Chapter 1
Introduction to Information Storage and Management

KEY CONCEPTS

- Data and Information
- Structured and Unstructured Data
- Storage Technology Architectures
- Core Elements of a Data Center
- Information Management
- Information Lifecycle Management

Information is increasingly important in our daily lives. We have become information dependents of the twenty-first century, living in an on-command, on-demand world that means we need information when and where it is required. We access the Internet every day to perform searches, participate in social networking, send and receive e-mails, share pictures and videos, and scores of other applications. Equipped with a growing number of content-generating devices, more information is being created by individuals than by businesses. Information created by individuals gains value when shared with others. When created, information resides locally on devices such as cell phones, cameras, and laptops. To share this information, it needs to be uploaded via networks to data centers. It is interesting to note that while the majority of information is created by individuals, it is stored and managed by a relatively small number of organizations. Figure 1-1 depicts this virtuous cycle of information.

The importance, dependency, and volume of information for the business world also continue to grow at astounding rates. Businesses depend on fast and reliable access to information critical to their success. Some of the business applications that process information include airline reservations, telephone billing systems, e-commerce, ATMs, product designs, inventory management, e-mail archives, Web portals, patient records, credit cards, life sciences, and global capital markets.

The increasing criticality of information to the businesses has amplified the challenges in protecting and managing the data. The volume of data that

business must manage has driven strategies to classify data according to its value and create rules for the treatment of this data over its lifecycle. These strategies not only provide financial and regulatory benefits at the business level, but also manageability benefits at operational levels to the organization.

Data centers now view information storage as one of their core elements, along with applications, databases, operating systems, and networks. Storage technology continues to evolve with technical advancements offering increasingly higher levels of availability, security, scalability, performance, integrity, capacity, and manageability.

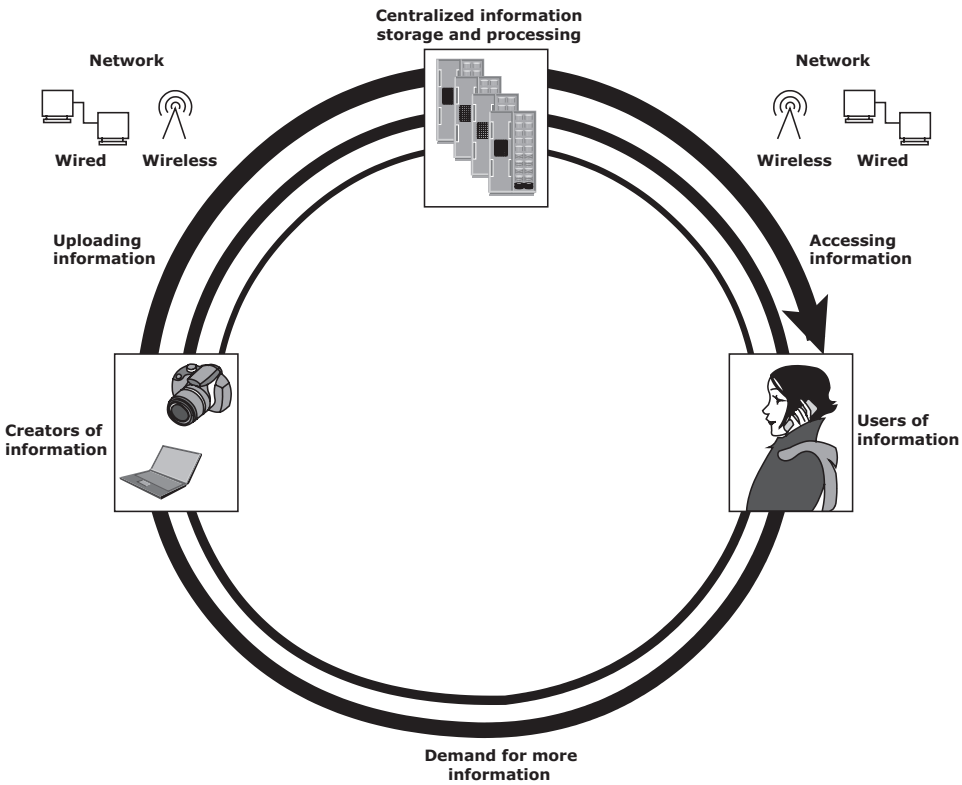


Figure 1-1: Virtuous cycle of information

This chapter describes the evolution of information storage architecture from simple direct-attached models to complex networked topologies. It introduces the information lifecycle management (ILM) strategy, which aligns the information technology (IT) infrastructure with business priorities.

