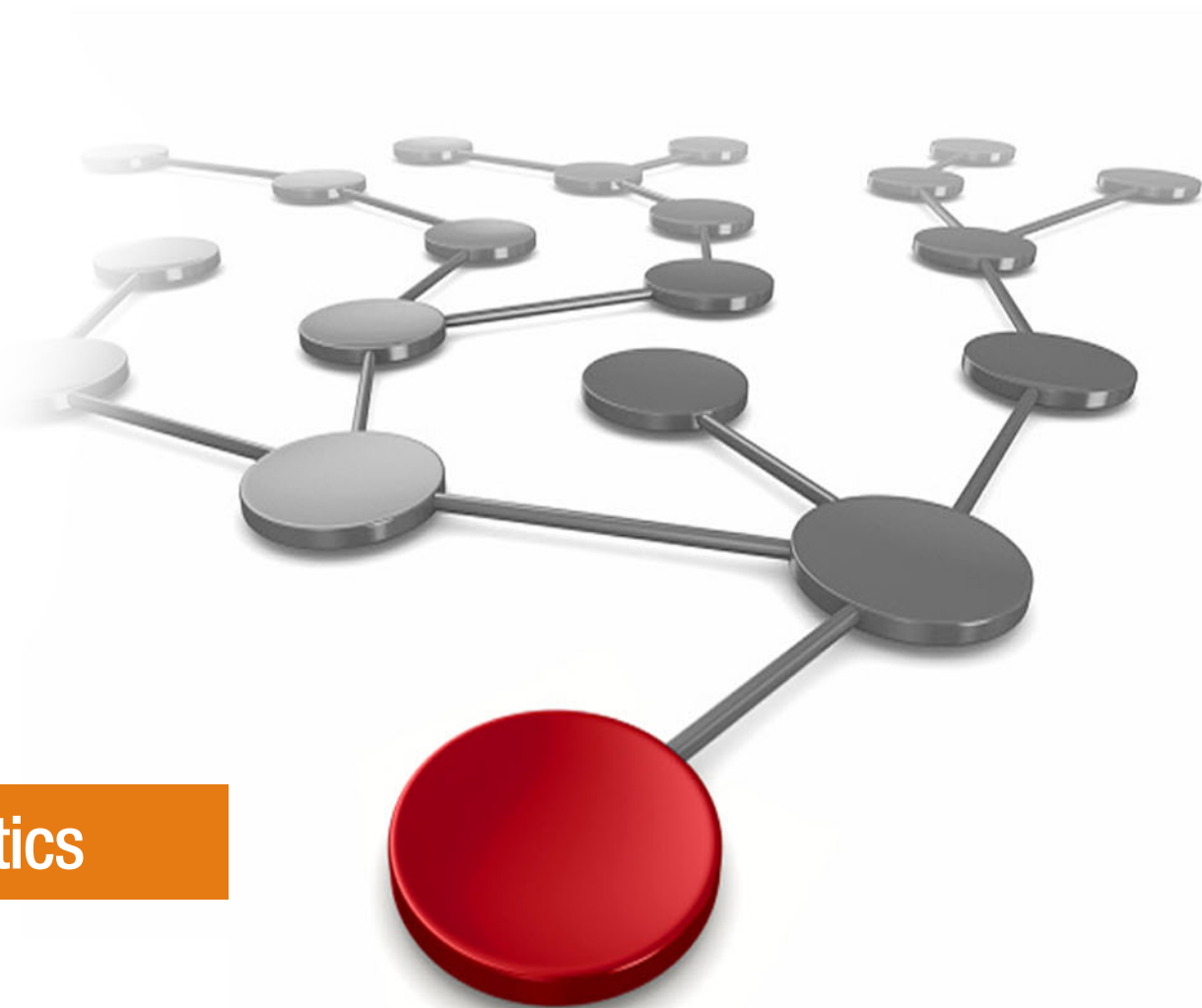


# The Journey Continues

## From Data Lake to Data-Driven Organization

Mandy Chessell  
Ferd Scheepers  
Maryna Strelchuk  
Ron van der Starre  
Seth Dobrin  
Daniel Hernandez



 Analytics





## Executive overview

In 2014, IBM® and ING published the IBM Redguide™ publication *Governing and Managing Big Data for Analytics and Decision Makers*, [REDP-5120](#). That publication laid out a vision of a governed *data lake* (referred to as the *data reservoir* at that time) and put the concept of a metadata and governance *catalog* at the heart of the data lake. The catalog controls the engines that manage the data within the data lake. It also defines the visibility and access that people and applications have to this data.

Further publications include *Designing and Operating a Data Reservoir*, [SG24-8274](#), and the [online blog](#), *Building a data reservoir to use big data with confidence*, both of which provide more details about how a governed data lake is built and operated.

Since then the concepts and design patterns have been successfully adopted by many organizations from different industries.

This new guide looks back on the key decisions that made the data lake successful and looks forward to the future. It proposes that the metadata management and governance approaches developed for the data lake can be adopted more broadly to increase the value that an organization gets from its data. Delivering this broader vision, however, requires a new generation of data catalogs and governance tools built on open standards that are adopted by a multivendor ecosystem of data platforms and tools.

Work is already underway to define and deliver this capability, and there are multiple ways to engage. This guide covers the reasons why this new capability is critical for modern businesses and how you can get value from it.

# Introduction

It was obvious from the start that the data lake was a different type of project. It was so much more than new data processing technology built around the Apache Hadoop open source platform. The data lake needs a new type of information governance, and this governance affects every aspect of the way an organization collects, processes, and governs their data—challenging traditional lines of control and ownership. However, when we began the partnership between IBM and ING, none of us realized the true extent of the impact it would have, both to an organization's operation and the way we design data driven solutions.

This guide covers details of ING's incredible transformation, driven by the idea that the way to deliver the best customer experience is to align the business around the data that supports them.

This is not just a single company's story. IBM is itself transforming to become data centric and is also working with visionary organizations from many industries on a similar journey. Despite the different industry processes and regulations and the variability in the types of data requiring focus, we have all found similar challenges and solutions. Thus, this guide is representative of many organization's experiences but told through the eyes of a global organization involved in its own transformation, and a technology and business innovation company supporting them.

## What is a data lake?

There are many definitions of a data lake used in the industry today, and so it is worth clarifying the definition used in this guide and some of the reasons for our design decisions.

The *data lake* consolidates an organization's data into a governed and well-managed environment that supports both analytics development and production workloads. It embraces multiple data platforms, such as relational data warehouses, Apache Hadoop clusters, and analytical appliances, and manages them together through a common governance program. These data platforms can be distributed geographically. Access to the data platforms is restricted to the data lake services and the engines that manage the data. Applications and people access the data through the data lake services.

The data lake allows organizations to innovate with data in a safe and properly governed way.

Figure 1 shows the ringed architecture of the data lake. The data lake repositories running on the data platforms are surrounded and protected by the data lake services that are underpinned by the information management and governance fabric.

