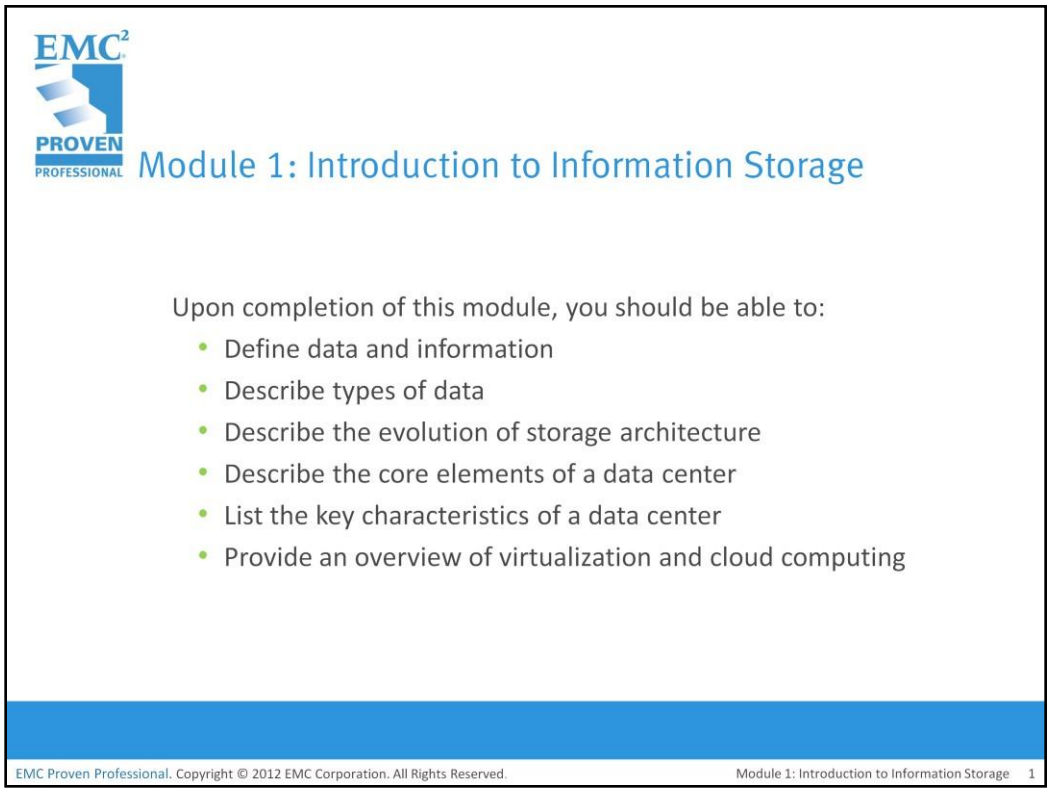
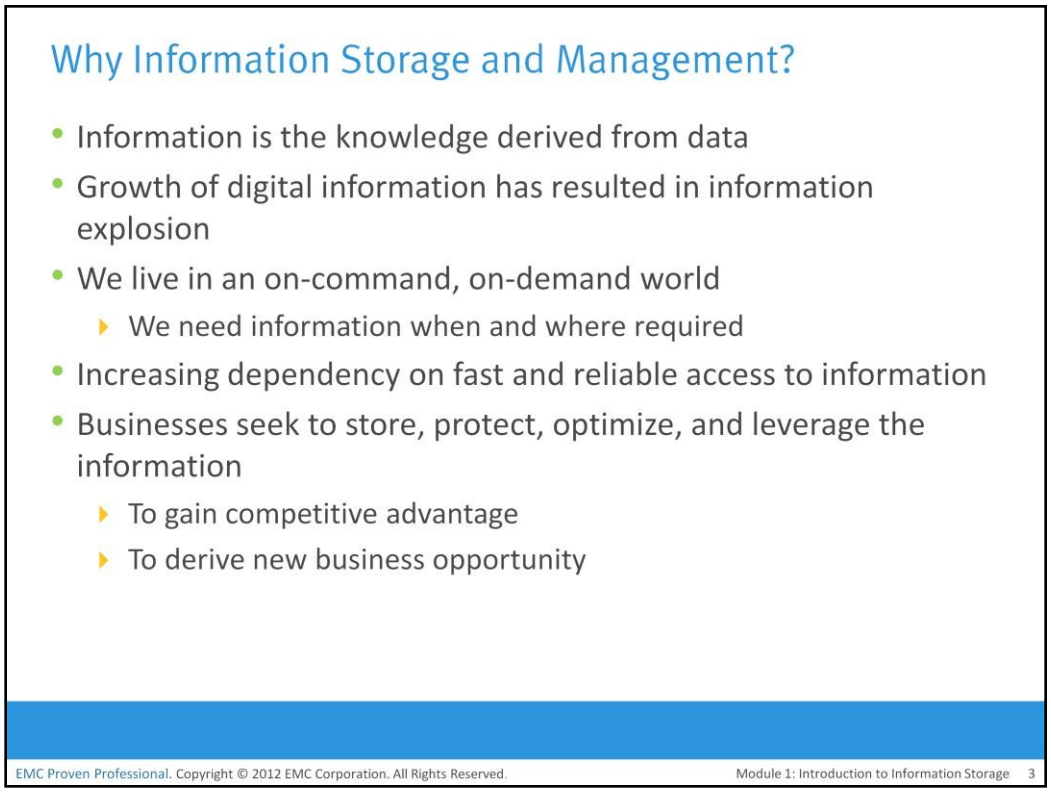
Module – 1

Introduction to Information Storage



This module focuses on the definition of data and information, types of data, and the evolution of storage architecture. It lists the five core elements of a data center and describes the key characteristics of a data center. This module also provides an overview of 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 repository (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.

