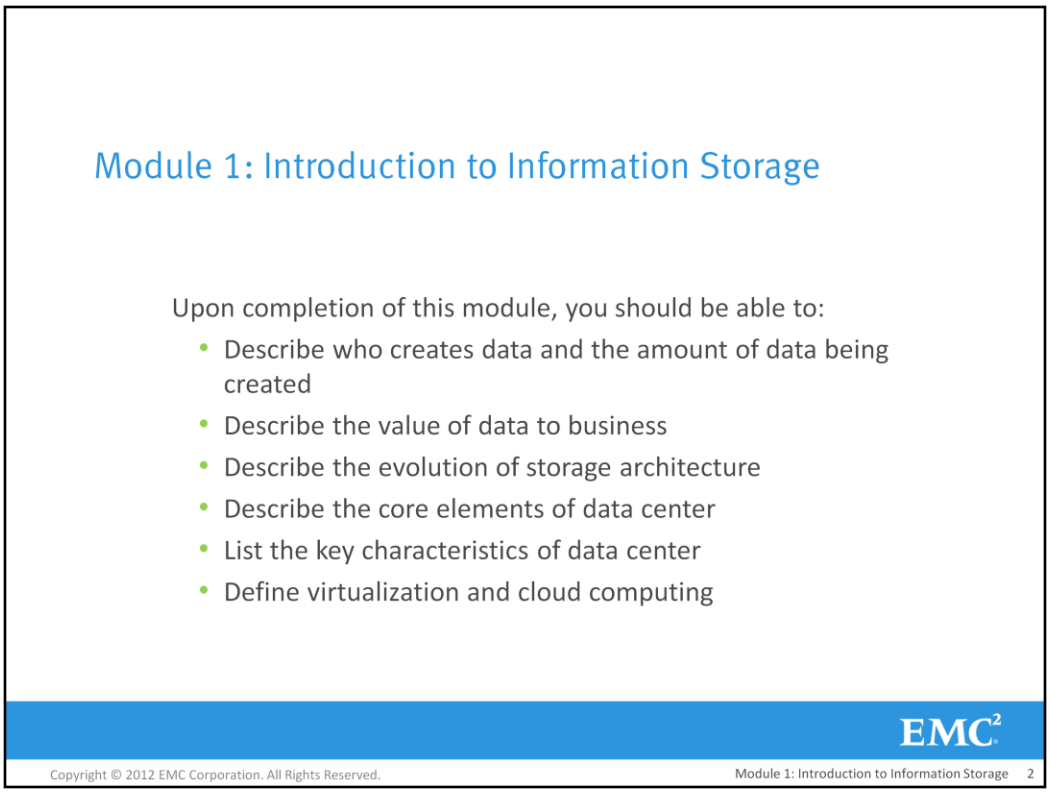
Module – 1

Introduction to Information Storage

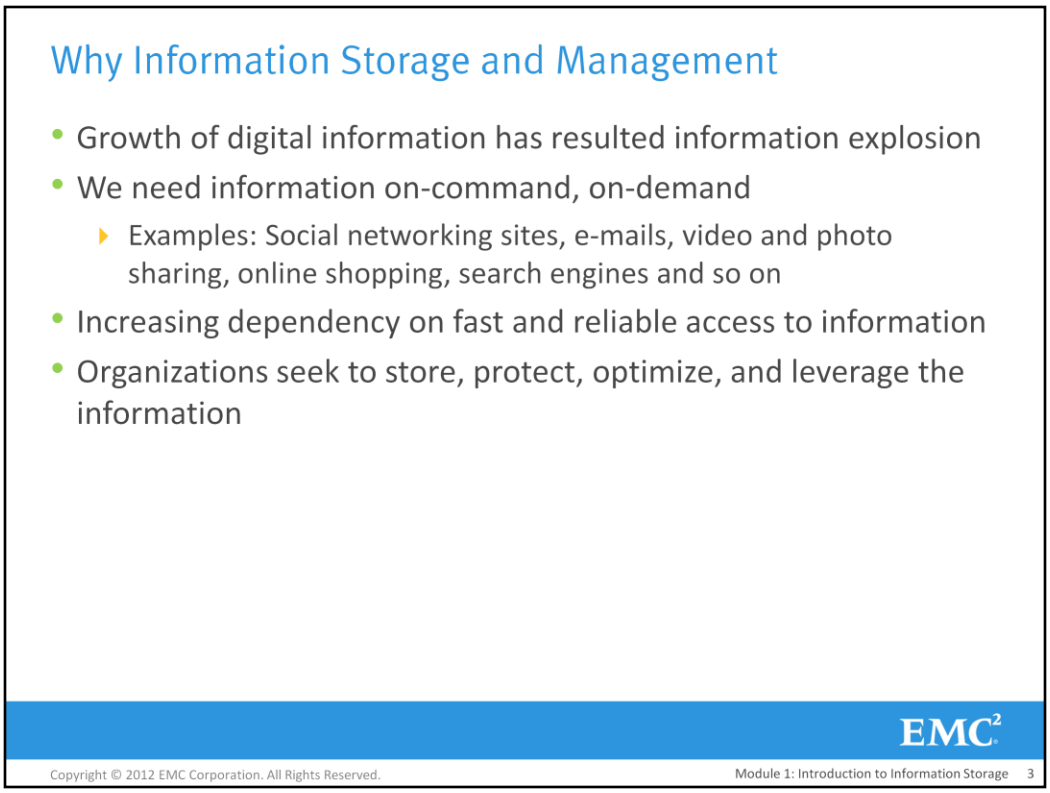


This module describes the amount of data is created by individuals and businesses. Both individuals and businesses process this data to derive meaningful information out of the data. Typically, businesses analyze data to identify meaningful trends. On the basis of these trends, a businesses plan or modify its strategy.

