

Principles of Management Information Systems

Lecture 5

MGM331

Database Systems and Big Data

Prepared by
Dr. Eng. Youssef Ntefeh

Chapter 5

- Organizations and individuals capture prodigious amounts of data from a myriad of sources every day.
- Where does all this data come from, where does it go, how is it safeguarded, and how can you use it to your advantage?
- In this chapter, you will learn about tools and processes that enable users to manage all this data so that it can be used to uncover new insights and make effective decisions.

Chapter 5

- For example, if you become a marketing manager, you can access a vast store of data related to the Web-surfing habits, past purchases, and even social media activity of existing and potential customers.
- You can use this information to create highly effective marketing programs that generate consumer interest and increased sales.

Chapter 5

- If you become a biologist, you may use big data to study the regulation of genes and the evolution of genomes in an attempt to understand how the genetic makeup of different cancers influences outcomes for cancer patients.
- If you become a human resources manager, you will be able to use data to analyze the impact of raises and changes in employee-benefit packages on employee retention and long-term costs.

Chapter 5

- Regardless of your field of study in school and your future career, using database systems and big data will likely be a critical part of your job.
- As you study this chapter, you will see how you can use databases and big data to extract and analyze valuable information to help you succeed.
- This chapter starts by introducing basic concepts related to databases and data management systems. Later, the topic of big data will be discussed along with several tools and technologies used to store and analyze big data.

Principles

- ✓ Data modeling is a key aspect of organizing data and information.
- ✓ A well-designed and well-managed database is an extremely valuable tool in supporting decision making.
- ✓ We have entered an era where organizations are grappling with a tremendous growth in the amount of data available and struggling to understand how to manage and make use of it.
- ✓ A number of available tools and technologies allow organizations to take advantage of the opportunities offered by big data.

Learning Objectives

- Identify and briefly describe the members of the hierarchy of data.
- Identify the advantages of the database approach to data management.
- Identify the key factors that must be considered when designing a database.
- Identify the various types of data models and explain how they are useful in planning a database.

Learning Objectives

- Describe the relational database model and its fundamental characteristics.
- Discuss the role of a database administrator and data administrator.
- Identify the common functions performed by all database management systems.
- Define the term big data and identify its basic characteristics.

Learning Objectives

- Explain why big data represents both a challenge and an opportunity.
- Define the term data management and state its overall goal.
- Define the terms data warehouse, data mart, and data lakes and explain how they are different.
- Outline the extract, transform, load process.
- Discuss the whole Hadoop computing environment and its various components.

Chapter 5 Database Systems and Big Data

Data Fundamentals

Hierarchy of Data

Data Entities, Attributes, and Keys

Data Modeling and Database Characteristics

Data Modeling

Relational Database Model

Manipulating Data

Data Cleansing

SQL Databases

Big Data

Sources of Big Data

Big Data Uses

Challenges of Big Data

Data Management

Technologies Used to Process Big Data

Data Warehouses, Data Marts, and Data Lakes

Hadoop

Introduction

- A *database* is a well-designed, organized, and carefully managed collection of data.
- A database can contribute to organizational success by providing managers and decision makers with timely, accurate, and relevant information built on data.
- Databases also help companies analyze information to reduce costs, increase profits, add new customers, track past business activities, and open new market opportunities.

Introduction

- A *database management system* (DBMS) consists of a group of programs used to access and manage a database as well as provide an interface between the database and its users and other application programs.
- A DBMS provides a single point of management and control over data resources, which can be critical to maintaining the integrity and security of the data.
- A database, a DBMS, and the application programs that use the data make up a database environment.

