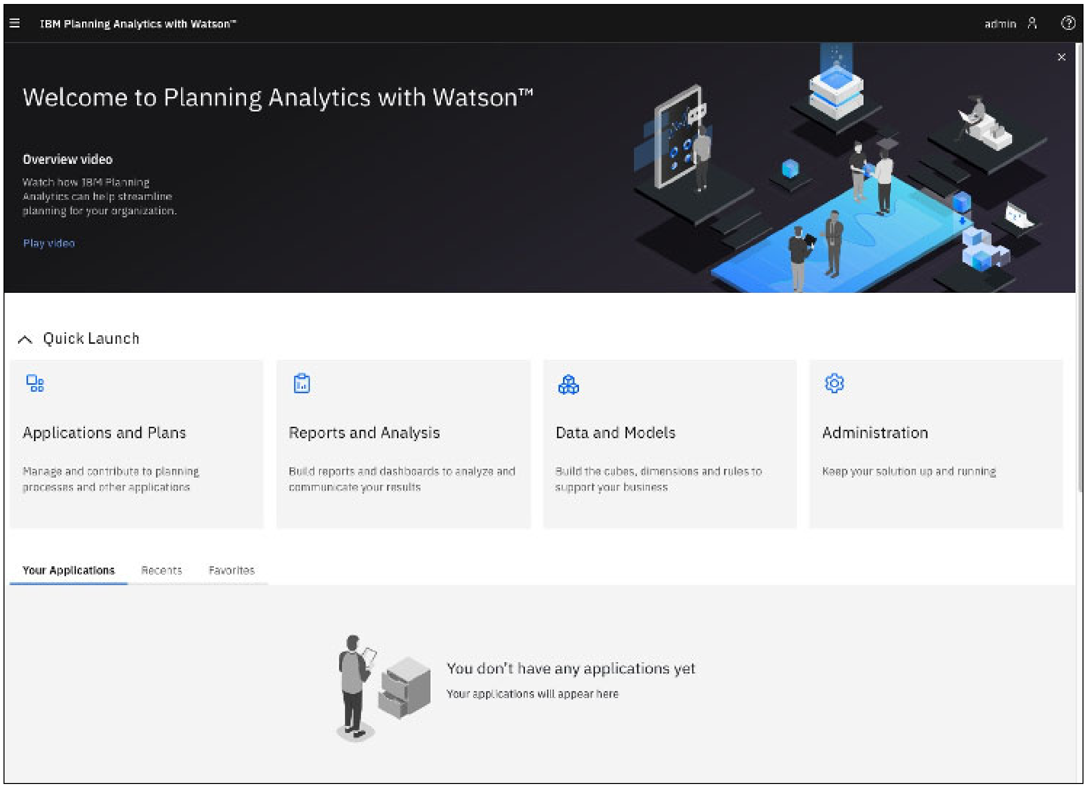

Figure 7-107 Planning Analytics with IBM Watson service home page

3. Initially after opening the Planning Analytics service in its own browser window (see Figure 7-108), select **Reports and Analysis** from the home page. From here, you can access the downloaded snapshot Deep Dive HOL to use to create your own customized visualization to answer the manager's question. Although much data is available to use, you know that you must use the Income Statement view.

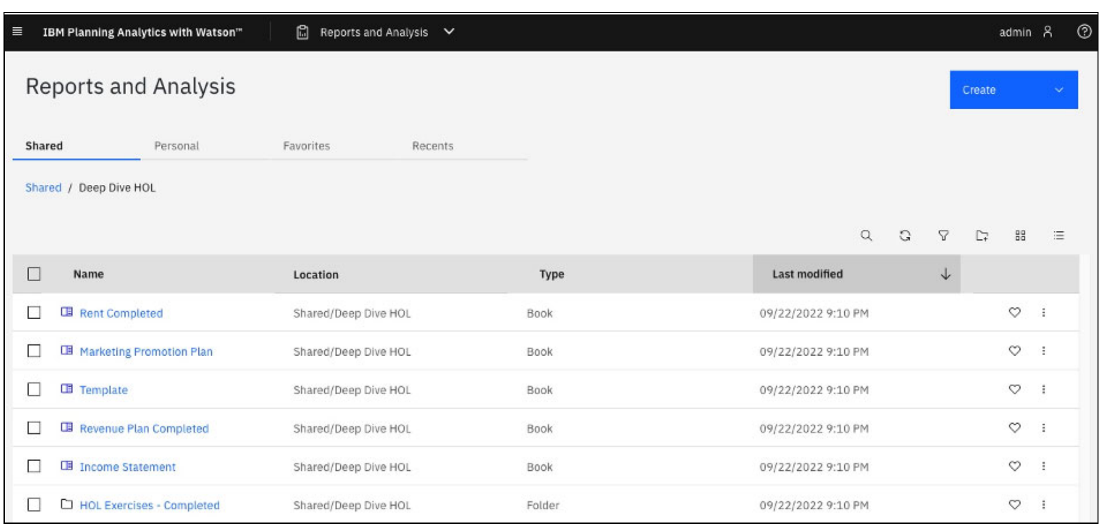

Figure 7-108 Reports and Analysis page where the object required is selected

You now see the opened Income Statement workbook view that was created with the income and expenses for the company, as shown in Figure 7-109.

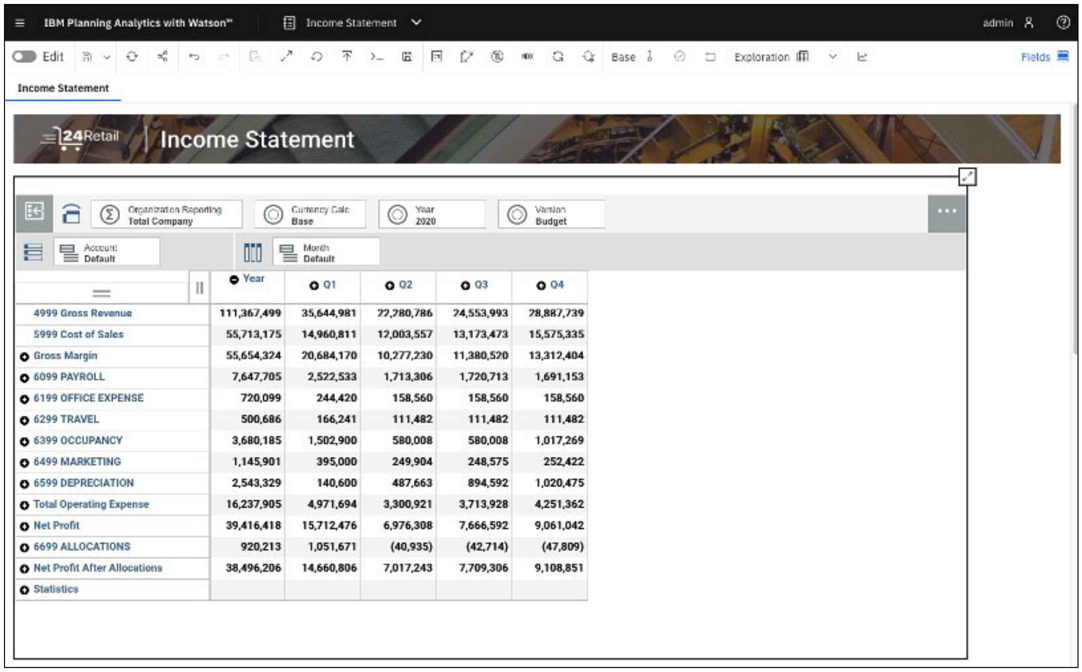

Figure 7-109 Income Statement view found in the sample 24Retail database

4. Because you want the Month dimension for Q4 by month, you can modify the Month dimension to choose **Q** for quarters, as shown in Figure 7-110.

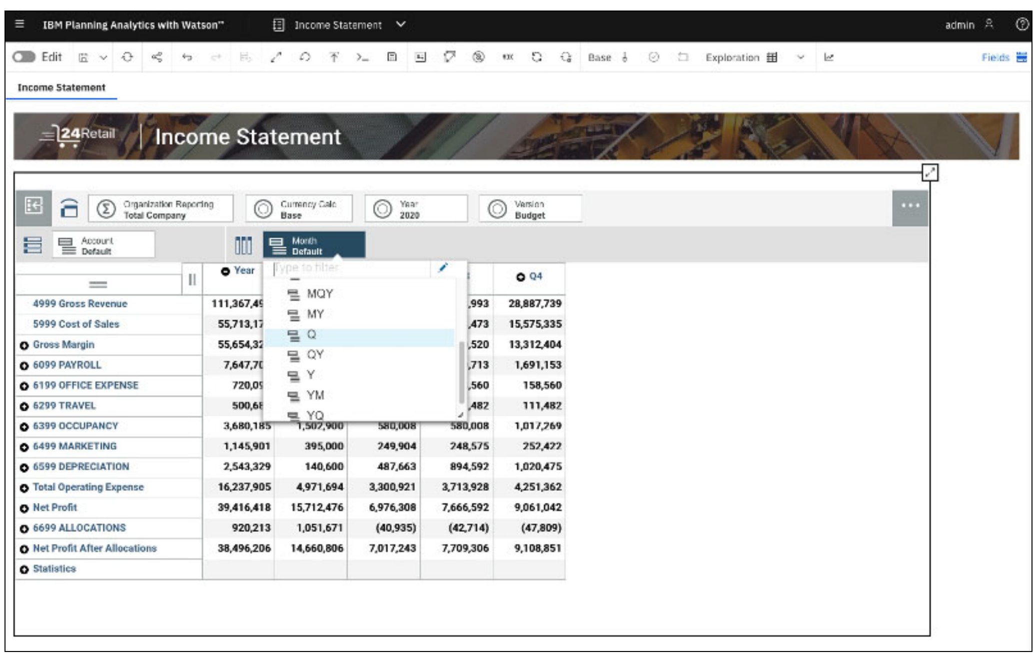

Figure 7-110 Selecting the correct dimensions for viewing by quarter

5. You also want to look at rent only; therefore, drill into the **Occupancy** dimension in the rows to get only facets of Occupancy, which includes Rent (and Utilities and Maintenance summaries), as shown in Figure 7-111.

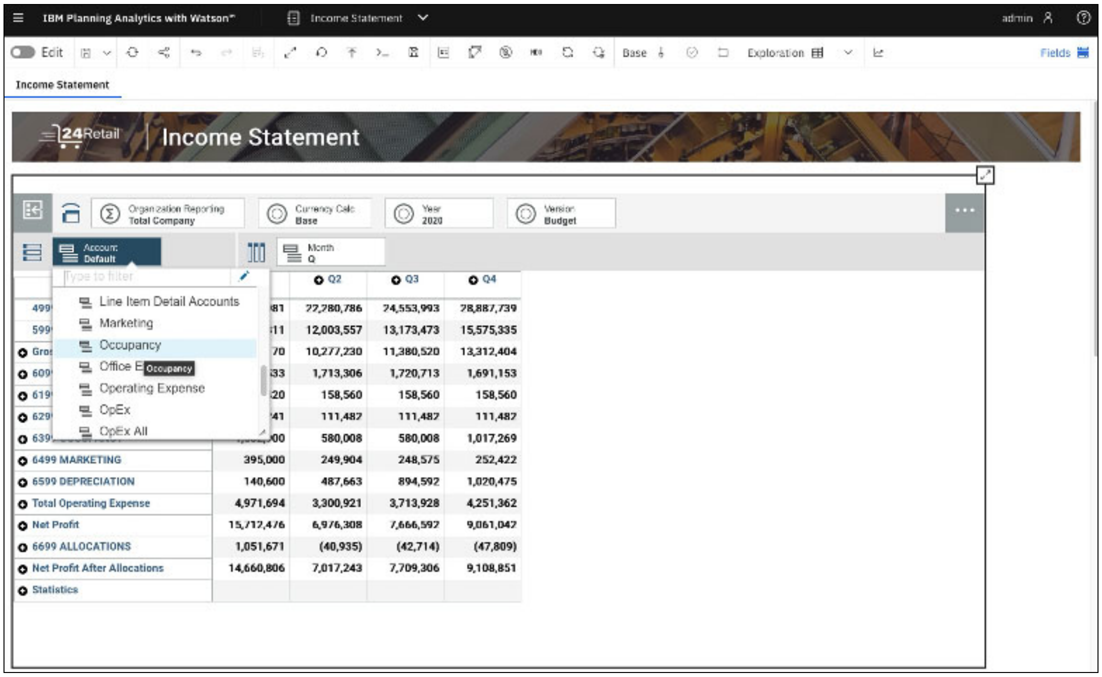

Figure 7-111 Refining the view to only Occupancy dimensions

The resulting view is reduced to Occupancy, including Rent by quarter, as shown in Figure 7-112.

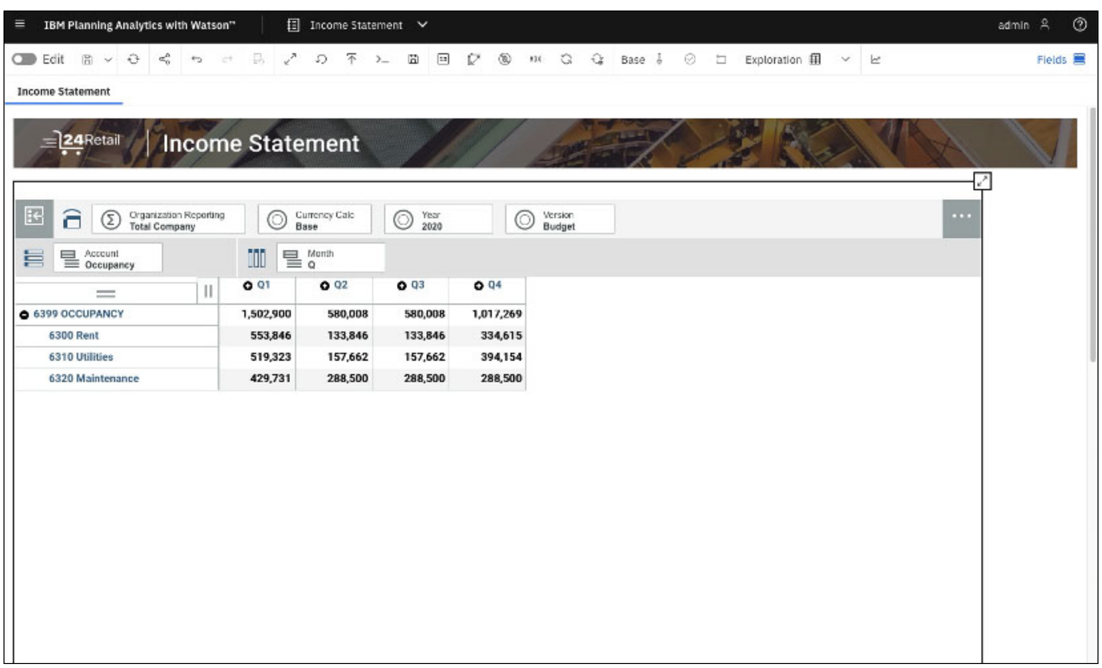

Figure 7-112 Occupancy figures by quarter, including Rent

6. The original request was to look at the quarterly Rent by Actual and Budget for the entire company. To do so, drag the **Version:Current** dimension down to the columns section of the dashboard. Also, shuffle the `Month:Q` dimension over to the rows section for easier viewing. The result looks like the example that is shown in Figure 7-113.

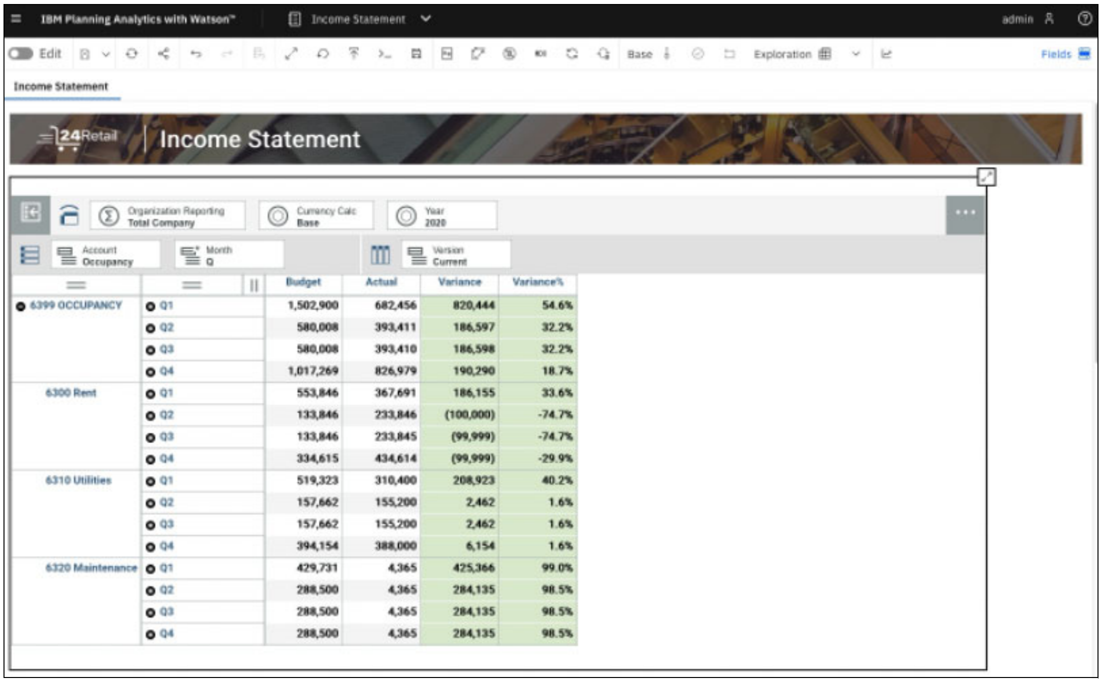

Figure 7-113 Rearranging the dimensions and fields for easier viewing

7. To review this data from a visualization and trends perspective, you can use the built-in visualization features to convert it into a line graphic for better interpretation. To do so, click the **Exploration** link in the top toolbar and select the **Line** option (see Figure 7-114).

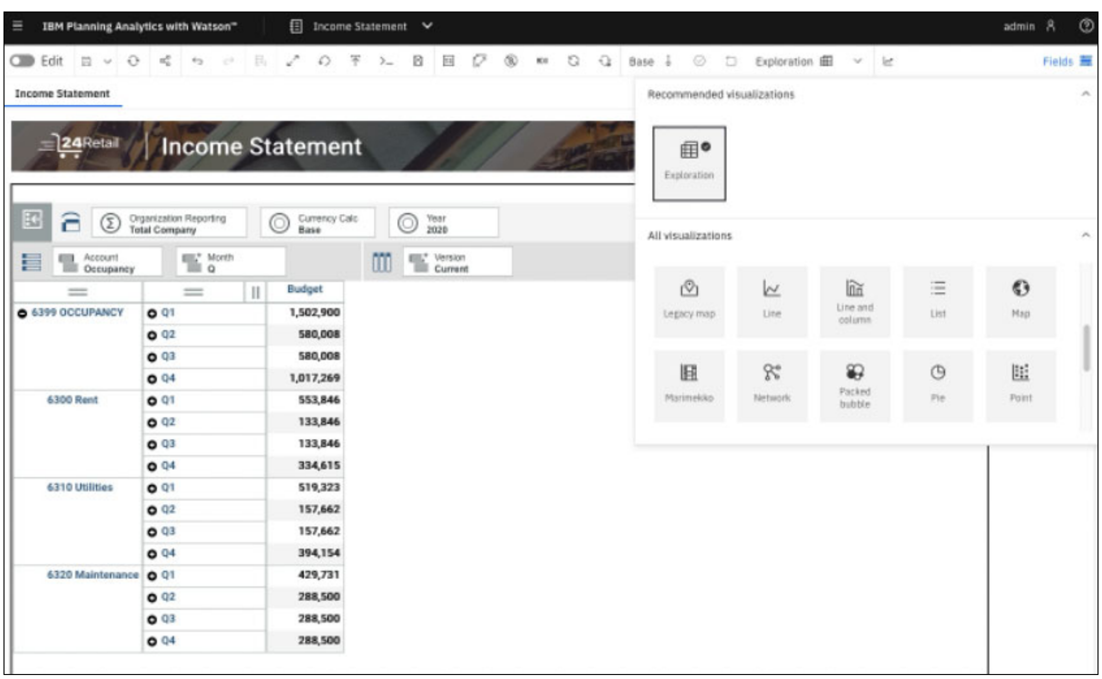

Figure 7-114 Visualization selection

The resultant line graph is shown in Figure 7-115.

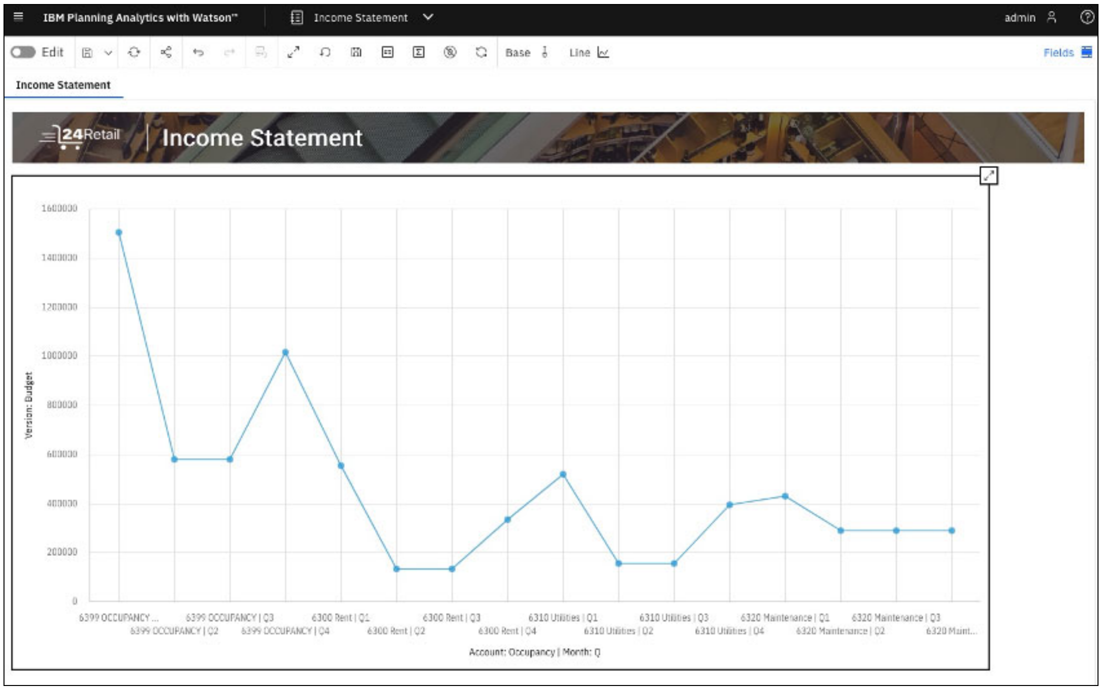

Figure 7-115 Line graph for the Rent data

8. Because the line chart does not look helpful as a visualization, you want to convert it to something more visually pleasing and interpretable. Click the **Fields** button that is in the upper right of the window to open a new view, as shown in Figure 7-116. This view includes a set of customizations that can be used to refine the graphic.

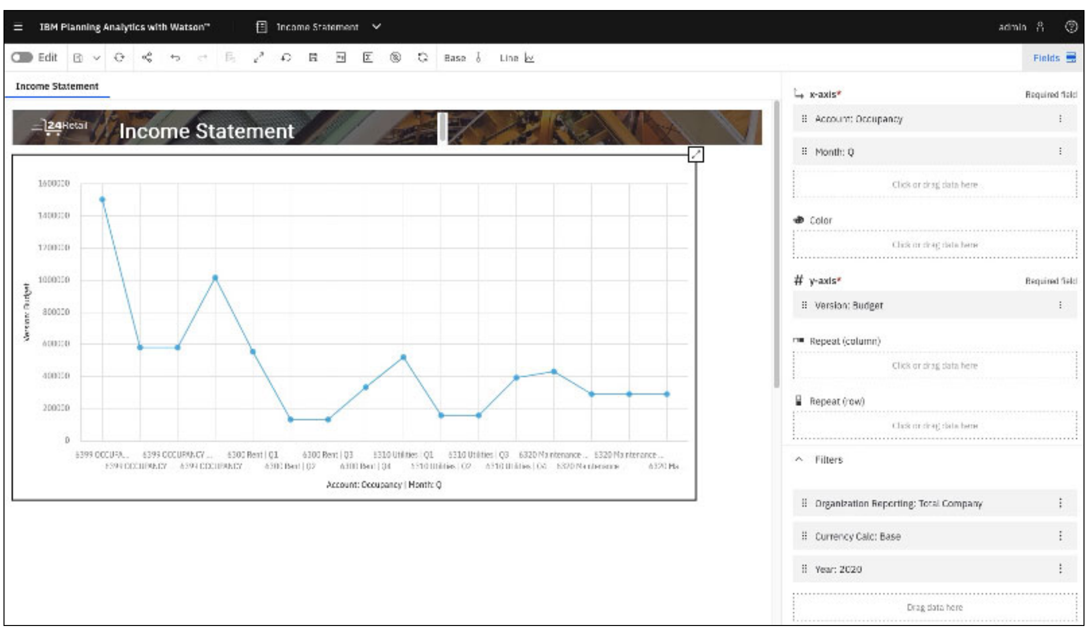

Figure 7-116 Activating the Fields interface to refine the view

9. With the Fields interface, you continue to drill down to make your visualization even more specific. For example, we refine to look at just Q4. Clear the **Fields** view on the right, click the blue **Occupancy** line and then, right-click the same line to open a drill-down menu, as shown in Figure 7-117.

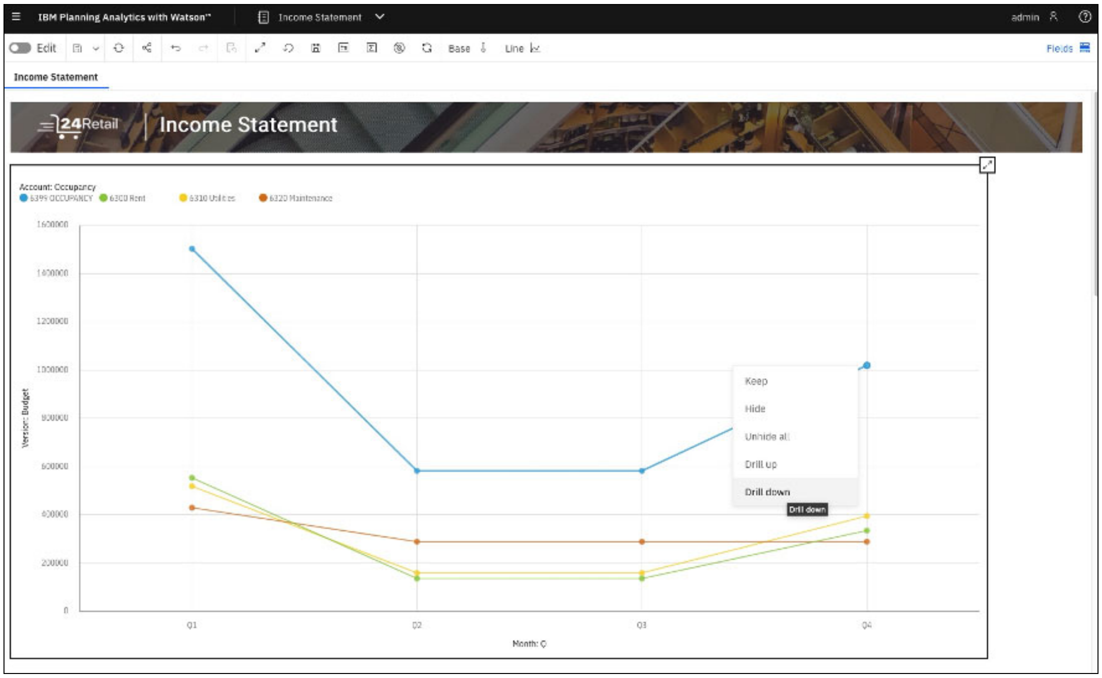

Figure 7-117 Drilling down into the suitable data scope

10. A dialog window opens. Select **Q4** and then, click **Drill Down** (see Figure 7-118).

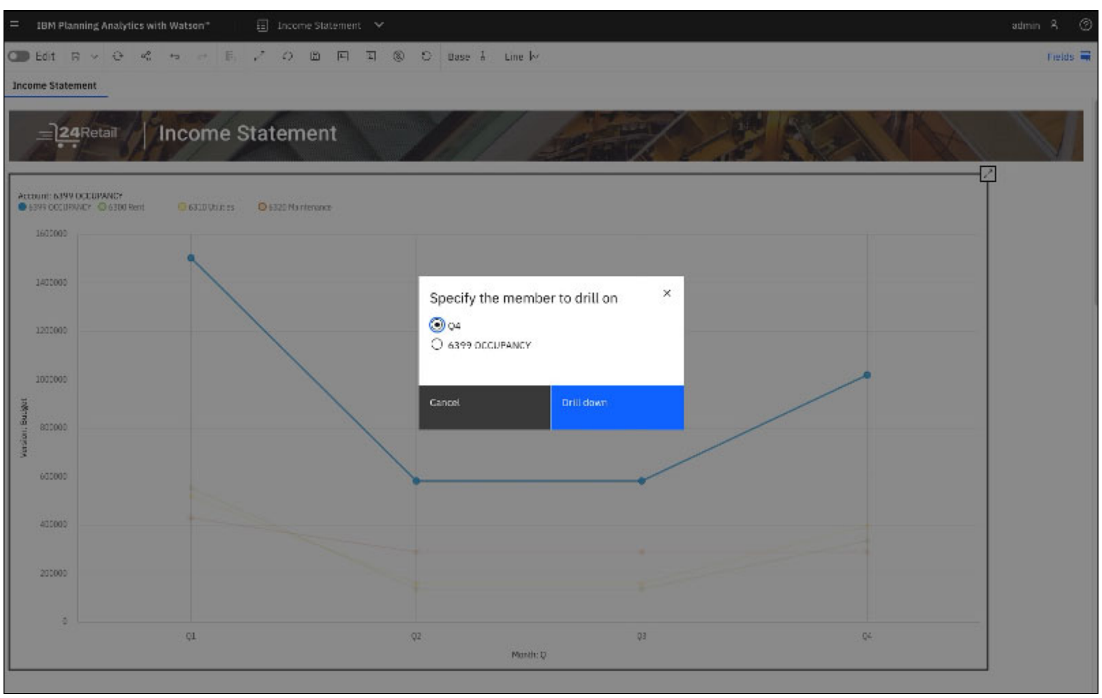

Figure 7-118 Drilling down into specific data values

11.The result looks like the example that is shown in Figure 7-119. Remove the total so that you can look at only the months in Q4, as shown with the selector dialog that is in the upper left.

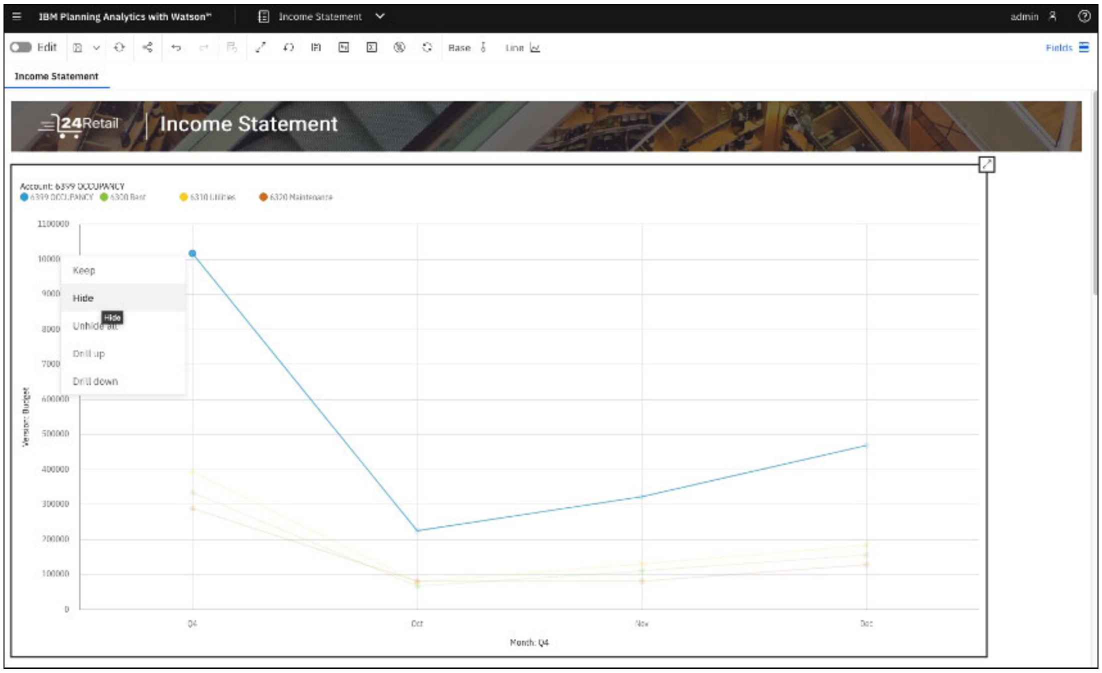

Figure 7-119 Line graph of Rent by quarter total and quarter, with filter dialog to remove total

12.Specify the member to hide (the Q4 total), as shown in Figure 7-120.

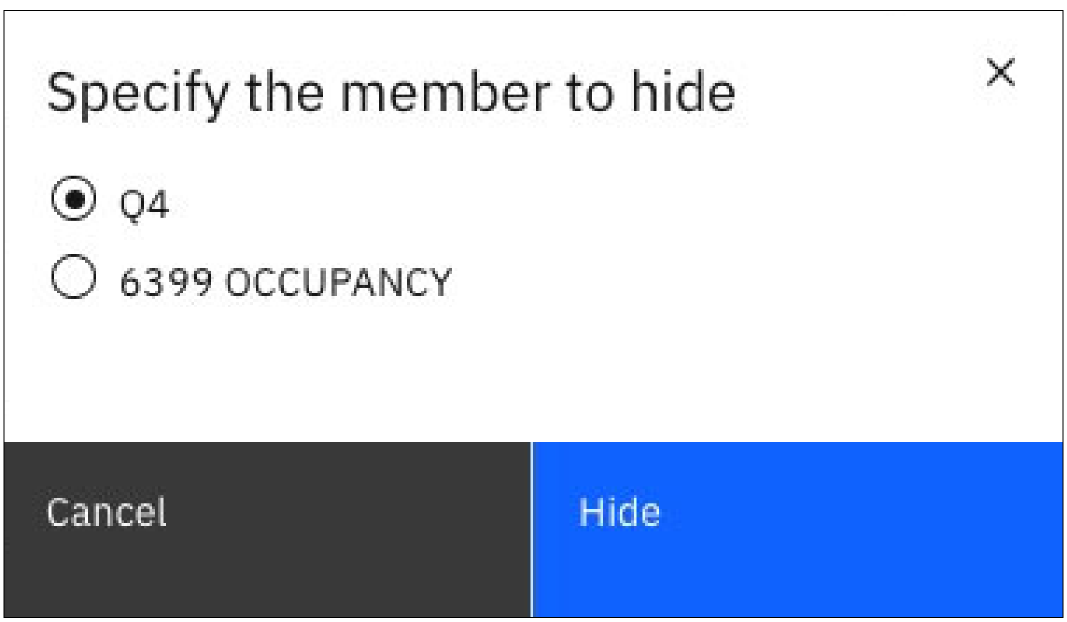

Figure 7-120 Choosing Q4 as the member category to hide in the data

13. Highlight only the **Rent in Q4**, as shown in Figure 7-121.

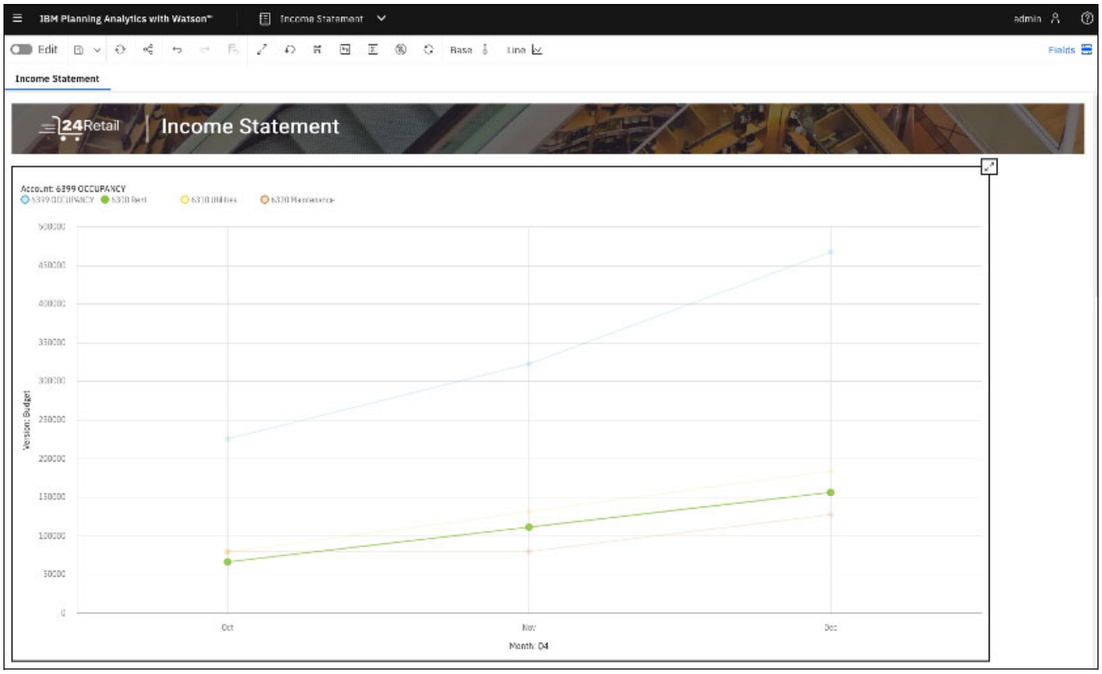

Figure 7-121 Rent by month during Q4

14. Because the visualization might be indicating that the rent is increasing during this period, flag it for sharing with a team member who might be able to take some action on it. To do so, revert to your Exploration view (by clicking the line on top and then, selecting **Exploration**) from the Line visualization and add a comment next to the Q4 Rent by right-clicking on it and selecting **Comments**, as shown in Figure 7-122.

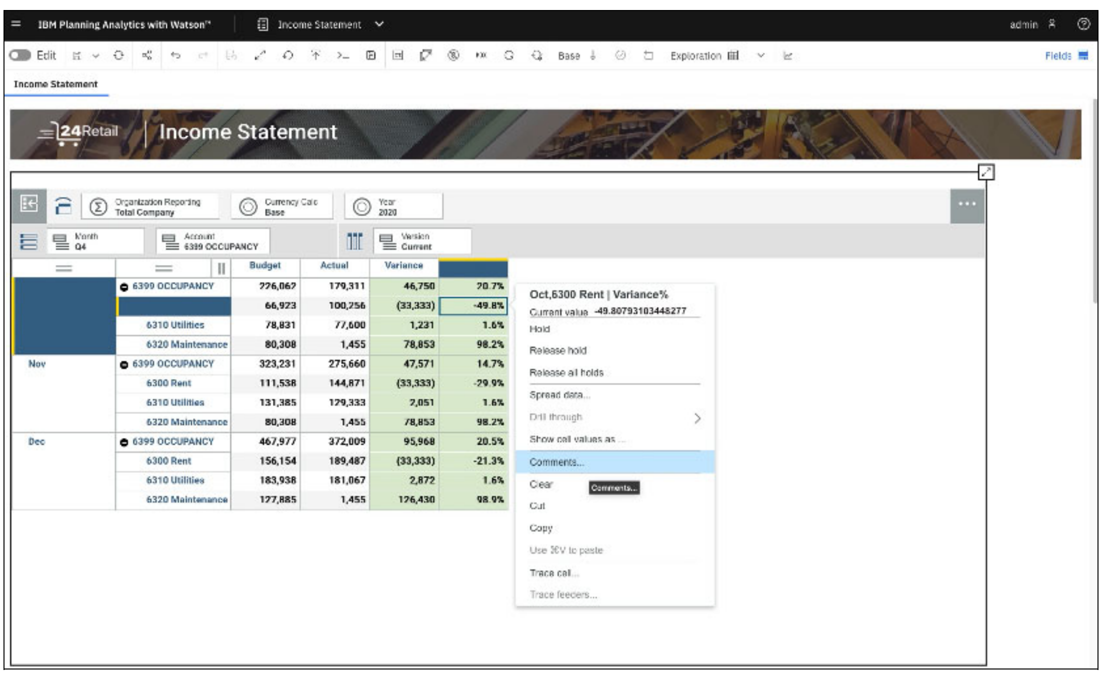

Figure 7-122 Adding a comment to the selected cell on the dashboard

15. After the Comments dialog is open, click **Add comment** and enter a comment, as shown in Figure 7-123. In this example, you notice an increasing trend and leave a note to that effect.

