MODULE - I

Introduction to Information Storage and Management

MODULE 1

Introduction to Information Storage and Management

Module Description

Individuals and businesses now have access to more and more content generating devices. The importance, volume and dependency of information continue to grow at astounding rates. Information, when created, resides locally on cameras, laptops, or mobile phones. This is then uploaded to data centres via networks and shared with others.

This module explains the concept of information storage and the evolution of storage infrastructure. By the end of this module, students will learn about data, information and storage. They will be able to identify the core elements and key requirements for data centres. In addition to these skills students will also able to discuss the major activities involved with the management of data centres.

Chapter Table of Contents

Chapter 1.1

Introduction to Information Storage and Management

Aim

To equip the students in the fundamentals of data storage

Instructional Objectives

After completing this chapter, you should be able to:

- Explain information storage
- Describe data and its various types
- Outline the evolution of storage technology and architecture
- Explain the core elements of data centre infrastructure
- Describe the key requirements of data centre elements

Learning Outcomes

At the end of this chapter, you are expected to:

- Identify different stages of information storage
- Summarise the core elements of data centre
- Outline the key characteristics of data centre elements
- Discuss data centre infrastructure
- Outline the activities involved in managing storage infrastructure

1.1.1 Introduction

Managing the ever-growing flood of data is one of the greatest challenges as far as storage is concerned. Not only is it important to store data appropriately, it is equally important to ensure its privacy. As the storage and privacy of data becomes increasingly intertwined with business operations, it affects not only major businesses and their data centres, but also smaller companies and SME businesses.

In this chapter, you will learn about the types of data, the distinction between data and information and the evolution of storage technologies over the years. In addition, you will study the core elements of data centres and their key requirements. You will also learn to identify and describe the major tasks and activities involved in the management of growing and complex data centres.

1.1.2 Information Storage

Information storage is a series of processes executed in an organised way to collect and arrange data so that they can be available for use at will. Since we are living in the age of information, each and every action of ours results in some form of data creation. While internet has fuelled the rate of data generation, the advent of social media has propelled it to another level altogether. Over 4 million Facebook posts and over 300 thousand tweets are being generated per minute. Similarly, a consulting firm such as Aon records 40 million transactions per day. Internet and easy to manage data storage devices have made the information portable. Each and every decision has an impact on almost every sector of business and every individual of the society is benefitted by the quick access of information. We access internet for our day to day needs like banking, travel planning, food ordering, news, social networking and exchange of personal and corporate e-mails. The information is stored locally on laptops, cameras, flash drives, mobiles and so on.

With the increase in the amount and availability of data, the need to create an intelligent analysis and access to data is becoming vital for businesses. In recent times, some business applications like Customer Relationship Management (CRM), Big Data and Enterprise Resource Planning (ERP) are aiming to address the need of organising this data in a useful pattern to enable easier decision-making for a business or an individual. Corporate intranets, e-mail, e-commerce, business-to-business, data warehousing, computer-aided design/ computer-aided manufacturing (CAD/CAM), voice/video/data convergence and many more business scores are benefiting from this organised data.

Businesses use data to derive information that is imperative to run their everyday operations. The information becomes intelligence when it is given a user friendly view and format and it gains value when shared with others. The storage of this data is the repository which enables users to store, protect and retrieve the information when required.

(i) Data

Data is a collection of raw facts from which the required output is extracted. A paper document, photograph, a movie on a disk, a message, a bank account holder's passbooks, buyer and seller agreements and papers of property are all examples of data. Earlier, the techniques implemented for the generation and sharing of data were inadequate for some forms, such as documents and movie in a reel. With the advancement of technology, the same form of data can now be converted into multiple forms such as pictures, videos, or e-mails. The rate of data generation and sharing has increased exponentially due to the advancement of computer, Internet and communication equipment.

Some of the factors that have contributed to the growth of digital data are as follows:

- **Flexible data processing technologies:** Data processing technologies should develop to fulfil the demanding business requirements. To further evolve the process of data generation and growth, old architectures, difficult and risky transitions and complex deployments should be eliminated.

