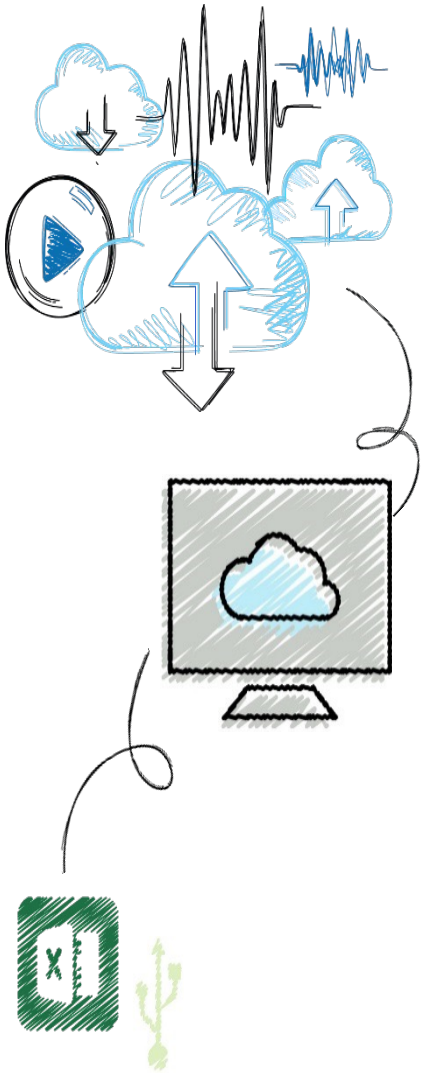
# Principles of Information Systems, Thirteenth Edition

*Chapter 5*

*Database System and Big Data*



# Objectives

After completing this chapter, you will be able to:

Identify the advantages of the database approach to data management

Identify the various types of data models and explain how they are useful in planning a database

Describe the relational database model

Define the role of the database schema, data definition language, and data manipulation language

Discuss the role of a database administrator and data administrator

Define the term big data

Define the terms data warehouse, and data lakes and explain how they are different

# Introduction

- Database: an organized collection of data
- A database management system (DBMS) is a group of programs that:
  - Manipulate the database
  - Provide an interface between the database and its users and other application programs

# Data Fundamentals

- Without data and the ability to process it:
  - An organization could not successfully complete most business activities
- Data consists of raw facts
- Data must be organized in a meaningful way to transform it into useful information

# Hierarchy of Data

- A bit (binary digit) represents a circuit that is either on or off
- A byte is made up of eight bits
  - Each byte represents a character
- Field: a name, number, or combination of characters that describes an aspect of a business object or activity
- Record: a collection of related data fields
- File: a collection of related records

# Hierarchy of Data

- Database: a collection of integrated and related files
- Hierarchy of data: bits, characters, fields, records, files, and databases

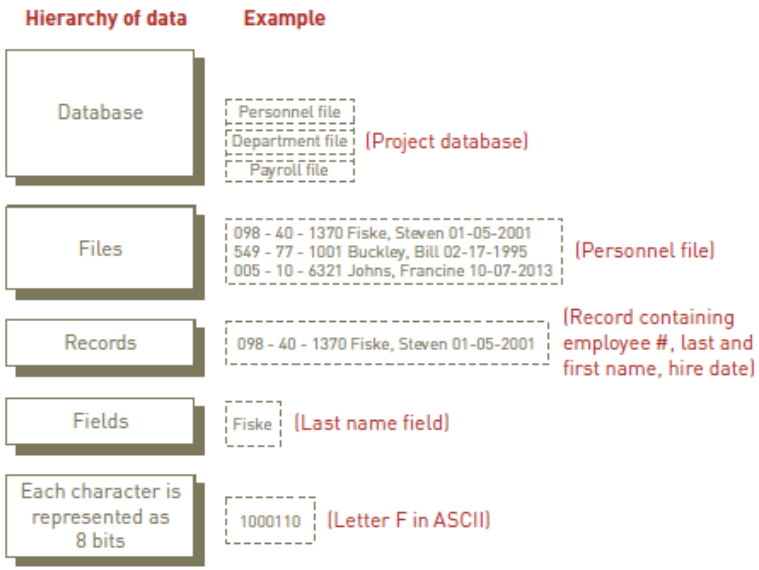



**FIGURE 5.1**
**Hierarchy of data**
Together, bits, characters, fields, records, files, and databases form the hierarchy of data.