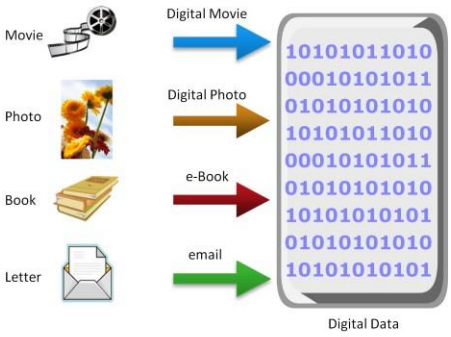
The importance, dependency, and volume of information for the business world also continue to grow at astounding rates. Businesses depend on fast and reliable access to information critical to their success. Examples of business processes or systems that rely on digital information include airline reservations, 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.

What is Data?

Data

It is a collection of raw facts from which conclusions may be drawn.

- Data is converted into more convenient form – digital data
- Factors for digital data growth are:
  - Increase in data-processing capabilities
  - Lower cost of digital storage
  - Affordable and faster communication technology
  - Proliferation of applications and smart devices



Data is a collection of raw facts from which conclusions may 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 examples that contain data.

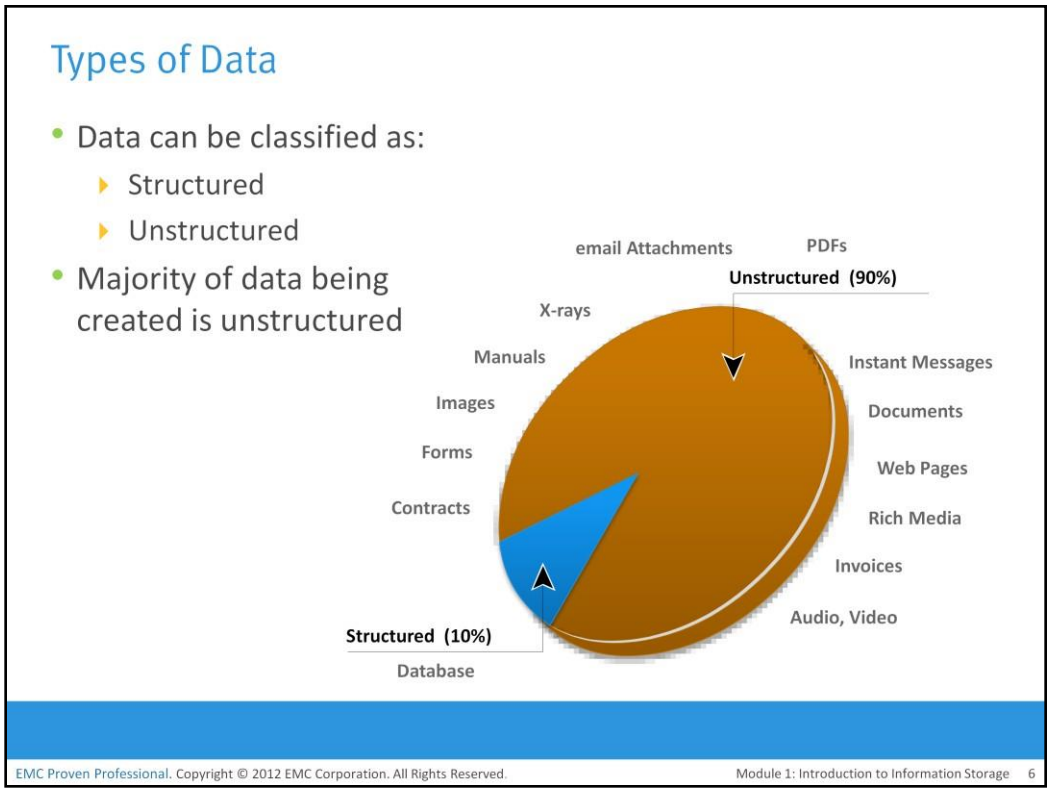
Before the advent of computers, the methods adopted for data creation and sharing were limited to fewer forms, such as paper and film. Today, the same data can be converted into more convenient forms, such as an e-mail message, an e-book, a digital image, or a digital movie. This data can be generated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

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.

Cont..

