CHAPTER

6 Foundations of Business Intelligence: Databases and Information Management

LEARNING OBJECTIVES

After reading this chapter, you will be able to answer the following questions:

6-1 What are the problems of managing data resources in a traditional file environment?

6-2 What are the major capabilities of database management systems (DBMS), and why is a relational DBMS so powerful?

6-3 What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?

6-4 Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?

6-5 How will MIS help my career?

CHAPTER CASES

Data Management Helps the Charlotte Hornets Learn More About Their Fans

Kraft Heinz Finds a New Recipe for Analyzing Its Data

Databases Where the Data Aren't There

How Reliable Is Big Data?

VIDEO CASES

Dubuque Uses Cloud Computing and Sensors to Build a Smarter City

Brooks Brothers Closes In on Omnichannel Retail

Maruti Suzuki Business Intelligence and Enterprise Databases

MyLab MIS

Discussion Questions: 6-5, 6-6, 6-7; **Hands-on MIS Projects** 6-8, 6-9, 6-10, 6-11; **Writing Assignments:** 6-17, 6-18; **eText with Conceptual Animations**

Data Management Helps the Charlotte Hornets Learn More About Their Fans

The NBA's Charlotte Hornets have millions of fans, but until recently they didn't know very much about them. The Charlotte, North Carolina–based basketball team had many millions of records of fan data—online ticket and team gear purchases, food and beverage purchases at games, and comments about the team on social media. Every time a fan performs one of these actions, more data about that fan are created. Three million records of food and beverage purchase transactions are generated during each Hornets game. There was too much unorganized customer data for decision makers to digest.

© Oleksii Sidorov/Shutterstock

All of this accumulating data, which came from many different sources, started to overtax the team's Microsoft Dynamics customer relationship management system. There were 12 to 15 different sources of data on Hornets fan behavior and they were maintained in separate data repositories that could not communicate with each other. It became increasingly difficult for the Hornets to understand their fans and how they were interacting with the organization.

Five years ago, Hornets management decided to improve its approach to data management. The team needed technology that could easily maintain data from many different sources and 12 different vendors and it needed to be able to combine and integrate what amounted to 12 different profiles on each fan into a single profile. This would enable the Hornets to understand each fan's behavior in much greater detail and offer them a more personalized experience.

Under the leadership of Chris Zeppenfeld, the Hornets' senior director of business intelligence, the team implemented a data warehouse that would consolidate all of the Hornets' customer data from its various data sources in a single location where the data could be easily accessed and analyzed by business users. The warehouse was based on a SAP HANA database optimized to process very large quantities of data at ultra-high speed and included Phizzle Fan Tracker™ software to cleanse, streamline, and combine millions of fan records to create a single profile for each Hornets fan. Phizzle Fan Tracker is a fan engagement platform designed to consolidate, analyze, and act on multiple data sources. The platform's data aggregation capabilities, innovative data visualization tools, and social listening solutions provide sports properties and brands the capability to gather and analyze digital, social, and real-world fan engagements. Fan Tracker works with the SAP HANA database to consolidate

customer profiles, analyze and act on real-time online behavior, and consolidate all existing data sources to uniquely identify fan records. The solution provides a unified overview and deeper understanding of each fan, allowing clubs to offer their fans a more personalized experience.

By using Fan Tracker and a unified data warehouse, the Hornets have compiled and synthesized 25 million fan and consumer interactions, saving over $1.5 million in consulting expenses. They now have a real-time data profile for every one of their 1.5 million fans, which includes up-to-the minute behavioral data on each fan from third-party applications as well as the Hornets' own sources. Each profile reveals detailed insights into a fan's behavior including sentiment, purchase history, interactions, and fan value across multiple points of contact. Zeppenfeld believes that better fan data management has helped the team rank among the top five NBA franchises for new full season ticket sales each year.