- **Affordable cost of digital storage:** The low cost of the storage devices has provided an affordable and easy solution to the data storage problem.

- **Faster communication technology:** The rate of sharing digital data is now much faster and it is growing every day. It now takes just a few seconds to capture a picture and share the same.

One of the biggest issues businesses are facing at present is the rapid growth of data. The amount of data that organisations store has grown exponentially over the past two decades. Data capacity in enterprises grows on average at the rate of 40% to 60% every year. Corporates of all sizes are struggling to deal with the increasing amount of data. However, the advancement of technology has increased the storage capacity of networks while at the same time decreased the cost of digital storage.

(ii) Types of Data

The classification of data is based on the way it is stored and managed. There are three types of data:

1. **Structured:** Structured data is very commonplace and often managed using Structured Query Language (SQL) – a programming language created for querying and managing data in relational database management systems (RDBMS). It has a relational key and can be easily planned into pre-calculated fields. Structured data is the most processed in development and the simplest way to manage information. However, structured data represents only 10% to 20% of all informatics data. Structured data has the advantage of being easily entered, stored and queried. Some examples of machine-generated structured data are Global Positioning System (GPS), medical devices, smart meters, data from networks, biometric devices and stock trading data. Some examples of human-generated structured data are phone numbers (and the phone book), census records (birth, income, employment, place), library catalogues (date, author, place, subject) and gaming related data.

2. **Unstructured:** Unstructured data includes data that is not classified and organised. It represents around 80% of data available in the world (as shown in Figure 1.1.1). It often includes text and multimedia content. E-mail messages, range of documents, Web Pages, audio files, photos, videos, presentations and many other types of business artefacts are examples of unstructured data. This type of data is internally structured, however, it is still considered unstructured because it does not fit in rows and columns of a database. Some examples of machine-generated unstructured data are seismic imagery, atmospheric data, high energy physics, photographs and videos. Some examples of human-generated unstructured data are text within documents, logs, survey results and e-mails. Data generated by users on social media platforms such as YouTube, Facebook, Twitter, LinkedIn, blogs and forums are all unstructured data.

3. **Semi-structured:** Semi-structured data is the intersection of structured and unstructured data. Although it is a type of structured data, it does not have a strict data model structure. Tags or other types of markers are used to recognise certain elements within the data but it does not have a firm structure. In semi-structured data, the entities typically belong to the same class and are grouped together, but may have different attributes. The order of these attributes is not considered important. There has been an increase in the occurrence of semi-structured data since the advent of the Internet. ***For example,*** e-mails have the sender,

recipient, date and time as the fixed fields whereas the body of the email messages is unstructured content. XML and other mark-up languages are often used to manage semi-structured data. This type of data has some organisational properties that make it easier to organise and retrieve the information. Electronic Data Interchange (EDI), e-mails and JavaScript Object Notation (JSON) are all forms of semi-structured data. Semi-structured data formats are particularly useful for hierarchical or nested data. With this kind of data, it is easier to avoid object-relational impedance mismatch and complicated translations of lists into a relational data model.