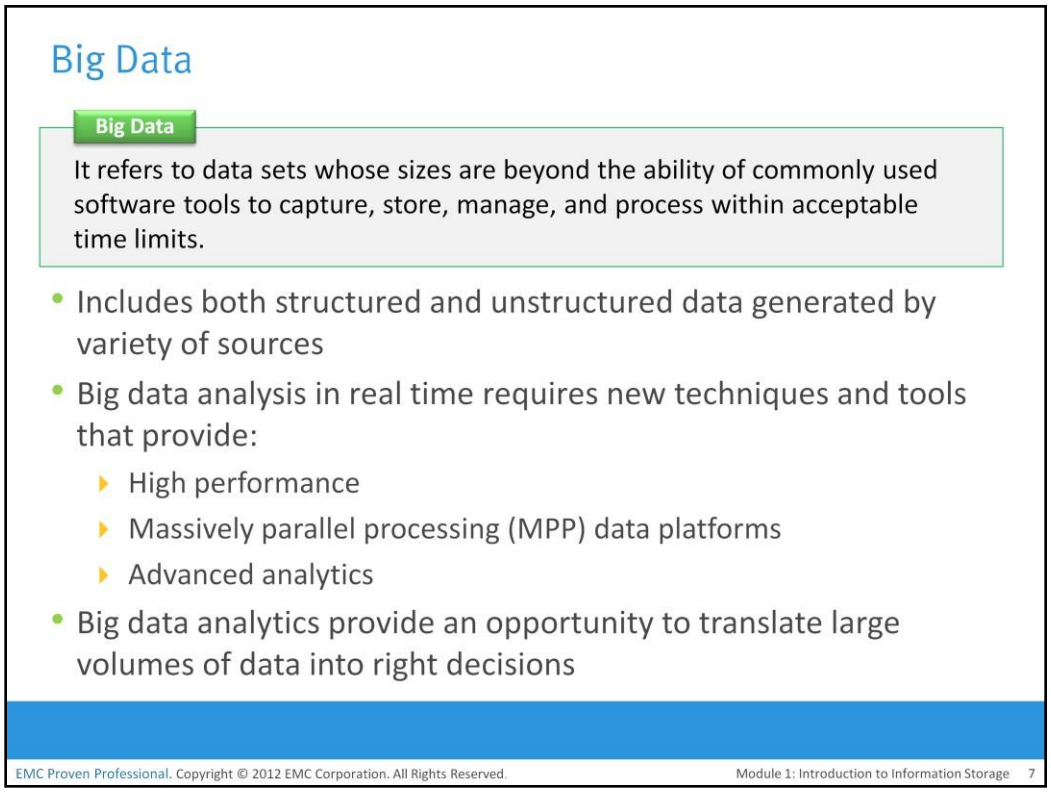
With the advancement of computer and communication technologies, the rate of data generation and sharing has increased exponentially. The following is a list of some of the factors that have contributed to the growth of digital data:

- **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.



Data can be classified as structured or unstructured 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).

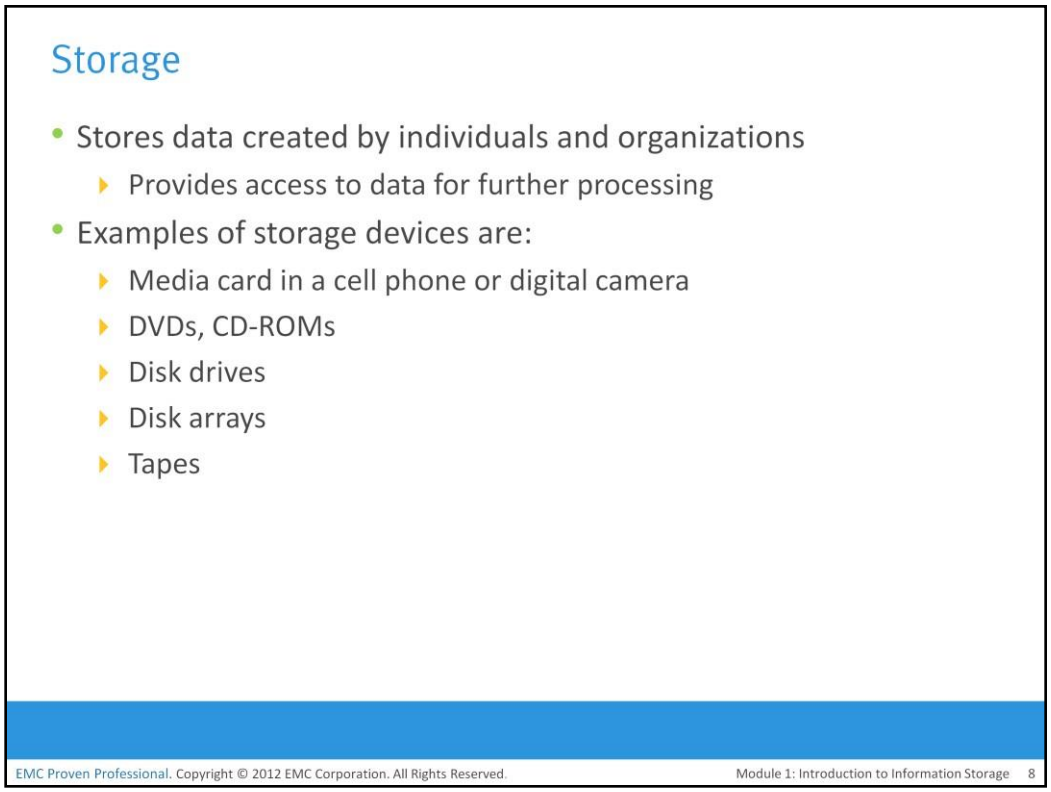
Data is unstructured if its elements cannot be stored in rows and columns, which makes it difficult to query and retrieve by applications. For example, customer contacts that are stored in various forms such as sticky notes, e-mail messages, business cards, or even digital format files, such as .doc, .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 new architectures, technologies, techniques, and skills to store, manage, analyze, and derive value from unstructured data from numerous sources.