1.1 Information Storage

Businesses use data to derive information that is critical to their day-to-day operations. Storage is a repository that enables users to store and retrieve this digital data.

1.1.1 Data

Data is a collection of raw facts from which conclusions may be drawn. Handwritten letters, a printed book, a family photograph, a movie on video tape, printed and duly signed copies of mortgage papers, a bank's ledgers, and an account holder's passbooks are all examples of data.

Before the advent of computers, the procedures and methods adopted for data creation and sharing were limited to fewer forms, such as paper and film. Today, the same data can be converted into more convenient forms such as an e-mail message, an e-book, a bitmapped image, or a digital movie. This data can be generated using a computer and stored in strings of `0`s and `1`s, as shown in Figure 1-2. Data in this form is called *digital data* and is accessible by the user only after it is processed by a computer.

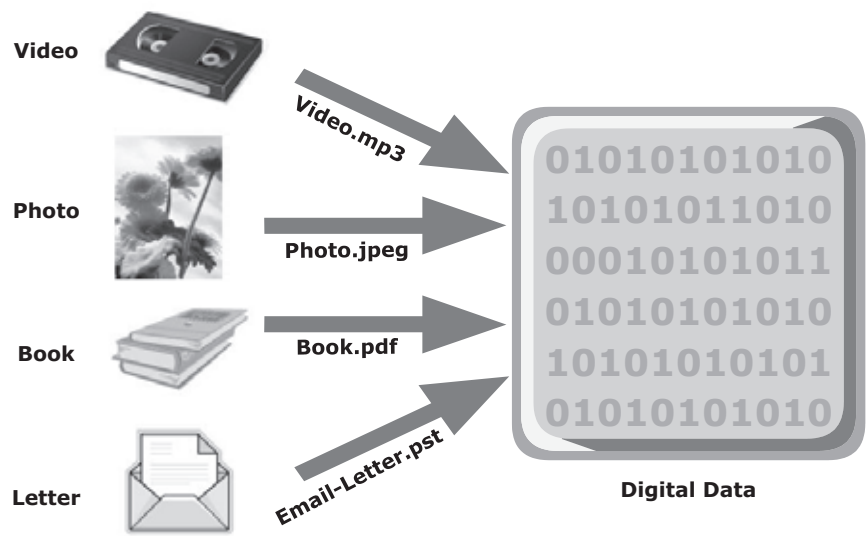


Figure 1-2: Digital data

With the advancement of computer and communication technologies, the rate of data generation and sharing has increased exponentially. The following is a list of some of the factors that have contributed to the growth of digital data:

- **Increase in data processing capabilities:** Modern-day computers provide a significant increase in processing and storage capabilities. This enables the conversion of various types of content and media from conventional forms to digital formats.
- **Lower cost of digital storage:** Technological advances and decrease in the cost of storage devices have provided low-cost solutions and encouraged the development of less expensive data storage devices. This cost benefit has increased the rate at which data is being generated and stored.
- **Affordable and faster communication technology:** The rate of sharing digital data is now much faster than traditional approaches. A handwritten letter may take a week to reach its destination, whereas it only takes a few seconds for an e-mail message to reach its recipient.

Inexpensive and easier ways to create, collect, and store all types of data, coupled with increasing individual and business needs, have led to accelerated data growth, popularly termed the *data explosion*. Data has different purposes and criticality, so both individuals and businesses have contributed in varied proportions to this data explosion.

The importance and the criticality of data vary with time. Most of the data created holds significance in the short-term but becomes less valuable over time. This governs the type of data storage solutions used. Individuals store data on a variety of storage devices, such as hard disks, CDs, DVDs, or Universal Serial Bus (USB) flash drives.

EXAMPLE OF RESEARCH AND BUSINESS DATA

- **Seismology:** Involves collecting data related to various sources and parameters of earthquakes, and other relevant data that needs to be processed to derive meaningful information.
- **Product data:** Includes data related to various aspects of a product, such as inventory, description, pricing, availability, and sales.
- **Customer data:** A combination of data related to a company's customers, such as order details, shipping addresses, and purchase history.
- **Medical data:** Data related to the health care industry, such as patient history, radiological images, details of medication and other treatment, and insurance information.

Businesses generate vast amounts of data and then extract meaningful information from this data to derive economic benefits. Therefore, businesses need to maintain data and ensure its availability over a longer period.