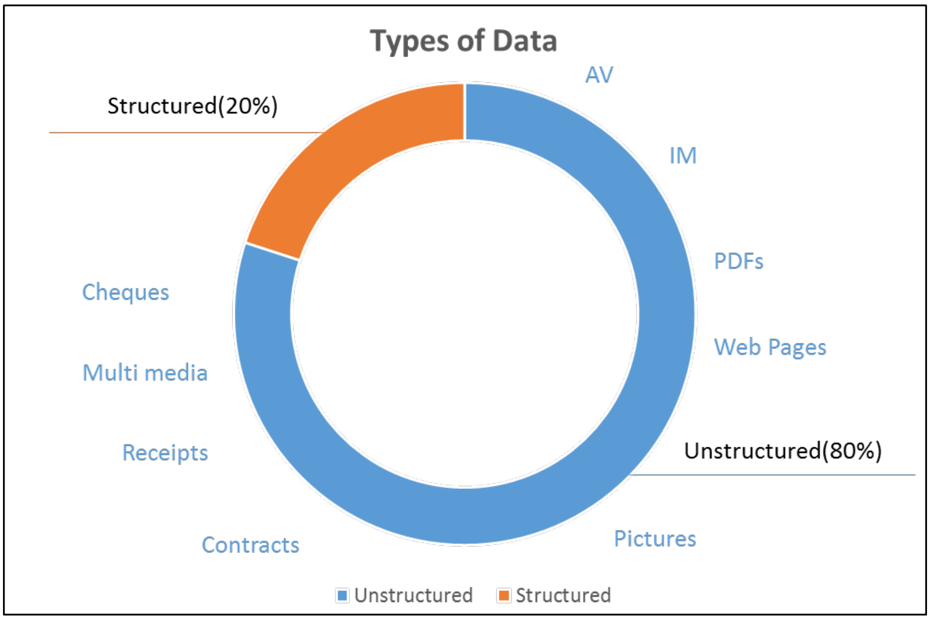

Figure 1.1.1: Types of Data

Did you Know?

SQL was developed by IBM in the early 1970s and later developed by Relational Software, Inc. which is now known as the Oracle Corporation.

(iii) Information

Information is the intelligence and knowledge that is derived from data. It is classified data which has some intrinsic value for the receiver. Businesses analyse raw data in order to identify meaningful trends on which decisions and actions are made. To form a productive business decision, the information available must be accurate and complete. Effective data analysis not only extends to existing businesses, but also creates the potential for new opportunities by using the information in a productive way. For instance, organisations post available positions on job search sites while job seekers post their resumes on websites offering job search facilities. Job-matching software then matches keywords from the resumes with the job postings, thereby using data and turning it into information for job seekers and employers.

Information provides value to businesses by:

- Identifying new business opportunities.
- Identifying buying/spending patterns.
- Reducing cost of product inventory, stocking, shipment and delivery.
- Identifying new services such as security alerts for "stolen" credit card purchases.
- Helping create targeted marketing campaigns.
- Creating a competitive advantage.

Difference between Data and Information

The major difference between data and information is that data is raw material that needs to be processed while information is the processed data. Individual pieces of data are usually not useful on their own. When these are put into context, they become information.

Table 1.1.1: Difference between Data and Information

Data	Information
Data is a collection of raw, unorganised facts. It is not specific unless it is being interpreted.	Information is processed, organised and structured data. It is the output that is specific enough to generate meaning.
Data can be numbers, characters, text, sound, pictures, video and figure that is not processed or put into context.	Information is usually formatted in a manner that can be understood by a human.
Data is the input language for a computer.	Information is the output language for humans. Computers use programming scripts, formulas, or software applications to turn data into information.
Example: Jones,W,1022,Circle,BN,QLD,4169, 61455840840	**Example:** W Jones 1022 Circle Brisbane, Queensland 4169 (61) 455840840

(iv) Storage

Storage, or storage devices, refers to the devices in a computing environment that are designed for storing data. This makes data easily accessible for further processing. The kind of storage used depends on the type of data and the frequency at which it is created and used. Businesses typically use internal hard disks, tapes and external disk arrays to store data. Devices such as hard disks in personal computers, DVDs, CD-ROMs and memory in a cell phone or digital camera are examples of storage devices, as shown in Figure 1.1.2.

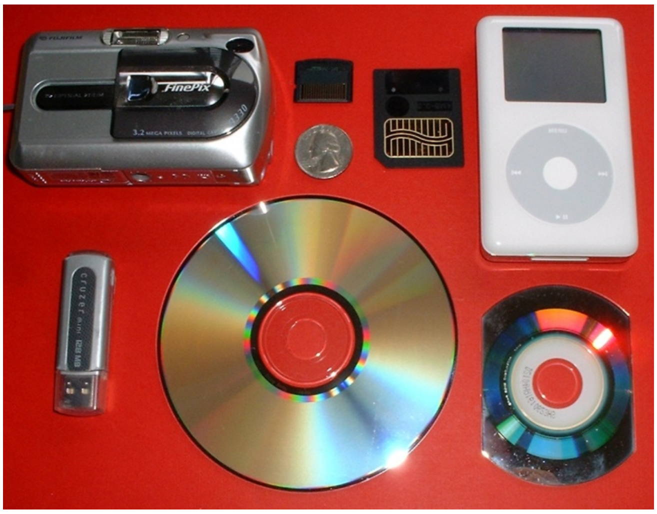

Figure 1.1.2: Examples of Storage Devices

Self-assessment Questions

1) Which of the following have contributed to the growth of digital data? (Choose all that apply.)