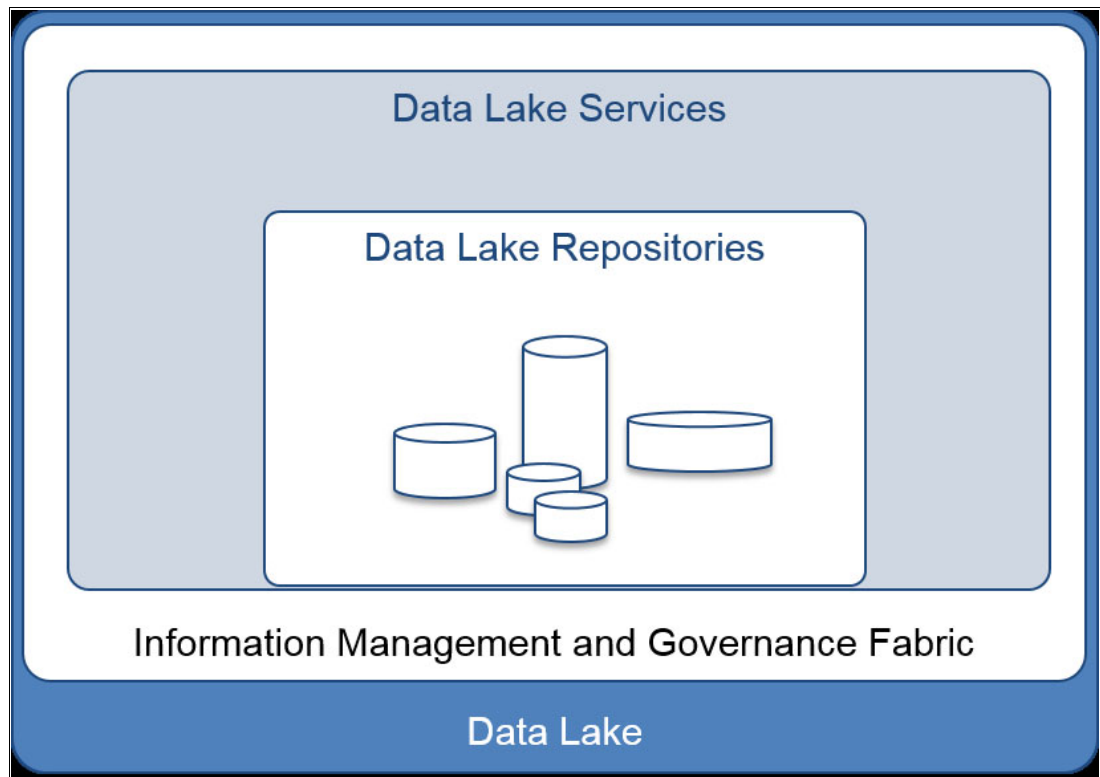


Figure 1 IBM data lake architecture

The following types of business drivers are supported by a data lake:

- ▶ Improving trust in data, for organizations where many decisions are made on gut feeling due to a lack of trust in the data presented
- ▶ The need for self-service business intelligence (BI), where new tools allow business users to produce smart reports quickly
- ▶ The need for advanced analytics, where new types of analytics demand a new approach on how to organize data of all types from both inside and outside of the organization

For the IT teams, the data lake provides the following opportunities:

- ▶ Complexity reduction: Over the years the analytical landscape might have become complex with numerous data warehouses and data marts with complex sets of interfaces. To provide agility and flexibility, these environments need to be aligned and made more consistent.
- ▶ Cost efficiency: As IT budgets came under increasing pressure, complex IT landscapes need to become more efficient and cheaper to run and maintain.
- ▶ Transparency: An ever-increasing regulatory pressure required a new approach on how to manage data and demanded an analytical platform that had governance by design.
- ▶ New sources of data: Existing structures of data were not ready for new semi- and unstructured data sources. Heterogeneous information virtualization is required to provisioning data in a simple way to the consumer.

Figure 2 shows the major groupings of data lake services.

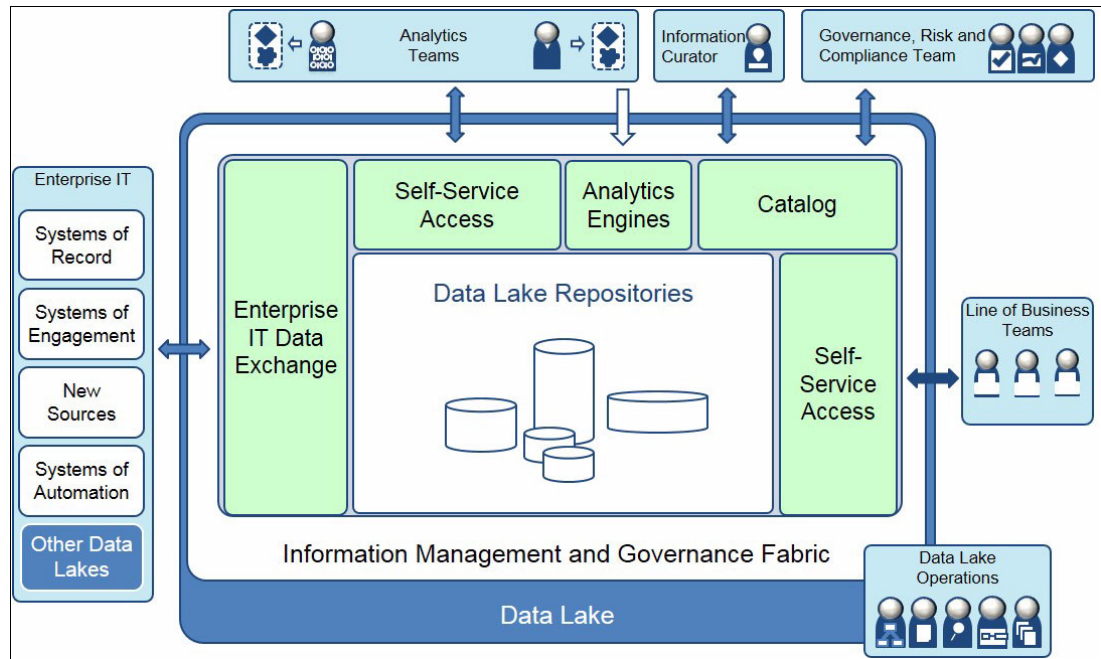


Figure 2 Key services within a data lake

The *catalog data lake service* is the heart of the data lake controlling what data people can find and access and controlling the processing of the various engines operating inside the data lake. The catalog consists of the following tightly integrated types of metadata:

- ▶ **Governance metadata:** Defines the governance program and the glossaries of business terminology that describes the types of data held and used by the organization.
- ▶ **Technical metadata:** Provides the inventory of the data assets of the organization. These data assets are used by numerous run times, such as applications, data movement and transformation engines, and databases and reporting platforms.
- ▶ **Operational metadata:** Provides transparency on the operation of the information supply chains as they copy data between the systems and data platforms, which is often referred to as *lineage*.

The Enterprise IT Data Exchange services enable data to flow in and out of the data through both batch and real-time interfaces. The data lake is a *hub*—not a data pit where data is thrown, never to be returned.

Finally many different types of people need self-service access to the data in the data lake. We typically divide these types of people into two broad groups. The data scientists and business analysts are building new analytics and executable rules that will be deployed into the production systems. They need access to raw data, just as it appears in the production systems so that they can produce analytics that work on real data. Other users tend to need data that has had some level of processing to make it simple to use in different tools. Thus, the data lake has two different self-service access points. The access points determine the scope of the data that the person can see. The metadata in the catalog determines exactly what a specific individual is allowed to see. The self-service access points enforce these restrictions.

Proper governance of the data managed within an enterprise requires more than technology. The organization's culture and operating procedures often need to change too. Strong support from senior stakeholders is required along with visible consequences for those who ignore the change. ING's story explains how the interaction with the enterprise and these senior stakeholders must dovetail with the rollout of the data lake technology.

## The value of the data lake to ING

Today, almost any company has the publicly stated ambition to become data driven. Everybody recognizes the value of data and the potential for data to both transform existing processes and to create new value. Despite famous success stories, such as Google and Facebook, who collectively make billions from data, many organizations are struggling to deliver on their strategy.

Why is it so challenging for companies to become data driven and to obtain value from their data?

### **Historical perspective: Data has never been managed as an asset**

For an organization with a history of IT that goes back decades, the move to become data driven is not an easy one. The existing IT landscape has evolved over many generations of technologies and design philosophies. From the mainframes of the 70s and 80s, the client-server architecture of the 90s, to the internet, then mobile applications and now the Internet of Things (IoT), the landscape has become larger and more varied. Integration approaches that began with component-based architecture, then object-oriented components, the service-oriented architecture (SOA), and now micro-services has added its own mix of new technologies. Then consider a history of mergers and acquisitions along with business change plus a reluctance to decommission obsolete systems. The result is a tremendously complex landscape with data flowing through hundreds or indeed thousands of applications.

The complexity of the IT landscape in many organizations is greater than any individual can understand. Almost any process in a given large-scale organization depends on dozens or more applications.

Figure 3 shows an example of the ING systems in use at the start of the data lake project for a single process in a single country.