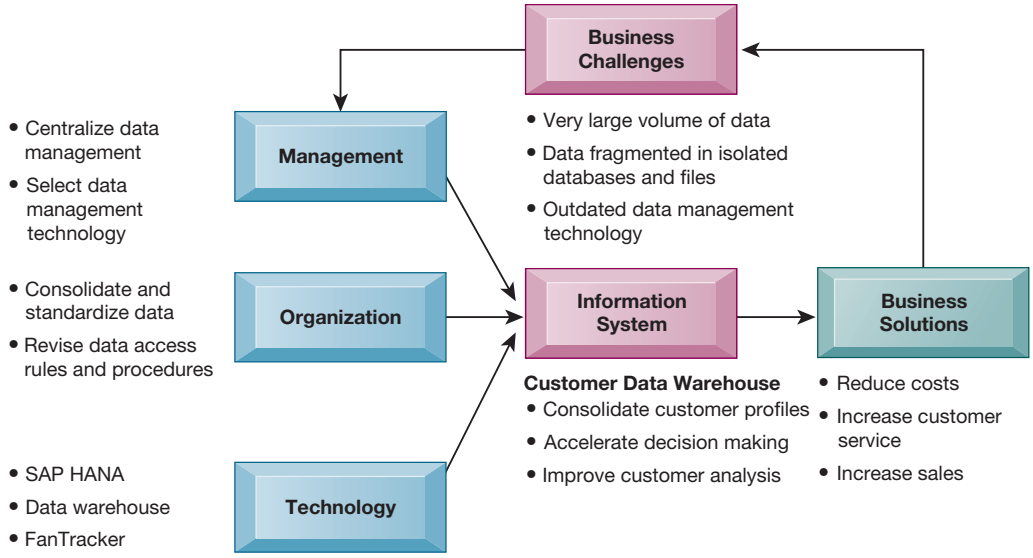
Sources: Jim O'Donnell, "Charlotte Hornets Use Phizzle Built on HANA to Analyze Fan Behavior," SearchSAPtechtarget.com, February 11, 2018; "NBA Team Charlotte Hornets/SAP Case Study," www.phizzle.com, accessed February 12, 2018; and Mark J. Burns, "Why The Charlotte Hornets Are Using Phizzle To Streamline Their Data Warehouse," *Sport Techie*, September 2016.

The experience of the Charlotte Hornets illustrates the importance of data management. Business performance depends on what a firm can or cannot do with its data. The Charlotte Hornets NBA basketball team was a thriving business, but both operational efficiency and management decision making were hampered by fragmented data stored in multiple locations that were difficult to access and analyze. How businesses store, organize, and manage their data has an enormous impact on organizational effectiveness.

The chapter-opening diagram calls attention to important points raised by this case and this chapter. The Charlotte Hornets had accumulated very large quantities of fan data from many different sources. Marketing campaigns and personalized offers to fans were not as effective as they could have been because it was so difficult to assemble and understand the data required to obtain a detailed understanding of each customer. The solution was to combine the Hornets' customer data from all sources in a data warehouse that provided a single source of data for reporting and analysis and use Fan Tracker software to consolidate disparate pieces of customer data into a single profile for each customer. The Hornets had to reorganize their data into a standard company-wide format; establish rules, responsibilities, and procedures for accessing and using the data; and provide tools for making the data accessible to users for querying and reporting.

The data warehouse integrated company data from all of its disparate sources into a single comprehensive database that could be queried directly. The data were reconciled to prevent multiple profiles on the same customer. The solution improved customer marketing, sales, and service while reducing costs. The Hornets increased their ability to quickly analyze very large quantities of data by using SAP HANA high-speed database technology.

The data warehouse boosted operational efficiency and decision making by making more comprehensive and accurate customer data available and by



making it easier to access all the business's data on each customer. By helping the Hornets understand their own customers better, the solution increased opportunities for selling to customers as well as the effectiveness of marketing and sales campaigns.

Here are some questions to think about: What was the business impact of the Hornets' data management problems? How did better use of the Hornets' customer data improve operational efficiency and management decision making?

6-1 What are the problems of managing data resources in a traditional file environment?

An effective information system provides users with accurate, timely, and relevant information. Accurate information is free of errors. Information is timely when it is available to decision makers when it is needed. Information is relevant when it is useful and appropriate for the types of work and decisions that require it.