    a) Flexible data processing technologies

    b) Globalisation of markets

    c) Faster communication technology

    d) Advances in financial theories

2) Which of the following is the correct definition of information?

    a) Intelligence and knowledge derived from data

    b) Setting up of protocols for communication

    c) Exchange of information between two or more devices

    d) A collection of raw facts from which required output is extracted

3) Which of the following statements are true for structured data? (Choose all that apply.)

    a) It represents around 80% of data available in the world.

    b) It represents only 10% to 20% of all informatics data.

    c) It the most processed in development and the simplest way to manage information.

    d) It is classified and organised but does not have a strict data model structure.

1.1.3 Evolution of Storage Technology and Architecture

Initially information was stored in floppy disks, tape reels and disk packs. These computer storage devices are now almost extinct. Individual storage devices now include external hard drives, pen drives and DVDs. Organisations, historically, had centralised computers (mainframe) in their data centre. The evolution of open systems made it affordable and easy for businesses to have their own servers and storage. In early implementations of open systems, the storage was usually internal to the server.

Storage technology has evolved through the following configurations:

- **Redundant Array of Independent Disks (RAID) –** The term RAID was originally defined as redundant array of inexpensive disks and now it refers to redundant array of independent disks. This type of storage was developed to address the cost, fault tolerance and performance and increase storage capacity in a system. It continues to evolve and is used in all storage architectures such as Direct Attached Storage, Storage Area Network and so on. The basic idea behind RAID is to combine multiple paces and disk drives into an array to achieve performance or redundancy objectives which was not attainable with one large and expensive drive. This array of drives appears to the computer as a single logical storage unit or drive, as shown in Figure 1.1.3.

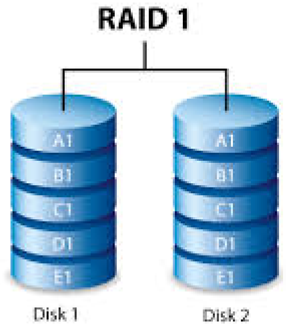

Figure 1.1.3: RAID

- **Direct-Attached Storage (DAS) –** This kind of storage connects directly to a computer or a group of servers. DAS can indicate a single drive or a group of drives that are connected together, as in a RAID array. Storage can be either internal or external to the server. External DAS reduces the limited capacity challenges of internal storage. The most common example of DAS is the internal hard drive in a laptop or desktop personal computer (PC). Multiple systems can use the same DAS device, as long as each server or PC has a separate connection to the storage device, as shown in Figure 1.1.4.