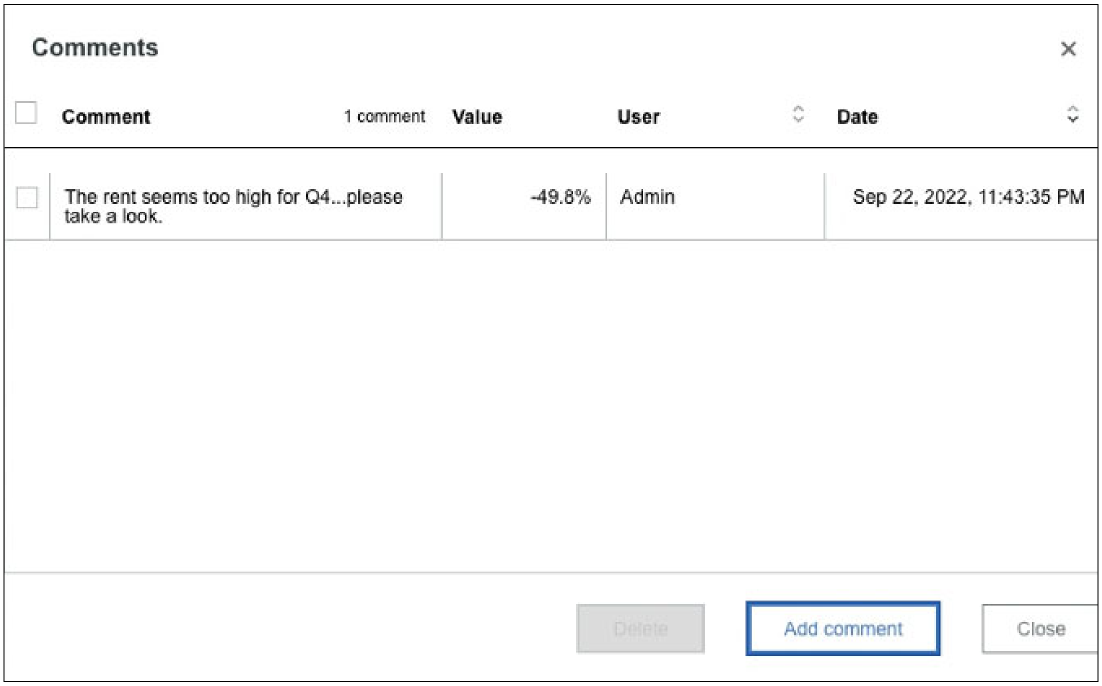

Figure 7-123 Adding a comment to a cell on the dashboard in the Comments dialog

That comment now persists in the dashboard by way of a small blue triangle in the upper right corner of the cell in which the comment was made, as shown in Figure 7-124.

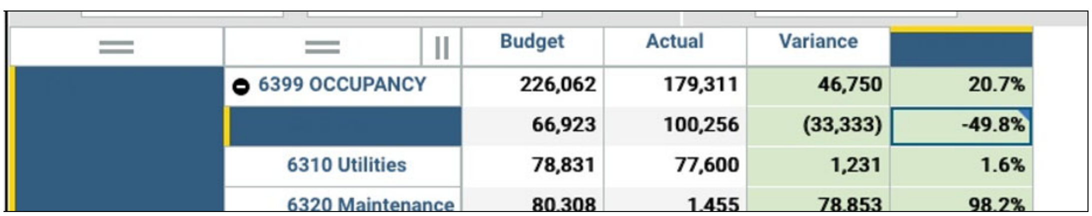

Figure 7-124 Highlighted cell with the small blue triangle in the upper right corner of the cell

16. Because we want to focus on only Rent and remove the Variance values, select the fields of primary interest (Budget and Actual) by right-clicking and selecting **Keep** to keep only those fields, as shown in Figure 7-125.

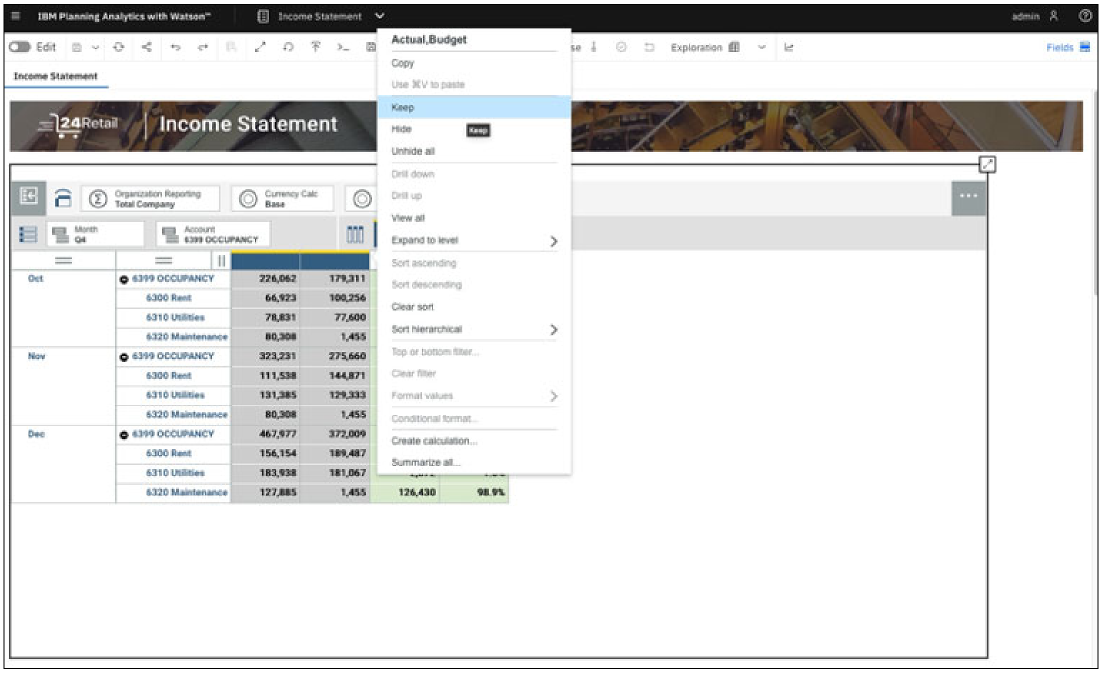

Figure 7-125 Opening the pop-up menu to select only required field of Rent

17. To make it a more focused view (see Figure 7-126), remove the Occupancy, Utilities and Maintenance features by clicking and dragging the **Rent** value to the top line of the workspace, which yields a result with only Rent showing for Q4 months.

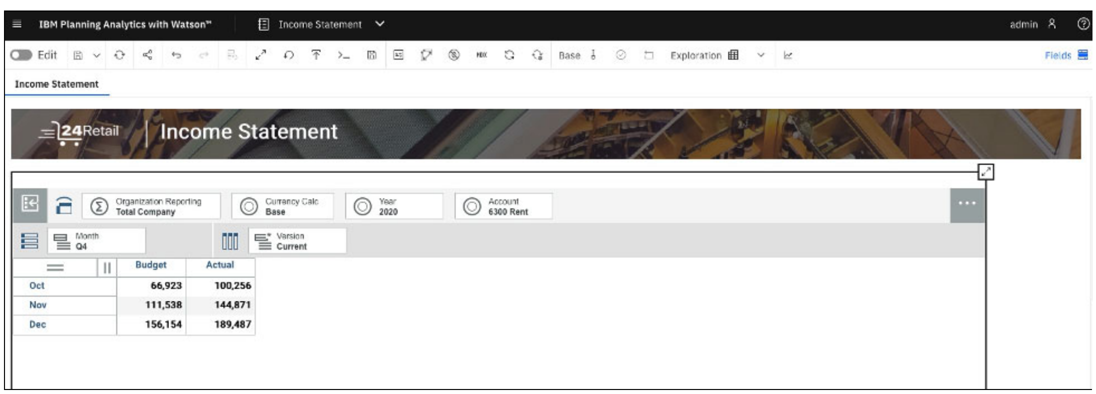

Figure 7-126 Focused view of only Rent Actual and Budget by month in Q4

18. Now that you refined your product to meet the initial requests, you can make it more visually pleasing by creating a column chart of this information and a final graphic that can be shared. To do so, you select **Column** from the Visualizations list on the top, where it says Exploration, to open the Visualizations selector dialog again, as shown in Figure 7-127.

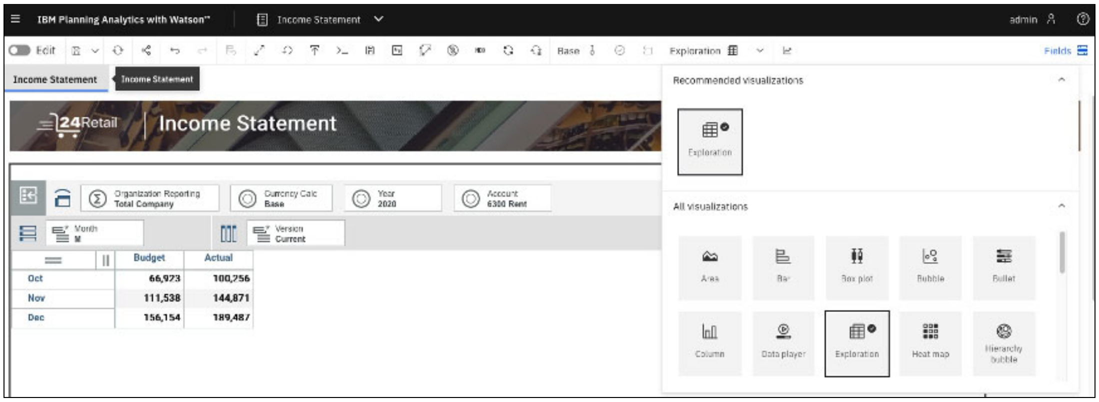

Figure 7-127 Visualization selector to select a Column graph

The Column graph that is shown in Figure 7-128 is displayed.

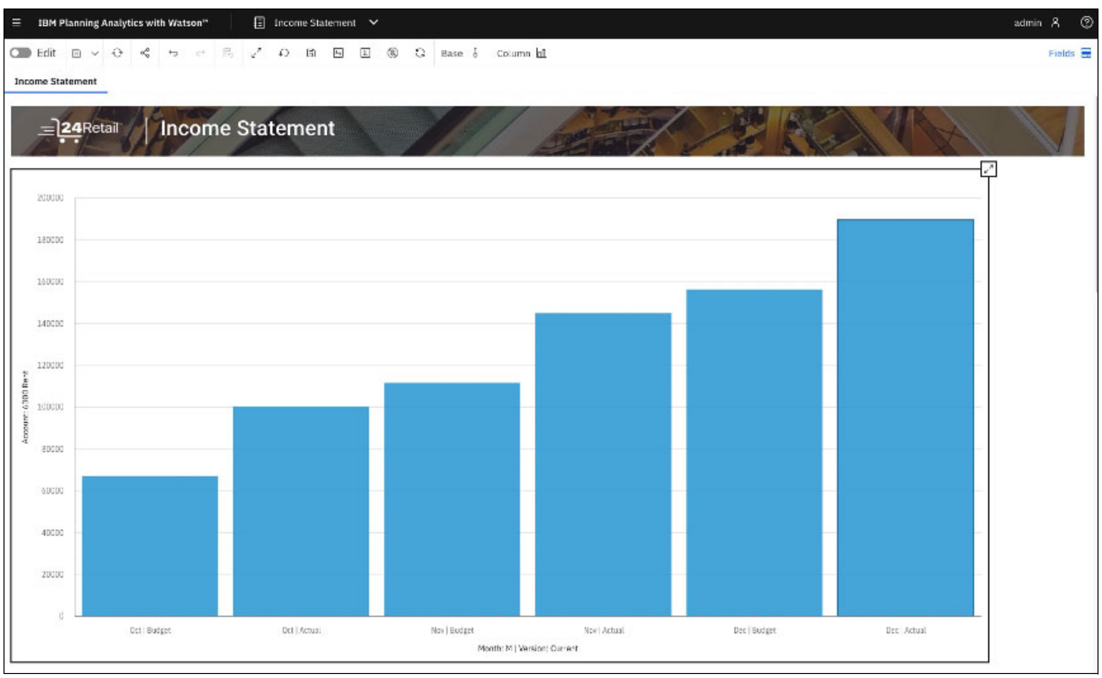

Figure 7-128 Column graph of Actual versus Budget Rent by Month

19.To make the colors of this graph more meaningful, several edits must be made. Activate the Fields dialogue box by clicking the **Fields** button in the upper right corner, coloring by Version to differentiate Budget and Actual (see Figure 7-129).

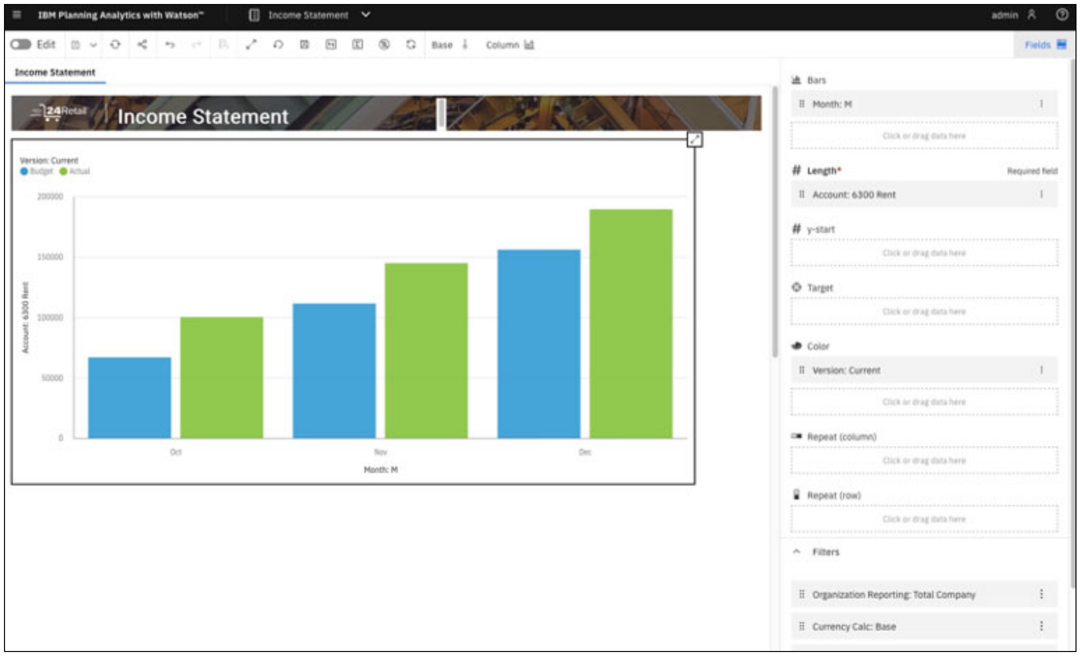

Figure 7-129 Adding colors through the Fields dialog

20.To create final customizations, edit this object by changing to Edit mode by clicking the switch in the upper left corner, which turns green with a check mark after it is turned on. Various functions are unlocked for completing visualizations, including adding a Properties button in the upper right that you can select to make several changes to the graphic, as shown in Figure 7-130.

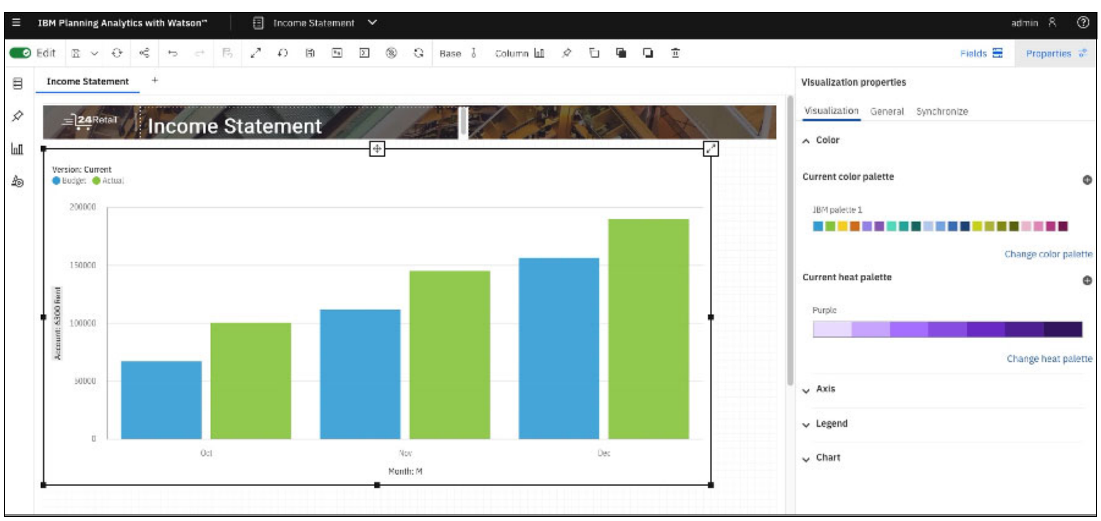

Figure 7-130 Customizing Properties in Edit mode

For example, if the color scheme does not default to the wanted colors, they can be customized by changing the color palette, as shown in Figure 7-130 on page 564. In this example, you keep the default colors. You do want to customize the name because this book is no longer the Income Statement, but now became a Rent Analysis through the dashboard functionality. The final result of making the title changes through the *Visualization Properties* functionality and using **Edit** mode is the following visualization, which is shown in Figure 7-131.

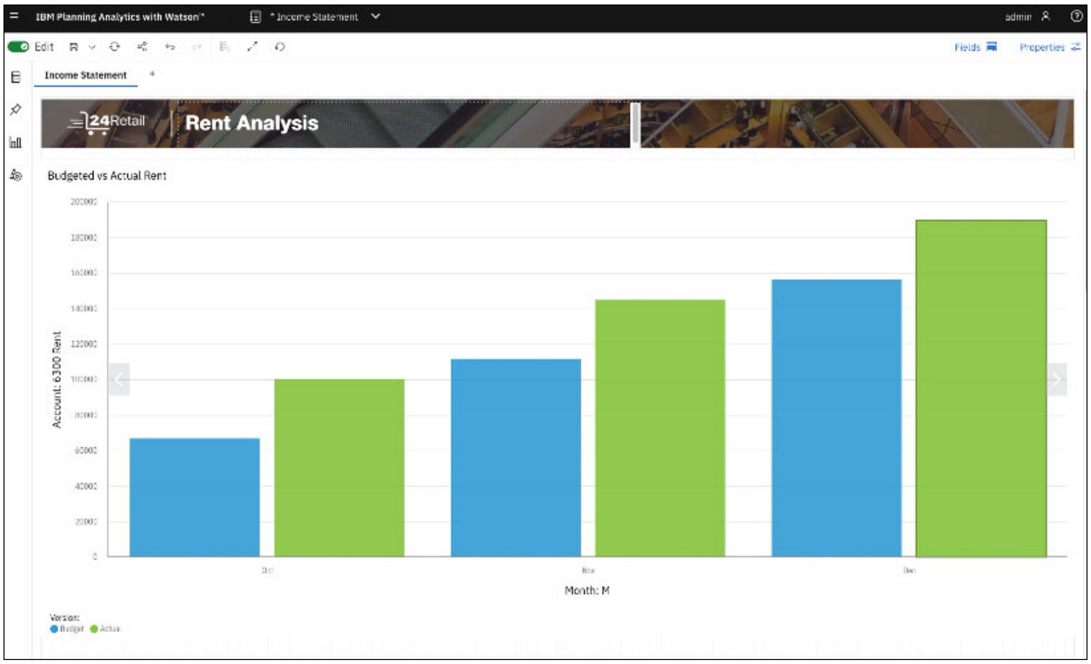

Figure 7-131 Final Column visualization showing Q4 Monthly Rent by Budget and Actual

21.Save this graph by clicking the **small floppy disk icon** in the upper left corner near the Edit option so that it is available in the future, as shown in Figure 7-132.

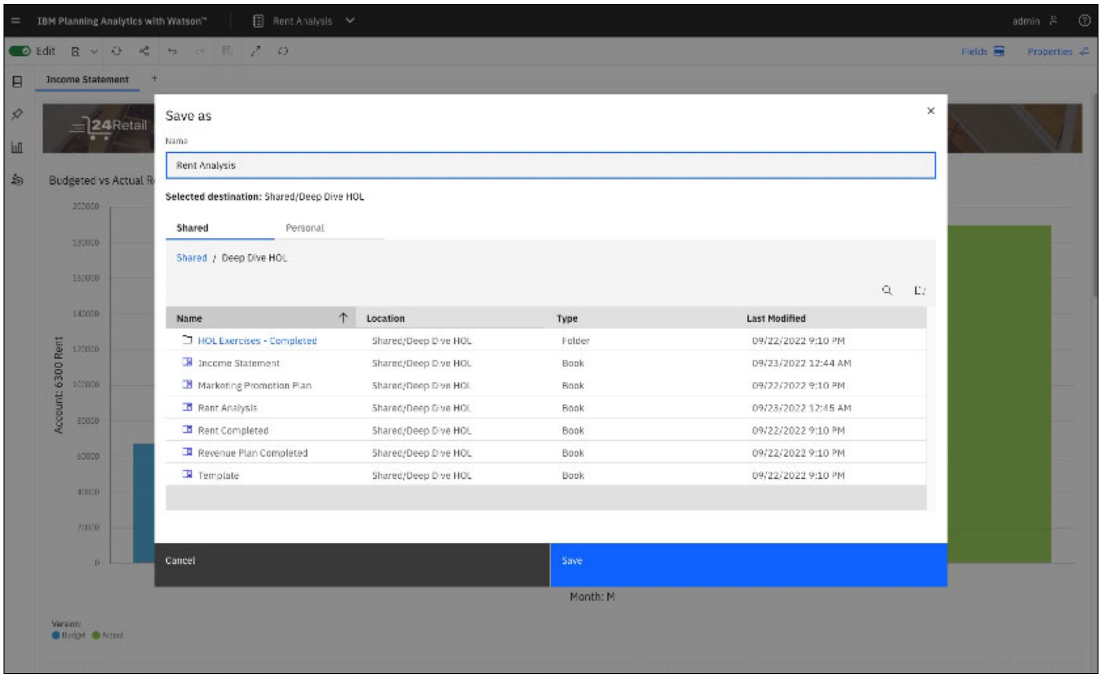

Figure 7-132 Saving the final result in the Shared folder for sharing

22.To retrieve the result, find the Rent Analysis that is available in the Shared folder under the Deep Dive HOL folder, as shown in Figure 7-132. It also is available on the home page under Recents, as shown in Figure 7-133.

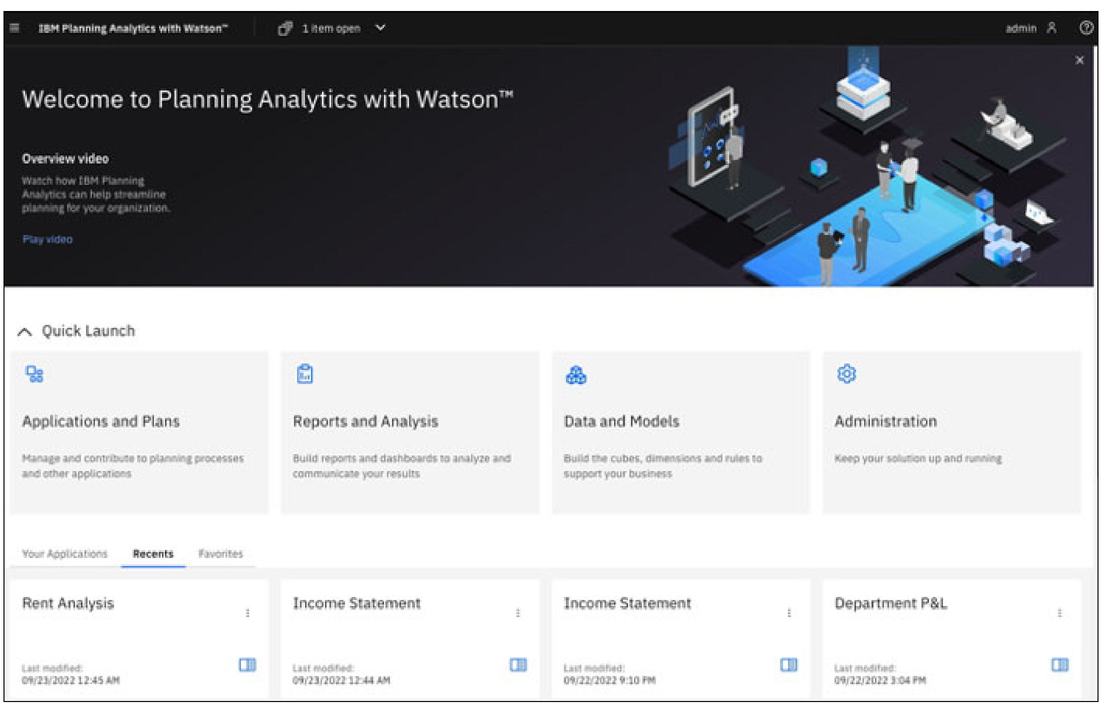

Figure 7-133 In the Home window, Recents feature the just edited visualization

The request is fulfilled for the Rent Analysis, and the manager can be referred to the shared folder for review and interaction with the results.

> **Note:** The data sets that were used in sections 7.3.1, “Use case #1: Visualizing disparate data sources” on page 496 - 7.3.3, “Use case #3: Creating a dashboard in Cognos Analytics” on page 534 were retrieved from the UCI Machine Learning Repository.
>
> *Dua, D. and Graff, C. (2019). UCI Machine Learning Repository* [`http://archive.ics.uci.edu/ml`]. *Irvine, CA: University of California, School of Information and Computer Science*.

8

IBM Cloud Pak for Data Operations

In this chapter, we discuss IBM Cloud Pak for Data Operations and provide several examples. This chapter includes the following topics:

8.1 Introduction and overview

Although the Cloud Pak for Data deployment process is becoming more simple as newer versions are being developed and released, some remaining considerations that, if performed incorrectly, can cause installations to fail in unexpected ways.

In this section, we discuss some of the key considerations and challenges that are encountered on the path to installing Cloud Pak for Data.

IBM Cloud Pak for Data Operations

When operating a Cloud Pak for Data environment, it is critical to reflect on the installation process and post-install operations on Day 1 and Day 2+.

In this chapter, we define the following four key categories to consider. These categories chart the various tasks that must be completed and architectural factors and decisions that are required to ensure an optimally performant Cloud Pak for Data cluster:

- Day 1 Operations

  Day 1 Operations (Day 1 Ops) include all of the tasks that are required to be carried out to install and configure an IBM Cloud Pak for Data cluster. This process includes planning and infrastructure provisioning, completing the installation and prerequisite tasks and post-install tasks. Day 1 Ops also incorporates any other tweaking or debugging to make sure that the cluster is running and operational.

- Day 2+ Operations: Business resilience

  Running any system (hardware, software, and storage) in production is a long-term commitment. It is not as simple as setting it up and letting it run automatically. Business and IT must work together to ensure this system, once stood up, stays up. It is here that Day 2+ Operations comes to the fore.

  Day 2+ Operations encompass a wide range of practices and strategies that are aimed at ensuring the system is consistently delivering on business outcomes by ensuring it meets SLAs and the requirements of the business it serves.

  After Day 1 Ops are complete, a key component of operationalizing any production system is ensuring the cluster remains available and resilient to failure and unexpected events. This idea of ensuring the resilience of a system is part of business continuity security, and we refer to it as *business resilience*.

  Day 2+ Operations (Day 2+ Ops) is a long-term commitment to continuous refinement of your business resilience, security, and observability regime. It means remaining focused on the strategies and measures that lead to impactful systems, such as Cloud Pak for Data remaining available and resilient to change and uncertainty.

- Day 2+ Operations: Observability

  Another facet of Day 2+ Ops is the concept of observability. Observability is more than just monitoring. Monitoring tells when you *when* something is wrong; observability helps you understand *why*.

  Observability is the cornerstone of robust IT Operations and proactive incident management and remediation. It allows the IT Ops or Site Reliability Engineering (SRE) team to gain a window into not only the system processes, but also an understanding of the causal event that provokes an incident or failure. This process, in turn, affords IT teams the opportunity to remediate the issue while also preempting future similar issues. Observability is a key component of intelligent Day 2+ Ops that are required for any production system.

- Security Operations

  Security Operations (SecOps) is the implementation of stringent security requirements in line with the operation of a production system. When we discuss SecOps in this chapter, we are primarily talking about implementing a Zero Trust framework alongside Cloud Pak for Data. Let us look briefly at the concept of Zero Trust with the following simple definition:

  "*Zero trust is a framework that assumes a complex network's security is always at risk to external and internal threats. It helps organize and strategize a thorough approach to counter those threats*." [1]

  Next, we explore how Zero Trust and SecOps ties in with ensuring that your Cloud Pak for Data System is highly secure and protected against threat and vulnerabilities.

8.2 Day 1 Operations

Day 1 Operations include all of the tasks that are required to be carried out to install and configure the IBM Cloud Pak for Data cluster. This process includes planning and infrastructure provisioning, carrying out the installation and prerequisite tasks, and postinstallation tasks. Day 1 Ops also incorporates any other tweaking or debugging to make sure that the cluster is running and operational.

8.2.1 Installation introduction

Critical decisions must be made for any successful Cloud Pak for Data deployment. Many of these factors have real business implications that must be discussed and addressed across diverse teams from IT and business.

Some points to consider include the following examples:

- On what infrastructure will Cloud Pak for Data be hosted and who or what teams own this decision?
- Is a cloud or on-premises installation to be used?
- What type of storage will be used?
- What container storage will be used to support the Cloud Pak for Data implementation and who or what teams own this decision?
- Security and networking considerations. How will networking protocols such as DNS, DHCP, and network firewalls and load balancing be performed?
- Business resilience and high availability disaster recovery (HADR) considerations.
- Monitoring and Day 2 Ops.
- Workload requirements.
- Skills and experience of the installation team.

These key decisions must be made to ensure success in deploying Cloud Pak for Data.

[1] https://www.ibm.com/topics/zero-trust

8.2.2 How the installation works

The Cloud Pak for Data installation process can be broken down into eight steps, as shown in Figure 8-1:

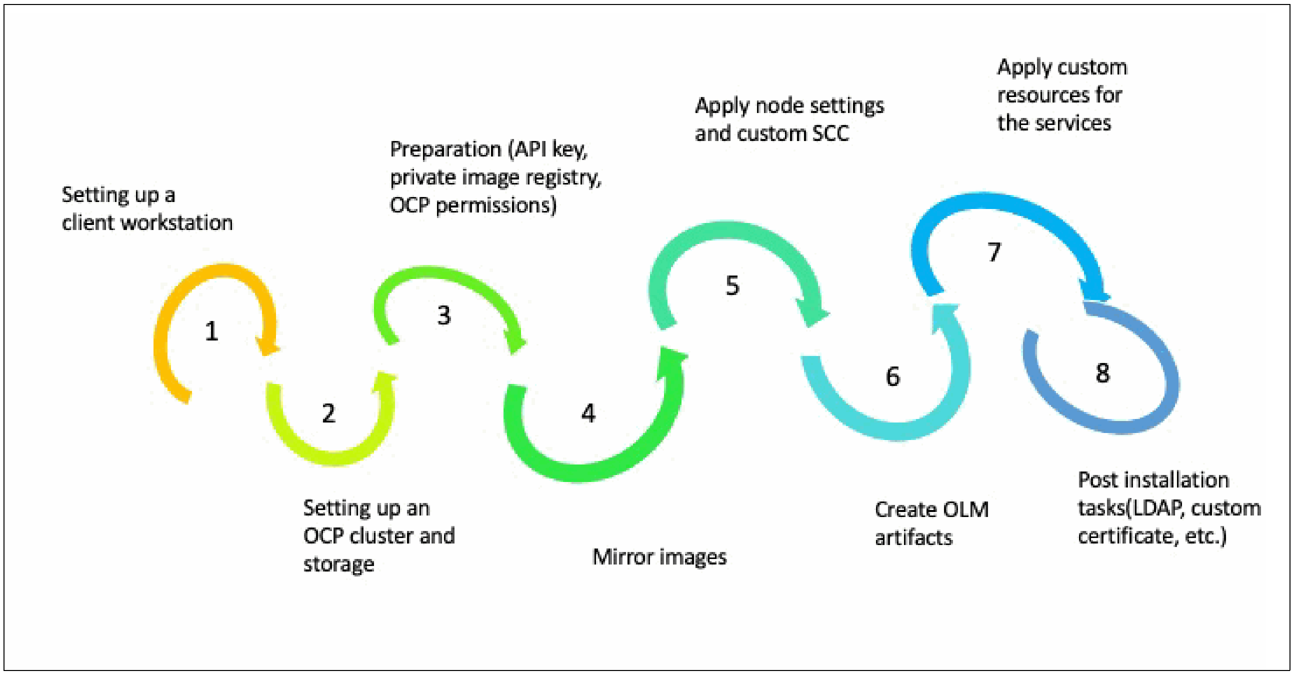

Figure 8-1 Cloud Pak for Data installation flow

1. Setting up a client workstation.
2. Setting up a Red Hat OpenShift Container Platform cluster and storage.
3. Preparation (API key, private image registry, Red Hat OpenShift Container Platform permissions).
4. Mirror images.
5. Apply node settings and custom SCC.
6. Create Operator Lifecycle Manager (OLM) artifacts.
7. Apply custom resources for the services.
8. Postinstallation tasks (LDAP, custom certificate, and so on).

These steps are described next.

Step 1: Setting up a client workstation

To install Cloud Pak for Data, you must have a client workstation that can connect to the Red Hat OpenShift Container Platform cluster. The workstation must meet the requirements that are discussed next.

Internet connection

Some tasks require a connection to the internet. If your cluster is in a restricted network, you can either:

- Move the workstation behind the firewall after you complete the tasks that require an Internet connection.
- Prepare a client workstation that can connect to the internet and a client workstation that can connect to the cluster and transfer any files from the internet-connected workstation to the cluster-connected workstation.

When the workstation is connected to the internet, the workstation must access the following sites:

- GitHub (@IBM)
- IBM Entitled Registry

Client workstation must be running a supported operating system

The supported operating system can be Linux, Mac, or Microsoft Windows. (It is recommended that the Linux operating system is used.) If Mac or Windows is used, consider the special requirements that are discussed at this IBM Documentation web page.

Container run time requirements

The workstation must include a supported container run time (Podman or Docker).

Required command-line interfaces

To install or upgrade the Cloud Pak for Data platform, the following command-line interfaces must be available:

- Red Hat OpenShift command-line interface (`oc`)
- Cloud Pak for Data command-line interface (`cpd-cli`)

Step 2: Setting up a Red Hat OpenShift Container Platform cluster and storage

In this section, we discuss how to set up a Red Hat OpenShift Container Platform cluster and storage.

Setting up Red Hat OpenShift Container Platform cluster

Cloud Pak for Data is deployed on a Red Hat OpenShift Container Platform cluster. If you do not have a cluster, complete the necessary steps to install Red Hat OpenShift on your environment.

The user who is the primary cluster administrator must install the Red Hat OpenShift Container Platform cluster. You can deploy Cloud Pak for Data on-premises or on the cloud.

The Red Hat OpenShift cluster can be managed Red Hat OpenShift or self-managed Red Hat OpenShift. Your deployment environment determines how you can install Red Hat OpenShift Container Platform.

Setting up Red Hat OpenShift Container Platform storage

Cloud Pak for Data supports dynamic storage provisioning. A Red Hat OpenShift cluster administrator must configure storage before Cloud Pak for Data is installed. Cloud Pak for Data supports and is optimized for several types of persistent storage, including:

- Red Hat OpenShift Data Foundation
- IBM Spectrum Fusion
- IBM Spectrum Scale Container Native
- Portworx
- Enterprise NFS
- Amazon Elastic Block Store (EBS)
- Amazon Elastic File System (EFS)
- IBM Cloud Block Storage
- IBM Cloud File Storage

For more information, see the "What storage options are supported for the platform?" section of this IBM Documentation web page.

Step 3: Preparing API key, private image registry, Red Hat OpenShift Container Platform permissions, and so on

All IBM Cloud Pak for Data images are accessible from the IBM Entitled Registry.

The IBM entitlement API key enables you to pull software images from the IBM Entitled Registry for installation or mirroring to a private container registry.

> **Note:** In most situations, it is strongly recommended that you mirror the necessary software images from the IBM Entitled Registry to a private container registry.

Cloud Pak for Data relies on a separation of roles and duties. Two administrative roles are identified and associated with a different level of permissions: Red Hat OpenShift cluster administrator and Project administrator. Corresponding installation tasks are associated with each administrative role.

Step 4: Mirroring images

If your client workstation can connect to the internet and the private container registry, you can mirror the images directly to your private container registry.

If your client workstation cannot connect to the internet and to the private container registry, you must mirror images to an intermediary container registry before you can mirror the images to your private container registry.

Step 5: Applying node settings and custom SCC

Some services that run on IBM Cloud Pak for Data require specific settings on the nodes in the cluster. To ensure that the cluster includes the required settings for these services, you must review and adjust the following settings on the suitable nodes in the cluster:

- Changing load balancer timeout: To prevent connections from being closed before processes complete, you might need to adjust the timeout settings on your load balancer node.
- Changing CRI-O container: To ensure that some services can run correctly, you must change settings that are required for the CRI-O container run time on the Red Hat OpenShift Container Platform.

- Changing kernel parameter: Db2 Universal Container (Db2U) is a dependency for some services. By default, Db2U runs with elevated privileges in most environments.

  Note: Db2U is open source and available at this Docker Hub web page.

- Changing Power settings On Power Systems: You must change the simultaneous multithreading (SMT) settings for Kernel-based Virtual Machine (KVM)-capable systems and large core, IBM PowerVM® capable systems.
- Red Hat OpenShift NVIDIA GPU Operator: The NVIDIA GPU Operator uses the operator framework within Kubernetes to automate the management of all NVIDIA software components that are required to provision GPU.

Step 6: Creating Operator Lifecycle Manager artifacts

A cluster administrator, or a user with the suitable permissions to install operators, must create the Operator Lifecycle Manager (OLM) objects, such as catalog sources and operator subscriptions that are required to install the IBM Cloud Pak for Data platform and services. The OLM objects that you create depend on the services that you plan to install.

Step 7: Applying custom resources for the services

A Red Hat OpenShift project administrator can create custom resources to install the IBM Cloud Pak for Data platform and services in a project (namespace). The custom resources that you create depend on the services that you plan to install.

Step 8: Postinstallation tasks (LDAP, custom certificate, and so on)

After you install IBM Cloud Pak for Data, complete the suitable tasks to secure your environment, set up the platform for users, and ensure that the platform and services can run smoothly.

The postinstallation tasks include, but are not limited to, the following examples:

- Changing the password for the default admin user.
- Customizing and securing the route to the platform.
- Configuring single sign-on.
- Managing certificates.
- Monitoring and reporting use against license terms.
- Setting up services after installation or upgrade.

Red Hat Operator Lifecycle Management process

The Red Hat OLM process is a fundamental paradigm that underpins the installation and implementation of IBM Cloud Pak for Data. Operators facilitate the delivery of software in the form of containers by bundling them in a way that seamlessly allows them to be deployed with the requisite services, deployments, deployment configurations, secrets, and other Red Hat OpenShift API resources.

The following terms are related to the Red Hat Operator Lifecycle Management process:

- Concepts of operators

  An *operator* is a set of Kubernetes-native resources that packages, deploys, and manages a Kubernetes application by extending the Kubernetes API. For more information about operators, see this Red Hat OpenShift Documentation web page.

- Catalog source

  A *catalog source* is a repository of cluster service versions (CSVs) that includes the name, version, icon and required resources, custom resource definitions (CRDs), and packages that comprise an application. For example, the `cpd-platform` catalog source defines the CSVs of the Cloud Pak for Data platform operator.

- Operator subscription

  An *operator subscription* tells the cluster where to install a specific operator and provides more information about the operator to the OLM. When you create an operator subscription, OLM receives the CSV for the operator. The CSV describes the operator, and OLM uses the CSV to introduce the CRD if it does not exist.

  Then, it sets up the operator's service accounts and starts the operator deployment. The `cpd-platform` operator subscription defines the `Ibmcpd` CRD, which defines the Cloud Pak for Data control plane.

- Custom resource

  When you create the custom resource, the corresponding operator processes the contents of the custom resource and starts the microservices that comprise the application or service. The `ibmcpd-cr` customer resource is an instance of the `Ibmcpd CRD` and is created in the Cloud Pak for Data Red Hat OpenShift project.

8.2.3 Planning an installation

A Cloud Pak for Data installation is designed with growth and evolution in mind. With a tight integration of services, a customer can start with a single service and expand to multiple complementary services. A Cloud Pak for Data System is likely to change and evolve over time.

The following two types of Cloud Pak for Data installations are available:

- Initial installation: The initial installation is a major milestone with significant planning and often, with future evolution in mind. The infrastructure is provisioned including storage, hardware, and software. The system tested, accepted, and delivered to the users. After this process is completed, the evolution of an installation varies from customer to customer.
- Upgrade and maintenance: The IBM Cloud Pak for Data Release Cycle features the following pattern:
  - IBM issues a major release (for example, 3.0 to 4.0) approximately every 2 years. These releases typically deliver significant functions and changes and often requires migration of services during the upgrade process.
  - IBM issues a minor release (for example, 4.0 to 4.5) approximately every 12 months. These releases often deliver new functions, but use the same code base as the major release.
  - IBM releases a patch point release (for example, 4.0.5 to 4.0.6) every month. When a new minor version is released (for example, 4.0 to 4.5), the monthly patches move to the new release.

  Based on these cycles and other factors, a customer can decide their own upgrade path. The following important factors must be considered:

  - Internal software release policies: Most companies have a policy that relates to the upgrade of software. This process can vary between ensuring all software is at the latest release to upgrading only when required and then, to only a proven mature release.

- Upgrade cycles: When planning upgrade cycles, a tradeoff exists between frequency and effort. Generally, moving between single patch releases requires minimal disruption and less change. It is also generally easier to migrate from the current release to the next release, even if that release is major.

  This cycle requires a monthly process and resource overhead to apply these patches. If a less frequent cycle is used (for example, only major releases), the change that is required is much larger with greater disruption but less frequent overhead. Choosing the best fit often depends on organizational structures.
- Functions: A primary motivation for moving to a new release is new or enhanced functions. Specific business needs might exist that are addressed by a specific release.
- Bug fixes and security patches: A new release can contain bug fixes that address specific issues that customers are experiencing. They might also address newly identified security threats.
- End of support: The version of Cloud Pak for Data, its supporting software (for example, Red Hat OpenShift), or some connecting system can reach the end of its support agreement and require an upgrade to maintain support. This process often can require integrated systems to be upgraded.
- Unprecedented or unforeseen events: Unprecedented or unforeseen events can occur that require an immediate fix pack to resolve.

All of these factors are considered and a suitable fit is defined.

> **Note:** A typical example is an upgrade policy of every 6 months to the latest available release. An exception policy also might exist for out-of-cycle updates if they are critical.

Other installations

With the broad set of services that is available in Cloud Pak for Data, it is likely that a customer initially starts with a specific set of services with a plan to expand that set after the initial business objective is achieved.