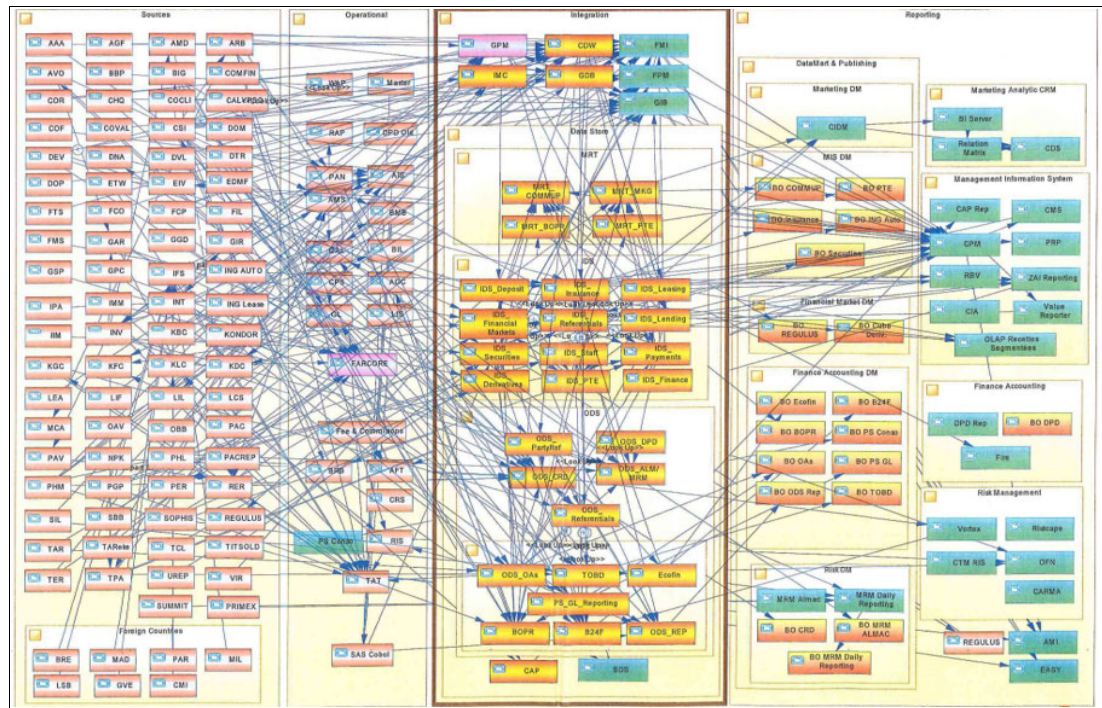


Figure 3 An example of the complexity of a single process in one country

Often these applications include a mix of purchased applications that are heavily modified to fit into the existing landscape and self-developed applications that have a history spanning decades. These application typically have their own internal data model, their own way of defining their interfaces, and their own set of experts.

### Not knowing what data is exchanged is a result of lacking documentation

One result of the evolution of the landscape is that it is not always well documented. The initial project often did deliver documentation, but subsequent changes are not well documented. Many times the true meaning of the different data elements that are exchanged is also not defined. Names used are cryptic, often following a naming standard that was based on an internal data model or that was kept down to short acronyms by choice. Full documentation that describes the data and the real definitions of the data elements rarely exists. Re-use of the interface is often impossible, which leads to the next issue.

### Many point-to-point connections built over time as “the fastest choice”

Every time a new interface is needed, it is easier to just start from scratch and build a dedicated interface than it is to build a reusable service that can cater to many use cases. So organizations end up with a huge number of point-to-point connections, all slightly different, even though the core of the data exchanged is often the same.

### Every new interface starts from scratch because we can’t re-use

As an organization begins working on a new interface, typically the data flowing through the previous interfaces is not well described and has not been made generic. Thus, we essentially must start from scratch. For ING, almost every interface also required a form of data transformation because none of the systems shared the same data model.



So, in these situations, a choice is made where to transform the data, and that choice is often determined by which project is paying or what resources are available, and a choice is made regarding what technology to use. None of these choices focus on reuse.

## **Conflicting definitions make it difficult to exchange data across domains and countries**

It also doesn't help when different domains have different definitions for the same thing.

A good example is the definition of a *customer*. For different domains, a customer can be defined differently. There is the obvious difference between a wholesale or retail customer but also regarding the moment somebody actually moves from being a prospect to a customer. Also consider does a family count as one customer or multiple customers? Must a customer be active (that is, have at least  $x$  products or transactions)? There are many other reasons the definition of a customer might not be the same throughout the system. However, after you start to exchange customer data between systems, the mismatch in definitions is not always known or understood. It can often lead to big issues in integration tests, bad data quality, and inconsistent data sent to regulators.

With this complex landscape in mind, it is good to think about the challenges you might face on top of just maintaining the existing complexity of the architecture.

## **The need to go to real time**

For most companies, the existing (legacy) data landscape is often still batch based. Many of the interfaces are using different technologies than just file transfers, but the pattern is one of the batches of data being exchanged, either once a day or at best in smaller batches. Reports are often based on old data.

## **The big data hype has led to more data that is less understood**

Even though big data is no longer the buzzword it was 5 years ago, big data is more of a reality in most companies today. The amount of data that companies generate and gather has gone up, and many companies have built big (Hadoop) clusters, both for exploration purposes and for production use cases. Many companies have a group of data scientists analyzing the data, understanding the many data sets they receive (often via emails), and examining ad hoc deliveries of all kinds of data. Some estimates suggest data scientists spend over 70% of their time on data preparation and making sense of data and less than 25% of their time on what they really need to focus on, which is creating new analytic models. And which company truly knows what data the data scientists have access to? Most data science occurs in a private sandbox, with data that is not under control or well understood.

## **Is governance a four-letter word in an agile world?**

Which bring us to *governance*, a word that seldom inspires. It is often rightfully associated with red tape, and overhead, in direct contrast with agile methodology and processes. *Agile* is all about speed, getting things in production faster, more power to the engineers, less focus on overhead. Every company aspires to be agile, to attract the best staff, and to not burden anyone with a huge governance overhead. But at the same time, regulators demand more control, especially on data.

Several years ago it was BCBS239, mainly a challenge for financial institutions. Now GDPR demands that every company with any European citizen's data must implement clear measures that control this data. And many more of such regulations will surely be coming.

## Creating the data lake architecture was a major step in our journey

ING realized this dilemma early on. The starting point for our data lake journey was the need to create an overall architecture that addressed simplifying the existing data landscape and enabled the change to a data-driven organization by giving access to data for all employees. But creating an architecture that is new and is not based on a reference that has already proved its merit is potentially high risk. Nobody wants to end up with a data architecture that is not in line with where the market is moving, that is not supported by technology in the market, and that leads to building everything yourself.

To avoid these issues, we decided to develop the architecture together with IBM, using the knowledge of IBM to challenge our thinking and to make sure we ended up with an architecture that can be mapped on real technology products but in such a way that the architecture itself is vendor agnostic. And from day one, we decided to publish our common architecture openly and not keep it to ourselves—again to make sure that we created something that is seen as a preferred practice, that can be adopted by others, and that can help steer the direction of the market.

## Selling a complex architecture is more work than creating it

After the architecture is created, it is necessary to convince all stakeholders that this is the best way forward. This task can be one of the most difficult parts of the journey. Although team members in IT might clearly understand the problem, on the business side, explaining the solution can be challenging.

Simply creating information and system architectural slides for the business stakeholders isn't always successful. It is essential to translate the story into the language of the stakeholder to make easier to understand and sell. We have found that using the analogy of a *library*, where data is represented by books, works well (as illustrated in Figure 4). The need to find your book using a catalog, the need to have common definitions to find information about the book, how to explain metadata using the analogy of a library card, all represent common methods that business stakeholders can relate to complex architectures.

