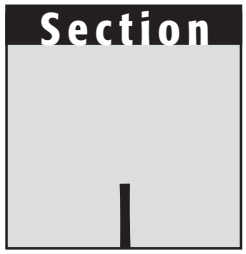
EMC²

Information Storage and Management

Second Edition

Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments

EMC Education Services



Storage System

In This Section

Chapter 1: Introduction to Information Storage
Chapter 2: Data Center Environment
Chapter 3: Data Protection: RAID
Chapter 4: Intelligent Storage Systems

Chapter 1
Introduction to Information Storage

KEY CONCEPTS

- Data and Information
- Structured and Unstructured Data
- Evolution of Storage Architecture
- Core Elements of a Data Center
- Virtualization and Cloud Computing

Information is increasingly important in our daily lives. We have become information-dependent in the 21st century, living in an on-command, on-demand world, which 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 use scores of other applications. Equipped with a growing number of content-generating devices, more information is created by individuals than by organizations (including business, governments, non-profits and so on). Information created by individuals gains value when shared with others. When created, information resides locally on devices, such as cell phones, smartphones, tablets, cameras, and laptops. To be shared, this information needs to be uploaded to central data repositories (data centers) via networks. Although the majority of information is created by individuals, it is stored and managed by a relatively small number of organizations.

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. Examples of business processes or systems that rely on digital information include airline reservations, telecommunications billing, Internet commerce, electronic banking, credit card transaction processing, capital/stock trading, health care claims processing, life science research, and so on. The increasing dependence of businesses on information has amplified the challenges in storing, protecting, and managing

data. Legal, regulatory, and contractual obligations regarding the availability and protection of data further add to these challenges.

Organizations usually maintain one or more data centers to store and manage information. A *data center* is a facility that contains information storage and other physical information technology (IT) resources for computing, networking, and storing information. In traditional data centers, the storage resources are typically dedicated for each of the business units or applications. The proliferation of new applications and increasing data growth have resulted in islands of discrete information storage infrastructures in these data centers. This leads to complex information management and underutilization of storage resources. Virtualization optimizes resource utilization and eases resource management. Organizations incorporate virtualization in their data centers to transform them into *virtualized data centers* (VDCs). Cloud computing, which represents a fundamental shift in how IT is built, managed, and provided, further reduces information storage and management complexity and IT resource provisioning time. Cloud computing brings in a fully automated request-fulfillment process that enables users to rapidly obtain storage and other IT resources on demand. Through cloud computing, an organization can rapidly deploy applications where the underlying storage capability can scale-up and scale-down, based on the business requirements.

This chapter describes the evolution of information storage architecture from a server-centric model to an information-centric model. It also provides an overview of virtualization and cloud computing.

1.1 Information Storage

Organizations process data to derive the information required for their day-to-day operations. Storage is a repository that enables users to persistently store and retrieve this digital data.

1.1.1 Data

Data is a collection of raw facts from which conclusions might be drawn. Handwritten letters, a printed book, a family photograph, printed and duly signed copies of mortgage papers, a bank's ledgers, and an airline ticket are all examples that contain data.

Before the advent of computers, the 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 digital image, or a digital movie. This data can be generated using a computer and stored as strings of binary numbers (0s and 1s), as shown in Figure 1-1. Data in this form is called *digital data* and is accessible by the user only after a computer processes it.

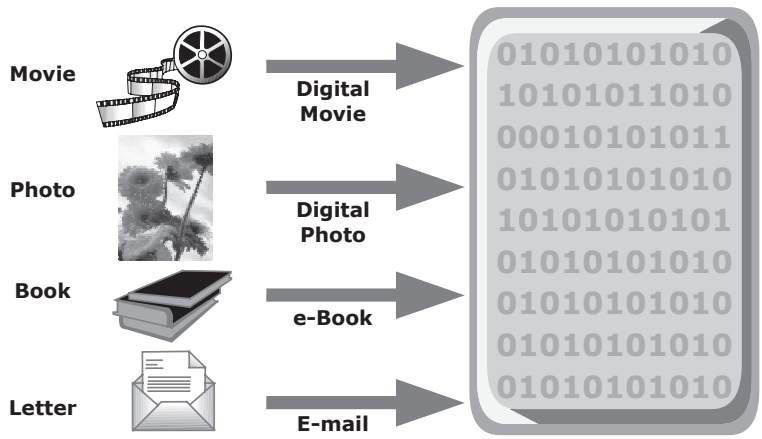


Figure 1-1: Digital data

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 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 the decrease in the cost of storage devices have provided low-cost storage solutions. This cost benefit has increased the rate at which digital data is generated and stored.
- **Affordable and faster communication technology:** The rate of sharing digital data is now much faster than traditional approaches. A handwritten letter might take a week to reach its destination, whereas it typically takes only a few seconds for an e-mail message to reach its recipient.
- **Proliferation of applications and smart devices:** Smartphones, tablets, and newer digital devices, along with smart applications, have significantly contributed to the generation of digital content.

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 *data explosion*. Both individuals and businesses have contributed in varied proportions to this data explosion.

The importance and value of data vary with time. Most of the data created holds significance for a short term but becomes less valuable over time. This governs the type of data storage solutions used. Typically, recent data which

has higher usage is stored on faster and more expensive storage. As it ages, it may be moved to slower, less expensive but reliable storage.

EXAMPLES OF RESEARCH AND BUSINESS DATA

Following are some examples of research and business data:

- **Customer data: Data related to a company's customers, such as order details, shipping addresses, and purchase history.**
- **Product data: Includes data related to various aspects of a product, such as inventory, description, pricing, availability, and sales.**
- **Medical data: Data related to the healthcare industry, such as patient history, radiological images, details of medication and other treatment, and insurance information.**
- **Seismic data: Seismology is a scientific study of earthquakes. It involves collecting data and processes to derive information that helps determine the location and magnitude of earthquakes.**

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 might 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 ensure the security and integrity of data over a long period of time. This requires high-performance and high-capacity storage devices with enhanced security and compliance that can retain data for a long period.