In some cases, the initial installation sizing is planned with this expansion in mind and the installation of subsequent services can be straight forward. In others, the extra installations must be planned to the same level as the initial installation.

8.2.4 Mirroring images and air-gapped environments

In this section, we discuss mirroring images and air gapped environments.

Mirroring images

Managing Cloud Pak for Data images is one of the key aspects of an installation. They are accessible from the IBM Entitled Registry. However, it is strongly recommended that you mirror the necessary images from the IBM Entitled Registry to a private container registry with the following considerations:

- Security: For enterprises with strict security requirements, they can run security scans against the software images before installing them on the Red Hat OpenShift cluster.
- Manageable: Multiple Cloud Pak for Data deployments with the same images, such as development or test environments and production environments.
- Stable: Predictable and reliable performance during the image mirror process.

- Validation: As Cloud Pak for Data images are key to the installation success, it is recommended to validate whether the images were mirrored as expected to the target registry.

  The Cloud Pak for Data utility `cpd-cli` can help identify the image manifests of a release and verify those images. This process helps avoid problems during installations or upgrades (especially with missing image manifests).
- Cleanup: After the Cloud Pak for Data cluster that is upgraded to a new release or some services is uninstalled, some images in the target registry become useless. To save the storage, you can remove the old images from the target registry that correspond to the selected release and set of components that are selected.
- Limitations: Currently, Cloud Pak for Data images are built with the support for multiple architectures, including X86_64, Power, and LinuxOne. This build causes the transfer of images to a private registry to take significantly longer. An enhancement request to repackage images for a single processor architecture was filed with IBM.

Air-gapped environments

At service installation time, the Docker images of the various services are downloaded from the remote image registry and automatically deployed to the selected worker node by Red Hat OpenShift. The image also is cached on the local disk of the target worker node.

Later on, when a pod is redeployed because of horizontal scaling or restarted after a crash, the new pod can be deployed to the same or a different worker node. Red Hat OpenShift first checks whether the related image is stored on the local storage of the selected node, and if found, that image can be immediately deployed. However, if the image is not found on the local cache, it is downloaded again from the remote image registry.

For security reasons, many environments, especially production environments, often restrict or control outbound traffic to the internet, potentially affecting the described process.

Connectivity setups include the following possibilities:

- Air-gapped: No node in the Red Hat OpenShift cluster can access the internet. In this setup, a local registry is mandatory because Red Hat OpenShift cannot download any image from remote registries.
- Proxy: All nodes in the Red Hat OpenShift cluster can access to the internet through a proxy, which must be configured to allow access to all the remote registries that are holding required images.
- Connected: All nodes in the Red Hat OpenShift cluster have full access to any site in the internet. Red Hat OpenShift can download images at any time from the remote registries.

Use of local registries

The use of local registries is highly recommended, while accessing remote registries generally is to be avoided. The latency to access the internet to download images might affect the user experience, which results in longer wait times compared to fetching the images from a local registry.

Especially on production environments, but also on test or development environments with slow internet access, the use of local registries ensures better performance during pod deployment.

Avoid installing the local registry on the bastion node

The use of a local registry is recommended because it speeds up pod deployments, and improves the user experience. However, avoid installing it on the bastion node because it becomes a single-point-of-failure.

In case of issues with the local registry, the system hangs until an operator fixed the issue, which results in much longer downtimes than the latency of a remote registry. Therefore, use only an enterprise grade local registry that is deployed to different servers and ensures high availability.

Use a bastion node that is connected to the internet

When local registries are used, the images must be downloaded from the remote registries on the internet during the installation and mirrored to the local registry. The recommended way is to have the bastion node connected to the internet directly or through a proxy. The bastion node can access both internet and the local registry. This configuration simplifies the initial mirroring of the images and image mirroring during upgrades.

Fully air-gapped environments

In fully air-gapped environments, no machine can access the internet, not even the bastion node. If no machine can access the internet and the local registry, the images must be downloaded to a portable registry that can access the internet.

The portable registry then must be transferred somehow (for example, by using FTP or even a USB stick) to a machine with access to the local registry and from there, pushed to the local registry. The Cloud Pak for Data documentation offers a well-described process of the use of `cpd-cli`, which helps you set up this portable registry and mirror the images to it and eventually to the local registry.

8.2.5 Continuous adoption

Over the past few years, a tremendous amount of thought leadership was conducted by the Academy of Technology that centered on the concepts of continuous adoption (CA) and delivery as it relates to IBM Cloud Paks.

The aim of this section is to build on what is a remarkable body of work on the topic, cites some seminal articles that are created on CA, and leaves the reader with some suggestions about how this process might work in practice. We also discuss into the practice of GitOps and Infrastructure-as-Code (IaC), including Ansible.

The what and why of continuous adoption

We define *continuous adoption* as the practice of keeping up to date with software that is released by a continuous delivery model.

The next questions that naturally arise: Why does continuous adoption matter? Why should we consider it at all?

Citing from Jeroen van der Schot's article[2]:

"The main reasons for considering Continuous Adoption are:

- *Delivering the latest innovations to developers for optimal productivity.*
- *Maintaining security compliance.*
- *Ensuring support availability from providers.*
- *Avoiding shadow IT.*
- *Reducing the risk of upgrades.*
- *Avoiding technical debt."*

[2] *Continuous Adoption: Keeping current with accelerating software innovations* by Jeroen van der Schot: https://jvdschot.medium.com/continuous-adoption-keeping-current-with-accelerating-software-innovations-33233461181a

Continuous Adoption Lifecycle

The Continuous Adoption Lifecycle consists of the following key steps:

1. Sourcing
2. Verification
3. Installation
4. Testing

These steps are described next.

Step 1: Sourcing

Citing from Frank Ketelaar's article:[3]

"Red Hat delivers patch releases (aka z-releases) for its Red Hat OpenShift Container Platform on an almost weekly basis, giving organizations peace of mind that the platform provides timely fixes for security and product issues. This is an example of sourcing software that delivered continuously."

Step 2: Verification

In larger enterprises that are running many Red Hat OpenShift clusters and Cloud Paks with mission-critical applications, direct and automatic upgrades from the internet are typically not feasible. Oftentimes, clusters cannot connect to the internet and every enhancement and patch first must be imported into a "demilitarized zone", where it is scanned for vulnerabilities and then, tested in different environments and under various conditions before being applied in production.

Step 3: Installation

In "GitOps principles" on page 581, we discuss GitOps and how that lends itself to implementing continuous adoption of the installation process by using Infrastructure-as-Code (IaC) and automation tools, such as Ansible.

Step 4: Testing

All pipelines that are implementing a continuous adoption framework must have testing at their heart. This process often consists of simple build and deploy tests that form part of a Continuous Integration/Continuous Deployment (CI/CD) pipeline.

> **Note:** *Continuous Deployment* is not to be confused with *Continuous Delivery*, which is a software release model that is used alongside CA.

Applying continuous adoption to Cloud Pak for Data

Applying continuous adoption to Cloud Pak for Data involves GitOps. *GitOps* is the harnessing of code that is stored in Git to install, manage, and maintain enterprise software deployments.

A brief history of GitOps

The GitOps concept originated from Weaveworks in 2017. The goal was to automate the operations of a Kubernetes (K8s) system by using a model external to the system as the source of truth.

[3] *Continuous Adoption logistics* by Frank Ketelaars published at https://frank-ketelaars.medium.com/continuous-adoption-logistics-ec5b1e60054e

GitOps principles

With the ever growing adoption of GitOps, the OpenGitOps project was started in 2021 to define a set of open source standards and best practices. These best practices help organizations adopt a standard and structured approach when implementing GitOps. GitOps is a Cloud Native Computing Foundation (CNCF) Sandbox project.

The GitOps Working Group released v0.1.0 of the following GitOps principles:

- The principle of declarative desired state: A system that is managed by GitOps must have its desired state expressed declaratively as data in a format writable and readable by humans and machines.
- The principle of immutable desired state versions: The desired state is stored in a way that supports versioning, immutability of versions, and retains a complete version history.
- The principle of continuous state reconciliation: Software agents continuously, and automatically, compare a system's actual state to its desired state. If the actual and desired states differ for any reason, automated actions are started to reconcile them.
- The principle of operations through declaration: The only mechanism through which the system is intentionally operated through these principles.

> **Note:** GitOps can be used to manage the infrastructure, services, and application layers of K8s-based systems. For more information about the use of the GitOps workflows to deploy IBM Cloud Paks on the Red Hat OpenShift platform, see this web page.

GitOps workflows

Various GitOps workflows are available. One of the views of how GitOps can be used is documented in the Cloud Pak production guides that were developed by IBM Customer Success Management.

It is an opinionated view about how GitOps can be used to manage the infrastructure, services, and application layers of K8s-based systems. It considers the various personas that interact with the system and accounts for separation of duties.

For more information about trying this GitOps workflow, see this web page.

The discussion of GitOps brings us to the topic of Infrastructure-as-Code (IaC) and the use of tools, such as Ansible and Terraform, to provision Cloud Pak for Data in an automated and robust fashion.

The *Cloud Pak Deployer Guide* is a another good example of how to bring GitOps and CA practices to life by using Cloud Pak for Data.

8.2.6 Troubleshooting 101

Some general tips are available for Cloud Pak for Data troubleshooting. These installations are classified into two basic phases, which are described next.

Phase I: Installing the OLM artifacts (including the operator)

The Cloud Pak for Data operators simplify the process of managing the services on Red Hat OpenShift Container Platform.

To install the service, you must install the IBM Watson Studio operator and create the OLM objects, such as the catalog source and subscription, for the operator.

During the operator installation, you might encounter the following problems:

- InstallPlans are not showing up.
- CSVs do not come up.
- Subscriptions have no meaningful condition messages.

Consider the following tips for resolving these issues:

- Check whether catalog sources are READY: If the Cloud Pak Foundational or Cloud Pak for Data catalog source is not READY, a problem exists. Also, check whether any failing pods exist in the `openshift-marketplace` namespace, especially the ones with image pull back-off status.
- Review OLM logs: Check the logs. This operator is in the Red Hat OpenShift `openshift-operator-lifecycle-manager` project.
- Check the OLM operator pod (`openshift-operator-lifecycle-manager`) logs: Check whether subscription and CSV issues exist. A usual problem that OLM gets stuck when one of the subscriptions is invalid. It seems to stop at the first error and does not proceed. So, you must locate the first failed subscription and then, clean up from that point.

Phase II: Installing the service

After the operator is installed, you can install the service. A specific custom resource must be created to install the service. If you encounter problems during the service installation, consider the following tips:

- If any operand fails to make an appearance, inspect the logs of the Operand Deployment Lifecycle Manager. You also might need to validate the existence of the operand request, operand registry, and operand configurations.
- Inspect the logs of the corresponding service operator.

Useful troubleshooting commands

The following troubleshooting commands can be useful:

> **Note:** Commands that use custom-columns are subject to change.

- Find any pods with issues:

  ```
  oc get po  -A| grep -Ev '([[:digit:]])/\1.*R' | grep -v 'Completed'
  ```

  This command is useful because it does not solely focus on the Cloud Pak for Data specific issues. It also includes storage (for example, ODF, Portworx) or other Red Hat OpenShift issues. On a healthy Red Hat OpenShift Cluster, this command does not return any pods.

- Quick pod deletion:

  ```
  oc get pods|   grep ImagePull | awk '{print $1}' | xargs oc delete pods
  ```

  Replace *`ImagePull`* with whatever you need. This command is useful when you resolved an image pull issue and want the pods to retry.

- List all catalog sources and their status:

  ```
  oc get catalogsource -o
  custom-columns=NAMESPACE:.metadata.namespace,NAME:.metadata.name,DESCRIPTION:sp
  ec.displayName,LATEST_STATES:status.connectionState.lastObservedState -A
  ```

- List all subscriptions, their state, and current version:

```
oc get sub -o
custom-columns=:metadata.name,:metadata.namespace,:status.state,:status.current
CSV -A
```

- List all ClusterService versions:

```
oc get csv -o
custom-columns=:metadata.name,:metadata.namespace,:status.lastUpdateTime,:.spec
.version,:spec.replaces,:.status.phase -A
```

- List all subscriptions and their ClusterService versions:

  If CSVs for a subscription are not created, this issue likely indicates that the OLM locked up. You can delete the subscriptions to avoid the blockage:

```
for i in $(oc get sub -n ibm-common-services
--sort-by=.metadata.creationTimestamp -o name);   do   oc get $i -n
ibm-common-services -o
jsonpath='{.metadata.name}{","}{.metadata.creationTimestamp}{","}{.metadata.lab
els}{","}{.status.installedCSV}{"\n"}'; done
```

All IBM custom resource definitions

Several variants of this command are available because a consistent convention does not exist for the information that is returned for CRDs. Also, some of the CRD have cpd in their names, but not all:

- Return all instances of CRDs where the CRD name contains `ibm` and `cpd`. Return the default information for each:

```
for i in $(oc get crd
-o=custom-columns=TYPE:.kind,NAME:.metadata.name,CR_NAME:.spec.names.singular |
grep ibm | grep cpd | awk '{print $2}') ; do oc get $i -A ; done
```

- Return all instances of CRDs where the CRD name contains `ibm`. Return the default information for each and print CRD name:

```
for i in $(oc get crd
-o=custom-columns=NAME:.metadata.name,CR_NAME:.spec.names.singular | grep ibm |
awk '{print $2}') ; do echo $i ; oc get $i -A ; done
```

- Return all instances of CRDs where the CRD name contains `ibm`. Return the custom information for each and print CRD name:

```
for i in $(oc get crd
-o=custom-columns=NAME:.metadata.name,CR_NAME:.spec.names.singular | grep ibm |
awk '{print $2}') ; do echo $i ; oc get $i  -o
custom-columns=CR:kind,CRD:metadata.name,VERSION:spec.version,FAILURES:status.c
onditions[*].ansibleResult.failures -A ; done
```

- Output and have the yaml for all custom resources:

```
for i in $(oc get crd
-o=custom-columns=NAME:.metadata.name,CR_NAME:.spec.names.singular | grep ibm |
grep cpd | awk '{print $2}') ; do echo $i ; oc get $i -A -o yaml > crd-$i.yaml
; done
```

- List all operand requests

  An *operand* is any resource that is acted upon by the operator. That is, if the operator acts on a resource, it is referred to as the operand. An *operand request* is a request for operands to be installed; for example, a request for operator objects to be installed somewhere within the Red Hat OpenShift cluster:

  ```
  oc get operandRequests -A
  -o=custom-columns=NAME:.metadata.name,NAMESPACE:.metadata.namespace,REGISTRY_NA
  ME:.spec.requests[*].registry,REGISTRY_NAMESPACE:.spec.requests[*].registryName
  space,STATUS:.status.phase,MEMBERS:.status.members[*].name
  ```

- List all operand registries

  An operand registry defines the OLM information, such as channel and catalog source, for each operator:

  ```
  oc get operandRegistry
  -o=custom-columns=NAME:.metadata.name,OPERATOR_NAME:.spec.operators[*].name,OPE
  RATOR_NAME:.spec.operators[*].channel,OPERATOR_NAMESPACE:.spec.operators[*].nam
  espace
  oc get operandRegistry
  -o=custom-columns=NAME:.metadata.name,STATUS_CHECK:.status.operatorsStatus
  ```

- Cloud Pak for Data installation logs

  Output the last 20 lines of every pod's log in current namespace to file (useful for operator logs to see the summary for each):

  ```
  for i in $(oc get pods | grep -v NAME  | awk '{print $1}') ; do oc logs $i |
  tail -20 > $i.log ; done
  ```

- Main Cloud Pak for Data operator logs:

  ```
  oc logs $(oc get pods --all-namespaces | grep -i cpd-platform-operator-manager
  | awk '{print $2}')  -n  $(oc get pods --all-namespaces | grep -i
  cpd-platform-operator-manager  | awk '{print $1}') -f
  ```

- Control plane:

  ```
  oc logs $(oc get pods --all-namespaces | grep -i ibm-zen-operator  | awk
  '{print $2}')  -n  $(oc get pods --all-namespaces | grep -i ibm-zen-operator  |
  awk '{print $1}') -f
  ```

- Watson Knowledge Catalog (WKC):

  ```
  oc logs $(oc get pods --all-namespaces | grep -i ibm-cpd-wkc-operator | grep -v
  catalog  | awk '{print $2}')  -n  $(oc get pods --all-namespaces | grep -i
  ibm-cpd-wkc-operator  | grep -v catalog  |  awk '{print $1}') -f
  ```

8.3 Day 2+ Operations: Business resilience

Running any system (hardware, software and storage) in production is a long-term commitment. It is not as simple as setting it up and letting it run automatically.

Business and IT must work together to ensure that this system, after is stood up, stays up. It is at this stage that Day 2+ Operations comes to the fore. Day 2+ Operations encompasses a wide range of practices and strategies that are aimed at ensuring that the system is consistently delivering on business outcomes by ensuring it meets SLAs and the requirements of the business it serves.

After Day 1 Operations are complete, a key component of operationalizing any production system is ensuring that the cluster remains available and resilient to failure and unexpected events. This idea of ensuring the resilience of a system is part of business continuity and is referred to as *business resilience*.

Day 2+ Operations (Day 2+ Ops) is a long-term commitment to continuous refinement of your business resilience, security, and observability regime. It means remaining focused on the strategies and measures that lead to impactful systems, such as Cloud Pak for Data, remain available and resilient to change and uncertainty.

8.3.1 Overview

We define *business resilience* as the ability of a business to withstand unexpected events, failures, or setbacks and to recover quickly from the event with minimal downtime and data loss. Business resilience in IT is the intersection of three core concepts: high availability (HA), disaster recovery (DR), and backup and restore (B/R) as shown in Figure 8-2. A business' resilience can be measured by its ability to competently implement HA, DR, and B/R.

Figure 8-2 Business resilience

Downtime and data loss can be caused by several factors, including the following examples:

- Failures of server, storage, site, and networks
- Cyberattacks or cyberevents
- Human or system errors
- Viruses and malware
- Hard disk drive damage
- Power outages
- Computer or laptop theft
- Liquid damage
- Natural disasters
- Software corruption

In this section, we focus on platform-level HA/DR that affects the breadth of applications and services that are running in Red Hat OpenShift Container Platform. In many cases, application-level HA/DR can be accomplished with the HA pattern that is based on an asynchronously replicated application or database cluster that is deployed across zones in a single site or region.

To provide DR for site failure, a Red Hat OpenShift Container Platform deployment can be stretched between two geographically different locations. To remain resilient to disaster, critical Red Hat OpenShift Container Platform services and storage must continue to run if one or more locations become partially or unavailable.

Next, we introduce several different approaches for HA, DR, and B/R.

8.3.2 Language of resilience

In this section, we describe the most common concepts and terminology that are used when discussing highly available systems.

Availability

The *availability* of a system is the likelihood that it is available to its users to do the work for which the system was designed. For example, a database might be 95% available or 99% available for database queries or updates. The availability of a system is calculated from its up-time that is divided by the sum of its uptime and downtime.

Ideally, we want to design systems that have 100% availability; we see that such systems incorporate resiliency and redundancy to achieve this goal.

Failover

Failover is the ability to switch automatically and seamlessly to a reliable backup system.

Five Nines

The term *Five Nines* is used to describe systems that are highly available to their users. It can be written as *Five Nines*, *5x9s*, or *99.999*, and describes the percentage of time that a system is available to its users:

Consider a year compromising 365 days, each of 24 hours:

- 5x9s or 99.999% translates to approximately 5 minutes of downtime.
- 4x9s or 99.99% translates to approximately 53 minutes of downtime.
- 3x9s or 99.9% translates to approximately 8.7 hours of downtime.
- 2x9s or 99% translates to approximately 3.5 days of downtime.

A Five Nines system is sometimes described as “always on”; therefore, it describes a gold standard, but comes at a cost. When designing a system, think of each higher order of availability as being 10 times more difficult to achieve than its lower-order neighbor.

Therefore, Five Nines (5 minutes per year) is 10X more difficult than Four Nines (53 minutes per year), which in turn, is 10x more difficult than Three Nines (530 minutes or 8.7 hours per year).

Planned and unplanned outages

We might say that a system had *eight-hour outage*, which meaning that it was unavailable to its users for eight hours.

However, it is important to make a distinction between planned and unplanned outages. In our example, if we plan to bring a system down over the weekend, when users do not require it, then it does not materially affect users’ availability.

However, if we suffer the same eight-hour unplanned outage during the working day, our users cannot conduct business. Therefore, unplanned outages generally are more serious than planned outages.

Service Level Agreement

The *Service Level Agreement* (SLA) of a system is the availability of the system that is agreed to between its owner and its users. For example, a system might have Five Nines or Three Nines availability.

The SLA might also specify restrictions, such as during working hours, or not including networking issues. In reality, an SLA description can vary from a short phrase, such as Five Nines, to a multi-page contractual agreement with restrictions and even penalty fees.

Recovery Time Objective

Recovery Time Objective (RTO) is a target that defines the amount time after which a system becomes available again after a failure. Consider the scenario that is shown in Figure 8-3.

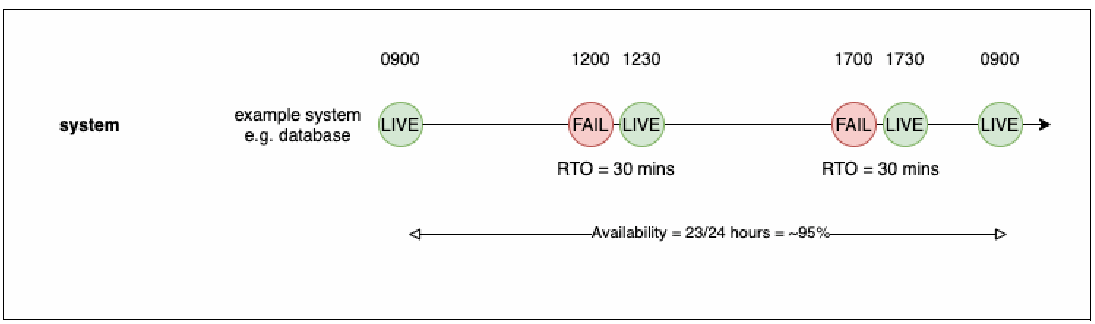

Figure 8-3 Recovery Time Objective

We can see an example timeline for a database system. At 0900, it is available for work; however, it fails at 1200. By design, the system has an RTO of 30 minutes if it is available again by 1230. Moreover, the system fails again at 1700, and is available again at 1730; again, an RTO of 30 minutes. Notice that over a 24-hour period, the accumulated one-hour downtime results in an availability of about 95%.

Synchronous replication

The most common mechanism to provide a highly available system is to replicate data between a primary location and a secondary location. If a failure occurs, the data at the secondary location is used to restart the system.

With synchronous replication, the data must be physically written to the primary and secondary locations before it can be considered logically complete.

Consider the scenario that is shown in Figure 8-4.

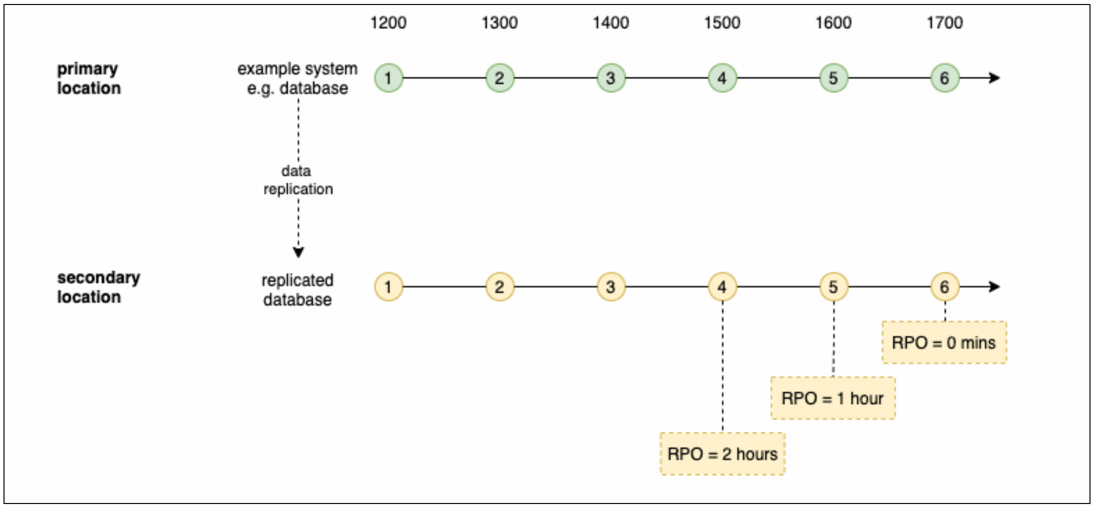

Figure 8-4 Synchronous replication

At 1200, data 1 is written to a database at the primary location. Until these data are safely stored at the secondary location, control is not returned to the writer. The remote operation increases the latency of the data write because the data must be stored locally and remotely before the write is considered complete.

Notice how, at each point on the timeline 1 - 6, a vertical line is drawn between the data that is written at the primary location and the secondary location. This vertical line emphasizes that this data is written synchronously; the data that is at the secondary location matches the data that is at the primary location always.

Asynchronous replication

Contrary to synchronous replication, with asynchronous replication, data is first written to the primary location, but written to the secondary location at any future point. Consider the scenario that is shown in Figure 8-5.

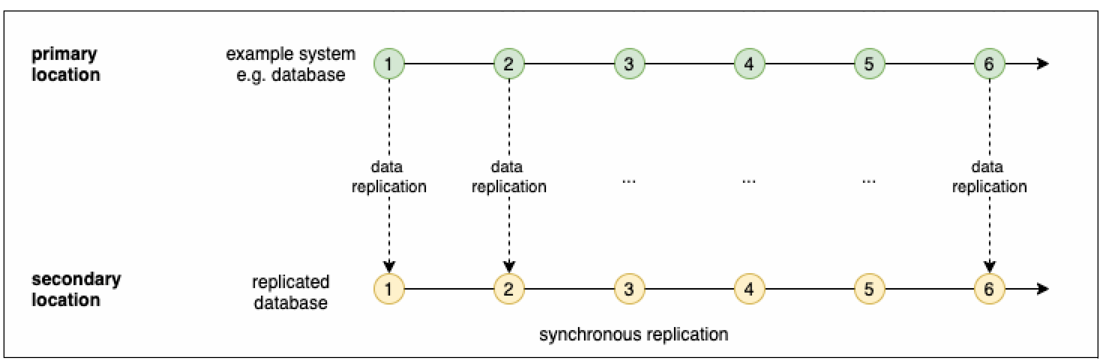

Figure 8-5 Asynchronous replication

At 1200, data 1 is written to the database at the primary location. Control is returned to the writer when this data is stored at the primary location. We can see that a period elapses until data 1 is written to the secondary location. This delayed secondary write does not increase the latency of the primary write.

Notice how, at each point on the timeline 1 - 6, an offset vertical line is drawn between the data that is written at the primary location and the secondary location. This offset emphasizes that this data is written asynchronously; the data at the secondary location is behind the data that is written to the primary location.

Recovery Point Objective

Recovery Point Objective (RPO) is a target that defines the amount of data loss in a system after it is available again after a failure. Consider the scenario that is shown in Figure 8-6.

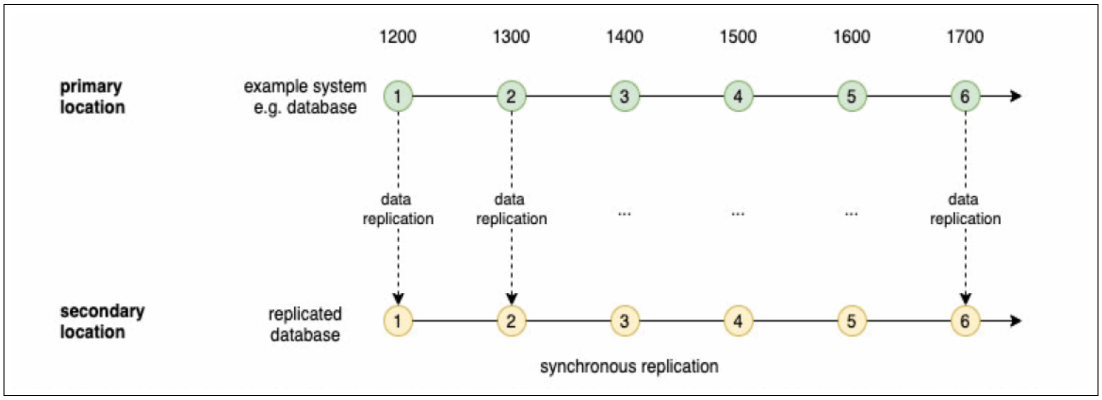

Figure 8-6 Recovery Point Objective (RPO)

We can see an example timeline for a database that is designed to run at a primary location, with a backup at a secondary location. If a failure occurs, the secondary location database becomes active. At the primary location, we can see that data 1 - 6 are written at hourly intervals and replicated to the secondary location so that they can be available to it if a failure occurs.

Imagine a failure occurs after 1700, and the secondary location database becomes active:

- If data up to 6 is available at the secondary location, the RPO is zero.
- If data up to 5 is available at the secondary location, the RPO is 1 hour.
- If data up to 4 is available at the secondary location, the RPO is 2 hours.

Systems with an RPO of zero require synchronous replication. This synchronous replication significantly affects performance because data must be written to the primary and remote secondary locations before a transaction can be considered durable.

The write to the secondary location can take a significant amount of time, which increases the latency of every transaction at the primary location such that synchronous replication is feasible up to 50 km (31 miles). Therefore, most systems are designed with asynchronous replication to minimize the performance effect at the primary location, and accept the potential data loss if a primary system failure occurs.

Persistent volumes

A *persistent volume* (PV) is a Kubernetes component that represents the external physical storage where long-term data is stored. Most importantly, in contrast to the default file storage that is available to a pod, data that is stored in a PV persists over a pod restart.

A persistent volume is accessed by a pod by using a persistent volume claim (PVC). A pod mounts a PVC at a named point in a container file system where it can be accessed to read or write data to the PV.

A persistent volume can represent different types of storage technology, including block, file, and object. Moreover, each PVC features an associated access type that defines the access concurrency that it supports, such as Read-Write-Once (RWO) or Read-Write-Many (RWX).

For more information about PV access modes, see this Kubernetes Documentation web page.

For a business resilience perspective, a PV is the unit to be replicated from one site to another. If the workload does not include a software replication feature (such as Kafka geo-replication or Db2 Data Replication), the replication of the PVs must provide a way to ensure that the data is present for the workload on another site. This replication can be synchronous or asynchronous.

Depending on the storage technology, the replication can be done by using a software approach or managed at the hardware level. The software approach relies on PV snapshot and resulting object copy. This approach can be implemented by storage software, such as Velero or Red Hat OpenShift Data Foundation. The hardware approach relies on the underlying capabilities of the storage hardware boxes and supporting product, such as IBM Spectrum Scale.

Data consistency

Data consistency refers to the following concerns:

- Whether data that is copied from a primary location to a secondary location remains the same.
- Whether the relationship between two or more different pieces of data is maintained after these data are copied between locations.

A good example of the first case is when a single file is copied from a primary location to a secondary backup location. If the file at the secondary location is different from the primary, we think of it as inconsistent with the primary. This issue often occurs because the primary file was updated after the copy was made; that is, data is now missing from the secondary location. If we use the secondary data to recover our system, we have a nonzero RPO.

A good example of the second case is exemplified by database file management. A database has two kinds of file: a data file that holds the database table definitions and data records, and a log file that records all the transactions performed on that database.

If a failure occurs, the database log file is used to restore the database data; therefore, it is crucial that these two files are consistent with each other. When we want to make a copy of a database at a secondary location, we must copy the log file *and* data file, such that they are consistent with each other, even though they both might be changing at the primary site as they are being copied.

We define the term consistency group to identify a set of files that, when replicated, maintain a consistent relationship to each other such that they can be used for recovery. The problem is not just a non-zero RPO; if these secondary files are not consistently replicated, the recovering database manager might not be able to restart because the log file and data file do not correspond to each other.

Another example of a consistency group can be found in Cloud Pak for Data. IBM Watson Knowledge Catalog consists of several microservices, such as the metastore, which is kept in a Db2 database, and other application data that is kept in file systems.

A consistency group for WKC in this case is a larger concept that spans multiple microservices, databases, and file systems. The consistency point must be reached by multiple players to avoid inconsistencies that lead to an ability to restart the WKC component at the secondary location.

8.3.3 High availability

In today's world, users of online retailers, banks, insurance companies, schools, colleges, and other establishments require services to be available around the clock, 24x7. We use the term *availability* to refer to the ability of a system to be accessed by its users to conduct useful work. The terms *uptime* and *downtime* are used to refer to periods when a system is available (uptime) or unavailable (downtime).

What is high availability?

High availability is one of the three essential components of a business resilience strategy. Specifically, it is helpful to think of HA as separate to B/R or DR, although all three are required in system with high business resilience.

We describe a system as having HA if it features significantly higher than normal uptime for its users. Specifically, HA is primarily the study of topology options for how we define a system with multiple redundant components such that if a subset of components fails, the overall uptime is not affected. HA does not worry about restoring the data for a restarted individual component, and the corresponding data integrity requirements.

Resiliency and redundancy

A system is described as highly available if it features the following qualities:

- Resilience

  A system is described as resilient when it can adapt to changing conditions and withstand and recover rapidly from disruptions. For example, if a server has a resilient power supply unit (PSU), fluctuations in the power supply voltage or amperage do not cause the server to fail (see Figure 8-7).

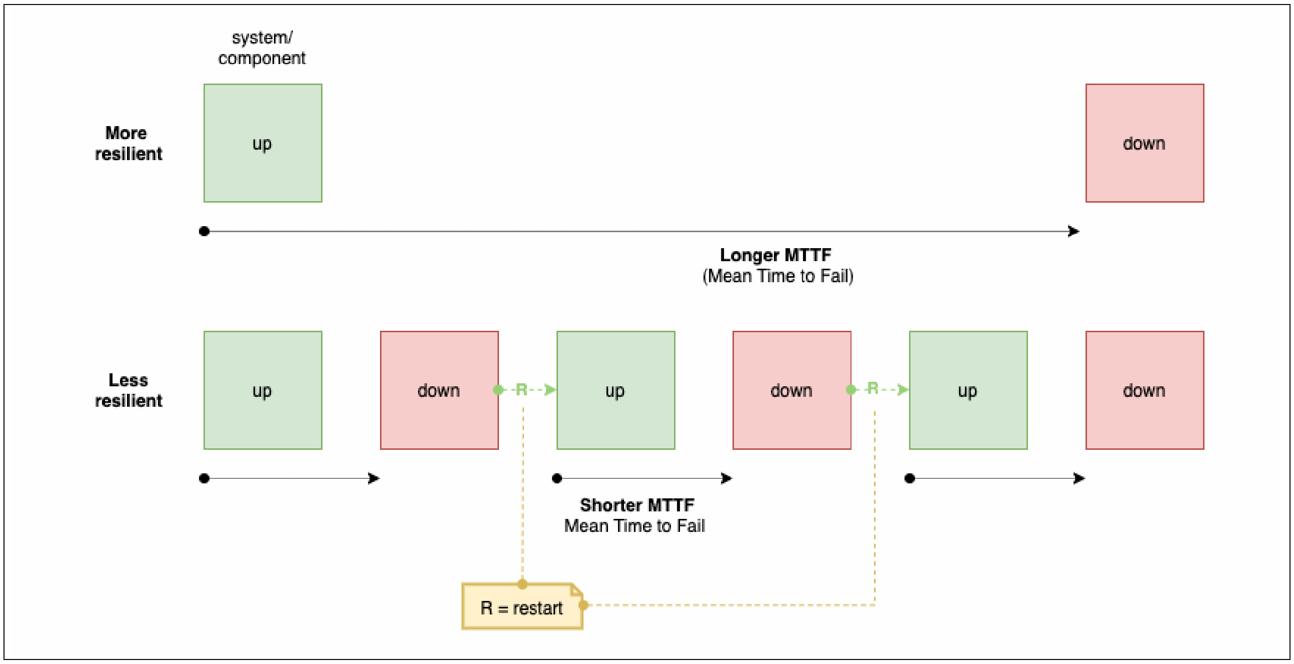

Figure 8-7 Resilience

As we can see in Figure 8-7, a system or component with a longer Mean Time to Failure (MTTF) features more resilience, which improves availability.

In a resilient software system, we expect a system component to be restarted after a failure, which might not be effective. However, the aim of a resilient component is for it to not fail.

- Redundancy

  A system or component is described as having (multiple) redundancy when it features replicated components such that each can continue processing when another fails.

  Multiple redundancy systems do not have a single point of failure. For example, a system with multiple redundant servers can process work (although at reduced capacity) if of the servers remained active (see Figure 8-8).

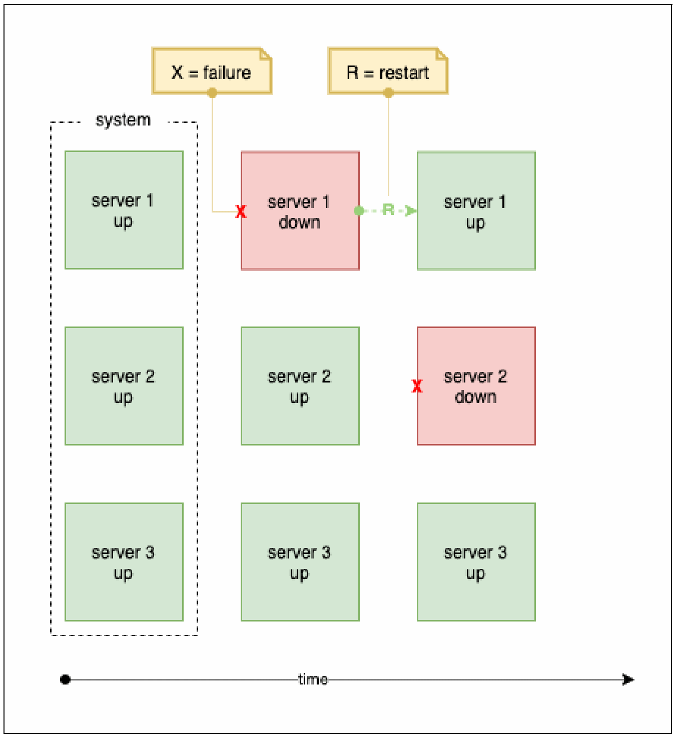

Figure 8-8 Redundancy

As we can see in Figure 8-8, a system with multiple servers can continue to process in the event of a server failure. Again, in redundant software systems, after a failure, we expect a system component to be restarted – but in contrast to a resilient system, we design redundancy into our systems because we expect failure.

Resiliency and redundancy complement and reinforce each other to create a highly available system. In our example, we improve HA by using a server with a resilient PSU, and then increase HA again by having multiple redundant servers.

Multiple sites

To further improve availability, a system might often be deployed across two or more sites. We think of a site as hosting an independent set of compute, storage, and network resources from which we build our system components.

We describe sites geographically close to each other as being in the same region, in contrast to remote sites in different regions (see Figure 8-9 on page 594).

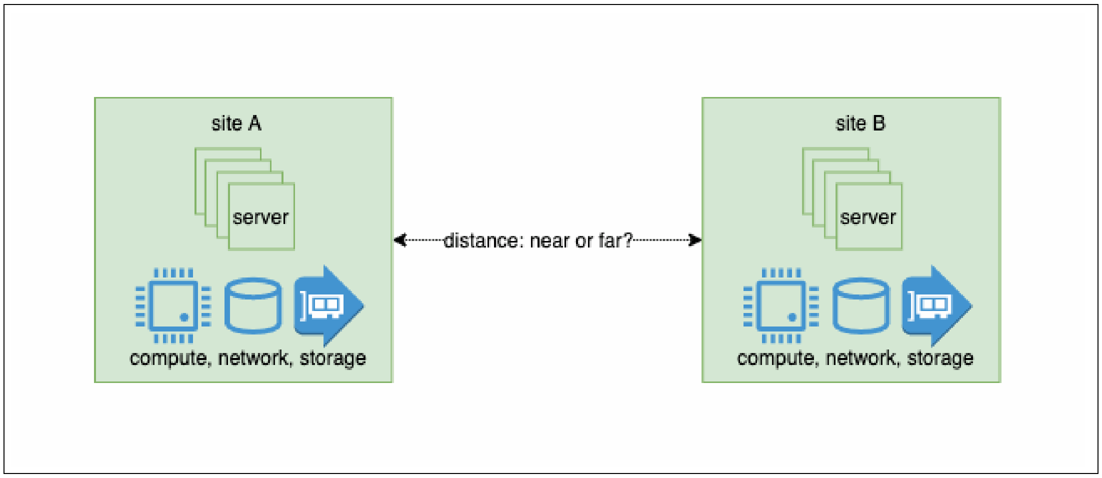

Figure 8-9 Multiple sites

Multiple sites are an essential part of HA because they use location independence. The extent of this independence depends on the proximity of the sites to each other. For example, sites on different sides of a large city are not affected by an electricity substation failure, whereas sites on different sides of a continent might not be affected by the same earthquake.

DCs and AZs

In the physical world, we think of a Data Center (DC) as a site; for example, a physical building. In cloud systems, the term *Availability Zone* (AZ) refers to a site. Think of these terms as the same thing; an AZ is a logical DC.

A cloud provider must host their AZs in physical locations; however, the key point is that they do it because an AZ (like a DC) has a set of independent physical resources, including compute, network, storage, and power, to protect against single site failures (see Figure 8-10).