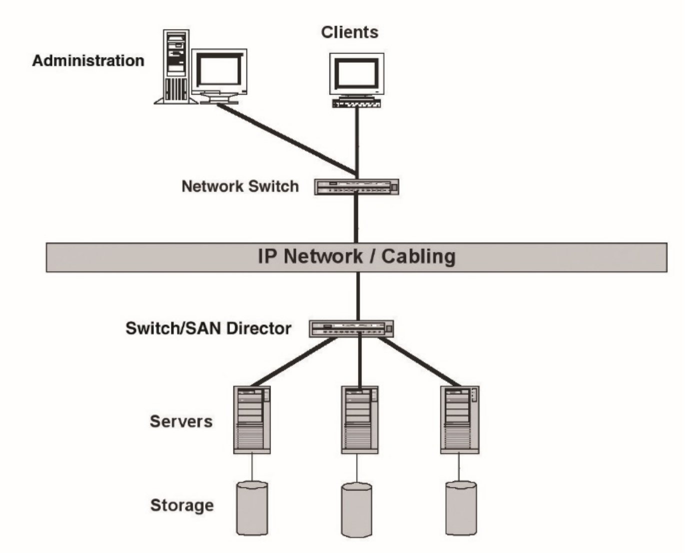

Figure 1.1.4: Direct-Attached Storage

- **Storage Area Network (SAN) –** This is a high-speed network of storage devices. It also links those storage devices with servers. Storage is segregated and assigned to a server for accessing its data. It offers block-level storage that can be retrieved by the applications running on any network server. SANs are particularly helpful in disaster recovery and backup settings. Within a SAN, it is possible to transfer data from one storage device to another without interacting with a server, as shown in Figure 1.1.5. In addition, many SANs utilise networking protocols or Fibre Channel technology that allows the networks to access longer distances geographically. That makes it more feasible for businesses to keep their backup data in remote locations.

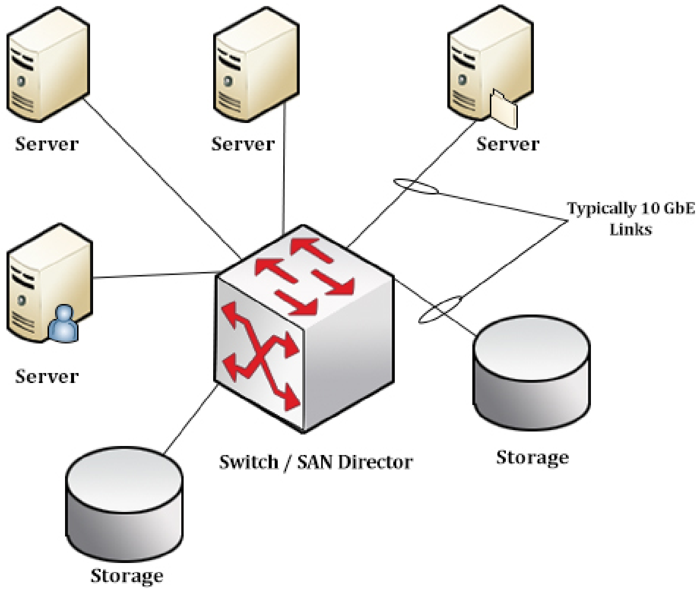

Figure 1.1.5: Storage Area Network

- **Network-Attached Storage (NAS) –** An NAS device is a storage device connected to a network that allows storage and retrieval of data from a centralised location, as shown in Figure 1.1.6. It is mainly required for file sharing. NAS devices are preferred for small businesses as they are simple to operate, safe and reliable. The advantage of centralising data storage is that it lowers costs, is easy to use for back up of data and is always accessible when needed. NAS does not provide any of the tasks that a server in a server-centric system usually provides, such as email, authentication, or file management.

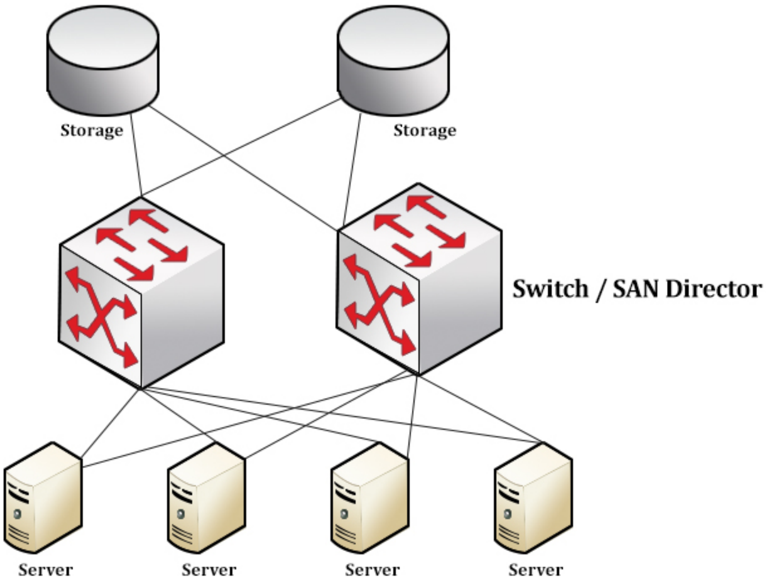

Figure 1.1.6: Network-Attached Storage

An NAS device does not need to be located within the server. It can exist anywhere in a local area network (LAN) and can be made up of multiple networked NAS devices.