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.

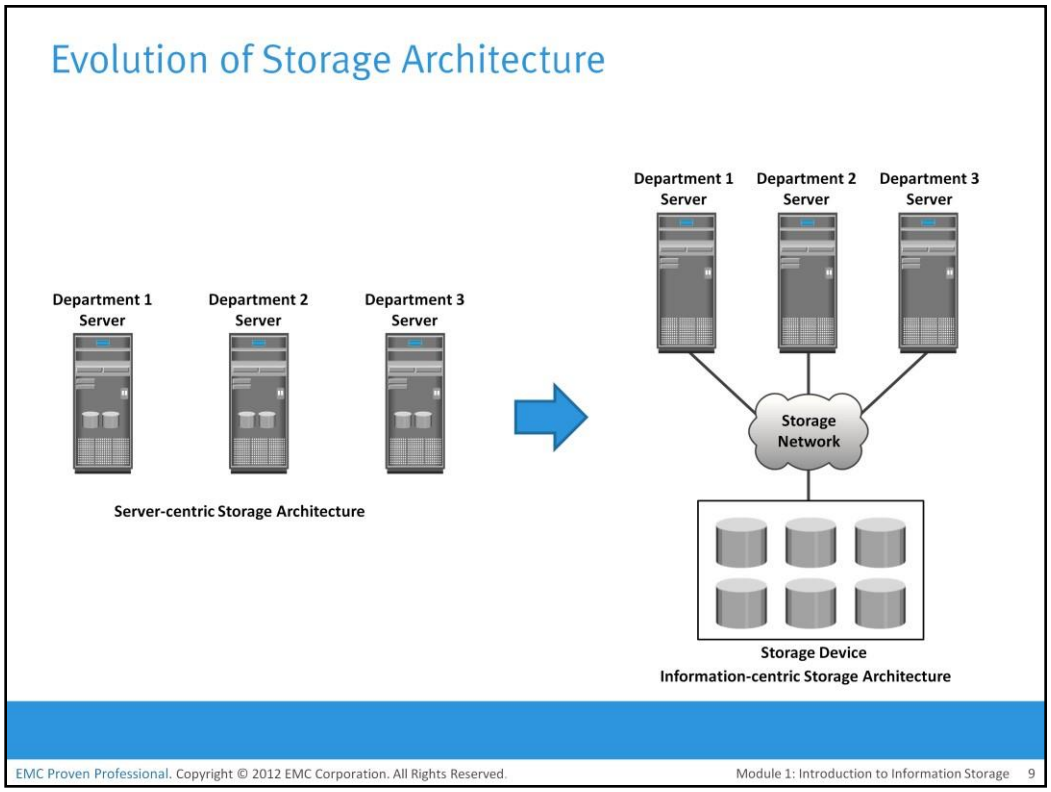
Traditional IT infrastructure and data processing tools and methodologies are inadequate to handle the volume, variety, dynamism, and complexity of big data. Analyzing big data in real time requires new techniques, architectures, and tools that provide high performance, massively parallel processing (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, healthcare, public administration, fraud detection, social media, banks, insurance companies, and other digital information-based entities that benefit from the analytics of big data. The storage architecture required for big data should be simple, efficient, and inexpensive to manage, yet provide access to multiple platforms and data sources simultaneously.



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.

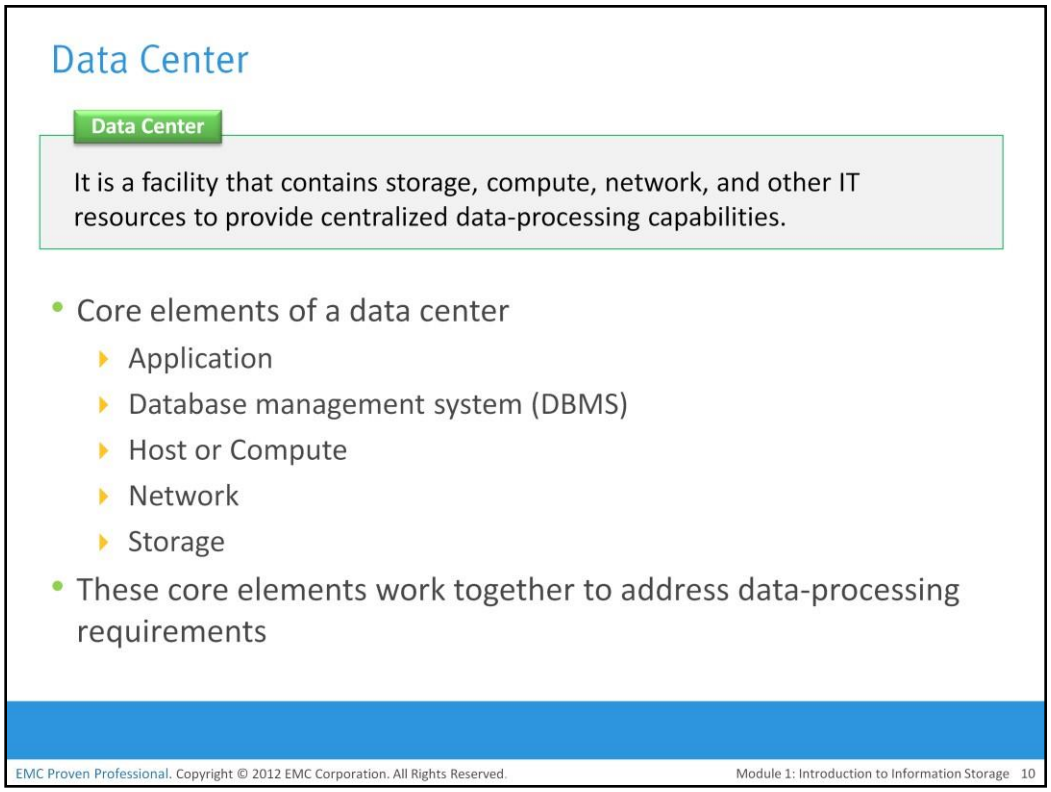
Businesses have several options available for storing data, including internal hard disks, external disk arrays, and tapes.