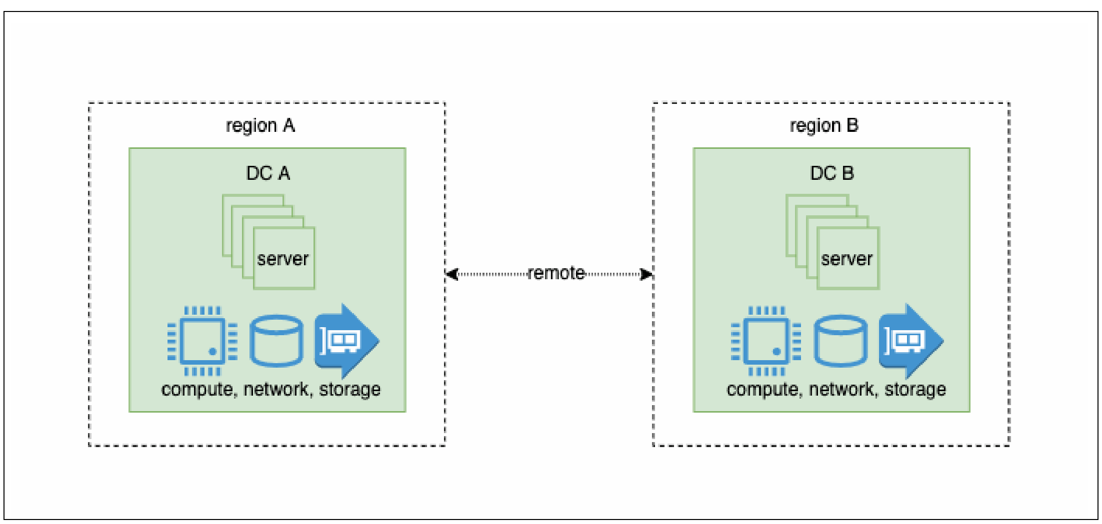

Figure 8-10 DCs and AZs

A topology that is based on AZs tends to have more sites than one that is based on physical data centers. This issue is economy of scale at work; cloud providers are in the business of providing multiple sites (AZs) within a geographical region, and having multiple geographical regions.

Moreover, they can share the cost of their sites across its entire user base, which makes it more cost effective. A physical data center tends to be owned by a single company; therefore, it is relatively rare to have more than two of them, usually in different regions. However, a tradeoff exists: when you use a cloud AZ, you are relying on a service provider, one who might provide a more cost-effective solution than you can provide.

Multiple regions

AZ or DC sites within the same region are in close proximity to each other, typically less than 30 kilometers (18.6 miles). This separation provides almost complete resilience against minor events, such as a burst water pipe or electricity substation failure.

Sites within different regions typically are hundreds, and sometimes thousands of kilometers apart. This separation provides almost complete resilience against minor and major events.

We can see that in the same region, sites are independently protected from minor failures, and can be synchronously linked by using high speed, low latency networks.

This second factor is important. It allows us to define a single Kubernetes cluster between same-region sites, which simplifies the development and operation of an HA system within a region.

Likewise, between different regions, sites are independently protected from minor and major failures and cannot be synchronously linked.

Equally, this second factor is important because sites in different regions are in separate Kubernetes clusters. This configuration increases the complexity of a multi-region HA system.

In Figure 8-11, we see a typical site topology for a physical deployment. Two data centers is usually considered the gold standard for location independence because DCs are expensive to build and operate. These DCs are in two different regions, which provides location independence against minor and major failures.

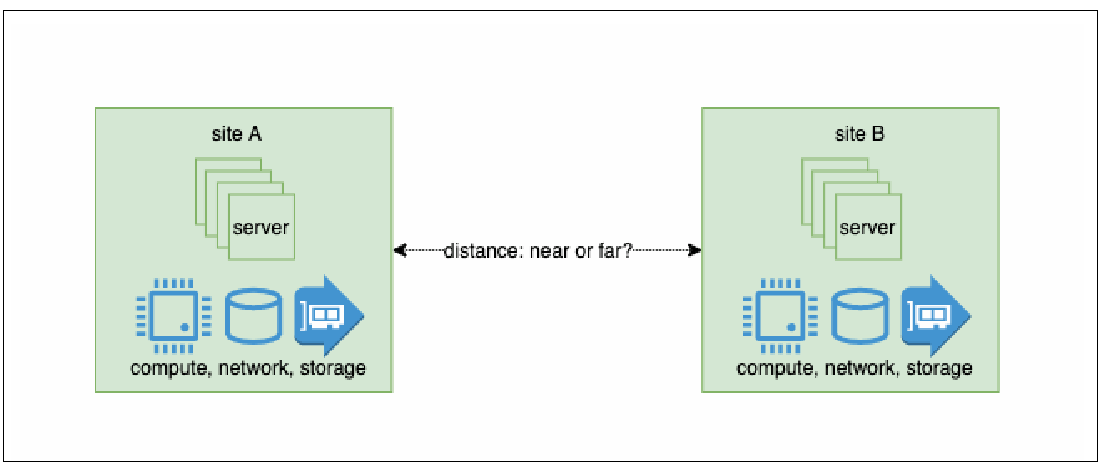

Figure 8-11 A typical cloud topology

In Figure 8-12, we see a typical cloud topology. Three availability zones are considered the gold standard. Moreover, it is common to have two regions, each with three AZs. This might configuration seem like an extravagance, but it is economically viable because cloud providers can share the cost of multiple AZs and regions across their entire user base.

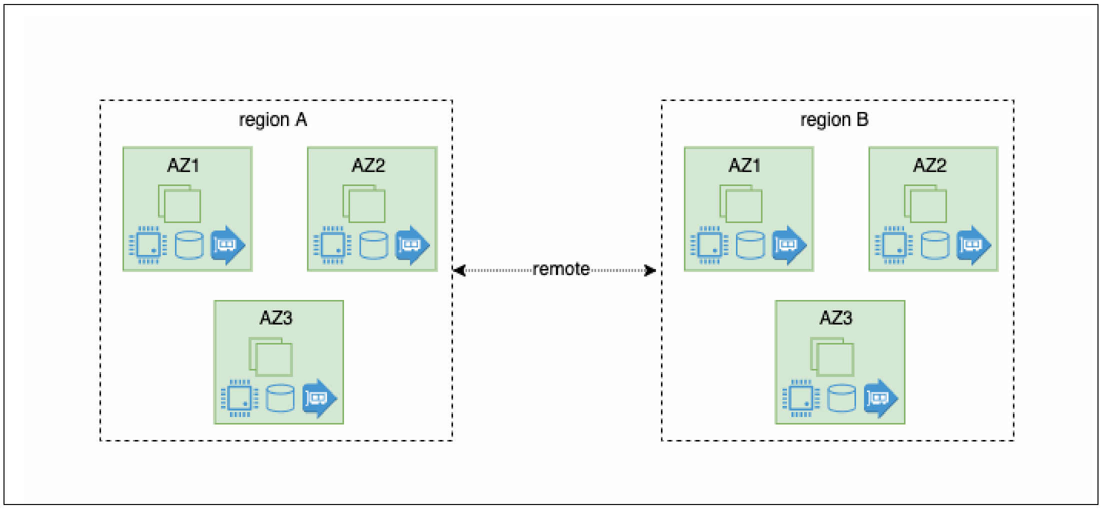

Figure 8-12 Three availability zones

Active-passive and active-active

When we design a redundant system with multiple servers and multiple sites, we have a choice as to which of these components are active by default. We describe the two basic approaches as active-passive and active-active.

Active-passive

In the simplest form of active-passive, we consider a system with two servers (see Figure 8-13).

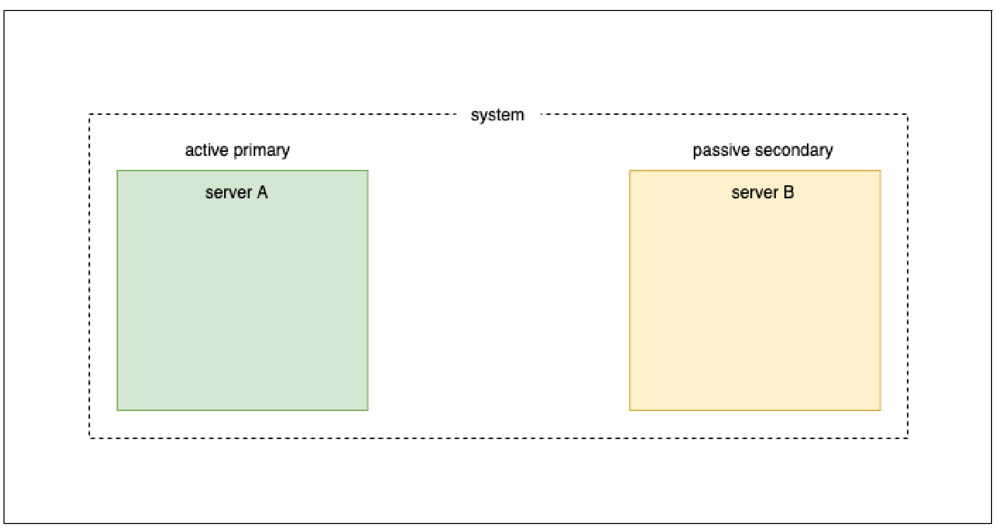

Figure 8-13 Active-passive servers

In this configuration, the primary is active (processing work requests) while the secondary is passive. In contrast to the primary, the secondary is not processing work requests, but it is available to do so if the primary fails.

We can also apply the active-passive concept to any elements of topology; for example, and active-passive servers, we can have active-passive sites.

Two versions of an active-passive configuration are available, as shown in Figure 8-14.

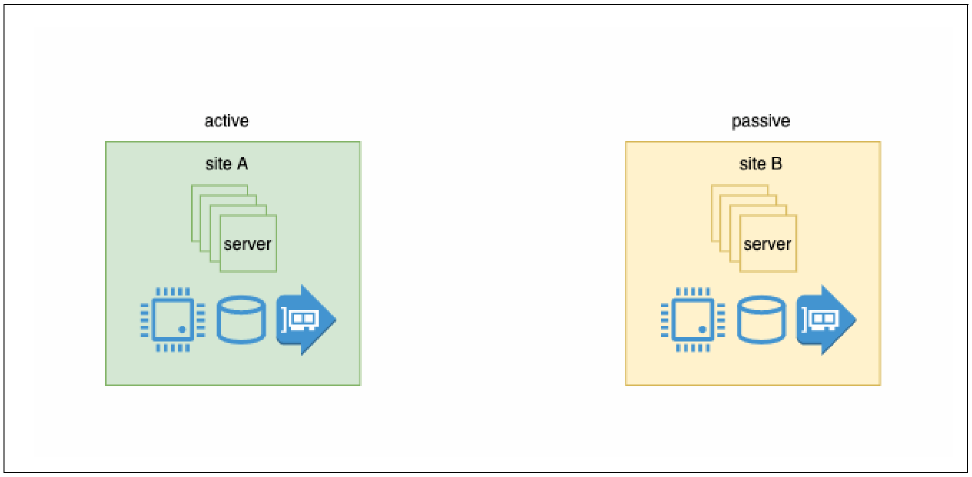

Figure 8-14 Active-passive configuration

In a hot standby configuration, the secondary is always running and ready to process work immediately if a failure occurs.

In contrast, a cold standby configuration requires the secondary to be started before it can replace the primary as the processor for work requests.

Active-active

In this topology, the primary *and* secondary are processing work at the same time. If a failure of the primary or secondary occurs, all work is processed the element that is unaffected by the failure.

For example, Figure 8-15 shows an active-active sites configuration.

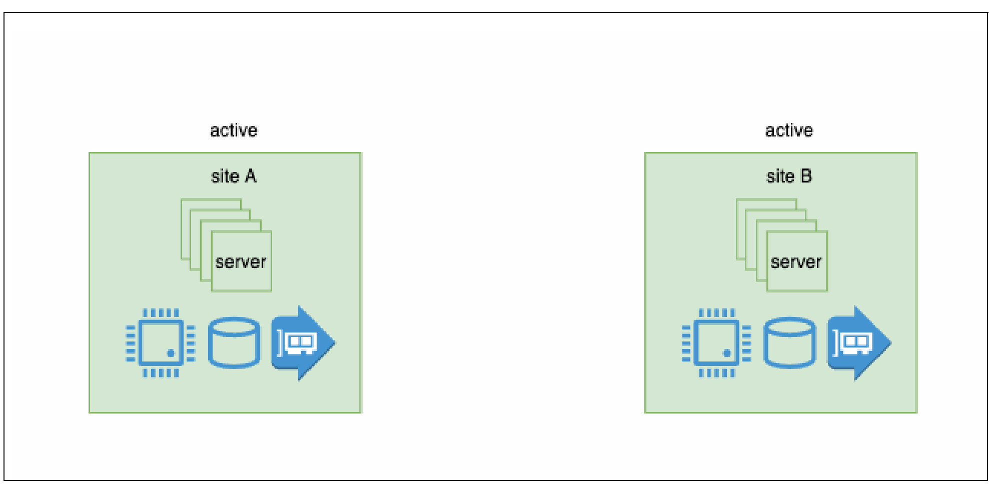

Figure 8-15 Active-active sites

As you can see, an active-active system requires more resources including electricity, compute, network, storage, and software licenses. This configuration represents the trade-off between cost and availability (highly available systems cost more to operate).

Continuous availability

The ability to continue processing after a failure without worrying about data that might be associated with a failed component is called continuous availability (CA).

In a CA system, work that is in-flight during a failure might be affected, but new work can be processed by a system component that did not fail. Although this new work is being processed, in-flight data that is affected by the failed system component is recovered in parallel by server restart, B/R, or DR. Therefore, continuous availability is a helpful way of thinking about a system that is available for new work after a failure.

Notice again how we make a clear distinction between HA and B/R or DR. Specifically, CA is not concerned with data recovery if failure occurs. This statement might seem strange; for example if a failed server includes vital customer data associated with it, we must be concerned. However, data restore and integrity is addressed when we discuss B/R or DR aspects of business resilience.

Scenarios

As we have seen, many aspects must be considered in the design of a HA system. However, some combinations of options are more common than others. We discuss these options in more detail in the remainder of this topic, but let us first enumerate them:

- For physical systems:
  - Single DC, single server
  - Single DC, multiple servers
  - Multiple regions DC, cold-standby
  - Multiple regions DC, active-active
- For cloud systems:
  - Single AZ, single server
  - Multiple AZ, single region
  - Multiple region AZ, active-active

We discuss each option next. Each configuration, whether physical or cloud, builds upon its previous topology, and adds an incremental degree of availability.

Single DC, single server

This configuration is the simplest. A single system component (such as a database server) runs in a single data center in a single region.

It is the typical developer configuration because it is easy to set up, all functions are available, and operational concerns (such as HA) are irrelevant (see Figure 8-16).

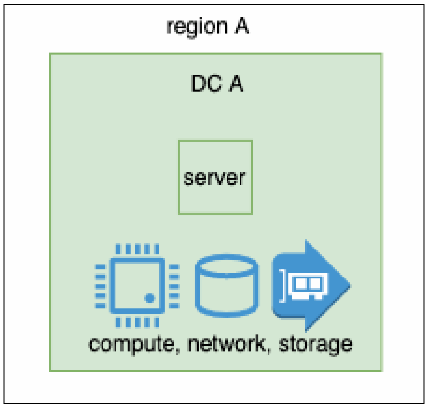

Figure 8-16 Single DC, single server

However, when used as a production configuration, it is typically accompanied by a component that restarts the system component automatically if a failure occurs.

In Kubernetes, system components run as a set of operating system processes that are inside a pod container, which are part of a deployment or stateful set. If one of these elements fails, Kubernetes restarts them automatically, which restores the component to its running state.

For systems that feature low importance or occasional usage, the single server, single DC configuration is adequate.

Single DC, multiple servers

This configuration extends the single-server option by adding one or more servers. It is still easy to configure because all components are provisioned in a single data center (see Figure 8-17).

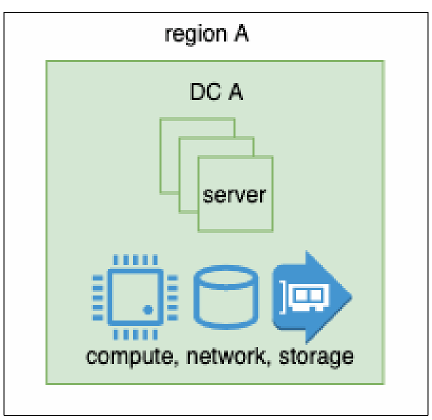

Figure 8-17 Single DC, multiple servers

This configuration is popular for on-premises deployments; a dual server approach provides significantly improved availability over the single-server configuration.

In Kubernetes environments, it is more typical to run three server instances to match the typical minimum number of compute nodes in a cluster. This configuration provides a highly available, two-instance system, even if one compute node fails.

As with a single-server configuration, Kubernetes provides the built-in restart mechanism for the stateful set or deployment containing the server.

Finally, because multiple servers act as a single logical provider of service, it is necessary to allocate incoming work to a suitable instance. Kubernetes provides the pod, service, and ingress components that enables incoming work to be balanced across multiple pods that are hosting the services (see Figure 8-18).

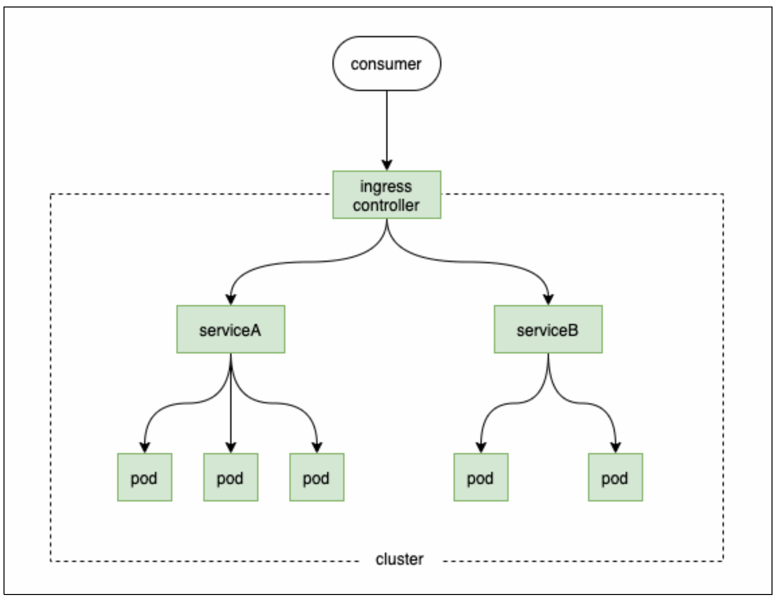

Figure 8-18 Kubernetes environment

In this example, work requests from a consumer application are routed by way of ingress to one or more services that refer to pods that host an active server instance. You can set different ingress rules to determine how work that is balanced across the different instances processes a request.

For more information about these features, see this Kubernetes Documentation web page.

Multiple region DC, cold-standby

This scenario maintains multiple servers, and adds the extra dimension of a data center site in a separate region. The data centers typically are hundreds and sometimes thousands of kilometers apart, which provides almost complete resilience against minor and major events; for example, we do not expect an earthquake to affect bot the eastern and western coasts of a large continent.

Notice how in this topology the remote secondary site is not normally active; it processes work only if the primary system fails (see Figure 8-19).

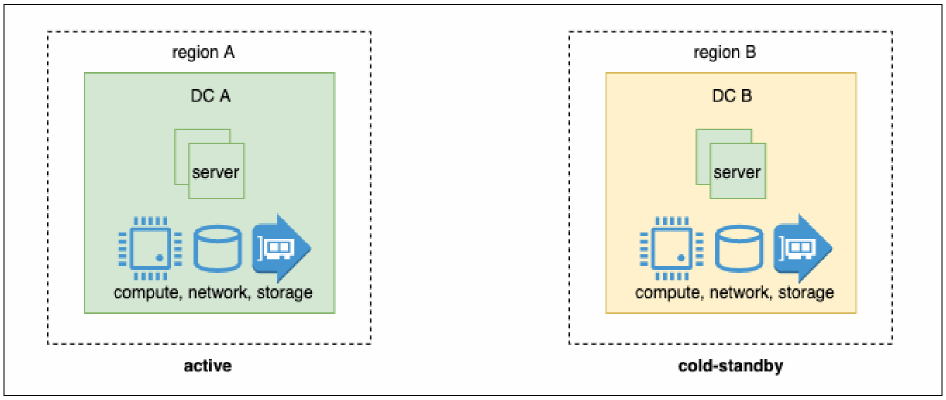

Figure 8-19 Multiple region DC, cold-standby

This configuration accepts that the remote site might take several hours to become the active site. For this reason, the secondary systems are in cold-standby mode; therefore, they must be restarted before they can process work. This configuration minimizes resource usage (compute, storage, network, and licenses) at the expense of an increased RTO.

Because this configuration features two data centers that are separated by a significant distance, we also must consider network load balancing between these two data centers (see Figure 8-20).

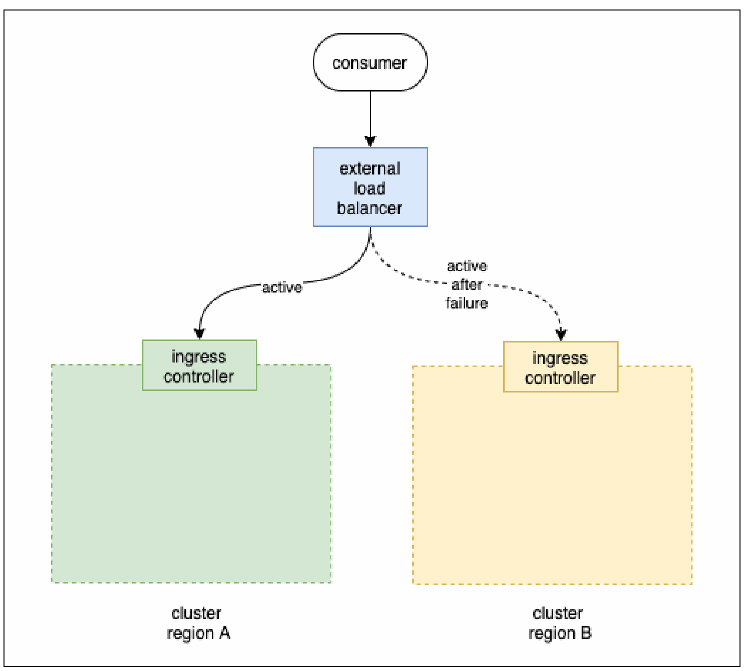

Figure 8-20 Network load balancing

In normal running, this configuration requires an external load balancer to route work requests from a consuming application to the active primary data center. If a failure occurs, the external load balancer directs work to the active secondary site after the cold-standby secondary site is made active.

Multiple region DC, active-active

This configuration represents the gold standard in HA for a physical deployment. The remote region data centers provide almost complete resilience against minor and major events. Moreover, these regions are configured in an active-active topology; that is, both can process work requests at the same time, and handle the failure of the other (see Figure 8-21).

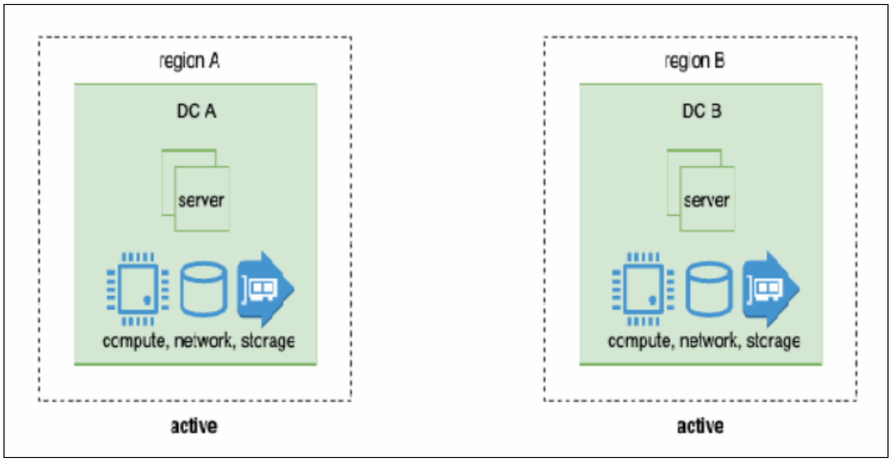

Figure 8-21 Multiple region DC, active-active

Again, this configuration features an external load balancer that routs work requests to the two active data centers (see Figure 8-22).

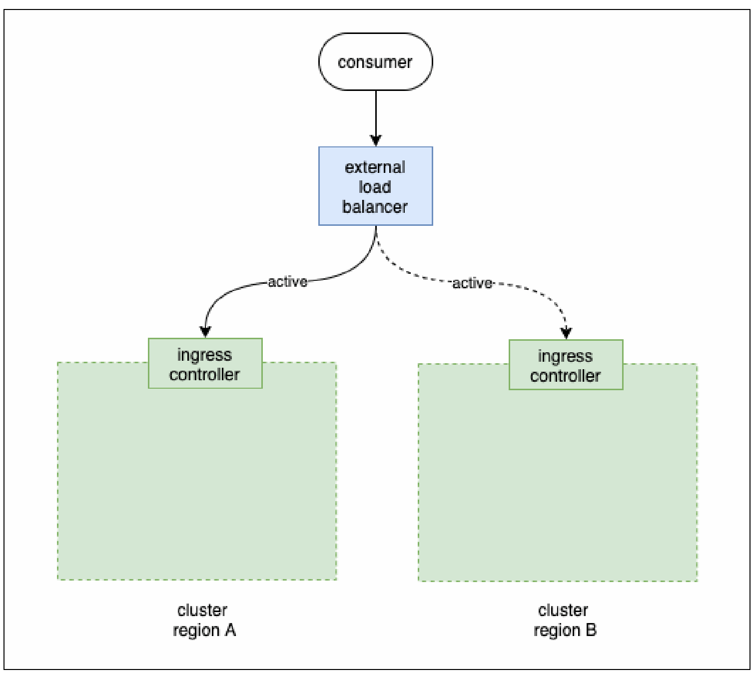

Figure 8-22 External load balancer routing work requests to both of the two active data centers

If either site fails, the external load balancer directs work to the remaining active secondary site.

Cloud systems: Single AZ, single server

This cloud configuration is the simplest. It is analogous to a single server in a single physical data center. A single component runs in a single AZ in a single region. Although cloud providers make it relatively easy to provision servers into multiple AZs, a single AZ is the typical cloud developer configuration because it is lowest cost, easy to set up, all functions are available, and operational concerns (such as HA) are irrelevant (see Figure 8-23).

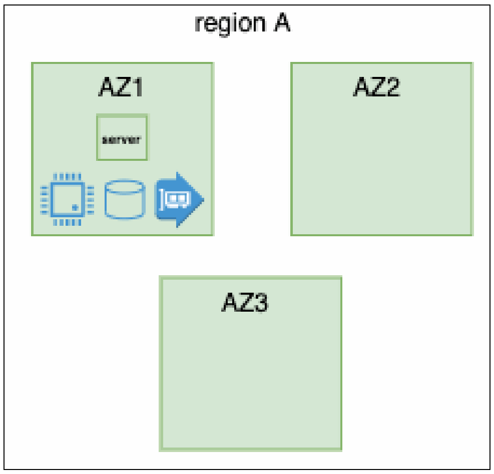

Figure 8-23 Single AZ, single server

When used as a production configuration, it is typically accompanied by a component that restarts the system component automatically is a failure occurs. Also, as with physical systems, Kubernetes restart the pod container within a deployment or stateful set automatically, which restores the component to its running state.

For systems that feature low importance or occasional usage, the single server, single AZ configuration is adequate.

Multiple AZs, single region

This configuration extends the single AZ topology by adding servers across an increased number of AZs, typically three. Therefore, our system operates across three sites within the same region, which is typically less than 30 kilometers (18.6 miles).

Cloud providers make this option easy to configure. It also is cost effective for the use because of the provider's economies of scale. The multiple AZs protect from minor events, such as a local power outage, though major events, such as an earthquake, which can impact all AZs in the same region (see Figure 8-24 on page 604).

Figure 8-24 Multiple AZs, single region

A single Kubernetes cluster can be stretched across multiple AZs within the same region because of the availability of high bandwidth and low latency of communications within a region. This configuration enables Kubernetes to restart work on a failed AZ elsewhere within the cluster.

Because AZs are configured in an active-active topology, every server can process work requests simultaneously (see Figure 8-25).

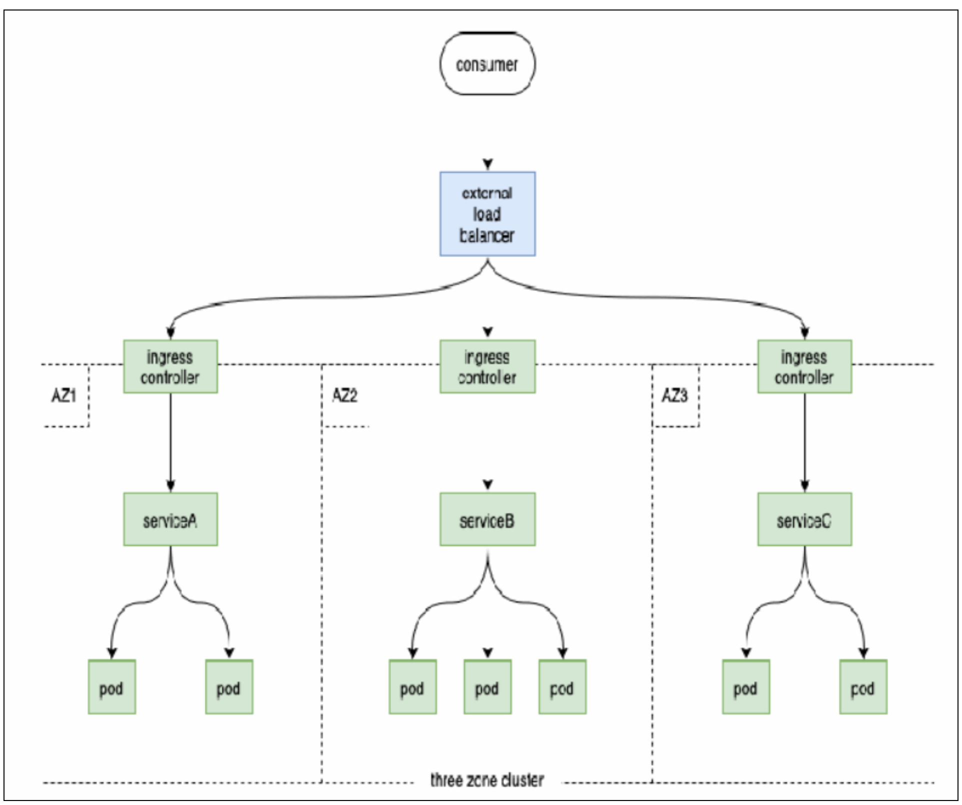

Figure 8-25 External load balancer routing work requests to one of the active AZs

This configuration features an external load balancer that routs work requests to one of the active AZs. If of any AZ fails, the external load balancer directs work to a remaining AZ.

Note: Cloud providers feature different ELB technology, which can be used to a greater or lesser degree by Kubernetes ingress objects.

Multiple region AZ, active-active

This configuration represents the gold standard in HA for a cloud deployment. In normal running, every AZ in each region is available to process work. The distance between remote regions provides almost complete resilience against minor and major events. If a failure occurs, whether of AZ or region, an AZ always is available to handle the failure (see Figure 8-26).

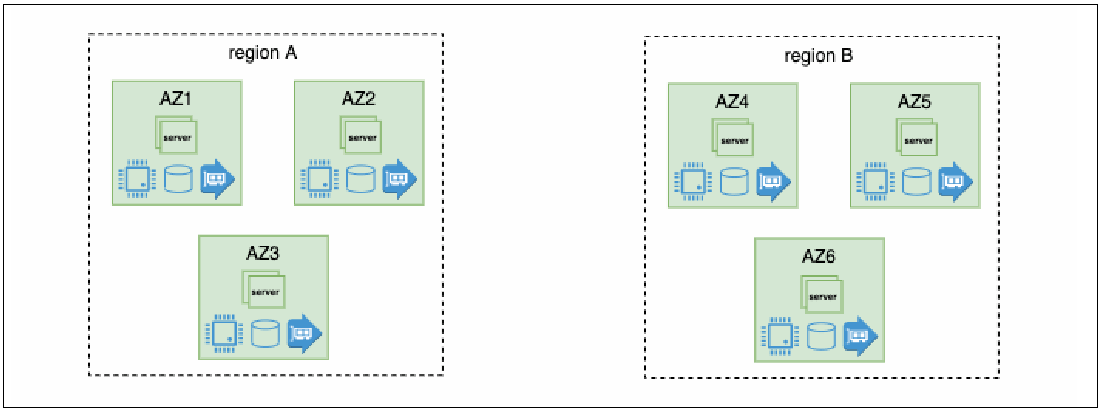

Figure 8-26 Multiple region AZ, active-active

Again, this configuration features an external load balancer that routs work requests to the two active data centers/

A single Kubernetes cluster can be stretched across multiple AZs within the same region, but a separate cluster is required in each region (see Figure 8-27 on page 606).

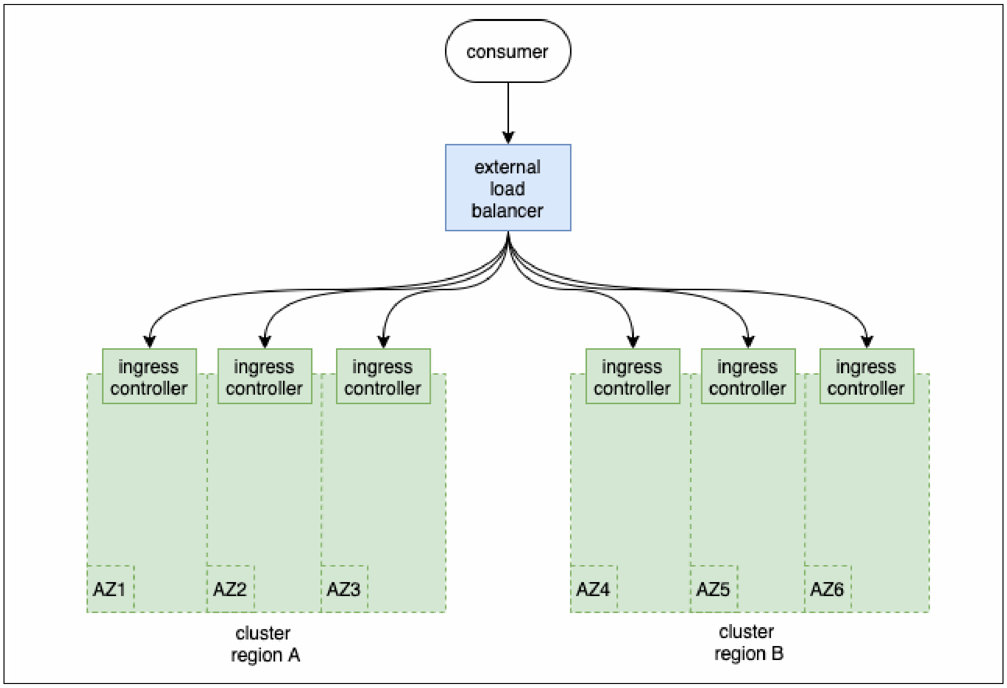

Figure 8-27 Kubernetes environment

If an AZ fails within a region, failed servers are restarted and the work is rebalanced to the remaining AZs within that region.

If either site fails, the external load balancer directs work to the remaining active secondary site.

Summary

In this topic, we discussed the major concepts in HA, including resiliency and redundancy. We reviewed topology options, including multiple sites, data centers, availability zones, and regions.

We also looked at active-active and active-passive options and cold and hot standby.

8.3.4 Disaster recovery

DR is a part of business continuity that deals with the immediate effect of an event. Recovering from a server outage, security breach, or natural disaster fall into this category. DR often includes several discreet steps in the planning stages, though those steps blur quickly during implementation because the situation during a crisis almost never exactly goes to plan.

DR involves stopping the effects of the disaster as quickly as possible and addressing the immediate aftermath. This process might include shutting down systems that were breached, evaluating which systems are effected by a flood or earthquake, and determining the best way to proceed.

DR in the context of IBM Cloud Pak for Data

In this section, we discuss some worked examples of DR in the context of IBM Cloud Pak for Data.

Metro-DR: Two Red Hat OpenShift Container Platform clusters and stretched storage

To achieve HA and DR, customers can consider having two Red Hat OpenShift clusters, preferably across two data centers (sites) in the same metro or cloud region.

The storage cluster must be stretched and data must be synchronously replicated across the two sites. With the same requirement for latency and bandwidth to support replication, the data centers and sites must be in the same region. The storage solutions also must support stretched cluster with an arbiter or quorum node in a third location (often known as *witness* or *tie-breaker data center*).

0

> **Tip:** IBM Spectrum Scale is a proven solution for stretched cluster in support of mission-critical workloads, such as database a sample use case of IBM Spectrum Scale stretch cluster for export protocol services. Customers also deployed stretched IBM Spectrum Scale clusters over data centers to support stretched a Red Hat OpenShift Container Platform cluster.

In this approach, if one data center fails or the Red Hat OpenShift Container Platform cluster becomes inoperative, the application can still be available in another environment.

Combining continuous deployment (CD) with multiple clusters yields a DR plan: If a Red Hat OpenShift cluster becomes unavailable (temporarily or permanently), we restore the cluster or deploy a new one. In the case of a new cluster, we use the CD process to redeploy all the applications and resources.

With this approach, we can still replicate the data by using the stretch cluster. The difference now is that two Red Hat OpenShift Container Platform clusters are available now for failover and DR.

Regional-DR: Two independent Red Hat OpenShift Container Platform clusters

If Red Hat OpenShift Container Platform clusters are separated by a distance greater than a metropolitan radius, use asynchronous replication between two independent clusters for DR. This configuration is known as *Regional DR*.

IBM and Red Hat teams are collaborating to develop regional DR solution for Red Hat OpenShift, project code named `Ramen`. You can access an internal update (replay and charts) on Project Ramen-DR from UDF BoF-19.

Regional-DR capability for Red Hat OpenShift Data Foundation (ODF) is expected to generally be available from 4.11.

> **Note:** IBM Spectrum Scale uses a feature that is called Active File Management, which is based Asynchronous Disaster Recovery (AFM-DR) to provide asynchronous DR.
>
> For more information, see this IBM Documentation web page.

8.3.5 Backup and restore

B/R refers to technology and practices that are used for making periodic copies of data and applications to a separate, secondary device. Then, those copies are used to recover the data and applications (and the business operations on which they depend) if the original data and applications are lost or damaged because of a power outage, cyber-attack, human error, disaster, or some other unplanned event.

B/R is an essential component of a business' DR and resilience strategy.

Developing a B/R strategy

Consider the following questions as you develop a B/R strategy:

- Will the database be recoverable if failure or outage occurs?
- How much time can be spent recovering Cloud Pak for Data?
- How much time can elapse between backup operations?
- How much storage space can be allocated for backup copies and archived logs?
- In Cloud Pak for Data, when we talk about backups, we are referring to:
  - Project backup
  - Deployment space backup
  - Catalog backup
  - Component-specific backup

  Therefore, will catalog backup suffice? Or, is a full cluster or Red Hat OpenShift project backup required here?
- Should a standby system be configured manually or through HADR?

Some other key points about B/R:

- A database backup is equivalent to any other data backup; that is, a copy of the data is taken and stored on a different medium in case of failure or damage.
- The simplest case of a backup involves shutting down the database to ensure that no further transactions occur and then, simply backing it up. Then, you can recreate the database if it becomes damaged or corrupted.
- *Recovery* is defined as the recreation of a database and its stored data and metadata.
- *Version recovery* is the restoration of a previous database version by using an image that was created during the backup operation.
- *Roll forward recovery* is the reapplication of transactions that were recorded in the database log files after the database or table space backup image was restored.

For more information, see this IBM Support web page.

Recovery log files and the recovery history file are created automatically when a database is created (see Figure 8-28). These log files are important if you must recover data that is lost or damaged.

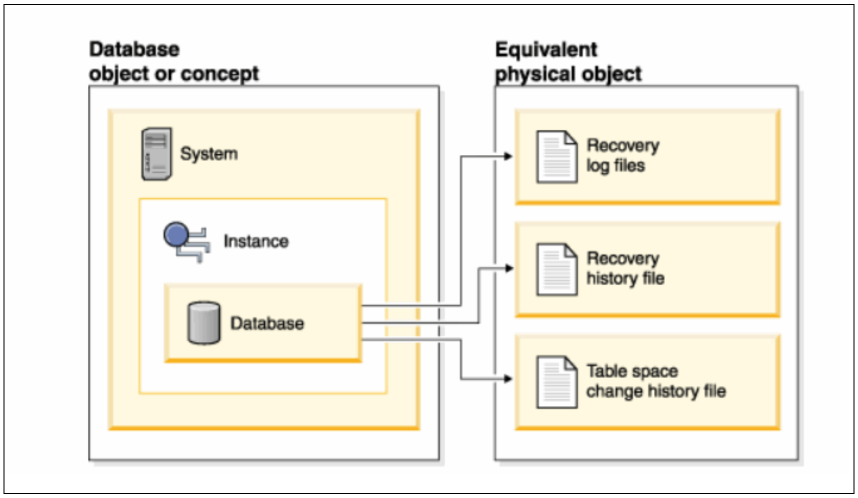

Figure 8-28 Recovery log files and the recovery history file are created automatically

The table space change history file, which is also in the database directory, contains information that can be used to determine which log files are required for the recovery of a particular table space.

Consider the following points about recoverable versus non-recoverable:

- Recoverable database backup operations can be performed offline or online.
- If the database is non-recoverable, database restore and roll forward operations must be performed offline.
- An online backup is one in which other applications can connect to the database during the backup operation.
- Because the database is recoverable, online table space restore and roll forward operations also can be performed.

Worked example: Db2 universal container

In this section, we discuss a worked example of Db2 universal container (Db2U). Consider the following points:

- Db2U is a set of containers for the Db2 product, which are based on a micro services architecture and optimized for Kubernetes.
- We use a database (Db2U) on Red Hat OpenShift and Cloud Pak for Data to demonstrate a simple B/R use case.

For a Db2U deployment, the method for the database backup is to use a Single System View (SSV) backup. This strategy helps you back up all database partitions simultaneously, including the catalog partition.

SSV backups provide a single time-stamp for all database partitions, which makes the recovery process simpler. The process includes the following steps:

1. Before you can run the **db2 backup** or **db2 restore** commands, you must identify the node where the Db2 database service is deployed. In this example, `cp4ba` is the namespace and the core Db2U pod is `-db2u-0`. Run the command that is shown in Example 8-1.