1.1.2 Types of Data

Data can be classified as structured or unstructured (see Figure 1-2) 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, which makes it difficult to query and retrieve by applications. For example, customer contacts that are 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 to its unstructured nature, it is difficult to retrieve this data using a traditional customer relationship management application. A vast majority of new data being created

today is unstructured. The industry is challenged with with new architectures, technologies, techniques, and skills to store, manage, analyze, and derive value from unstructured data from numerous sources.

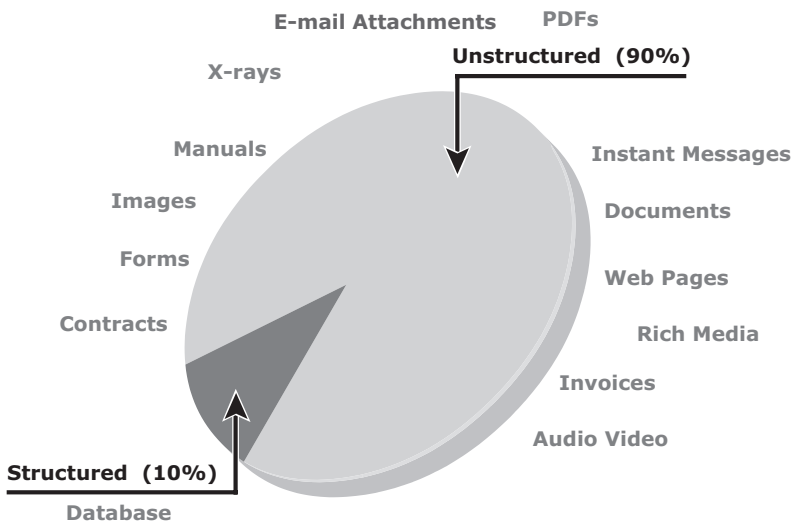


Figure 1-2: Types of data

1.1.3 Big Data

Big data is a new and evolving concept, which refers to data sets whose sizes are beyond the capability of commonly used software tools to capture, store, manage, and process within acceptable time limits. It includes both structured and unstructured data generated by a variety of sources, including business application transactions, web pages, videos, images, e-mails, social media, and so on. These data sets typically require real-time capture or updates for analysis, predictive modeling, and decision making.

Significant opportunities exist to extract value from big data. The big data ecosystem (see Figure 1-3) consists of the following:

1. Devices that collect data from multiple locations and also generate new data about this data (metadata).
2. Data collectors who gather data from devices and users.
3. Data aggregators that compile the collected data to extract meaningful information.
4. Data users and buyers who benefit from the information collected and aggregated by others in the data value chain.