Furthermore, the data can vary in criticality and may require special handling. For example, legal and regulatory requirements mandate that banks maintain account information for their customers accurately and securely. Some businesses handle data for millions of customers, and ensures the security and integrity of data over a long period of time. This requires high-capacity storage devices with enhanced security features that can retain data for a long period.

1.1.2 Types of Data

Data can be classified as structured or unstructured (see Figure 1-3) based on how it is stored and managed. Structured data is organized in rows and columns in a rigidly defined format so that applications can retrieve and process it efficiently. Structured data is typically stored using a database management system (DBMS).

Data is unstructured if its elements cannot be stored in rows and columns, and is therefore difficult to query and retrieve by business applications. For example, customer contacts may be stored in various forms such as sticky notes, e-mail messages, business cards, or even digital format files such as .doc, .txt, and .pdf. Due its unstructured nature, it is difficult to retrieve using a customer relationship management application. Unstructured data may not have the required components to identify itself uniquely for any type of processing or interpretation. Businesses are primarily concerned with managing unstructured data because over 80 percent of enterprise data is unstructured and requires significant storage space and effort to manage.

1.1.3 Information

Data, whether structured or unstructured, does not fulfill any purpose for individuals or businesses unless it is presented in a meaningful form. Businesses need to analyze data for it to be of value. *Information* is the intelligence and knowledge derived from data.

Businesses analyze raw data in order to identify meaningful trends. On the basis of these trends, a company can plan or modify its strategy. For example, a retailer identifies customers' preferred products and brand names by analyzing their purchase patterns and maintaining an inventory of those products.

Effective data analysis not only extends its benefits to existing businesses, but also creates the potential for new business opportunities by using the information in creative ways. Job portal is an example. In order to reach a wider set of prospective employers, job seekers post their résumés on various websites offering job search facilities. These websites collect the résumés and post them on centrally accessible locations for prospective employers. In addition, companies post available positions on job search sites. Job-matching software matches keywords from

résumés to keywords in job postings. In this manner, the job search engine uses data and turns it into information for employers and job seekers.

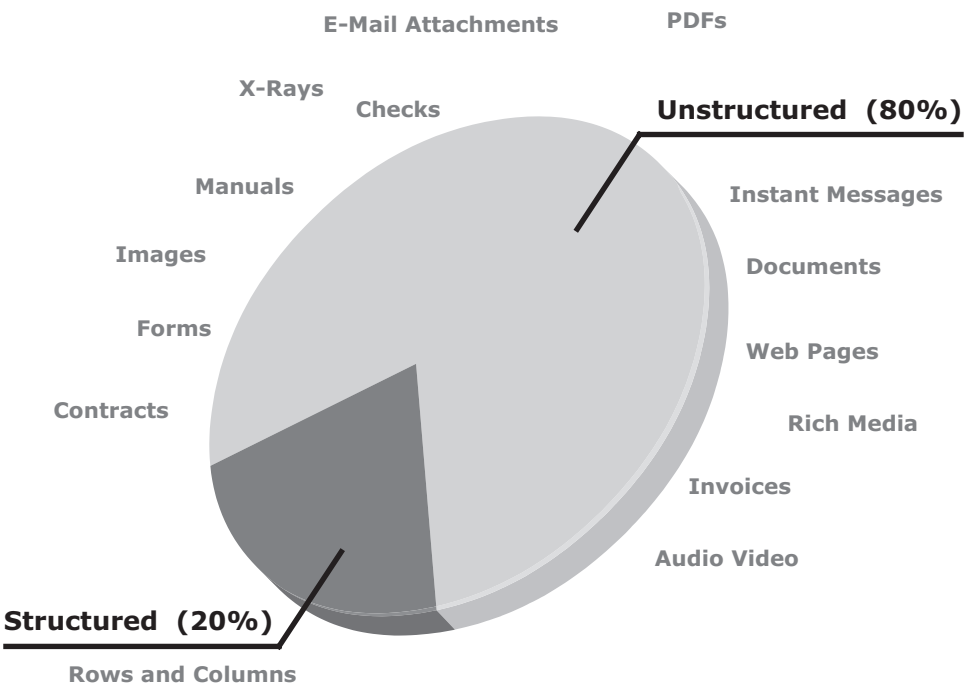


Figure 1-3: Types of data

Because information is critical to the success of a business, there is an ever-present concern about its availability and protection. Legal, regulatory, and contractual obligations regarding the availability and protection of data only add to these concerns. Outages in key industries, such as financial services, telecommunications, manufacturing, retail, and energy cost millions of U.S. dollars per hour.