Example 8-1 Identify the node where the Db2 database service is deployed

```
# get pods -n cp4ba | grep db2u
  c-db2ucluster-cp4ba-db2u-0                   1/1     Running     0      11h
  c-db2ucluster-cp4ba-etcd-0                   1/1     Running     0      11h
  c-db2ucluster-cp4ba-instdb-fq9pv             0/1     Completed   0      11h
  c-db2ucluster-cp4ba-ldap-7d57d4d478-nz8q9    1/1     Running     0      11h
  c-db2ucluster-cp4ba-restore-morph-r9zr2      0/1     Completed   0      10h
  db2u-operator-manager-c6897d5f8-m9lp9        1/1     Running     0      11h
```

2. Choose the SMP db2u pod by using the following command:

   ```
   # kubectl exec -it -n cp4ba c-db2ucluster-cp4ba-db2u-0 bash
   ```

3. Back up the Db2u database online (see Example 8-2).

Example 8-2 Backing up a Db2u database

```
$ su - db2inst1
$ db2 LIST ACTIVE DATABASES
     Database name                              = ODMDB
     Applications connected currently           = 0
     Database path                              =
/mnt/bludata0/db2/databases/db2inst1/NODE0000/SQL00001/MEMBER0000/
   $ cd /tmp
   $ mkdir -p backup/backup_odmdb
   $ db2 connect to ODMDB
   $ db2 backup db ODMDB on all dbpartitionnums online to /tmp/backup/backup_odmdb  include logs
without prompting

    Part  Result
    ----  ------------------------------------------------------------------------
    0000  DB20000I  The BACKUP DATABASE command completed successfully.
    Backup successful. The timestamp for this backup image is : 20220727025637
```

4. If you want to restore a Db2 database from an online backup, use the following command:

   ```
   # Kubectl exec -it -n cp4ba c-db2ucluster-cp4ba-db2u-0 bash
   ```

5. To restore up a Db2u database online (DBNAME is ODMDB), run the following command (see Example 8-3).

Example 8-3 Restore up a Db2u database online: DBNAME is ODMDB

```
$ su - db2inst1
$ db2 list history backup all for ODMDB
    Op Obj Timestamp+Sequence Type Dev Earliest Log Current Log  Backup ID
    -- --- ------------------ ---- --- ------------ ------------ --------------
    B  D  20220727025637001   N    D  S0000012.LOG S0000012.LOG
    ---------------------------------------------------------------------------
    Contains 3 tablespace(s):

    00001 SYSCATSPACE
```

```
    00002 USERSPACE1
    00003 SYSTOOLSPACE
    -----------------------------------------------------------------------------
    Comment: Db2 BACKUP ODMDB ONLINE
    Start Time: 20220727025637
    End Time: 20220727025647
    Status: A
    -----------------------------------------------------------------------------
    EID: 19 Location: /tmp/backup/backup_odmdb

       $ cd /tmp/backup/backup_odmdb
       $ db2 connect to ODMDB
       $ db2 list applications

        Auth Id  Application    Appl.      Application Id
DB     # of agent
        -------- -------------- ----------
------------------------------------------------------ -------- -----
        DB2INST1 db2bp          139       *LOCAL.db2inst1.220727133626
ODMDB    1

       $ db2 force application all
       $ db2 deactivate database ODMDB
       $ db2stop force
       $ ipclean -a
       $ db2set -null DB2COMM
       $ db2start admin mode restricted access
       Run the restore operation: db2 RESTORE DATABASE BLUDB FROM backup_dir TAKEN AT
backup_image_timestamp INTO BLUDB REPLACE EXISTING WITHOUT ROLLING FORWARD
(backup_image_timestamp = 20220727025637)
       $ db2 RESTORE DATABASE ODMDB FROM /tmp/backup/backup_odmdb TAKEN AT 20220727025637 INTO
ODMDB REPLACE EXISTING WITHOUT ROLLING FORWARD
       $ db2 rollforward db ODMDB to end of backup on all dbpartitionnums and stop
       $ db2stop force
       $ ipclean -a
       $ db2set DB2COMM=TCPIP,SSL
       $ db2start
       $ db2 activate db ODMDB
       $ db2 connect to ODMDB
```

IBM Cloud Pak for Data worked example

IBM clients that are running IBM Cloud Pak for Data must gracefully handle planned and unexpected outages. Outages must not cause substantial data loss or have a significant business impact. It is expected that Cloud Pak for Data provides a utility for performing B/R operations nondisruptively.

Specifically, clients must be able to perform B/R on:

- All the services in a specified namespace
- All the services that are associated with an installation of Cloud Pak for Data (including the primary namespace and any tethered namespaces)

B/R use cases include the following examples:

- Crash recovery from latest online backup: Online B/R requires CSI Snapshot support. Examples of such storage options include ODF and IBM Spectrum Scale.

> **Note:** A *snapshot* represents a point-in-time copy of a volume.

- Offline B/R: Offline backups support a range of storage choices because it uses Restic backups. Restic is a widely used Open Source file backup storage solution.

> **Note:** For more information about Restic, see this website.

- Point-in-time restores: Cloud Pak for Data applications support restart from point of failure. Data is then collected during an application checkpoint. Checkpointing data is metadata that an application saves to its data volumes B/Rs during checkpointing.
- Nondisruptive online backups of selected namespaces: This use case is challenging and yet can be achieved by using various strategies.

When the **`cpd-cli oadp`** command is used, online and offline B/R require an S3-compatible object store for backup storage location. An S3-compatible object store can be IBM Cloud Object Storage, Amazon Web Services (AWS) S3, MinIO, or NooBaa (see Figure 8-29).

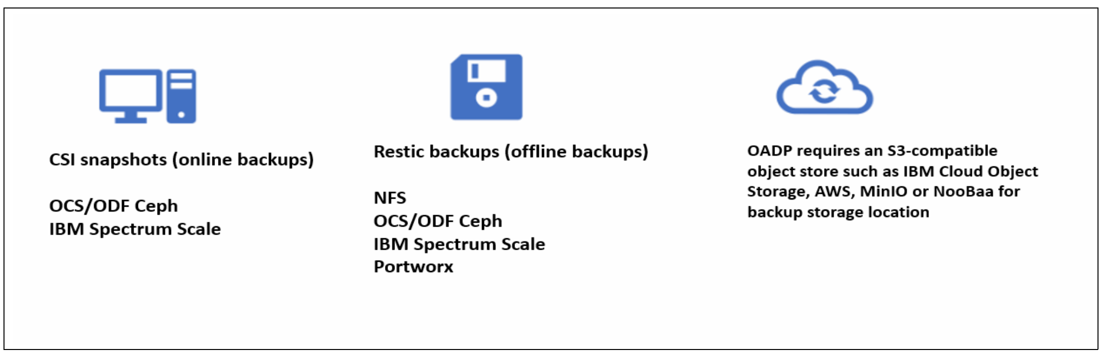

Figure 8-29 Using cpd-cli oadp command

As of this writing, different types of B/R approaches are supported for Cloud Pak for Data services. For more information, see this IBM Documentation web page.

All services support offline backups. The **`cpdbr`** utility that implements the backup uses a quiesce step to bring services to a consistent state. Each service stops the use of its data volumes before the storage layer backup is started and then, resumes when the backup is finished. Effectively, this means that external operations for Cloud Pak for Data services are not available for the entire duration of the backup. A full offline backup for a Cloud Pak for Data instance typically takes a few hours to complete.

IBM Cloud Pak for Data B/R approach

In general, the following B/R methods are available for online, offline, and volume B/R:

- Online B/R with OADP: This method is a key feature of Cloud Pak for Data 4.5. It uses the Checkpoint mechanism and can help minimize disruption to the production cluster during backup.
- Cloud Pak for Data OADP backup REST service: This method also is known as *CPDRB API Service with OADP*. This feature also was introduced by Cloud Pak for Data 4.5. By using this feature, you can schedule a backup job easily without having to worry about timeout that is caused by a lengthy backup process. For more information, see this IBM Documentation web page.
- Offline B/R with OADP: This method is supported since Cloud Pak for Data 4.0.2 by way of OADP.
- Volume B/R: It is available since Cloud Pak for Data 3.0. It backs up data volumes only. Kubernetes objects (such as secrets, configuration maps, and pods) are not part of B/R.

Figure 8-30 shows the IBM Cloud Pak for Data B/R approach.

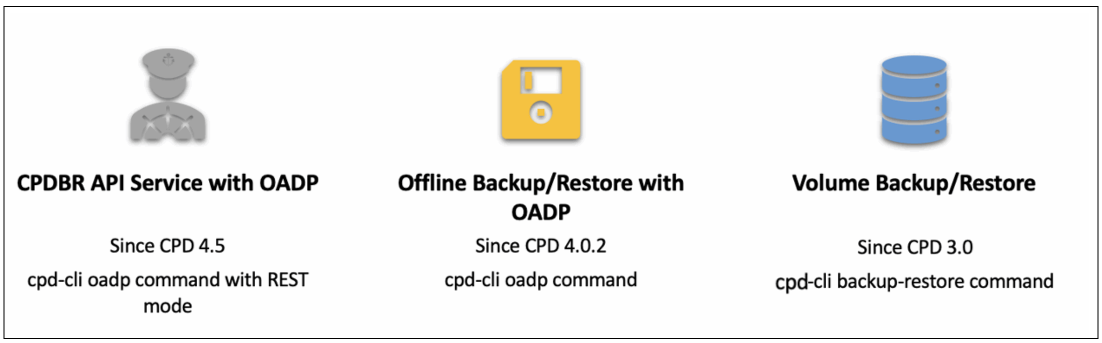

Figure 8-30 IBM Cloud Pak for Data - B/R approach

What is Red Hat OpenShift APIs for Data Protection?

Red Hat OpenShift APIs for Data Protection (OADP) is a Golang operator that was created by Red Hat. It provides a set of APIs for B/R of Red Hat OpenShift clusters. OADP can be used to back up application persistent data and cluster metadata.

OADP includes the following key features:

- Services can provide snapshots while the application is running.
- Supported Red Hat operator.
- Ease of installation process by way of the OperatorHub.

OADP includes the following basic components:

- Velero (backs up Kubernetes objects and volumes)
- Restic (DaemonSet for file backup)
- Red Hat OpenShift Velero B/R plug-ins
- Volume Snapshot plug-ins and CSI plug-in

Cloud Pak for Data B/R utilities

The following utilities are included in Cloud Pak for Data B/R:

- Online B/R: Install the Cloud Pak for Data OADP B/R utility.
- Offline B/R: Install one of the following utilities:
  - For B/R of Kubernetes metadata and persistent volumes in a Cloud Pak for Data instance namespace, install the OADP B/R utility.
  - To B/R only volume data, install the Cloud Pak for Data volume B/R utility (cpdbr).
- Cloud Pak for Data OADP backup REST service: You can run the Cloud Pak for Data OADP B/R utility in Representational State Transfer (REST) mode.

Online B/R

Complete the following steps to perform an online backup:

1. Install and configure the OADP B/R utility:
   a. Log in as user with cluster-admin privileges.
   b. In the Red Hat OpenShift Container Platform web console, click **Operators** → **OperatorHub**.
   c. Use the Filter by keyword field to find the OADP Operator.
   d. Select the **OADP Operator** and click **Install**.
   e. Click **Install** to install the Operator in the `openshift-adp` project.
   f. Click **Operators** → **Installed Operators** to verify the installation.
   g. Obtain the **`cpd-cli`** downloadable and install (see Example 8-4).

Example 8-4 Installing the cpd-cli downloadable

```
wget
https://github.com/IBM/cpd-cli/releases/download/v11.0.0/cpd-cli-linux-EE-11.0.0.t
gz
tar xvf cpd-cli-linux-EE-11.0.0.tgz
cd cpd-cli-linux-EE-11.0.0-20
chmod +x cpd-cli
```

2. Access to an S3-compatible object storage is needed. MinIO is used in the example. You also can choose AWS S3, IBM Cloud Object Storage, or Ceph Object Gateway. Complete the following steps:
   a. In the Red Hat OpenShift Container Platform web console, click **Operators** → **OperatorHub**.
   b. Use the Filter by keyword field to find the MinIO Operator.
   c. Select the **MinIO Operator** (IBM provided) and click **Install**.
   d. Click **Install** to install the Operator in the `openshift-adp` recommended project.
   e. Click **Operators** → **Installed Operators** to verify the installation.

3. The **cpd-operators.sh** requires the jq JSON command-line utility; therefore, run the command that is shown in Example 8-5.

Example 8-5 jq JSON command-line utility

```
jq --version
wget -O jq https://github.com/stedolan/jq/releases/download/jq-1.6/jq-linux64
chmod +x ./jq
cp jq /usr/bin
```

4. Configure the client to set the OADP operator namespace and CPD (control plane) namespace, respectively (see Example 8-6).

Example 8-6 Configuring the client to set the OADP operator namespace and CPD

```
cpd-cli oadp client config set namespace=oadp-operator
cpd-cli oadp client config set cpd-namespace=zen
```

5. Create a checkpoint by using the following command:

```
cpd-cli oadp checkpoint create --include-namespaces=zen --log-level=debug
--verbose
```

6. Create a backup of Cloud Pak for Data volumes and provide an ID for the backup (see Example 8-7).

Example 8-7 Creating a backup of Cloud Pak for Data volumes, and provide an ID for the backup

```
cpd-cli oadp backup create <ckpt-backup-id1> \
--include-namespaces ${PROJECT_CPD_INSTANCE} \
--hook-kind=checkpoint \
--include-resources='ns,pvc,pv,volumesnapshot,volumesnapshotcontent' \
--selector='icpdsupport/empty-on-nd-backup notin
(true),icpdsupport/ignore-on-nd-backup notin (true)' \
--snapshot-volumes \
--log-level=debug --verbose
```

7. Create a backup of Kubernetes resources and provide an ID for the backup (see Example 8-8).

Example 8-8 Creating a backup of Kubernetes resources and providing an ID for the backup

```
cpd-cli oadp backup create <ckpt-backup-id2> \
--include-namespaces ${PROJECT_CPD_INSTANCE} \
--hook-kind=checkpoint \
--exclude-resources='pod,event,event.events.k8s.io' \
--selector='icpdsupport/ignore-on-nd-backup notin (true)' \
--snapshot-volumes=false \
--skip-hooks=true \
--log-level=debug \
--verbose
```

8. Check the status of the backup by using the following command:

```
cpd-cli oadp backup status --details <backup_name>
```

9. Check that snapshots are created and READYTOUSE is true by using the following command:

```
oc get volumesnapshots -n ${PROJECT_CPD_INSTANCE}
```

Performing an online restore

Restoring a Cloud Pak for Data instance from an online backup on the same cluster involves the following steps:

- Delete the instance of Cloud Pak for Data on the cluster.
- Restore the instance of Cloud Pak for Data on the cluster.

Complete the following steps to delete the instance of Cloud Pak for Data on the cluster:

1. Log in to Red Hat OpenShift Container Platform as a user with sufficient permissions to complete the task:

   ```
   oc login OpenShift_URL:port
   ```

2. If the Cloud Pak for Data instance was configured with `iamintegration: true`, delete clients in Cloud Pak for Data projects:

   ```
   oc delete client -n ${PROJECT_CPD_INSTANCE} --all
   ```

3. Delete service-specific finalizers from service custom resources (CRs), as shown in Example 8-9.

Example 8-9 Deleting service-specific finalizers from service custom resources

```
# Get CCS CR, delete finalizers and delete CR
oc get ccs -n ${PROJECT_CPD_INSTANCE}
oc patch ccs <ccs_name> -n ${PROJECT_CPD_INSTANCE} -p
'{"metadata":{"finalizers":[]}}' --type=merge
oc delete ccs -n ${PROJECT_CPD_INSTANCE} --all
```

4. Delete finalizers from Zen-Service Operand Request (see Example 8-10).

Example 8-10 Delete finalizers from Zen-Service Operand Request

```
oc get operandrequest -n ${PROJECT_CPD_INSTANCE}
oc patch operandrequest <zen_service_operandrequest_name> -n
${PROJECT_CPD_INSTANCE} -p '{"metadata":{"finalizers":[]}}' --type=merge
```

5. Delete OperandRequests and ZenServices (see Example 8-11).

Example 8-11 Delete OperandRequests and ZenServices

```
oc delete operandrequests -n ${PROJECT_CPD_INSTANCE} --all
oc delete zenservice -n ${PROJECT_CPD_INSTANCE} --all
```

6. Delete finalizers from the admin RoleBinding (see Example 8-12).

Example 8-12 Delete finalizers from the admin RoleBinding

```
oc patch rolebinding admin -n ${PROJECT_CPD_INSTANCE} -p
'{"metadata":{"finalizers":[]}}' --type=merge
```

7. Delete the Cloud Pak for Data instance projects (see Example 8-13).

Example 8-13 Delete the Cloud Pak for Data instance projects

```
oc delete project ${PROJECT_CPD_INSTANCE}
oc get project ${PROJECT_CPD_INSTANCE} -o jsonpath="{.status}"
```

8. If finalizers remain, repeat these substeps to locate resources and delete the finalizers.

Complete the following to restore the instance of Cloud Pak for Data on the same cluster:

1. Check that the backup is available and that it completed with no errors (see Example 8-14).

Example 8-14 Checking the backup

```
cpd-cli oadp backup ls
NAME           STATUS      ERRORS  WARNINGS  CREATED     EXPIRES  STORAGE LOCATION  SELECTOR
<backup_name>  Completed     0        0     <timestamp>   358d       default         <none>
```

2. Restore volume data from the online backup by entering the volume backup ID and specifying a restore ID (see Example 8-15).

Example 8-15 Restoring volume data from the online backup

```
cpd-cli oadp restore create --from-backup=<ckpt-backup-id1> <chkpt-restore-id1>
--skip-hooks=true --log-level=debug --verbose
```

3. Restore resources, except pod-generating resources and operandrequests (see Example 8-16).

Example 8-16 Restoring resources

```
cpd-cli oadp restore create --from-backup=<ckpt-backup-id2>
--exclude-resources='clients,ImageTag,deploy,rs,dc,rc,sts,ds,cj,jobs,controllerrevisions,po,opre
q' <chkpt-restore-id2> --include-cluster-resources=true --skip-hooks --log-level=debug --verbose
```

4. Restore pod-generating resources (see Example 8-17).

Example 8-17 Restoring pod generating resources

```
cpd-cli oadp restore create --from-backup=<ckpt-backup-id2>
--include-resources='deploy,rs,dc,rc,sts,ds,cj,jobs,controllerrevisions' <chkpt-restore-id3>
--preworkloadhooks=true --posthooks=true --log-level=debug --verbose
```

4. Restore operandrequests (see Example 8-18).

Example 8-18 Restore operandrequests

```
cpd-cli oadp restore create --from-backup=<ckpt-backup-id2> --include-resources='opreq'
<chkpt-restore-id4> --skip-hooks --log-level=debug --verbose
```

5. To check the status of a restore, run the command that is shown in Example 8-19.

Example 8-19 Check the status of a restore

```
cpd-cli oadp restore status --details <restore_name>
cpd-cli oadp restore ls
cpd-cli oadp restore logs <restore_name>
```

Creating an offline B/R

You can create an offline backup of an entire Cloud Pak for Data deployment by using the Red Hat OpenShift APIs for Data Protection (OADP) B/R utility.

You can perform a back restore of B/R the Cloud Pak for Data control plane and services by using one of the following methods:

- If Cloud Pak for Data is installed on Ceph Container Storage Interface (CSI) volumes, create volume snapshots. Snapshots are typically much faster than file copying, by using copy-on-write techniques to save changes instead of doing a full copy.
- Create Restic backups on an S3-compatible object store.

Offline B/R includes the following features:

- Restore to the same cluster or to a different cluster
- B/R of multiple projects (namespaces)

The Cloud Pak for Data volume B/R utility can perform backup and restore of a file system by using one of the following methods:

- If you use Portworx storage, you can create volume snapshots. Portworx snapshots are atomic, point-in-time snapshots.
- You can create volume backups on a separate PVC or on an S3-compatible object store. Volume backups work with any storage type.

  Use this method if you have Cloud Pak for Data services that use different storage types, such as NFS (configured with `no_root_squash`), Portworx, and Red Hat OpenShift Data Foundation.

The Cloud Pak for Data volume B/R utility supports only offline volume B/R. The utility does not provide application-level B/R that re-creates your Kubernetes resources, such as configuration maps, secrets, PVCs, PVs, pods, deployments, and StatefulSets.

A typical use case is backing up and restoring all volumes in the same Red Hat OpenShift project, if the same Kubernetes objects still exist. For some Cloud Pak for Data services, you must run scripts before and after you run B/R operations.

A note on nondisruptive backup: Nondisruptive backup is difficult to achieve. Consider the following points:

- Depending on the storage type and methods that are used for creating storage layer snapshots, the creation of volume snapshots for “in-use” PVs can cause a temporary freeze of its associated storage data volumes (to preserve the volume's data consistency during the snapshot), which causes a temporary suspension of the service’s IO write operations to the data volumes.
- Service state checkpointing methods and crash recovery implementation depend on the application.
- Consistent crash recovery is difficult to test and to trust such validation because the crash might occur under various external and internal conditions (including internal states of the application, and so on).
- Crash recovery RPO depends on service-specific internal implementation. For each service, we must ensure that its recovery point is within an externally required RPO time. This issue varies for the same service under different customer cases.
- Multiple B/R and DR products and external solutions have different B/R “consistency” interfaces (pre-/post- scripts, and so on) or do not provide any such interfaces at all, while expecting from a service to provide reliable crash recovery from any crash event.
- Many data stores (such as EDB, Elasticsearch, and Minio) do not support live file system-based storage snapshots. Rather, they require the use of snapshotting APIs that are specific to that component. In this case, the nondisruptive backup is supported only for data stores, which snapshotting APIs can be called when the data store is online without causing data store downtime.
- A service can include multiple pods and use multiple persistent volumes. Therefore, during the nondisruptive backup process, the storage layer can create its snapshots of the dependent persistent volumes at slightly different times (because volume group backup is not generally supported by most storage layer types). In another scenario in which a service’s components crashed (potentially not all at the same time), its persistent data also might be left in non-consistent states within each volume and across the volumes.

8.4 Day 2+ Operations: Observability

Another facet of Day 2+ Ops is the concept of observability. Observability is more than just monitoring. Monitoring tells when you when something is wrong; observability helps you understand why.

Observability is the cornerstone of robust IT Operations and proactive incident management and remediation. It enables the IT Ops or SRE team to gain insight into not only the system processes, but also develop an understanding of the causal event that provokes an incident or failure. This process enables IT teams to help remediate the issue while also preempting future similar issues. Observability is a key component of intelligent Day 2+ Ops that is required for any production system.

After IBM Cloud Pak for Data is deployed and running on Red Hat OpenShift, Day 2 configurations on observability are important to ensure the application remains in a running important to ensure the application remains in a running state, and potential issues can be identified as early as possible.

Red Hat OpenShift already provides an observability framework that IBM Cloud Pak for Data can use. This framework enables a centralized collection of monitoring and logging assets, which makes it easier to forward these assets to external applications and to create dashboards.

Because of this dependency, it is important to start at the Red Hat OpenShift level with the following components to configure observability on IBM Cloud Pak for Data:

- Red Hat OpenShift Cluster Auditing
- Red Hat OpenShift Cluster Logging
- Red Hat OpenShift Cluster Monitoring

After the following components are configured, IBM Cloud Pak for Data can make use of it, if needed:

- IBM Cloud Pak for Data Auditing
- IBM Cloud Pak for Data Logging
- IBM Cloud Pak for Data Monitoring
- IBM Cloud Pak for Data Notifications

8.4.1 Cloud Pak for Data: Monitoring

IBM Cloud Pak for Data provides a monitoring and alerting framework that with which you can monitor the state of the platform and set up events to alert when action is needed based on a thresholds configuration. This framework is implemented by the zen-watchdog, which is part of the Cloud Pak for Data Control plane.

Monitoring metrics are exposed as a Prometheus endpoint, which allows for these metrics to be used by the Red Hat OpenShift Monitoring framework.

Default set of monitors

When IBM Cloud Pak for Data is installed, a default set of monitors is deployed (see Table 8-1). These monitors focus on the deployment and availability of the Cloud Pak for Data components.

Table 8-1 Default set of monitors

Monitor type	Description
Deployment Status check	For each Deployment part of the IBM Cloud Pak for Data installation, a monitor of type Deployment status check is created. This check confirms that each Deployment has the correct number of replicas available.
Statefulset status check	For each StatefulSet part of the IBM Cloud Pak for Data installation, a monitor of type Statefulset status check is created. This check confirms that each StatefulSet has the correct number of replicas available.
PVC Status check	A persistent volume claim (PVC) is a request for storage that meets specific criteria, such as a minimum size or a specific access mode. This monitor checks the state of the PVC and whether it has run out of available storage.
Quota Status check	An administrator set a vCPU quota and a memory quota for the service or for the platform. A critical state indicates that the service has insufficient resources to fulfill requests. The service cannot create pods if the new pods push the service over the memory quota or the vCPU quota. These pods remain in a pending state until sufficient resources are available.

Monitor type	Description
Service status check	A service (for example IBM Watson Studio or IBM Watson Machine Learning) consists of pods and one or more service instances. The state of the service depends on the state of these pods and service instances. A critical state indicates that a service instance is in a failed state or a pod is in a failed or unknown state.
Service instance status check	A service instance (for example, a Cognos Analytics or DataStage instance) consists of one or more pods. The state of the service instance depends on the state of these pods. A critical state indicates that one or more pods that are associated with the instance are in a failed or unknown state.

Registering and executing monitors

Cloud Pak for Data includes an alerting framework (called zen-watchdog) in which standard monitors are bundled. Custom monitors can be added by using an API. The zen-watchdog cron job updates its configuration every 10 minutes, which enables new monitors to be registered to the framework.

All default monitors are run in a single cron job (the Diagnostics cron job). This job collects all metrics part of the default set of monitors and sends these metrics to the zen-watchdog, where they are stored in the influxdb metastore (see Figure 8-31).

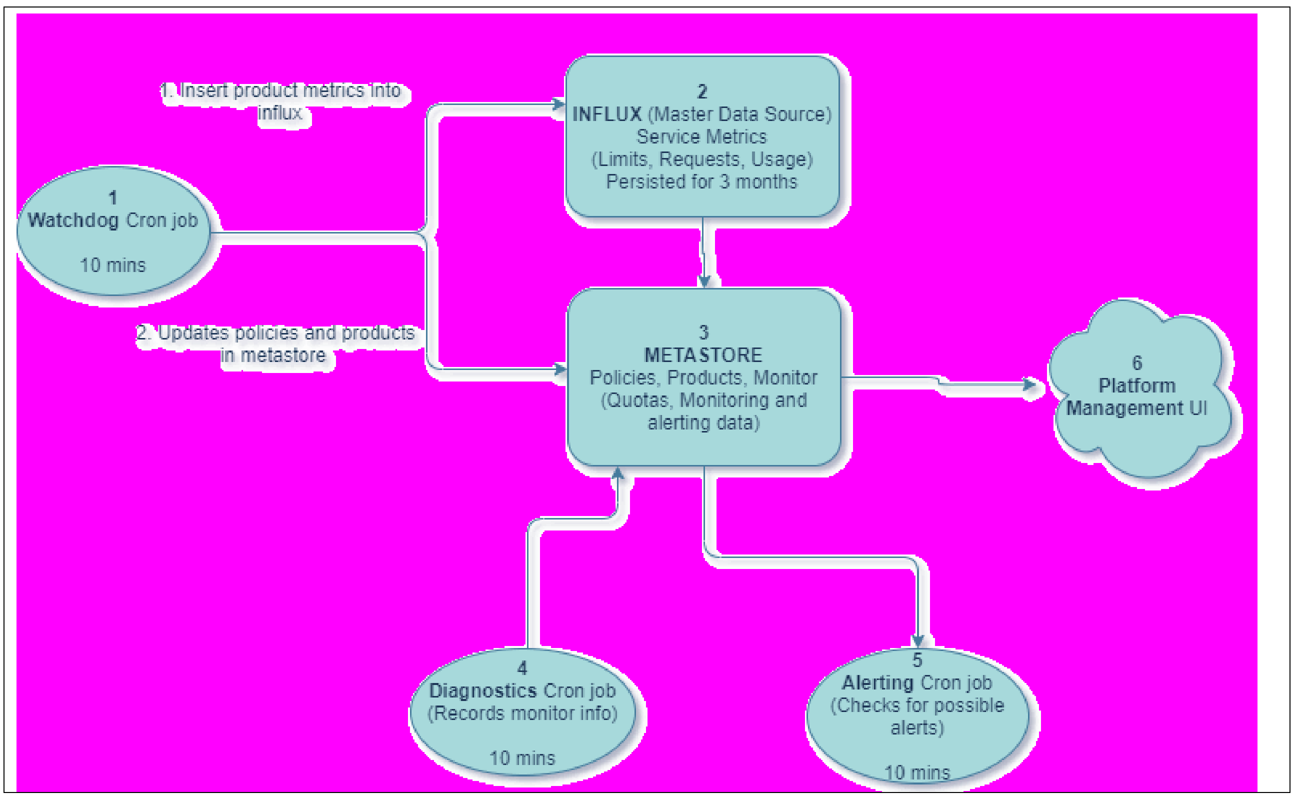

Figure 8-31 Diagnostics cron job

Monitor execution and alert triggering

You can monitor the execution of the monitors. The results can be found in the IBM Cloud Pak for Data web console by clicking **Administration** → **Monitoring** and then, selecting the **Events** page (View all Events and Alerts).

Cloud Pak for Data Prometheus data endpoint

The Prometheus endpoint data can be accessed at `https://<CP4D_URL>/zen/metrics` (a valid login session to IBM Cloud Pak for Data is required).

To fetch the metrics by using a script, a bearer token must be created first to collect the metrics (see Example 8-20).

Example 8-20 Fetching the metrics by using a script

```
export CP4D_URL=https://<CP4D_URL>
export CP4D_USERNAME=admin
export CP4D_PASSWORD=*********
export CP4D_BEARER_TOKEN=$(curl -k -X POST ${CP4D_URL}/icp4d-api/v1/authorize -H 'cache-control: no-cache' -H 'content-type: application/json' -d
"{\"username\":\"${CP4D_USERNAME}\",\"password\":\"${CP4D_PASSWORD}\"}" | jq -r '. | .token')
curl -v -k -H "Authorization: Bearer ${CP4D_BEARER_TOKEN}" ${CP4D_URL}/zen/metrics
# HELP available
# TYPE available gauge
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="asset-files-api"} 1
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="ax-environments-api-deploy"} 1
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="ax-environments-ui-deploy"} 1
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="catalog-api"} 2
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="dap-dashboards-api"} 1
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="datastage-ibm-datastage-apidoc"} 1
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="datastage-ibm-datastage-assets"} 1
available{event_type="check-deployment-status",monitor_type="diagnostics",reference="datastage-ibm-datastage-canvas"} 1
```

Alerting rules

The Cloud Pak for Data monitors report their status as monitor events. When the monitor is run, it generates a monitor event, which summarizes the state of the objects it monitors.

> **Note:** Each monitor can have a different schedule when it runs; for example, every 10 minutes, 30 minutes, or 2 hours).

These events can be informational (no issues), warning (potential issue), or critical (immediate attention needed). For warning and critical events, alerts can be triggered to specify when to forward a specific alert to the user. These alerts are configured by using alerting rules.

The following default alerting rules are set:

- For critical events, a condition persists for 30 minutes when three consecutive critical events are recorded during monitor runs. When the condition is alerted, it is snoozed for 12 hours.
- For warning events, 5 warning events are recorded during the last 20 monitor runs with a snooze period of 24 hours.

For more information about creating alerting rules by using the API endpoint, see this IBM Documentation web page.

Building a custom monitor

You can add custom monitors by using the IBM Cloud Pak for Data zen-watchdog framework. Custom monitors become part of the zen-watchdog configuration and are displayed in the Events and Alert page.

In addition, the metrics for these custom monitors are added to the Prometheus data endpoint. Custom monitors are built by using a custom image and running the monitor code. The custom monitor code posts its events to the zen-watchdog API endpoint, so they are processed.

For more information about an example of building a custom monitor, see the following resources:

- This IBM Documentation web page
- This GitHub web page

This repository contains a set of functional monitors including detailed information about how to deploy and integrate the custom monitors, update the monitors if changes are applied to the source, and how to reset the Cloud Pak for Data metrics configuration, and InfluxDB if an issue corrupted the monitor events.

8.4.2 Cloud Pak for Data: Logging

In IBM Cloud Pak for Data, each pod generates logs when actions are performed. If an error occurs, the logs contain information about the error, which can be used to find a solution to resolve the error.

To capture the log of a pod, the following command can be used:

```
export CP4D_PROJECT=zen
oc get logs <podname> -n ${CP4D_PROJECT}
```

If the pod contains multiple containers, use the `-c` flag to specify for which container to fetch the logs:

```
export CP4D_PROJECT=zen
oc get logs <podname> -c <containername> -n ${CP4D_PROJECT}
```

This method of collecting logs can be used for isolated incidents in which the logs are requests.

For a more structural solution, it is recommended to collect the logs of all IBM Cloud Pak for Data pods and forward them to the Red Hat OpenShift Logging framework. This process is implemented by creating a `ClusterLogging` instance to set up Red Hat OpenShift Logging and then, creating a `ClusterLogForwarder` for each instance of Cloud Pak for Data. All logs of the project in which IBM Cloud Pak for Data is deployed are then forwarded.

From there. the logs can be forwarded to an external Log collector in which it can be stored and analyzed.

For more information about configuring Red Hat OpenShift Logging and creating a ClusterLogForwarder for IBM Cloud Pak for Data, see this Red Hat OpenShift Documentation web page.

Alternative logging method

Logs also can be obtained by way of the Cloud Pak for Data interface by completing the following steps:

1. Click **Administration** → **Monitoring**.
2. Find the individual pods, or the pods that belong to a service or instance.
3. View and search the logs.

8.4.3 Cloud Pak for Data: Auditing

IBM Cloud Pak for Data generates audit events that can be used to monitor the use of the platform and identify potential misuse of the platform. For example, many failed login attempts by using the admin user can indicate a potential attempt to access the system with Administrative permissions.

The types of audit events that are generated by IBM Cloud Pak for Data depends on which cartridges are installed:

- IBM Cloud Pak for Data platform events
- IBM Watson Knowledge Catalog events

For more information about sample events that are generated by IBM Cloud Pak for Data, see this IBM Documentation web page.

Exporting IBM Cloud Pak for Data audit messages

IBM Cloud Pak for Data audit messages can be forwarded to a Security Information and Event Management (SIEM) solution, including Splunk, LogDNA, and IBM QRadar®. The auditing framework is deployed by default; therefore, only the forwarders must be configured. You can connect each instance of Cloud Pak for Data to one or more SIEM systems.

The Audit Logging Service uses Fluentd output plug-ins to forward and export audit records. When you enable the Audit Logging Service, you specify the external SIEM system to which you want to forward the audit records.

The Audit Logging Service explicitly supports the following SIEM solutions:

- Splunk
- LogDNA
- QRadar

Forwarding audit logs to stdout

It is possible to forward the audit records to the stdout of the zen-audit pod of Cloud Pak for Data for debugging purposes. Although this process is not recommended for production environments, it can be used to validate and test the audit messages.

Forwarding audit logs to stdout also is used in combination with the Red Hat OpenShift ClusterLogForwarder to send audit logs to an external log store by using one of the following methods:

- Editing the zen-audit-config ConfigMap
- Creating a ConfigMap and patching the changes to the Cloud Pak for Data zenService instance lite-cr

Editing zen-audit-config ConfigMap

Complete the following steps:

1. Create the ConfigMap for the auditing configuration:

```
export CP4D_PROJECT=zen
oc edit configmap zen-audit-config -n ${CP4D_PROJECT}
```

2. Add the following `<store>` entry to the ConfigMap and save it as shown in Example 8-21.

Example 8-21 Editing zen-audit-config ConfigMap

```
apiVersion: v1
metadata:
  name: zen-audit-config
data:
  fluent.conf: |-
    <match export export.** records records.** syslog syslog.**>
      @type copy
      <store>
        @type stdout
      </store>
    </match>
In order for changes to be put if effect, the zen-audit pods need to be restarted
export CP4D_PROJECT=zen
oc delete po -l component=zen-audit -n ${CP4D_PROJECT}
```

3. After the pod is restarted, it contains the Cloud Pak for Data audit events when accessing the logs of the zen-adit pod.

Creating a ConfigMap

Complete the following steps to create a ConfigMap:

1. Create a ConfigMap by using the auditing configuration (see Example 8-22).

Example 8-22 Creating ConfigMap

```
export CP4D_PROJECT=zen
export AUDIT_CM_NAME=my-custom-audit-config

cat <<EOF | oc apply -f -
apiVersion: v1
kind: ConfigMap
metadata:
  name: ${AUDIT_CM_NAME}
  namespace: ${CP4D_PROJECT}
data:
  fluent.conf: |-
    <match export export.** records records.** syslog syslog.**>
      @type copy
      <store>
```

```
        @type stdout
      </store>
    </match>
EOF

oc patch zenService lite-cr \
  -n ${CP4D_PROJECT} \
  --type="json" \
  --patch '[{"op": "replace", "path":"/spec/zen_audit_config_custom", "value":
'${AUDIT_CM_NAME}' }]'
```

2. Restart the zen-audit pods so that the changes are put into effect:

```
export CP4D_PROJECT=zen
oc delete po -l component=zen-audit -n ${CP4D_PROJECT}
```

3. After the pod is restarted, it contains the Cloud Pak for Data audit events when the logs of the zen-adit pod are accessed.

Forwarding audit logs to Splunk

Cloud Pak for Data uses the Splunk HTTP Event Collector Output plug-in to integrate with Splunk. To export the audit records to Splunk, configure a new HTTP event collector in Splunk for Cloud Pak for Data. Then, store the token, which is used in the Splunk forward configuration in Cloud Pak for Data.

Complete the following steps to forward audit logs to Splunk:

1. Edit zen-audit-config ConfigMap by using the following commands:

```
export CP4D_PROJECT=zen
oc edit configmap zen-audit-config -n ${CP4D_PROJECT}
```

2. Add the following `<store>` entry to the ConfigMap and save it. This example uses the recommended `use_ssl true` value. For non-production environments, this value can be set to `false`, where the `ca.pem section` and the `ca_file property` can be omitted (see Example 8-23 on page 626).

Example 8-23 Editing the ConfigMap

```
apiVersion: v1
metadata:
  name: zen-audit-config
data:
  ca.pem: |
    -----BEGIN CERTIFICATE-----
     <XXXXXXXXX>
    -----END CERTIFICATE-----
  fluent.conf: |-
    <match export export.** records records.** syslog syslog.**>
      @type copy
      <store>
        @type splunk_hec
        host <SPLUNK-HOST>    # Replace <SPLUNK-HOST> with the address of the
Splunk host
        port <SPLUNK-PORT>    # Replace <SPLUNK-PORT>. The default port is 8088
        token <SPLUNK-TOKEN>  # Replace <SPLUNK-TOKEN> with the token you
generated
        flush_interval 10s                    # Recommended value
        use_ssl true                          # Required to use TLS
```

```
        ca_file /fluentd/config/ca.pem       # Required to use TLS; specify the
cert in the ca.pem section
      </store>
    </match>
```

3. Restart the zen-audit pods so that the changes are put into effect:

   ```
   export CP4D_PROJECT=zen
   oc delete po -l component=zen-audit -n ${CP4D_PROJECT}
   ```

4. After the pod is restarted, confirm in Splunk that IBM Cloud Pak for Data audit events are being processed.

Creating a ConfigMap

Complete the following steps:

1. Create a ConfigMap. The example that is shown in Example 8-24 uses the recommended `use_ssl true` value. For non-production environments, this value can be set to `false`, where the `ca.pem` section and the `ca_file property` can be omitted (see Example 8-24).

Example 8-24 Creating ConfigMap

```
export CP4D_PROJECT=zen
export AUDIT_CM_NAME=my-custom-audit-config
export TLS_CERT=<certificate>
export SPLUNK_HOST=<hostname>
export SPLUNK_PORT=<port>
export SPLUNK_TOKEN=<token>

cat <<EOF | oc apply -f -
apiVersion: v1
kind: ConfigMap
metadata:
  name: ${AUDIT_CM_NAME}
  namespace: ${CP4D_PROJECT}
data:
  ca.pem: |
    -----BEGIN CERTIFICATE-----
     ${TLS_CERT}
    -----END CERTIFICATE-----
  fluent.conf: |-
    <match export export.** records records.** syslog syslog.**>
      @type copy
      <store>
        @type splunk_hec
        host ${SPLUNK-HOST}
        port ${SPLUNK-PORT}
        token ${SPLUNK-TOKEN}
        flush_interval 10s                   # Recommended value
        use_ssl true                         # Required to use TLS.
        ca_file /fluentd/config/ca.pem       # Required to use TLS.
      </store>
    </match>
EOF

oc patch zenService lite-cr \
  -n ${CP4D_PROJECT} \
  --type="json" \
```

```
  --patch '[{"op": "replace", "path":"/spec/zen_audit_config_custom", "value": 
'${AUDIT_CM_NAME}' }]'
```

2. Restart the zen-audit pods so that the changes are put into effect:

   ```
   export CP4D_PROJECT=zen
   oc delete po -l component=zen-audit -n ${CP4D_PROJECT}
   ```

3. After the pod is restarted, confirm in Splunk that IBM Cloud Pak for Data audit events are being processed.

8.4.4 Cloud Pak for Data: Notifications (SMTP)

By configuring an SMTP connection, IBM Cloud Pak for Data can send email notifications to users. To send these email notifications, the IBM Watson Studio or IBM Watson Knowledge Catalog must be installed.

Email notifications can be enabled by way of the web client of IBM Cloud Pak for Data. Browse to **Administer** → **Configure Platform**. On the SMTP settings page, specify the SMTP connection information.

8.4.5 IBM Foundational Services: License Service

The IBM Foundational Services: License Service collects and measures the license use of IBM products at the cluster level. This data can be retrieved upon request for monitoring and compliance.

License Service overview

The IBM Foundational Services: License Service is required for monitoring and measuring license use of the IBM Cloud Paks and IBM stand-alone containerized software in accord with the pricing rule for containerized environments. Manual license measurements are not allowed.