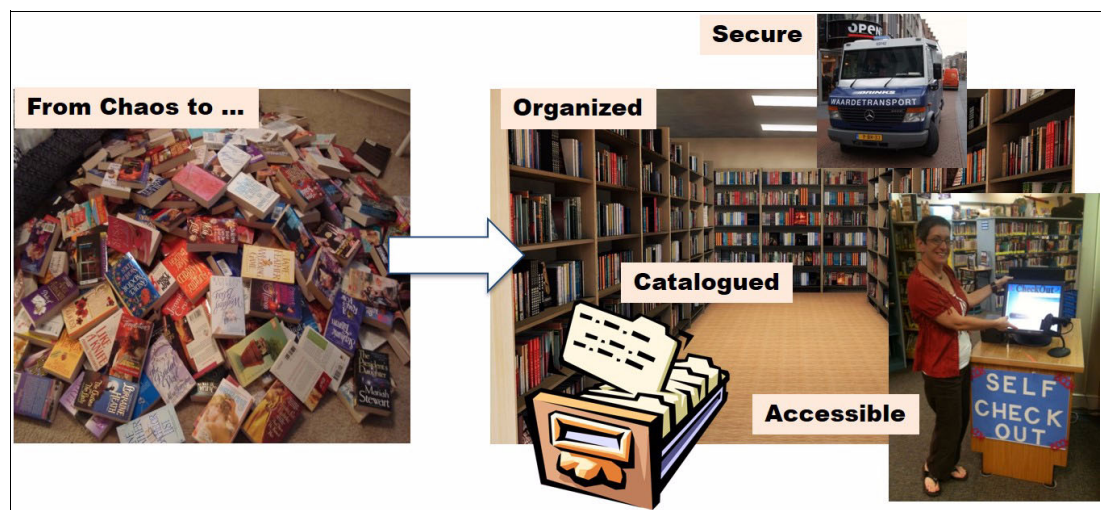


Figure 4 The library analogy to explain the role of the catalog

Another challenge is the scale of the problem and, thus, the scale of the solution. It is a journey of many years, and the real benefits come after an investment of both time and money. Helping stakeholders to understand that this is not a quick journey that will pay off in a few months or even a year is critical. The low hanging fruits are limited, and the impact on the organization is huge. But so are the benefits.

For ING, the promise of lineage, the ability to address many of the regulatory requirements through this architecture, and the need to start describing, and agreeing on, the definition of data across many domains and countries were the arguments that convinced our different stakeholders. What also helped was creating short news articles and informative videos that explained the different aspects of the architecture in easy-to-understand terms. A set of videos that explained common language, the architecture, and the need for data quality—all in the same simple way—made a huge impact on selling the story to a broad audience.

### **Building and rolling out the ING Data Lakes as a multi-year journey**

With an agreed architecture and budget, the longest part of the journey started—building the ING Data Lakes and rolling them out. To prove each key aspect of the data lake, a technology or concept was delivered first. People understand technology much better when they can see it running and experiment with it. The data lake team set up a regular delivery of demonstrations of the following different aspects of the data lake:

- ▶ The catalog as central data management point of the data lake
- ▶ The definition of glossary terms and how they relate to data assets
- ▶ Lineage across the different Data Lake technologies
- ▶ Lineage for data flowing between data lakes
- ▶ Support for ING specific metadata, such as “in the Data Lake” indicator
- ▶ Support for masking in ingestion jobs based on confidentiality data classification level

The technology proof of concept made one thing clear to the organization. The benefits of using a governed data lake with a central catalog depends strongly on a consistent implementation throughout all domains and countries. Using different technologies and standards would make it impossible to have a single metadata and governance view and lineage view.

To make sure we created this consistent data lake throughout all countries and delivered our first value quickly, we made the following decisions:

- ▶ Scope the initial delivery down to what we called the *Data Lake Foundation (DLF)*. The initial delivery included the minimal foundational components of the data lake that we needed for governance.
- ▶ Build this DLF together with a team of engineers from multiple domains and countries. This team ensured that all the knowledge of ING worldwide was brought together and created buy in from everyone in the organization from day one.

In our internal sales pitch, the delivery of this team was called the *Data lake CD*, effectively the basic data lake technology on a CD. This technology embedded the governance rules that were based on metadata settings in the catalog.

Each country took the CD and deployed their data lake from it, supported by a central team that travelled to all countries to help with the rollout. Specific places in the ingestion jobs allowed a country to customize the software, and these customizations were checked. The result was a consistent rollout of a data lake in each country, and a consistent rollout for the group. The country data lakes feed a subset of data to the group data lake, this subset determined by the (growing) demand from group level, initially contained the data that was necessary for the Financial and Regulatory Reporting.

### **After a multiyear journey, ING is benefitting from this investment**

ING's journey is not complete. There is still work to do, but we are already seeing benefits from our investment. Many countries and domains are now delivering the data from their Systems of Record to the data lakes, initially mostly focused on the data that is necessary for Finance and Risk, which is already leading to a simplification of the landscape. The number

of individual point-to-point connections per system is being reduced, and we can indeed show lineage across the entire chain from system of record (SoR) to the reports.

A second benefit we have already seen is a huge increase in the speed of delivery for changes to reports. As we have all the data in the data lakes, changes that in the past took many months to implement can now be implemented in days.

For existing systems that already are connected to the data lakes, having all the data available outside of the SoR makes decommissioning these systems also a lot easier. This process is now something we are using to facilitate a faster migration to the target landscape.

An added benefit is that we are building common platform enabled teams from different countries to collaborate and share technology, data definitions, and glossaries. This collaboration formed spontaneously as the teams recognized their shared challenges.

### **Originally an IT journey, ING is making organizational changes as well**

As part of the journey that initially was an IT-driven journey, ING appointed Chief Data Officers (CDOs) in the different domains and a global CDO that leads the Global Data Management organization. Many new data roles have also been established, business ownership of data is agreed upon, and across the organization, people recognize the value of data and are developing new business opportunities through the use of the data lake services.

A first important deliverable of this new Global Data Management organization was to establish a common language for ING, something we call the *ING Esperanto*. Previously, when we exchanged data between systems, information received from the sender in the original language had to be translated by the receiving party to be further consumed. This process was not a scalable solution, because each country would need to “know” multiple languages. In addition, assuring consistency and reliability was a challenge. Understanding these issues led to the realization that we needed to have single data-exchange language (*ING Esperanto*). *ING Esperanto* is a global glossary that describes the business terms that we use frequently within entities and creates one consistent language throughout ING.

Switching to one common language for all of ING’s countries and entities from day one would not be a feasible task. Instead, it was decided to focus on key terms and definitions that must be consistent throughout multiple locations, different divisions, and systems. This way, each ING country or entity can use their local language and configuration on local systems and, at the same time, use *ING Esperanto* for the information that must be shared throughout the entities.

Using *ING Esperanto* addressed the data-exchange problem and also helped to ensure that we understand the data, have a single source of truth, and can guarantee that reporting is accurate, consistent, and timely.

Starting with a set of common business terms, the *ING Esperanto* now is much more. We have a glossary of more than 1000 agreed upon business terms, a logical data model of these terms and their relationships, and a (flattened) physical data model to exchange data between the data lakes. Currently, we are working on a physical data model for the information warehouse in the data lake, and we are using the same definitions also to structure our APIs and events.

Today, the data lakes are a key part of delivering ING’s data-centric strategy, but there is more to do. Customer, product, and payment information used in the bank’s day-to-day operation must be trusted if the bank is to respond rapidly to requests and opportunities. Building on the success of the data lake, ING’s enterprise architecture team is expanding the use of the catalog and the governance program to include all key operational data and processes. This expansion is again raising the bar on ING’s governance and data management capability.

## The 5-level model of governance maturity

It is no secret that information governance is a complex undertaking for any organization. Potentially, it impacts the roles people perform, the tools they use, plus how and where data is processed and stored. This impact extends from deep in its internal operations, out to the touch-points where the organization interacts with its customers and business partners.

An organization's attitude to data management and its competence to execute are often plainly visible to its customers and business partners, because poor data management manifests itself as lost orders, incorrect payments, and inconsistent customer service. An increasing use of digital services makes data management a front-line capability and the information governance program a critical driver of an organization's success.

### Structuring the governance program

The data lake taught us that it is not practical to apply a rigid set of processes over all data. Information governance must be targeted, automated, and focused on delivering value to the organization.

In fact this focus on value begins with the *data strategy*. Its alignment with the business strategy ensures the correct focus is given to the types of data that is targeted along with the people and processes that are affected. Regulations affecting the organization often require accurate reporting of the business activity or a demonstration that a particular type of data is being managed and used appropriately. So they are often relevant to the governance program. These regulations along with the data strategy combine to define the governance requirements.

Aligning the data strategy with the regulatory requirements creates synergies rather than conflicts between the cost of doing business and the desire to deliver value. These synergies are reflected in the following common types of governance responses:

- ▶ Governance principles that guide the way that data should be managed
- ▶ Governance obligations that define the regulatory and corporate rules around managing and using data
- ▶ Governance approaches that provide the stake-in-ground decisions around specific best practices, tools, and related methods that are agreed upon across the organization

Governance responses define “what” the organization is going to do. Governance controls and measures then start to flesh out the “how.” They define which types of processing, rules, data collection, and procedures need to be performed on the organization's data. Often they are divided into *technical* and *organizational* controls. Technical controls are implemented by technology, and organizational controls are implemented through staff roles, culture, and procedures.