- **Internet Protocol SAN (IP-SAN) –** IP was developed as an open standard with a complete design of components. The storage protocols designed to shift block-based data between a host server and storage array include the Fibre Channel over IP (FCIP), Internet Fibre Channel Protocol (iFCP) and Internet Small Computer Systems Interface (iSCSI). The most common type of IP SAN uses iSCSI to summarise SCSI commands and assemble data into packets for transfer between the storage devices and host servers. One of the latest evolutions in storage architecture, IP-SAN is a merging of technologies used in SAN and NAS.

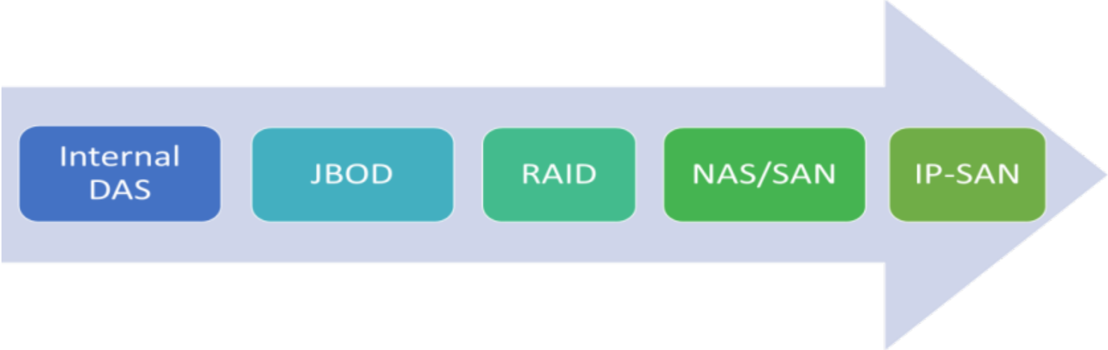

Figure 1.1.7: Evolution of Data Storage

Did you Know?

Using NAS, it is possible to add more hard disk storage space to a network that already utilises servers without shutting them down for maintenance and upgrades.

Self-assessment Questions

4) What is RAID?

   a) It connects directly to a computer or a group of servers.

   b) It is a high-speed network of storage devices.

   c) It combines multiple paces and disk drives into an array.

   d) It is connected to a network that allows retrieval of data from a centralised location.

5) Which of the following statements is true for NAS devices?

   a) They do not enable server-centric tasks such as email, authentication, or file management.

   b) They use iSCSI to summarise SCSI commands and assemble data into packets.

   c) They offer block-level storage that can be retrieved by the applications running on any network server.

   d) They can be either internal or external to the server.

6) Which of the following storage technologies facilitates transfer of data from one storage device to another without interacting with a server?

   a) IP-SAN b) SAN

   c) DAS d) NAS

1.1.4 Data Centre Infrastructure

Organisations maintain data centres to provide integrated data processing facility across the system. Data centres store and manage a large volume of mission-critical data. The data centre infrastructure includes storage systems, computers, dedicated power backups, network devices and environmental controls. Large organisations frequently maintain more than one data centre. This enables them to provide backups in the event of a disaster and distribute data processing workloads. A combination of various storage architectures is used to meet the storage requirements of a data centre.

(i) Core Elements

The following five core elements are critical for the basic functionality of a data centre:

1. **Application:** This is a computer program which provides the logic for computing operations. Applications can be layered on a database. The database, in turn, can use operating system services to carry out read/write operations to storage devices.

2. **Database:** A database management system (DBMS) offers a structured way to store data in logically organised, interconnected tables. A DBMS facilitates the storage and retrieval of data.

3. **Server and operating system:** This is a computing platform that runs databases and applications.

4. **Network:** This is a data path that enables communication between servers and storage and between customers and servers.

5. **Storage array:** This is a device that stores data for consequent use.

These elements are usually treated as separate entities; however, to address data processing requirements, all five elements must work collectively.