# Data Entities, Attributes, and Keys

- Entity: a person, place, or thing for which data is collected, stored, and maintained
- Attribute: a characteristic of an entity
- Data item: the specific value of an attribute
- Primary key: a field or set of fields that uniquely identifies the record

# Data Entities, Attributes, and Keys

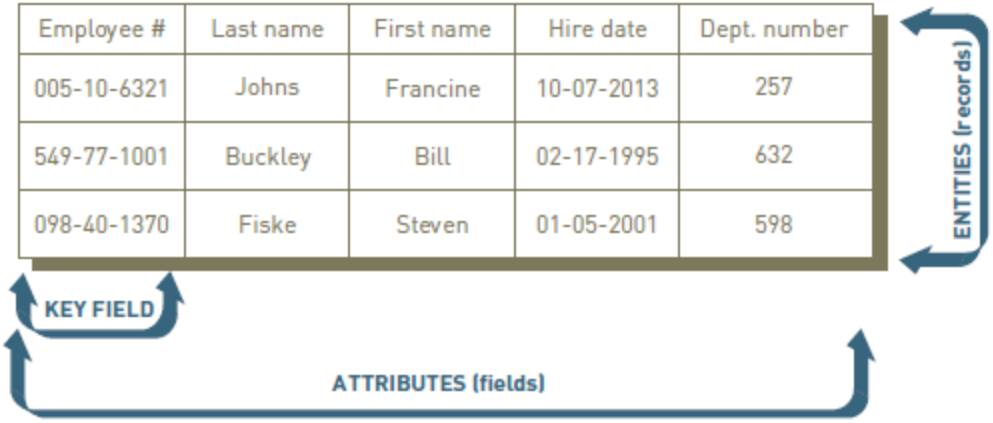



**FIGURE 5.2**

**Keys and attributes**

The key field is the employee number. The attributes include last name, first name, hire date, and department number.

# The Database Approach

- Traditional approach to data management
  - Each distinct operational system used data files dedicated to that system
- Database approach to data management
  - Information systems share a pool of related data
  - Offers the ability to share data and information resources
  - A database management system (DBMS) is required

# The Database Approach

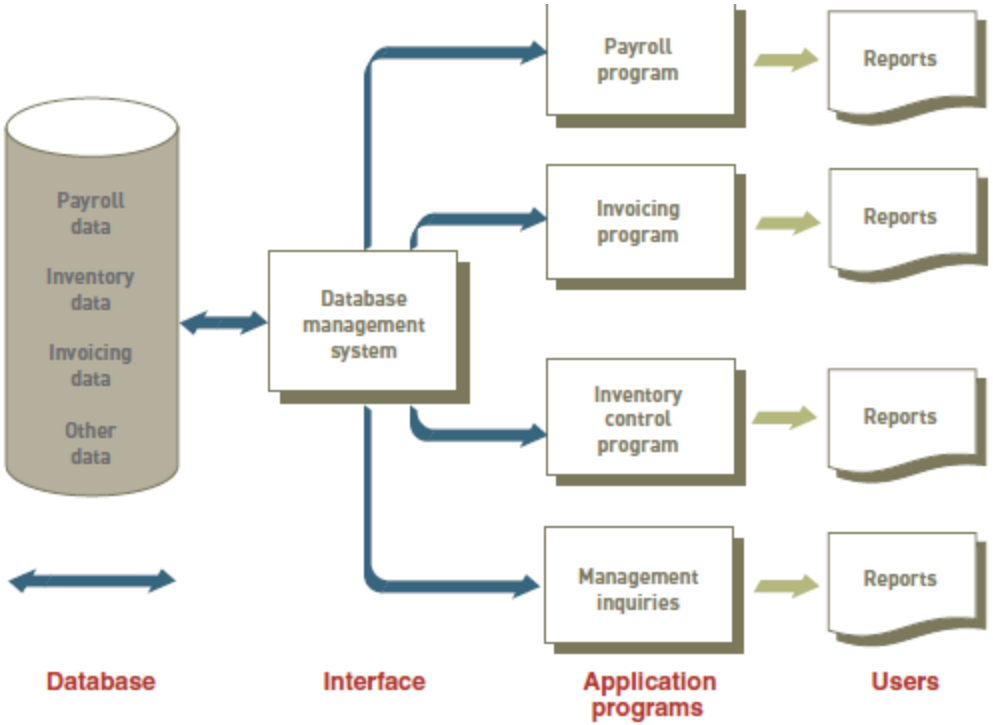



**FIGURE 5.4**
**Database approach to data management**
In a database approach to data management, multiple information systems share a pool of related data.

# Data Modeling and Database Characteristics

- Considerations when building a database
  - Content: what data should be collected? cost?
  - Access: what data should be provided to which users and when?
  - Logical structure: how should data be arranged so that it makes sense?
  - Physical organization: where should data be physically located?
  - Archiving: how long to store?
  - Security: how can data be protected?