Hierarchy of Data

- Data is generally organized in a hierarchy that begins with the smallest piece of data used by computers (a bit), progressing up through the hierarchy to a database.
- A *bit* is a binary digit (i.e., 0 or 1)
- A *byte* is typically eight bits. Each byte represents a character, which is the basic building block of most information.
- A *character* can be an uppercase letter (A, B, C, ..., Z), a lowercase letter (a, b, c, ..., z), a numeric digit (0, 1, 2, ..., 9)

Hierarchy of Data

- Characters are put together to form a field.
- A *field* is typically a name, a number, or a combination of characters that describes an aspect of a business object (such as an employee, a location, or a plant) or activity (such as a sale).
- In addition to being entered into a database, fields can be computed from other fields.
- Computed fields include the total, average, maximum, and minimum value

Hierarchy of Data

- A collection of data fields all related to one object, activity, or individual is called a *record*.
- By combining descriptions of the characteristics of an object, activity, or individual, a record can provide a complete description of it.
- For instance, an employee record is a collection of fields about one employee, one field includes the employee's name, another field contains the address, and still others the phone number, pay rate, earnings made to date, and so forth.
- A collection of related records is a *file*, for example, an employee file is a collection of all company employee records.
- At the highest level of the data hierarchy is a *database*, a collection of integrated and related files.

Data Entities, Attributes, and Keys

- An *entity* is a person, place, or thing (object) for which data is collected, stored, and maintained. Examples of entities include employees, products, and customers.
- An *attribute* is a characteristic of an entity. For example, employee number, last name, first name, hire date, and department number are attributes for an employee.
- Each attribute can be constrained to a range of allowable values called its *domain* which is what values can be placed in each column.
- The domain for a particular attribute indicates A *primary key* is a field or set of fields that uniquely identifies the record. No other record can have the same primary key.

Chapter 5 Database Systems and Big Data

Data Fundamentals

Hierarchy of Data

Data Entities, Attributes, and Keys

Data Modeling and Database Characteristics

Data Modeling

Relational Database Model

Manipulating Data

Data Cleansing

SQL Databases

Big Data

Sources of Big Data

Big Data Uses

Challenges of Big Data

Data Management

Technologies Used to Process Big Data

Data Warehouses, Data Marts, and Data Lakes

Hadoop

Data Modeling and Database Characteristics

Because today's businesses must keep track of and analyze so much data, they must keep the data well organized so that it can be used, when building a database, an organization must carefully consider the following questions:

- **Content:** What data should be collected and at what cost?
- **Access:** What data should be provided to which users and when?
- **Logical structure:** How should data be arranged so that it makes sense to a given user?
- **Physical organization:** Where should data be physically located?
- **Archiving:** How long must this data be stored?
- **Security:** How can this data be protected from unauthorized access?

Data Modeling

- A *data model* is a diagram of entities and their relationships.
- Data modeling usually involves developing an understanding of a specific business problem and then analyzing the data and information needed to deliver a solution.
- When done at the level of the entire organization, this procedure is called enterprise data modeling.
- An *enterprise data model*: A data model that provides a roadmap for building database and information systems by creating a single definition and format for data that can ensure data compatibility and the ability to exchange and integrate data among systems.

Data Modeling

- *Entity-relationship (ER) diagram*, uses basic graphical symbols to show the relationships between data.
- Boxes in ER diagrams indicate data items or entities contained in data tables, and lines show relationships between entities.
- In other words, ER diagrams show data items in tables (entities) and the ways they are related.
- ER diagrams help ensure that the relationships among the data entities in a database are correctly structured so that any application programs developed are consistent with business operations and user needs.

Data Modeling

- Figure 5.6 shows an ER diagram for an order database.
- One salesperson serves many customers. one-to-many relationship,
- Each customer can place one-to-many orders
- One order generates one invoice one-to-one relationships