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 server-centric storage architecture. 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.

To overcome these challenges, storage evolved from server-centric to information-centric architecture. 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.



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.

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.

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

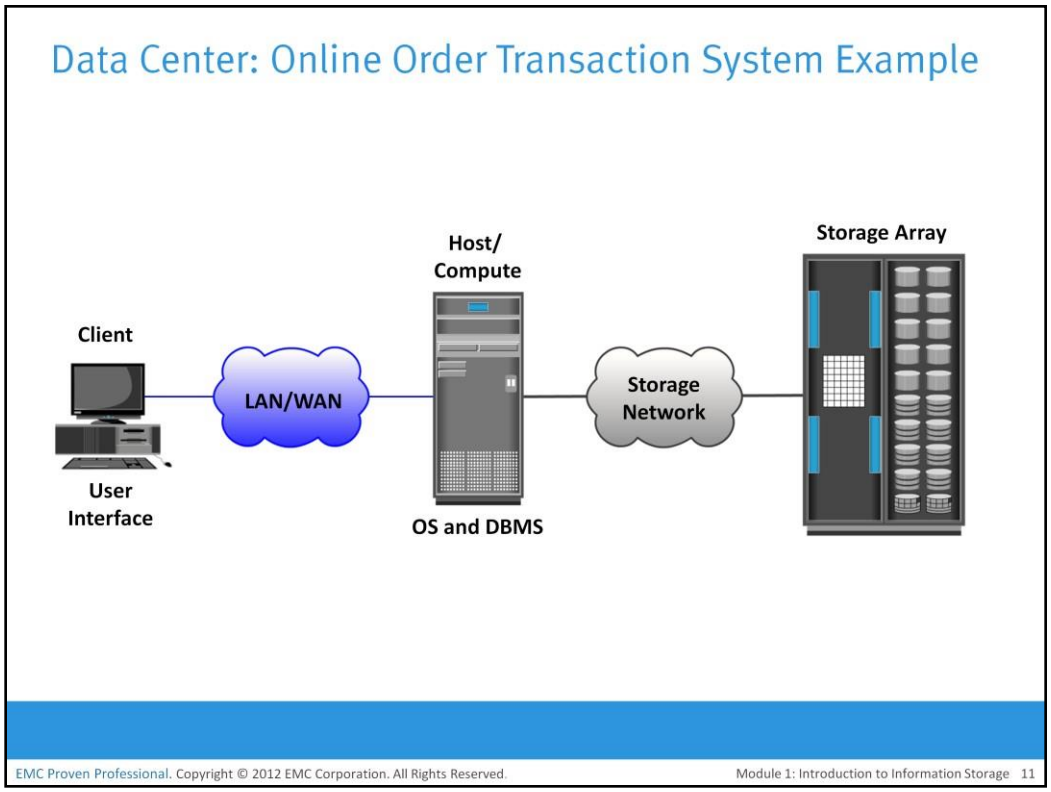
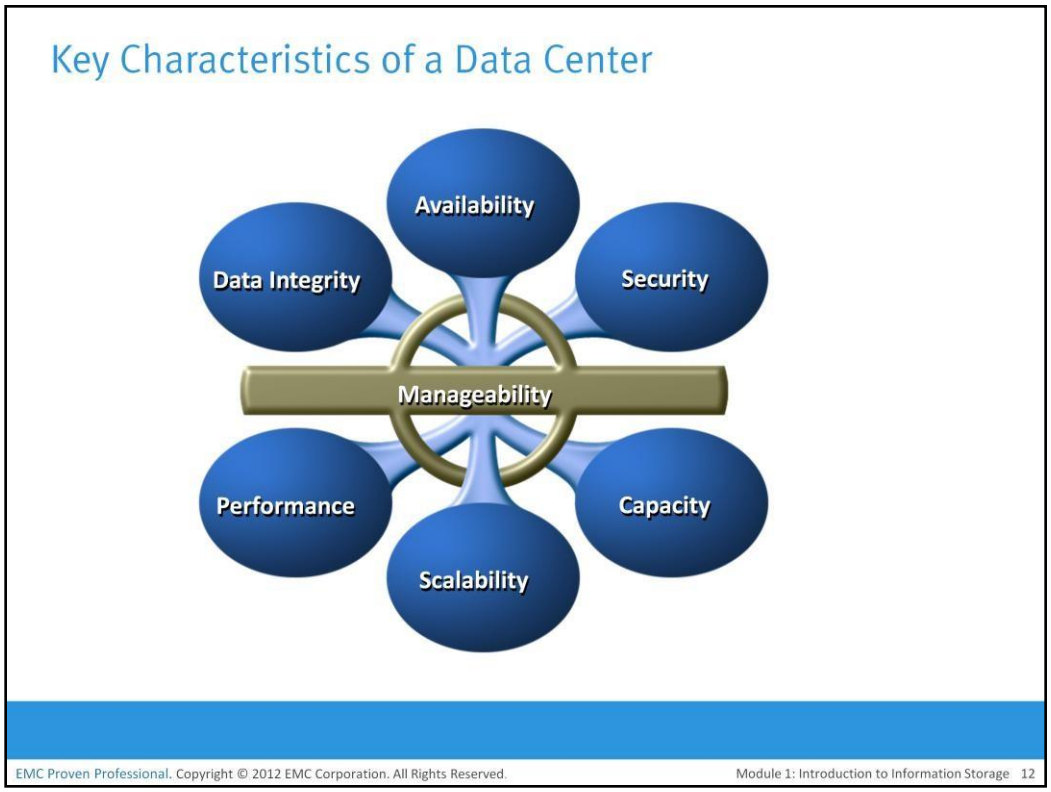