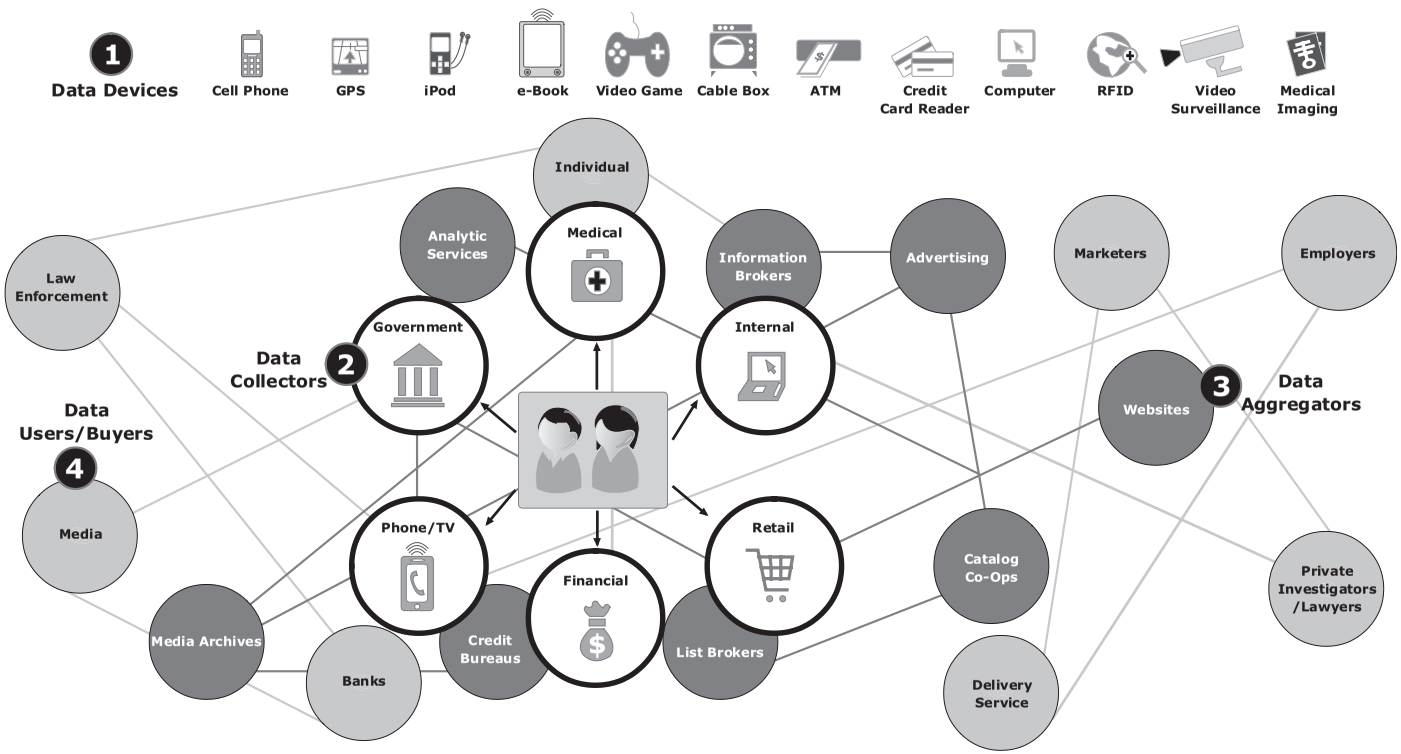


Figure 1-3: Big data ecosystem

Traditional IT infrastructure and data processing tools and methodologies are inadequate to handle the volume, variety, dynamism, and complexity of big data. Analyzing big data in real time requires new techniques, architectures, and tools that provide high performance, massively parallel processing (MPP) data platforms, and advanced analytics on the data sets.

Data science is an emerging discipline, which enables organizations to derive business value from big data. Data science represents the synthesis of several existing disciplines, such as statistics, math, data visualization, and computer science to enable data scientists to develop advanced algorithms for the purpose of analyzing vast amounts of information to drive new value and make more data-driven decisions.

Several industries and markets currently looking to employ data science techniques include medical and scientific research, health care, public administration, fraud detection, social media, banks, insurance companies, and other digital information-based entities that benefit from the analytics of big data.

1.1.4 Information

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

Businesses analyze raw data 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.

1.1.5 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 a media card in a cell phone or digital camera, DVDs, CD-ROMs, and disk drives 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 Architecture

Historically, organizations had centralized computers (mainframes) and information storage devices (tape reels and disk packs) in their data center. The evolution of open systems, their affordability, and ease of deployment 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. These storage devices could not be shared with any other servers. This approach is referred to as *server-centric storage architecture* (see Figure 1-4 [a]). In this architecture, each server has a limited number of storage devices, and any administrative tasks, such as maintenance of the server or increasing storage capacity, might result in unavailability of information. The proliferation of departmental servers in an enterprise resulted in unprotected, unmanaged, fragmented islands of information and increased capital and operating expenses..

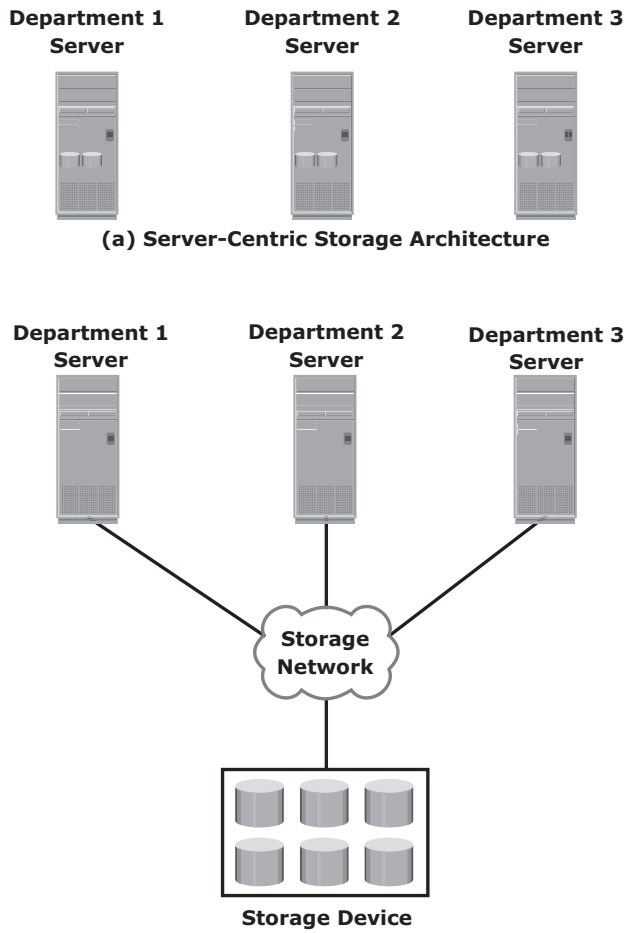


Figure 1-4: Evolution of storage architecture

To overcome these challenges, storage evolved from server-centric to *information-centric architecture* (see Figure 1-4 [b]). In this architecture, storage devices are managed centrally and independent of servers. These centrally-managed

storage devices are shared with multiple servers. When a new server is deployed in the environment, storage is assigned from the same shared storage devices to that server. The capacity of shared storage can be increased dynamically by adding more storage devices without impacting information availability. In this architecture, information management is easier and cost-effective.

Storage technology and architecture continue 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 house and manage large amounts of data. The data center infrastructure includes hardware components, such as computers, storage systems, network devices, and power backups; and software components, such as applications, operating systems, and management software. It also includes environmental controls, such as air conditioning, fire suppression, and ventilation.

Large organizations often maintain more than one data center to distribute data processing workloads and provide backup if a disaster occurs.

1.3.1 Core Elements of a Data Center

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

- **Application:** A computer program that provides the logic for computing operations
- **Database management system (DBMS):** Provides a structured way to store data in logically organized tables that are interrelated
- **Host or compute:** A computing platform (hardware, firmware, and software) that runs applications and databases
- **Network:** A data path that facilitates communication among various networked devices
- **Storage:** 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.

In this book, host, compute, and server are used interchangeably to represent the element that runs applications.