For effectiveness, the governance controls are expressed in terms of the organization's data classifications. Each data classification describes a well-defined characteristic of the organization's data. Often there are classifications that describe how confidential a data set is, the different levels of quality or confidence that should be given to the data, any legal retention requirements, and the subject area that the data set describes. For example, consider a data classification called “sensitive personal data,” which describes a data classification for particularly private data about an individual. The governance controls then describes how sensitive personal data is managed in different situations. For example, there might be a control that states that sensitive personal information must be encrypted if copied

to a portable storage device and another control that says only individuals with a specific business need must be given access to the data.

Defining governance controls using classifications has the following benefits:

- ▶ Provided the data classifications are clear and limited in number, it makes the governance program easy for employees to understand and learn.
- ▶ The governance program is stable despite new sources of data being continually introduced. New data sets simply need to be classified using the data classifications for them to find their place in the governance program.

The governance controls identify where change is necessary in the IT systems and organization. They help to scope implementation projects that encode the governance program requirements into the roles, processes and technology across the organization.

Figure 5 summarizes this structure and the concepts typically used by governance leaders. It is drawn as a pyramid to reflect the increasing scope and cost of the governance program as it rolls out over the organization. Arrows show a natural flow downwards, although a governance program is built out iteratively and relies on feedback from the more detailed activities to tune its operation to maximize business value.

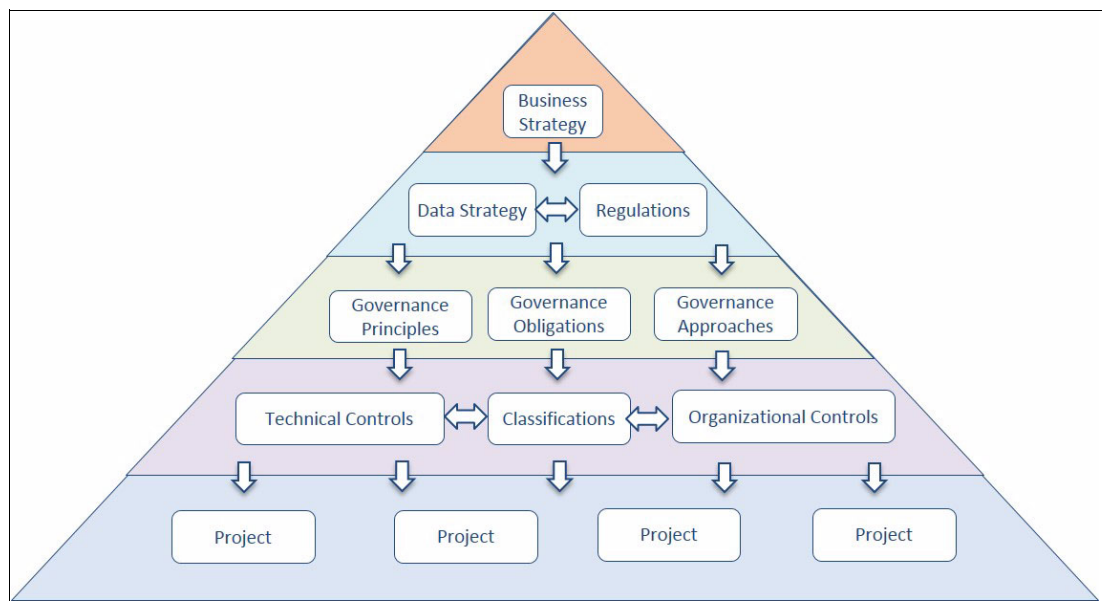


Figure 5 Structure of a governance program

## Governance rollout to create a data-centric organization

The model of governance shown in Figure 5 reflects the structure of the governance program. The skill in delivering the governance program is in ensuring the rollout of the governance activities continuously builds the data maturity of the organization, while ensuring that the organization realizes value in the changes that the governance program brings. In a modern organization, this value must be achieved on a continually shifting, and typically growing, landscape of data.

Many organizations have an information governance program, typically around their data warehouse. This program is often experienced by the business as restrictive and bureaucratic because the data warehouse needs consistent and accurate data from processes that feed

the data warehouse. Often the teams incurring the cost of governance are not those benefitting from the improved data coming from the warehouse.

Only a mandatory regulation enables such an asymmetric distribution of cost and benefit in an organization. If an organization is to be truly data centric, the balance must be restored so that activities around data occur throughout the organization, and any investment a team needs to make in managing data is returned to them through increased efficiency and capability to their operations.

Forrester characterizes this type of operation as *data citizenship*.<sup>1</sup> Their vision describes the types of job roles that people in a data-centric organization perform and the way their work delivers value to the organization. For many organizations, although they might not use the term “data citizenship,” the capabilities that Forrester describes are closely aligned with their vision. Enabling data citizenship requires a focus on data that challenges a traditional siloed, process-centric operation. It also changes the way IT systems are developed and operated.

To understand this concept, it is necessary to look behind the tools that support the new data roles to the IT technology and the processes that feed them.

## **An organization’s data landscape**

An organization typically has many types of data. First, from a business perspective, its data covers many different topics, or subject areas. For example, there is data about its customers, employees, assets, the processes it operates, its finances and risk, plus the data that is specific to its industry. Some of these subject areas appear in most systems—information about people, such as customers and employees is an obvious example. Other subject areas are heavily focused within a particular area of the business, although they receive data from other parts of the business. Financial data, for example, tends to be focused in the finance systems, but receives data from sales and procurement to balance the books.

From a technical perspective, data is formatted and stored in different ways, and that affects the type of technology needed to manage it. For example, we talk about *structured* data that is stored in databases and *unstructured* data (text, documents, and videos) that is stored in file systems and document stores.

Many IT organizations are structured around the types of technology they support. The data for a single subject area is typically distributed across this technology.

This complexity presents challenges for each phase of the governance rollout, both technically and organizationally. Each move to broaden scope of the program—either by bringing in new parts of the organization or adding new types of data—typically expands the list of technology and cultures that need to be integrated.

---

<sup>1</sup> Brief: Establish A Data Governance Journey Toward Data Citizenship”, 27th April 2016, <https://www.forrester.com/report/Brief+Establish+A+Data+Governance+Journey+Toward+Data+Citizenship/-/E-RES132243>

## Governance technology underpinning the maturity model

Governance brings its own technology into the IT landscape. This technology is typically coordinating the collection, management, and distribution of metadata. There are also new tools that support the people with specific responsibilities relating to the management of the data landscape itself. We characterize the technology and the associated activity into the following levels of maturity:

- ▶ Cataloging data.
- ▶ Defining the governance requirements and linking them to the data descriptions in the catalog.
- ▶ Automating the governance requirements in the operational environment.
- ▶ Enabling the business to manage the governance requirements for data and any changes are reflected automatically in the operational environment.
- ▶ Supporting data citizenship roles across the organization.

Each of these levels is covered in more detail in the sections that follow.

### Maturity level 1: Cataloging

Cataloging builds a list or inventory of all of the data stores and data feeds that are used by the organization. With a catalog in place, an organization knows what data it has and where that data is located.

Cataloging is a multi-stage process.

Initially the details recorded about each data store are simple, covering its name, location, short description, and owner. Ideally this cataloging is automated in the data stores and engines that create the data stores.

Next is a process of looking at the data itself and recording details, such as the structure of the data and any characteristics of the data itself. This process is called *metadata discovery* and is another area where automation is both possible and desirable. Metadata discovery can also include an initial assessment of the quality of the data. For example, are all of the values completed and do they match the agreed valid values.

The final stage of cataloging is classification, which needs human input. Typically the owner of the data or a subject matter expert needs to classify the data in the data store with glossary terms that define the data element's meaning and the data classifications that are used in the definition of the governance controls. Analytical functions can help to identify the logical type and candidate classifications for the data, but the human expert is needed to confirm the classification is correct based on their understanding of the business context around the data.

### Maturity level 2: Defining governance

The next level of maturity is where the governance drivers, requirements, and controls are linked to the catalog of data stores, typically via the classifications. This process creates an authoritative definition of how data should be governed.