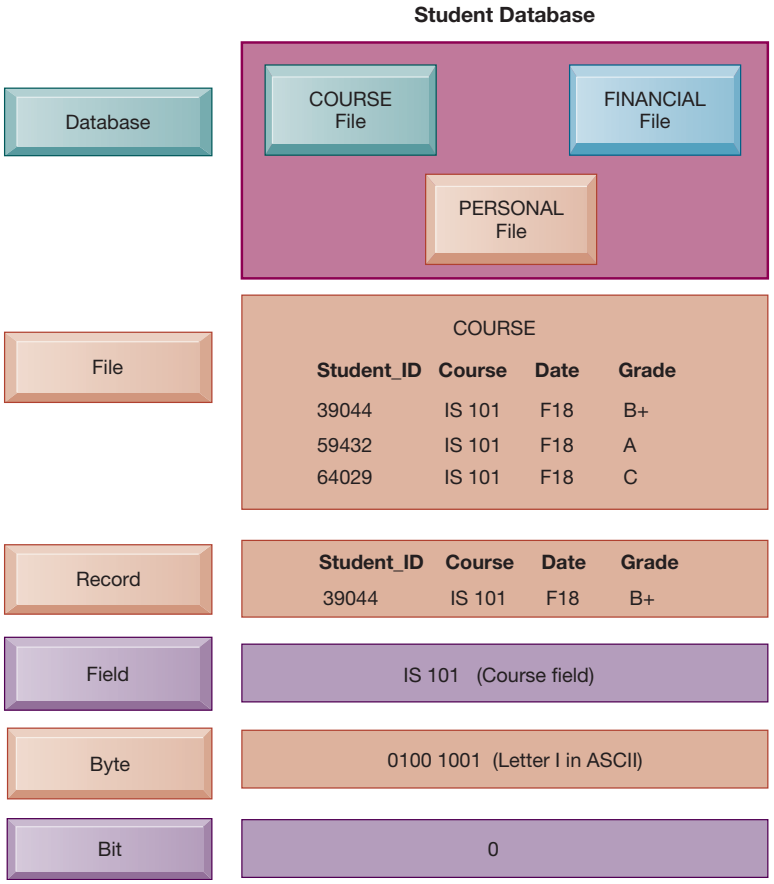
You might be surprised to learn that many businesses don't have timely, accurate, or relevant information because the data in their information systems have been poorly organized and maintained. That's why data management is so essential. To understand the problem, let's look at how information systems arrange data in computer files and traditional methods of file management.

File Organization Terms and Concepts

A computer system organizes data in a hierarchy that starts with bits and bytes and progresses to fields, records, files, and databases (see Figure 6.1). A **bit** represents the smallest unit of data a computer can handle. A group of bits, called a **byte**, represents a single character, which can be a letter, a number, or another symbol. A grouping of characters into a word, a group of words, or a complete number (such as a person's name or age) is called a **field**. A group of related fields, such as the student's name, the course taken, the date, and the grade, comprises a **record**; a group of records of the same type is called a **file**.

FIGURE 6.1 THE DATA HIERARCHY

A computer system organizes data in a hierarchy that starts with the bit, which represents either a 0 or a 1. Bits can be grouped to form a byte to represent one character, number, or symbol. Bytes can be grouped to form a field, and related fields can be grouped to form a record. Related records can be collected to form a file, and related files can be organized into a database.



For example, the records in Figure 6.1 could constitute a student course file. A group of related files makes up a database. The student course file illustrated in Figure 6.1 could be grouped with files on students' personal histories and financial backgrounds to create a student database.

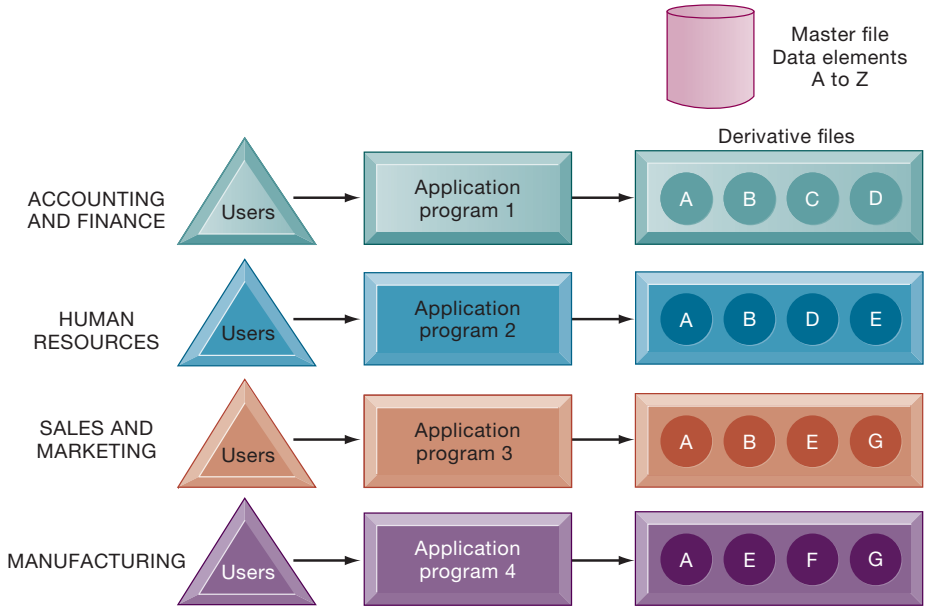
A record describes an entity. An **entity** is a person, place, thing, or event on which we store and maintain information. Each characteristic or quality describing a particular entity is called an **attribute**. For example, Student_ID, Course, Date, and Grade are attributes of the entity COURSE. The specific values that these attributes can have are found in the fields of the record describing the entity COURSE.