This module also describes the evolution of storage architecture from server centric architecture to information centric architecture.

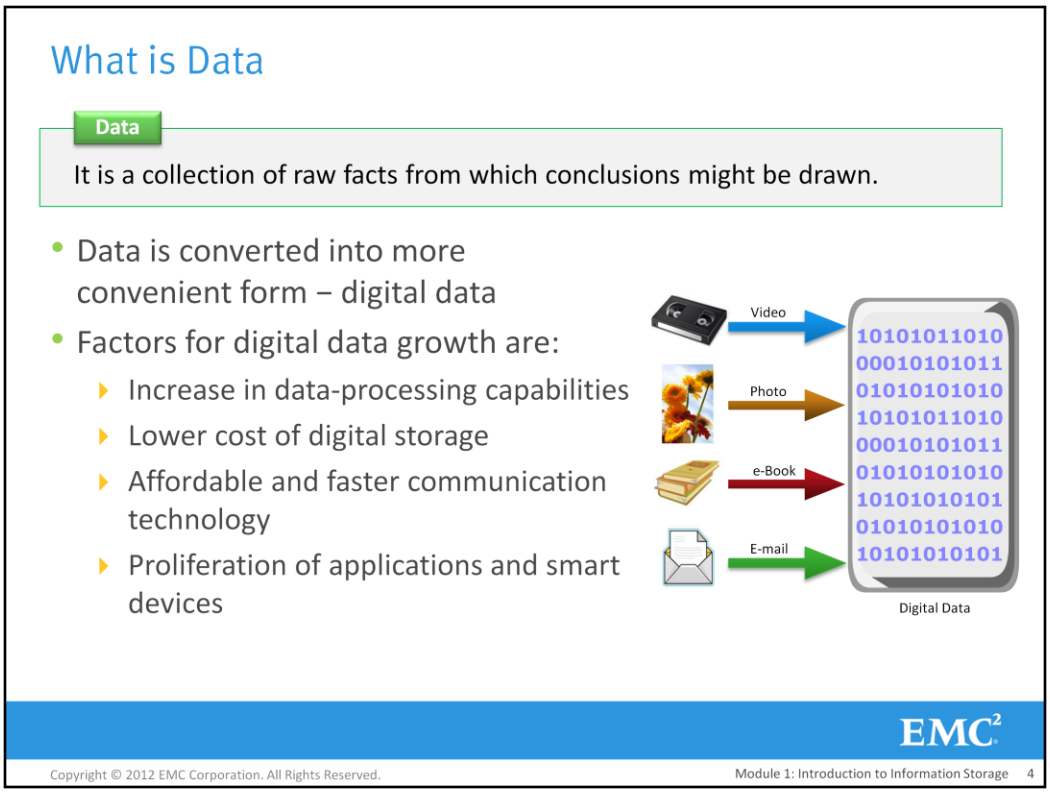
Further, this module also describes the five core elements of data centers and lists the key characteristics od data center.

Finally, this module introduces the concept of virtualization and cloud computing.



Information is increasingly important in our daily lives. We have become information dependents of 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 scores of other applications. Equipped with a growing number of content-generating devices, more information is being created by individuals than by businesses. Information created by individuals gains value when shared with others. When created, information resides locally on devices, such as cell phones, smart phones, tablets, cameras, and laptops. To share this information, it needs to be uploaded via networks to data centers. It is interesting to note that while the majority of information is created by individuals, it is stored and managed by a relatively small number of organizations.

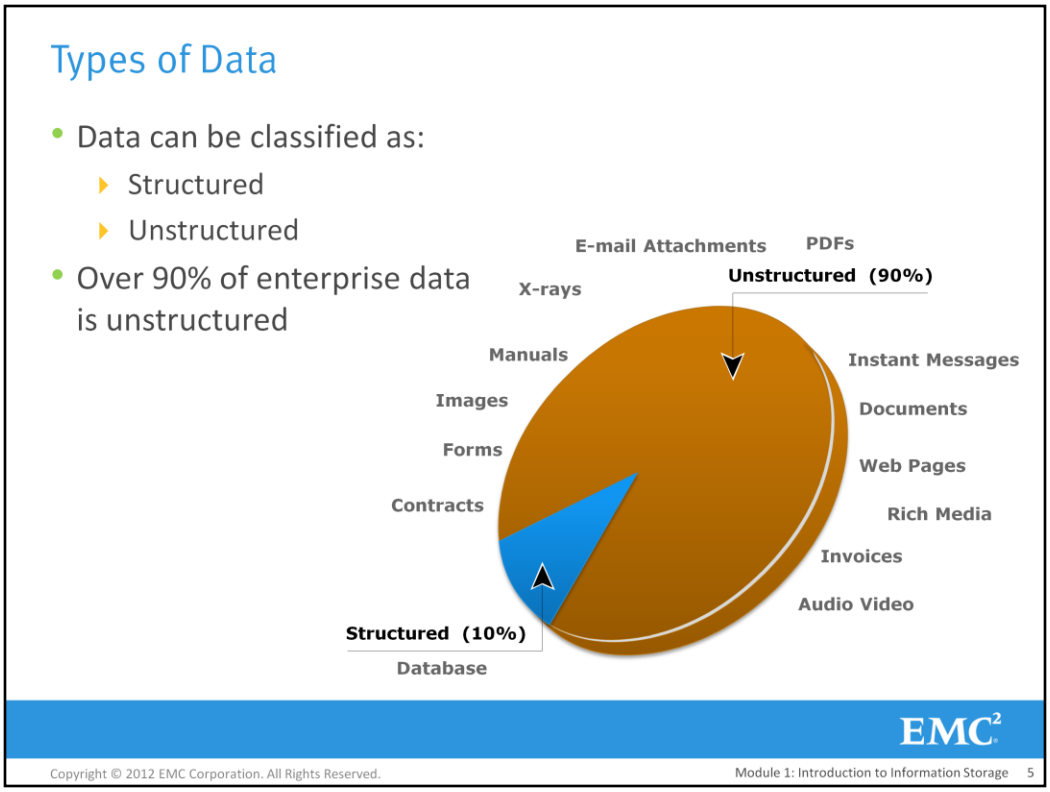
The importance, dependency, and volume of information for the business world also continue to grow at astounding rates. Businesses depend on fast and reliable access to information critical to their success. Some of the business applications that rely on information include airline reservations, telecom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