The License Service of the IBM Foundational Services provides the following capabilities:

- Collects and measures the license use of Virtual Processor Core (VPC) metric at the cluster level.
- Collects and measures the license use of IBM Cloud Paks and their bundled products that are enabled for reporting and licensed with the Managed Virtual Server (MVS) license metric.
- Collects and measures the license use of Virtual Processor Core (VPC) and Processor Value Unit (PVU) metrics at the cluster level of IBM stand-alone containerized software that is deployed on a cluster and is enabled for reporting.

As of this writing, License Service refreshes the data every 5 minutes. With this frequency, you can capture changes in a dynamic cloud environment. License Service stores the historical licensing data for the last 24 months. However, the frequency and the retention period might be subject to change in the future.

License Service includes the following features:

- Provides the API that you can use to retrieve data that outlines the highest license usage on the cluster.
- Provides the API that you can use to retrieve an audit snapshot that lists the highest license use values for the requested period for products that are deployed on a cluster.
- Supports hyperthreading on worker nodes, which also is referred to as Simultaneous multi-threading (SMT) or Hyperthreading (HT).

> **Note:** Only one instance of the License Service is deployed per cluster, regardless of the number of IBM Cloud Paks and containerized products that are installed on this cluster.

Installing the License Service

The License Service is part of the IBM Foundational Services. The License Service can be deployed by creating an `OperandRequest`, which handle the deployment (see Example 8-25).

Example 8-25 Installing the License Service

```
export FS_PROJECT=ibm-common-services
--
cat << EOF | oc apply -f -
apiVersion: operator.ibm.com/v1alpha1
kind: OperandRequest
metadata:
  name: common-service
  namespace: ${FS_PROJECT}
spec:
  requests:
  - operands:
    - name: ibm-licensing-operator
    registry: common-service
EOF
--
```

After the Deployment is completed, a route is created that can be used to start the License Service API calls.

Obtaining the License Service API token

When installed, the License Service API can be accessed by using an API token set as a parameter in the URL. Optionally, this token can be disabled and Service Accounts used instead, which is more secure method. (The API token is not part of the URL requesting the information.)

After the License Service is installed, a secret is available, which is named `ibm-licensing-token`, which contains the token value:

```
export FS_PROJECT=ibm-common-services
export FS_API_TOKEN=$(oc get secret ibm-licensing-token -n ${FS_PROJECT} -o
jsonpath={.data.token} | base64 -d)
```

Using the License Service API endpoint

The Foundational Services License Service provides the API endpoints that are listed in Table 8-2.

Table 8-2 API endpoints

API description	API URL
Retrieving an audit snapshot	< License Service URL >/snapshot?token=< token >
Retrieving license usage of products	< License Service URL >/products?token=< token >
Retrieving license usage of bundled products	< License Service URL >/bundled_products?token=< token >
Retrieving contribution of services	< License Service URL >/services?token=< token >
Retrieving information about License Service health	< License Service URL >/health?token=< token >
Obtaining the status page (HTML page)	< License Service URL >/status?token=< token >
Version information	< License Service URL >/version

Examples of the use of the License Service API endpoint

The examples in this section use the `FS_LIC_SERVER_URL`, which is set to the License Service URL. This URL is read from the route that is named `ibm-licensing-service-instance` from the IBM Foundational Services project (typically, `ibm-common-services`), as shown in Example 8-26.

Example 8-26 ibm-licensing-service-instance route

```
export FS_PROJECT=ibm-common-services
export FS_LIC_SERVER_URL=https://$(oc get route ibm-licensing-service-instance -n
${FS_PROJECT} -o jsonpath={.spec.host})
```

For the `< token >`, the `FS_API_TOKEN` is used, which was defined as described in “Obtaining the License Service API token” on page 629.

Acquiring License Service audit snapshot

The API that is shown in Example 8-27, retrieves a snapshot of the auditing information.

Example 8-27 Acquire License Service audit snapshot

```
curl -k ${FS_LIC_SERVER_URL}/snapshot?token=${FS_API_TOKEN} --output output.zip

unzip -v output.zip

Archive:  output.zip
 Length   Method    Size  Cmpr    Date    Time   CRC-32   Name
--------  ------  ------- ---- ---------- ----- --------   ----
     317  Defl:N      183  42% 08-24-2022 11:04 d009eb4e
bundled_products_2022-07-25_2022-08-24_10.213.0.132.csv
     144  Defl:N      131   9% 08-24-2022 11:04 8b4f49c1
products_2022-07-25_2022-08-24_10.213.0.132.csv
     368  Defl:N      194  47% 08-24-2022 11:04 4999faf8
bundled_products_daily_2022-07-25_2022-08-24_10.213.0.132.csv
```

```
     157  Defl:N      144   8% 08-24-2022 11:04 869a2bb5
products_daily_2022-07-25_2022-08-24_10.213.0.132.csv
     386  Defl:N      244  37% 08-24-2022 11:04 b65c375e  data_condition.txt
       0  Defl:N        2   0% 08-24-2022 11:04 00000000
unrecognized-apps-2022-07-25-2022-08-24.csv
     256  Defl:N      261  -2% 08-24-2022 11:04 c3557500  signature.rsa
     445  Defl:N      359  19% 08-24-2022 11:04 b86a7776  pub_key.pem
     678  Defl:N      348  49% 08-24-2022 11:04 5a665342  checksum.txt
--------          -------  ---                            -------
    2751             1866  32%                            9 files
```

The `output.zip` file is created, which contains the snapshot information. The various CSV files that are output show use statistics from the various services. For example, `bundled_products_daily_2022-07-25_2022-08-24_10.213.0.132.csv` shows the use stats for the various bundled products in IBM Cloud Pak for Data.

Acquiring License Service version information

The API that is shown in Example 8-28 returns the license service version.

Example 8-28 Acquire License Service version information

```
curl -k ${FS_LIC_SERVER_URL}/version

{
  "version": "1.16.0",
  "buildDate": "Thu Jun 2 13:21:47 UTC 2022",
  "commit": "69604a3"
}
```

Acquiring License Service license usage of products

The API that is shown in Example 8-29 returns the license usage of all products.

Example 8-29 Acquire License Service license usage of products

```
curl -k ${FS_LIC_SERVER_URL}/products?token=${FS_API_TOKEN}

[
  {
    "name": "IBM Cloud Pak for Data",
    "id": "eb9998dcc5d24e3eb5b6fb488f750fe2",
    "metricPeakDate": "2022-08-24",
    "metricName": "VIRTUAL_PROCESSOR_CORE",
    "metricQuantity": 24
  }
]
```

Acquiring License Service license usage of bundled products

The API that is shown in Example 8-30 returns the license use for all bundled products.

Example 8-30 Acquire License Service license usage of bundled products

```
curl -k ${FS_LIC_SERVER_URL}/bundled_products?token=${FS_API_TOKEN}

[
  {
    "productName": "IBM Cloud Pak for Data",
    "productId": "eb9998dcc5d24e3eb5b6fb488f750fe2",
    "cloudpakId": "eb9998dcc5d24e3eb5b6fb488f750fe2",
    "cloudpakMetricName": "VIRTUAL_PROCESSOR_CORE",
    "metricName": "VIRTUAL_PROCESSOR_CORE",
    "metricPeakDate": "2022-08-24",
    "metricMeasuredQuantity": 24,
    "metricConversion": "1:1",
    "metricConvertedQuantity": 24
  }
]
```

Acquiring License Service information about License Service health

Example 8-31 shows how to check whether the license service is healthy by using a REST call.

Example 8-31 Acquire License Service information about License Service health

```
curl -k ${FS_LIC_SERVER_URL}/health?token=${FS_API_TOKEN}

{
  "incompleteAnnotations": {
    "count": 0,
    "pods": []
  }
}
```

8.5 Security Operations

Security Operations (SecOps) is the implementation of stringent security requirements in line with the operation of a production system. When we talk about SecOps, we focus primarily on implementing a Zero Trust framework alongside Cloud Pak for Data.

In this section, we discuss how Zero Trust and SecOps ties in with ensuring your Cloud Pak for Data system is highly secure and protected against threat and vulnerabilities.

Security Operations in Cloud Pak for Data

When deploying Cloud Pak for Data on a Red Hat OpenShift cluster, the product inherits many security aspects of the cluster, such as the hiding compute nodes in a private network, firewalls, and proxies. In many cases, images that make up the product are pulled from a private container registry and vulnerability scans were performed before an image is stored in the registry.

However, not all security measures are controlled at the Red Hat OpenShift level. It remains important to consider other aspects that are specific to Cloud Pak for Data. Some of these aspects, such as identity and access management (IAM) and certificate life-cycle management, can be configured in the Cloud Pak Foundational Services (formerly known as IBM Common Services). Also, IAM configuration can be applied to Cloud Pak for Data instead of the use of the Foundational Services.

We identify the following categories when defining security measures for Cloud Pak for Data:

- Perimeter security: Protection from unauthorized access to the application by external users or applications and unauthorized access to the outside world by Cloud Pak for Data users and processes. Examples of perimeter security are firewalls, load balancers, and identity management, which also is referred to as *authentication*.
- Internal security: Protection from unauthorized access by internal applications. Access management to capabilities in Cloud Pak for Data and Red Hat OpenShift network policies to limit inter-namespace communications are examples of internal security.
- Ongoing security validation: Repeating processes to check that security measures meet the defined guardrails. Cloud Pak for Data microservices interact with each other by way of a trusted contract, which also manages encrypting traffic.

  The trusted contract is implemented by using TLS certificates that typically are recycled regularly, such as monthly, quarterly, or yearly. Also, because security vulnerabilities (common vulnerabilities and exposures or CVEs)[4] are published, organizations might want to reevaluate their security posture.

When planning the implementation of Cloud Pak for Data, it is important to start conversations with the organization's security team as early as possible. Changes, such as opening firewall ports or integration with LDAP or Identity Providers, often require several levels of approval and can take considerable time, sometimes even weeks.

8.5.1 Identity and Access Management (SSO)

In this section, we discuss Identity and Access Management.

Access Management

Access management to the platform is covered in a few layers. To access the user interface or any of the constituents (services) that are installed in the Cloud Pak for Data platform, users must be registered in the platform's user registry. This requirement also applies for API access to the platform or any underlying service.

Access control is based on the identity of a user (such as the username or email address) who is requesting to perform an operation on an object. For example, a user might request permission to create an analytics project within the interface. This request requires the user to create project permissions.

After access to the platform is attained and users can authenticate, the second layer of protection is the access to individual services. Services and service instances have their own registration of access that users or user groups also use.

[4] https://csrc.nist.gov/glossary/term/common_vulnerabilities_and_exposures

Role-Based Access Control

In Cloud Pak for Data, individual permissions are grouped into roles. Roles are defined to organize access controls that typically are used in combination with each other, such as the Data Scientist role, which includes permissions to create projects and work with deployment spaces.

Users are assigned a role, also known as Role-Based Access Control (RBAC), to control what they can do in Cloud Pak for Data. By assigning more than one role, permissions can be combined. In small organizations, it is usual that a member of the data science team has the Data Scientist and Data Engineer roles to allow them to perform more tasks.

Access in Cloud Pak for Data is solely controlled by using roles. After installing, the User, Administrator, Data Scientist, Data Quality Analyst and other roles are available to assign to users. The User role is the least-capable role and allows someone to create only deployment spaces and projects. The Data Scientist role extends this ability with accessing catalogs and governance artifacts. On the other side of the spectrum is the Administrator role, which holds a superset of all permissions by other roles, including platform administration management to set permissions.

Other roles with other combinations of permissions also can be defined.

User groups

In most organizations, access control is defined by using groups (for example: Finance, Human Resources, and Marketing). People are organized into groups, which defines their access to systems, applications, and what they can do in the applications based on responsibility and their job.

LDAP or Active Directory often holds the registry of all users in the organization, their email address, phone numbers, and work address and (most importantly) the groups of which they are a member.

In Cloud Pak for Data, users in the Finance department can be granted edit access to all projects that are owned by this department by creating a Cloud Pak for Data user group *finance* that references all these users and then, granting the suitable access at the project level.

Individual users can be made part of multiple Cloud Pak for Data user groups if their job in the organization requires. A scenario might be that James, who works in the Finance department, has edit access to the data quality projects of his department by way of the `finance-dq-editors` user group. Claire in the Chief Data Officer (CDO) organization has access to all projects across the organization and can be made a member of the `finance-dq-viewers` and `marketing-dq-viewers` groups to retain overall visibility.

To reduce administrative overhead, Cloud Pak for Data user groups often are aligned with LDAP groups, but Cloud Pak for Data user groups can consist of a combination of individual users and LDAP groups. If an LDAP server is defined (preferably through the Identity and Access Management Service (IAM) service in Cloud Pak Foundational Services), user groups can be assembled by adding LDAP groups and LDAP users.

Attribute-Based Access Control

Attribute-based access control (ABAC) is a special method that is used for defining dynamic Cloud Pak for Data user groups. It provides more flexibility to group users based on other attributes the directory service.

Note: ABAC is available only when identity and access management is delegated to Cloud Pak Foundational Services IAM.

Attribute-based access control is controlled dynamically by using the following specific Active Directory attributes:

- Location
- Nationality
- Organization
- User type

Rules can be composed of conditions with AND and OR operators. For example, an `Administrator` role can be assigned to someone in the `Head Quarters` location having a `User type` of `admin`. Any user in the company's directory service that matches these attributes can administer Cloud Pak for Data.

LDAP and single sign-on

It is recommended to delegate identity and access management within Cloud Pak for Data to the Cloud Pak Foundational Services because it supports the federation of multiple LDAPs and authentication to an external identity provider (IdP) by way of Security Assertion Markup Language (SAML). It also can receive a standardized user identity, such as name, email address, and group information by using the System for Cross-domain Identity Management (SCIM) protocol.

Foundational Services acts as the interface between external IdPs and the Cloud Pak for Data user management service. It keeps individual identities and user groups up to date with changes that are made in the IdP; for example, federate the IdP.

To implement single sign-on (SSO) for users, an external IdP that exposes the SAML/OpenID Connect must be configured in Foundational Services IAM. Alternatively, one or more LDAP servers can be configured for authentication and access control.

When a user wants to access a service within the context of Cloud Pak for Data, authentication is handled by the IAM service (see Figure 8-32).

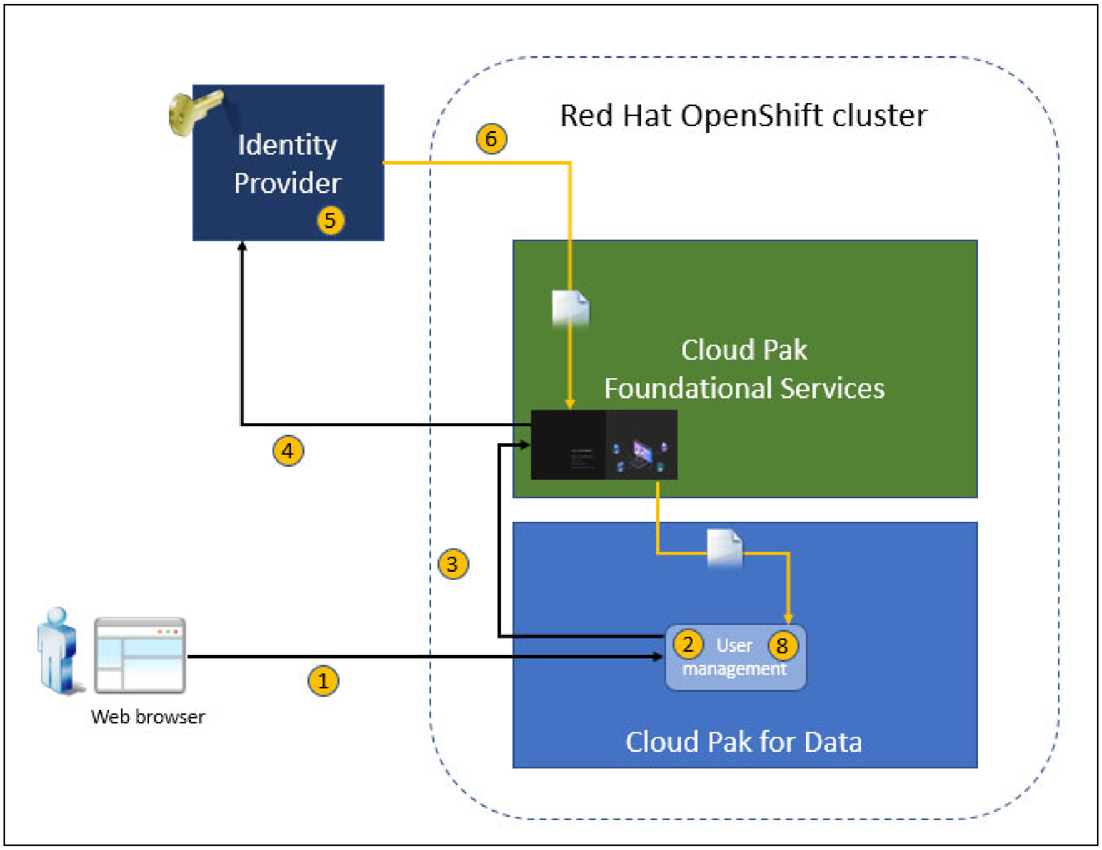

Figure 8-32 Authentication flow

The authentication flow for an user of Cloud Pak for Data features the following steps:

1. The user accesses the Cloud Pak for Data home page through the web browser.
2. The Cloud Pak for Data user management service verifies whether an authentication token is present, meaning that the user is logged in to Cloud Pak for Data or another service that uses the same identity provider.
3. If no authentication token is present, Cloud Pak for Data redirects the browser to the identity provider URL. (Foundational Services IAM).
4. Foundational Services IAM presents a login page in which the user can select the method for authentication. User selects Enterprise SAML and is redirected to the identity provider.
5. The identity provider presents a login page to the URL. Depending on the configuration of the IdP, a user can log in by using a username and password. The IdP can request more proof though multi-factor authentication (MFA).
6. If the user logs in successfully, the identity provider redirects the user back to Cloud Pak Foundational Services IAM, including an encrypted SAML response (assertion) with the user's information.
7. Foundational Services IAM returns the user information to the Cloud Pak for Data user management service.
8. Cloud Pak for Data generates an authentication token and grants access to the services the user requested (if allowed by way of Cloud Pak for Data access control).

This process represents a standard authentication flow, which also is referred to as a *SAML* or *OAuth dance*. When Foundational Services IAM is configured for Cloud Pak for Data, instead of redirecting the browser to an external IdP, Cloud Pak for Data redirects to Foundational Services IAM, which then manages redirecting to the external IdP. The “dance” includes a few other steps. If the federation is configured for the external IdP, even more intermediate steps are added to the “dance”.

8.5.2 Configuring Cloud Pak for Data authentication by using Foundational Services

In this section, we discuss how to configure Cloud Pak for Data authentication by using Foundational Services.

Delegating authentication to Foundational Services IAM

When a default installation of Cloud Pak for Data is used, an `admin` user is created. This user can be used to configure roles, groups, and users on the platform through the internal user registry. However, is not recommended to use this user registry other than for testing.

Complete the following steps to delegate authentication to Foundational Services IAM:

1. To enable authentication by using an external IdP, you must first enable IAM on the `Ibmcpd` custom resource. This process is be done by editing the resource and changing the properties, or directly by patching by way of the Red Hat OpenShift client:

   ```
   CP4D_PROJECT=cpd-instance
   oc edit Ibmcpd -n ${CP4D_PROJECT} ibmcpd-cr
   ```

2. As shown in Example 8-32, we assume that cpd-`instance` is the Red Hat OpenShift project of your Cloud Pak for Data instance.

Example 8-32 Editing the Ibmcpd custom resource

```
apiVersion: cpd.ibm.com/v1
kind: Ibmcpd
metadata:
  name: ibmcpd-cr
  namespace: cpd-instance
spec:
…
  cloudpakfordata: true
  iamIntegration: true
```

   Not all attributes of the `Ibmcpd` custom resource are shown in Example 8-32, only the two properties that are applicable to activating authentication by using IAM.

   Note: Enabling IAM integration cannot be reversed without help from IBM support.

   The Cloud Pak for Data platform operator ensures that all required services within Foundational Services are started. This takes approximately 20 - 30 minutes.

3. Run the following command to check the status of `ZenService`:

   ```
   oc get ZenService -n ${CP4D_PROJECT} lite-cr -o
   jsonpath='{.status.zenStatus}{"\n"}'
   ```

The `zenStatus` property changes to `InProgress` after a few minutes and when finished, the property is set to `Completed`. Meanwhile, you can monitor the pods in the `ibm-common-services` project to ensure that the IAM services are started:

```
oc get pods -n ibm-common-services -w
```

4. If you need the Cloud Pak for Data host name, retrieve it as shown in the following example:

```
CP4D_HOST=$(oc get route -n ${CP4D_PROJECT} cpd \
-o jsonpath='{.spec.host}{"\n"}') && echo $CP4D_HOST
```

Obtaining Cloud Pak Foundational Services properties

Complete the following steps to obtain Cloud Pak Foundational Services properties:

1. Obtain the URL of the Foundational Services route by using the following command:

```
FS_CONSOLE_HOST=$(oc get route -n ibm-common-services cp-console -o
jsonpath='{.spec.host}') && echo $FS_CONSOLE_HOST
```

The output that is returned resembles the following example:

```
cp-console.itzroks-270001318b-0vtwtn-6ccd7f378ae819553d37d5f2ee142bd6-0000.eu-g
b.containers.appdomain.cloud
```

The Foundational Services console is available at the following URL:

https://cp-console.itzroks-270001318b-0vtwtn-6ccd7f378ae819553d37d5f2ee142bd6-0000.eu-de.containers.appdomain.cloud

2. Extract the admin password as shown in Example 8-33.

Example 8-33 Extracting the admin password

```
FS_ADMIN_PASSWORD=$(oc extract -n ibm-common-services
secret/platform-auth-idp-credentials --to=- --keys=admin_password) && echo
$FS_ADMIN_PASSWORD
```

Output:

```
# admin_password
8CuJ4MCAt2AKz9nwBLbSYvPgqnR45QjH
```

3. Obtain the bearer token of the current console session (see Example 8-34). This access token is needed for various tasks.

Example 8-34 Obtain the bearer token

```
FS_ACCESS_TOKEN=$(curl -k -s -H "Content-Type:
application/x-www-form-urlencoded;charset=UTF-8" -d
"grant_type=password&username=admin&password=${FS_ADMIN_PASSWORD}&scope=openid"
"https://${FS_CONSOLE_HOST}/idprovider/v1/auth/identitytoken" | jq -r
.access_token) && echo ${FS_ACCESS_TOKEN}
```

8.5.3 Configuring Foundational Services IAM for Azure AD using SAML

For more information about this process, see this IBM Documentation web page.

In this section, figures demonstrate what to expect and how a user that is registered in Active Directory can log in to Cloud Pak for Data and attain the correct permissions.

Also, it is assumed that the Red Hat OpenShift client (`oc`) is installed on the server or workstation on which you run these steps.

Retrieving the SAML metadata file from Cloud Pak Foundational Services

Cloud Pak Foundational Services IAM connects as a client to a registered application in Azure AD. To enable this connection, the SAML client metadata of the IAM service must be extracted.

Complete the following steps:

1. Enable SAML:

   ```
   curl -k -X PUT --header "Authorization: Bearer ${FS_ACCESS_TOKEN}" \
   --header 'Content-Type: application/json' -d '{"enable": true}' \
   https://${FS_CONSOLE_HOST}/idmgmt/v1/saml/management
   ```

2. Export the SAML client metadata file:

   ```
   curl -k -s -X GET --header "Authorization: Bearer ${FS_ACCESS_TOKEN}"
   https://${FS_CONSOLE_HOST}/idprovider/v3/saml/metadata/defaultSP >
   /tmp/fs-iam-client.xml
   ```

   The output resembles the example that is shown in Example 8-35.

Example 8-35 SAML client metadata file

```
<?xml version="1.0" encoding="UTF-8"?><md:EntityDescriptor
xmlns:md="urn:oasis:names:tc:SAML:2.0:metadata"
entityID="https://cp-console.itzroks-270001318b-
5bxqln-6ccd7f378ae819553d37d5f2ee142bd6-0000.eu-de.containers.appdomain.cloud/ibm/
saml20/defaultSP
.
.
output suppressed
.
.
</md:EntityDescriptor>
```

Now, you can continue to set up Active Directory on Azure.

Creating or using an Active Directory on Azure

If you ran the Active Directory service on Azure and want to connect Cloud Pak for Data to your current tenant, you can skip this step.

Complete the following steps to configure a tenant for Azure AD and register some groups and users:

1. Go to the Azure portal.
2. In the search bar, enter `Active Directory` and then. select the **Azure Active Directory** service. If you signed up for Azure, you can use the default tenant of the account. If you have a directory and only want to try out SAML, you can use the existing tenant or create a tenant to isolate your user registry.

   In this example we use the `redbookorg` tenant (see Figure 8-33).

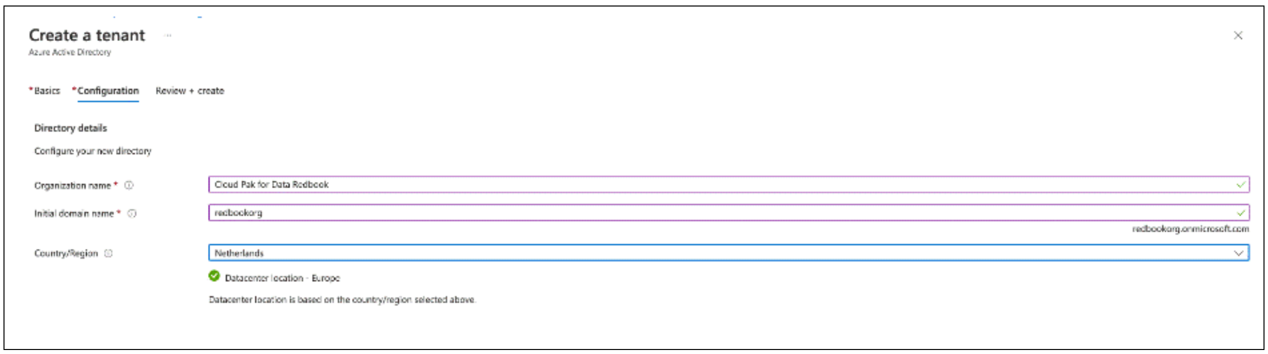

Figure 8-33 Creating an Active Directory tenant

Creating or identifying groups and users

In Active Directory (AD), you can create or find groups and individual users who can access Cloud Pak for Data. AD users or AD groups then identify the roles (permissions) they have within the platform.

In our example, we created the following groups (see Figure 8-34):

- AD CP4D Admins
- AD Data Engineers
- AD Data Scientists

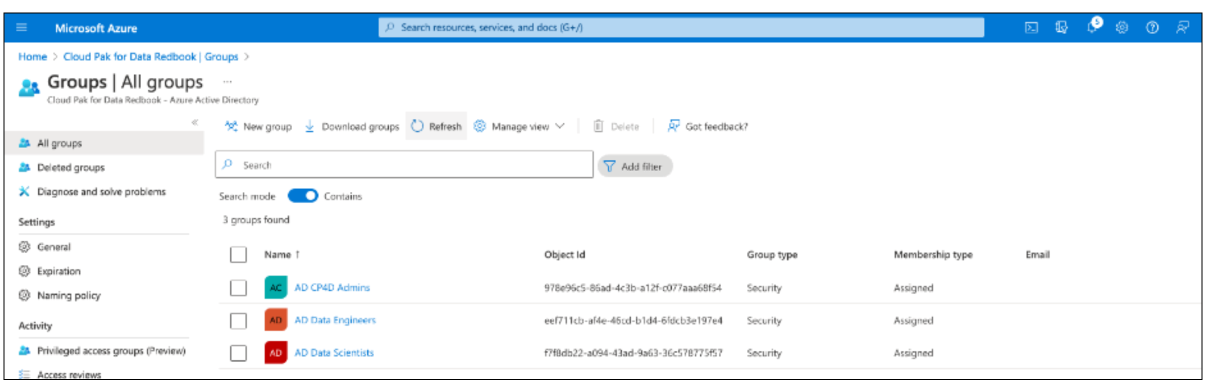

Figure 8-34 Active Directory groups

The AD CP4D Admins group has one member: Tara Toussaint.

The Data Engineers group has two members:

- Rosa Ramones
- Shelly Sharpe

The Data Scientists group has three members (see Figure 8-35):

- Paco Primo
- Shelly Sharpe
- Rico Roller

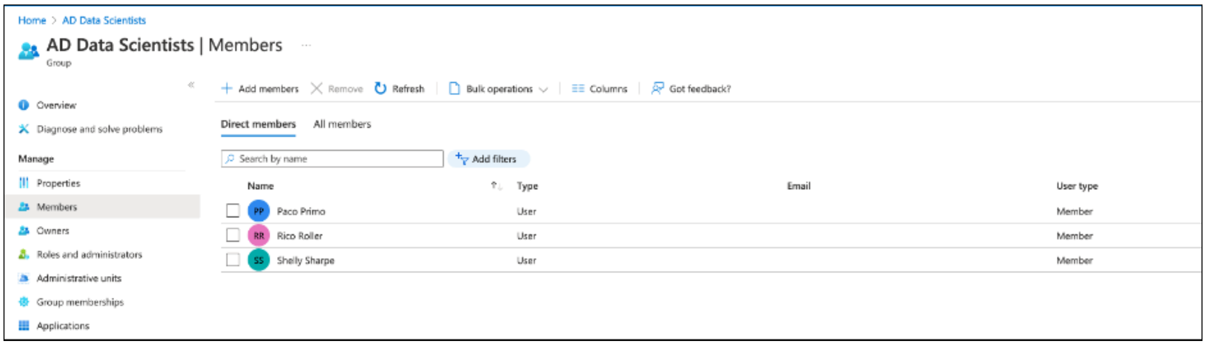

Figure 8-35 Data Scientists group members

We use the email addresses of the users to authenticate them to AD and to Cloud Pak for Data. Now that users and groups were created or identified, we can create an Enterprise Application to register the Cloud Pak Foundational Services client.

Creating an Active Directory application

Complete the following steps to create an Active Directory application:

1. Browse to your directory by selecting **Azure Active Directory** from the Azure menu. Then, select **Enterprise applications** from the left menu (see Figure 8-36).

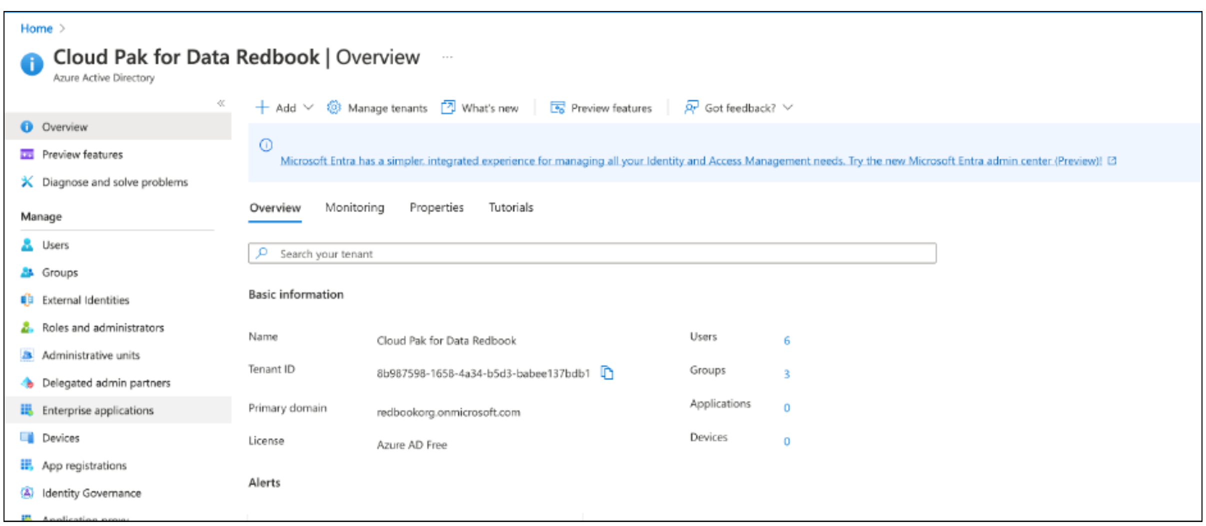

Figure 8-36 Enterprise applications

2. Add an application by clicking **New application** and then, selecting **Create your own application**.

3. Enter the name for the application (in our example, we used `cp4d-redbook`). Click **Create** (see Figure 8-37).

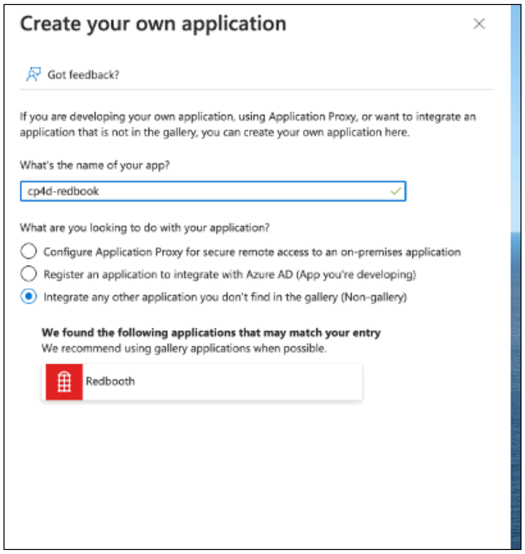

Figure 8-37 Creating an application

4. Click the **Assign users and groups** tile (see Figure 8-38).

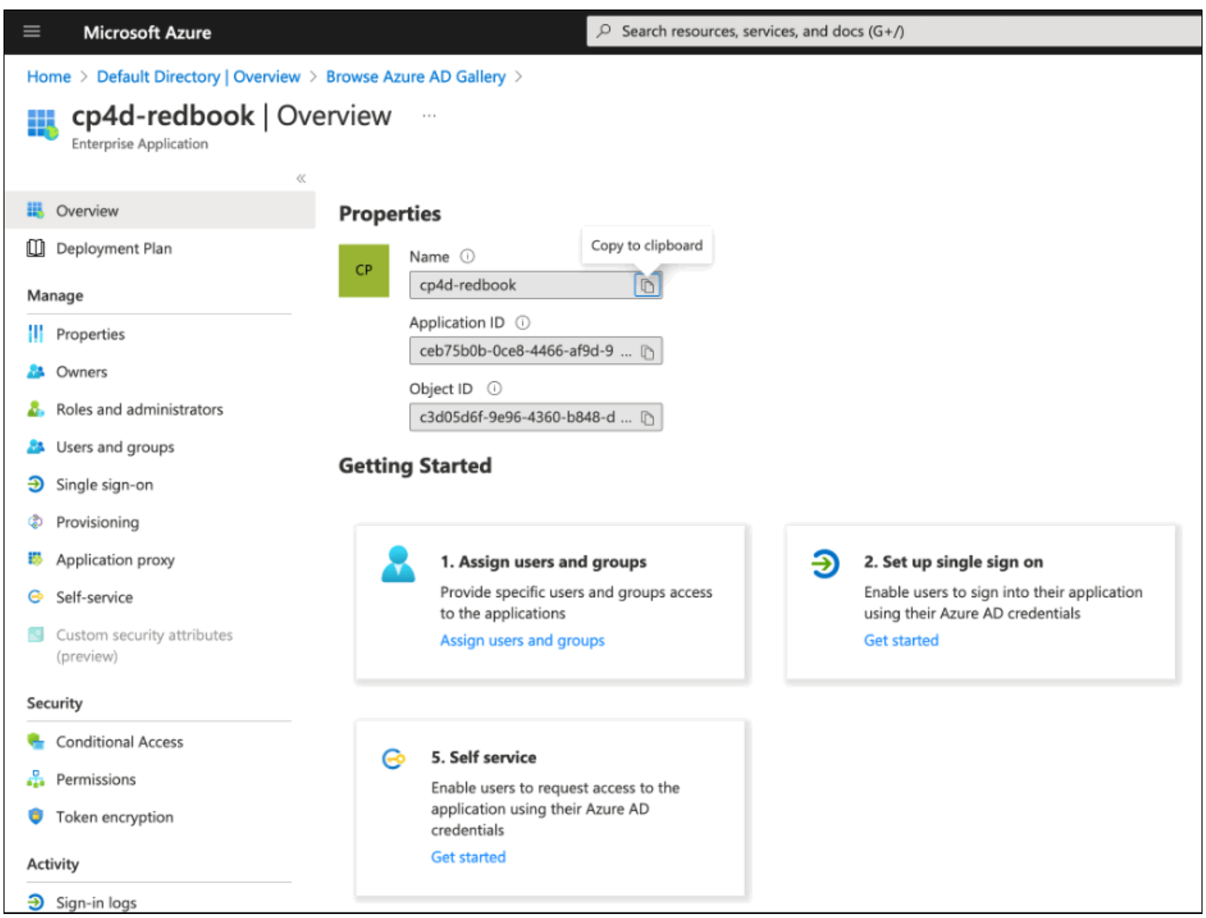

Figure 8-38 Assigning users

5. Click **Add user/group** to add assignments. You must click the **None selected** link that is under the Users header to assign users. On the right side of the window, a list of users appears. You can assign groups to the application only if you have a paid or premium plan for Active Directory.
6. Click all of the users that must access the `cp4d-redbook` application and then, click **Select** (see Figure 8-39).

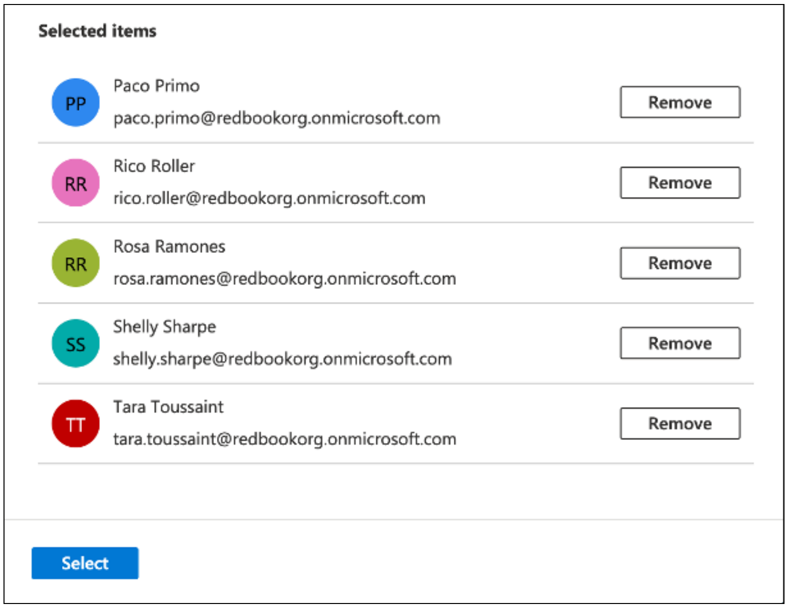

Figure 8-39 Selecting users for the application

7. Click **Assign** in the Add assignment window to assign these users to the application. Your window likely resembles the example that is shown in Figure 8-40.

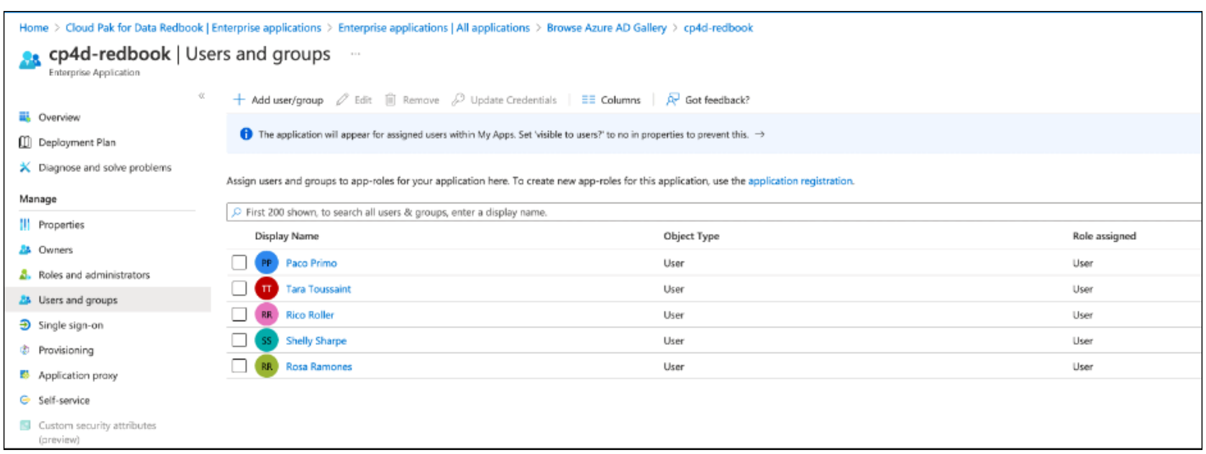

Figure 8-40 Users and groups for application

Setting up single sign-on

Complete the following steps to connect the Foundational Services IAM service as a client to the `cp4d-redbook` application:

1. Click the **Single sign-on** link in left menu and then, click the **SAML** tile. Active Directory shows the SAML properties of the application (see Figure 8-41).