Governance definitions should be created by the governance team through tools that are linked to the metadata repository that hosts the catalog of data stores.

### Maturity level 3: Operational governance

The next level is implementing the governance controls in enforcement and validations points within the IT infrastructure and providing what are called *control points* for human decision makers to confirm particular actions.



An enforcement point ensures a governance control is enforced automatically in the IT systems. For example, an enforcement point can ensure certain types of data are masked before they are displayed to particular users. Verification points test whether a particular condition is true and raise an exception record if not. They are often used for verifying data quality or to test that data coming from another organization meets agreed standards.

Control points are places where a decision or judgment needs to be made to resolve an anomaly or error in the data. A person with appropriate responsibility chooses the best course of action and their decision is recorded in an audit log, potentially approved by a colleague and then actioned.

Automation in the operational landscape provides greater coverage and consistency in the implementation. There are many runtime engines that can be configured to execute governance rules at particular points in the processing, supporting both the enforcement points and verification points. These engines include security tools, data movement engines, and data access APIs. Workflow tools support the automation of exception management and control points.

Enterprise-wide programs to use common platforms and solution platforms can reduce the cost of implementing operational governance. Nevertheless, it requires changes to the way IT solutions are budgeted for and rolled out to ensure the operational governance implementations are included.

#### **Maturity level 4: Business-controlled governance**

Business-controlled governance is where the operational governance behavior is controlled directly from the governance definitions and classifications in the catalog. So if a governance control is changed or if the classification of a data element changes in the catalog, the appropriate enforcement points, verification points, and control points automatically pick up and honor the new settings.

Many business run their digital services continuously, so the metadata for operational governance must also be continuously available. Digital service need to execute efficiently so that they cannot make multiple calls to a centralized metadata server, even if it was guaranteed to always be available.

Business-driven governance requires the systematic distribution of consistent metadata to all of the enforcement points, verification points, and control points irrespective of the platform they are operating on.

There is no single technology or technology vendor that can deliver business-controlled governance at an enterprise scale, yet if the business is to take charge of the use and governance of data, it is a critical requirement. Organizations that have achieved this level of maturity have done this through extensive investment and focus across the organization.

#### **Maturity level 5: Data-driven enterprises**

Finally, the data-driven enterprise is one where decisions are routinely made using a wide variety of data, from both inside and outside of the organization.

Effectively a data-driven enterprise expands the users of business-driven governance from a few trained and trusted individuals to everyone in the organization. Individuals own and manage data and develop analysis and visualizations of data that is sharable. They collaborate around the use of data, provide feedback opinions, and share knowledge about the data they are using.

This process means that the catalog and access to data needs to be a part of the toolset that is available to all employees. Appropriate governance controls can enforce data protection

and auditing for ad hoc data requests and usage. The organization has a culture of data custodianship and ownership, and individual employees are trained in appropriate use and governance of data.

This process also requires that most digital services and tools are used by the organization to make use of the same metadata to locate, access, and govern data. The challenges for an organization to reach this level of maturity are similar to level 4, but they are magnified by the increased variety of technology that needs to support the shared metadata.

## Driving the governance program

As an organization drives up its level of governance maturity or broadens the coverage of its data, the breadth and depth of the technologies that need to work with a consistent view of metadata grows, as illustrated in Figure 6.

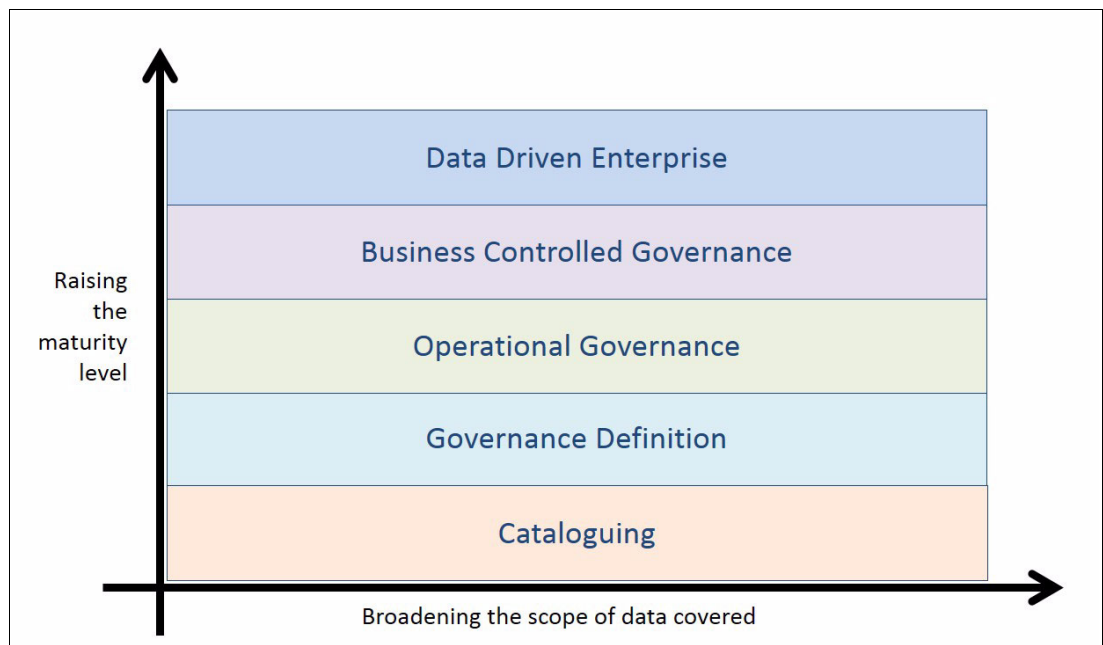


Figure 6 The expanding scope of governance

Each maturity level also expands the number of people who need to be a part of the governance program. This growth takes influence and leadership to make the necessary organizational changes.

## The open metadata and governance project

The governance maturity model demonstrates how rapidly the variety and breadth of technology that needs to be integrated with the information governance program grows. This integration needs to exchange metadata covering the following topics:

- ▶ Facts about the data stores and the condition of the data within them
- ▶ Instructions about how to process each type of data
- ▶ Details regarding any actions taken to improve or correct the data or processing

Making this work requires a common set of APIs, protocols, and message formats for technology to implement.

The standards organizations recognized this need and have been busy. There are literally thousands of defined metadata standards. However, each standard covers a small aspect of the data landscape—typically focusing on a particular type of data or style of processing. Even standards for metadata repositories cover only the capture of details about the data sets and the tagging with business terms. We need to build upon the excellent set of standards we have today and knit them together into a metadata and governance fabric that addresses all of the layers in the maturity model.

Paper documents of this integrated standard are necessary but not sufficient. Organizations need working software to deploy immediately.

A proprietary solution from a single vendor is not going to get the industry traction, but open source, with an open, commercially friendly license can provide the base for open, data-centric technology.

The open software for metadata must be deployable into a wide range of operational environment from IoT solutions that operate in remote locations and physical assets, such as cars, to edge servers and cloud platforms and across the wide range of technology already deploying in enterprises today. It must not rely on a central point of control or central server for metadata. The protocol must be peer-to-peer, supporting the specific needs of a local community of data users and their use cases, whilst embracing the greater needs of the enterprise. Finally it must accept that there is a wide variety of technology that must play in this ecosystem, many from different, competing vendors. It must allow market differentiation for these vendors and innovation whilst ensuring metadata interoperability.

A challenging set of requirements, but they form the basis of our strategy for open metadata and governance.

## Choosing Apache Atlas

Teaming up with Hortonworks, we chose the Apache Atlas open source software project to build the open metadata and governance software. Apache Atlas already has an extensible metadata type system and deep integration with key components in a Hadoop cluster to automatically gather metadata about the Hadoop data files and the processing around them. This architecture makes it an excellent base for the open metadata and governance reference implementation.

The initial software development phase includes the following parts:

- ▶ The Open Metadata Repository Services (OMRS)

OMRS provides federated queries and peer-to-peer metadata exchange among a group of metadata repositories (called a *cohort*). The metadata repositories can come from different vendors that can each still offer their own proprietary APIs and tools. Typically these repositories are owned and deployed by different parts of the organization for their specific project needs.