Problems with the Traditional File Environment

In most organizations, systems tended to grow independently without a companywide plan. Accounting, finance, manufacturing, human resources, and sales and marketing all developed their own systems and data files. Figure 6.2 illustrates the traditional approach to information processing.

FIGURE 6.2 TRADITIONAL FILE PROCESSING

The use of a traditional approach to file processing encourages each functional area in a corporation to develop specialized applications. Each application requires a unique data file that is likely to be a subset of the master file. These subsets of the master file lead to data redundancy and inconsistency, processing inflexibility, and wasted storage resources.



Each application, of course, required its own files and its own computer program to operate. For example, the human resources functional area might have a personnel master file, a payroll file, a medical insurance file, a pension file, a mailing list file, and so forth, until tens, perhaps hundreds, of files and programs existed. In the company as a whole, this process led to multiple master files created, maintained, and operated by separate divisions or departments. As this process goes on for 5 or 10 years, the organization is saddled with hundreds of programs and applications that are very difficult to maintain and manage. The resulting problems are data redundancy and inconsistency, program-data dependence, inflexibility, poor data security, and an inability to share data among applications.

Data Redundancy and Inconsistency

Data redundancy is the presence of duplicate data in multiple data files so that the same data are stored in more than one place or location. Data redundancy occurs when different groups in an organization independently collect the same piece of data and store it independently of each other. Data redundancy wastes storage resources and also leads to **data inconsistency**, where the same attribute may have different values. For example, in instances of the entity COURSE illustrated in Figure 6.1, the Date may be updated in some systems but not in others. The same attribute, Student_ID, might also have different names in different systems throughout the organization. Some systems might use Student_ID and others might use ID, for example.

Additional confusion can result from using different coding systems to represent values for an attribute. For instance, the sales, inventory, and manufacturing systems of a clothing retailer might use different codes to represent clothing size.

One system might represent clothing size as "extra large," whereas another might use the code "XL" for the same purpose. The resulting confusion would make it difficult for companies to create customer relationship management, supply chain management, or enterprise systems that integrate data from different sources.

Program-Data Dependence

Program-data dependence refers to the coupling of data stored in files and the specific programs required to update and maintain those files such that changes in programs require changes to the data. Every traditional computer program has to describe the location and nature of the data with which it works. In a traditional file environment, any change in a software program could require a change in the data accessed by that program. One program might be modified from a five-digit to a nine-digit ZIP code. If the original data file were changed from five-digit to nine-digit ZIP codes, then other programs that required the five-digit ZIP code would no longer work properly. Such changes could cost millions of dollars to implement properly.

Lack of Flexibility

A traditional file system can deliver routine scheduled reports after extensive programming efforts, but it cannot deliver ad hoc reports or respond to unanticipated information requirements in a timely fashion. The information required by ad hoc requests is somewhere in the system but may be too expensive to retrieve. Several programmers might have to work for weeks to put together the required data items in a new file.

Poor Security

Because there is little control or management of data, access to and dissemination of information may be out of control. Management might have no way of knowing who is accessing or even making changes to the organization's data.

Lack of Data Sharing and Availability

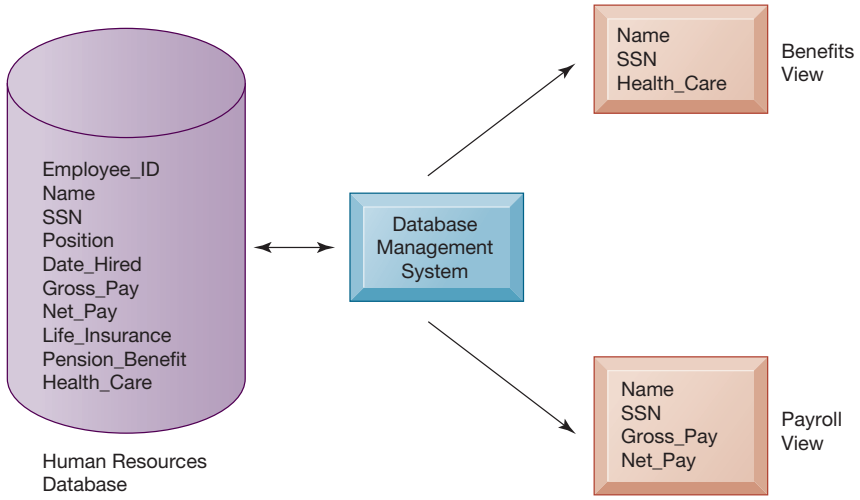
Because pieces of information in different files and different parts of the organization cannot be related to one another, it is virtually impossible for information to be shared or accessed in a timely manner. Information cannot flow freely across different functional areas or different parts of the organization. If users find different values for the same piece of information in two different systems, they may not want to use these systems because they cannot trust the accuracy of their data.