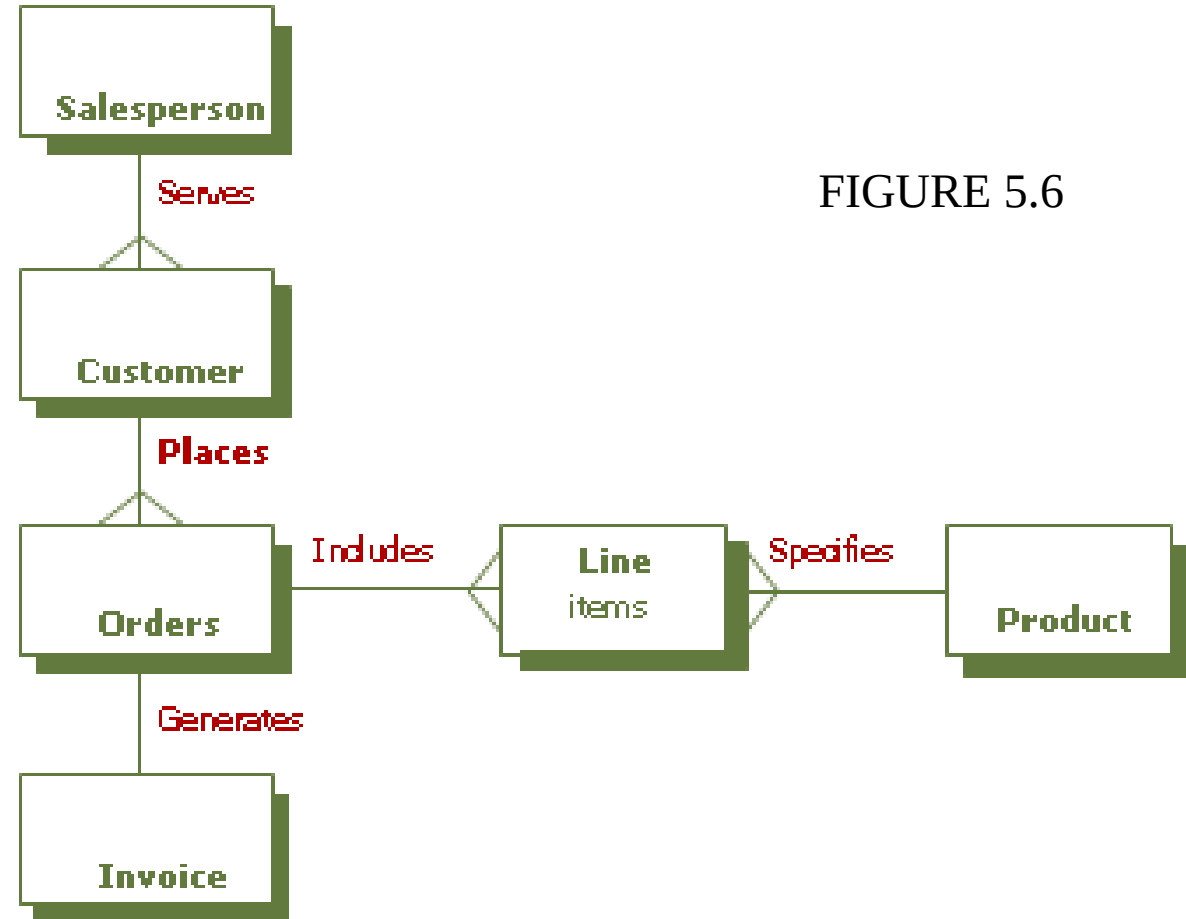


FIGURE 5.6

Relational Database Model

- In the *relational model*, data is placed in two dimensional tables, or relations. As long as they share at least one common attribute, these relations can be linked to provide output useful information.
- After entering data into a relational database, users can make inquiries and analyze the data.

Manipulating Data

- Basic data manipulations include **selecting**, **projecting**, and **joining**.
- *Selecting* involves eliminating rows according to certain criteria. Suppose the department manager of a company wants to use an employee table that contains the project number, description, and department number for all projects a company is performing.
- Using selection, the manager can eliminate all rows except the one for Project 226 and see that the department number for the department completing the sales manual project is 598.

Manipulating Data

Projecting involves eliminating columns in a table.