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

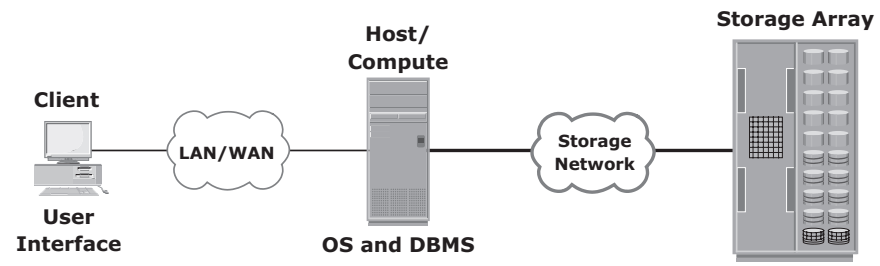


Figure 1-5: Example of an online order transaction system

A customer places an order through a client machine connected over a LAN/WAN to a host running an order-processing application. The client accesses the DBMS on the host through the application to provide order-related information, such as the customer name, address, payment method, products ordered, and quantity ordered.

The DBMS uses the host operating system to write this data to the physical disks in the storage array. The storage networks provide the communication link between the host and the storage array and transports the request to read or write data between them. The storage array, after receiving the read or write request from the host, performs the necessary operations to store the data on physical disks.

1.3.2 Key Characteristics of a Data Center

Uninterrupted operation of data centers is critical to the survival and success of a business. Organizations must have a reliable infrastructure that ensures that data is accessible at all times. Although the characteristics shown in Figure 1-6 are applicable to all elements of the data center infrastructure, the focus here is on storage systems. This book covers the various technologies and solutions to meet these requirements.

- **Availability:** A data center should ensure the availability of information when required. Unavailability of information could cost millions of dollars per hour to businesses, such as financial services, telecommunications, and e-commerce.
- **Security:** Data centers must establish policies, procedures, and core element integration to prevent unauthorized access to information.

- **Scalability:** Business growth often requires deploying more servers, new applications, and additional databases. Data center resources should scale based on requirements, without interrupting business operations.
- **Performance:** All the elements of the data center should provide optimal performance based on the required service levels.
- **Data integrity:** Data integrity refers to mechanisms, such as error correction codes or parity bits, which ensure that data is stored and retrieved exactly as it was received.
- **Capacity:** Data center operations require adequate resources to store and process large amounts of data, efficiently. When capacity requirements increase, the data center must provide additional capacity without interrupting availability or with minimal disruption. Capacity may be managed by reallocating the existing resources or by adding new resources.
- **Manageability:** A data center should provide easy and integrated management of all its elements. Manageability can be achieved through automation and reduction of human (manual) intervention in common tasks.

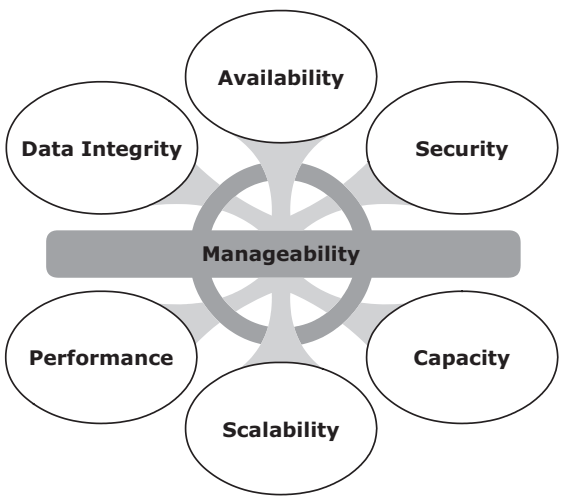


Figure 1-6: Key characteristics of a data center

1.3.3 Managing a Data Center

Managing a data center involves many tasks. The key management activities include the following:

- **Monitoring:** It is a continuous process of gathering information on various elements and services running in a data center. The aspects of a data

center that are monitored include security, performance, availability, and capacity.

- **Reporting:** It 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:** It is a process of providing the hardware, software, and other resources required to run a data center. Provisioning activities primarily include resources management to meet capacity, availability, performance, and security requirements.

Virtualization and cloud computing have dramatically changed the way data center infrastructure resources are provisioned and managed. Organizations are rapidly deploying virtualization on various elements of data centers to optimize their utilization. Further, continuous cost pressure on IT and on-demand data processing requirements have resulted in the adoption of cloud computing.

1.4 Virtualization and Cloud Computing

Virtualization is a technique of abstracting physical resources, such as compute, storage, and network, and making them appear as logical resources. Virtualization has existed in the IT industry for several years and in different forms. Common examples of virtualization are virtual memory used on compute systems and partitioning of raw disks.

Virtualization enables pooling of physical resources and providing an aggregated view of the physical resource capabilities. For example, storage virtualization enables multiple pooled storage devices to appear as a single large storage entity. Similarly, by using compute virtualization, the CPU capacity of the pooled physical servers can be viewed as the aggregation of the power of all CPUs (in megahertz). Virtualization also enables centralized management of pooled resources.

Virtual resources can be created and provisioned from the pooled physical resources. For example, a virtual disk of a given capacity can be created from a storage pool or a virtual server with specific CPU power and memory can be configured from a compute pool. These virtual resources share pooled physical resources, which improves the utilization of physical IT resources. Based on business requirements, capacity can be added to or removed from the virtual resources without any disruption to applications or users. With improved utilization of IT assets, organizations save the costs associated with procurement and

management of new physical resources. Moreover, fewer physical resources means less space and energy, which leads to better economics and green computing.

In today's fast-paced and competitive environment, organizations must be agile and flexible to meet changing market requirements. This leads to rapid expansion and upgrade of resources while meeting shrinking or stagnant IT budgets. *Cloud computing*, addresses these challenges efficiently. Cloud computing enables individuals or businesses to use IT resources as a service over the network. It provides highly scalable and flexible computing that enables provisioning of resources on demand. Users can scale up or scale down the demand of computing resources, including storage capacity, with minimal management effort or service provider interaction. Cloud computing empowers self-service requesting through a fully automated request-fulfillment process. Cloud computing enables consumption-based metering; therefore, consumers pay only for the resources they use, such as CPU hours used, amount of data transferred, and gigabytes of data stored.