1.1.4 Storage

Data created by individuals or businesses must be stored so that it is easily accessible for further processing. In a computing environment, devices designed for storing data are termed *storage devices* or simply *storage*. The type of storage used varies based on the type of data and the rate at which it is created and used. Devices such as memory in a cell phone or digital camera, DVDs, CD-ROMs, and hard disks in personal computers are examples of storage devices.

Businesses have several options available for storing data including internal hard disks, external disk arrays and tapes.

1.2 Evolution of Storage Technology and Architecture

Historically, organizations had centralized computers (mainframe) and information storage devices (tape reels and disk packs) in their data center. The evolution of open systems and the affordability and ease of deployment that they offer made it possible for business units/departments to have their own servers and storage. In earlier implementations of open systems, the storage was typically internal to the server.

The proliferation of departmental servers in an enterprise resulted in unprotected, unmanaged, fragmented islands of information and increased operating cost. Originally, there were very limited policies and processes for managing these servers and the data created. To overcome these challenges, storage technology evolved from non-intelligent internal storage to intelligent networked storage (see Figure 1-4). Highlights of this technology evolution include:

- **Redundant Array of Independent Disks (RAID):** This technology was developed to address the cost, performance, and availability requirements of data. It continues to evolve today and is used in all storage architectures such as DAS, SAN, and so on.
- **Direct-attached storage (DAS):** This type of storage connects directly to a server (host) or a group of servers in a cluster. Storage can be either internal or external to the server. External DAS alleviated the challenges of limited internal storage capacity.
- **Storage area network (SAN):** This is a dedicated, high-performance *Fibre Channel (FC)* network to facilitate *block-level* communication between servers and storage. Storage is partitioned and assigned to a server for accessing its data. SAN offers scalability, availability, performance, and cost benefits compared to DAS.
- **Network-attached storage (NAS):** This is dedicated storage for *file serving* applications. Unlike a SAN, it connects to an existing communication network (LAN) and provides file access to heterogeneous clients. Because it is purposely built for providing storage to file server applications, it offers higher scalability, availability, performance, and cost benefits compared to general purpose file servers.
- **Internet Protocol SAN (IP-SAN):** One of the latest evolutions in storage architecture, IP-SAN is a convergence of technologies used in SAN and NAS. IP-SAN provides block-level communication across a local or wide area network (LAN or WAN), resulting in greater consolidation and availability of data.

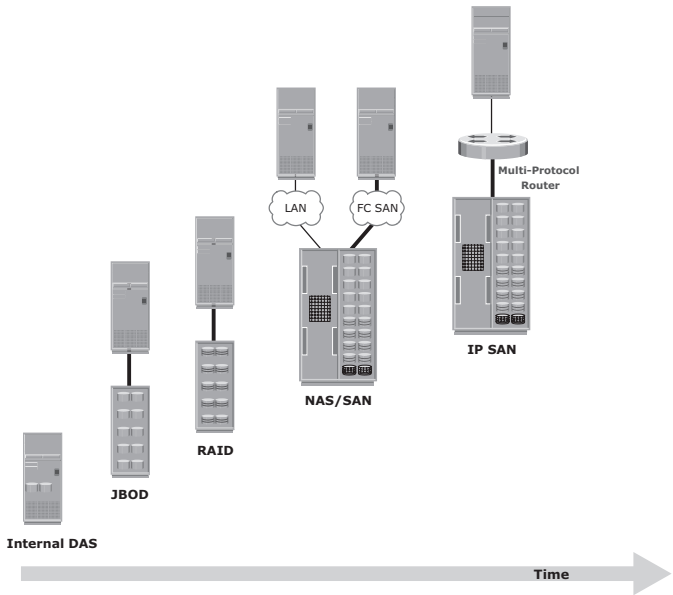


Figure 1-4: Evolution of storage architectures

Storage technology and architecture continues to evolve, which enables organizations to consolidate, protect, optimize, and leverage their data to achieve the highest return on information assets.

1.3 Data Center Infrastructure

Organizations maintain data centers to provide centralized data processing capabilities across the enterprise. Data centers store and manage large amounts of mission-critical data. The data center infrastructure includes computers, storage systems, network devices, dedicated power backups, and environmental controls (such as air conditioning and fire suppression).