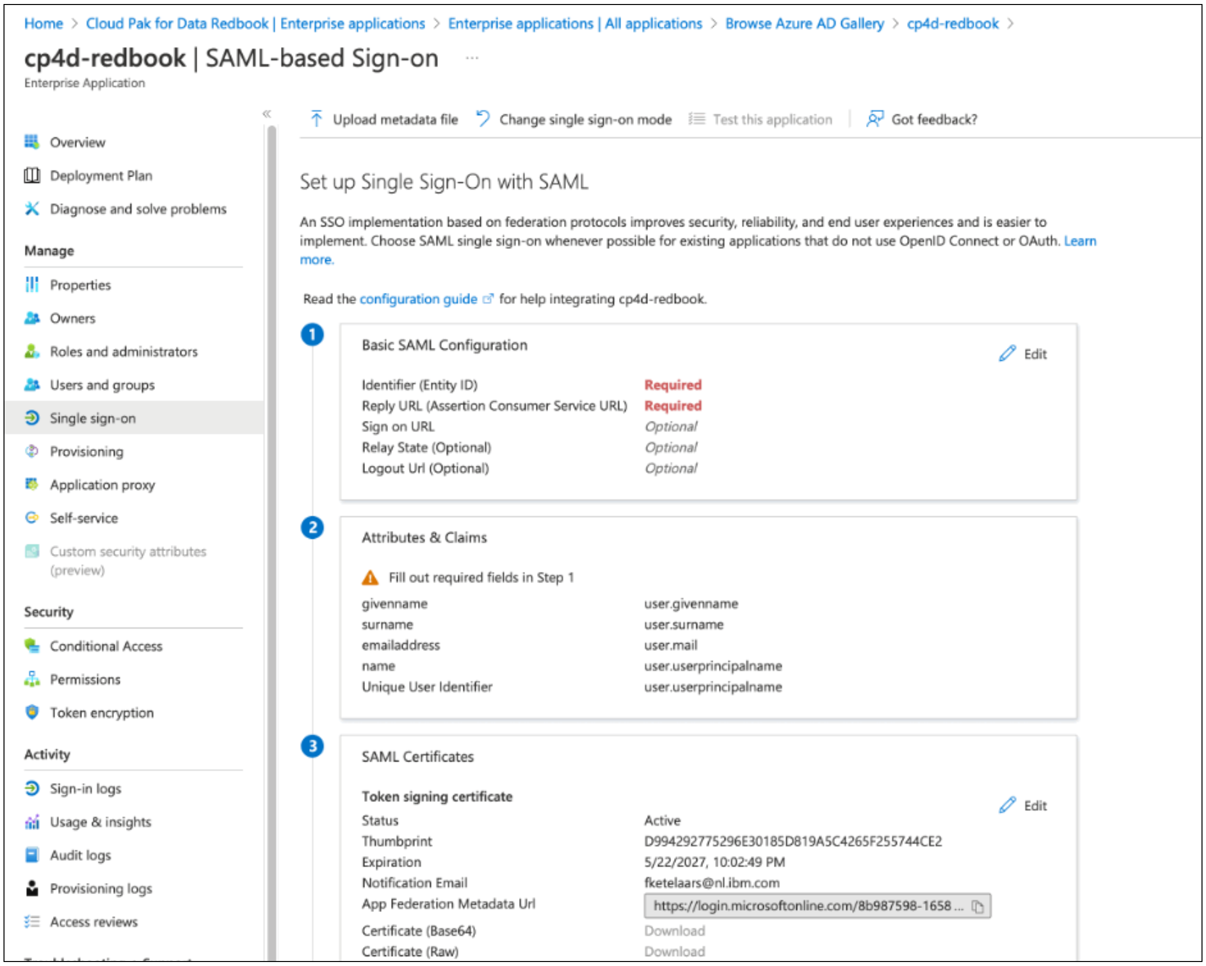

Figure 8-41 SAML configuration

2. Click **Upload metadata file** at the top of the window and then, select the **fs-iam-client.xml** file that you extracted. Click **Add**. A window with Basic SAML Configuration opens (see Figure 8-42).

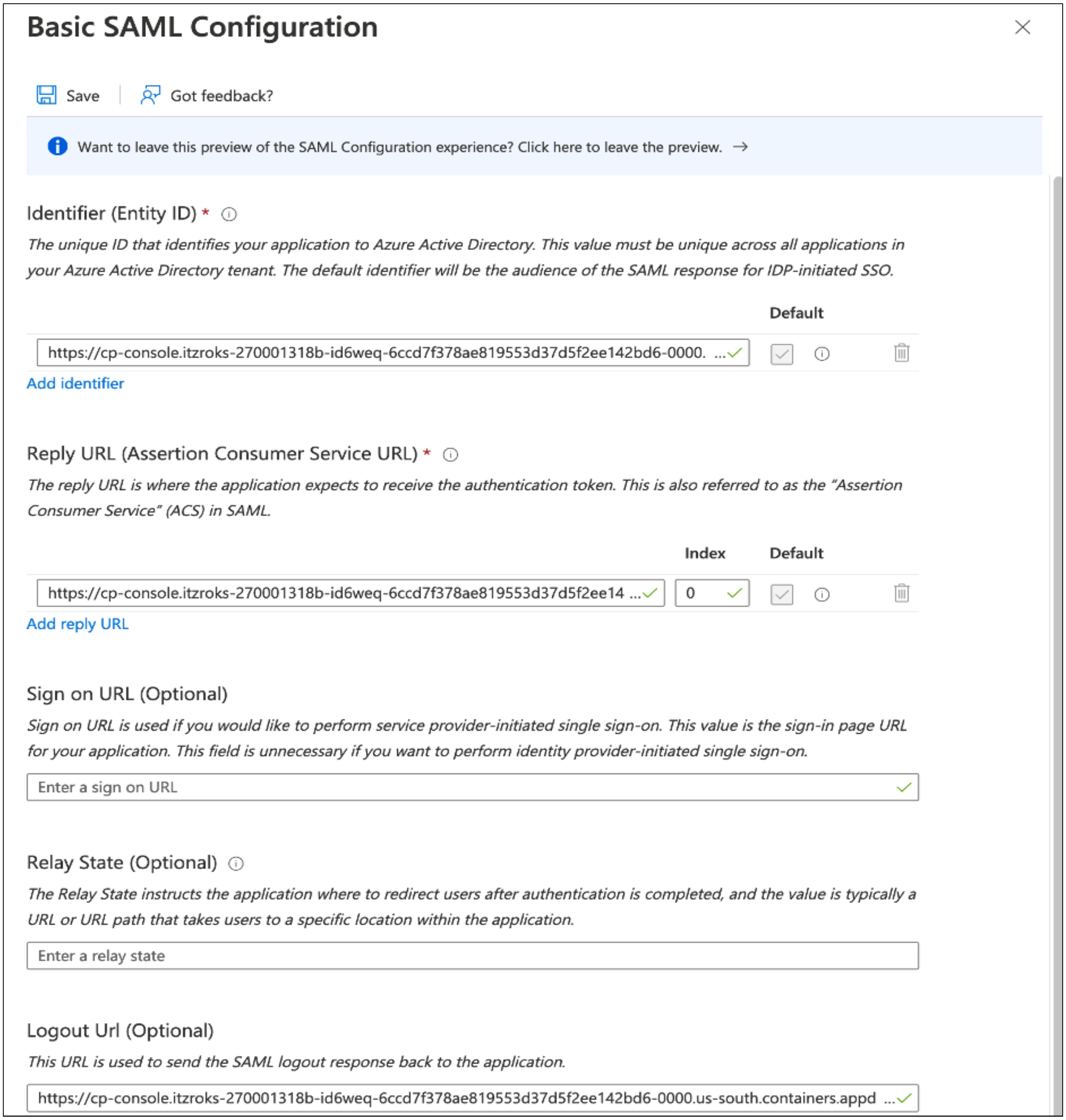

Figure 8-42 Basic SAML configuration

3. Click **Save** and close the window. Now, the SAML Configuration, such as Entity ID, Reply URL, and Logout URL are populated.
4. On the **Attributes & Claims** tile, click **Edit**.

5. Click **Add a group claim** and select the **All groups** option. Click **Save** (see Figure 8-43).

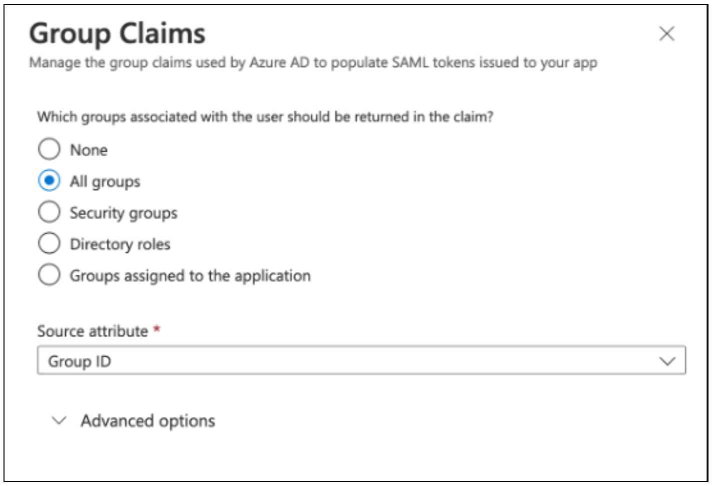

Figure 8-43 Add group claim

When a user successfully authenticates, Active Directory sends the properties and all group memberships back to Cloud Pak Foundational Services.

6. Return to the Set up Single Sign-on with SAML page and download the Federation Metadata XML from the **SAML Certificates** tile.

   In the example, a file that is named `cp4d-redbook.xml` is downloaded, which is the application definition that we upload into the Foundational Services IAM configuration.

Converting the IdP metadata

Before adding the IdP registration to Foundational Services IAM, base64 encode the XML file:

```
IDP_METADATA=$(cat /tmp/cp4d-redbook.xml | base64 -w0)
```

Creating the IdP registration with the directory attribute mapping

This process configures how Active Directory attributes (`https://schemas.../claims/attribute`) are mapped to the IAM attributes.

Example 8-36 shows how to post the attribute mapping.

Example 8-36 Posting the attribute mapping

```
curl -k -X POST "https://${FS_CONSOLE_HOST}/idprovider/v3/auth/idsource" \
--header 'Content-Type: application/json' \
--header "Authorization: Bearer $FS_ACCESS_TOKEN" --data-raw '{
  "name": "cp4d_redbook_azure",
  "description": "Azure SAML configuration",
  "protocol": "saml",
  "type": "default",
  "idp_config": {
    "token_attribute_mappings": {
    "sub": "http://schemas.xmlsoap.org/ws/2005/05/identity/claims/name",
    "given_name":
"http://schemas.xmlsoap.org/ws/2005/05/identity/claims/givenname",
    "family_name":
"http://schemas.xmlsoap.org/ws/2005/05/identity/claims/surname",
```

```
    "groups": "http://schemas.microsoft.com/ws/2008/06/identity/claims/groups",
    "email": "http://schemas.xmlsoap.org/ws/2005/05/identity/claims/emailaddress",
    "first_name":
"http://schemas.xmlsoap.org/ws/2005/05/identity/claims/givenname",
    "last_name": "http://schemas.xmlsoap.org/ws/2005/05/identity/claims/surname"
    }, "idp_metadata": "'${IDP_METADATA}'"
  }, "jit": true }'
```

The result resembles the following example:

```
{"status":"success","message":"Identity provider {cp4d_redbook_azure} is
successfully registered with unique identifier defaultSP"}
```

Check if SAML is active:

```
curl -k -X GET --header "Authorization: Bearer ${FS_ACCESS_TOKEN}"
https://${FS_CONSOLE_HOST}/idmgmt/v1/saml/status
```

Assigning AD groups to Cloud Pak for Data user groups

Finally, we need to create the Cloud Pak for Data user groups and add the Active Directory groups to these user groups. In our example, we intentionally chose to differ between AD group names and Cloud Pak for Data user group names to emphasize that these names do not have to match.

It also is possible to add users to the Cloud Pak for Data user groups; however, in most organizations with sizable number of users, these groups are likely difficult to maintain.

In this version of Cloud Pak for Data, no graphical user interface is used to assign groups from the IdP to Cloud Pak for Data user groups. Therefore, all configuration work must by done by using APIs.

Complete the following steps:

1. Retrieve t the Cloud Pak for Data token that we need for the remainder of the APIs calls, as shown in Example 8-37.

Example 8-37 Getting the Cloud Pak for Data token

```
CP4D_TOKEN=$(curl -s -k -X GET "https://${CP4D_HOST}/v1/preauth/validateAuth"
--header 'username: admin' --header "iam-token: ${FS_ACCESS_TOKEN}" | jq -r
.accessToken) && echo ${CP4D_TOKEN}
```

2. Create the three groups, each with their own roles. The group ID must be kept because it is needed when Active Directory groups are assigned to the user groups (see Example 8-38).

Example 8-38 Creating three groups

```
CP4D_ADMIN_GID=$(curl -s -k -X POST "https://${CP4D_HOST}/usermgmt/v2/groups" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d  '{"name": "cp4d-admins",
"description": "Cloud Pak for Data Admins",
"role_identifiers": ["zen_administrator_role"]
}' | jq -r .group_id) && echo ${CP4D_ADMIN_GID}

CP4D_DE_GID=$(curl -s -k -X POST "https://${CP4D_HOST}/usermgmt/v2/groups" \
```

```
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d  '{"name": "cp4d-data-engineers",
"description": "Cloud Pak for Data Data Engineers",
"role_identifiers": ["zen_data_engineer_role"]
}' | jq -r .group_id) && echo ${CP4D_DE_GID}

CP4D_DS_GID=$(curl -s -k -X POST "https://${CP4D_HOST}/usermgmt/v2/groups" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d  '{"name": "cp4d-data-scientists",
"description": "Cloud Pak for Data Data Scientists",
"role_identifiers": ["wkc_data_scientist_role"]
}' | jq -r .group_id) && echo ${CP4D_DS_GID}
```

3. If the groups were created previously, you can use the API that is shown in Example 8-39 to retrieve the group details, including the group ID.

Example 8-39 Retrieving the group details

```
curl -s -k -X GET "https://${CP4D_HOST}/usermgmt/v2/groups" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" | jq -r .
```

Now that the Cloud Pak for Data user groups are created, we can assign the Active Directory groups as members. When authenticating with AD, the Azure SAML response contains the Object IDs of all groups to which the user belongs, as configured here.

> **Note:** We assigned all uses as described in "Creating or using an Active Directory on Azure" on page 640

4. Go to the Azure portal to find the Object IDs of the Azure groups that we want to use (see Figure 8-44).

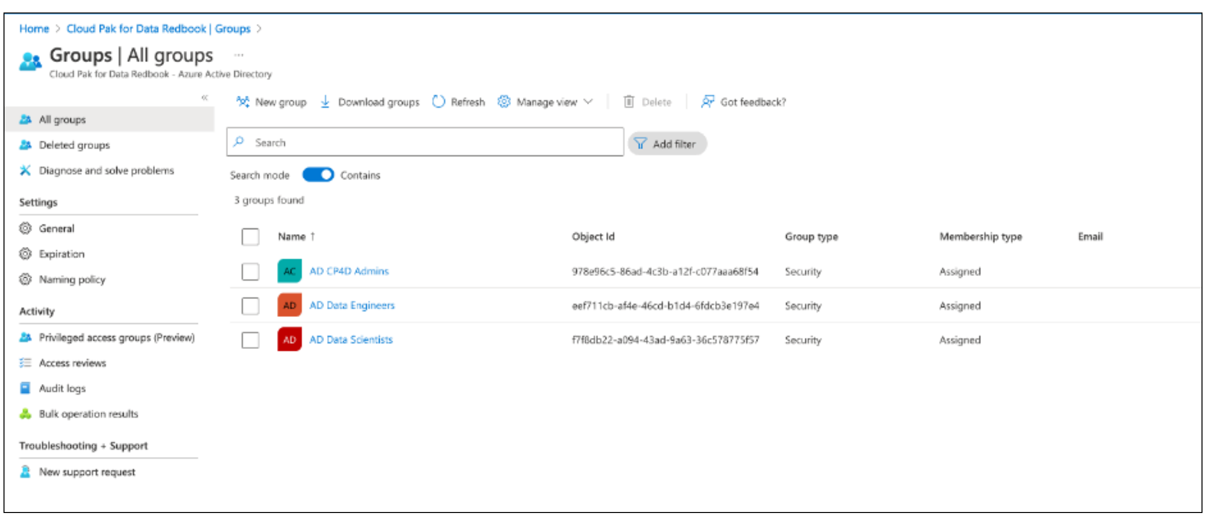

Figure 8-44 Active Directory groups

The following object IDs are used:

```
AD_ADMIN_OID="978e96c5-86ad-4c3b-a12f-c077aaa68f54"
AD_DE_OID="eef711cb-af4e-46cd-b1d4-6fdcb3e197e4"
AD_DS_OID="f7f8db22-a094-43ad-9a63-36c578775f57"
```

5. Add the Azure groups into the user groups, first for the Cloud Pak for Data Administrators (see Example 8-40).

Example 8-40 Adding the Azure groups into the user groups

```
curl -k -s  -X POST
"https://${CP4D_HOST}/usermgmt/v2/groups/${CP4D_ADMIN_GID}/members" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d '{
"ldap_groups": ["'${AD_ADMIN_OID}'"]
}'
```

6. Repeat this step for the other two groups (see Example 8-41).

Example 8-41 Add the Azure groups into the user groups

```
curl -k -s  -X POST
"https://${CP4D_HOST}/usermgmt/v2/groups/${CP4D_DE_GID}/members" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d '{
"ldap_groups": ["'${AD_DE_OID}'"]
}'

curl -k -s  -X POST
"https://${CP4D_HOST}/usermgmt/v2/groups/${CP4D_DS_GID}/members" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d '{
"ldap_groups": ["'${AD_DS_OID}'"]
}'
```

All API calls should have a response the resembles the following example:

{"group_id":10003,"_messageCode_":"success","message":"success"}

Checking groups in Cloud Pak for Data

Complete the following steps to confirm that the user groups were created and the AD groups are members:

1. Go to the Cloud Pak for Data home page:

   ```
   echo $CP4D_HOST
   ```

2. At the log-in window, select **IBM provided credentials** (see Figure 8-45).

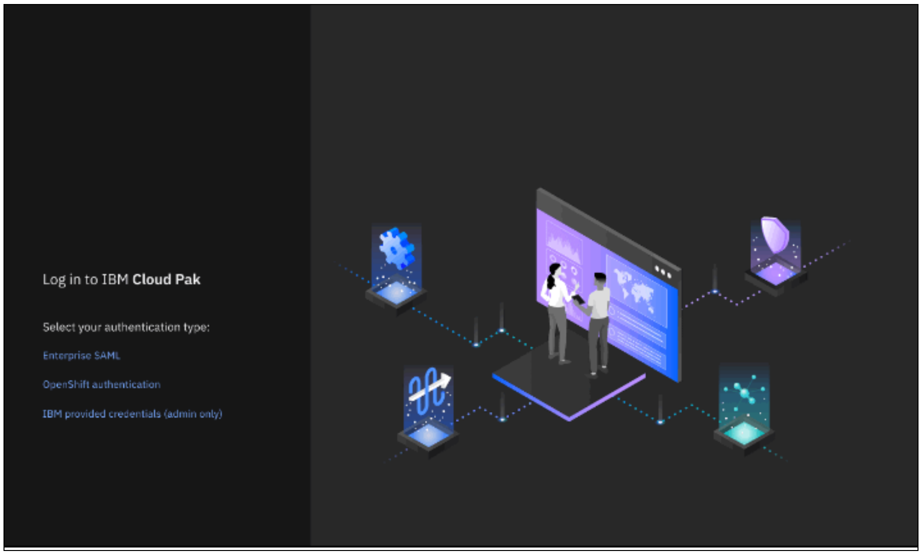

Figure 8-45 Platform login

3. Log in as `admin` by using the Foundational Services admin password. After you are logged in, click the navigation menu and select **Access Control**. Then, click the **User groups** tab (see Figure 8-46).

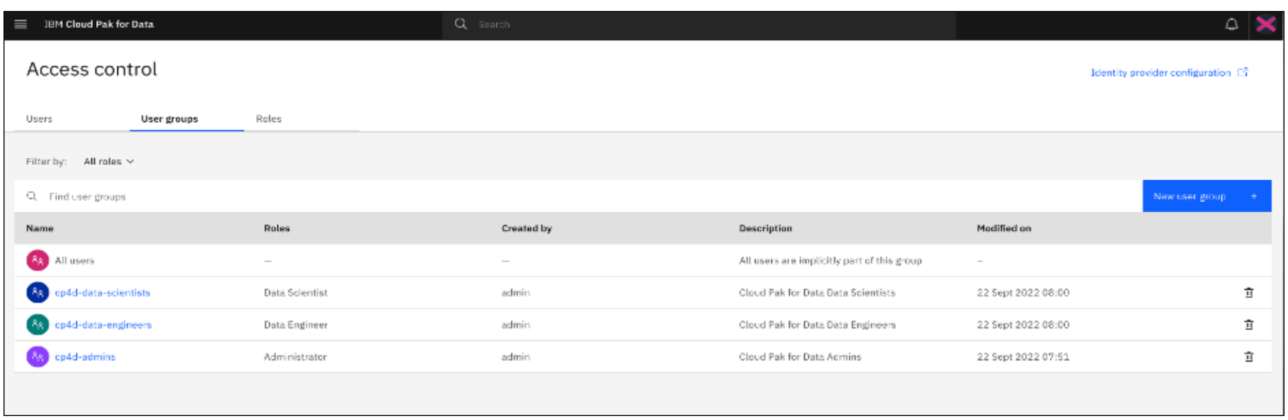

Figure 8-46 Cloud Pak for Data user groups

4. You can see the three user groups. Click the **cp4d-admins** group to see the associated AD group (see Figure 8-47).

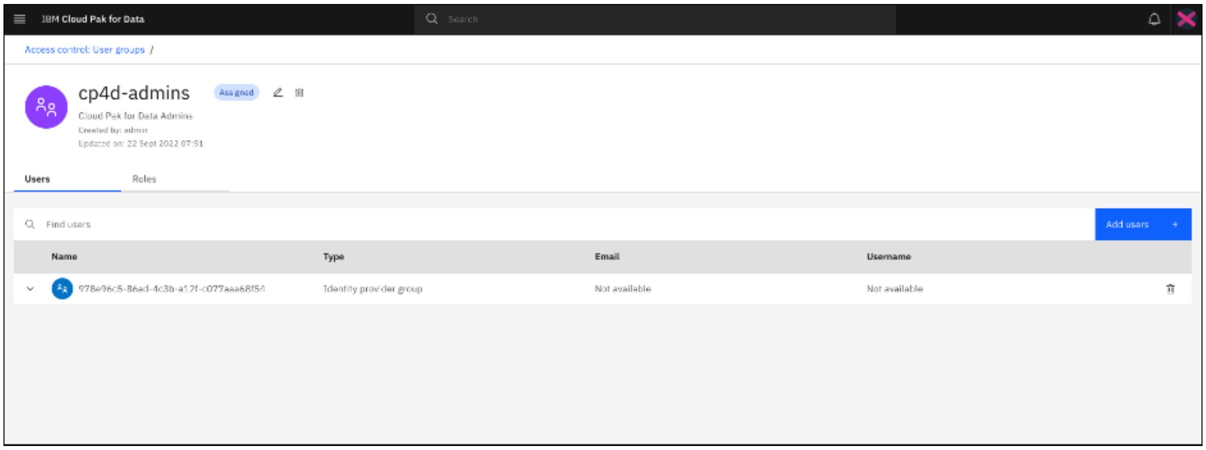

Figure 8-47 Active Directory groups for Cloud Pak for Data user group

5. Click the **Roles** tab. You see that this user group was assigned the `Administrator` role (see Figure 8-48).

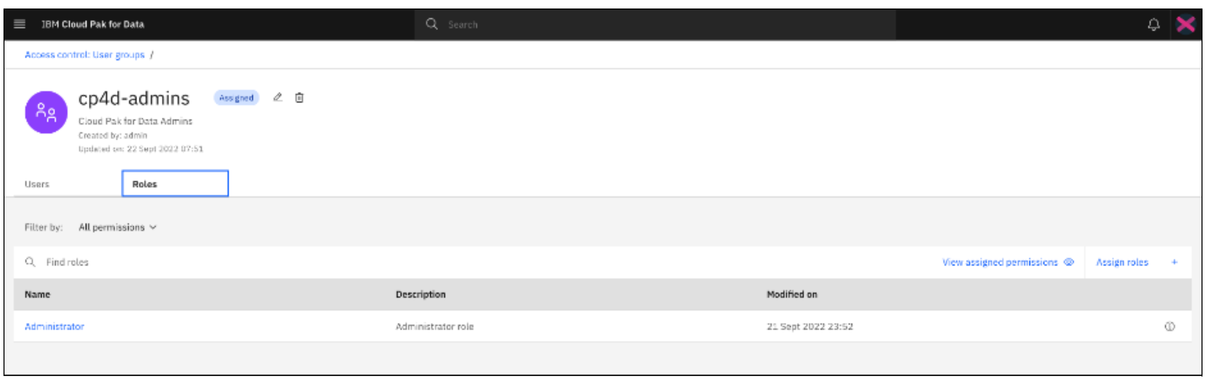

Figure 8-48 Roles for user group

6. Repeat these steps for the `cp4d-data-engineers` and `cp4d-data-scientists` user groups. Check that the correct AD group is a member and that the correct roles were assigned to the user group.
7. Log out from Cloud Pak for Data and click **Log in** on the logout page. Then, click **Change your authentication method**.

You are directed to the generic Cloud Pak log-in page (see Figure 8-49).

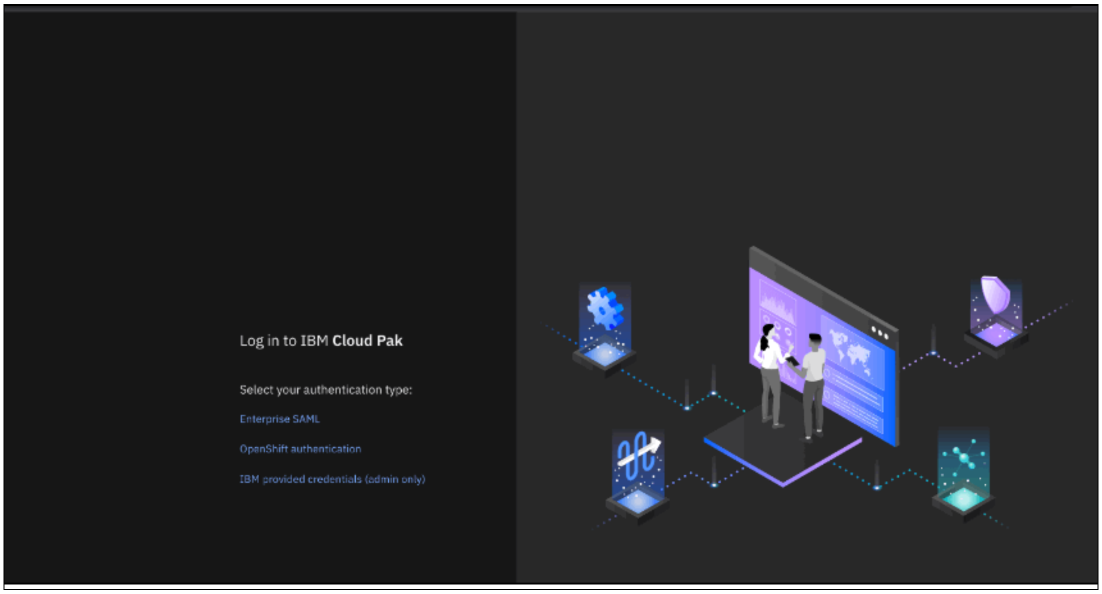

Figure 8-49 Platform login page

8. Click **Enterprise SAML**, which redirects the browser to the Microsoft Sign in page (see Figure 8-50).

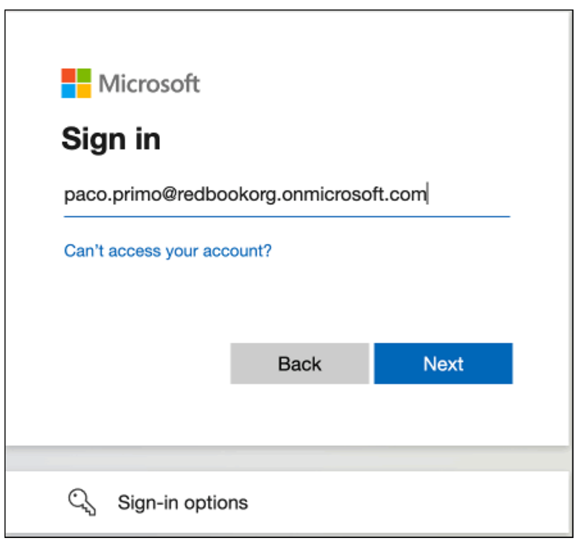

Figure 8-50 Azure Sign in window

9. Log in as one of the users. After you are logged in, you are directed to the Cloud Pak for Data home page (see Figure 8-51).

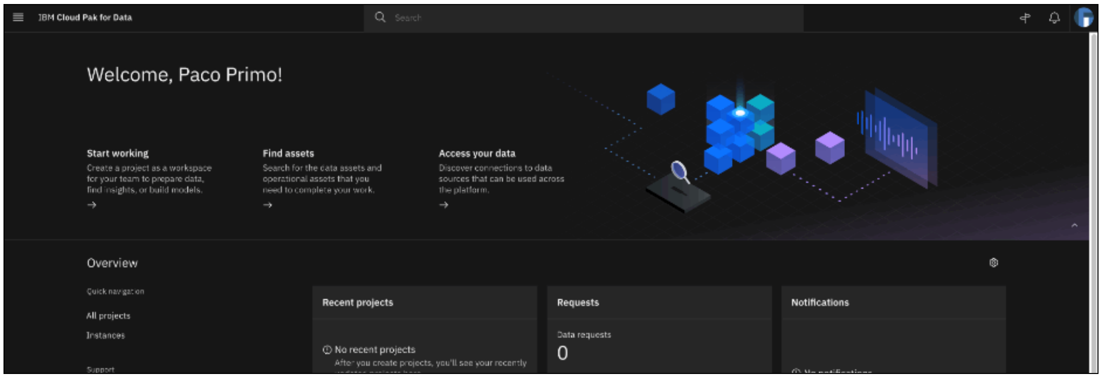

Figure 8-51 Welcome Paco

8.5.4 Configuring Foundational Services IBM Security Verify by way of OIDC

For more information about this process, see this IBM Documentation web page.

This section adds some figures to demonstrate what to expect and also how a user registered in IBM Security® Verify can log in to Cloud Pak for Data and attain the correct permissions.

It is assumed that the Red Hat OpenShift client (`oc`) is installed on the server or workstation on which you run these steps.

Complete the following steps:

1. Create or use IBM Security Verify (ISV)

   If you use IBM Security Verify (ISV) and want to connect Cloud Pak for Data to your current identity provider, you can skip this step. To register for a free trial of IBM Security Verify, see this web page. Click **Try Verify Now** and enter your account information. If you have an IBM account, enter the email addresses that is associated with that account to sign up.

2. Set up your tenant. In our examples, we set up tenant `cp4d-redbook` (see Figure 8-52.

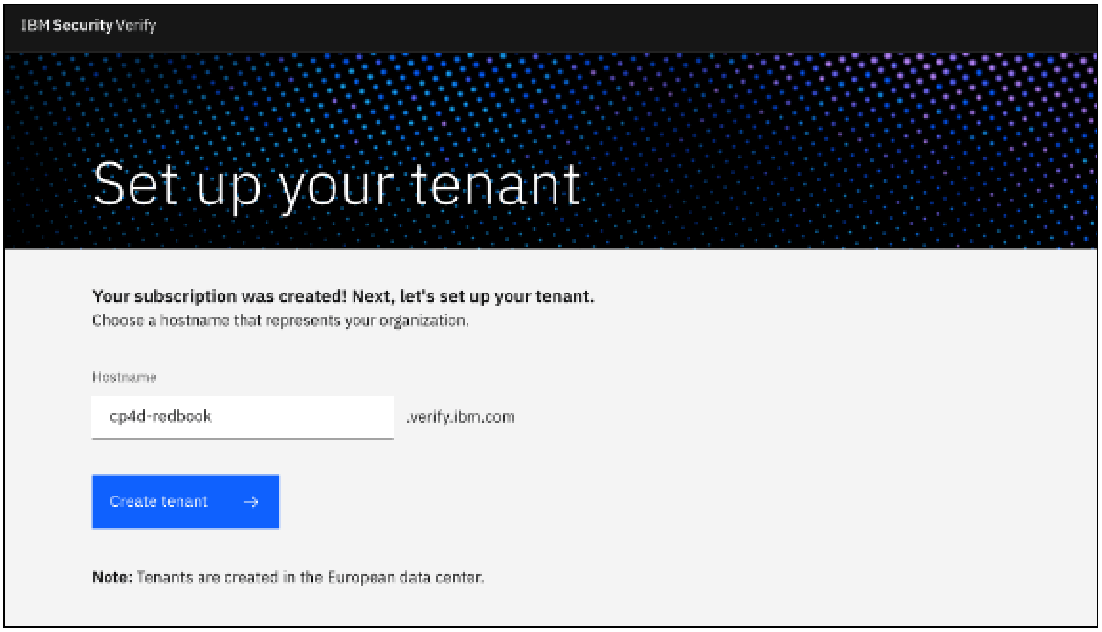

Figure 8-52 Creating an ISV tenant

3. If you use a tenant, you can access it by using the following link:

   `https://<your-tenant-name>.verify.ibm.com`

   In the example, the link is `https://cp4d-redbook.verify.ibm.com/`

Creating or identifying groups and users

In ISV Directory, you can create or find groups and individual users who can access Cloud Pak for Data. Users or Groups eventually are assigned to Cloud Pak for Data groups that then depict the roles (permissions) they have within the platform.

We are creating the following Cloud Directory users for our exercise:

- Paco Primo (paco.primo@cp4d-redbook.com)
- Shelly Sharpe (shelly.sharpe@cp4d-redbook.com)
- Rosa Ramones (rosa.ramones@cp4d-redbook.com)
- Rico Roller (rico.roller@cp4d-redbook.com)
- Tara Toussaint (tara.toussaint@cp4d-redbook.com)

When adding users, ensure you scroll down to specify the preferred email address (see Figure 8-53).

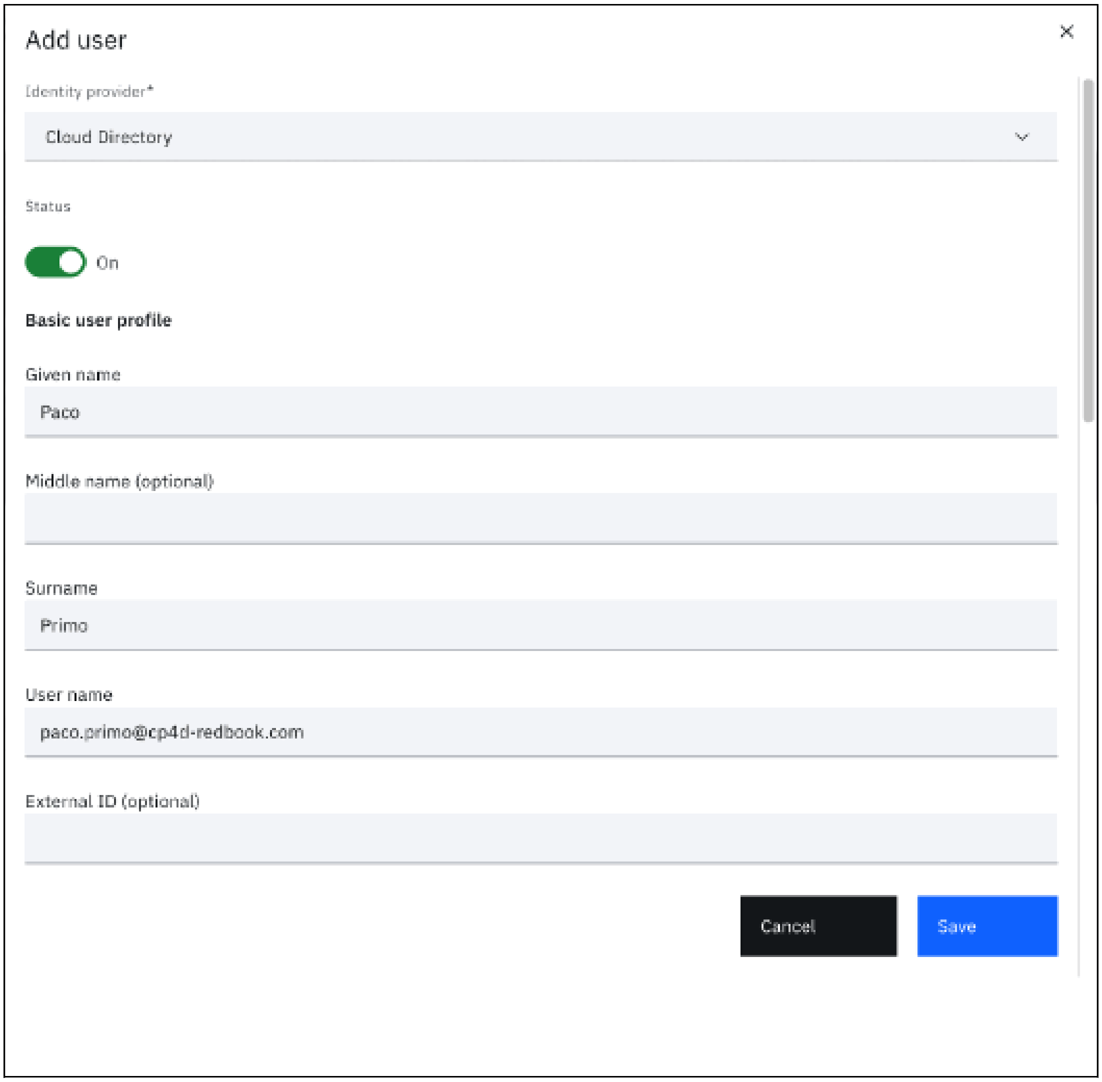

Figure 8-53 Adding a directory user

After this process is completed, the users that are shown in Figure 8-54 are in your directory.

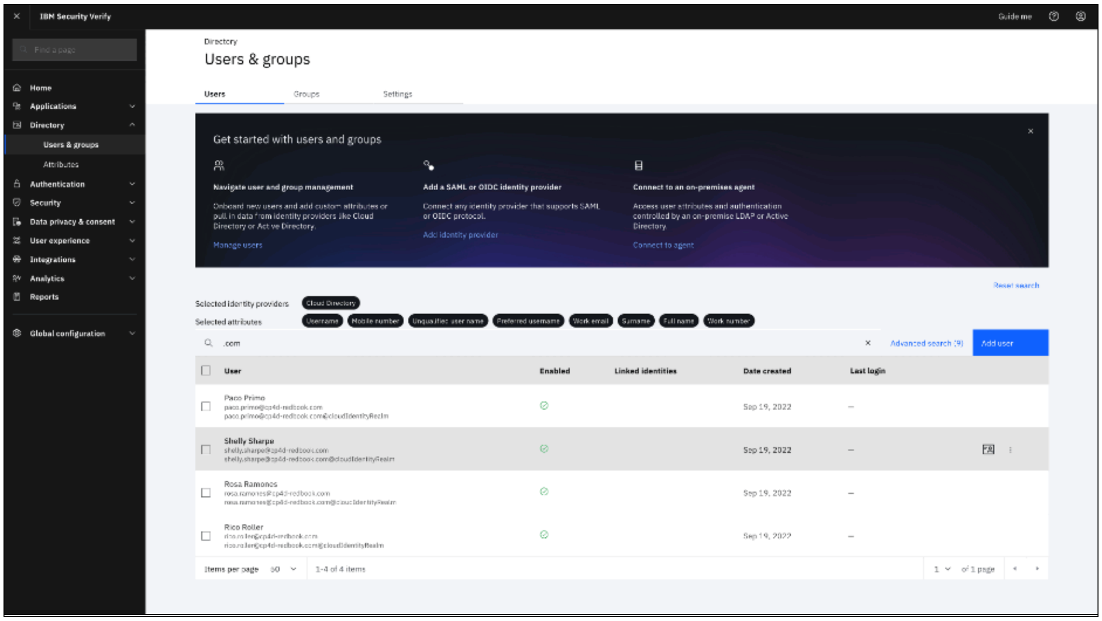

Figure 8-54 ISV Directory users

In our example, we created the groups that are shown in Figure 8-55.

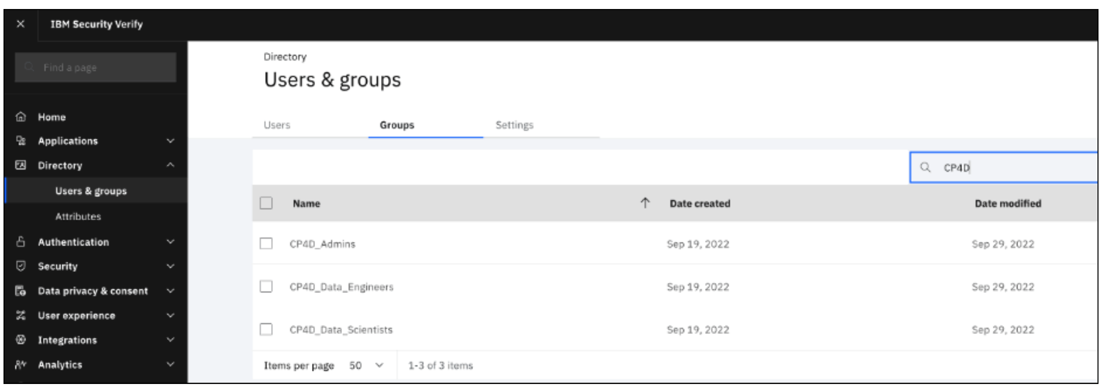

Figure 8-55 Directory groups

The CP4D_Admins group has one member: Tara Toussaint

The CP4D_Data_Engineers AD group has two members:

- Rosa Ramones
- Shelly Sharpe

The CP4D_Data_Scientists group has three members:

- Paco Primo
- Shelly Sharpe
- Rico Roller

Note: ISV groups must not contain spaces because such spaces cause issues later when members are added to the Cloud Pak for Data user groups.

Figure 8-56 shows an example for the `CP4D_Data Scientists` group.