Figure in the slide shows an example of an online order transaction system that involves the
five core elements of a data center and illustrates their functionality in a business process.

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.

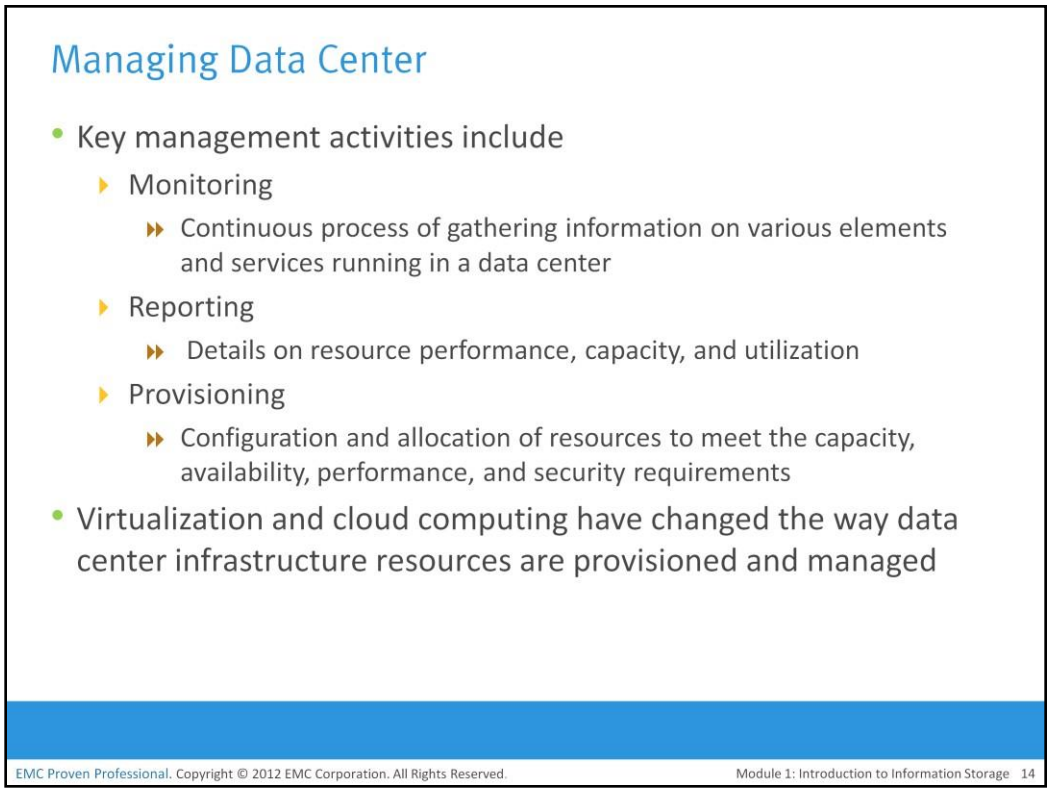


Uninterrupted operation of data centers is critical to the survival and success of a business. Although the characteristics shown in the slide are applicable to all elements of the data center infrastructure, the focus here is on storage systems.

- **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.

Cont..