Large organizations often maintain more than one data center to distribute data processing workloads and provide backups in the event of a disaster. The storage requirements of a data center are met by a combination of various storage architectures.

1.3.1 Core Elements

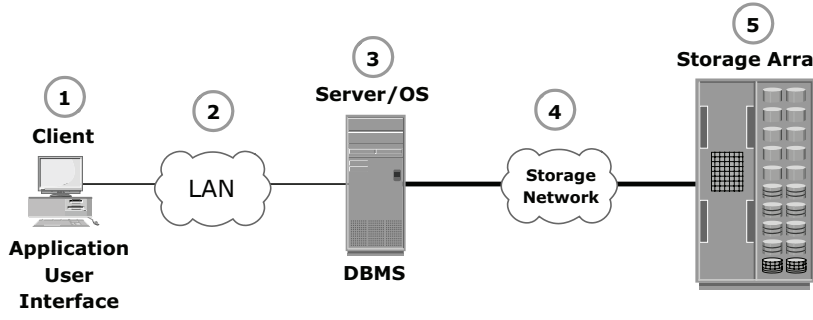
Five core elements are essential for the basic functionality of a data center:

- **Application:** An application is a computer program that provides the logic for computing operations. Applications, such as an order processing system, can be layered on a database, which in turn uses operating system services to perform read/write operations to storage devices.

- **Database:** More commonly, a database management system (DBMS) provides a structured way to store data in logically organized tables that are interrelated. A DBMS optimizes the storage and retrieval of data.
- **Server and operating system:** A computing platform that runs applications and databases.
- **Network:** A data path that facilitates communication between clients and servers or between servers and storage.
- **Storage array:** A device that stores data persistently for subsequent use.

These core elements are typically viewed and managed as separate entities, but all the elements must work together to address data processing requirements.

Figure 1-5 shows an example of an order processing system that involves the five core elements of a data center and illustrates their functionality in a business process.



1. A customer places an order through the AUI of the order processing application software located on the client computer.
2. The client connects to the server over the LAN and accesses the DBMS located on the server to update the relevant information such as the customer name, address, payment method, products ordered, and quantity ordered.
3. The DBMS uses the server operating system to read and write this data to the database located on physical disks in the storage array.
4. The Storage Network provides the communication link between the server and the storage array and transports the read or write commands between them.
5. The storage array, after receiving the read or write commands from the server, performs the necessary operations to store the data on physical disks.

Figure 1-5: Example of an order processing system

1.3.2 Key Requirements for Data Center Elements

Uninterrupted operation of data centers is critical to the survival and success of a business. It is necessary to have a reliable infrastructure that ensures data is accessible at all times. While the requirements, shown in Figure 1-6, are applicable to all elements of the data center infrastructure, our focus here is on storage

systems. The various technologies and solutions to meet these requirements are covered in this book.