For example, the governance team might have bought a governance tool that includes a metadata repository. The team that builds and runs the data lake will have their own metadata repository that supports the data lake catalog, and another team might have a metadata repository that is integrated with their ETL tools that is used to supply data to the data lake.

Figure 7 shows the integration of these three metadata repositories through the open metadata technology. Each repository has a deployment of the OMRS, which might be embedded in the repository's server or can run as a repository proxy along side the metadata repository. The deployed OMRS components communicate through a topic to exchange details of each other's capabilities and REST API network address.

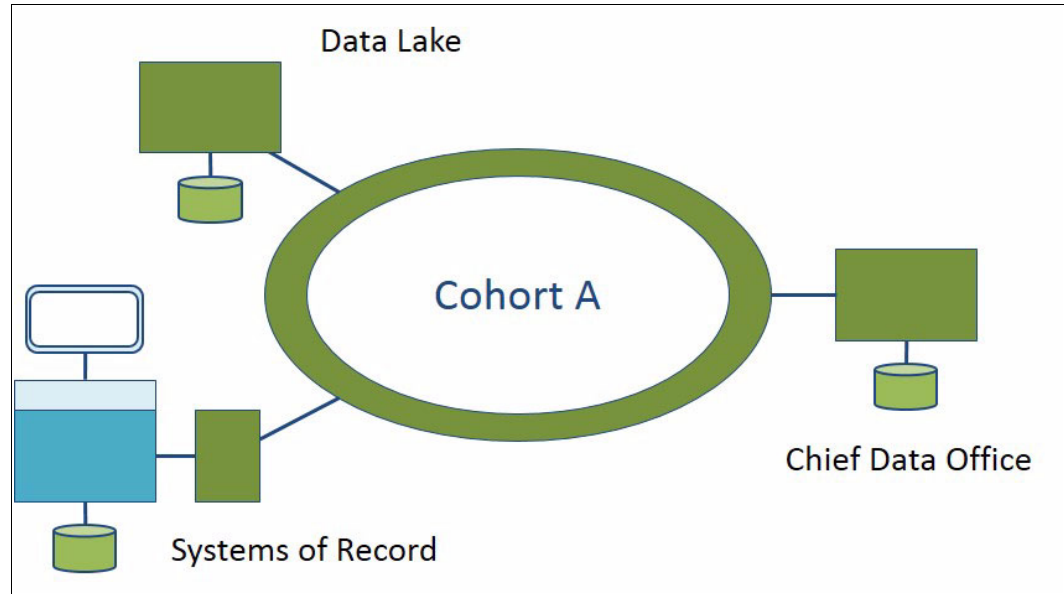


Figure 7 Simple cohort

After the registration process is complete, each OMRS can issue queries across all of the repositories through a single request. In the background, the metadata repositories are also exchanging copies of some metadata to ensure a copy of the metadata used frequently by their local users is stored locally. This exchange of metadata is designed to improve performance and availability of metadata for all users.

► The Open Metadata Access Services (OMAS)

The OMAS provides specialized APIs and events for data tools, data processing engines, and governance tools. For example, there is a specific OMAS for the catalog, another for data movement tools, another for data science tools, and another for security enforcement points. Each OMAS has a REST API and an event exchange protocol for asynchronous integration.

These services simplify the access to the broad range of technologies needed to support the higher levels of governance maturity needed by a data centric organization.

Each OMAS retrieves and stores open metadata through the OMRS so it has access to all connected metadata repositories. The OMRS can be running in its own server (called a *repository proxy*) or embedded in a vendor's metadata server.

Figure 8 shows the OMAS enabled for the data lake catalog and the governance team. The ETL tool supporting data movement to and from the systems of record (shown in blue) are using the user interfaces (UIs) and services from the proprietary tool.

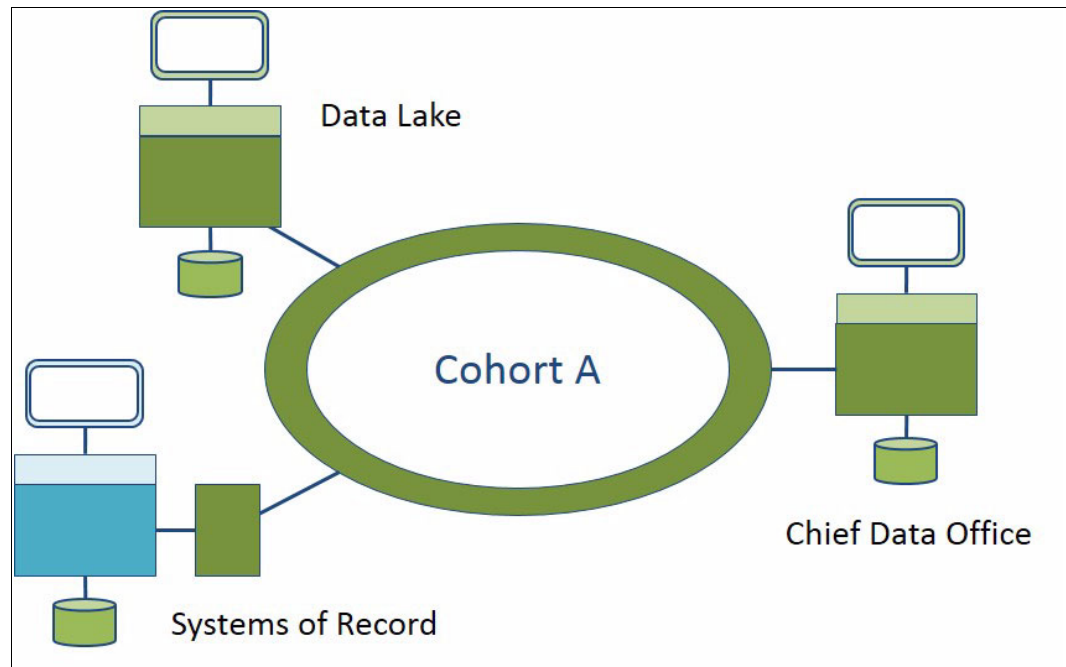


Figure 8 Simple cohort with OMAS

Many organizations are decentralized and siloed to allow them to scale. The OMRS enables a server to connect to multiple cohorts. Thus, separate cohorts are running in decentralized parts of the organization and can be linked to create an enterprise view for, say, the enterprise Chief Data Office.

Figure 9 shows the metadata repository for the Chief Data Office linked to two cohorts—one for each division of the enterprise. The OMAS services for the Chief Data Office can potentially access all metadata from both divisions. Whereas the other deployed metadata repositories can see metadata only from their own cohort.

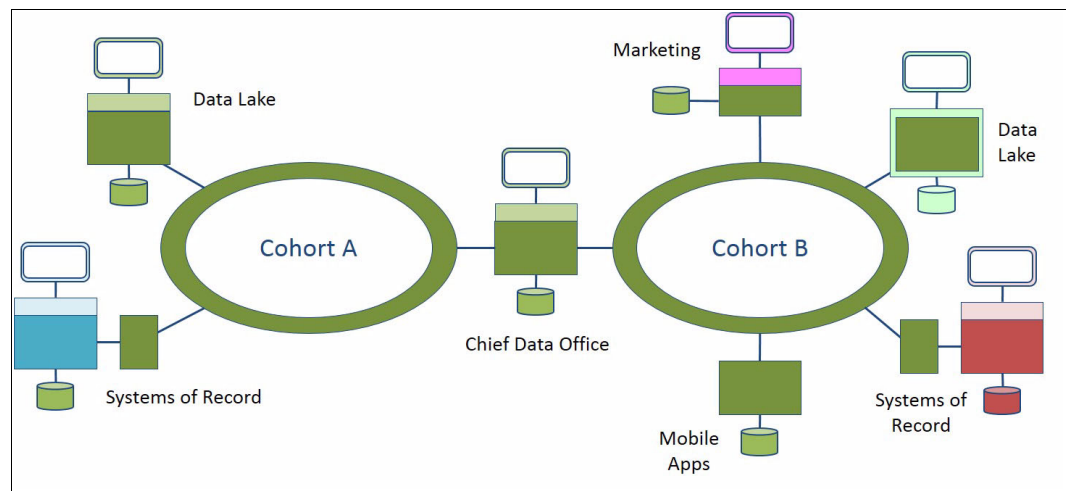


Figure 9 Multiple cohorts

## Future developments

The OMRS and OMAS focus on providing access to open metadata. After they are complete, the next phase focuses on automated metadata discovery and open governance services, such as stewardship and exception management.