Data is a collection of raw facts from which conclusions might be drawn. Handwritten letters, a printed book, a family photograph, a movie on video tape, printed and duly signed copies of mortgage papers, a bank's ledgers, and an account holder's passbooks are all examples that contain data. Before the advent of computers, the procedures and methods adopted for data creation and sharing were limited to fewer forms, such as paper and film. Today, the same data can be converted into more convenient forms, such as an e-mail message, an e-book, a bitmapped image, or a digital movie. This data can be generated using a computer and stored in strings of 0s and 1s, as shown in the slide. Data in this form is called *digital data* and is accessible by the user only after it is processed by a computer.

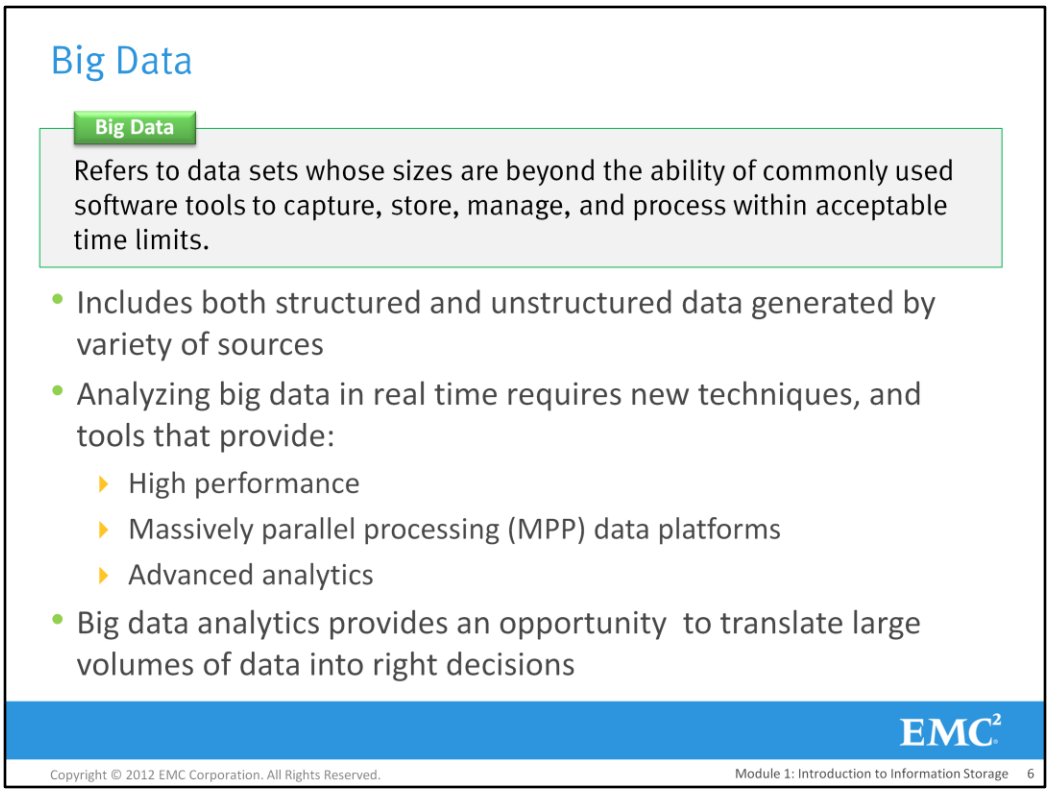
With the advancement of computer and communication technologies, the rate of data generation and sharing has increased exponentially. The following is a list of some of the factors that have contributed to the growth of digital data:

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

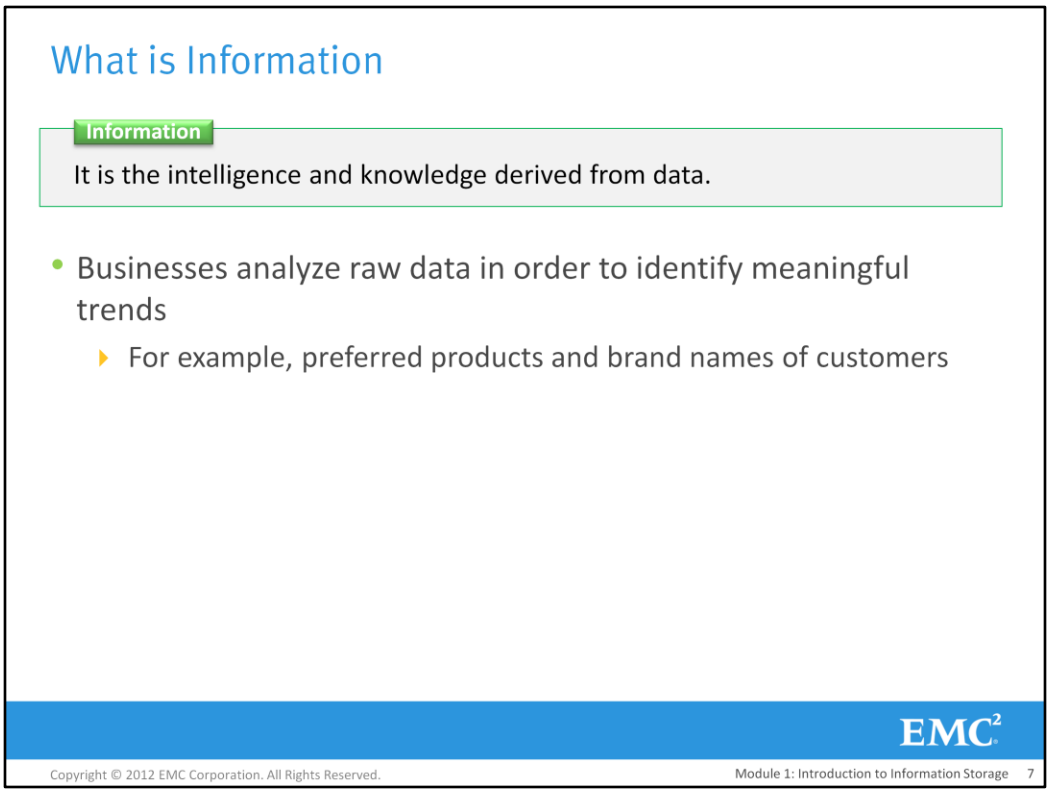
Data is unstructured if its elements cannot be stored in rows and columns, and is therefore difficult to query and retrieve by business applications. For example, customer contacts may be stored in various forms such as sticky notes, e-mail messages, business cards, or even digital format files, such as .doc, .txt, and .pdf. Due to its unstructured nature, it is difficult to retrieve this data using a traditional customer relationship management application. Businesses are primarily concerned with managing unstructured data because over 90 percent of enterprise data is unstructured and require significant storage space and effort to manage.



'*Big data*' is a new and evolving concept, which 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. It includes both structured and unstructured data generated by variety of sources, including business application transactions, Web pages, videos, images, emails, social media, and so on. These data sets typically require real time capture or updates for analysis, predictive modeling, and decision making.

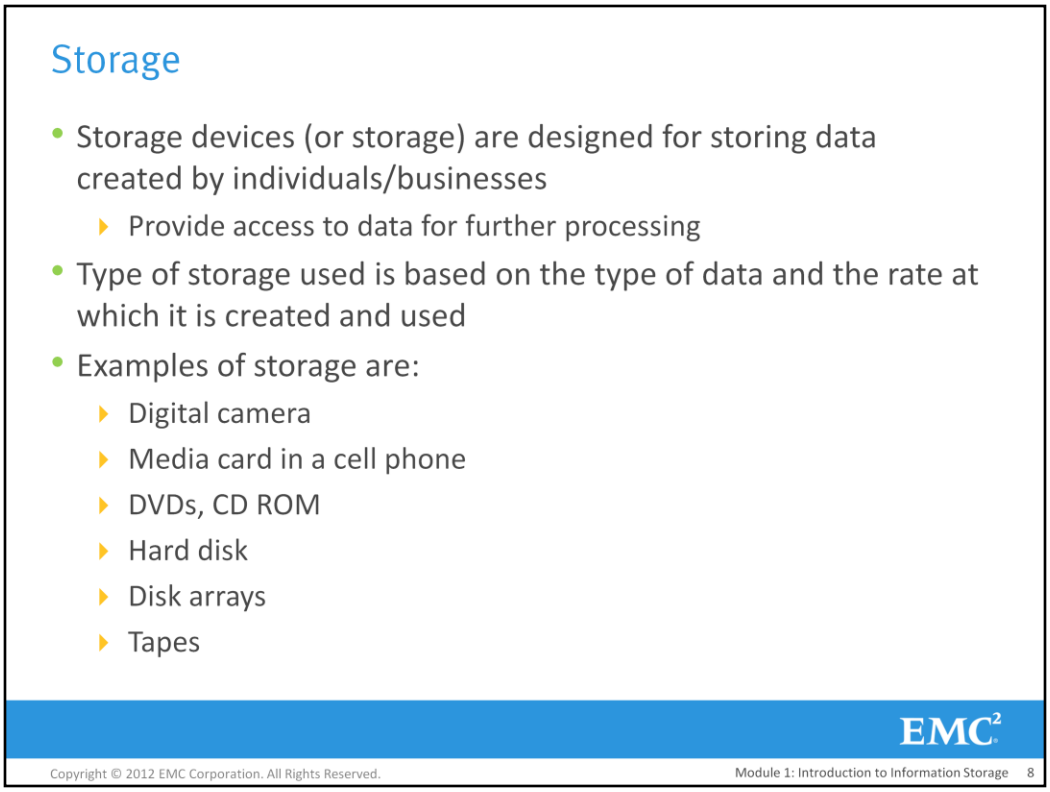
Traditional IT infrastructure and data processing tools and methodologies are inadequate to handle the volume, variety, dynamism, and complexity of big data. Analyzing big data in real time requires new techniques, architectures, and tools that provide high performance, massively parallel processing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

Organizations faced challenges translating large volumes of information into the right decisions. Big data analytics provides an opportunity to find insight in new and emerging data types, spot business trends, improve customer acquisition, drive product and service strategies, optimize business operations, and create competitive advantages in a dynamic global marketplace. Medical and scientific research, healthcare, public administration, fraud detection, social media, banks, insurance companies, and other digital information based entities benefit from big data analytics. 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.