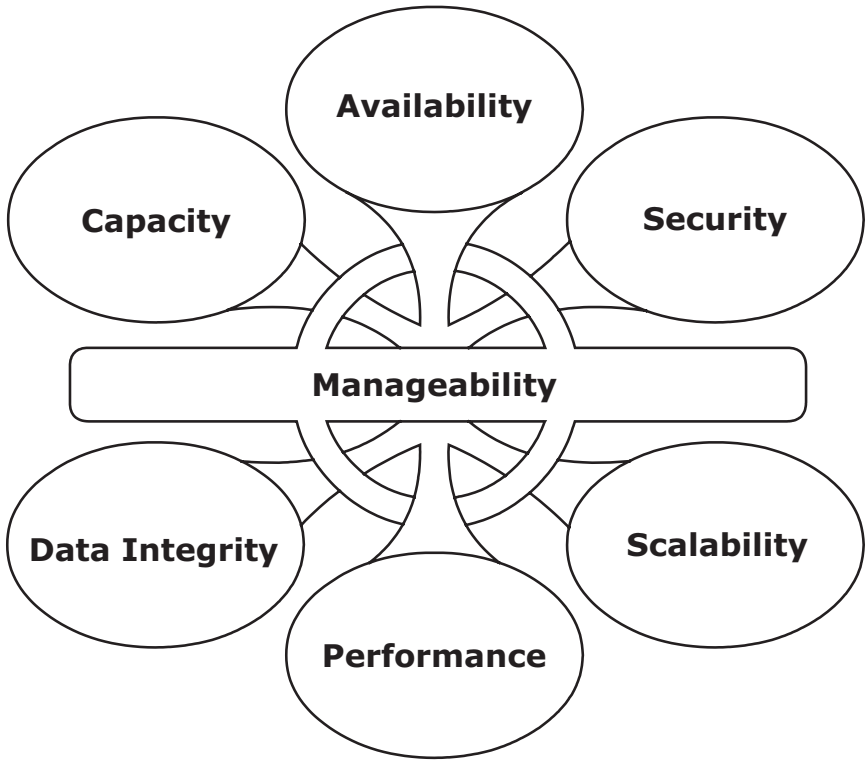


Figure 1-6: Key characteristics of data center elements

- **Availability:** All data center elements should be designed to ensure accessibility. The inability of users to access data can have a significant negative impact on a business.
- **Security:** Polices, procedures, and proper integration of the data center core elements that will prevent unauthorized access to information must be established. In addition to the security measures for client access, specific mechanisms must enable servers to access only their allocated resources on storage arrays.
- **Scalability:** Data center operations should be able to allocate additional processing capabilities or storage on demand, without interrupting business operations. Business growth often requires deploying more servers, new applications, and additional databases. The storage solution should be able to grow with the business.

- **Performance:** All the core elements of the data center should be able to provide optimal performance and service all processing requests at high speed. The infrastructure should be able to support performance requirements.
- **Data integrity:** Data integrity refers to mechanisms such as error correction codes or parity bits which ensure that data is written to disk exactly as it was received. Any variation in data during its retrieval implies corruption, which may affect the operations of the organization.
- **Capacity:** Data center operations require adequate resources to store and process large amounts of data efficiently. When capacity requirements increase, the data center must be able to provide additional capacity without interrupting availability, or, at the very least, with minimal disruption. Capacity may be managed by reallocation of existing resources, rather than by adding new resources.
- **Manageability:** A data center should perform all operations and activities in the most efficient manner. Manageability can be achieved through automation and the reduction of human (manual) intervention in common tasks.

1.3.3 Managing Storage Infrastructure

Managing a modern, complex data center involves many tasks. Key management activities include:

- *Monitoring* is the continuous collection of information and the review of the entire data center infrastructure. The aspects of a data center that are monitored include security, performance, accessibility, and capacity.
- *Reporting* is done periodically on resource performance, capacity, and utilization. Reporting tasks help to establish business justifications and chargeback of costs associated with data center operations.
- *Provisioning* is the process of providing the hardware, software, and other resources needed to run a data center. Provisioning activities include capacity and resource planning. *Capacity planning* ensures that the user's and the application's future needs will be addressed in the most cost-effective and controlled manner. *Resource planning* is the process of evaluating and identifying required resources, such as personnel, the facility (site), and the technology. Resource planning ensures that adequate resources are available to meet user and application requirements.

For example, the utilization of an application's allocated storage capacity may be monitored. As soon as utilization of the storage capacity reaches a critical

value, additional storage capacity may be provisioned to the application. If utilization of the storage capacity is properly monitored and reported, business growth can be understood and future capacity requirements can be anticipated. This helps to frame a proactive data management policy.

1.4 Key Challenges in Managing Information

In order to frame an effective information management policy, businesses need to consider the following key challenges of information management:

- **Exploding digital universe:** The rate of information growth is increasing exponentially. Duplication of data to ensure high availability and repurposing has also contributed to the multifold increase of information growth.
- **Increasing dependency on information:** The strategic use of information plays an important role in determining the success of a business and provides competitive advantages in the marketplace.
- **Changing value of information:** Information that is valuable today may become less important tomorrow. The value of information often changes over time.

Framing a policy to meet these challenges involves understanding the value of information over its lifecycle.

1.5 Information Lifecycle

The *information lifecycle* is the "change in the value of information" over time. When data is first created, it often has the highest value and is used frequently. As data ages, it is accessed less frequently and is of less value to the organization. Understanding the information lifecycle helps to deploy appropriate storage infrastructure, according to the changing value of information.

For example, in a sales order application, the value of the information changes from the time the order is placed until the time that the warranty becomes void (see Figure 1-7). The value of the information is highest when a company receives a new sales order and processes it to deliver the product. After order fulfillment, the customer or order data need not be available for real-time access. The company can transfer this data to less expensive secondary storage with lower accessibility and availability requirements unless or until a warranty claim or another event triggers its need. After the warranty becomes void, the company can archive or dispose of data to create space for other high-value information.