6-2 What are the major capabilities of database management systems (DBMS), and why is a relational DBMS so powerful?

Database technology cuts through many of the problems of traditional file organization. A more rigorous definition of a **database** is a collection of data organized to serve many applications efficiently by centralizing the data and controlling redundant data. Rather than storing data in separate files for each application, data appear to users as being stored in only one location. A single database services multiple applications. For example, instead of a corporation storing employee data in separate information systems and separate files for personnel, payroll, and benefits, the corporation could create a single common human resources database (see Figure 6.3).

FIGURE 6.3 HUMAN RESOURCES DATABASE WITH MULTIPLE VIEWS

A single human resources database provides many different views of data, depending on the information requirements of the user. Illustrated here are two possible views, one of interest to a benefits specialist and one of interest to a member of the company's payroll department.



Database Management Systems

A **database management system (DBMS)** is software that enables an organization to centralize data, manage them efficiently, and provide access to the stored data by application programs. The DBMS acts as an interface between application programs and the physical data files. When the application program calls for a data item, such as gross pay, the DBMS finds this item in the database and presents it to the application program. Using traditional data files, the programmer would have to specify the size and format of each data element used in the program and then tell the computer where they were located.

The DBMS relieves the programmer or end user from the task of understanding where and how the data are actually stored by separating the logical and physical views of the data. The *logical view* presents data, as they would be perceived by end users or business specialists, whereas the *physical view* shows how data are actually organized and structured on physical storage media.

The database management software makes the physical database available for different logical views required by users. For example, for the human resources database illustrated in Figure 6.3, a benefits specialist might require a view consisting of the employee's name, social security number, and health insurance coverage. A payroll department member might need data such as the employee's name, social security number, gross pay, and net pay. The data for all these views are stored in a single database, where they can be more easily managed by the organization.

How a DBMS Solves the Problems of the Traditional File Environment

A DBMS reduces data redundancy and inconsistency by minimizing isolated files in which the same data are repeated. The DBMS may not enable the organization to eliminate data redundancy entirely, but it can help control redundancy. Even if the organization maintains some redundant data, using a DBMS eliminates data

inconsistency because the DBMS can help the organization ensure that every occurrence of redundant data has the same values. The DBMS uncouples programs and data, enabling data to stand on their own. The description of the data used by the program does not have to be specified in detail each time a different program is written. Access and availability of information will be increased and program development and maintenance costs reduced because users and programmers can perform ad hoc queries of the database for many simple applications without having to write complicated programs. The DBMS enables the organization to centrally manage data, their use, and security. Data sharing throughout the organization is easier because the data are presented to users as being in a single location rather than fragmented in many different systems and files.

Relational DBMS

Contemporary DBMS use different database models to keep track of entities, attributes, and relationships. The most popular type of DBMS today for PCs as well as for larger computers and mainframes is the **relational DBMS**. Relational databases represent data as two-dimensional tables (called relations). Tables may be referred to as files. Each table contains data on an entity and its attributes. Microsoft Access is a relational DBMS for desktop systems, whereas DB2, Oracle Database, and Microsoft SQL Server are relational DBMS for large mainframes and midrange computers. MySQL is a popular open source DBMS.

Let's look at how a relational database organizes data about suppliers and parts (see Figure 6.4). The database has a separate table for the entity SUPPLIER and a table for the entity PART. Each table consists of a grid of columns and rows of data. Each individual element of data for each entity is stored as a separate field, and each field represents an attribute for that entity. Fields in a relational database are also called columns. For the entity SUPPLIER, the supplier identification number, name, street, city, state, and ZIP code are stored as separate fields within the SUPPLIER table and each field represents an attribute for the entity SUPPLIER.

The actual information about a single supplier that resides in a table is called a row. Rows are commonly referred to as records, or in very technical terms, as **tuples**. Data for the entity PART have their own separate table.

The field for Supplier_Number in the SUPPLIER table uniquely identifies each record so that the record can be retrieved, updated, or sorted. It is called a **key field**. Each table in a relational database has one field that is designated as its **primary key**. This key field is the unique identifier for all the information in any row of the table and this primary key cannot be duplicated. Supplier_Number is the primary key for the SUPPLIER table and Part_Number is the primary key for the PART table. Note that Supplier_Number appears in both the SUPPLIER and PART tables. In the SUPPLIER table, Supplier_Number is the primary key. When the field Supplier_Number appears in the PART table, it is called a **foreign key** and is essentially a lookup field to look up data about the supplier of a specific part.