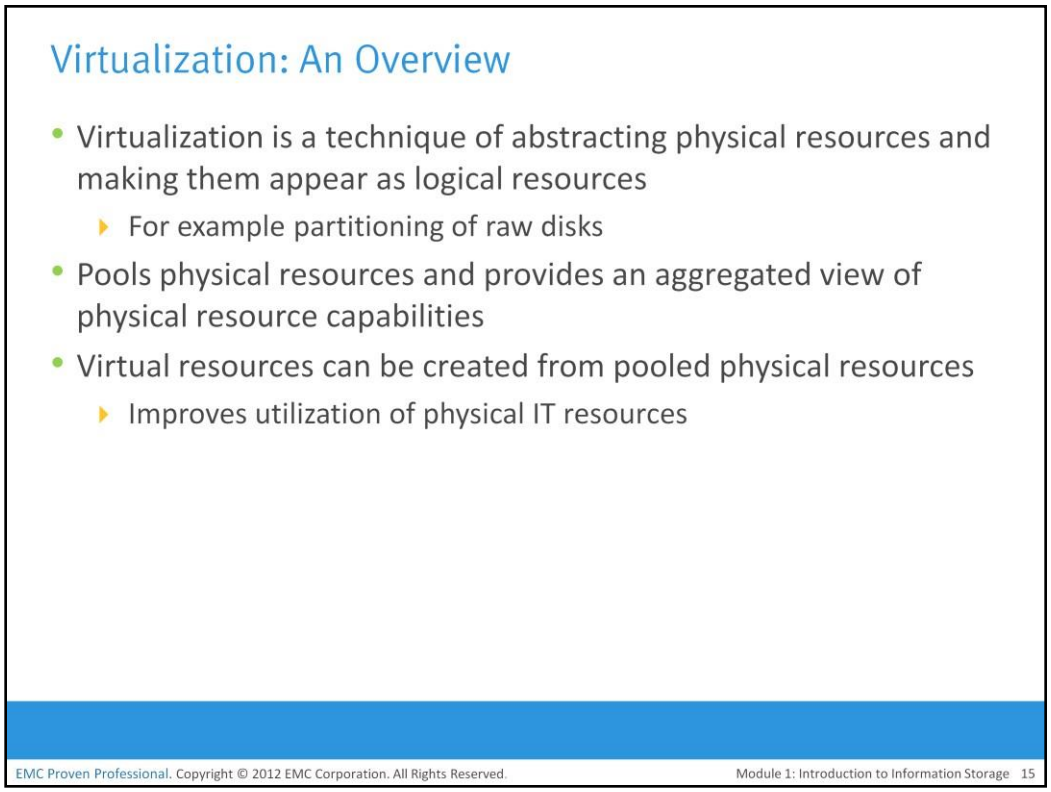
- **Capacity:** Data center operations require adequate resources to store and process large amounts of data, efficiently. When capacity requirements increase, the data center must provide additional capacity without interrupting availability or with minimal disruption. Capacity may be managed by reallocating 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.



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.

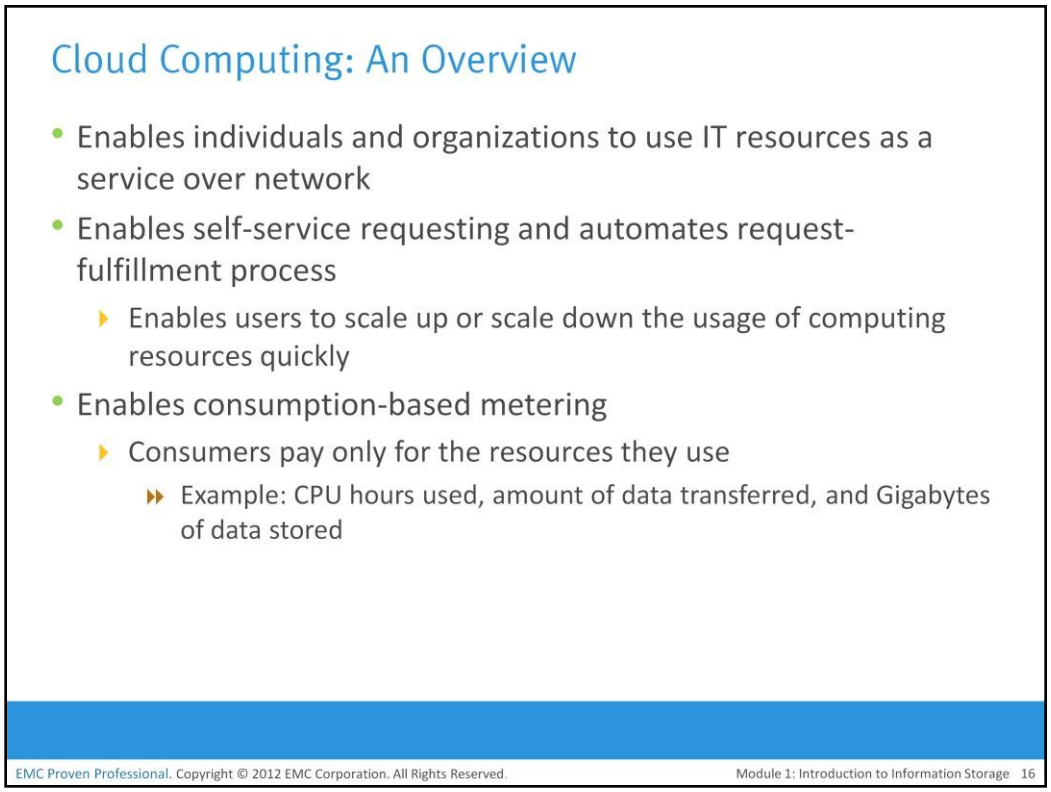
Virtualization and cloud computing have dramatically changed the way data center infrastructure resources are provisioned and managed. Organizations are rapidly deploying virtualization on various elements of data centers to optimize their utilization. Further, continuous cost pressure on IT and on-demand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



Virtualization is a technique of abstracting physical resources, such as compute, storage, and network, and making them appear as logical resources. Virtualization 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.