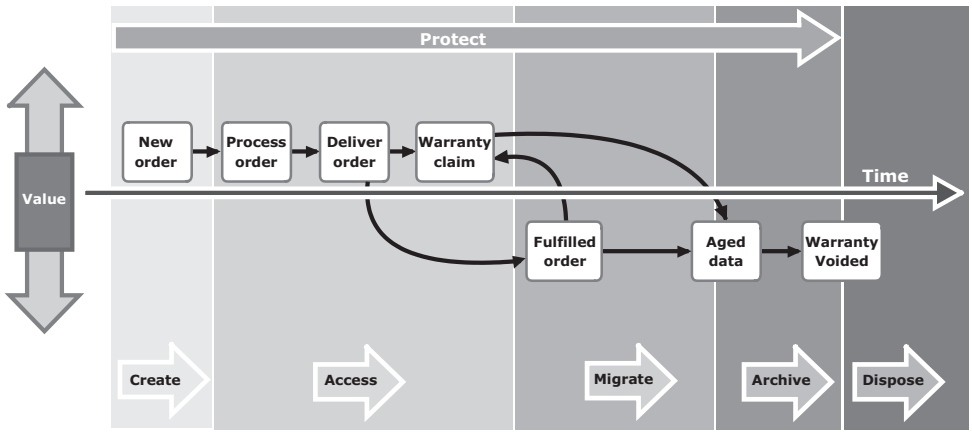


Figure 1-7: Changing value of sales order information

1.5.1 Information Lifecycle Management

Today's business requires data to be protected and available 24 × 7. Data centers can accomplish this with the optimal and appropriate use of storage infrastructure. An effective information management policy is required to support this infrastructure and leverage its benefits.

Information lifecycle management (ILM) is a proactive strategy that enables an IT organization to effectively manage the data throughout its lifecycle, based on predefined business policies. This allows an IT organization to optimize the storage infrastructure for maximum return on investment. An ILM strategy should include the following characteristics:

- **Business-centric:** It should be integrated with key processes, applications, and initiatives of the business to meet both current and future growth in information.
- **Centrally managed:** All the information assets of a business should be under the purview of the ILM strategy.
- **Policy-based:** The implementation of ILM should not be restricted to a few departments. ILM should be implemented as a policy and encompass all business applications, processes, and resources.
- **Heterogeneous:** An ILM strategy should take into account all types of storage platforms and operating systems.
- **Optimized:** Because the value of information varies, an ILM strategy should consider the different storage requirements and allocate storage resources based on the information's value to the business.

TIERED STORAGE

Tiered storage is an approach to define different storage levels in order to reduce total storage cost. Each tier has different levels of protection, performance, data access frequency, and other considerations. Information is stored and moved between different tiers based on its value over time. For example, mission-critical, most accessed information may be stored on Tier 1 storage, which consists of high performance media with a highest level of protection. Medium accessed and other important data is stored on Tier 2 storage, which may be on less expensive media with moderate performance and protection. Rarely accessed or event specific information may be stored on lower tiers of storage.

1.5.2 ILM Implementation

The process of developing an ILM strategy includes four activities—classifying, implementing, managing, and organizing:

- *Classifying* data and applications on the basis of business rules and policies to enable differentiated treatment of information
- *Implementing* policies by using information management tools, starting from the creation of data and ending with its disposal
- *Managing* the environment by using integrated tools to reduce operational complexity
- *Organizing* storage resources in tiers to align the resources with data classes, and storing information in the right type of infrastructure based on the information's current value

Implementing ILM across an enterprise is an ongoing process. Figure 1-8 illustrates a three-step road map to enterprise-wide ILM.

Steps 1 and 2 are aimed at implementing ILM in a limited way across a few enterprise-critical applications. In Step 1, the goal is to implement a storage networking environment. Storage architectures offer varying levels of protection and performance and this acts as a foundation for future policy-based information management in Steps 2 and 3. The value of tiered storage platforms can be exploited by allocating appropriate storage resources to the applications based on the value of the information processed.

Step 2 takes ILM to the next level, with detailed application or data classification and linkage of the storage infrastructure to business policies. These classifications and the resultant policies can be automatically executed using tools for one or more applications, resulting in better management and optimal allocation of storage resources.

Step 3 of the implementation is to automate more of the applications or data classification and policy management activities in order to scale to a wider set of enterprise applications.