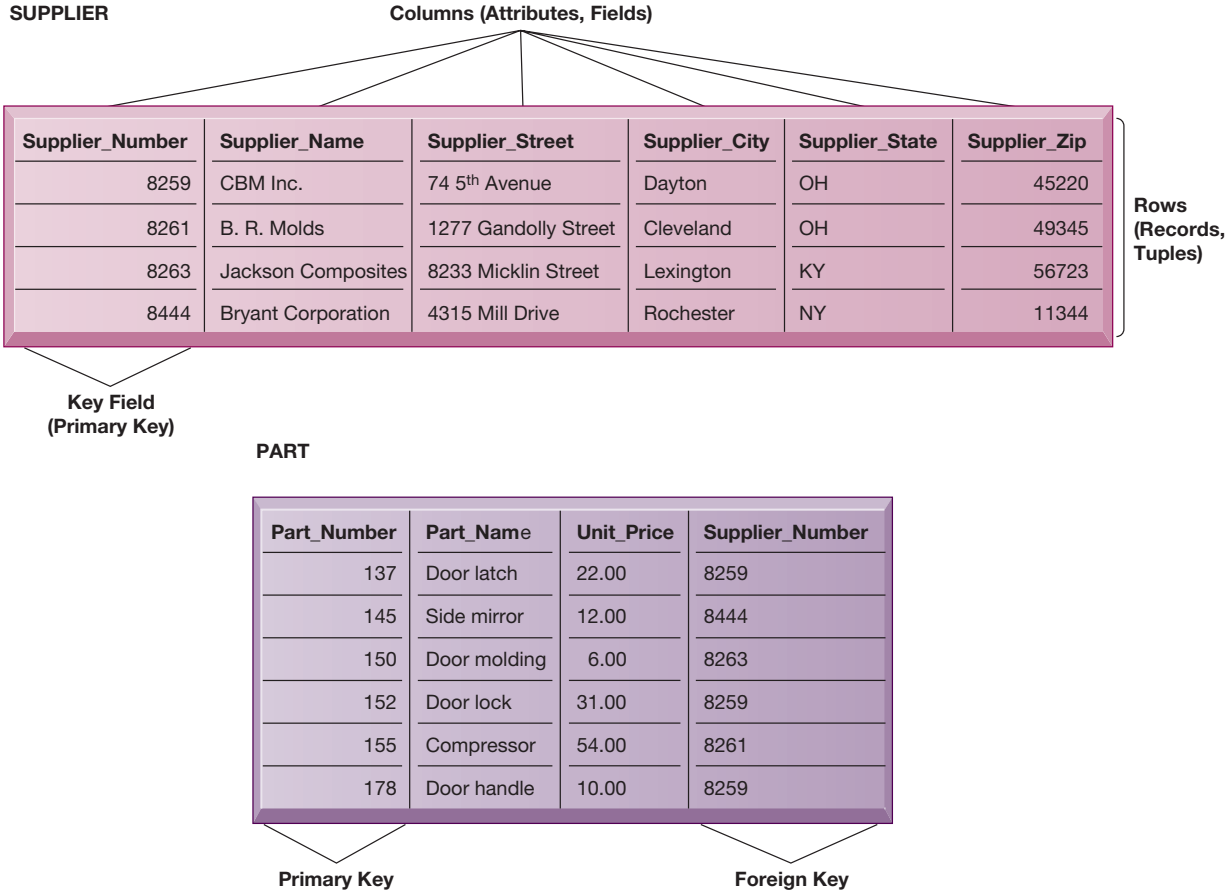
Operations of a Relational DBMS

Relational database tables can be combined easily to deliver data required by users, provided that any two tables share a common data element. Suppose we wanted to find in this database the names of suppliers who could provide us with part number 137 or part number 150. We would need information from two tables: the SUPPLIER table and the PART table. Note that these two files have a shared data element: Supplier_Number.

In a relational database, three basic operations, as shown in Figure 6.5, are used to develop useful sets of data: select, join, and project. The *select*

FIGURE 6.4 RELATIONAL DATABASE TABLES

A relational database organizes data in the form of two-dimensional tables. Illustrated here are tables for the entities SUPPLIER and PART showing how they represent each entity and its attributes. Supplier_Number is a primary key for the SUPPLIER table and a foreign key for the PART table.



operation creates a subset consisting of all records in the file that meet stated criteria. Select creates, in other words, a subset of rows that meet certain criteria. In our example, we want to select records (rows) from the PART table where the Part_Number equals 137 or 150. The *join* operation combines relational tables to provide the user with more information than is available in individual tables. In our example, we want to join the now-shortened PART table (only parts 137 or 150 will be presented) and the SUPPLIER table into a single new table.