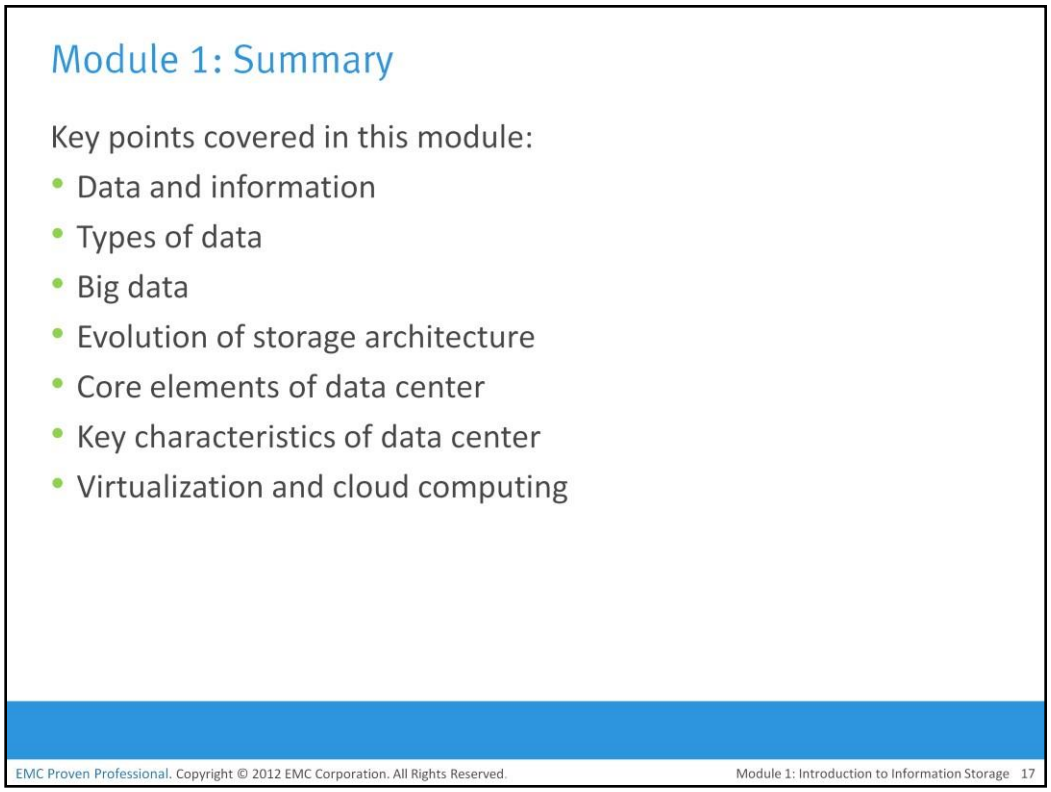
Virtualization enables pooling of physical resources and providing an aggregated view of the physical resource capabilities. For example, storage virtualization enables multiple pooled storage devices to appear as a single large storage entity. Similarly, by using compute virtualization, the CPU capacity of the pooled physical servers can be viewed as 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 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 this course.



This module covered the definition of data and information. Data is a collection of raw facts from which conclusions may be drawn and 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.

Data can be classified as structured and unstructured. Big data refers to data sets whose sizes are beyond the ability of commonly used software tools to capture, store, manage, and process within acceptable time limits.

Information-centric architecture is commonly deployed in today's data center. It helps to overcome the challenges of server-centric storage architecture.

A data center has five core elements such as application, database management system
(DBMS), host, network, and storage.

The key characteristics of data are availability, security, scalability, performance, data integrity, capacity, and manageability.

Virtualization is a technique of abstracting physical resources, such as compute, storage, and network, and making them appear as logical resources.

Cloud computing enables individuals or businesses to use IT resources as a service over the network.

Check Your Knowledge – 1

- Which is an example of structured data?
  - A. Image
  - B. PDF document
  - C. Database
  - D. Web page

- Which is true about big data?
  - A. Includes only unstructured data
  - B. Includes data from a single source
  - C. Captured efficiently using traditional software tools
  - D. Data size is beyond the capability of traditional software to process

Check Your Knowledge – 2

- Which is a feature of information-centric architecture?
  - A. Storage is internal to the servers
  - B. Prevents sharing of storage among servers
  - C. Consists of server, network, and storage in a single system
  - D. Storage is managed centrally and independent of servers

- What accurately describes virtualization?
  - A. Provides on-demand, metered services
  - B. Abstracts physical resources into logical resources
  - C. Pools logical resources to provide data integrity
  - D. Enables decentralized management across data centers

Check Your Knowledge – 3

- Which requirement refers to the ability of a storage solution to grow with the business?
  - A. Availability
  - B. Manageability
  - C. Integrity
  - D. Scalability