# Data Modeling

- Data model: a diagram of data entities and their relationships
- Entity-relationship (ER) diagrams: data models that use basic graphical symbols to show the organization of and relationships between data

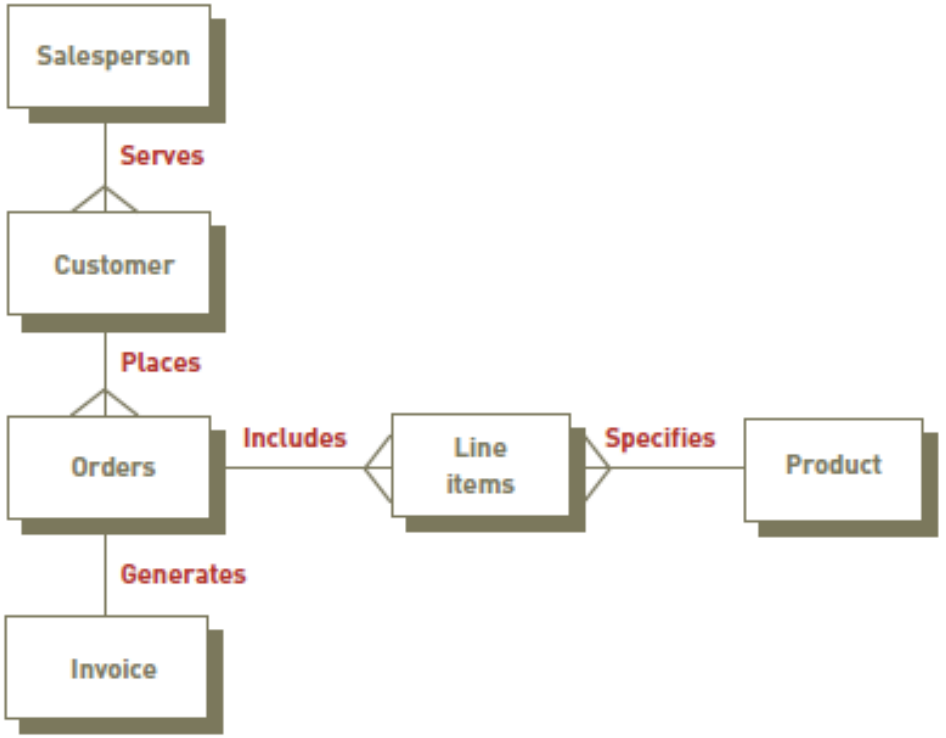



**FIGURE 5.6**
**Entity-relationship (ER) diagram for a customer order database**
Development of ER diagrams helps ensure that the logical structure of application programs is consistent with the data relationships in the database.

# Relational Database Model

- Relational model: a simple but highly useful way to organize data into collections of two-dimensional tables called relations
  - Each row in the table represents an entity
  - Each column represents an attribute of that entity

# Relational Database Model

Data Table 1: Project Table

| Project | Description | Dept. number |
|---|---|---|
| 155 | Payroll | 257 |
| 498 | Widgets | 632 |
| 226 | Sales manual | 598 |

Data Table 2: Department Table

| Dept. | Dept. name | Manager SSN |
|---|---|---|
| 257 | Accounting | 005-10-6321 |
| 632 | Manufacturing | 549-77-1001 |
| 598 | Marketing | 098-40-1370 |

Data Table 3: Manager Table

| SSN | Last name | First name | Hire date | Dept. number |
|---|---|---|---|---|
| 005-10-6321 | Johns | Francine | 10-07-2013 | 257 |
| 549-77-1001 | Buckley | Bill | 02-17-1995 | 632 |
| 098-40-1370 | Fiske | Steven | 01-05-2001 | 598 |

**FIGURE 5.7**

**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. In this example, all three tables include the Dept. number attribute.

# Manipulating Data

- Selecting: eliminating rows according to certain criteria
- Projecting: eliminating columns in a table
- Joining: combining two or more tables
- Linking: combining two or more tables through common data attributes to form a new table with only the unique data attributes

# Manipulating Data

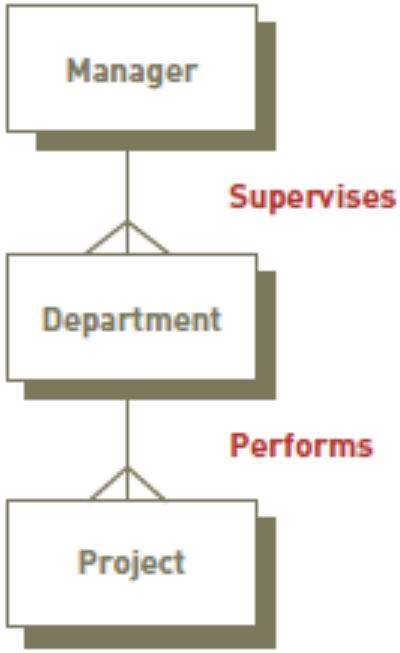



**FIGURE 5.8**

**Simplified ER diagram**

This diagram shows the relationship among the Manager, Department, and Project tables.