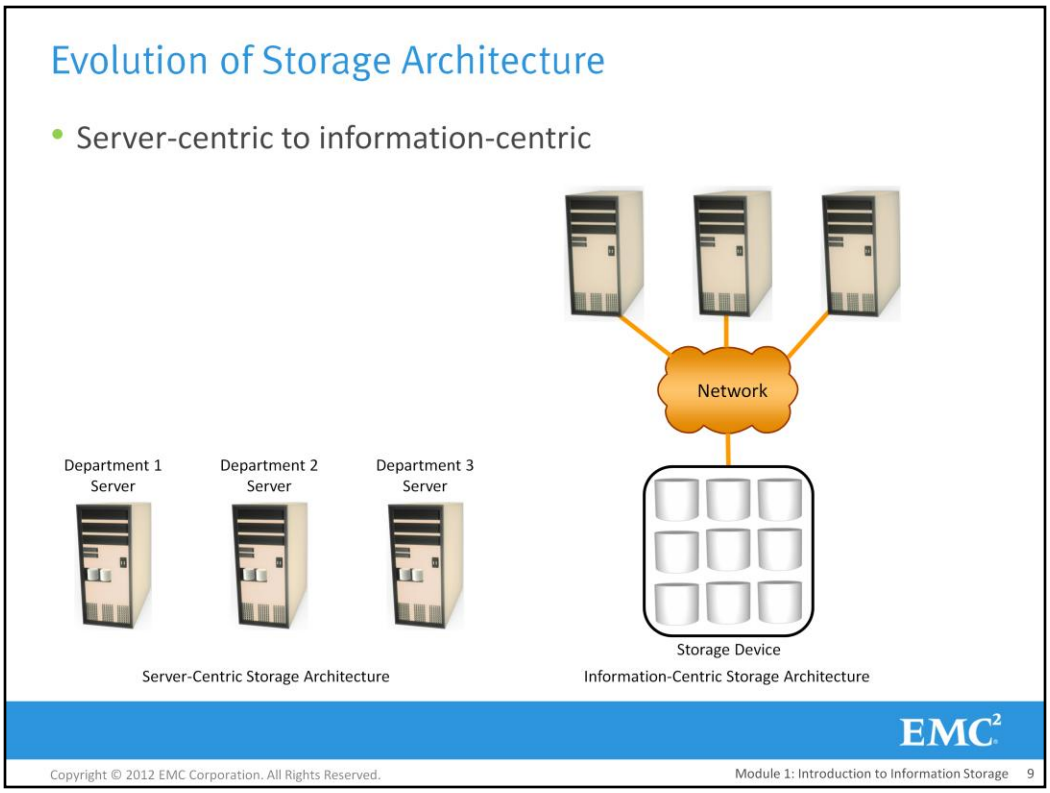
Data, whether structured or unstructured, does not fulfill any purpose for individuals or businesses unless it is presented in a meaningful form. *Information* is the intelligence and knowledge derived from data.

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.



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.

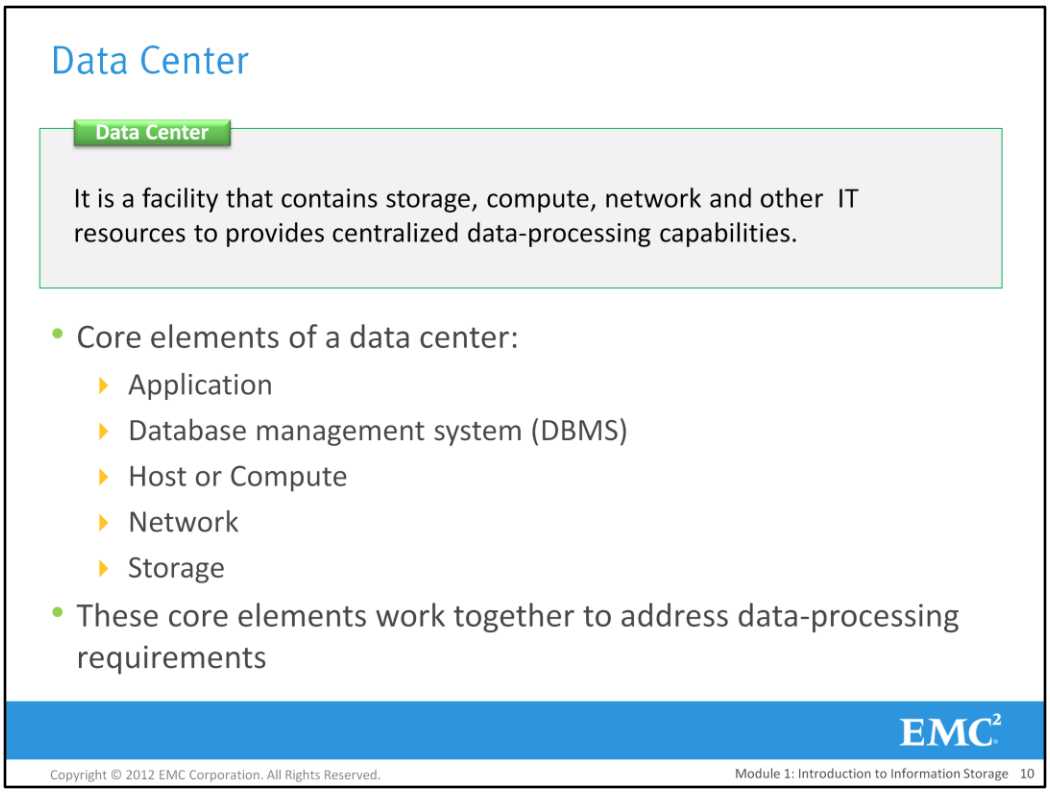
Businesses have several options available for storing data, including internal hard disks, external disk arrays, and tapes.