- For example, a department table might contain the department number, department name, and Social Security number (SSN) of the manager in charge of the project.
- A sales manager might want to create a new table that contains only the department number and the Social Security number of the manager in charge of the sales manual project.
- The sales manager can use projection to eliminate the department name column and create a new table containing only the department number and Social Security number.

Manipulating Data

Joining involves combining two or more tables.

- For example, you can combine the project table and the department table to create a new table with the project number, project description, department number, department name, and Social Security number for the manager in charge of the project.

Data Cleansing

- Data used in decision making must be accurate, complete, economical, flexible, reliable, relevant, simple, timely, verifiable, accessible, and secure.
- *Data cleansing* (data cleaning or data scrubbing) is the process of detecting and then correcting or deleting incomplete, incorrect, inaccurate, or irrelevant records that reside in a database.
- The goal of data cleansing is to improve the quality of the data used in decision making. The “bad data” may have been caused by user data-entry errors or by data corruption during data transmission or storage.

SQL Databases

- *SQL* (Structured Query Language) is a special-purpose programming language for accessing and manipulating data stored in a relational database.
- `SELECT ClientName, Debt FROM Client WHERE Debt > 1000`
- This query displays clients (ClientName) and the amount they owe the company (Debt) from a database table called Client; the query would only display clients who owe the company more than \$1,000 (WHERE Debt > 1000).

Chapter 5 Database Systems and Big Data

Data Fundamentals

Hierarchy of Data

Data Entities, Attributes, and Keys

Data Modeling and Database Characteristics

Data Modeling

Relational Database Model

Manipulating Data

Data Cleansing

SQL Databases

Big Data

Sources of Big Data

Big Data Uses

Challenges of Big Data

Data Management

Technologies Used to Process Big Data

Data Warehouses, Data Marts, and Data Lakes

Hadoop

Big Data

- *Big data* is the term used to describe data collections that are so enormous (terabytes or more) and complex (from sensor data to social media data) that traditional data management software, hardware, and analysis processes are incapable of dealing with them.
- *Volume* of data that exists in the digital universe
- The *velocity* at which data is currently coming at us
- Data today comes in a variety of formats.

Big Data

- *Structured data*: its format is known in advance, and it fits nicely into traditional databases
- *Unstructured data*: meaning that it is not organized in any predefined manner. comes from sources such as word-processing documents, social media, email, photos, surveillance video, and phone messages.

Sources of Big Data

- Organizations collect and use data from a variety of sources, including business applications, social media, sensors and controllers that are part of the manufacturing process, systems that manage the physical environment in factories and offices, media sources (including audio and video broadcasts), machine logs that record events and customer call data, public sources (such as government Web sites), and archives of historical records of transactions and communications.

Big Data Uses

- Here are just a few examples of how organizations are employing big data to improve their day-to-day operations, planning, and decision making:
 - Retail organizations monitor social networks such as Facebook, Google, LinkedIn, Twitter, and Yahoo to engage brand advocates, identify brand adversaries (and attempt to reverse their negative opinions), and even enable passionate customers to sell their products.
 - Advertising and marketing agencies track comments on social media to understand consumers' responsiveness to ads, campaigns, and promotions.

Big Data Uses

- Hospitals analyze medical data and patient records to try to identify patients likely to need readmission within a few months of discharge, with the goal of engaging with those patients in the hope of preventing another expensive hospital stay.
- Consumer product companies monitor social networks to gain insight into customer behavior, likes and dislikes, and product perception to identify necessary changes to their products, services, and advertising.

Big Data Uses

- Financial services organizations use data from customer interactions to identify customers who are likely to be attracted to increasingly targeted and sophisticated offers.
- Manufacturers analyze minute vibration data from their equipment, which changes slightly as it wears down, to predict the optimal time to perform maintenance or replace the equipment to avoid expensive repairs or potentially catastrophic failure

Challenges of Big Data

The challenge is manifold, with a variety of questions that must be answered, including:

- How to choose what subset of data to store
- Where and how to store the data
- How to find those nuggets of data that are relevant to the decision making at hand
- How to derive value from the relevant data
- How to identify which data needs to be protected from unauthorized access.