Cloud infrastructure is usually built upon virtualized data centers, which provide resource pooling and rapid provisioning of resources. Information storage in virtualized and cloud environments is detailed later in the book.

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 attributes of a data center.

The evolution of storage architecture and the core elements of a data center covered in this chapter provided the foundation for information storage and management. The emergence of virtualization has provided the opportunity to transform classic data centers into virtualized data centers. Cloud computing is further changing the way IT resources are provisioned and consumed.

The subsequent chapters in the book provide comprehensive details on various aspects of information storage and management in both classic and virtualized environments. It begins with describing the core elements of a data center with a focus on storage systems and RAID (covered in Chapters 2, 3, and 4). Chapters 5 through 8 of this book detail various storage networking technologies, such as storage area network (SAN), network attached storage (NAS), and object-based and unified storage. Chapters 9 through 12 cover various business continuity solutions, such as backup and replication, along with archival technologies. Chapter 13 introduces cloud infrastructure and services. Chapters 14 and 15 describe securing and managing storage in traditional and virtualized environments.

EXERCISES

1. What is structured and unstructured data? Research the challenges of storing and managing unstructured data.
2. Discuss the benefits of information-centric storage architecture over server-centric storage architecture.
3. What are the attributes of big data? Research and prepare a presentation on big data analytics.
4. Research how businesses use their information assets to derive competitive advantage and new business opportunities.
5. Research and prepare a presentation on personal data management.

Contents

Foreword		**xxvii**
Introduction		**xxix**
Section I	**Storage System**	**1**
Chapter 1	**Introduction to Information Storage**	**3**
	1.1 Information Storage	4
	1.1.1 Data	4
	1.1.2 Types of Data	6
	1.1.3 Big Data	7
	1.1.4 Information	9
	1.1.5 Storage	9
	1.2 Evolution of Storage Architecture	9
	1.3 Data Center Infrastructure	11
	1.3.1 Core Elements of a Data Center	11
	1.3.2 Key Characteristics of a Data Center	12
	1.3.3 Managing a Data Center	13
	1.4 Virtualization and Cloud Computing	14
	Summary	15
Chapter 2	**Data Center Environment**	**17**
	2.1 Application	18
	2.2 Database Management System (DBMS)	18
	2.3 Host (Compute)	19
	2.3.1 Operating System	19
	Memory Virtualization	20
	2.3.2 Device Driver	20
	2.3.3 Volume Manager	20
	2.3.4 File System	22
	2.3.5 Compute Virtualization	25

2.4 Connectivity 27
2.4.1 Physical Components of Connectivity 27
2.4.2 Interface Protocols 28
IDE/ATA and Serial ATA 28
SCSI and Serial SCSI 29
Fibre Channel 29
Internet Protocol (IP) 29
2.5 Storage 29
2.6 Disk Drive Components 31
2.6.1 Platter 32
2.6.2 Spindle 32
2.6.3 Read/Write Head 32
2.6.4 Actuator Arm Assembly 33
2.6.5 Drive Controller Board 33
2.6.6 Physical Disk Structure 34
2.6.7 Zoned Bit Recording 35
2.6.8 Logical Block Addressing 36
2.7 Disk Drive Performance 36
2.7.1 Disk Service Time 37
Seek Time 37
Rotational Latency 37
Data Transfer Rate 38
2.7.2 Disk I/O Controller Utilization 39
2.8 Host Access to Data 40
2.9 Direct-Attached Storage 41
2.9.1 DAS Benefits and Limitations 42
2.10 Storage Design Based on Application Requirements and Disk Performance 43
2.11 Disk Native Command Queuing 45
2.12 Introduction to Flash Drives 46
2.12.1 Components and Architecture of Flash Drives 47
2.12.2 Features of Enterprise Flash Drives 48
2.13 Concept in Practice: VMware ESXi 48
Summary 49

Chapter 3 Data Protection: RAID 51
3.1 RAID Implementation Methods 52
3.1.1 Software RAID 52
3.1.2 Hardware RAID 52
3.2 RAID Array Components 53
3.3 RAID Techniques 53
3.3.1 Striping 53
3.3.2 Mirroring 55
3.3.3 Parity 55
3.4 RAID Levels 57
3.4.1 RAID 0 57
3.4.2 RAID 1 58

3.4.3 Nested RAID 59
3.4.4 RAID 3 62
3.4.5 RAID 4 63
3.4.6 RAID 5 63
3.4.7 RAID 6 64
3.5 RAID Impact on Disk Performance 64
3.5.1 Application IOPS and RAID Configurations 66
3.6 RAID Comparison 66
3.7 Hot Spares 68
Summary 68

Chapter 4 Intelligent Storage Systems 71
4.1 Components of an Intelligent Storage System 72
4.1.1 Front End 72
4.1.2 Cache 72
Structure of Cache 73
Read Operation with Cache 73
Write Operation with Cache 75
Cache Implementation 75
Cache Management 76
Cache Data Protection 77
4.1.3 Back End 78
4.1.4 Physical Disk 78
4.2 Storage Provisioning 79
4.2.1 Traditional Storage Provisioning 79
LUN Expansion: MetaLUN 80
4.2.2 Virtual Storage Provisioning 82
Comparison between Virtual and Traditional Storage Provisioning 82
Use Cases for Thin and Traditional LUNs 84
4.2.3 LUN Masking 84
4.3 Types of Intelligent Storage Systems 85
4.3.1 High-End Storage Systems 85
4.3.2 Midrange Storage Systems 86
4.4 Concepts in Practice: EMC Symmetrix and VNX 87
4.4.1 EMC Symmetrix Storage Array 87
4.4.2 EMC Symmetrix VMAX Component 88
4.4.3 Symmetrix VMAX Architecture 89
Summary 91

