Managing Data in Today’s Information Overloaded World

An IBM® Redbooks®
Point-of-View publication by the
IBM Academy of Technology

By Dan Wolfson, IBM Distinguished Engineer, **Thomas Pflueger**, IBM Distinguished Engineer, and **Vincent Hsu**, IBM Fellow

Highlights

As the amount of data for your business grows, your enterprise can stay competitive by addressing the following trends:

- Increased capacity for memory and devices at a lower cost
- Hardware virtualization that allocates and reallocates hardware resources in response to workload demands
- Software that takes advantage of the evolution in hardware technologies
- Big data that combines emerging technologies to support massively distributed computations

Emerging trends in information infrastructure technology

The landscape of today’s information infrastructure is continually changing. The amount of data that is being collected is increasing, and organizations are finding new ways to analyze, evaluate, and use the information that they collect. In addition, hardware trends are making it cost effective to capture new information sources and to automate data management.

The ability to derive new meaning from data and to use this insight within the enterprise is perhaps the most important trend that is shaping the information infrastructure. But how will this trend continue to play out?

The *Internet of Things* continues to automate the collection of huge quantities of new kinds of information. The number of mobile devices continues to proliferate and provide access to both existing and novel applications and data. In addition, businesses continue to employ analytics and artificial intelligence to understand and use this cornucopia of information. As data workloads continue to change and grow, systems will evolve to accommodate them.

A growing trend is that this overabundance of information is being processed and managed by various types of software and hardware. The expectation is that these different types of software and hardware must work together to accumulate, preserve, manage, protect, and use the information that is now the lifeblood of the enterprise. This combination of systems presents both challenges and opportunities. When these systems are combined well, they provide an organization with reliable, safe, trustworthy, and fast access to information.

Thus, storage and information management are at the core of many emerging trends that are driving change. Understanding these trends is key to gaining a competitive advantage for today’s businesses. Information management capabilities continue to evolve across the full spectrum of software and hardware. They offer increased performance at a lower cost and change what is affordable in commercial computing systems.

In the hardware arena especially, falling prices and increasing capacity for memory devices are reshaping the storage industry. Solid-state drives (SSDs), large amounts of RAM, and new I/O technologies influence raw system performance. The traditional paradigm that a system structure is built by a processor, cache, and global memory (dynamic RAM (DRAM)) is now being challenged by new trends.

In addition, system virtualization can improve the utilization of hardware, software, networking, and storage resources. It can treat the physical resources as pools of virtual resources, which can in turn be dynamically allocated and reallocated in response to workload demands.

Trends in storage efficiency using virtualization and data reduction

Storage virtualization is a critical element in managing the costs of data growth. Virtual storage functions, such as thin provisioning, space efficient copies, and automated provisioning, provide high levels of utilization, with allocation levels of over 100 percent possible in certain application environments.

In addition, as processor power increases, storage controllers can perform computations on data to squeeze out redundancy. This processing power can be accomplished through lossless compression algorithms, such as Lempel-Zev class 1 (LZ1). It can also be accomplished on a small window of data as it is written, through block or file deduplication, or by using a combination of the two techniques on the same data.

Organizations can use compression and deduplication techniques for backup and long-term archiving of data. New data reduction techniques for online data provide the ability for transforms to be performed directly in the storage controllers.

For more information about how data compression and deduplication technologies work, see the article “IBM Data Footprint Reduction Technology for Data Compression and Deduplication” by IBM expert, Tony Pearson, at:

http://www.ibm.com/systems/storage/resource/technology-topics/data_footprint_reduction_techology.html

Also, enterprise data can be created anywhere in the world, at an analyst’s workstation, at a cash register in a store, on a street corner by a camera monitoring traffic, or at a supplier’s factory. Moving the required data through an automated, policy-driven process is an important area of innovation.

Trends in information management software

Software continues to take advantage of (and often drives) the evolution in hardware technologies, from emerging features of databases to analytical systems.

Software capabilities are being implemented to address new issues, such as the need to perform analytics on real-time data streams. For example, using predictive analytics can help you to understand and predict trends and behaviors in operational real time.