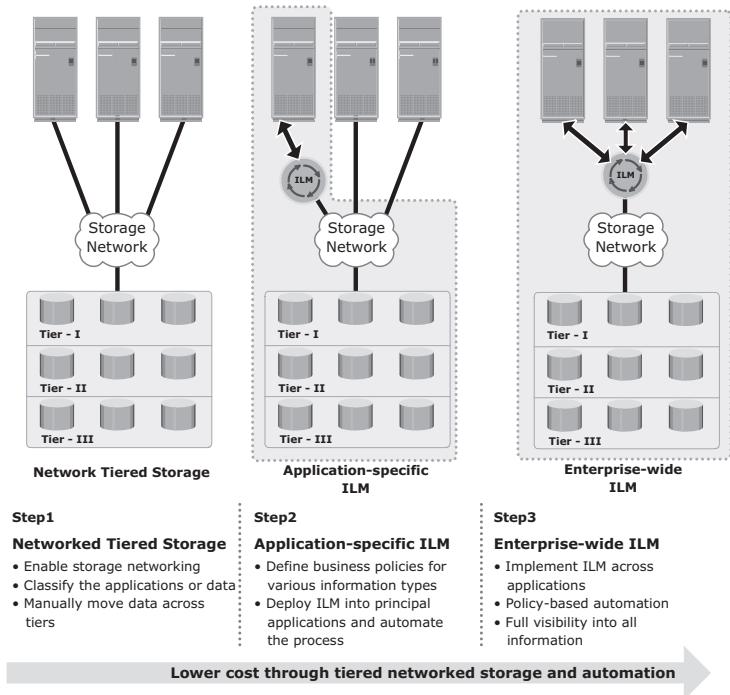


Figure 1-8: Implementation of ILM

1.5.3 ILM Benefits

Implementing an ILM strategy has the following key benefits that directly address the challenges of information management:

- *Improved utilization* by using tiered storage platforms and increased visibility of all enterprise information.
- *Simplified management* by integrating process steps and interfaces with individual tools and by increasing automation.
- *A wider range of options* for backup, and recovery to balance the need for business continuity.
- *Maintaining compliance* by knowing what data needs to be protected for what length of time.

- *Lower Total Cost of Ownership* (TCO) by aligning the infrastructure and management costs with information value. As a result, resources are not wasted, and complexity is not introduced by managing low-value data at the expense of high-value data.

Summary

This chapter described the importance of data, information, and storage infrastructure. Meeting today's storage needs begins with understanding the type of data, its value, and key management requirements of a storage system.

This chapter also emphasized the importance of the ILM strategy, which businesses are adopting to manage information effectively across the enterprise. ILM is enabling businesses to gain competitive advantage by classifying, protecting, and leveraging information.

The evolution of storage architectures and the core elements of a data center covered in this chapter provided the foundation on information storage. The next chapter discusses storage system environment.

EXERCISES

1. A hospital uses an application that stores patient X-ray data in the form of large binary objects in an Oracle database. The application is hosted on a UNIX server, and the hospital staff accesses the X-ray records through a Gigabit Ethernet backbone. Storage array provides storage to the UNIX server, which has 6 terabytes of usable capacity.
   - Explain the core elements of the data center. What are the typical challenges the storage management team may face in meeting the service-level demands of the hospital staff?
   - Describe how the value of this patient data might change over time.
2. An engineering design department of a large company maintains over 600,000 engineering drawings that its designers access and reuse in their current projects, modifying or updating them as required. The design team wants instant access to the drawings for its current projects, but is currently constrained by an infrastructure that is not able to scale to meet the response time requirements. The team has classified the drawings as "most frequently accessed," "frequently accessed," "occasionally accessed," and "archive."
   - Suggest a strategy for design department that optimizes the storage infrastructure by using ILM.
   - Explain how you will use "tiered storage" based on access frequency.
   - Detail the hardware and software components you will need to implement your strategy.
   - Research products and solutions currently available to meet the solution you are proposing.
3. The marketing department at a mid size firm is expanding. New hires are being added to the department and they are given network access to the department's files. IT has given marketing a networked drive on the LAN, but it keeps reaching capacity every third week. Current capacity is 500 gigabytes (and growing), with hundreds of files. Users are complaining about LAN response times and capacity. As the IT manager, what could you recommend to improve the situation?
4. A large company is considering a storage infrastructure—one that is scalable and provides high availability. More importantly, the company also needs performance for its mission-critical applications. Which storage topology would you recommend (SAN, NAS, IP SAN) and why?