Historically, organizations had centralized computers (mainframe) 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 architecture is called *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, result in unavailability of information. The proliferation of departmental servers in an enterprise resulted in unprotected, unmanaged, fragmented islands of information and increased operating cost.

To overcome these challenges, storage architecture 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. 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.



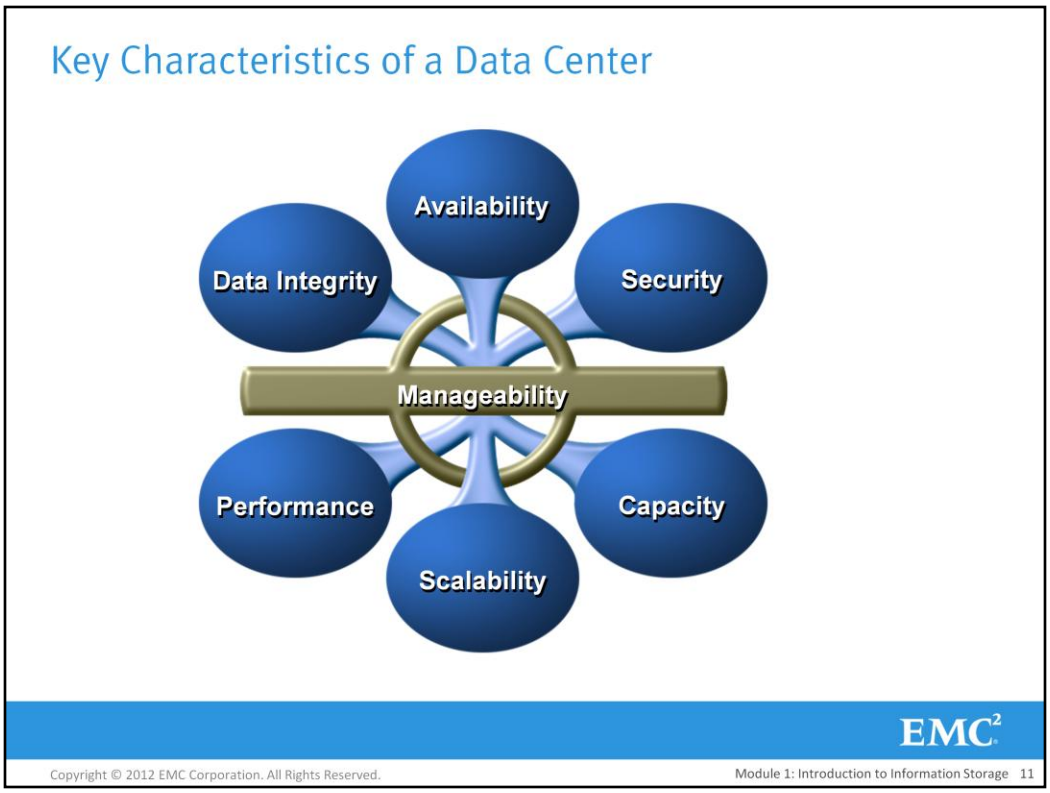
Organizations maintain data centers to provide centralized data-processing capabilities across the enterprise. Data centers store and manage large amounts of data. The data center infrastructure includes hardware components, such as computers, storage systems, network devices, and power backups, and environmental controls, such as air conditioning, fire suppression, and ventilation. It also includes the number of software, such as applications, operating systems, and management software. Large organizations often maintain more than one data center to distribute data processing workloads and provide backup in the event of a disaster.

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

- Application: A computer program that provides the logic for computing operations.
- Database management system (DBMS): It provides a structured way to store data in logically organized tables that are interrelated.
- Host or Compute: A computing platform 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.

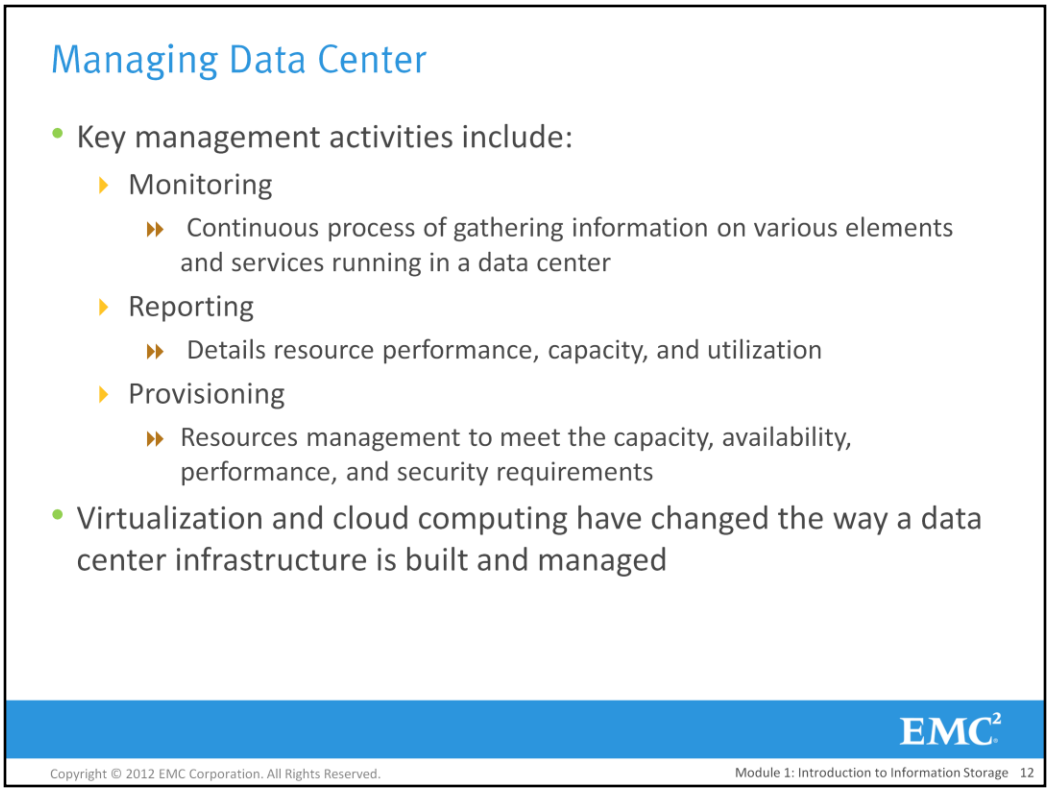
These core elements are typically viewed and managed as separate entities, but all the elements must work together to address data-processing requirements.