The rise of Internet scale systems, such as Google, Yahoo, and Facebook, is in part because of the ready availability of commodity hardware. These computers, when used in large distributed systems, can support collection and computation over vast amounts of information. This data is used to support searches, to target ads to the most appropriate user at a particular time, or to understand (and monetize) other forms of user behavior.

The following trends have emerged to support the storage, query, and analysis of these increasing volumes and varieties of information:

- *Big data* combines emerging technologies that support massively distributed computations with existing technologies to derive insight and meaning from all kinds, amounts, and sources of information. Big data encompasses more than analytics. It represents a full ecosystem that gathers information from many sources. It prepares that information for use, processes the information to derive insight, and then delivers that insight throughout the enterprise.
- The desire for trusted information drives the requirements for *information integration and governance,* which provides the following benefits:
  - Ensures the highest quality information
  - Governs data throughout its lifecycle
  - Protects and secures all information
  - Integrates all data for a common view
  - Ensures a single understanding and set of knowledge
- Raw information is delivered into the beginning of an *information supply chain*. Information supply chain patterns for analytical systems are becoming well documented. They use a range of techniques, such as extract, transform, and load (ETL) and data replication to cleanse, combine, and load information into data warehouses and data marts. Managing information supply chains to deliver the

right information to the right consumers at the right time is critical to the successful enterprise.

- NoSQL databases, such as MongoDB, HBase, Cassandra, and MemcacheD, give up much of the richness and robustness of modern relational databases for massive scalability (by using sharding) and performance. As the movement has matured, the term *NoSQL* has come to mean *not-only SQL*. This meaning recognizes that the capabilities of SQL are often still required. It also recognizes that the requirements and simpler approach that many NoSQL databases address can be increasingly accommodated within existing relational databases.
- Consumers are looking for better ways to optimize hardware systems for application workloads. *Workload-optimized systems* can improve both the performance and manageability of systems. With this demand comes a new category of *expert integrated systems*. These systems combine the flexibility of general-purpose systems, the elasticity of cloud computing, and the simplicity of an appliance that is tuned to the workload.

The journey to an efficient information infrastructure

IT demands on storage continue to evolve and grow as the number of data sources grows. Innovation is required to ensure that storage systems can accommodate this growth and can remain scalable and manageable. Storage requirements are growing faster than the cost reduction of storage technology, meaning that the cost to IT is increasing even as clients are challenged to reduce expenses. Enterprises must find ways to control and reduce costs through technical innovations.

With storage solutions from IBM, you can generate insight from vast quantities of data, fundamentally changing the way you use information. In addition, with IBM's innovative solutions, you can capitalize on the growing volume, variety, and velocity of data, without adding complexity. These tailored solutions can help you select the best mix of technology, services, and financing for your enterprise.

When considering storage solutions for your environment, keep in mind the following key areas:

- Workload optimized systems
- Expert integrated systems
- Storage efficiency
- Global data access and movement

Workload optimized systems

Realistically, different workloads require a different combination of processing, memory, and storage. Begin by optimizing your hardware systems for the application workloads. Workload-optimized systems use specialized hardware to improve the performance and manageability of systems. The popularity of workload-optimized systems is driving the pre-integration of specialized capabilities by using both appliance structures and hardware virtualization.

Workloads determine the access patterns for storage, the mix of reads and writes, the performance objectives, the need to create copies, and the required types of data protection. The following types of workloads provide online data access:

- *Online transaction processing (OLTP) and database* workloads focus on high transaction rates and small record writes. The data in this type of workload is usually mission-critical, and storage replication is used to mirror updates to recovery sites.
- *Business applications* workloads usually include database applications but with lower I/O rates and fewer writes. Servers are often virtualized, have less frequent replication, and have less expensive data protection than when OLTP is used.
- *High-performance computing (HPC) and analytics* workloads historically include engineering and scientific applications. But, more recently, they are dominated by business analytics.
- *Web, collaboration, and infrastructure* workloads often involve detecting insight from vast quantities of data or multiple copies of fixed content, such as movie or song streaming services. These applications are often deployed on virtual servers and are some of the earliest applications to be deployed on cloud environments.

IBM covers all significant workload categories with its portfolio of storage products:

- IBM provides high performance and optimal capacity utilization at minimal client expense.
- IBM provides global data placement and automated data placement, including information lifecycle management (ILM) functions.
- IBM provides advanced virtualization technology that matches storage to public and private cloud uses.