# Manipulating Data

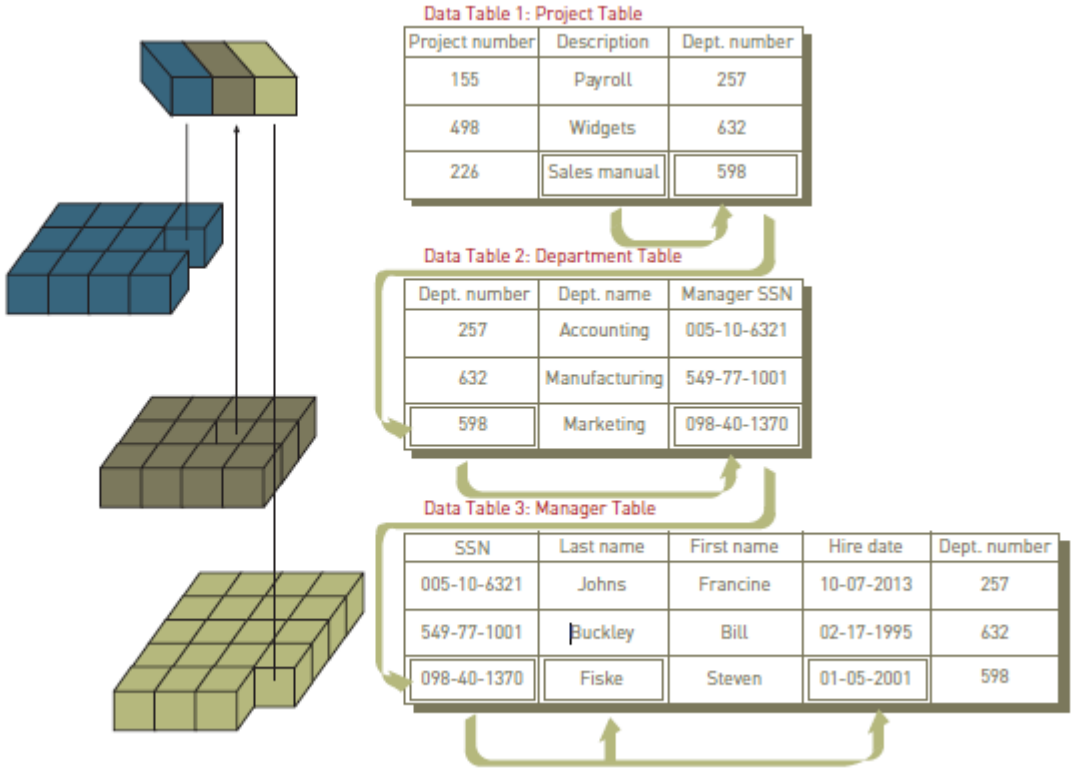



**FIGURE 5.9**

**Linking data tables to answer an inquiry**

To find the name and hire date of the manager working on the sales manual project, the president needs three tables: Project, Department, and Manager. The project description (Sales manual) leads to the department number (598) in the Project table, which leads to the manager's Social Security number (098-40-1370) in the Department table, which leads to the manager's last name (Fiske) and hire date (01-05-2001) in the Manager table.

# Data Cleansing

- Also called data cleaning
- The process of detecting and then correcting or deleting incomplete, incorrect, inaccurate, irrelevant records that reside in a database
- The cost of performing data cleansing can be quite high
- Different from data validation
  - Which involves the identification of “bad data” and its rejection at the time of data entry

# Data Cleansing

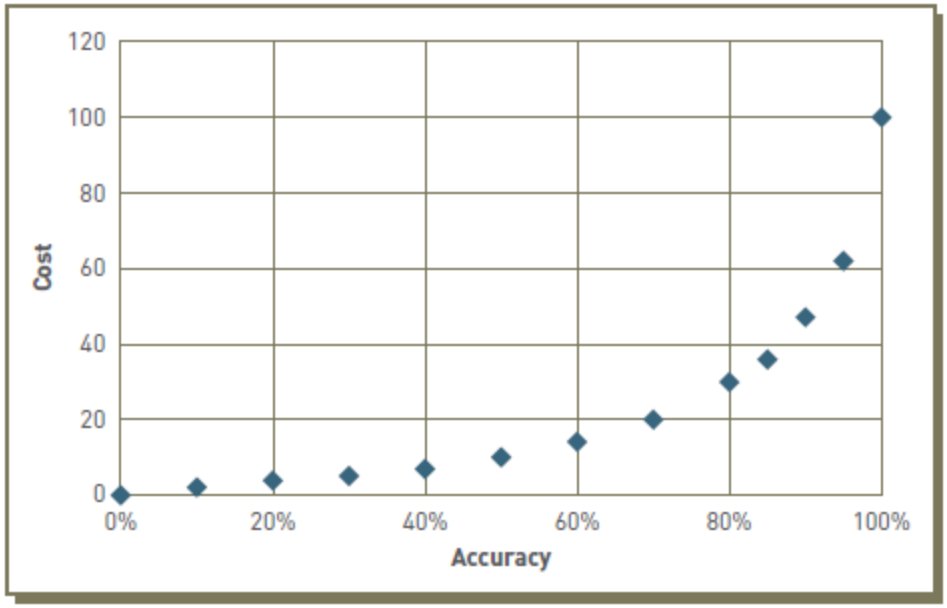



**FIGURE 5.11**

**Tradeoff of cost versus accuracy**

The cost of performing data cleansing to achieve 100 percent database accuracy can be prohibitively expensive.

# Structured Query Language (SQL)

- SQL: a special-purpose programming language for accessing and manipulating data stored in a relational database

TABLE 5.1 Examples of SQL commands

| SQL Command | Description |
|---|---|
| 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). |

# SQL Databases

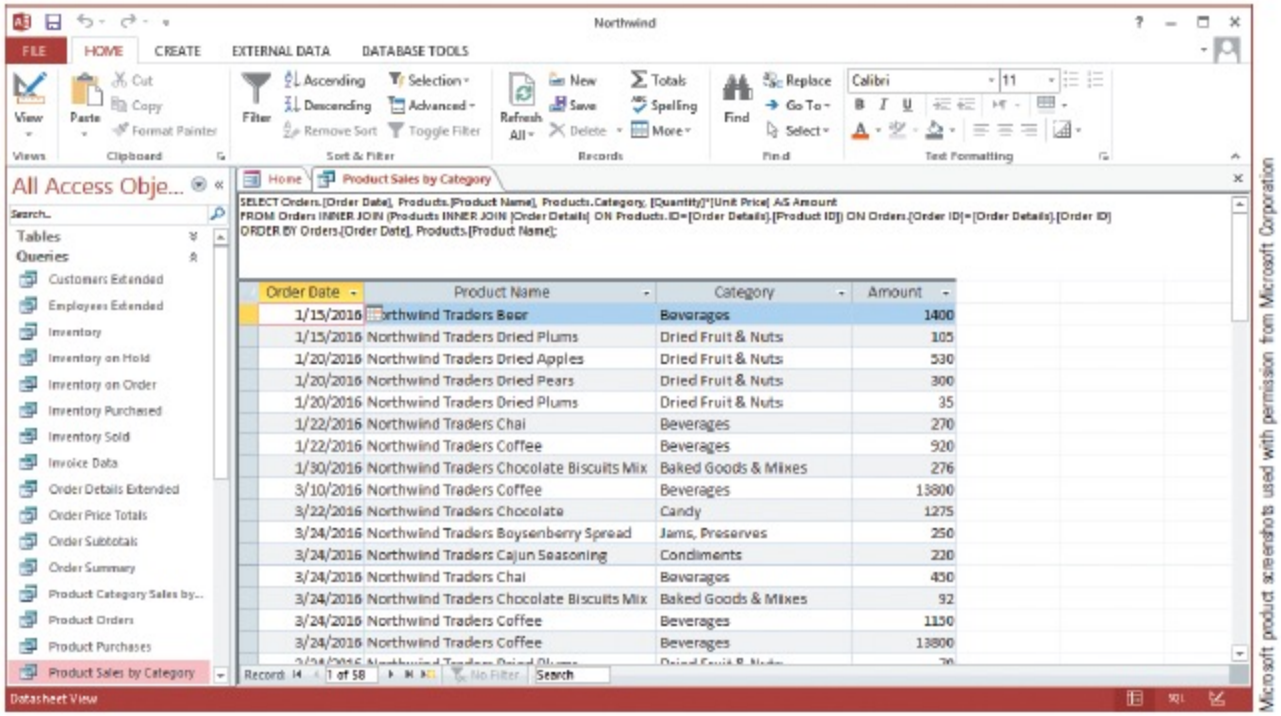



**FIGURE 5.12**

**Structured Query Language (SQL)**

SQL has become an integral part of most relational databases, as shown by this example from Microsoft Access 2013.

# Schema and Data Dictionary

- Schema: a description of the entire database
  - The DBMS can reference a schema to find where to access the requested data in relation to another piece of data
- Data dictionary: a detailed description of all the data used in the database

NORTHWESTERN MANUFACTURING

| | |
|---|---|
| PREPARED BY: | D. BORDWELL |
| DATE: | 04 AUGUST 2016 |
| APPROVED BY: | J. EDWARDS |
| DATE: | 13 OCTOBER 2016 |
| VERSION: | 3.1 |
| PAGE: | 1 OF 1 |
| DATA ELEMENT NAME: | PARTNO |
| DESCRIPTION: | INVENTORY PART NUMBER |
| OTHER NAMES: | PTNO |
| VALUE RANGE: | 100 TO 5000 |
| DATA TYPE: | NUMERIC |
| POSITIONS: | 4 POSITIONS OR COLUMNS |

**FIGURE 5.14**
**Data dictionary entry**
A data dictionary provides a detailed description of all data used in the database.

# Database Activities

- Providing a user view of the database
- Adding and modifying data
- Storing and retrieving data
- Manipulating the data and generating reports

# Storing and Retrieving Data

- When an application program needs data, it requests the data through the DBMS
- Concurrency control deals with the situation in which two or more users or applications need to access the same record at the same time

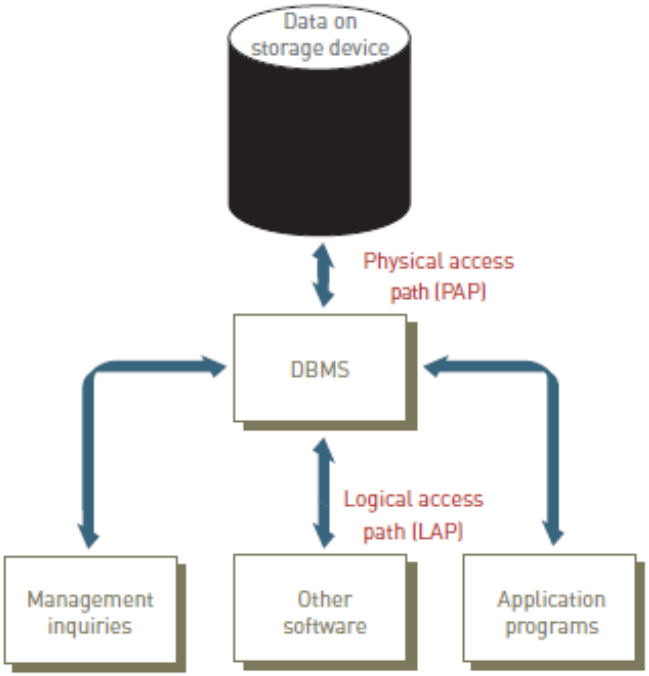



**FIGURE 5.15**
**Logical and physical access paths**
When an application requests data from the DBMS, it follows a logical access path to the data. When the DBMS retrieves the data, it follows a path to the physical access path to the data.

# Manipulating Data and Generating Reports

- A DBMS can produce a wide variety of documents, reports, and other output that can help organizations achieve their goals


**Figure 5.15**

Database Output

A database application offers sophisticated formatting and organization options to produce the right information in the right format.

# Database Administration

- Database administrators (DBAs): skilled and trained IS professionals
  - Works with users to define their data needs
  - Applies database programming languages to craft a set of databases to meet those needs
  - Assures that data is secure from unauthorized access
- Data administrator: a nontechnical position responsible for defining and implementing consistent principles for a variety of data issues
  - Including setting data standards and data definitions that apply across all the databases in an organization
    - Ensuring compliance with regulatory requirements and standards, such as GDPR, HIPAA, and PCI-DSS
      - PCI-DSS: https://www.youtube.com/watch?v=szVmMxWORBc
  - The data administrator can be a high-level position reporting to top-level managers

# Popular Database Management Systems

- Open-source relational DBMS:
  - MySQL
  - PostgreSQL
- Relational DBMS for Enterprise:
  - Oracle
  - Microsoft SQL Server
  - IBM DB2

# Popular Database Management Systems

- Database as a Service (DaaS)
  - A cloud-based service that provides database storage, management, and administration
  - The database is hosted on the service provider's servers and is accessed by the client over a network, typically the Internet
- Example of DaaS:
  - Amazon Relational Database Service (Amazon RDS)
  - Microsoft Azure SQL Database

# Using Databases with Other Software

- DBMSs can act as front-end or back-end applications
  - Front-end applications interact directly with people
  - Back-end applications interact with other programs or applications
- Example:
  - The Library of Congress (LOC) provides a back-end application that allows Web access to its databases, which include references to books and digital media in the LOC collection

# Big Data

- Extremely large and complex data collections
  - Traditional data management software, hardware, and analysis processes are incapable of dealing with them
- Three characteristics of big data
  - Volume
  - Velocity
  - Variety

# Sources of Big Data

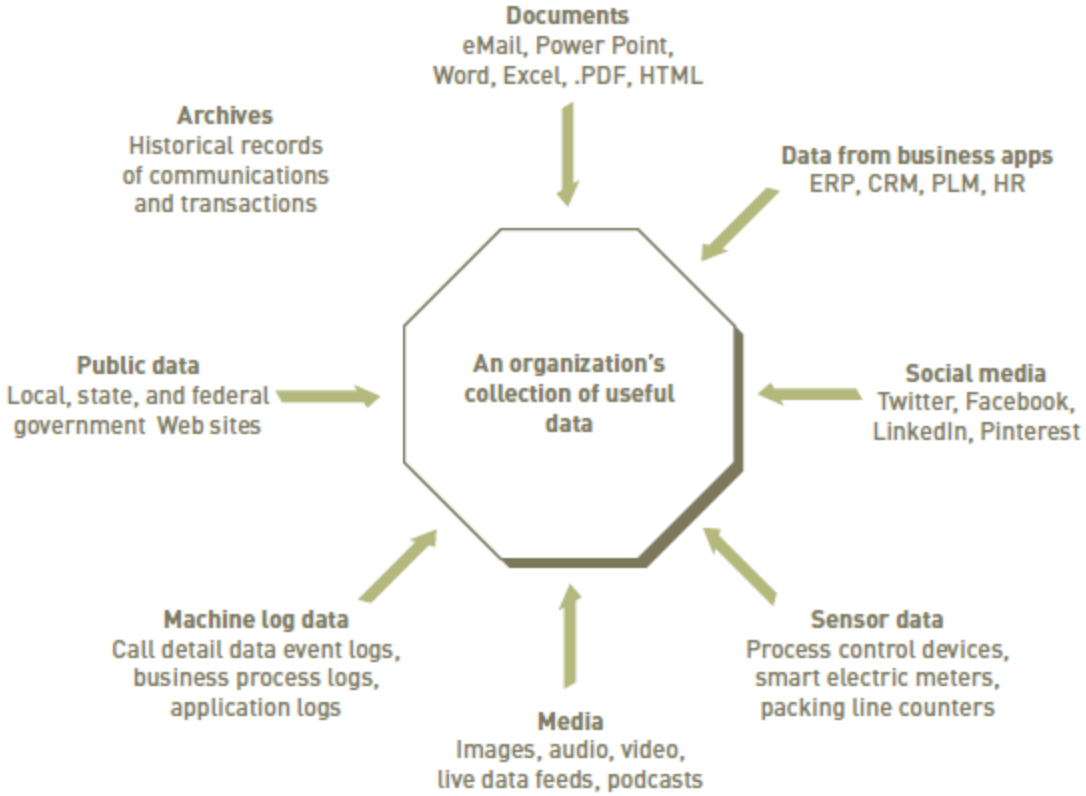



**FIGURE 5.20**

**Sources of an organization's useful data**

An organization has many sources of useful data.

# Big Data Uses

- Examples:
  - Retail organizations monitor social networks to engage brand advocates, identify brand adversaries
  - Advertising and marketing agencies track comments on social media
  - Hospitals analyze medical data and patient records
  - Consumer product companies monitor social networks to gain insight into consumer behavior
  - Financial service organizations use data to identify customers who are likely to be attracted to increasingly targeted and sophisticated offers

# Challenges of Big Data

- How to choose what subset of the data to store
- Where and how to store the data
- How to find the 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

# Data Warehouses and Data Lakes

- Data warehouse: a large database that collects historical business information, typically spanning multiple years (e.g., 3, 5, or 10 years), from many sources in the enterprise in support of management decision making
- ETL process
  - Extract
  - Transform
  - Load