The *project* operation creates a subset consisting of columns in a table, permitting the user to create new tables that contain only the information required. In our example, we want to extract from the new table only the following columns: Part_Number, Part_Name, Supplier_Number, and Supplier_Name.

Capabilities of Database Management Systems

A DBMS includes capabilities and tools for organizing, managing, and accessing the data in the database. The most important are its data definition language, data dictionary, and data manipulation language.

FIGURE 6.5 THE THREE BASIC OPERATIONS OF A RELATIONAL DBMS

The select, join, and project operations enable data from two different tables to be combined and only selected attributes to be displayed.

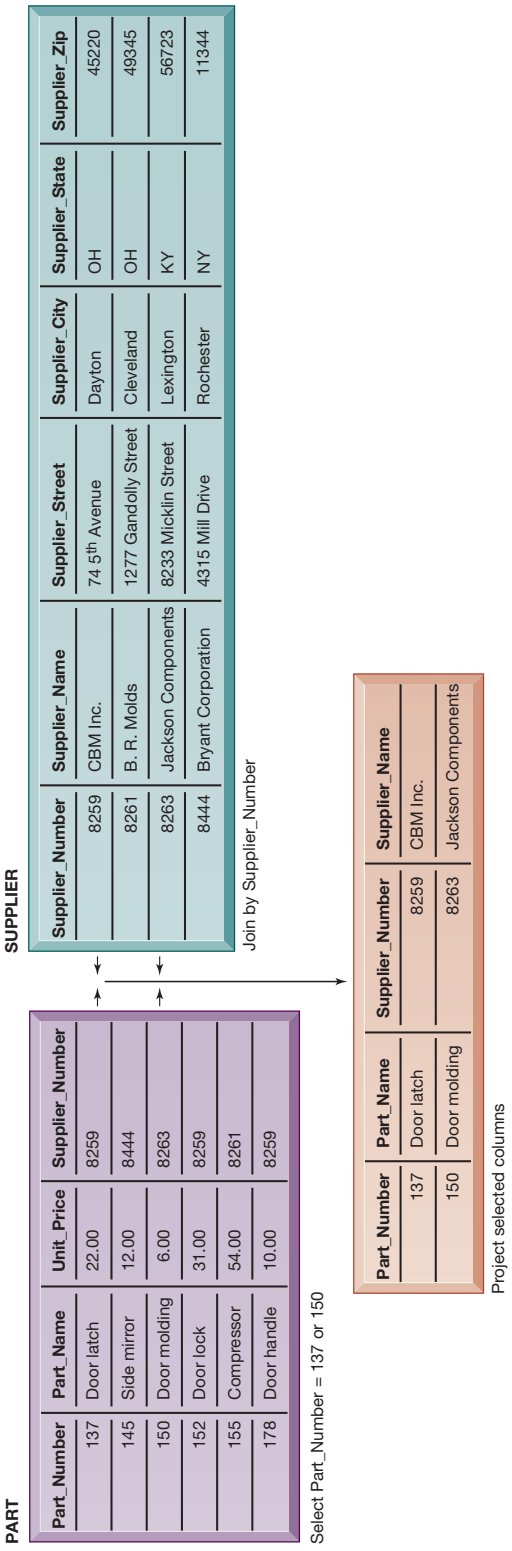
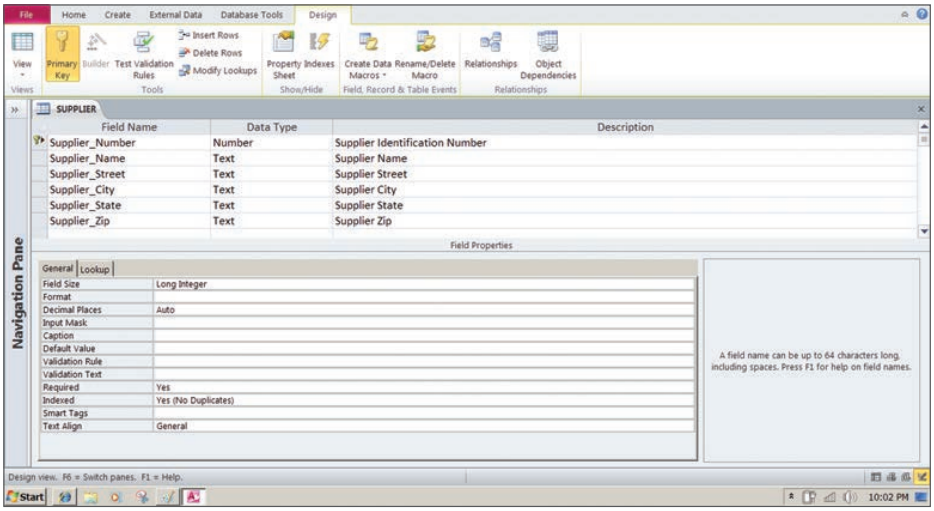


FIGURE 6.6 ACCESS DATA DICTIONARY FEATURES

Microsoft Access has a rudimentary data dictionary capability that displays information about the size, format, and other characteristics of each field in a database. Displayed here is the information maintained in the SUPPLIER table. The small key icon to the left of Supplier_Number indicates that it is a key field.