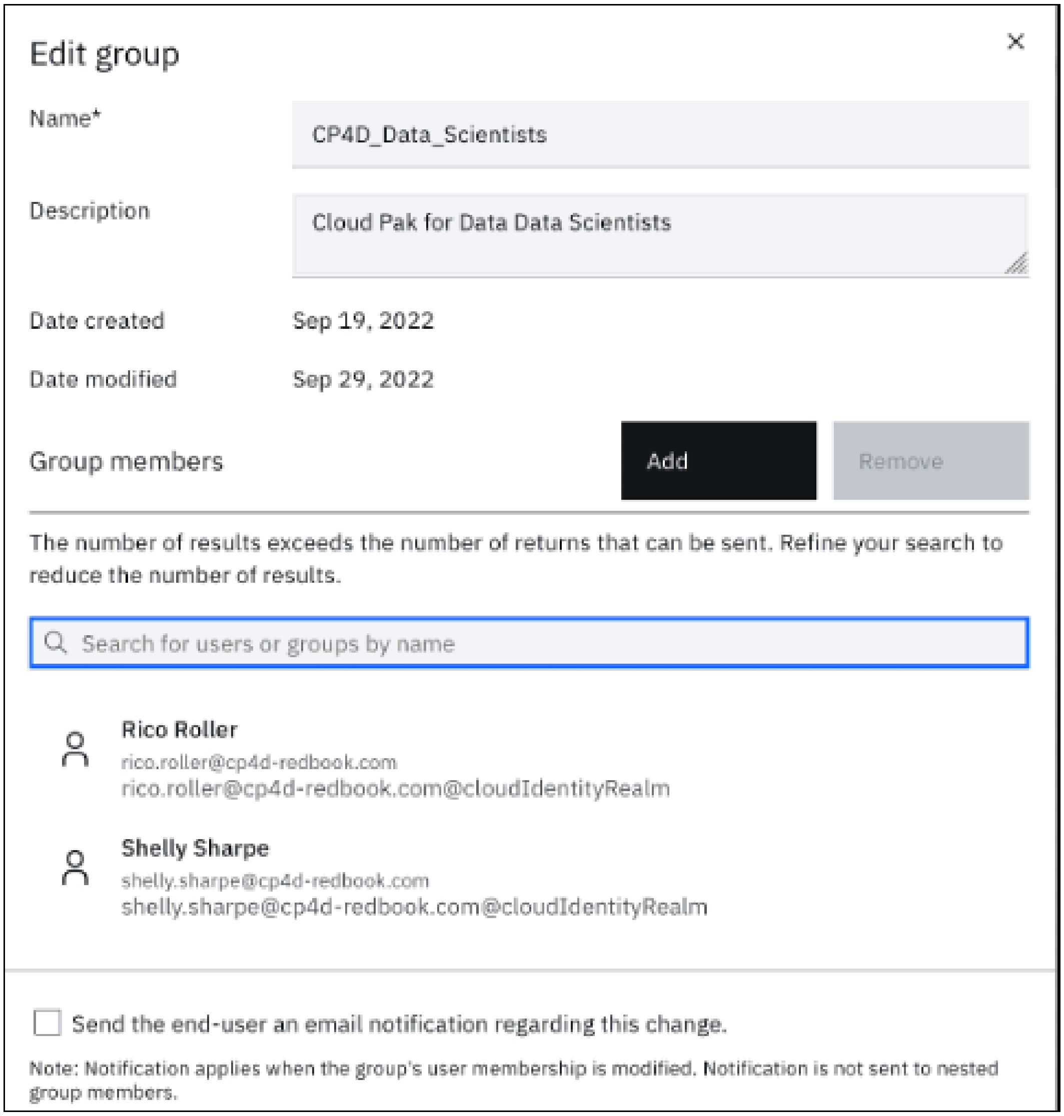

Figure 8-56 Data scientists group

We use the email addresses of the users to authenticate to ISV and to Cloud Pak for Data.

Now that you created or identified users and groups, we can create an Application to register the Cloud Pak Foundational Services client.

In the left menu, click **Applications** and then, click **Add an application**. For the Application Type, select **Custom Application** and then, click **Add application** (see Figure 8-57).

Figure 8-57 Adding ISV application

Enter the name of the application (`cp4d-redbook`), an optional description, and the company name (see Figure 8-58).

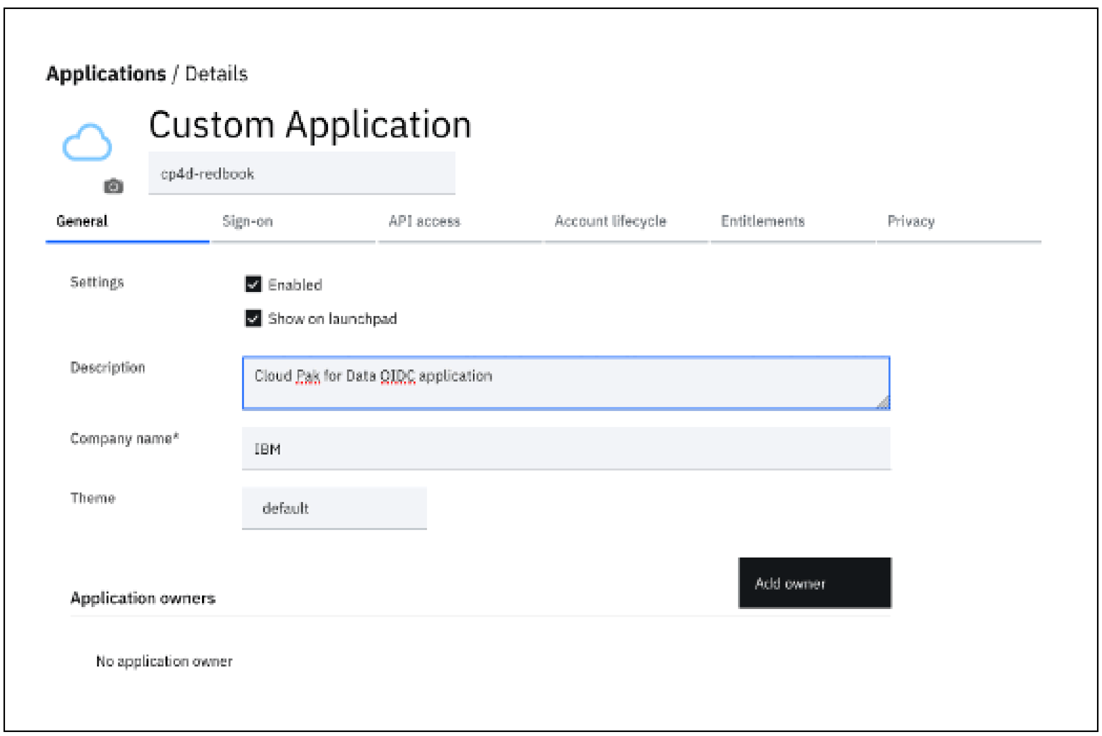

Figure 8-58 ISV custom application

Click the **Sign-on** tab to specify the Sign-on method and other attributes. In this tab, complete the following steps (see Figure 8-59 on page 659 - Figure 8-61 on page 661):

1. Select **OpenID Connect 1.0** as the sign-on method.
2. Enter the **Cloud Pak Foundational Services console** (retrieved earlier) as the Application URL.

3. Select **Authorization code** and **Implicit** Grant types.
4. Clear the **Require proof key for code exchange (PKCE)** checkbox.
5. For the Redirect URIs, use the Foundational Services console URL and append `/ibm/api/social-login/redirect/oidc_isv`, where `oidc_isv` is the name of the IdP registration you create later.
6. Scroll down and select **Generate refresh token**.
7. Select **Server** for Signing certificate.
8. Select **Send all known user attributes in the ID token**.
9. Ensure that **Allow all enterprise identity providers that are enabled for end users (2 providers)** is selected.
10. For User consent, select **Do not ask for consent**.
11. Clear the **Restrict Custom Scopes** checkbox.

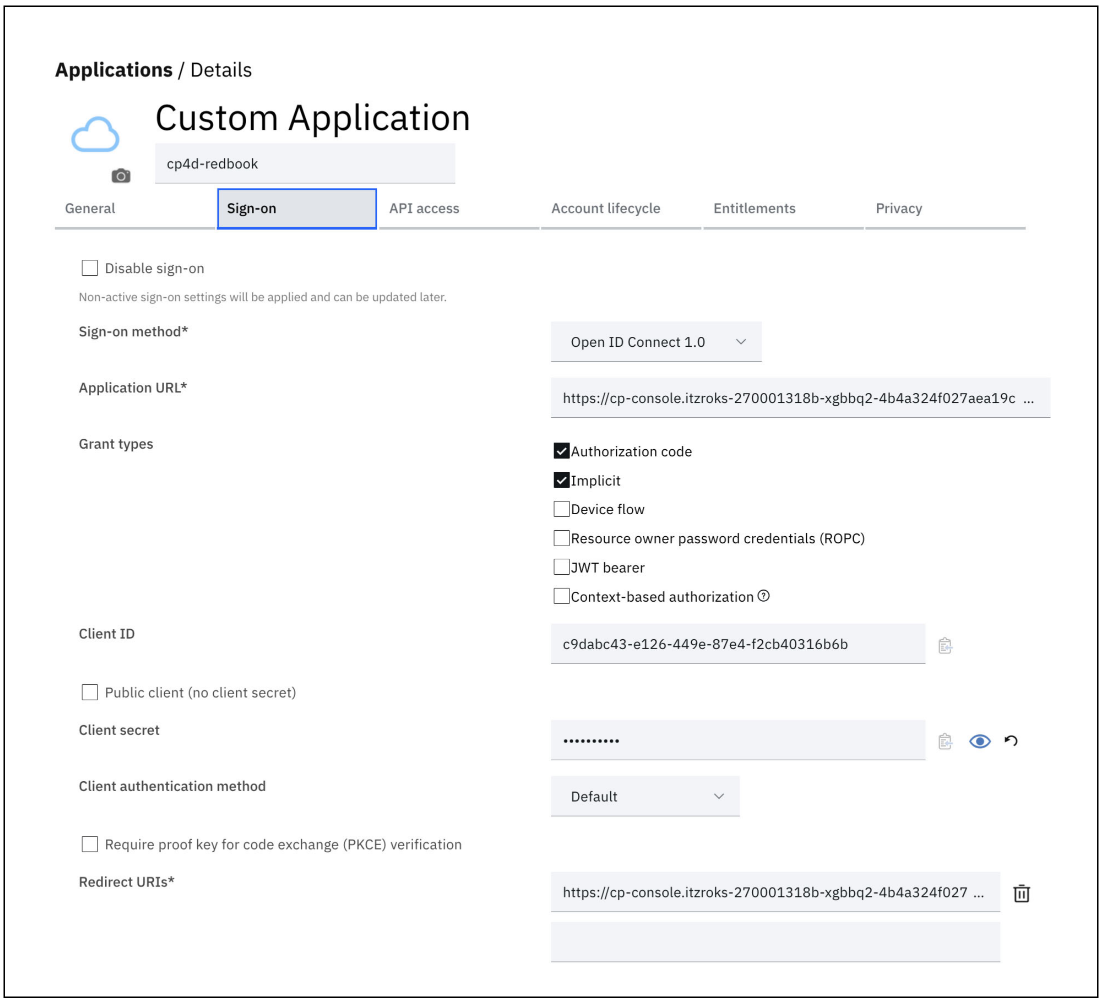

Figure 8-59 App details: Part 1

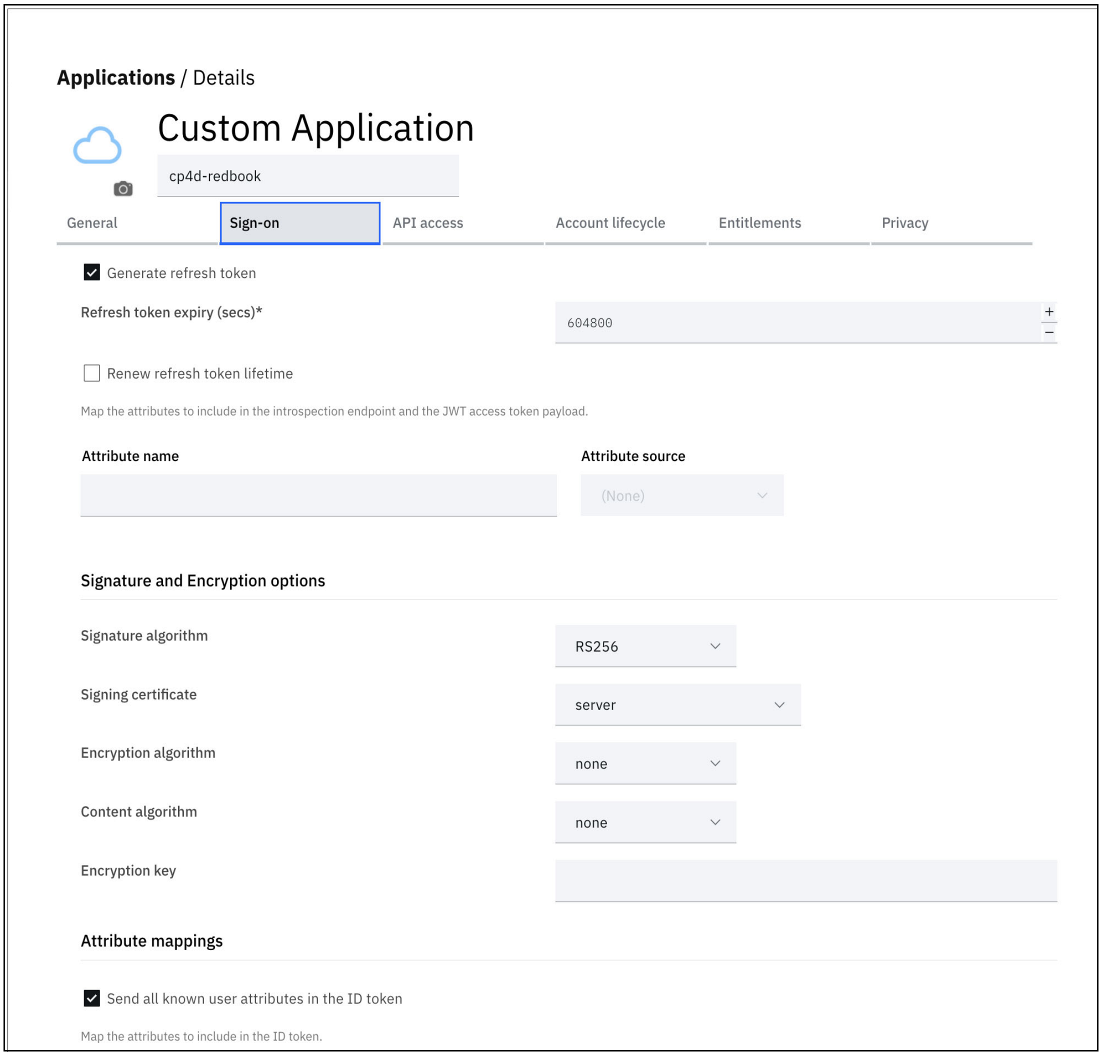

Figure 8-60 App details: Part 2

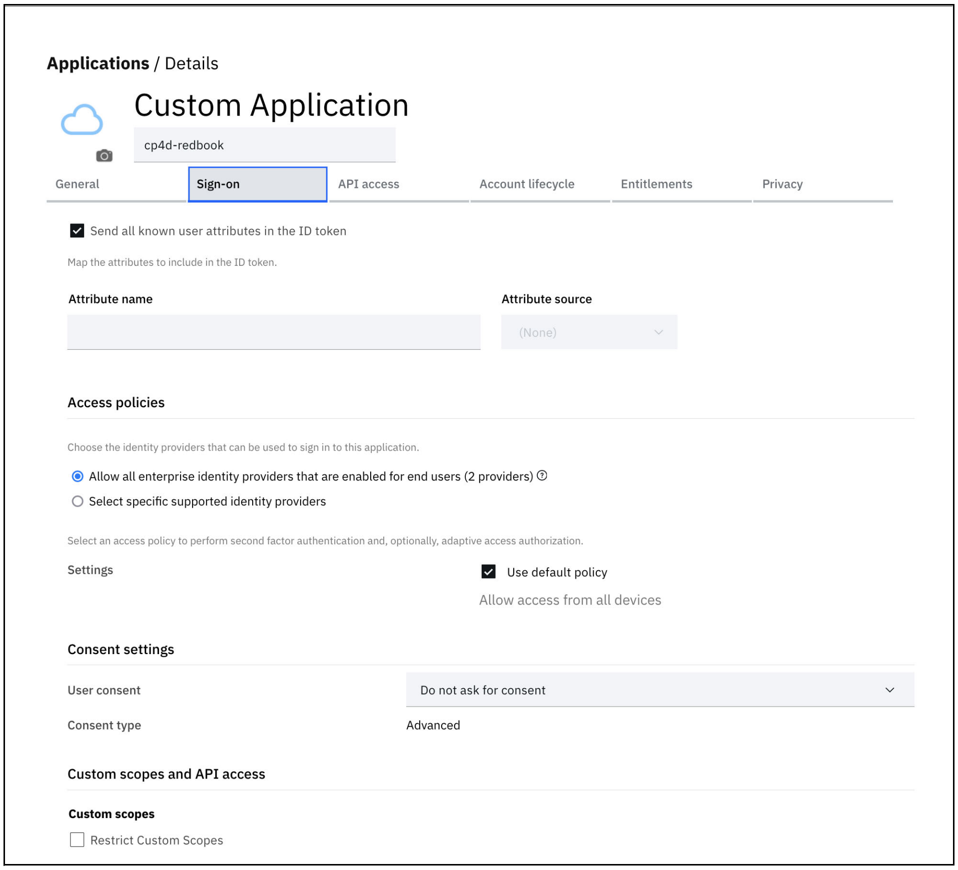

Figure 8-61 App details: Part 3

Click **Save** to persist the custom application. If you return to the **Sign-on** tab, you see that a Client ID and a Client secret were generated. Copy this information for later use. Also, if you scroll down in the frame on the right-hand side, you see the IBM Security Verify endpoint. Copy this information as well (see Figure 8-62).

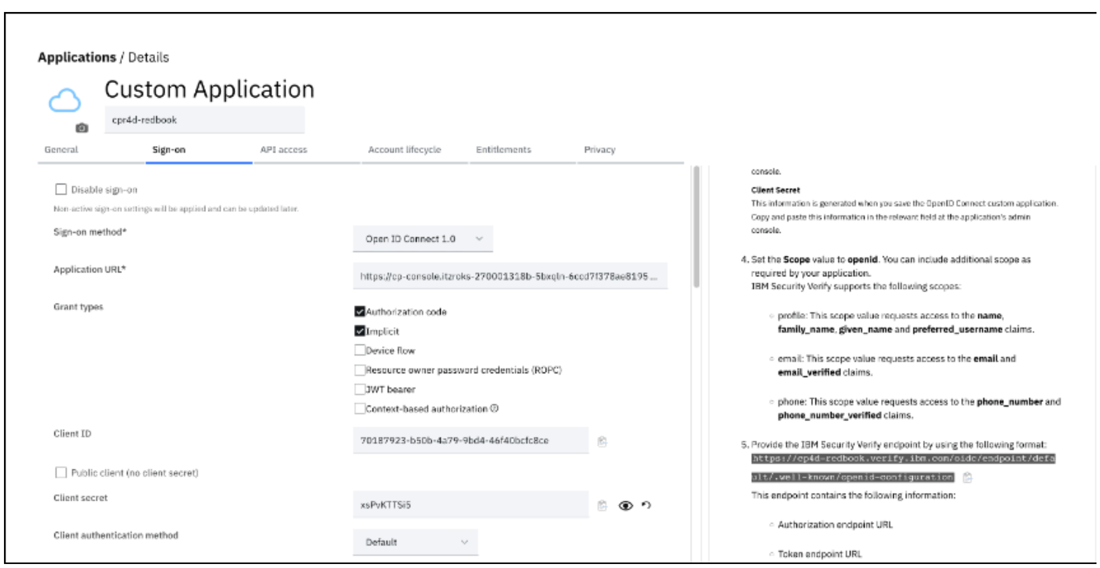

Figure 8-62 ISV endpoint

On the **Entitlements** tab, select **Automatic access for all users and groups** and click **Save**.

After creating and configuring the `cp4d-redbook` application, you are ready register the identity provider in Cloud Pak Foundational Services.

Configuring IdP in Cloud Pak Foundational Services

In this step, you connect Foundational Services as a client to the `cp4d-redbook` application. This connection is possible by using the IdP V3 APIs (see Example 8-42).

Example 8-42 Connecting Foundational Services as a client to the cp4d-redbook application

```
ISV_ENDPOINT="<your-isv-endpoint>"
OIDC_CLIENT_ID="<your-client-id>"
OIDC_CLIENT_SECRET="<your-client-secret>"

curl -k -X POST "https://${FS_CONSOLE_HOST}/idprovider/v3/auth/idsource/" \
--header 'Authorization: Bearer ${FS_ACCESS_TOKEN}' \
--header 'Content-Type: application/json' \
--data-raw \
'{
    "name": "oidc_isv",
    "description": "cp4d-redbook application via OIDC",
    "protocol": "oidc",
    "type": "IBMVerify",
    "idp_config": {
        "discovery_url": "'${ISV_ENDPOINT}'",
        "client_id": "'${OIDC_CLIENT_ID}'",
        "client_secret": "'${OIDC_CLIENT_SECRET}'",
        "token_attribute_mappings": {
            "groups": "groupIds",
            "given_name": "given_name",
            "family_name": "family_name",
            "first_name": "given_name",
            "last_name": "family_name",
            "sub": "uid",
            "email": "email"
        }
    }
}'
```

The response likely resembles the following example:

```
{
    "status": "success",
    "message": "Identity provider {oidc_isv} is successfully registered with
unique identifier 39faWJ4kSqBgA6IWKAPtB"
}
```

Although you cannot find the IdP registration from within the Foundational Services console, you can list all registrations by way of an API (see Example 8-43).

Example 8-43 Listing all registrations by way of an API

```
curl -k -X GET "https://${FS_CONSOLE_HOST}/idprovider/v3/auth/idsource/" \
--header "Authorization: Bearer ${FS_ACCESS_TOKEN}" \
--header 'Content-Type: application/json' | jq -r .
Response:
{
  "idp": [
    {
      "name": "oidc_isv",
      "description": "cp4d-redbook application via OIDC",
      "protocol": "oidc",
      "type": "IBMVerify",
      "idp_config": {
        "discovery_url":
"https://cp4d-redbook.verify.ibm.com/oidc/endpoint/default/.well-known/openid-conf
iguration",
        "client_id": "c9dabc43-e126-449e-87e4-f2cb40316b6b",
        "token_attribute_mappings": {
          "groups": "groupIds",
          "given_name": "given_name",
          "family_name": "family_name",
          "first_name": "given_name",
          "last_name": "family_name",
          "sub": "uid",
          "email": "email"
        }
      },
      "uid": "39faWJ4kSqBgA6IWKAPtB"
    }
  ]
}
```

Assigning ISV groups to Cloud Pak for Data user groups

The users in IBM Security Verify cannot log in to Cloud Pak for Data until they are assigned a role (User at the minimum). We must create Cloud Pak for Data user groups and add the ISV groups to these user groups.

In our example, we deliberately chose to differ between ISV group names and Cloud Pak for Data user group names to emphasize that these names do not have to match.

It is also possible to add users to the Cloud Pak for Data user groups. However, in most organizations with sizable number of users, this process be difficult to maintain.

In this version of Cloud Pak for Data, no graphical user interface is used to assign groups from the IdP to Cloud Pak for Data user groups. Therefore, all configurations must be done by using APIs.

Complete the following steps:

1. Retrieve the Cloud Pak for Data token that we need for the remainder of the APIs calls (see Example 8-44).

Example 8-44 Getting the Cloud Pak for Data token

```
CP4D_TOKEN=$(curl -s -k -X GET "https://${CP4D_HOST}/v1/preauth/validateAuth"
--header 'username: admin' --header "iam-token: ${FS_ACCESS_TOKEN}" | jq -r
.accessToken) && echo ${CP4D_TOKEN}
```

2. Create the three groups, each with their own roles. We must keep the group ID because this information is needed when Active Directory groups are assigned to the user groups (see Example 8-45).

Example 8-45 Creating the three groups

```
CP4D_ADMIN_GID=$(curl -s -k -X POST "https://${CP4D_HOST}/usermgmt/v2/groups" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d  '{"name": "cp4d-admins",
"description": "Cloud Pak for Data Admins",
"role_identifiers": ["zen_administrator_role"]
}' | jq -r .group_id) && echo ${CP4D_ADMIN_GID}

CP4D_DE_GID=$(curl -s -k -X POST "https://${CP4D_HOST}/usermgmt/v2/groups" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d  '{"name": "cp4d-data-engineers",
"description": "Cloud Pak for Data Data Engineers",
"role_identifiers": ["zen_data_engineer_role"]
}' | jq -r .group_id) && echo ${CP4D_DE_GID}

CP4D_DS_GID=$(curl -s -k -X POST "https://${CP4D_HOST}/usermgmt/v2/groups" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d  '{"name": "cp4d-data-scientists",
"description": "Cloud Pak for Data Data Scientists",
"role_identifiers": ["wkc_data_scientist_role"]
}' | jq -r .group_id) && echo ${CP4D_DS_GID}
```

3. If the groups already exist created, use the API that is shown in Example 8-46 to retrieve the group information, including the group ID.

Example 8-46 Retrieving the group details

```
curl -s -k -X GET "https://${CP4D_HOST}/usermgmt/v2/groups" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" | jq -r .
```

4. Now that the Cloud Pak for Data user groups are created, assign the ISV groups as members. When authenticating with ISV, the OpenID Connect (OIDC) document contains the names of all groups to which the user belongs, as configured here.

Note: We specified application attributes as described in “Creating or using an Active Directory on Azure” on page 640.

Recall that we created the following ISV groups:

– CP4D_Admins
– CP4D_Data_Engineers
– CP4D_Data_Scientists

```
ISV_ADMIN_ID="CP4D_Admins"
ISV_DE_ID="CP4D_Data_Engineers"
ISV_DS_ID="CP4D_Data_Scientists"
```

5. Add the ISV groups into the user groups: first for the Cloud Pak for Data Administrators (see Example 8-47).

Example 8-47 Adding the ISV groups into the user groups

```
curl -k -s  -X POST
"https://${CP4D_HOST}/usermgmt/v2/groups/${CP4D_ADMIN_GID}/members" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d '{
"ldap_groups": ["'${ISV_ADMIN_ID}'"]
}'
```

6. Repeat this process for the other two groups (see Example 8-48).

Example 8-48 Add the ISV groups into the user groups

```
curl -k -s  -X POST
"https://${CP4D_HOST}/usermgmt/v2/groups/${CP4D_DE_GID}/members" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d '{
"ldap_groups": ["'${ISV_DE_ID}'"]
}'

curl -k -s  -X POST
"https://${CP4D_HOST}/usermgmt/v2/groups/${CP4D_DS_GID}/members" \
--header 'Content-Type: application/json'   \
--header 'Accept: application/json'   \
--header "Authorization: Bearer ${CP4D_TOKEN}" \
-d '{
"ldap_groups": ["'${ISV_DS_ID}'"]
}'

All API calls should have a response like below.
{"group_id":10003,"_messageCode_":"success","message":"success"}
```

Logging in by using OIDC

Complete the following steps to log in to Cloud Pak for Data:

1. When you attempt to access the home page, you are directed to the generic Cloud Pak login page (see Figure 8-63).

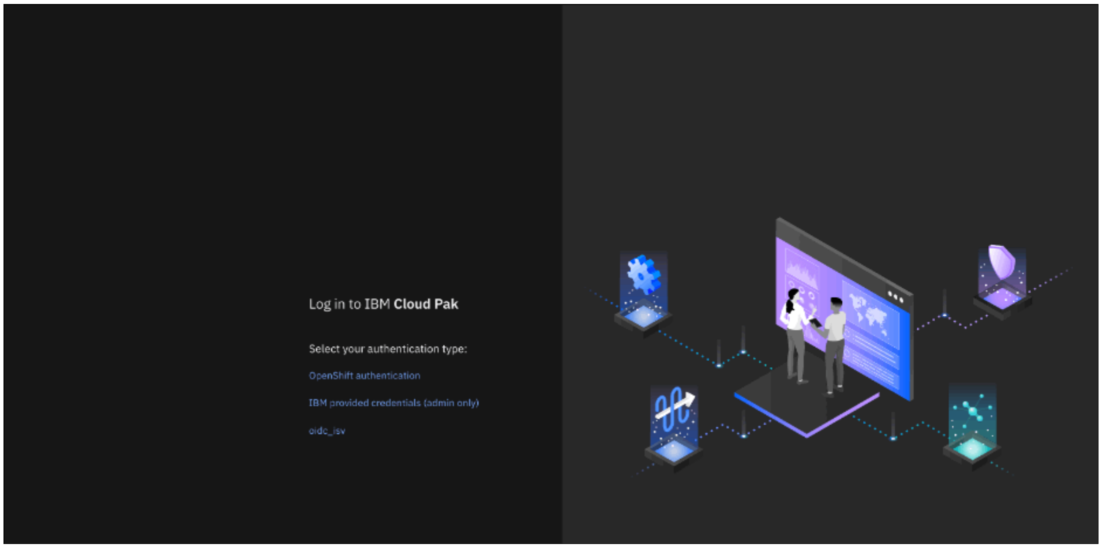

Figure 8-63 Platform login

2. Click **oidc_isv**, which redirects the browser to the IBM Security Verify sign in page (see Figure 8-64).

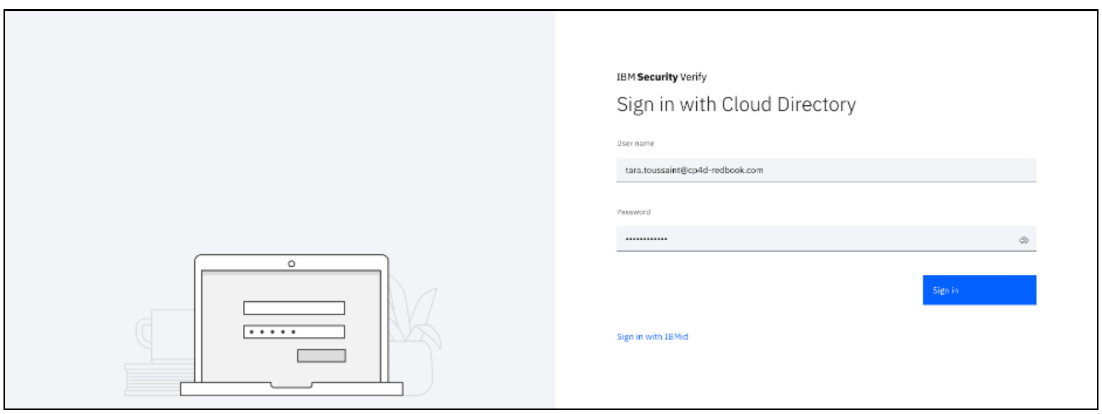

Figure 8-64 ISV login

3. Log in as Tara Toussaint, who is a member of the `CP4D_Admins` ISV group; therefore, they also are a member of the `cp4d-admins` user group (see Figure 8-65).

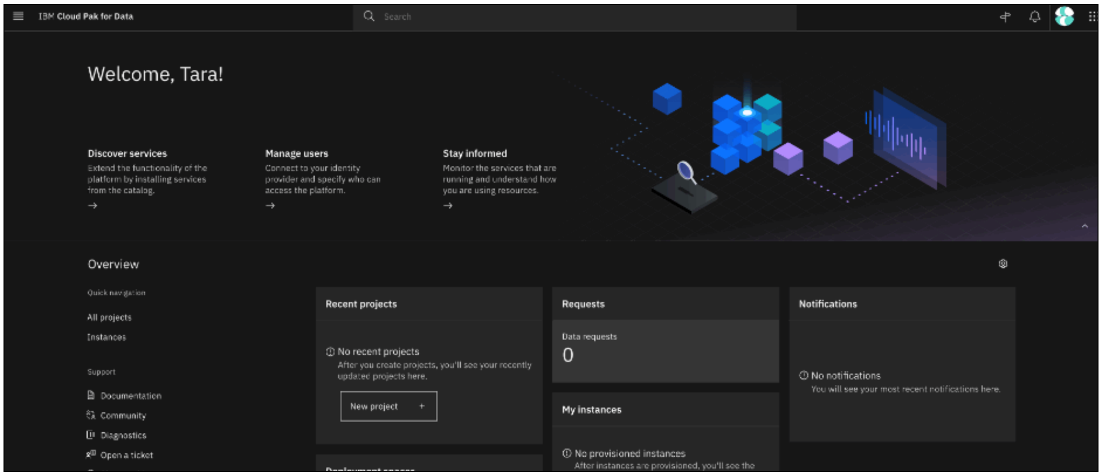

Figure 8-65 Welcome Tara

4. Because Tara is a Cloud Pak for Data administrator, you can view the user group members by clicking on navigation menu and then, clicking the User groups tab. If you click the **cp4d-admins** user group, you can see that the ISV `CP4D_Admins` group is a member (see Figure 8-66).

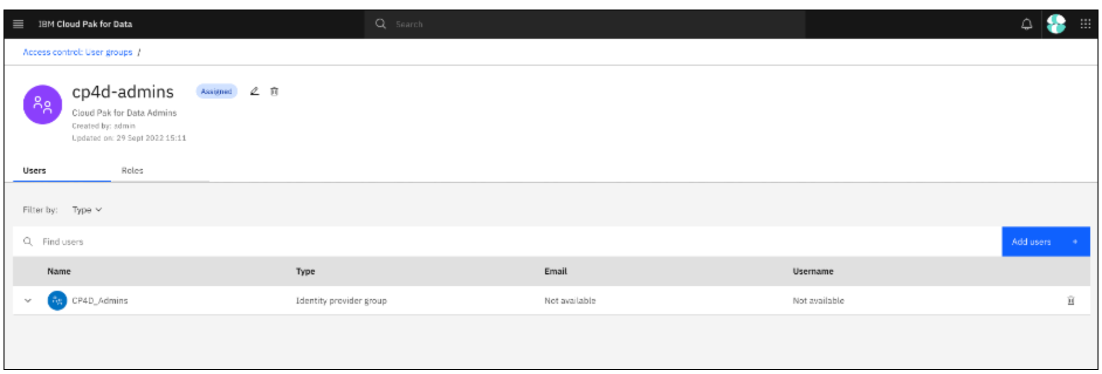

Figure 8-66 cp4d-admins user group

8.5.5 Network firewalls

A network firewall is a common security precaution that is used by most networks. It sits between the local network and the internet and monitors all of the data that is sent and received. It acts as a specialized filter to allow or deny packets.

A web application firewall (WAF) helps protect web applications by filtering and monitoring HTTP traffic between a web application and the internet. We apply the logic of a WAF to IBM Cloud Pak for Data, which is a web application.

Note: WAF typically protects web applications from attacks, such as cross-site forgery, cross-site-scripting (XSS), file inclusion, and SQL injection. A WAF is a protocol layer 7 defense (in the OSI model) and is not designed to defend against all types of attacks. This method of attack mitigation often is part of a suite of tools that creates a holistic defense against a range of attack vectors.

By deploying a WAF alongside Cloud Pak for Data, a protective shield is created between the web application (CP4D) and the internet.

WAFs can be deployed by using one of the following methods:

- Network WAF
- Host WAF
- Cloud WAF

An example of a Cloud WAF is deployed IBM CIS (Cloud Internet Services) WAF that uses CloudFlare.

8.5.6 Managing certificates

By default, the IBM Cloud Pak for Data route is secured with a self-signed TLS certificate, which is done for several reasons, including the following examples:

- The self-signed certificate is the easiest approach to configure during installation.
- The self-signed certificate is free and created by using OpenSSL for instance.
- The route is less secure in comparison to a CA-approved certificate. The self-signed certificate is used to enable HTTPS connections.

Many clients that are implementing IBM Cloud Pak for Data opt for the CA certificate track which, although it requires a domain, provides an extra layer of security for the CP4D Web Console. It is strongly recommended that you replace the self-signed certificate with a client CA certificate.

For more information about completing these tasks, see the following IBM Documentation web pages:

- Using a custom TLS certificate for HTTPS connections to the platform
- Using a CA certificate to connect to internal servers from the platform

Note: If the IBM Cloud Pak for Data instance uses a CA certificate, you must create a secret in Red Hat OpenShift that contains that certificate.

SSL certificates in the Cloud Pak for Data Web Console

By default, Cloud Pak for Data is configured with a self-signed certificate; that is, a public key certificate that is issued by Cloud Pak for Data on its own behalf.

When accessing the Cloud Pak for Data's web console by using the secure HTTPS protocol, the web browser receives the self-signed certificate as a means for the Cloud Pak for Data server to ensure that the response actually is from Cloud Pak for Data.

However, the browser cannot trust the certificate because self-signed certificates are easy to create and can be used to perform man-in-the-middle attacks. Because the browser cannot validate the certificate against any certificate authority (CA), the user must accept a security risk when accessing the web console.

Self-signed certificates generally are not acceptable on production systems, and many customers do not allow them, even on test or development environments.

Important: Use only certificates signed by a trusted CA.

The Cloud Pak for Data dashboard is a web application that is exposed through a Red Hat OpenShift route that is configured by using TLS Passthrough. Replace the self-signed certificate that is used by the web server with a PEM-encoded certificate and private key that is signed by the customer's trusted CA. This process prevents man-in-the-middle attacks and increases the overall security of the environment.

Also, because the browsers are configured with the main certificate authorities, they silently validate and accept the certificate that is presented by the Cloud Pak for Data dashboard and immediately access the web page without prompting the user for any security exception.

For more information, see this IBM Documentation web page.

Service-specific certificates

Some services can be accessed from outside the cluster; for example, Db2 and Db2 Warehouse though JDBC connections.

Configuring secure routes to services

It is recommended to create secure Red Hat OpenShift routes to the services, setting the TLS termination mode to edge (communication encrypted up to the infrastructure node that is serving the route, and unencrypted inside the cluster between the route and the service) or, if supported by the service, the pass-through mode (full encryption from the client application to the service).

Both modes require a certificate. As advised for the CPD Web Console, use only certificates that are suitable signed by a trusted CA (certificate authority).

For more information, see this IBM Documentation web page.

Certificate expiration

Certificates do not last forever (often only between a few months to one or two years at most). Certificates must be renewed before their expiry date to ensure secure communication.

Automating certificate renewal

Attempt to automate the certificate renewal process, if possible. Also, set up a notification *before* the expiry date to ensure not to end up with invalid certificates.

Distributing certificate expiry dates

Avoid the stress of having to renew multiple certificates at the same time by selecting different expiry dates for every service or cluster.

Trust customer's back-end services

Some Cloud Pak for Data services access back-end servers of the customer; for example, as platform connections to data sources or as connections in an analytics project or in a catalog. When the secure TLS/SSL protocols is used, these back-end servers identify themselves by using certificates that are signed by the customer's trusted CA.

Configuring Cloud Pak for Data with the CA's certificate

Cloud Pak for Data must be configured by using the CA's certificate to validate and trust the back-end server's certificates. Otherwise, the communication still is encrypted, but with no assurance that the response is coming from the expected back-end service.

For more information, see this IBM Documentation web page.

If the Red Hat OpenShift cluster is equipped with a CA-signed certificate for the ingress controller, you can consider exporting the PEM certificate and key from the secret that is referenced by the ingress controller and then, apply the certificates to Cloud Pak for Data.

8.5.7 Vault integration

The following types of vaults are available:

- Internal vault

  This vault is the default setting when IBM Cloud Pak for Data is installed and configured. By default, the internal vault is accessible through only the Credentials and Secrets API.

  For more information, see this IBM Documentation web page.

- External vault

  An external, enterprise-grade vault is advisable for running IBM Cloud Pak for Data in production. Some examples of enterprise-grade vault technology include CyberArk and HashiCorp vault technologies.

The recommended architecture pattern for enterprise deployments of IBM Cloud Pak for Data is to use an enterprise-grade vault to store sensitive data. However, the internal vault likely works sufficiently for a dev instance and no stringent data privacy requirements that are associated with the data are stored.

Disabling and enabling vaults

The deployment pattern with vault integration is to disable the internal default vault in favor of an enterprise-grade vault. However, when you disable the internal vault, the secrets that are stored in the vault also are deleted. Therefore, it is critical that you save your changes to a file, which is the product-configmap file in the official documentation.

For more information about vault integrations with IBM Cloud Pak for Data, see the following guides:

- Enabling vaults for the Cloud Pak for Data web client
- Disabling the internal vault for the Cloud Pak for Data web client

A

Additional material

This appendix refers to additional material that can be downloaded from the internet as described in the following sections.

Locating the GitHub material

The web material that is associated with this book is available in softcopy on the internet from the IBM Redbooks GitHub web page:

https://github.com/IBMRedbooks/SG248522-Hands-on-with-IBM-Cloud-Pak-for-Data

Cloning the GitHub material

To clone the GitHub repository for this book, complete the following steps:

1. Download and install the **git** client if it is not installed from Git.
2. Clone the GitHub repository by running the following command:

```
git clone
https://github.com/IBMRedbooks/SG248522-Hands-on-with-IBM-Cloud-Pak-for-Data.gi
t.
```


Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this book.

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *IBM Cloud Pak for Data with IBM Spectrum Scale Container Native,* REDP-5652
- *SingleStore Database on High Performance IBM Spectrum Scale Filesystem with Red Hat OpenShift and IBM Cloud Pak for Data*, REDP-5689

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Other publications

These publications are also relevant as further information sources:

- *Whitepaper for basic online backup and restore with Spectrum Scale*

  `https://community.ibm.com/community/user/cloudpakfordata/viewdocument/whitepaper-for-basic-online-backup`
- *Whitepaper for Cloud Pak for Data disaster recovery using IBM Spectrum Fusion*

  `https://community.ibm.com/community/user/cloudpakfordata/viewdocument/whitepaper-for-cloud-pak-for-data-d`

Online resources

These websites are also relevant as further information sources:

- IBM Cloud Pak for Data documentation

  `https://www.ibm.com/docs/en/cloud-paks/cp-data`
- IBM Cloud Pak for Data Redbooks domain

  `https://www.redbooks.ibm.com/domains/cloudpaks`

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

Redbooks

IBM Cloud Pak for Data Version 4.5

A practical, hands-on guide with best

SG24-8522-00

ISBN 0738460907

IBM

(1.0” spine)
0.875”<->1.498”
460 <-> 788 pages

SG24-8522-00

ISBN 0738460907

Printed in U.S.A.

Get connected

ibm.com/redbooks