(ii) Key Requirements for Data Centre Elements

For the survival and success of a business, it is imperative for data centres to operate uninterrupted. The following factors need to be considered during the installation of a data centre:

- **Availability:** All data centre elements should be designed to enable accessibility. The ability of users to access data when necessary can have a positive impact on a business.

- **Location and Facility:** Location can have a considerable impact and could serve as an opportunity to enhance the company's ability to address network expectancy, disaster recovery and data control. When determining the location of a data centre, it is also necessary to consider the geographical location with respect to the service provider. This helps to ensure continuous service availability and uptime.

- **Uptime:** Organisations can lose a lot of money during the downtime due to the inaccessibility of necessary information and resources. The service provider's track record of uptime should be tested and it should be determined whether the provider has been able to maintain 100% availability for electrical and mechanical systems in the past. In addition, it is important to know about prior events related to planned or unplanned downtime and the methods by which these instances were addressed by the service provider.

- **Cooling:** The establishment of a data centre involves a combination of utilities, generators and a range of power supplies that emit excessive heat. It is essential to have a robust cooling infrastructure to ensure cost efficiencies and a viable environment to establish the data centre. Organisations need to understand all the facets including redundancy power architecture, electrical and cooling capacity (watts per square foot), power distribution and backup systems.

- **Security:** The security of the data centre is equally important. It is necessary to establish policies and procedures and ensure proper integration of the data centre core elements in order to prevent unauthorised access to information. To maximise protection, the facility should be fabricated with military-grade security measures. This should include onsite security workforces and video surveillance and recording especially at the entrances, exits, roof access and equipment areas.

- **Facility Maintenance:** During the installation of a data centre, it is important to consider the maintenance and management aspects. The provider should have sufficient insight of the facility and its efficiencies to ensure that the entire system is being observed and any issues can be quickly addressed before they cause any outage. Providers must have documented, comprehensive and continually tested procedures to ensure efficient operation.

- **Scalability:** Data centre operations should be able to allocate additional on-demand storage or processing capabilities, without interrupting business operations. For the growth of a business, it is often necessary to install more servers, additional databases and new applications. The storage solution should also able to grow with the business. ***For example,*** an organisation might need just 10 servers today for storage of business tasks including virtualisation, redundancy, file services, email, databases and analytics. However, the need for servers might change with increase in workforce, projects, clients and other data. It is therefore essential to have a suitable sized data centre with an ample expansion capacity to increase power, network, physical space and storage. Established organisations actively track and report on this concept.

- **Change Management:** Proper guidelines for change management ensure that there is no alteration in the data centre that has not been planned, scheduled, discussed and approved. It is also necessary to provide back out steps or a Plan B. Whether bringing new amendments to the system or making old practices obsolete, data centres must follow the change management process.

Did you Know?

Planning for scalability is an ongoing process.

(iii) Managing Storage Infrastructure

Due to the increase in use of information technology for both professional and personal purposes, the concerns over management of data have gained considerable attention. Modern marketing success is dependent on the ability to measure, monitor and track the activities performed as well as the ability to react quickly and adjust accordingly. In particular, data centres are playing a crucial role in the modern society. They are helping make information available anytime and anywhere. Managing growing and complex data centres involves many tasks and some of the major activities involved are as follows:

- **Monitoring:** It is the methodical and routine collection of information from projects and programs for the following purposes:

    - To gain knowledge from the ongoing projects and improve practices and activities in the future assignments.

    - To take informed decisions on the future projects.

    - To promote empowerment of beneficiaries of the initiative.

  Typically, the security, performance, accessibility and capacity of a data centre are monitored.

- **Reporting:** Reports on utilisation, capacity and resource performance are prepared periodically for data centres. Reporting makes data comprehensible and ready for efficient, accurate and easy analysis. Reporting consists of facts in the form of data that can only be adjusted to a certain extent without compromising the integrity of the information.

- **Provisioning:** This is the process of providing the software, hardware and other resources necessary for operating a data centre. Provisioning activities include resource and capacity planning. Resource planning is the process of assessing and identifying required resources, such as facility, technology and personnel. It ensures that adequate resources are available and user and application requirements can be met. Capacity planning ensures that the application's and user's imminent needs will be addressed in the most controlled and cost-effective manner.