Section II Storage Networking Technologies 93

Chapter 5 Fibre Channel Storage Area Networks 95
5.1 Fibre Channel: Overview 96
5.2 The SAN and Its Evolution 97
5.3 Components of FC SAN 98
5.3.1 Node Ports 99
5.3.2 Cables and Connectors 99

5.3.3 Interconnect Devices 100
5.3.4 SAN Management Software 101
5.4 FC Connectivity 102
5.4.1 Point-to-Point 102
5.4.2 Fibre Channel Arbitrated Loop 102
5.4.3 Fibre Channel Switched Fabric 103
FC-SW Transmission 105
5.5 Switched Fabric Ports 106
5.6 Fibre Channel Architecture 106
5.6.1 Fibre Channel Protocol Stack 107
FC-4 Layer 108
FC-2 Layer 108
FC-1 Layer 108
FC-0 Layer 108
5.6.2 Fibre Channel Addressing 109
5.6.3 World Wide Names 109
5.6.4 FC Frame 110
5.6.5. Structure and Organization of FC Data 112
5.6.6 Flow Control 112
BB_Credit 112
EE_Credit 112
5.6.7 Classes of Service 113
5.7 Fabric Services 113
5.8 Switched Fabric Login Types 114
5.9 Zoning 115
5.9.1 Types of Zoning 116
5.10 FC SAN Topologies 118
5.10.1 Mesh Topology 118
5.10.2 Core-Edge Fabric 119
Benefits and Limitations of Core-Edge Fabric 119
5.11 Virtualization in SAN 122
5.11.1 Block-level Storage Virtualization 122
5.11.2 Virtual SAN (VSAN) 124
5.12 Concepts in Practice: EMC Connectrix and EMC VPLEX 125
5.12.1 EMC Connectrix 125
Connectrix Switches 126
Connectrix Directors 126
Connectrix Multi-purpose Switches 126
Connectrix Management Tools 127
5.12.2 EMC VPLEX 127
VPLEX Family of Products 128
Summary 128

Chapter 6 IP SAN and FCoE 131
6.1 iSCSI 132
6.1.1 Components of iSCSI 132
6.1.2 iSCSI Host Connectivity 133

6.1.3 iSCSI Topologies 133
Native iSCSI Connectivity 133
Bridged iSCSI Connectivity 135
Combining FC and Native iSCSI Connectivity 135
6.1.4 iSCSI Protocol Stack 135
6.1.5 iSCSI PDU 136
6.1.6 iSCSI Discovery 138
6.1.7 iSCSI Names 138
6.1.8 iSCSI Session 140
6.1.9 iSCSI Command Sequencing 141
6.2 FCIP 142
6.2.1 FCIP Protocol Stack 142
6.2.2 FCIP Topology 144
6.2.3 FCIP Performance and Security 144
6.3 FCoE 145
6.3.1 I/O Consolidation Using FCoE 145
6.3.2 Components of an FCoE Network 147
Converged Network Adapter 148
Cables 148
FCoE Switches 149
6.3.3 FCoE Frame Structure 150
FCoE Frame Mapping 151
6.3.4 FCoE Enabling Technologies 152
Priority-Based Flow Control (PFC) 153
Enhanced Transmission Selection (ETS) 154
Congestion Notification (CN) 154
Data Center Bridging Exchange Protocol (DCBX) 154
Summary 155

Chapter 7 Network-Attached Storage 157
7.1 General-Purpose Servers versus NAS Devices 158
7.2 Benefits of NAS 159
7.3 File Systems and Network File Sharing 160
7.3.1 Accessing a File System 160
7.3.2 Network File Sharing 160
7.4 Components of NAS 162
7.5 NAS I/O Operation 163
7.6 NAS Implementations 163
7.6.1 Unified NAS 164
7.6.2 Unified NAS Connectivity 164
7.6.3 Gateway NAS 164
7.6.4 Gateway NAS Connectivity 165
7.6.5 Scale-Out NAS 166
7.6.6 Scale-Out NAS Connectivity 167
7.7 NAS File-Sharing Protocols 168
7.7.1 NFS 169
7.7.2 CIFS 170

7.8 Factors Affecting NAS Performance 171
7.9 File-Level Virtualization 174
7.10 Concepts in Practice: EMC Isilon and EMC VNX Gateway 175
7.10.1 EMC Isilon 175
7.10.2 EMC VNX Gateway 176
Summary 177

Chapter 8 Object-Based and Unified Storage 179
8.1 Object-Based Storage Devices 180
8.1.1 Object-Based Storage Architecture 181
8.1.2 Components of OSD 182
8.1.3 Object Storage and Retrieval in OSD 183
8.1.4 Benefits of Object-Based Storage 184
8.1.5 Common Use Cases for Object-Based Storage 185
8.2 Content-Addressed Storage 187
8.3 CAS Use Cases 188
8.3.1 Healthcare Solution: Storing Patient Studies 188
8.3.2 Finance Solution: Storing Financial Records 189
8.4 Unified Storage 190
8.4.1 Components of Unified Storage 190
Data Access from Unified Storage 192
8.5 Concepts in Practice: EMC Atmos, EMC VNX, and EMC Centera 192
8.5.1 EMC Atmos 193
8.5.2 EMC VNX 194
8.5.3 EMC Centera 195
EMC Centera Architecture 196
Summary 197