Expert integrated systems

Building and tuning an information infrastructure is one of the most complex tasks in IT. You need the correct combination of the following systems:

- A server, with considerations such as processor strength, large memory, parallel execution, and internal adapter bandwidth
- A network, with considerations such as optimization for input/output operations per second (IOPS) or latency, and which technologies to use
- Storage, with considerations such as IOPS or latency, streaming or random access, read only, 50/50 read/write, change frequency, compression, and SSD or hard disk drive (HDD)

Because the storage aspect is application-dependent, this consideration requires an amount of expertise. Also, you need the software to take advantage of the available hardware features, and you need to ensure the availability and integrity of the data.

Experiences have shown that setting up a successful analytics system can take up to a year to go through all the necessary decisions to build and then tune the system. Expert integrated systems can reduce this time by allowing you to deploy prebuilt patterns that reflect experience and expertise in designing and deploying systems. Expert integrated systems are the building blocks of capability. When intelligence and knowledge are built directly into your systems, your team does not waste time devising, testing, or tuning their integrated solutions. Instead, your team can create capabilities with new levels of efficiency and speed.

Using systems with integrated expertise can help you achieve greater agility so that you can adapt to workload spikes and deliver new business capabilities. You can increase efficiency by consolidating IT resources and raising productivity. In addition, you can improve simplicity for ease of management, deployment, and integration.

With IBM PureData™ System, the steps to set up an integrated system have already been executed by experts in that area. IBM offers the following optimized IBM PureData System models:

- IBM PureData System for Transactions
- IBM PureData System for Analytics
- IBM PureData System for Operational Analytics

Storage efficiency

After you address the application workloads, consider the storage hierarchies as an important strategy for performance-sensitive applications. Faster storage devices are more expensive (in dollars per gigabyte) than slower devices. By using a modest amount of fast storage and automation to place only the most active data on the faster storage device, you can have a system with dramatic performance benefits at an affordable cost.

Storage efficiency is a key focus item, because technology cost reductions (dollars per gigabyte) are not keeping pace with data growth. Many new applications are highly cost sensitive and might not be viable without low storage costs. Storage virtualization is a critical element in expense management. Virtual storage functions of thin provisioning, space-efficient copies, and automated provisioning allow high levels of utilization, with allocation levels over 100 percent possible in certain application environments. As processor power increases, storage controllers can perform computations on data to squeeze out redundancy through compression algorithms, through block or file de-duplication, or a combination of these techniques on the same data.

For some time, IBM has led the industry in the delivery of replication functions for block storage devices that can now support block storage mirroring at distances up to 300 kilometers (km). And most recently the company announced the IBM Active Cloud Engine™, which is a geographic distribution capability.

With policies established through Active Cloud Engine, an organization can replicate file data between IBM Scale Out Network Attached Storage (SONAS) or IBM Storwize® V7000 instances that are cached on demand by a remote user or that are distributed to a collection of participating nodes. Peer-to-peer and hub-and-spoke topologies are supported, as illustrated in Figure 1 on page 5.

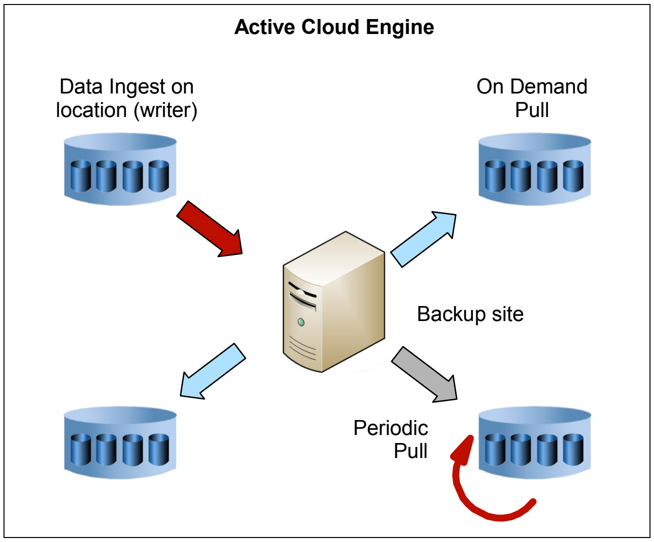

Figure 1   IBM Active Cloud Engine

When combined with local storage efficiency optimization, and virtualization functions, Active Cloud Engine is a powerful tool in creating an optimized, resilient, and global infrastructure.