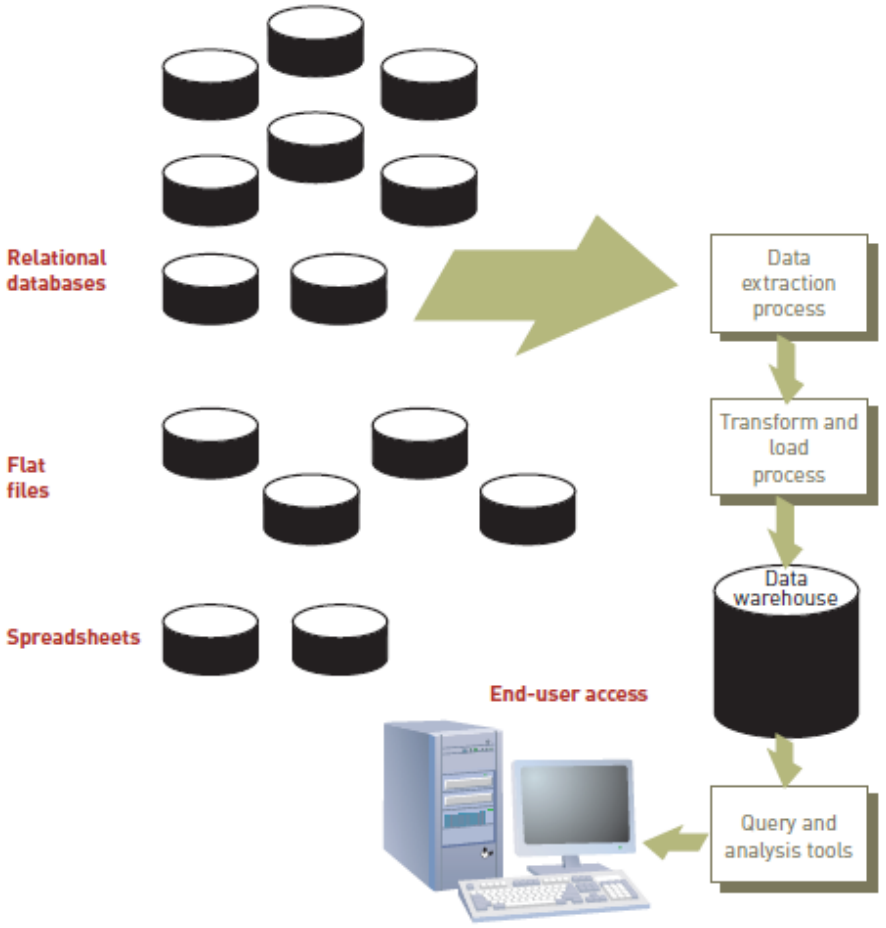



**FIGURE 5.23**
**Elements of a data warehouse**
A data warehouse can help managers and executives relate information in innovative ways to make better decisions.

# Data Warehouses and Data Lakes

- Data lake: takes a “store everything” approach to big data, saving all the data in its raw and unaltered form
  - Also called an enterprise data hub
  - Raw data is available when users decide just how they want to use the data
  - Only when the data is accessed for a specific analysis is it extracted from the data lake

# Business Intelligence (BI) and Business Analytics

- Business intelligence (BI)
  - Includes a wide range of applications, practices, and technologies for the extraction, transformation, integration, visualization, analysis, interpretation, and presentation of data to support improved decision making
  - BI tools and techniques can be used to analyze data from a data warehouse to identify trends, patterns, and anomalies
  - Using data for description
  - Examples of BI software:
    - Tableau
      - https://public.tableau.com/app/profile/jan.kaiser1651/viz/HelpDeskRealWorldFakeDataRWFD/Overview_page
      - https://public.tableau.com/app/profile/utkarsha.shinde/viz/BuyerPersona_CaseStudy/Dashboard1
    - Microsoft Power BI
- Business analytics
  - The extensive use of data and quantitative analysis to support fact-based decision making within organizations
  - Using data for prediction

# NoSQL Databases

- NoSQL is a type of database that does not use a relational table structure
- NoSQL databases have a flexible schema that can be dynamically changed
- NoSQL databases can store data in a variety of ways, including key-value pairs and graphs
- NoSQL databases are often used for applications that need to be scalable

# Types of NoSQL Databases

- Document databases
  - Document databases store data in semi-structured documents, typically using formats like JSON and others
  - Each document can have a different structure, and there is no fixed schema
  - Commonly used for content management systems and web applications
  - Example: MongoDB
- Column-oriented databases
- Graph Databases
- And others

# In-Memory Databases

- In-memory database (IMDB)
  - A database management system that stores the entire database in random access memory (RAM)
  - Provides access to data at rates much faster than storing data on some form of secondary storage

# Summary

- The database approach to data management has become broadly accepted
- 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
- Recently, organizations deal with a tremendous growth in the amount of data available and struggling 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