With so much data available, business users can have a hard time finding the information they need to make decisions, and they may not trust the validity of the data they can access.

Challenges of Big Data

- Trying to deal with all this data from so many different sources, much of it from outside the organization, can also increase the risk that the organization fails to comply with government regulations or internal control.
- Violation of these regulations can lead not only to government investigations but also to dramatic drops in stock prices

Challenges of Big Data

- Big data also introduces security concerns.
- ❑ Are organizations able to keep big data secure from competitors and malicious hackers?
- Some experts believe companies that collect and store big data could be open to liability suits from individuals and organizations.
- Even with these potential disadvantages, many companies are rushing into big data due to the lure of a potential treasure trove of information and new applications.

Data Management

- *Data management* is an integrated set of functions that defines the processes by which data is obtained, certified fit for use, stored, secured, and processed in such a way as to ensure that the accessibility, reliability, and timeliness of the data meet the needs of the data users within an organization.
- The need for data management is driven by a variety of factors, including the need to meet external regulations designed to manage risk associated with financial misstatement, the need to avoid the inadvertent release of sensitive data, or the need to ensure that high data quality is available for key decisions.

Data Management

- *Data steward* is an individual responsible for the management of critical data elements, including identifying and acquiring new data sources; creating and maintaining consistent reference data and master data definitions; and analyzing data for quality and data issues.
- *Data lifecycle management* (DLM) is a policy-based approach to managing the flow of an enterprise's data, from its initial acquisition or creation and storage to the time when it becomes outdated and is deleted.

Chapter 5 Database Systems and Big Data

Data Fundamentals

Hierarchy of Data

Data Entities, Attributes, and Keys

Data Modeling and Database Characteristics

Data Modeling

Relational Database Model

Manipulating Data

Data Cleansing

SQL Databases

Big Data

Sources of Big Data

Big Data Uses

Challenges of Big Data

Data Management

Technologies Used to Process Big Data

Data Warehouses, Data Marts, and Data Lakes

Hadoop

Technologies Used to Process Big Data

- A *data warehouse* is a database that holds business information from many sources in the enterprise, covering all aspects of the company's processes, products, and customers.
- Data warehouses allow managers to “drill down” to get greater detail or “roll up” to generate aggregate or summary reports.
- The primary purpose is to relate information in innovative ways and help managers and executives make better decisions.
- A data warehouse stores historical data that has been extracted from operational systems and external data sources.

Data warehouses

- Data warehouses are continuously refreshed with huge amounts of data from a variety of sources so the probability that some of the sources contain “dirty data” is high
- The **ETL** (extract, transform, load) process takes data from a variety of sources, edits and transforms it into the format used in the data warehouse, and then loads this data into the warehouse.
- **Extract:** Source data for the data warehouse comes from many sources and may be represented in a variety of formats. The goal of this process is to extract the source data from all the various sources and convert it into a single format suitable for processing. During the extract step, data that fails to meet expected patterns or values may be rejected from

Data warehouses

- **Transform:** During this stage of the ETL process, a series of rules or algorithms are applied to the extracted data to derive the data that will be stored in the data warehouse. to reduce the processing time required to create anticipated reports. For example, total sales may be accumulated by store or sales district.
- **Load:** During this stage of the ETL process, the extracted and transformed data is loaded into the data warehouse. As the data is being loaded into the data warehouse, new indices are created and the data is checked against the constraints defined in the database schema to ensure its quality. As a result, the data load stage for a large data warehouse can take days.

Data Marts Data Lakes

- A ***data mart*** is a subset of a data warehouse. Rather than store all enterprise data in one monolithic database, data marts contain a subset of the data for a single aspect of a company's business for example, finance, inventory, or personnel.
- A ***data lake*** (also called an enterprise data hub) takes a “store everything” approach to big data, saving all the data in its raw and unaltered form. The raw data residing in a data lake is available when users decide just how they want to use the data to glean new insights. Only when the data is accessed for a specific analysis is it extracted from the data lake, classified, organized, edited, or transformed. Thus a data lake serves as the definitive source of data in its original, unaltered form.