Section III Backup, Archive, and Replication 199

Chapter 9 Introduction to Business Continuity 201
9.1 Information Availability 202
9.1.1 Causes of Information Unavailability 202
9.1.2 Consequences of Downtime 203
9.1.3 Measuring Information Availability 204
9.2 BC Terminology 205
9.3 BC Planning Life Cycle 207
9.4 Failure Analysis 210
9.4.1 Single Point of Failure 210
9.4.2 Resolving Single Points of Failure 211
9.4.3 Multipathing Software 212
9.5 Business Impact Analysis 213
9.6 BC Technology Solutions 213
9.7 Concept in Practice: EMC PowerPath 214
9.7.1 PowerPath Features 214
9.7.2 Dynamic Load Balancing 215

I/O Operation without PowerPath 215
I/O Operation with PowerPath 216
9.7.3 Automatic Path Failover 217
Path Failure without PowerPath 218
Path Failover with PowerPath: Active-Active Array 218
Path Failover with PowerPath: Active-Passive Array 219
Summary 221

Chapter 10 Backup and Archive 225
10.1 Backup Purpose 226
10.1.1 Disaster Recovery 226
10.1.2 Operational Recovery 226
10.1.3 Archival 226
10.2 Backup Considerations 227
10.3 Backup Granularity 228
10.4 Recovery Considerations 231
10.5 Backup Methods 231
10.6 Backup Architecture 233
10.7 Backup and Restore Operations 234
10.8 Backup Topologies 236
10.9 Backup in NAS Environments 239
10.9.1 Server-Based and Serverless Backup 239
10.9.2 NDMP-Based Backup 240
10.10 Backup Targets 242
10.10.1 Backup to Tape 243
Physical Tape Library 243
Limitations of Tape 245
10.10.2 Backup to Disk 245
10.10.3 Backup to Virtual Tape 246
Virtual Tape Library 246
10.11 Data Deduplication for Backup 249
10.11.1 Data Deduplication Methods 249
10.11.2 Data Deduplication Implementation 250
Source-Based Data Deduplication 250
Target-Based Data Deduplication 250
10.12 Backup in Virtualized Environments 252
10.13 Data Archive 254
10.14 Archiving Solution Architecture 255
10.14.1 Use Case: E-mail Archiving 256
10.14.2 Use Case: File Archiving 257
10.15 Concepts in Practice: EMC NetWorker, EMC Avamar, and EMC Data Domain 257
10.15.1 EMC NetWorker 258
10.15.2 EMC Avamar 258
10.15.3 EMC Data Domain 259
Summary 260

Chapter 11 Local Replication 263
11.1 Replication Terminology 264
11.2 Uses of Local Replicas 264
11.3 Replica Consistency 265
11.3.1 Consistency of a Replicated File System 265
11.3.2 Consistency of a Replicated Database 266
11.4 Local Replication Technologies 269
11.4.1 Host-Based Local Replication 269
LVM-Based Replication 269
Advantages of LVM-Based Replication 269
Limitations of LVM-Based Replication 270
File System Snapshot 271
11.4.2 Storage Array-Based Local Replication 272
Full-Volume Mirroring 273
Pointer-Based, Full-Volume Replication 274
Pointer-Based Virtual Replication 277
11.4.3 Network-Based Local Replication 278
Continuous Data Protection 279
CDP Local Replication Operation 280
11.5 Tracking Changes to Source and Replica 281
11.6 Restore and Restart Considerations 282
11.7 Creating Multiple Replicas 283
11.8 Local Replication in a Virtualized Environment 284
11.9 Concepts in Practice: EMC TimeFinder, EMC SnapView, and EMC RecoverPoint 285
11.9.1 EMC TimeFinder 285
TimeFinder/Clone 286
TimeFinder/Snap 286
11.9.2 EMC SnapView 286
SnapView Snapshot 286
SnapView Clone 287
11.9.3 EMC RecoverPoint 287
Summary 287

Chapter 12 Remote Replication 289
12.1 Modes of Remote Replication 289
12.2 Remote Replication Technologies 292
12.2.1. Host-Based Remote Replication 292
LVM-Based Remote Replication 293
Host-Based Log Shipping 294
12.2.2 Storage Array-Based Remote Replication 295
Synchronous Replication Mode 295
Asynchronous Replication Mode 296
Disk-Buffered Replication Mode 297
12.2.3 Network-Based Remote Replication 298
CDP Remote Replication 298
12.3 Three-Site Replication 300

12.3.1 Three-Site Replication — Cascade/Multihop 300
Synchronous + Asynchronous 300
Synchronous + Disk Buffered 302
12.3.2 Three-Site Replication — Triangle/Multitarget 302
12.4 Data Migration Solutions 304
12.5 Remote Replication and Migration in a Virtualized Environment 306
12.6 Concepts in Practice: EMC SRDF, EMC MirrorView, and EMC RecoverPoint 307
12.6.1 EMC SRDF 308
12.6.2 EMC MirrorView 308
12.6.3 EMC RecoverPoint 308
Summary 309

Section IV Cloud Computing 311

Chapter 13 Cloud Computing 313
13.1 Cloud Enabling Technologies 314
13.2 Characteristics of Cloud Computing 314
13.3 Benefits of Cloud Computing 316
13.4 Cloud Service Models 316
13.4.1 Infrastructure-as-a-Service 316
13.4.2 Platform-as-a-Service 317
13.4.3 Software-as-a-Service 318
13.5 Cloud Deployment Models 318
13.5.1 Public Cloud 318
13.5.2 Private Cloud 319
13.5.3 Community Cloud 320
13.5.4 Hybrid Cloud 321
13.6 Cloud Computing Infrastructure 322
13.6.1 Physical Infrastructure 322
13.6.2 Virtual Infrastructure 323
13.6.3 Applications and Platform Software 324
13.6.4 Cloud Management and Service Creation Tools 324
13.7 Cloud Challenges 326
13.7.1 Challenges for Consumers 326
13.7.2 Challenges for Providers 327
13.8 Cloud Adoption Considerations 327
13.9 Concepts in Practice: Vblock 329
Summary 330