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



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

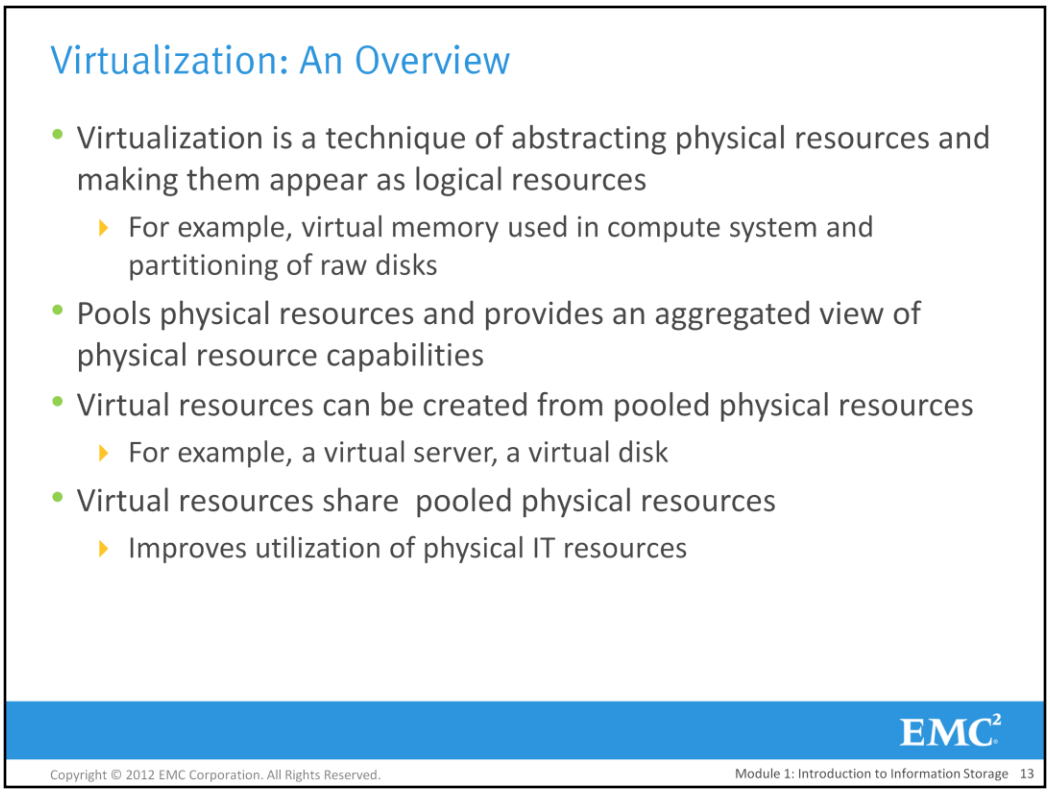
- Availability: A data center should ensure the availability of information, when required. Unavailability of information could cost millions of dollars per hour to businesses, such as financial services, telecommunications, and e-commerce.
- Security: Polices, procedures, and proper integration of the data center core elements, which will prevent unauthorized access to information must be established. In addition to the security measures for client access, specific mechanisms must enable servers to access only their allocated resources on storage arrays.
- Scalability: Data center resources should be able to scale based on requirements, without interrupting business operations. Business growth often requires deploying more servers, new applications, and additional databases. The storage solution should be able to grow with the business.
- Performance: All the elements of the data center should be able to 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 written to the disk 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 be able to 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 modern, complex data center involves many tasks. The key management activities include:

- *Monitoring* is the 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, accessibility, and capacity.
- *Reporting* is done periodically on resource performance, capacity, and utilization. Reporting tasks help to establish business justifications and chargeback of costs associated with data center operations.
- *Provisioning* is the process of providing the hardware, software, and other resources required to run a data center. Provisioning activities primarily include resources management to meet the capacity, availability, performance, and security requirements.