Technologies Used to Process Big Data

Hadoop

- Hadoop is an open-source software framework that includes several software modules that provide a means for storing and processing extremely large data sets,
- Hadoop has two primary components: a data processing component and a distributed file system for data storage.
- Hadoop divides data into subsets and distributes the subsets onto different servers for processing

Summary

- ❖ Data is one of the most valuable resources that a firm possesses.
- ❖ It is organized into a hierarchy that builds from the smallest element to the largest.
- ❖ The smallest element is the bit, a binary digit. A byte (a character such as a letter or numeric digit) is made up of eight bits.
- ❖ A group of characters, such as a name or number, is called a field (an object).

Summary

- ❖ A collection of related fields is a record; a collection of related records is called a file.
- ❖ The database, at the top of the hierarchy, is an integrated collection of records and files.
- ❖ An entity is a generalized class of objects (such as a person, place, or thing) for which data is collected, stored, and maintained.
- ❖ An attribute is a characteristic of an entity. Specific values of attributes—called data items—can be found in the fields of the record describing an entity.
- ❖ A primary key uniquely identifies a record, while a secondary key is a field in a record that does not uniquely identify the record.

Summary

- ❖ When building a database, an organization must consider content, access, logical structure, physical organization, archiving, and security of the database.
- ❖ A data model is a map or diagram of entities and their relationships.
- ❖ Enterprise data modeling involves analyzing the data and information needs of an entire organization and provides a roadmap for building database and information systems by creating a single definition and format for data that can ensure compatibility and the ability to exchange and integrate data among systems

Summary

- ❖ Entity-relationship (ER) diagrams can be used to show the relationships among entities in the organization.
- ❖ The relational database model places data in two-dimensional tables.
- ❖ Each row in a relational database table represents a record, and each column represents an attribute (or field).
- ❖ The allowable values for each attribute are called the attribute's domain.

Summary

- ❖ Basic data manipulations include selecting, projecting, joining.
- ❖ Data cleansing is the process of detecting and then correcting or deleting incomplete, incorrect, inaccurate, or irrelevant records that reside in the database. The goal of data cleansing is to improve the quality of the data used in decision making.

Summary

- ❖ A database management system (DBMS) is a group of programs used as an interface between a database and its users and between a database and other application programs.
- ❖ A type of specialized data manipulation language is the query language, the most common being Structured Query Language (SQL).
- ❖ SQL is used in several popular database packages today and can be installed in PCs and mainframes.

Summary

- ❖ We have entered an era where organizations are grappling with a tremendous growth in the amount of data available and struggling how to manage and make use of it.
- ❖ “Big data” is the term used to describe data collections that are so enormous and complex that traditional data management software, hardware, and analysis processes are incapable of dealing with them.

Summary

- ❖ There are many challenges associated with big data, including how to choose what subset of data to store, where and how to store the data, how to find those nuggets of data that are relevant to the decision making at hand, how to derive value from the relevant data, and how to identify which data needs to be protected from unauthorized access.
- ❖ An extract, transform, load process takes data from a variety of sources, edits and transforms it into the format to be used in the data warehouse, and then loads the data into the warehouse.

Summary

- ❖ Data marts are subdivisions of data warehouses and are commonly devoted to specific purposes or functional business areas.
- ❖ A data lake (also called an enterprise data hub) takes a “store everything” approach to big data, saving all the data in its raw and unaltered form.
- ❖ Hadoop is software that provide a means for storing and processing extremely large data sets.
- ❖ Hadoop has two primary components a data processing component and a distributed file for data storage.
- ❖ Hadoop divides data into subsets and distributes the subsets onto different servers for processing.

Reference