## Building a broader consortium

Apache Atlas provides the software platform for the open metadata and governance capability. However, there is more to do to encourage adoption of these standards.

The ODPi is an organization with experience in driving standardization across multiple Hadoop vendors. They are turning their attention to open metadata and governance. There is a new ODPi Data Governance PMC that is driving the following initiatives:

- ▶ Compliance tests and integration guidance for vendors who adopt the open metadata and governance standards.
- ▶ A community of governance practitioners who are sharing experiences to build open metadata content packs that describe common regulations, glossaries, and approaches using the open metadata formats. These packs will be consumable from vendor tools who adopt the open metadata standards. The aim of the packs is to accelerate the development and rollout of an organization's governance program.

Both of these work streams help to build the adoption ecosystem around the Apache Atlas base.

Finally we need the buy-in of the standards organizations that have defined the base standards that underpin open metadata. Our first set of discussions has been with The Open Group. Their Open Platform 3.0 standards are closely aligned with the data-oriented solutions that open metadata and governance is focused on. One of their new standards is particularly important. The Open Group's Open Data Element Framework (O-DEF) standard provides a data naming framework plus numeric identifier for data items based on their meaning. The resulting O-DEF tag can be used to identify how data elements from different sources map to one another.

By extending the scope of the defined O-DEF tags to cover the full scope of the open metadata model we have a standardized way to show how each of the existing open metadata standards maps to the open metadata data model. The practical value of this work is in simplifying the effort of building bridging software needed to link tools that use the existing metadata standards with the wider open metadata ecosystem. However, it also exposes the overlaps and gaps in existing standards landscape that can lead to further standards work in the future.

## Progress to date

To date we have a significant portion of the open metadata capability coded and available in Apache Atlas. There is a full-time cross-company team that is focused on its delivery and maintenance going forward. Through our partnerships and progress in developing the software in Apache Atlas, the understanding and buy-in to the open metadata and governance strategy is growing.

IBM, ING, and Hortonworks are key players in all aspects of both Apache Atlas and the ODPi, and we have significant strategic initiatives built on their success.

## Unified governance

Open metadata and governance is the cornerstone of IBM's unified governance capability.<sup>2</sup> Effective data and governance underpins both cloud and cognitive computing (also known as artificial intelligence).

Unified governance is distributed to reach all systems that work with data; from the small IoT monitoring station to the popular global public cloud service. It is also connected into a single logical platform that enables the catalog to draw from metadata collected in many different systems and the governance to be optimized and distributed to the data access points and management engines.

IBM's unified governance offers the ability to govern data consistently, irrespective of how it is stored or structured. It supports data in an organization's data centers and in IBM cloud services. Its integration and internal use of the open metadata and governance standards extends its reach beyond IBM software and platforms to encompass all of an organization's data.

IBM's unified governance drives to support organizations as they raise their governance maturity on the journey to becoming a data driven organization.

## Conclusions

A data-driven organization is more agile and responsive to its customer needs. The transformation from a traditional hierarchical organization to a data-driven organization is on many CEOs' agenda, but it is a challenging undertaking, affecting processes, organizational roles and careers, and long cherished customs and practices.

Technology can provide a major boost to this transformation if it is backed by comprehensive and proven open standards. These open standards need to integrate into the technology platforms, governance engines, and repositories as well as the data-driven tools that are used day-to-day in the business.

This guide laid out the strategic approach for open metadata and governance that has emerged from our experiences in transforming organizations to become data centric. It also covered the activities in progress to make this strategy a reality.

At this stage, although much of the vision and base capability is in place, there are plenty of areas that need specific focus and contribution. Open metadata and governance is an ecosystem that will grow in strength through contribution. It will enable organizations to work confidently with data, making conscious choices in their use of data and the direction of the data strategy.

As digital technology transforms every aspect of our lives, data volumes will continue to grow. Every organization and individual is impacted with both new opportunities and obligations. Strong information governance is in everyone's interest but it should not be as hard as it is today. Open standards embedded in tools and technology platforms that automate the proper cataloging, classification, and use of data is key to maximizing the value we can get from this data. We have laid out the road map and started the journey. The door is open for you to join.

---

<sup>2</sup> IBM's Unified Governance, <https://public.dhe.ibm.com/common/ssi/ecm/im/en/imw14949usen/analytics-analytics-platform-im-white-paper-external-imw14949usen-20170719.pdf>

## Next steps

If you are interested in learning more:

- ▶ Talk to an IBM representative about IBM's contributions to driving open metadata and governance through its Unified Governance strategy.
- ▶ Sign up developers to contribute to the [Apache Atlas](#) open source project.
- ▶ Join the [ODPi](#) to:
  - Drive the adoption of the open metadata and governance standards by your vendors.
  - Work with fellow subject matter experts in developing governance collateral to accelerate an organization's transformation to a data-centric organization.
- ▶ Join [The Open Group](#) to influence the open standards related to open metadata and governance.

## Authors

This guide was produced by a team of specialists from around the world working at the International Technical Support Organization (ITSO).

**Mandy Chessell** is an IBM Distinguished Engineer, Master Inventor, and Fellow of the Royal Academy of Engineering. Mandy is a trusted advisor to executives from large organizations, working with them to develop their strategy and architecture relating to the governance, integration, and management of information. She is also driving IBM's strategic move to open metadata and governance through the Apache Atlas open source project. Mandy is the leader of the ODPi Data Governance PMC.

**Ferd Scheepers** is the Chief Information Architect for ING. He has worked for ING since 1995 and in that time has held many different roles, ranging from Lead Architect for Development Environments, Business Intelligence and Middleware, to Enterprise Architect for Payments, Transaction Services, and Customer Centricity. In his current role, Ferd is responsible for the Data and Information Architecture for ING Global.

**Maryna Strelchuk** is an Information Architect at ING, leading a team of developers who are building key components in Apache Atlas. She is enrolled on the ING International Talent Program (ITP) and has experience as a Data Scientist as well as in software development.

**Ron van der Starre** is an Information Management Architect for the IBM Software Group in the Netherlands. He has over 20 years of experience in IT with the last 15 years within IBM. He started within the services organization as an IT Specialist for packaged solutions and after a couple of years switched to work as a consultant in the area of business process redesign.

**Seth Dobrin** is Vice President and Chief Data Officer for IBM Analytics. He is responsible for internal data, data science, and digital transformation as well as influencing offering and sharing experiences with clients. Seth is also on the board of the ODPi. Previously he was the Director of Digital Strategies at Monsanto.

**Daniel Hernandez** is Vice President and Offering Management leader in IBM Analytics. He leads a talented team that builds, operates, and continuously improves the products and services in IBM's data, unified governance and analytics portfolio.



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

- ▶ Patrick van der Drift IBM, Netherlands
- ▶ Graham Wallis, IBM, United Kingdom
- ▶ Nigel Jones, IBM, United Kingdom
- ▶ David Radley, IBM, United Kingdom
- ▶ Dan Wolfson, IBM, USA
- ▶ Christopher Grote, IBM, United Kingdom
- ▶ Mike Ruland, IBM, USA
- ▶ Susan Malika, IBM, USA
- ▶ Mike Nicpan, ING, Netherlands
- ▶ Kees van de Fliert, ING, Netherlands
- ▶ Yao Li, ING, Netherlands
- ▶ Bogdan Mihail Sava, ING, Romania
- ▶ Daniela Valentina Otelea, ING, Romania
- ▶ Madhan Neethiraj, Hortonworks, USA
- ▶ Srikanth Venkat, Hortonworks, USA
- ▶ Alan Gates, Hortonworks, USA

## 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 ITSO 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](http://ibm.com/redbooks/residencies.html)

## Stay connected to IBM Redbooks

- ▶ Find us on Facebook:  
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:  
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:  
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new IBM 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>



# 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

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at <http://www.ibm.com/legal/copytrade.shtml>

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

Redbooks (logo) ®  
IBM®

Redbooks®  
Redguide™

The following terms are trademarks of other companies:

Other company, product, or service names may be trademarks or service marks of others.





REDP-5486-00

ISBN 0738456667

Printed in U.S.A.

Get connected