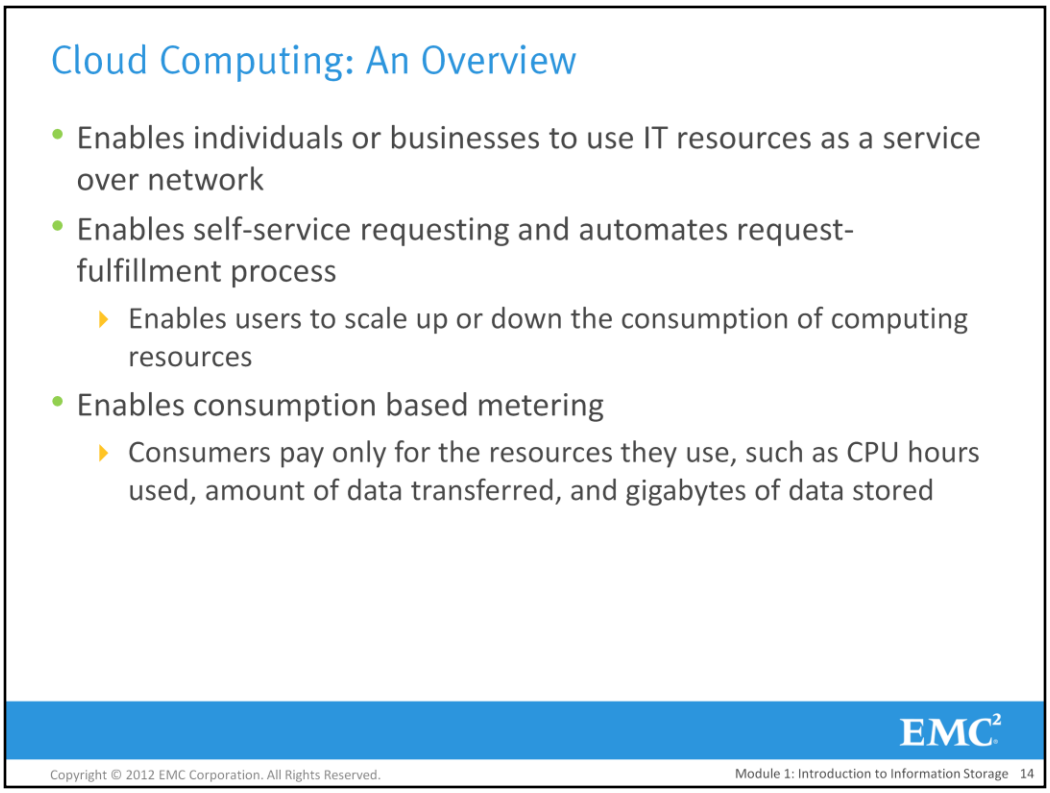
Virtualization and cloud computing have dramatically changed the way a data center infrastructure is built and managed. Organizations are rapidly deploying virtualization on various elements of data centers to optimize their utilization. Further, continuous cost pressure on IT and on-demand data processing 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 has existed in the IT industry for several years and in different forms. Common examples of virtualization are virtual memory used in compute system and partitioning of raw disks.

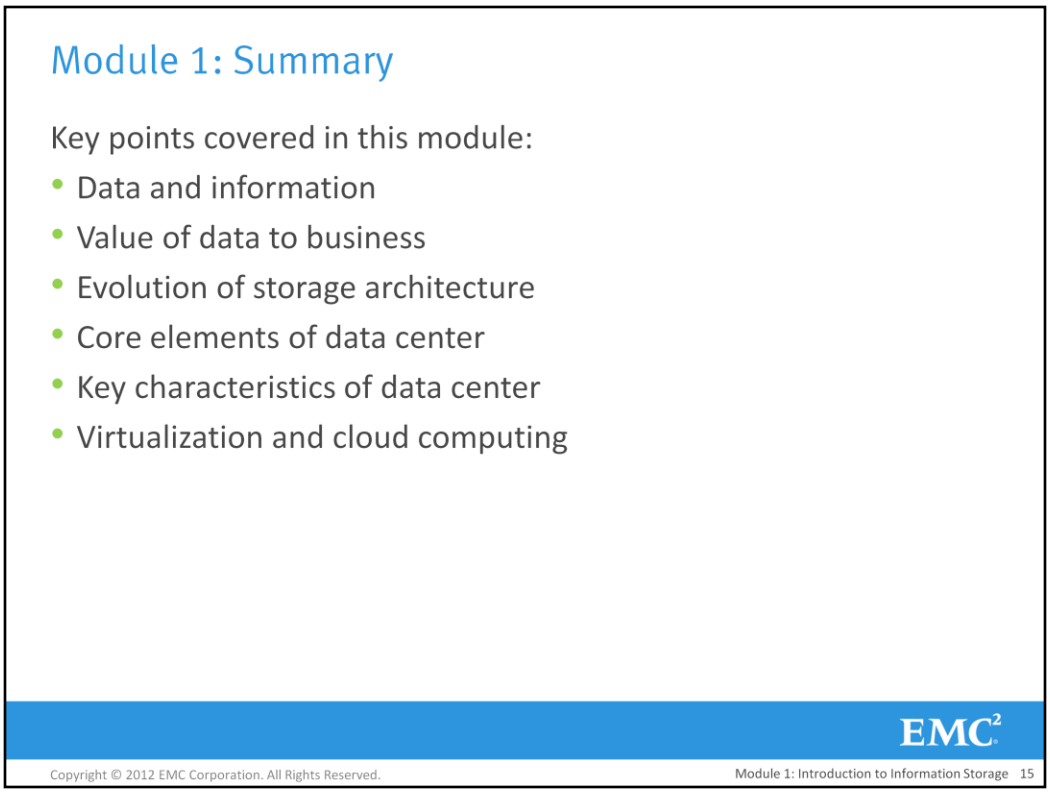
Virtualization enables pooling of physical resources and providing an aggregated view of the physical resource capabilities. For example, storage virtualization enables multiple pooled storage devices to appear as a single large storage. Similarly, by using compute virtualization, the CPU capacity of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

Virtual resources can be created and provisioned from the pooled physical resources. For example, a virtual disk of 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 cost associated with procurement and management of new physical resources. Moreover, less resources means less space and energy, which in turn 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 IT budgets. *Cloud computing*, a next generation style of 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 down the consumption 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 in the background. Cloud computing enables consumption based metering; hence, 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 scaling of resource capabilities. Information storage in virtualized and cloud environments is detailed later in the course.



This module defined data and information. Data is a collection of raw facts from which conclusions might 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.

Information centric architecture is commonly deployed in todays 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

- What are the two types of data?
- What is Big Data?
- What are the five core elements of data center?
- What are the key characteristics of data center?
- Define virtualization and cloud computing.