IBM storage infrastructure management software can deliver the following benefits:

- It helps to reduce the management complexity of your storage environments by centralizing, simplifying, and automating storage tasks. These tasks are often associated with storage hardware, replication services, data protection services, capacity management, and security compliance reporting.
- It improves service levels by notifying administrators of problems faster and manages more of the environment from fewer consoles.
- It controls operating expenses by helping storage administrators to manage more information with less effort.

Global data access and movement

Finally, keep in mind that, with enterprise data being created anywhere in the world, your data must have global access and movement and must move through an automated, policy-driven process.

The volume and velocity of business data is growing at an accelerated rate, and more of this data is being identified as mission critical. At the same time, data availability and response time expectations continue to rise. Deploying and running reliable and highly scalable transactional database systems can be expensive and time consuming. A new approach to delivering fast, reliable, and scalable data services is needed.

In today's business environment, your information management strategies must support big data to help you store it securely and access it efficiently. You need information infrastructure solutions that can filter vast quantities of data from almost any connected device. You also need to analyze the data while it is still in motion to decide what, if any, data must be stored and then virtually integrated with traditional data warehouses.

Why IBM?

Businesses are ready for a new approach to IT. Advances in IT, such as virtualization or the cloud, can help drive a company forward. But with the information explosion showing no signs of slowing down, any gains could start and stop at the data center. For IT to give real value throughout a business, a smarter approach to computing is required.

As the amount of information grows, the requirements for finding, managing, and properly using this information will drive a continued stream of innovations throughout the enterprise. Some of these innovations might result in unique combinations of existing technologies to provide new capabilities. In addition, other innovations might reflect more fundamental enhancements in the spectrum of technologies.

Storage and workload optimized systems will enable and lead the proliferation of new cost-effective information systems. In-memory processing, ever faster and larger storage, and higher density computational power will help make commercially feasible what today is barely possible.

IBM information infrastructure solutions help you to keep up with big data and address the challenges around storage efficiency or data protection. Using IBM storage solutions can lead your company toward improved productivity, service delivery, and reduced risk while streamlining costs.

Resources for more information

For more information about the concepts highlighted in the paper, see the following resources:

- IBM Information Management solutions

  http://www.ibm.com/software/data/

- IBM Information Management Solution Portal

  https://www.ibm.com/developerworks/wikis/display/im/Information+Management+Solution+Portal

- Technical resources for IBM Information Management software on IBM developerWorks®

  http://www.ibm.com/developerworks/data

- IBM big data platform

  http://www.ibm.com/software/data/bigdata/enterprise.html

- Information Integration for Big Data

  http://www.ibm.com/software/data/infosphere/information-integration-big-data/index.html

- A Smarter Approach to IT: IBM PureSystems™

  http://www.ibm.com/ibm/puresystems/us/en/op-ad.html

- IBM PureApplication™ System

  http://www.ibm.com/ibm/puresystems/us/en/pf_pureapplication.html

- IBM PureData System

  http://www.ibm.com/ibm/puresystems/us/en/pf_puredata.html

- *Overview of IBM PureSystems*, TIPS0892

  http://www.redbooks.ibm.com/abstracts/tips0892.html?Open

Notices

This document, REDP-4945-00, was created or updated on December 13, 2012.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation in the United States, other countries, or both. These and other IBM trademarked terms are marked on their first occurrence in this information with the appropriate symbol (® or ™), indicating US registered or common law trademarks owned by IBM at the time this information was published. Such trademarks may also be registered or common law trademarks in other countries. A current list of IBM trademarks is available on the Web at `http://www.ibm.com/legal/copytrade.shtml`

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Active Cloud Engine™
developerWorks®
IBM®
PureApplication™
PureData™
PureSystems™
Redbooks®
Redbooks (logo) ®
Storwize®

The following terms are trademarks of other companies:

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