Section V Securing and Managing Storage Infrastructure 331

Chapter 14 Securing the Storage Infrastructure 333
14.1 Information Security Framework 334
14.2 Risk Triad 334
14.2.1 Assets 335
14.2.2 Threats 336

14.2.3 Vulnerability 337
14.3 Storage Security Domains 338
14.3.1 Securing the Application Access Domain 339
Controlling User Access to Data 340
Protecting the Storage Infrastructure 341
Data Encryption 342
14.3.2 Securing the Management Access Domain 342
Controlling Administrative Access 344
Protecting the Management Infrastructure 344
14.3.3 Securing Backup, Replication, and Archive 345
14.4 Security Implementations in Storage Networking 346
14.4.1 FC SAN 346
FC SAN Security Architecture 347
Basic SAN Security Mechanisms 347
LUN Masking and Zoning 349
Securing Switch Ports 349
Switch-Wide and Fabric-Wide Access Control 350
Logical Partitioning of a Fabric: Virtual SAN 350
14.4.2 NAS 350
NAS File Sharing: Windows ACLs 351
NAS File Sharing: UNIX Permissions 352
NAS File Sharing: Authentication and Authorization 353
Kerberos 354
Network-Layer Firewalls 355
14.4.3 IP SAN 357
14.5 Securing Storage Infrastructure in Virtualized and Cloud Environments 358
14.5.1 Security Concerns 359
14.5.2 Security Measures 359
Security at the Compute Level 359
Security at the Network Level 360
Security at the Storage Level 361
14.6 Concepts in Practice: RSA and VMware Security Products 361
14.6.1 RSA SecureID 362
14.6.2 RSA Identity and Access Management 362
14.6.3 RSA Data Protection Manager 362
14.6.4 VMware vShield 363
Summary 363

Chapter 15 Managing the Storage Infrastructure 365
15.1 Monitoring the Storage Infrastructure 366
15.1.1 Monitoring Parameters 366
15.1.2 Components Monitored 367
Hosts 367
Storage Network 368
Storage 369

15.1.3 Monitoring Examples 369
Accessibility Monitoring 369
Capacity Monitoring 370
Performance Monitoring 372
Security Monitoring 374
15.1.4 Alerts 375
15.2 Storage Infrastructure Management Activities 376
15.2.1 Availability Management 376
15.2.2 Capacity Management 376
15.2.3 Performance Management 377
15.2.4 Security Management 377
15.2.5 Reporting 378
15.2.6 Storage Infrastructure Management in a Virtualized Environment 378
15.2.7 Storage Management Examples 380
Example 1: Storage Allocation to a New Server/Host 380
Example 2: File System Space Management 381
Example 3: Chargeback Report 382
15.3 Storage Infrastructure Management Challenges 384
15.4 Developing an Ideal Solution 384
15.4.1 Storage Management Initiative 385
15.4.2 Enterprise Management Platform 386
15.5 Information Lifecycle Management 386
15.6 Storage Tiering 388
15.6.1 Intra-Array Storage Tiering 388
15.6.2 Inter-Array Storage Tiering 390
15.7 Concepts in Practice: EMC Infrastructure Management Tools 391
15.7.1 EMC ControlCenter and Prosphere 391
15.7.2 EMC Unisphere 392
15.7.3 EMC Unified Infrastructure Manager (UIM) 393
Summary 393

Appendix A Application I/O Characteristics 395
Random and Sequential 395
Reads and Writes 395
I/O Request Size 396

Appendix B Parallel SCSI 399
SCSI Standards Family 400
SCSI Client-Server Model 401
Parallel SCSI Addressing 402

Appendix C SAN Design Exercises 405
Exercise 1 405
Solution 405
Exercise 2 406
Solution 406

Appendix D Information Availability Exercises **409**
Exercise 1 409
Solution 409
Exercise 2 410
Solution 410

Appendix E Network Technologies for Remote Replication **411**
DWDM 411
CWDM 412
SONET 412

Appendix F Acronyms and Abbreviations **413**

Glossary **427**

Index **465**

Information Storage and Management Second Edition

Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments

More than ever, the IT industry is challenged with employing and developing highly skilled technical professionals with storage technology expertise across classic, virtualized, and cloud environments. This book covers concepts, principles, and deployment considerations across technologies that are used for storing and managing information.

Key Technology Strategies for Classic, Virtualized, and Cloud Environments:

- Challenges and Solutions for Data Storage and Management
- Intelligent Storage, Object-Based Storage, and Unified Storage
- Storage Networking, Federation, and Protocols
- Backup, Recovery, Deduplication, and Archive
- Business Continuity and Disaster Recovery
- Cloud Computing and Converged Infrastructure
- Storage Security and Managing Storage Infrastructure

EMC Proven™ Professional is a leading education and certification program in the IT industry, providing comprehensive coverage of information storage technologies, virtualization, cloud computing, data science/big data analytics, and more.

Being Proven means investing in yourself and formally validating your expertise!

This book prepares you for the Information Storage and Management exam E10-001 leading to EMC Proven Professional Information Storage Associate v2 certification.

Visit http://education.EMC.com for details.

EMC is a global leader in enabling businesses and service providers to transform their operations and deliver IT as a service. Fundamental to this transformation is cloud computing. Through innovative products and services, EMC accelerates the journey to cloud computing, helping IT departments to store, manage, protect and analyze their most valuable asset — information — in a more agile, trusted and cost-efficient way. Additional information about EMC can be found at www.EMC.com

ISBN 978-1-118-09483-9
90000
9 781118 094839

WILEY
wiley.com

Also available as an e-book

COMPUTERS/
Networking/General