Self-assessment Questions

7) What are the core elements of a data centre infrastructure? (Choose all that apply.)

   a) Location
   b) Security
   c) Storage array
   d) Server and operating system

8) Which of the following are key requirements for data centre elements? (Choose all that apply.)

   a) Activity tracking
   b) Facility maintenance
   c) Change management
   d) Performance monitoring

9) Which of the following activities are involved in the management of data centres? (Choose all that apply.)

   a) Monitoring security, performance, accessibility and capacity
   b) Ensuring data centre is close to the service provider
   c) Ensuring utilities do not emit excessive heat
   d) Reporting on utilisation, capacity and resource performance

Summary

- Information storage is a series of processes executed in an organised way to collect and arrange data so that they can be available for use at will.

- Data is a collection of raw facts from which the required output is extracted.

- The factors that have contributed to the growth of digital data include flexible data processing technologies, affordable cost of digital storage and faster communication technology.

- The three types of data are structured data, unstructured data and semi-structured data.

- Information is the intelligence and knowledge that is derived from data.

- The major difference between data and information is that data is raw material that needs to be processed while information is the processed data.

- Storage technology has evolved through configurations such as RAID, DAS, SAN, NAS and IP-SAN.

- The core elements of a data centre are application, database, server and operating system, network and storage array.

- The key requirements for a data centre are availability, location and facility, uptime, cooling, security, facility maintenance, scalability and change management.

- The management of a storage infrastructure include monitoring the security, performance, accessibility and capacity of a data centre, reporting on utilisation, capacity and resource performance and provisioning for the software, hardware and other necessary resources.

Terminal Questions

1. What are the different types of data?

2. What is the difference between data and information?

3. What are the core elements of a data centre?

4. What are the major activities involved in the management of data centres?

Answer Keys

Self-assessment Questions	
Question No.	Answer
1	a, c
2	a, d
3	b, c
4	c
5	a
6	b
7	c, d
8	b, c
9	a, d

Activity

Activity Type: Online/Offline **Duration:** 30 Minutes

Description:

Perform an online research on the evolution of storage technology and architecture and write an article of 300 words.

Case Study

An organisation is considering a storage infrastructure that provides high availability and is scalable. It is also important for the storage infrastructure to have a high performance for the organisation's mission-critical applications. Which storage topology (SAN, NAS, IP-SAN) will you recommend and why?

Bibliography

e-References

- Network Magazine, Indian Express Newspapers (Bombay) Limited. (September 2003). *Cover Story: Storage Management* Retrieved 12 July 2016, from http://www.networkmagazineindia.com/200309/coverstory01.shtml

- TechTarget. (December 2013). *Storage infrastructure management is still elusive.* Retrieved 12 July 2016, from http://searchstorage.techtarget.com/opinion/Storage-infrastructure-management-is-still-elusive

External Resources

- Somasundaram, G. & Shrivastava, A. (2009) *Information Storage and Management* - Storing, Managing and Protecting Digital Information. *Indianapolis, Ind.: Wiley Pub.*

- *Kimball, R. &Ross, M. (2013).* The data warehouse toolkit - The Definitive Guide to Dimensional Modeling. *Indianapolis, Ind.: Wiley Pub.*

Video Links

Topic	Link
Information Storage and Management	https://www.youtube.com/watch?v=IxSyLQd_nfM https://www.youtube.com/watch?v=RSP07rU_Swo&list=PL7TlPYbQ9d5GeIUO6PIr-tJGiB_r2S5-B
Evolution and Architecture of Storage Technology	https://www.youtube.com/watch?v=eE7Bfw9lFfs https://www.youtube.com/watch?v=YWaspUZa0Nk
Data Centre Infrastructure	https://www.youtube.com/watch?v=zJvRRYsXVSA https://www.youtube.com/watch?v=DlA2p5L_osM
Managing Storage Infrastructure	https://www.youtube.com/watch?v=RSP07rU_Swo&list=PL7TlPYbQ9d5GeIUO6PIr-tJGiB_r2S5-B https://www.youtube.com/watch?v=5hZ2RJ1a0pI

Notes: