



សាកលវិទ្យាល័យភូមិន្ទភ្នំពេញ

ROYAL UNIVERSITY OF PHNOM PENH

CHAPTER

5

Database Systems and Big Data

MIS

Chea Daly



Did You Know?

- ❑ The amount of data in the digital universe is expected to increase to 44 zettabytes (44 trillion gigabytes) by 2020.
- ❑ The majority of data generated between now and 2020 will not be produced by humans, but rather by machines as they talk to each other over data networks.



Why Learn about Database Systems and Big Data?

- ❑ 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.



Why Learn about Database Systems and Big Data?

- ❑ 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.
- ❑ 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.



Database

- ❑ A **database** is a well-designed, organized, and carefully managed collection of data.
- ❑ Databases help companies analyze information to reduce costs, increase profits, add new customers, track past business activities, and open new market opportunities.



Data Management

- ❑ Most organizations have many databases; however, without good data management, it is nearly impossible for anyone to find the right and related information for accurate and business-critical decision making.
- ❑ For data to be transformed into useful information, it must first be organized in a meaningful way.



Data Entry and Input

- Data can be human-readable or machine-readable.
- “Human-readable data” means data that people can read and understand.
- An example of machine-readable data is the universal barcode on many grocery and retail items that indicates the stock-keeping identification number for that item.





Data Entry and Input

- Getting data into the computer system is a two-stage process:
 - Data entry: Converts human-readable data into machine-readable form.
 - Data input: Transfers machine-readable data into system.



Data Entry and Input

- Data entry:
 - Is the process of transferring data from manual records to a digital database.
 - you make information available to computer systems in a manner and form where it can be understood.
 - The information by itself is perhaps relevant and useful in many different ways but that can only be realized if it becomes computer readable.



Data Entry and Input

- Data input
 - Is the process of providing relevant information to a software program so that it can produce output.
 - The input enables the computer to do what is designed to do and produce an output. Thus, the word or phrase that you type into the text box of your search engine is the input which it will process and produce an output for you.



The Hierarchy of Data

- ❑ A **Bit** is a binary digit (i.e., 0 or 1) that represents a circuit that is either on or off. Bits can be organized into units called bytes.
- ❑ A **Byte** is made up of eight bits. Each byte represents a character.
- ❑ A **Character** is a basic building block of information. Characters are put together to form a field.
- ❑ A **Field** is name, number, or combination of characters that describes an aspect of a business object (such as an employee) or activity (such as a sale).



The Hierarchy of Data

- A **Record** is a collection of related fields.
 - For example, an employee record is a collection of fields about one employee.
- A **File** is a collection of related records.
 - for example, an employee file is a collection of all company employee records.
- A **Database** is collection of integrated and related files.



The Hierarchy of Data

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



The Hierarchy of Data

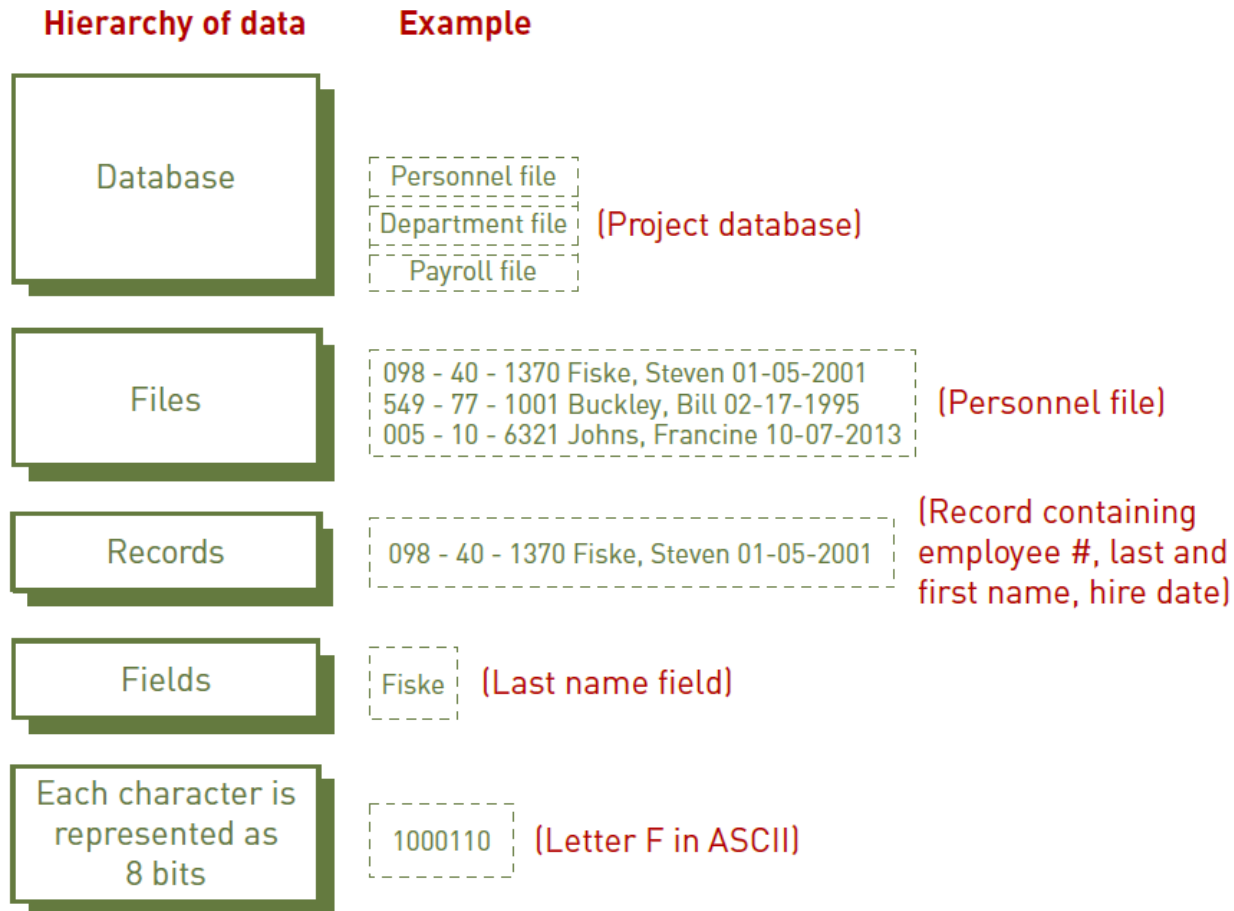


FIGURE 5.1

Hierarchy of data

Together, bits, characters, fields, records, files, and databases form the hierarchy of data.



Organizations and Databases

- ❑ Many organizations create databases of to store data needed to run their day-to-day operations.



Case Study: U.S. Wireless Service Providers

- ❑ Major U.S. wireless service providers have implemented a stolen-phone database to report and track stolen 3G and 4G/LTE phones.
- ❑ The providers use the database to check whether a consumer's device was reported lost or stolen.
- ❑ If a device has been reported lost or stolen, it will be denied service on the carrier's network. Once the device is returned to the rightful owner, it may be reactivated.



Data Entities, Attributes, and Keys

- ❑ Entities, attributes, and keys are important database concepts.
- ❑ An **Entity** is a person, place, or thing (object) for which data is collected and stored.
 - ❑ Examples of entities include employees, products, and customers.
- ❑ Most organizations organize and store data as entities.



Data Entities, Attributes, and Keys

- 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.
- The specific value of an attribute, called a **data item**.



Keys and Attributes

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.

Employee #	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

KEY FIELD



ATTRIBUTES (fields)

ENTITIES (records)



Data Entities, Attributes, and Keys

- A **Primary key** is a field or set of fields that uniquely identifies the record.
- No other record can have the same primary key. Primary keys ensure that each record in a file is unique.


[Go](#)
[Buy](#)
[Sell](#)
[My eBay](#)
[Community](#)
[Help](#)



[CATEGORIES](#)
[ELECTRONICS](#)
[FASHION](#)
[MOTORS](#)
[TICKETS](#)
[DEALS](#)
[CLASSIFIEDS](#)


[Parts & Accessories](#)
[Cars & Trucks](#)
[Motorcycles](#)
[Powersports, Boats & More](#)
[MY VEHICLES](#)
[TIRE CENTER](#)
[LIGHT CENTER](#)

[Back to search results](#) | [eBay Motors](#) > [Cars & Trucks](#) > [Chevrolet](#) > [Malibu](#)
[Add to Watch list](#)

2013 Chevrolet Malibu Eco

Eco Hybrid-electric New 2.4L CD Preferred Equipment Group 1SA AM/FM radio [Research 2013 Chevrolet Malibu](#)



Item Location: 

Advertised price: **US \$26,160.00**
[Make Offer](#)

Phone: **(888) 468-2047**
[Add to Watch list](#)

[Order an independent inspection](#)

Coverage: This vehicle is eligible for up to \$50,000 in Vehicle Purchase Protection when your transaction is completed online through eBay. To qualify you must be the winning bidder on an auction or click the Buy It Now button directly on the eBay site. [Restrictions Apply](#). (Not eligible for eBay Buyer Protection)

Seller info

(32 ★)
 100% Positive feedback

[Ask a question](#)
[Save this seller](#)
[See other items](#)

Other item info

Item number: **110868309963**
 Item condition: **New**
 Sells to: **Local pick-up only**




Share:    [Print](#) | [Report item](#)

FIGURE 5.3

Primary key

eBay assigns an Item number as a primary key to keep track of each item in its database.



Data Management

- ❑ **Traditional approach to data management:**
At one time, information systems referenced specific files containing relevant data.
 - ❑ For example, a payroll system would use a payroll file.
 - ❑ Each distinct operational system used data files dedicated to that system.



Data Management

- Today, most organizations use the **database approach to data management**, where multiple information systems share a pool of related data.



The Database Approach

- A database offers the ability to share data and information resources.
 - For example, Federal databases often include the results of DNA tests as an attribute for convicted criminals. The information can be shared with law enforcement officials around the country.



The Database Approach

- Often, distinct yet related databases are linked to provide enterprise-wide databases.
 - For example, many Walgreens stores include in-store medical clinics for customers. Walgreens uses an electronic health records database that stores the information of all patients across all stores. The database provides information about customers' interactions with the clinics and pharmacies.



The Database Approach

- ❑ To use the database approach to data management, a database management system (DBMS) is required.
- ❑ DBMS consists of a group of programs used to access and manage a database.
- ❑ DBMS provides an interface between the database and its users and other application programs.

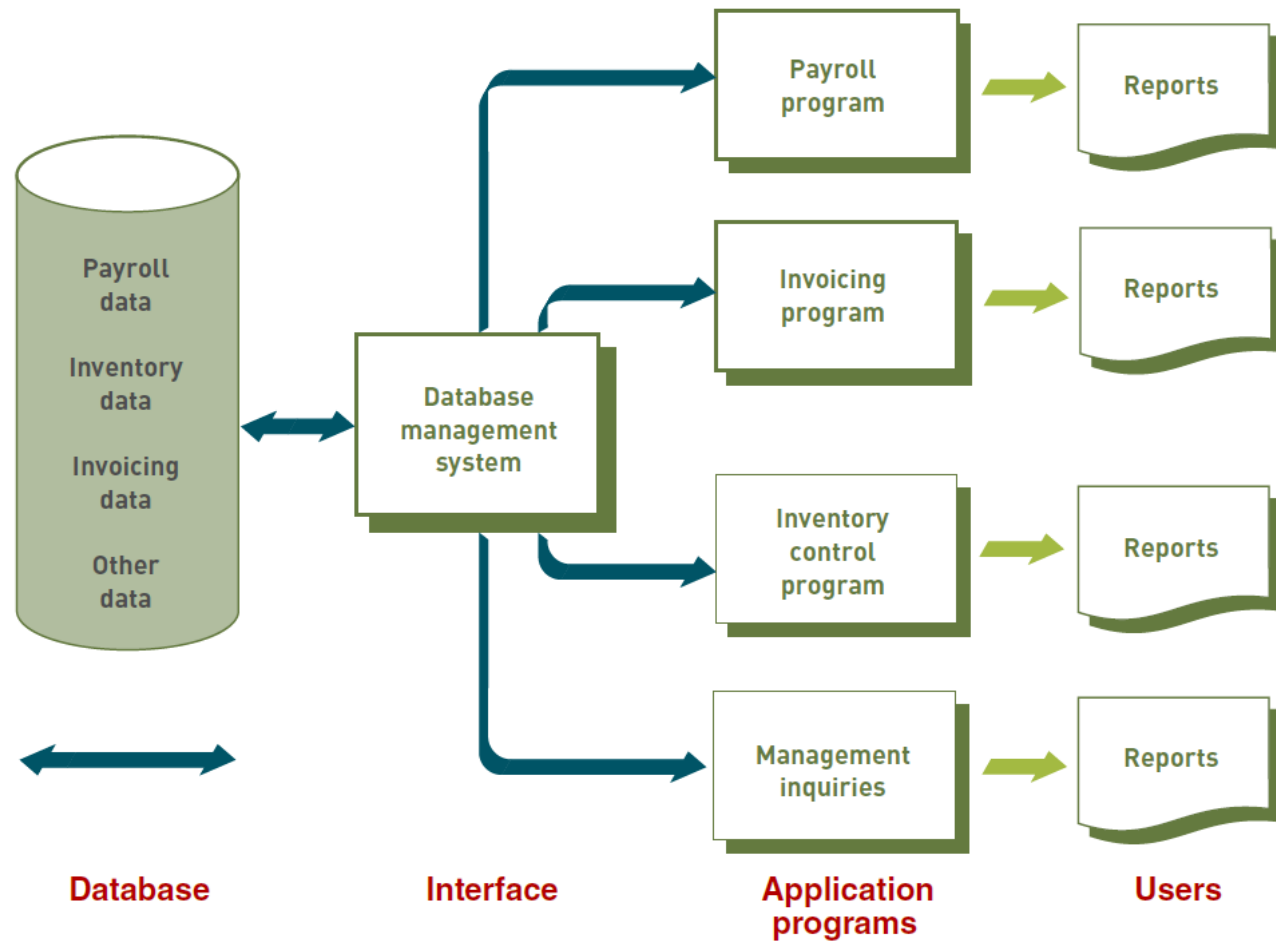


FIGURE 5.4

Database approach to data management

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



Database Management System

- ❑ DBMSs are becoming even more important to organizations as they deal with rapidly increasing amounts of information.
- ❑ Database management systems come in a wide variety of types and capabilities, ranging from small inexpensive software packages to sophisticated systems costing hundreds of thousands of dollars.



Advantages of The Database Approach

Advantages	Explanation
Improved strategic use of corporate data	Accurate, complete, up-to-date data can be made available to decision makers where, when, and in the form they need it. The database approach can also give greater visibility to the organization's data resources.
Reduced data redundancy	Data is organized by the DBMS and stored in only one location. This results in a more efficient use of system storage space.
Improved data integrity	With the traditional approach, some changes to data were not reflected in all copies of the data. The database approach prevents this problem because no separate files are maintained.
Easier modification and updating	The DBMS coordinates data modifications and updates. Programmers and users do not have to know where the data is physically stored. Data is stored and modified once. Modification and updating is also easier because the data is commonly stored in only one location.
Data and program independence	The DBMS organizes the data independently of the application program, so the application program is not affected by the location or type of data. Introduction of new data types not relevant to a particular application does not require rewriting that application to maintain compatibility with the data file.
Better access to data and information	Most DBMSs have software that makes it easy to access and retrieve data from a database. In most cases, users give simple commands to get important information. Relationships between records can be more easily investigated and exploited, and applications can be more easily combined.
Standardization of data access	A standardized, uniform approach to database access means that all application programs use the same overall procedures to retrieve data and information.
A framework for program development	Standardized database access procedures can mean more standardization of program development. Because programs go through the DBMS to gain access to data in the database, standardized database access can provide a consistent framework for program development. In addition, each application program need address only the DBMS, not the actual data files, reducing application development time.
Better overall protection of the data	Accessing and using centrally located data is easier to monitor and control. Security codes and passwords can ensure that only authorized people have access to particular data and information in the database, thus ensuring privacy.
Shared data and information resources	The cost of hardware, software, and personnel can be spread over many applications and users. This is a primary feature of a DBMS.



Disadvantages of The Database Approach

Disadvantages	Explanation
More complexity	DBMSs can be difficult to set up and operate. Many decisions must be made correctly for the DBMS to work effectively. In addition, users have to learn new procedures to take full advantage of a DBMS.
More difficult to recover from a failure	With the traditional approach to file management, a failure of a file affects only a single program. With a DBMS, a failure can shut down the entire database.
More expensive	DBMSs can be more expensive to purchase and operate than traditional file management. The expense includes the cost of the database and specialized personnel, such as a database administrator, who is needed to design and operate the database. Additional hardware might also be required.

Table 5.2

Disadvantages of the Database Approach



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 effectively.
- A database should be designed to store all data relevant to the business and to provide quick access and easy modification.



Data Modeling and Database Characteristics

- When building a database, an organization must consider:
 - **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

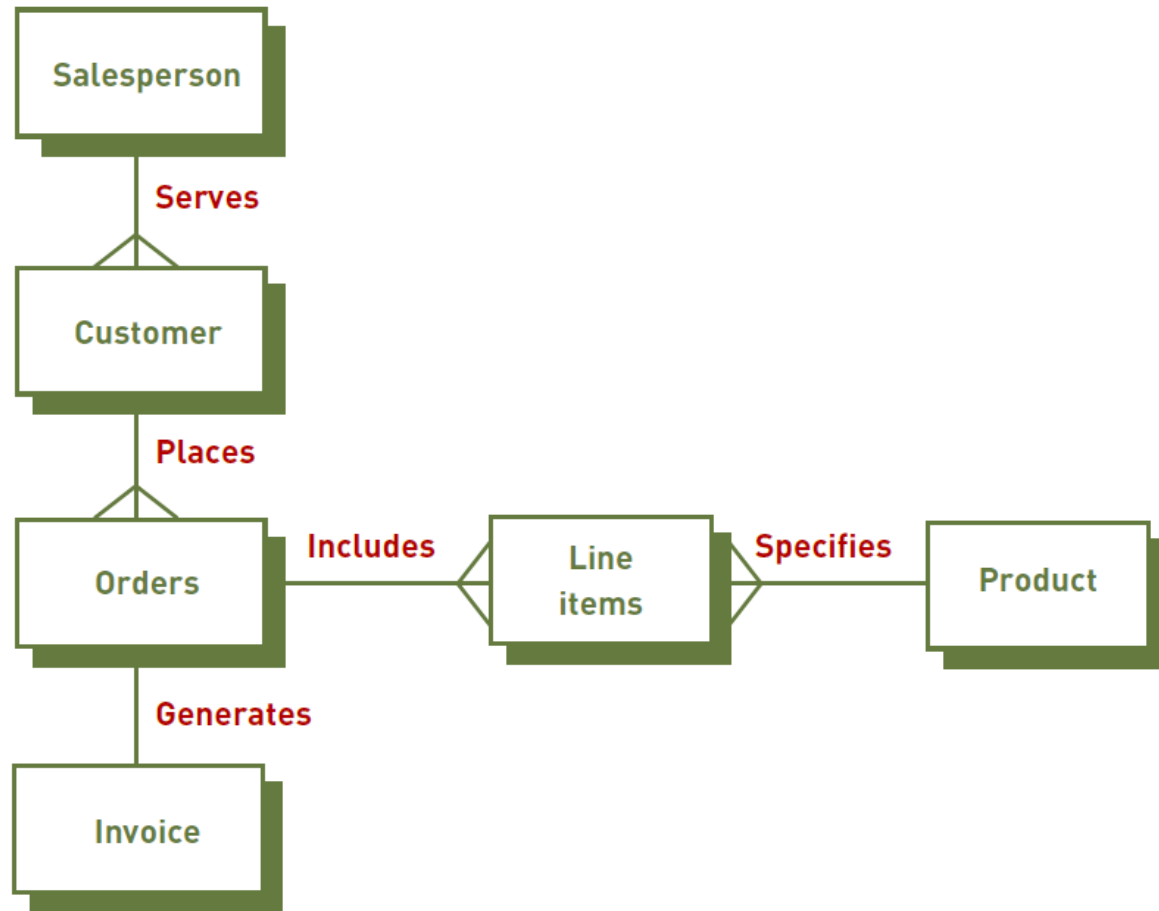
- Data model:
 - Diagram of data entities and their relationships.
- Enterprise data modeling:
 - Starts by investigating the general data and information needs of the organization at the strategic level. and then moves on to examine more specific data and information needs for the functional areas and departments within the organization.
- Entity-relationship (ER) diagrams:
 - Data models that use basic graphical symbols to show the organization of and relationships between data.



Entity-relationship Diagram

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

- **The relational database model:** A simple but highly useful way to organize data into collections of two-dimensional tables called **relations**.



Relational Database Model

- Relational model:
 - Describes data using a standard tabular format
 - Each row of a table represents an **entity** (record)
 - Each column represents an **attribute** (fields) of that entity.
 - **Domain**: The range of allowable values for a data attribute.

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

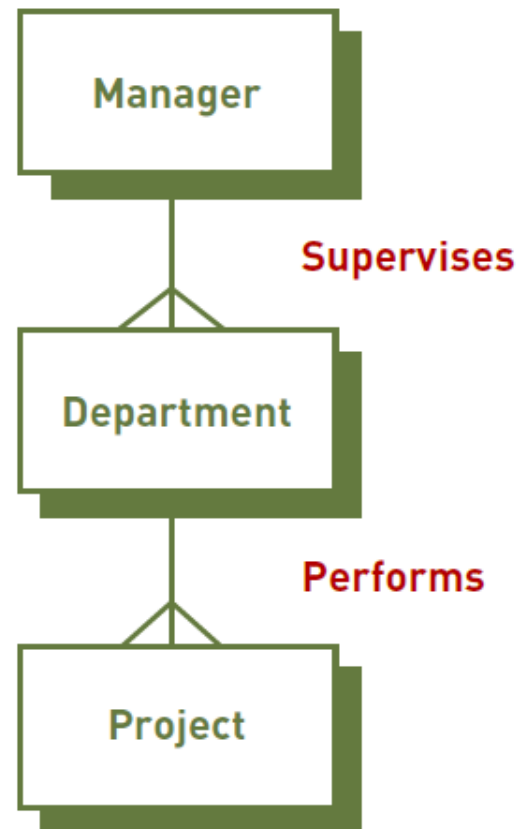
Relational database model

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.

FIGURE 5.8

Simplified ER diagram

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





Manipulating Data

- After entering data into a relational database, users can make inquiries and analyze the data.
- Basic data manipulations include **selecting**, **projecting**, **joining** and **linking**.



Manipulating Data

- ❑ **Selecting:**
 - ❑ Eliminates rows according to certain criteria.
- ❑ **Projecting:**
 - ❑ Eliminates columns in a table.
- ❑ **Joining:**
 - ❑ Combines two or more tables.
- ❑ **Linking:**
 - ❑ Combine two or more tables through common data attributes to form a new table with only the unique data attributes.

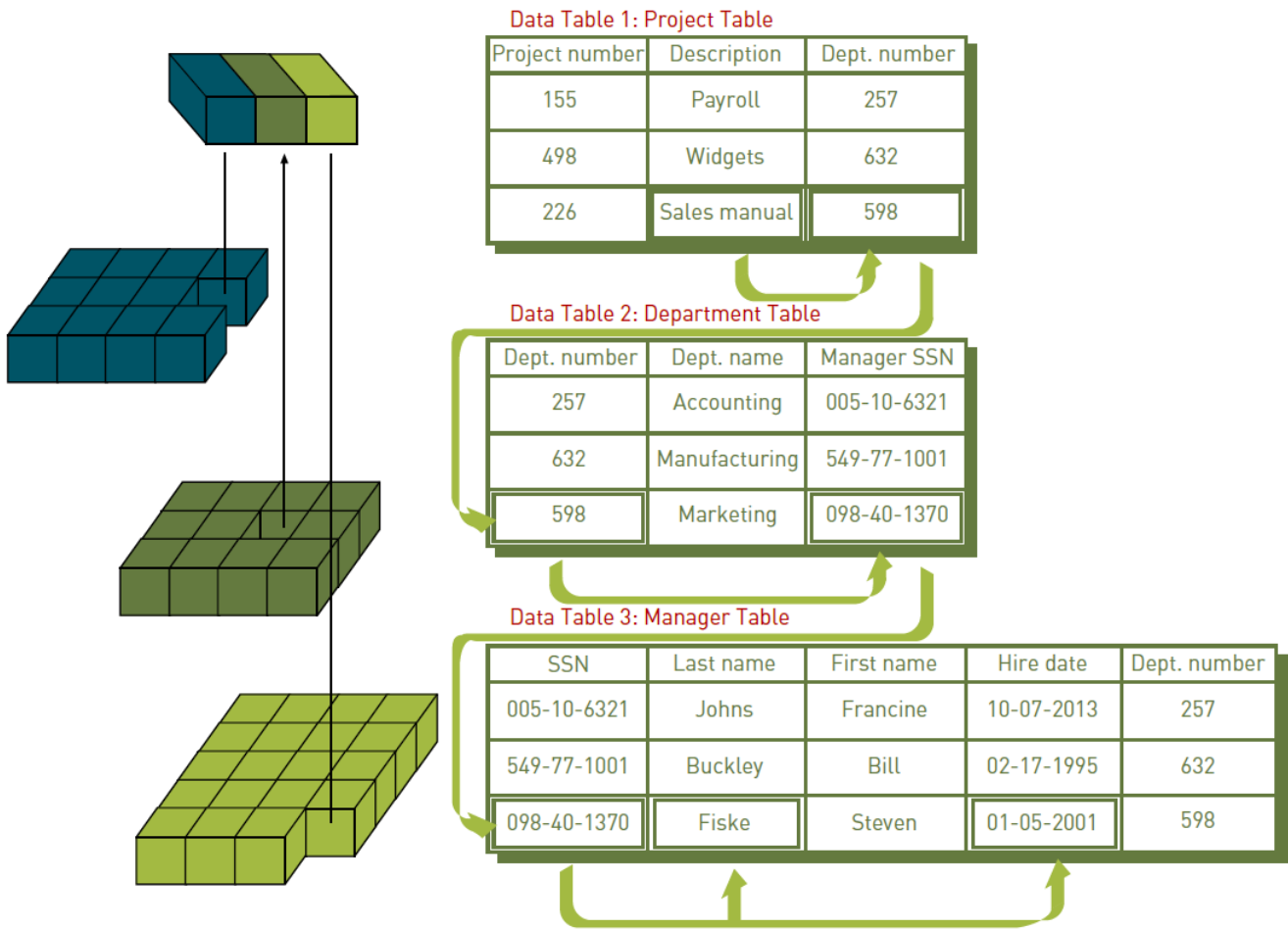


Linking

- As long as the tables share at least one common data attribute, the tables in a relational database can be linked to provide useful information.
- Suppose the president of a company wants to find out the name of the manager of the sales manual project as well as the length of time the manager has been with the company.

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.



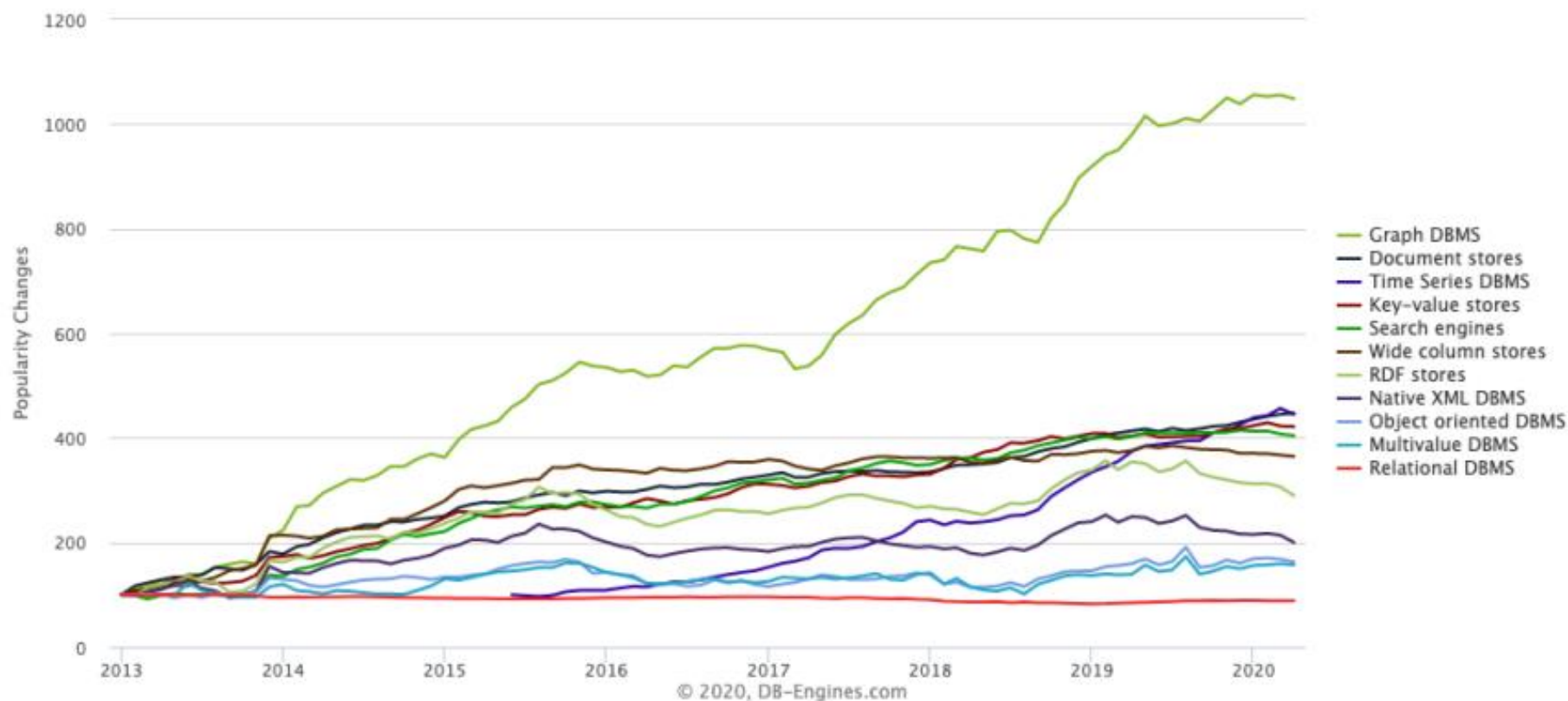


Relational Database Model

- ❑ Databases based on the relational model include Oracle, IBM DB2, Microsoft SQL Server, Microsoft Access, MySQL, and others.
- ❑ The relational database model has been an outstanding success and is dominant in the commercial world today, although many organizations are beginning to use new non-relational models to meet some of their business needs.



Databases Popularity Ranking





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.



Data Cleansing

- ❑ One data cleansing solution is to identify and correct data by crosschecking it against a validated data set.
- ❑ For example, street number, street name, city, state, and zip code entries in an organization's database may be cross-checked against the United States Postal Zip Code database.
- ❑ Data cleansing may also involve standardization of data, such as the conversion of various possible abbreviations (St., St, st., st) to one standard name (Street).



Case Study: Banco Popular

- ❑ Banco Popular is the largest bank in Puerto Rico. 3,000 bank employees in 200 branches use a customer database to obtain a complete view of 5.7 million personal and business accounts.
- ❑ The bank uses a data cleansing process to eliminate duplicate records by identifying how many account holders live at the same address to eliminate duplicate mailings to the same household, thus saving over \$840,000 in mailing expenses each year.



SQL Databases

- ❑ **Structured Query Language (SQL)** is a special-purpose programming language for accessing and manipulating data stored in a relational database.
- ❑ In 1986, the American National Standards Institute (ANSI) adopted SQL as the standard query language for relational databases.



SQL Databases

- Programmers and database users also find SQL valuable because SQL statements can be embedded into many programming languages, such as the widely used C++ and Java.
- Because SQL uses standardized and simplified procedures for retrieving, storing, and manipulating data, many programmers find it easy to understand and use—hence, it is popular.



Database Activities

- ❑ Providing a User View
- ❑ Creating and Modifying the Database
 - ❑ Data Definition Language (DDL) allows the database's creator to describe the data and relationships.
- ❑ Storing and Retrieving Data
- ❑ Manipulating Data and Generating Reports
 - ❑ Data Manipulation Language (DML) allows managers and other database users to access and modify the data, to make queries, and to generate reports.



Overview of Database Types

- ❑ Flat file
 - ❑ Simple database program whose records have no relationship to one another.
- ❑ Single user
 - ❑ Only one person can use the database at a time.
 - ❑ Examples: Access, FileMaker Pro, and InfoPath
- ❑ Multiple users
 - ❑ Allow dozens or hundreds of people to access the same database system at the same time
 - ❑ Examples: Oracle, Sybase, and IBM.



Creating and Modifying the Database

- Data definition language (DDL):
 - Collection of instructions and commands used to define and describe data and relationships in a specific database
- Data dictionary:
 - Detailed description of all the data used in the database.



Storing and Retrieving Data

- ❑ When an application program needs data:
 - ❑ It requests the data through the DBMS
- ❑ Concurrency control:
 - ❑ Method of dealing with a situation in which two or more users or applications need to access the same record at the same time



Manipulating Data and Generating Reports

- ❑ Data manipulation language (DML):
 - ❑ Commands that manipulate the data in a database
- ❑ Structured query language (SQL):
 - ❑ Adopted by the American National Standards Institute (ANSI) as the standard query language for relational databases
- ❑ Once a database has been set up and loaded with data:
 - ❑ It can produce reports, documents, and other outputs.



Database Administration

- ❑ Database administrators (DBAs):
 - ❑ skilled and trained IS professionals who hold discussions with business users to define their data needs;
 - ❑ apply database programming languages to craft a set of databases to meet those needs;
 - ❑ test and evaluate databases; implement changes to improve their performance;
 - ❑ and assure that data is secure from unauthorized access.



Database Administration

- Database administrators (DBAs):





Database Administration

- ❑ Database systems require a skilled database administrator, who must have a clear understanding of the fundamental business of the organization, be proficient in the use of selected database management systems, and stay abreast of emerging technologies and new design approaches.
- ❑ The role of the DBA is to plan, design, create, operate, secure, monitor, and maintain databases.



Database Administration

- Some organizations have also created a position called the **data administrator**.
- A **data administrator** is an individual responsible for defining and implementing consistent principles for a variety of data issues.



Database Administration

- ❑ For example, the data administrator would ensure that a term such as “customer” is defined and treated consistently in all corporate databases.
- ❑ The data administrator also works with business managers to identify who should have read or update access to certain databases. This information is then communicated to the database administrator for implementation.
- ❑ The data administrator can be a high-level position reporting to top-level managers.



Popular Database Management Systems

TABLE 5.2 Popular database management systems

Open-Source Relational DBMS	Relational DBMS for Individuals and Workgroups	Relational DBMS for Workgroups and Enterprise
MySQL	Microsoft Access	Oracle
PostgreSQL	IBM Lotus Approach	IBM DB2
MariaDB	Google Base	Sybase Adaptive Server
SQL Lite	OpenOffice Base	Teradata
CouchDB		Microsoft SQL Server
		Progress OpenEdge



Database as a Service (DaaS)

- With **DaaS**, the database is stored on a service provider's servers and accessed by the service subscriber over the Internet, with the database administration handled by the service provider.



Database as a Service (DaaS)

- ❑ More than a dozen companies are now offering DaaS services, including Amazon, Google, IBM, Microsoft, etc.
- ❑ Amazon Relational Database Service (Amazon RDS) is a DaaS that enables organizations to set up and operate their choice of a MySQL, Microsoft SQL, or PostgreSQL relational database in the cloud. The service automatically backs up the database and stores those backups based on a user-defined retention period.



Case Study: TinyCo

- TinyCo, a mobile gaming firm, stores its data on stored in the Amazon Relational Database Service (Amazon RDS) for MySQL. Which enables it to support the rapid growth in the number of its users without having to devote constant time and effort to organize and configure its information systems infrastructure.
- This arrangement has allowed the company to focus its resources on developing and marketing its new games.



Selecting a Database Management System

- ❑ Important characteristics of databases to consider:
 - ❑ Database size
 - ❑ Database cost
 - ❑ Concurrent users
 - ❑ Performance
 - ❑ Integration
 - ❑ Vendor



Using Databases with Other Software

- Database management systems are often used with other software and to interact with users over the Internet.
- A DBMS can act as a front-end application or a back-end applications:
 - Front-end applications interact directly with people.
 - Back-end applications interact with other programs or applications.



Data Warehouses, Data Marts, and Data Mining

- ❑ Data warehouse
 - ❑ Database that holds business information from many sources in the enterprise
- ❑ Data mart
 - ❑ Subset of a data warehouse
- ❑ Data mining
 - ❑ Information-analysis tool that involves the automated discovery of patterns and relationships in a data warehouse.



Data Warehouses, Data Marts, and Data Mining

- Predictive analysis:
 - Form of data mining that combines historical data with assumptions about future conditions to predict outcomes of events
 - Used by retailers to upgrade occasional customers into frequent purchasers
 - Software can be used to analyze a company's customer list and a year's worth of sales data to find new market segments.



Data Warehouses, Data Marts, and Data Mining

Application	Description
Branding and positioning of products and services	Enable the strategist to visualize the different positions of competitors in a given market using performance (or other) data on dozens of key features of the product and then to condense all that data into a perceptual map of only two or three dimensions.
Customer churn	Predict current customers who are likely to switch to a competitor.
Direct marketing	Identify prospects most likely to respond to a direct marketing campaign (such as a direct mailing).
Fraud detection	Highlight transactions most likely to be deceptive or illegal.
Market basket analysis	Identify products and services that are most commonly purchased at the same time (e.g., nail polish and lipstick).
Market segmentation	Group customers based on who they are or on what they prefer.
Trend analysis	Analyze how key variables (e.g., sales, spending, promotions) vary over time.

Table 5.8

Common Data-Mining Applications



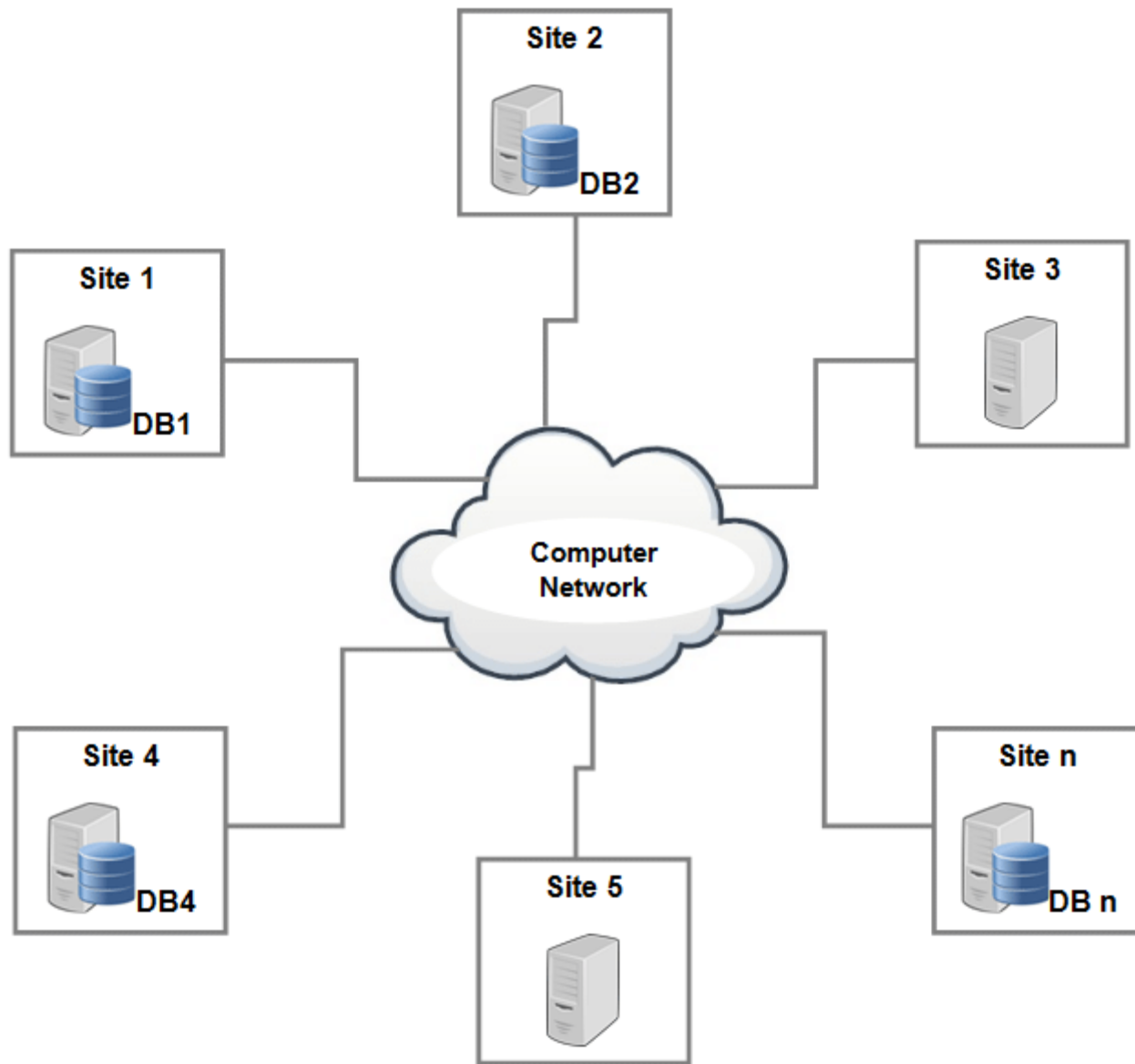
Business Intelligence

- Involves gathering enough of the right information:
 - In a timely manner and usable form and analyzing it to have a positive impact on business strategy, tactics, or operations



Distributed Databases

- Distributed database:
 - Database in which the data may be spread across several smaller databases connected via telecommunications devices
 - Gives corporations more flexibility in how databases are organized and used.





Object-oriented and Object-relational Database Management Systems

- ❑ Object-oriented database:
 - ❑ Uses an object-oriented database management system (OODBMS) to provide a user interface and connections to other programs
- ❑ Object-relational database management system (ORDBMS)
 - ❑ Provides the ability for third parties to add new data types and operations to the database.



Characteristics of Big Data

- ❑ Computer technology analysts associated the three characteristics of **volume**, **velocity**, and **variety** with big data.



Characteristics of Big Data: Volume

- ❑ **Volume** refers to the size of the data sets that need to be analyzed and processed, which are now frequently larger than terabytes and petabytes.
- ❑ The data sets in Big Data are too large to process with a regular laptop or desktop processor.
- ❑ An example of a high-volume data set would be all credit card transactions on a day within Europe.



Characteristics of Big Data:

Velocity

- ❑ **Velocity** refers to the speed with which data is generated.
 - ❑ An example of a data that is generated with high velocity would be Twitter messages or Facebook posts.
- ❑ The velocity at which data is currently coming at us exceeds 5 trillion bits per second.
- ❑ This rate is accelerating rapidly, and the volume of digital data is expected to double every two years.



Characteristics of Big Data:

Variety

Data today comes in a variety of formats. **Structured data** is a kind of data whose format is known in advance, and it fits nicely into traditional databases.

- For example, the data generated by the well-defined business transactions that are used to update many corporate databases containing customer, product, inventory, financial, and employee data is generally structured data.



Characteristics of Big Data: Variety

However, most of the data that an organization must deal with is **unstructured data**, meaning that it is not organized in any predefined manner.

- **Unstructured data** comes from sources such as word-processing documents, social media, email, photos, CCTV audio/video files, and phone messages.



Sources of Big Data

- ❑ Organizations collect and use data from a variety of sources, including business applications, social media, public sources (such as government Web sites), and archives of historical records of transactions and communications, etc.

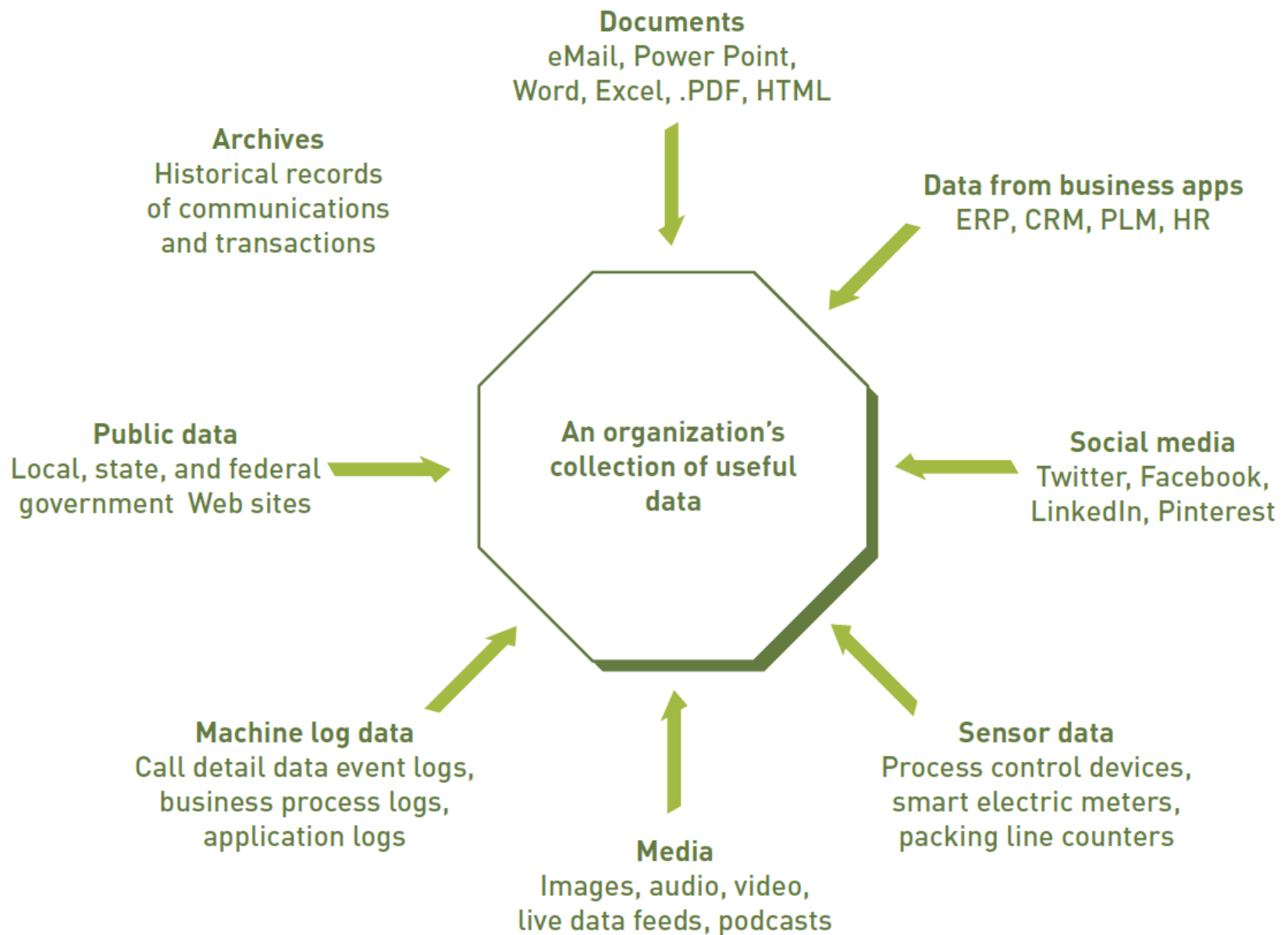


FIGURE 5.20

Sources of an organization's useful data

An organization has many sources of useful data.

TABLE 5.3 Portals that provide access to free sources of useful big data

Data Source	Description	URL
Amazon Web Services (AWS) public data sets	Portal to a huge repository of public data, including climate data, the million song data set, and data from the 1000 Genomes project.	http://aws.amazon.com/datasets
Bureau of Labor Statistics (BLS)	Provides access to data on inflation and prices, wages and benefits, employment, spending and time use, productivity, and workplace injuries	www.bls.gov
CIA World Factbook	Portal to information on the economy, government, history, infrastructure, military, and population of 267 countries	https://cia.gov/library/publications/the-world-factbook
Data.gov	Portal providing access to over 186,000 government data sets, related to topics such as agriculture, education, health, and public safety	http://data.gov
Facebook Graph	Provides a means to query Facebook profile data not classified as private	https://developers.facebook.com/docs/graph-api
FBI Uniform Crime Reports	Portal to data on Crime in the United States, Law Enforcement Officers Killed and Assaulted, and Hate Crime Statistics	https://www.fbi.gov/about-us/cjis/ucr/ucr/
Justia Federal District Court Opinions and Orders database	A free searchable database of full-text opinions and orders from civil cases heard in U.S. Federal District Courts	http://law.justia.com/cases/federal/district-courts/
Gapminder	Portal to data from the World Health Organization and World Bank on economic, medical, and social issues	www.gapminder.org/data



Big Data Uses

Here are a few examples of how organizations are employing big data to improve their day-to-day operations, planning, and decision making:

- Hospitals analyze medical data and patient records to identify patients who likely need readmission within a few months of discharge, with the goal of preventing another expensive hospital stay.



Big Data Uses

- ❑ 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.



Data Lifecycle Management

- 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.

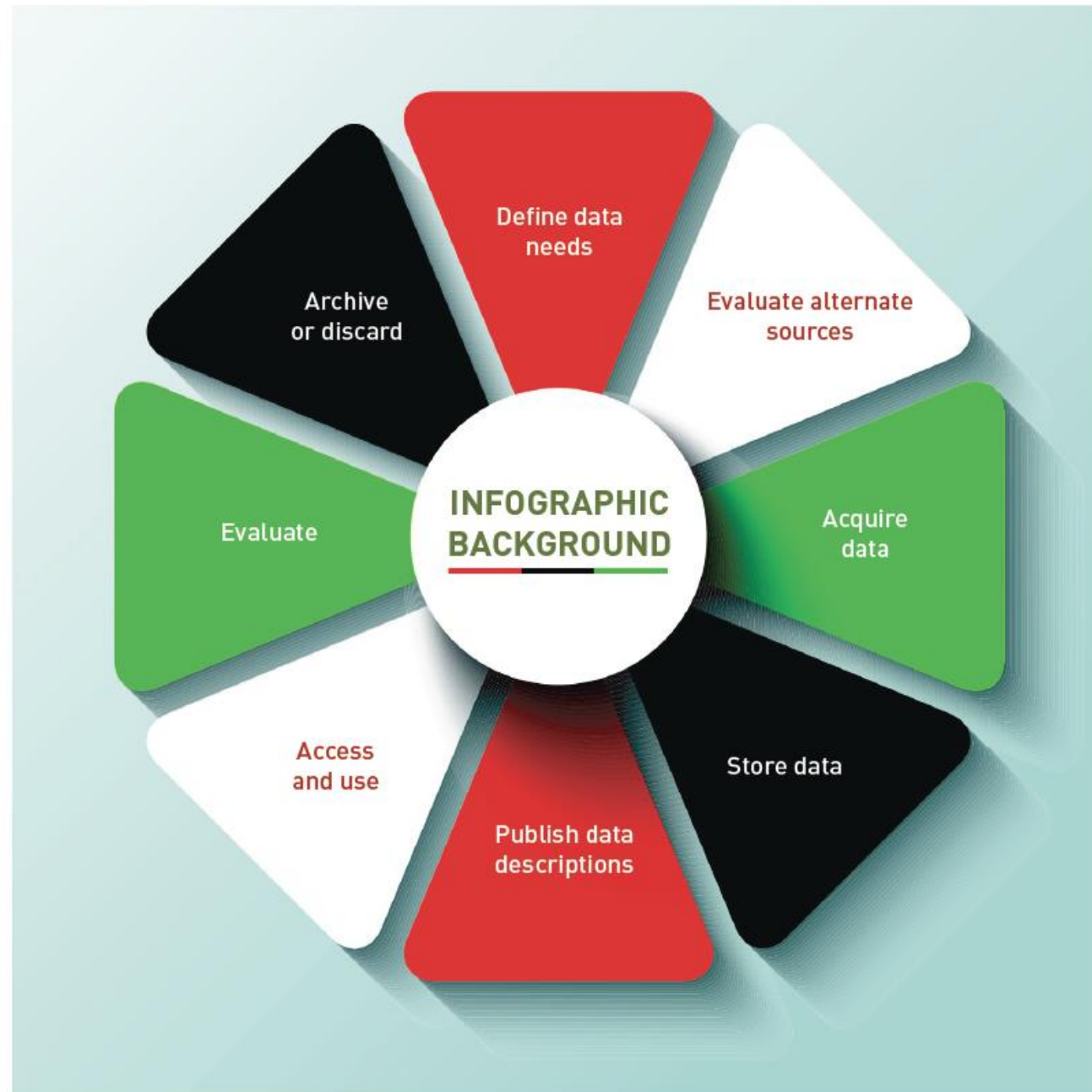


FIGURE 5.22

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 Lakes

- ❑ A traditional data warehouse is created by extracting (and discarding some data in the process), transforming (modifying), and loading incoming data for predetermined and specific analyses and applications.
- ❑ This process can be lengthy and computer intensive, taking days to complete.



Data Lakes

- ❑ 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.



Data Lakes

- ❑ 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. Its contents can include business transactions, clickstream, sensor data, server logs, social media, videos, and more.



NoSQL Databases

- ❑ A 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.
- ❑ Such databases are being used to deal with the variety of data found in big data and Web applications.



NoSQL Databases

- A major advantage of NoSQL databases is the ability to spread data over multiple servers so that each server contains only a subset of the total data. This horizontal scaling capability enables hundreds or even thousands of servers to operate on the data, providing faster response times for queries and updates.
- Most RDBMSs have problems with such horizontal scaling and instead require large, powerful, and expensive proprietary servers and large storage systems.



NoSQL Databases

- ❑ Often, the data structures used by NoSQL databases are more flexible than relational database tables and, in many cases, they can provide improved access speed and redundancy.



NoSQL Databases

The four main categories of 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.



NoSQL Databases

- 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.



NoSQL Databases

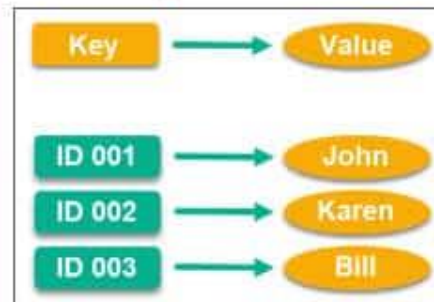
SQL Databases

Table

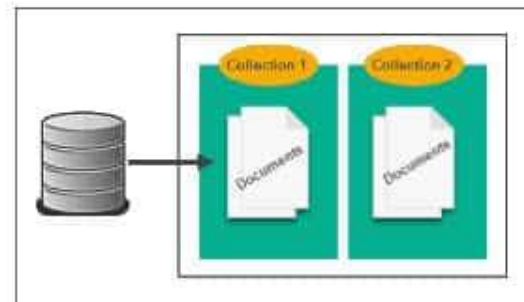
ID	Name	Grade	GPA
001	John	Senior	4.00
002	Karen	Freshman	3.67
003	Bill	Junior	3.33

NoSQL Databases

Key-value



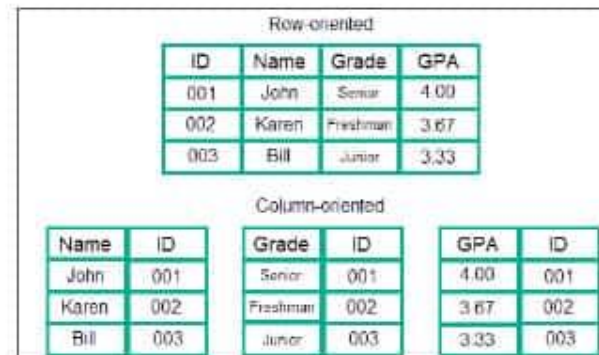
Document



Graph



Wide-column





NoSQL Databases

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



In-Memory Databases

- An in-memory database (IMDB) is a database management system that stores the entire database in random access memory (RAM).
- This approach provides access to data at rates much faster than storing data on some form of secondary storage (e.g., a hard drive or flash drive) as is done with traditional database management systems.



In-Memory Databases

- ❑ Of course, we can only use an in-memory database in applications and scenarios where data does not need to be persisted or for the purpose of executing tests faster.
- ❑ These databases are created when a process starts and discarded when the process ends.



In-Memory Databases

- ❑ IMDBs enable the analysis of big data and other challenging data-processing applications, and they have become feasible because of the increase in RAM capacities and a corresponding decrease in RAM costs.
- ❑ In-memory databases perform best on multiple multicore CPUs that can process parallel requests to the data, further speeding access to and processing of large amounts of data.



In-Memory Databases

TABLE 5.6 IMDB providers

Database Software Manufacturer	Product Name	Major Customers
Altibase	HDB	E*Trade, China Telecom
Oracle	Times Ten	Lockheed Martin, Verizon Wireless
SAP	High-Performance Analytic Appliance (HANA)	eBay, Colgate
Software AG	Terracotta Big Memory	AdJuggler



References

- ▣ **Reynolds, George Walter, Stair, Ralph M.**
“Principles of information systems”, 13e – 2017