ISD100-Introduction to Systems & Informatics

Database Systems and Big Data

Dr. Riham Moharam

Faculty of Information Technology & Computer Science

Sinai University

North Sinai, Egypt

Introduction

- Data consists of raw facts
- Data must be organized in a meaningful way to transform it into useful information.
- **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.
- **Database administrator (DBA):**
  - Skilled IS professional who directs all activities related to an organization's database.

Introduction

Data Management:

- Without data and the ability to process the data:
  - An organization could not successfully complete most business activities.
- To transform data into useful information:
  - It must first be organized in a meaningful way.

The Hierarchy of Data

- **Bit (a binary digit):**
- **Byte:** typically made up of eight bits.
- **Character:** basic building block of information.
- **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.
- **Database:** collection of integrated and related data.
- **Hierarchy of data:** bits, characters, fields, records, files, and databases.

The Hierarchy of Data

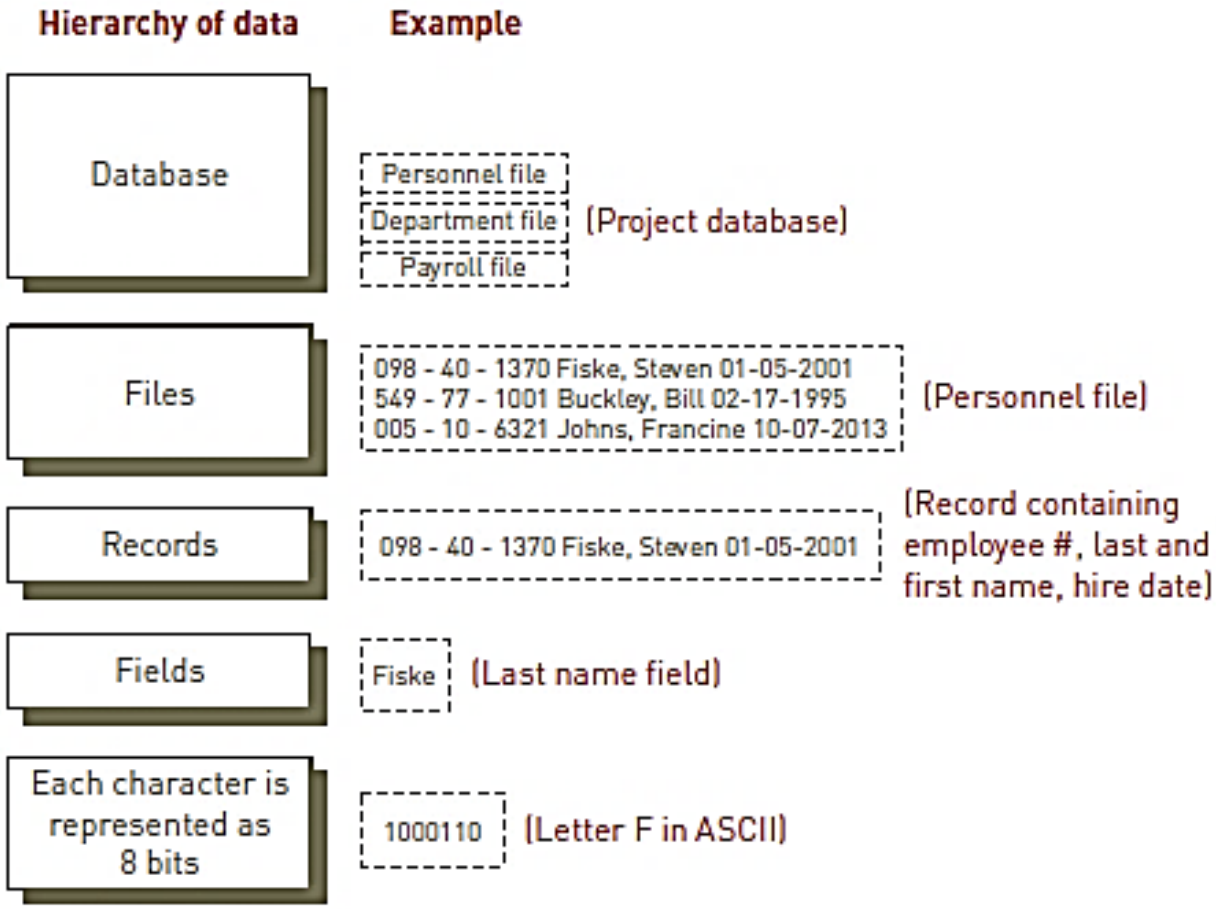


FIGURE 5.1

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

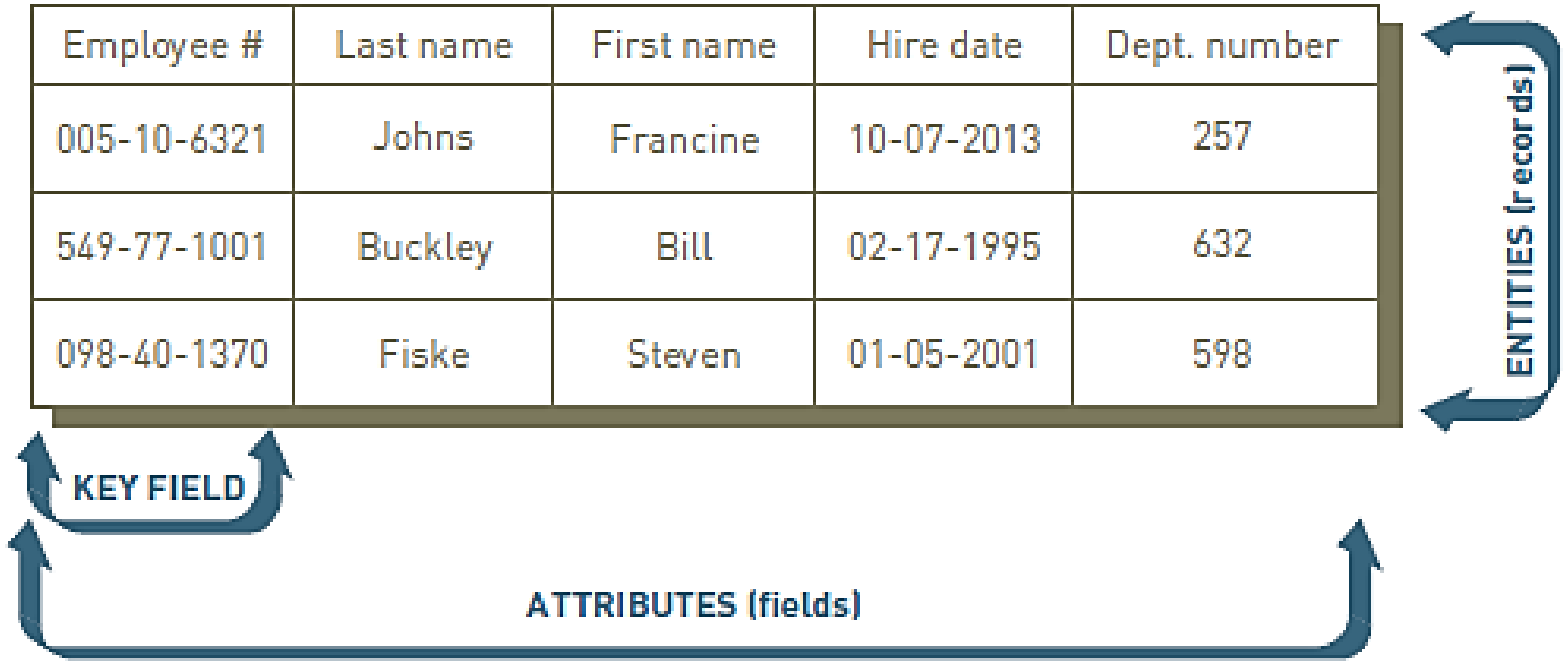


FIGURE 5.2

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

❖ **Two approaches of data management:**

- 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

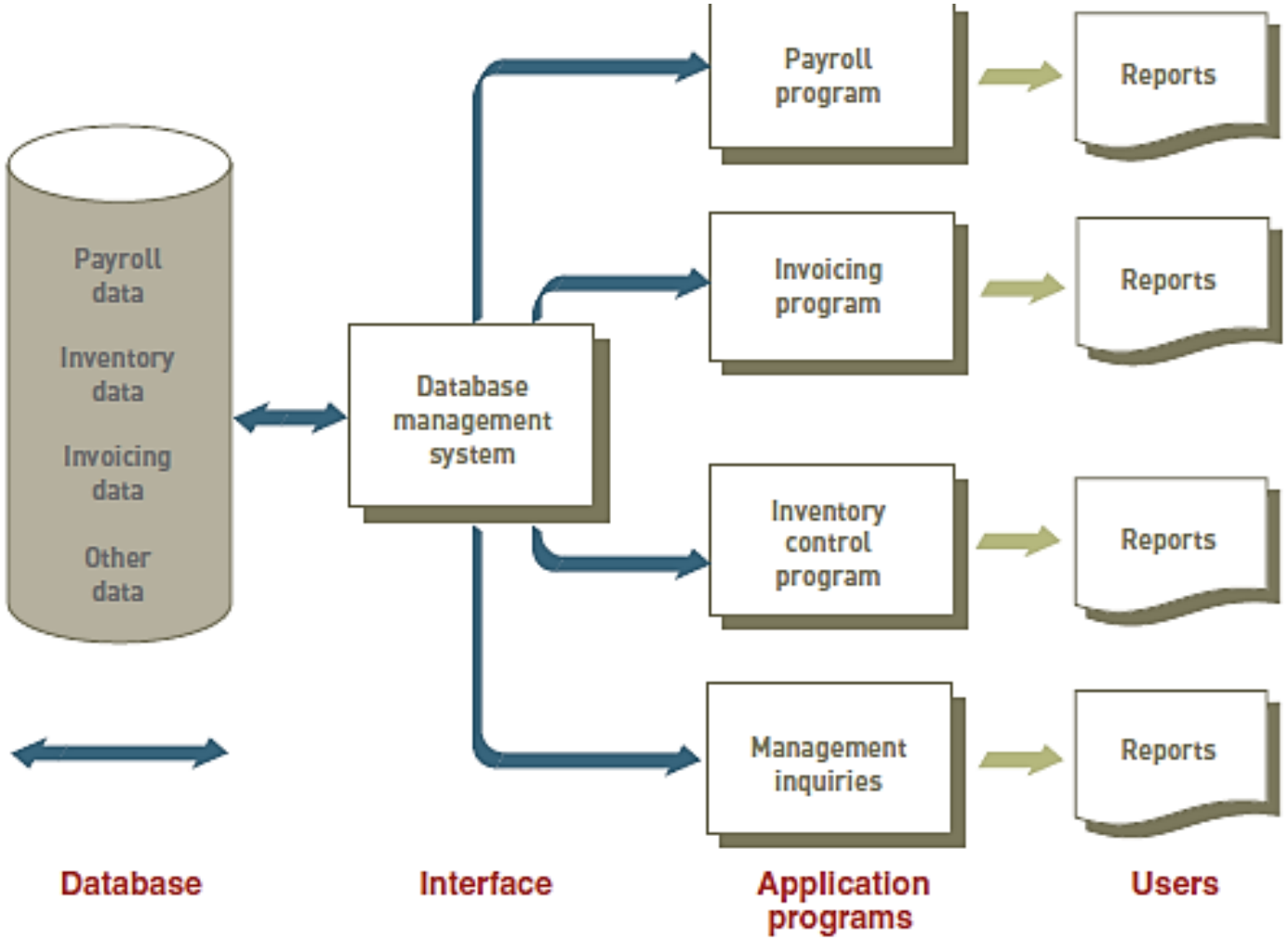


FIGURE 5.4

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?
  - **Security:** how can data be protected?

Data Modeling

- **Data model**: a diagram of data entities and their relationships.
- **Enterprise data modeling**: starts by investigation the general data and information needs of the organization at the **strategic level**.
- **Entity-relationship (ER) diagrams**: data models that use basic graphical symbols to show the organization entities of and relationships between data.

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 a data entity (record).
  - Each column represents an attribute of that entity (fields).
- **Domain**: range of allowable values for a data attribute.

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

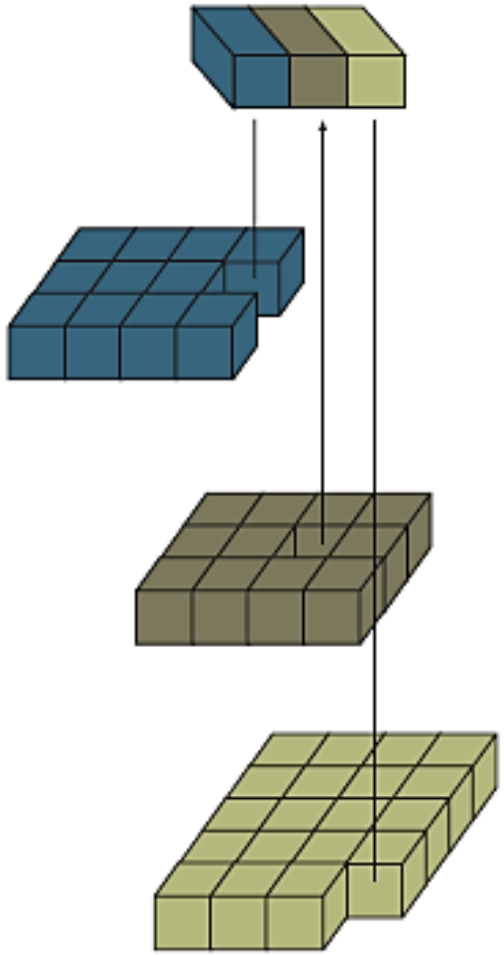
Relational database model

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



Data Table 1: Project Table

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

Data Table 2: Department Table

Dept. number	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 5.9

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

- **Data cleanup:**
  - The process of detecting and then correcting or deleting incomplete, incorrect, inaccurate, irrelevant records that reside in a database.
  - Eliminate redundancies and anomalies (problems in data).
  - The cost of performing data cleansing can be quite high.

Data Center

- **Climate-controlled building or set of buildings that:**
  - Houses database servers and the systems that deliver mission-critical information and services.
- **Traditional data centers:**
  - Consist of warehouses filled with row upon row of server racks and powerful cooling systems.

Database Types

- **Flat file:**
  - Simple database program whose records have no relationship to one another.
- **Single user:**
  - Only one person can use a database at a time.
  - Ex: Access
- **Multiple users:**
  - Allow dozens or hundreds of people to access the same database system at the same time.
  - Ex: SQL Server and Oracle

Providing a User View

- **Schema**:
  - A description of the entire database.
  - A schema can be part of the database or a separate schema file.
- **DBMS**:
  - Can reference a schema to find where to access the requested data in relation to another piece of data.

Creating and Modifying the Database

- **Data definition language (DDL)**
  - A collection of instructions and commands used to define and describe data and relationships in a specific database.
  - Allows the database's creator to describe data and relationships that are to be contained in the schema.
- **Data dictionary**: a detailed description of all the data used in the database
  - Can also include a description of data flows, information about the way records are organized, and the data-processing requirements.

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.

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.

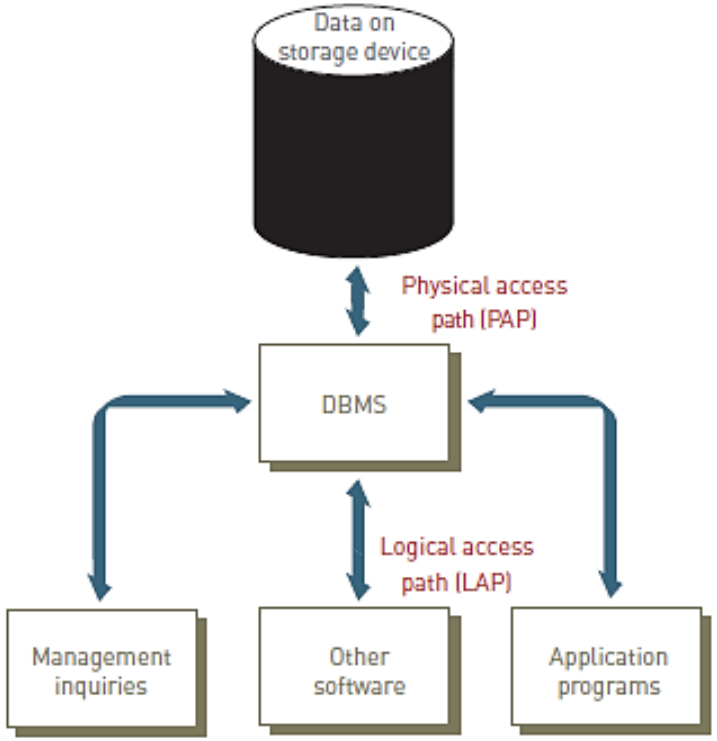


FIGURE 5.15
Logical and physical access paths
When an application requests data from the DBMS, it follows a logical access path to the data. When the DBMS retrieves the data, it follows a path to the physical access path to the data.

Manipulating Data and Generating Reports

- **Data manipulation language (DML)**:
  - A specific language, provided with a DBMS
  - Allows users to access and modify the data, to make queries, and to generate reports.
- **Structured Query Language (SQL):**
  - Adopted by the American National Standards Institute (ANSI) as the standard query language for relational databases.
- A DBMS can produce a wide variety of documents, reports, and other output that can help organizations achieve their goals.

Popular Database Management Systems

- **Database as a Service (DaaS):**
  - The database is stored on a service provider's servers.
  - The database is accessed by the client over a network, typically the Internet.
  - Database administration is handled by the service provider.
- **Example of DaaS:**
  - Amazon Relational Database Service (Amazon RDS).

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.

Big Data

- Big Data is the current term for the enormous datasets generated by Web and mobile applications such as search tools (for example, Google and Bing), Web 2.0 social networks (for example, Facebook, LinkedIn and Twitter), and scientific data collection tools.
- Big data is referring to terabytes and petabytes of data.

					Big data		
Bytes	10^2	10^4	10^6	10^8	10^{10}	10^{12}	$10^{>12}$
Size	tiny	small	medium	large	huge	monster	Very large

Sources of Big Data

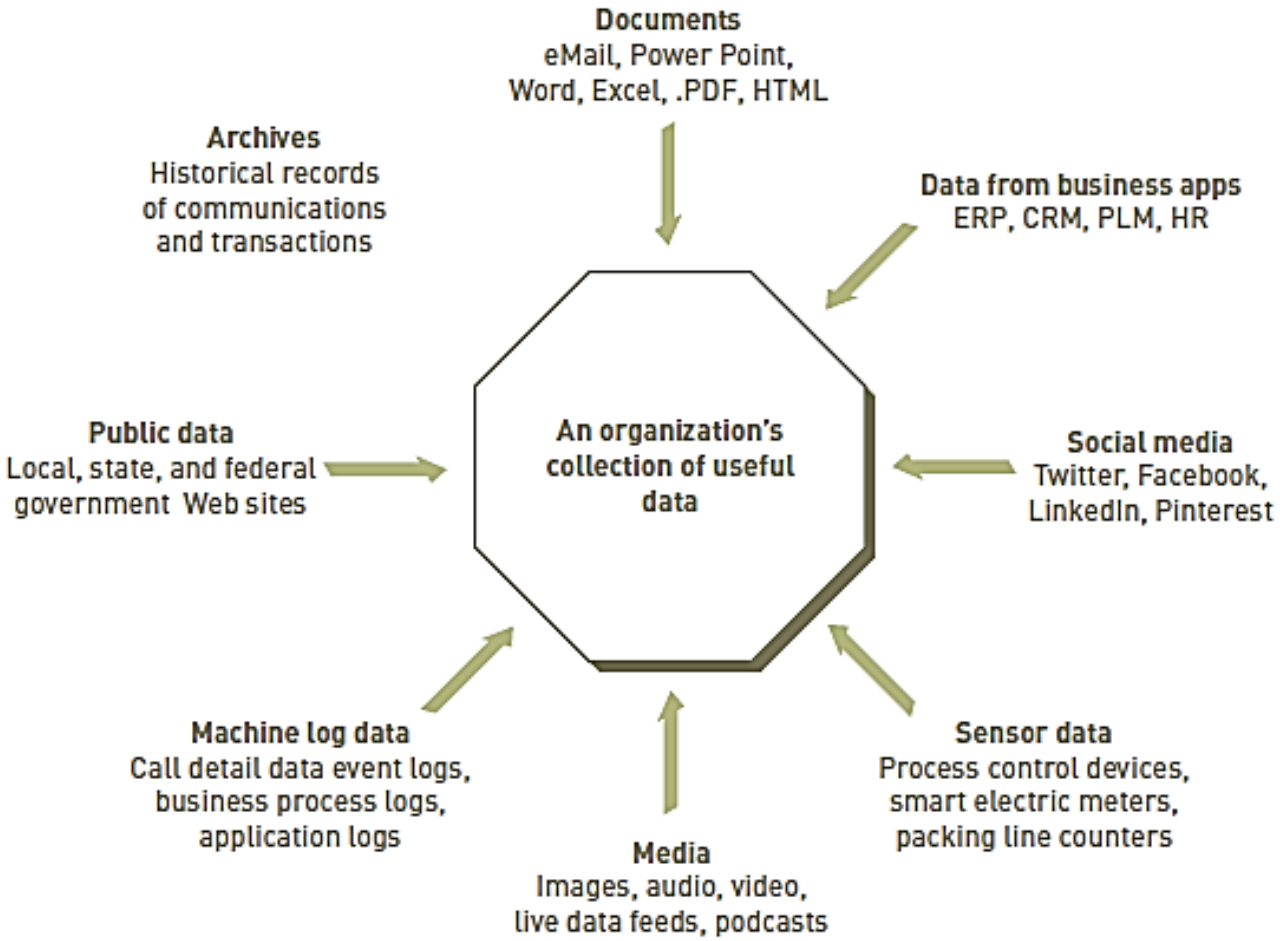
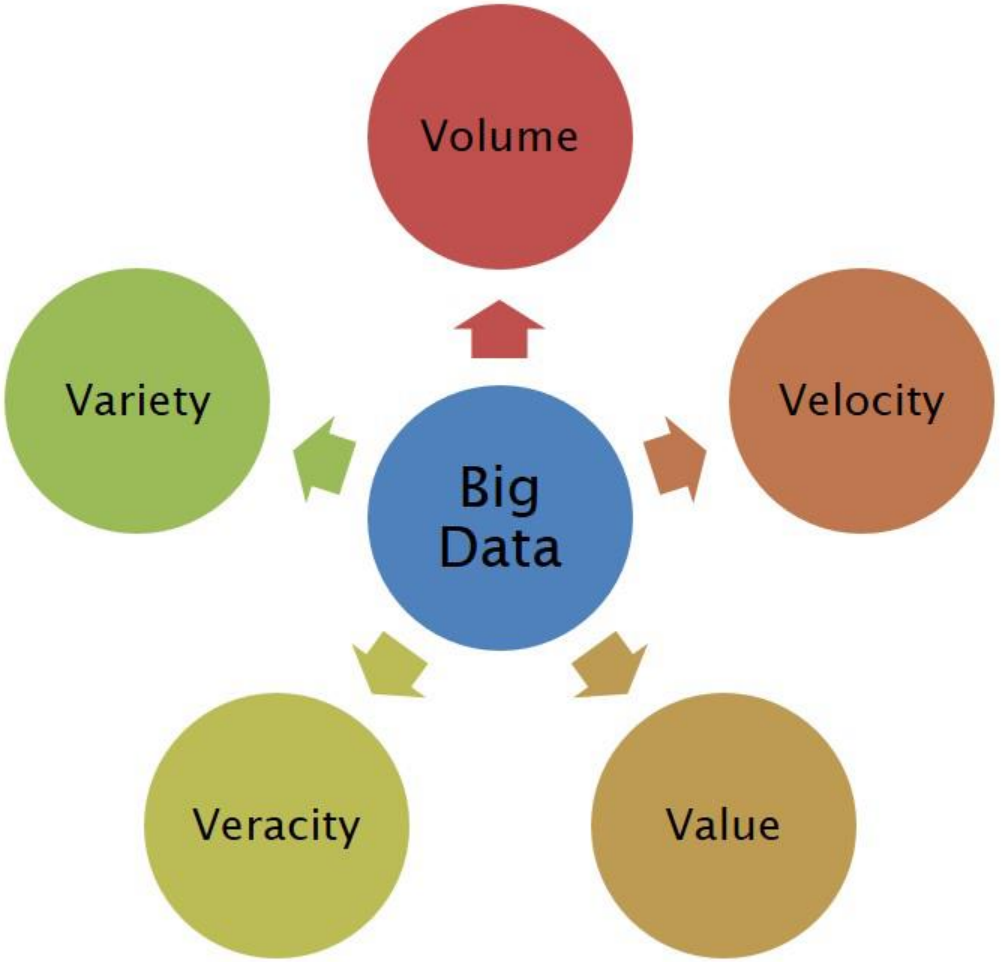


FIGURE 5.20

Sources of an organization's useful data

An organization has many sources of useful data.

Characteristics of Big Data (5Vs)



Characteristics of Big Data (5Vs)

- Volume:
  - It is indicates to the size of data. Analyzing data with very large volume to extract valuable information is one of important challenges of big data.
- Velocity:
  - The term velocity is referring to the speed of data. Flooding of data is very high speed and it has to be dealt with in appropriate time.

Characteristics of Big Data (5Vs)

- Variety:
  - The data is very diverse and has many types as it comes from different sources with different structures such as: social data, audio, video unstructured data, email and etc.
- Value:
  - Another challenging issue is to convert the data into values to understand and discover hidden values.

Characteristics of Big Data (5Vs)

- Veracity:
  - Data veracity, in general, is how accurate or truthful a data set may be. More specifically, when it comes to the accuracy of big data, it's not just the quality of the data itself but how trustworthy the data source, type, and processing of it is.

Challenges of Big Data

- Big data is facing many challenges such as data capture, storage, visualization, analysis, and updating data securely.
- Analyzing data with very large volume to extract valuable information is one of important challenges of big data.
- Extracting or mining valuable information from huge amounts of data is referred by data mining methods.

Data Management

- **Data management:**
  - 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.
- **Data governance:**
  - Defines the roles, responsibilities, and processes for ensuring that data can be trusted and used by an entire organization.

Data Management

FIGURE 5.21

Data management

The Data Management Association (DAMA) International has identified 10 basic functions associated with data management.

Source: "Body of Knowledge," DAMA International, *https://www.dama.org/content/body-knowledge*. Copyright DAMA International.

Data Management

- Data lifecycle management (DLM)
  - A policy-based approach for managing the flow of an enterprise's data.

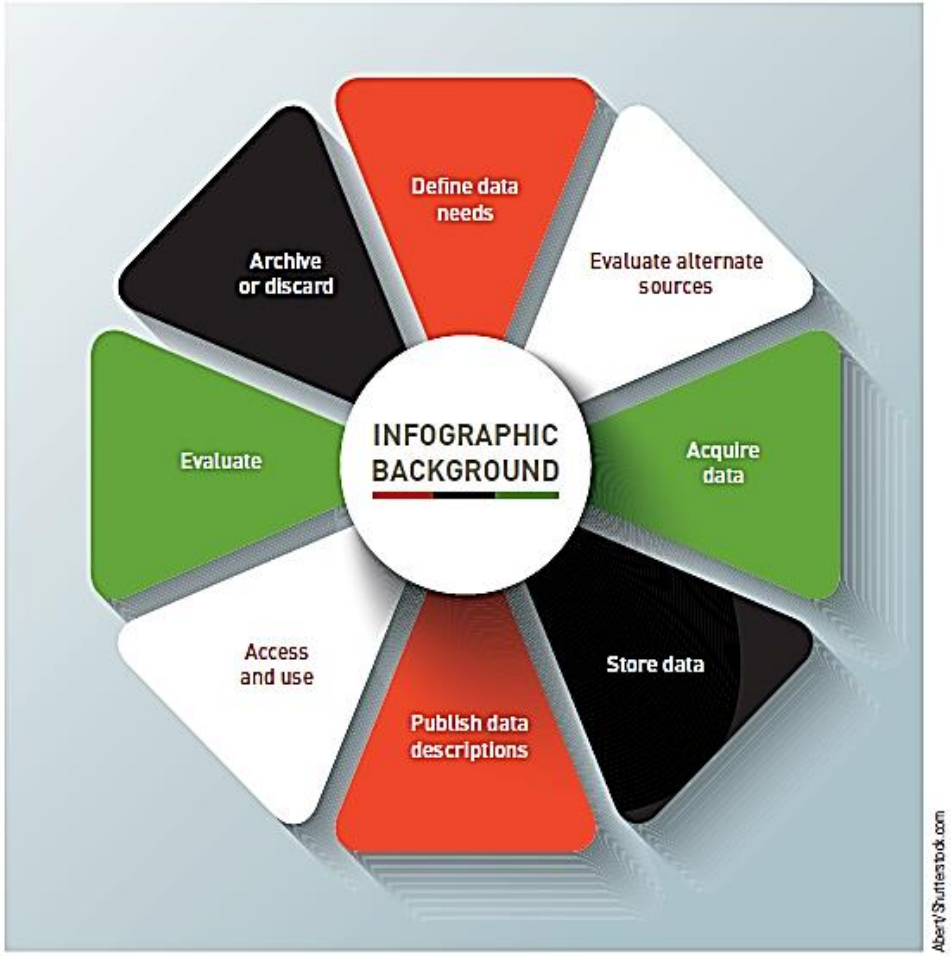


FIGURE 5-22
The big data life cycle
A policy-based approach to managing the flow of an enterprise's data, from its initial acquisition or creation and storage to the time when it becomes outdated and is deleted.

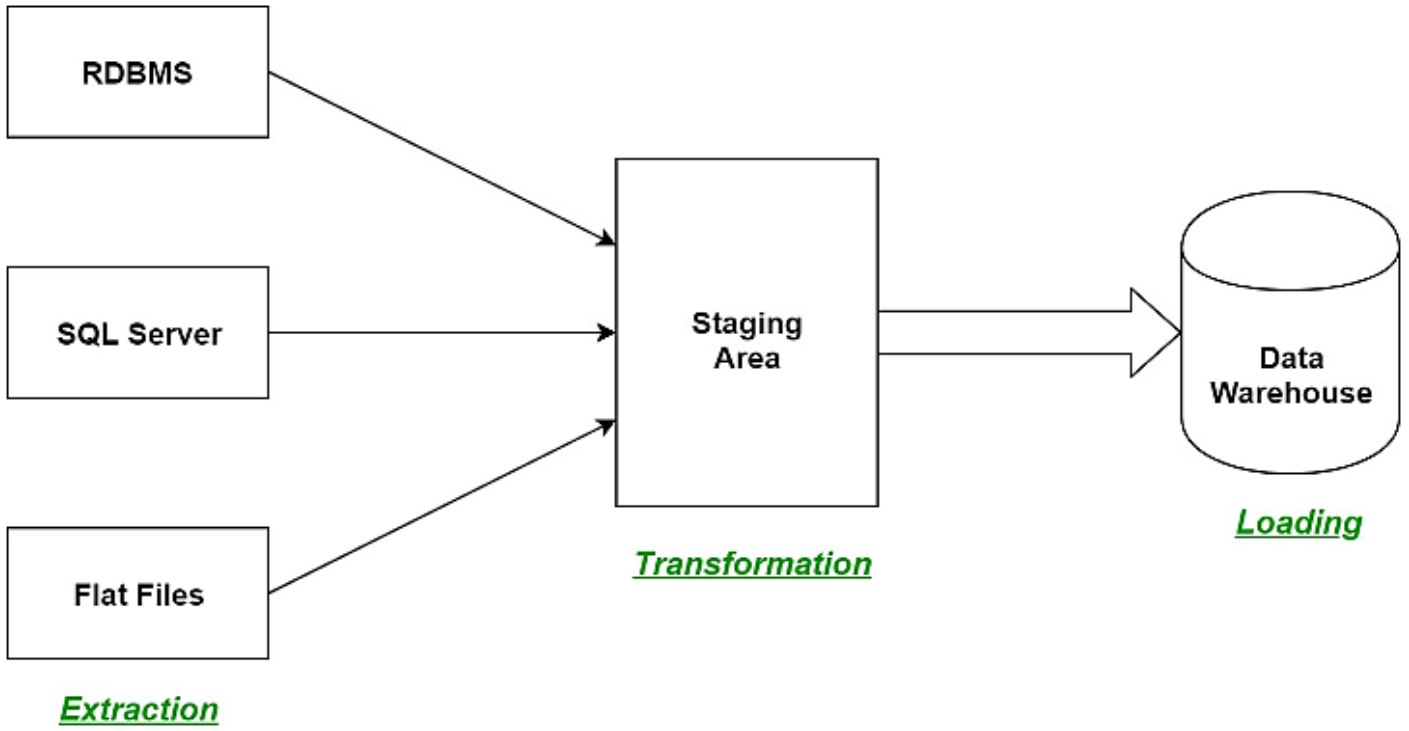
Data Warehouses and Data Marts

- **Data warehouse** is a type of data management system.
- **Data warehouse**: a large database that collects business information from many sources in the enterprise in support of management decision making.
- **ETL process:**
  1. Extract
  2. Transform
  3. Load

➢ ETL stands for Extract, Transform, Load and it is a process used in data warehousing to extract data from various sources, transform it into a format suitable for loading into a data warehouse, and then load it into the warehouse.

Data Warehouses and Data Marts

- The ETL process is an iterative process that is repeated as new data is added to the warehouse. The process is important because it ensures that the data in the data warehouse is accurate, complete, and up-to-date.



Data Warehouses and Data Marts

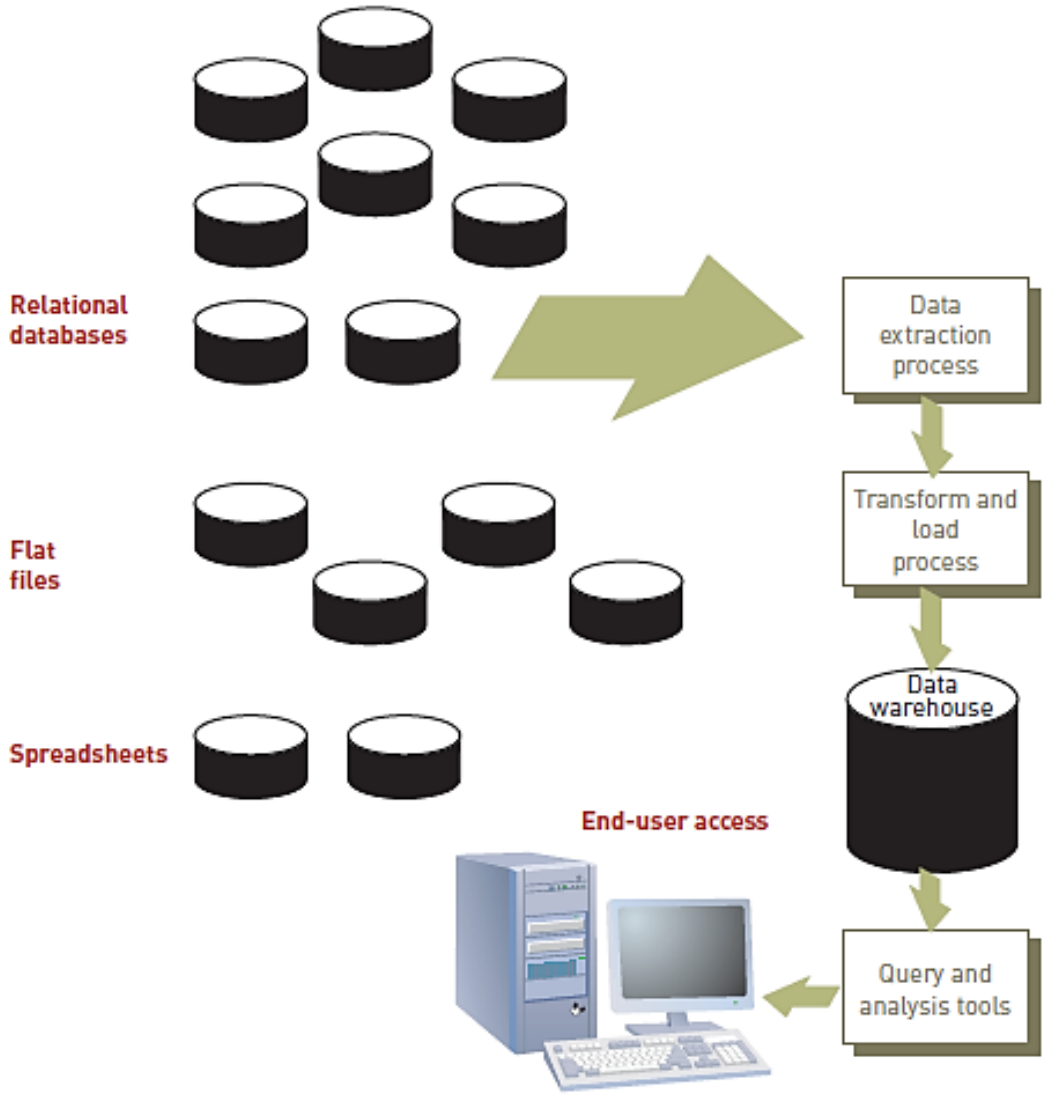


FIGURE 5.23
Elements of a data warehouse
A data warehouse can help managers and executives relate information in innovative ways to make better decisions.

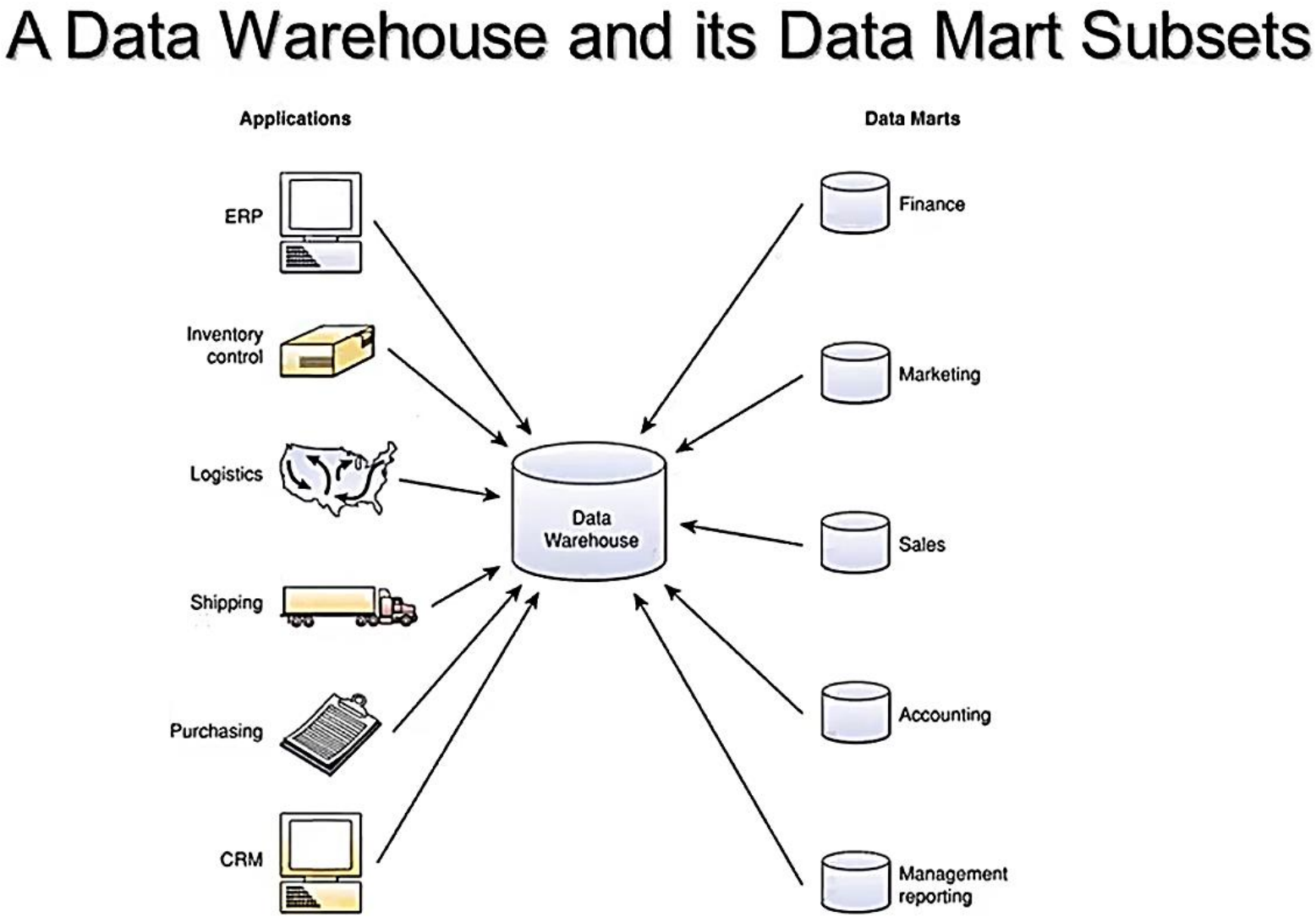
Data Warehouses and Data Marts

- A typical data warehouse often includes the following elements:
  - A relational database to store and manage data.
  - An extraction, loading, and transformation (ELT) solution for preparing the data for analysis.
  - Analysis tools, reporting, and data mining capabilities.
  - Client analysis tools for visualizing and presenting data to business users.

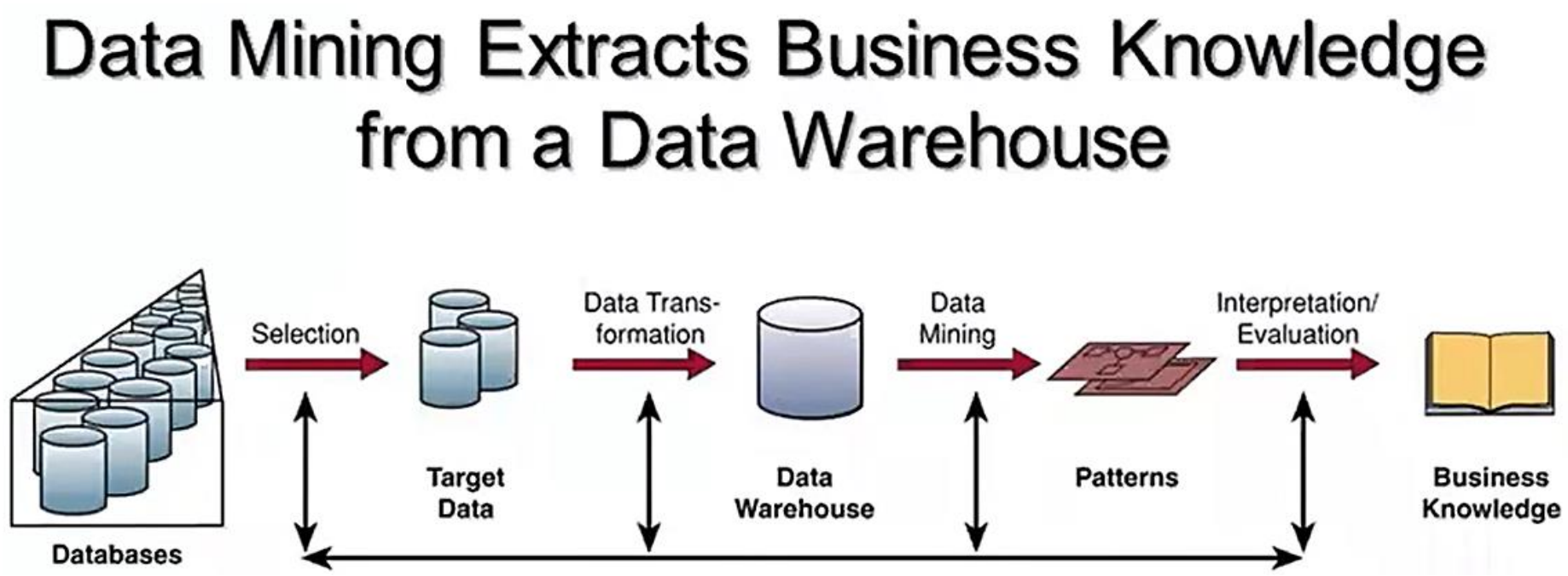
Data Warehouses and Data Marts

- **Data mart**: a subset of a data warehouse that is used by small- and medium-sized businesses and departments within large companies to support decision making.
- A specific area in the data mart might contain greater detailed data than the data warehouse.

Data Warehouses and Data Marts



Data Warehouses and Data Mining



NoSQL Databases

- Big Data datasets are often stored in **non-relational databases**, which are often referred to as **NoSQL** databases. The term **NoSQL**, however, is really a bit of a misnomer. It means, literally, a database that doesn't use SQL.
- **NoSQL database**
  - Provides a means to store and retrieve data that is modeled using some means other than the simple two-dimensional tabular relations used in relational databases.
- **Advantages:**
  - Ability to spread data over multiple servers so that each server contains only a subset of the total data.
  - Do not require a predefined schema.

NoSQL Databases

- **Key–value NoSQL** databases are similar to SQL databases, but have only two columns ("key" and "value"), with more complex information sometimes stored within the "value" columns.
- **Document NoSQL** databases are used to store, retrieve, and manage document-oriented information, such as social media posts and multimedia, also known as semi-structured data.
- **Graph NoSQL** databases are used to understand the relationships among events, people, transactions, locations, and sensor readings and are well suited for analyzing interconnections such as when extracting data from social media.
- **Column NoSQL** databases store data in columns, rather than in rows, and are able to deliver fast response times for large volumes of data.

TABLE 5.5 Popular NoSQL database products, by category

Key–Value	Document	Graph	Column
HyperDEX	Lotus Notes	Allegro	Accumulo
Couchbase Server	Couchbase Server	Neo4J	Cassandra
Oracle NoSQL Database	Oracle NoSQL Database	InfiniteGraph	Druid
OrientDB	OrientDB	OrientDB	Vertica
	MongoDB	Virtuoso	HBase

Hadoop

- Hadoop
  - An open-source software framework that includes several software modules that provide a means for storing and processing extremely large data sets.
- Has two primary components:
  - A data processing component (**MapReduce**).
  - A distributed file system (**Hadoop Distributed File System, HDFS**).

Hadoop

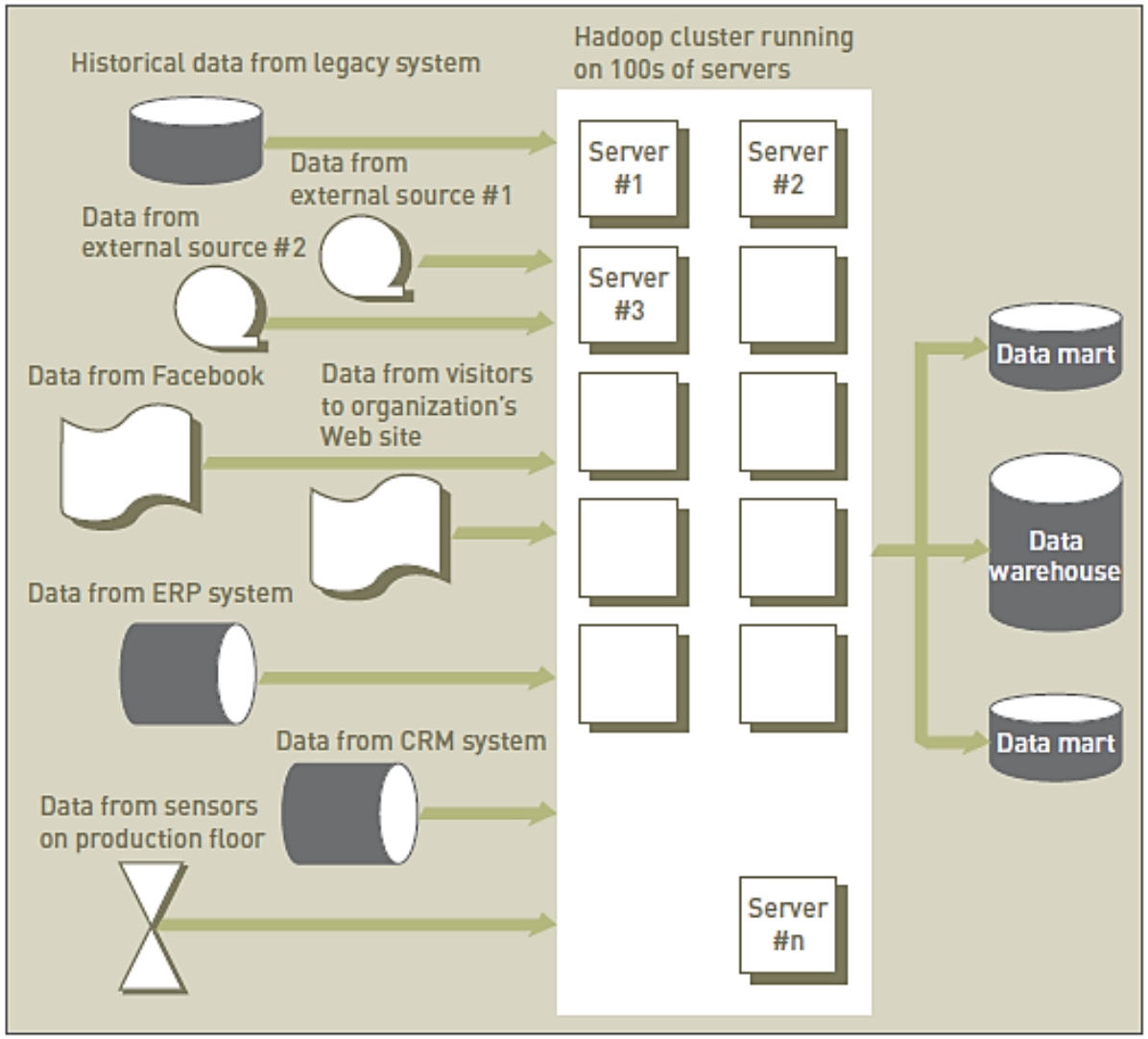


FIGURE 5.24

Hadoop environment

Hadoop can be used as a staging area for data to be loaded into a data warehouse or data mart.