Courtesy of Microsoft Corporation



DBMS have a **data definition** capability to specify the structure of the content of the database. It would be used to create database tables and to define the characteristics of the fields in each table. This information about the database would be documented in a data dictionary. A **data dictionary** is an automated or manual file that stores definitions of data elements and their characteristics.

Microsoft Access has a rudimentary data dictionary capability that displays information about the name, description, size, type, format, and other properties of each field in a table (see Figure 6.6). Data dictionaries for large corporate databases may capture additional information, such as usage, ownership (who in the organization is responsible for maintaining the data), authorization, security, and the individuals, business functions, programs, and reports that use each data element.

Querying and Reporting

DBMS includes tools for accessing and manipulating information in databases. Most DBMS have a specialized language called a **data manipulation language** that is used to add, change, delete, and retrieve the data in the database. This language contains commands that permit end users and programming specialists to extract data from the database to satisfy information requests and develop applications. The most prominent data manipulation language today is **Structured Query Language**, or **SQL**. Figure 6.7 illustrates the SQL query

FIGURE 6.7 EXAMPLE OF AN SQL QUERY

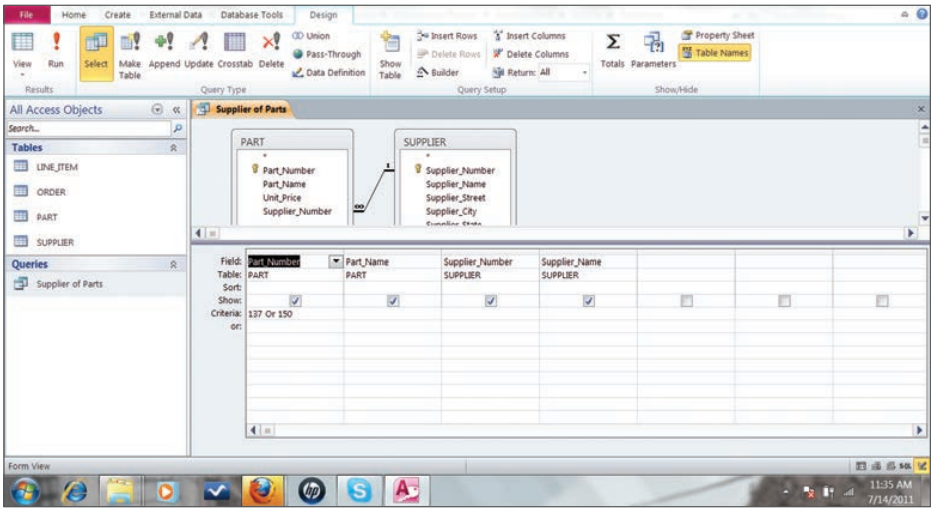
Illustrated here are the SQL statements for a query to select suppliers for parts 137 or 150. They produce a list with the same results as Figure 6.5.

```
SELECT PART.Part_Number, PART.Part_Name, SUPPLIER.Supplier_Number,
SUPPLIER.Supplier_Name
FROM PART, SUPPLIER
WHERE PART.Supplier_Number = SUPPLIER.Supplier_Number AND
Part_Number = 137 OR Part_Number = 150;
```

FIGURE 6.8 AN ACCESS QUERY

Illustrated here is how the query in Figure 6.7 would be constructed using Microsoft Access query-building tools. It shows the tables, fields, and selection criteria used for the query.

Courtesy of Microsoft Corporation



that would produce the new resultant table in Figure 6.5. You can find out more about how to perform SQL queries in our Learning Tracks for this chapter.

Users of DBMS for large and midrange computers, such as DB2, Oracle, or SQL Server, would employ SQL to retrieve information they needed from the database. Microsoft Access also uses SQL, but it provides its own set of user-friendly tools for querying databases and for organizing data from databases into more polished reports.

In Microsoft Access, you will find features that enable users to create queries by identifying the tables and fields they want and the results and then selecting the rows from the database that meet particular criteria. These actions in turn are translated into SQL commands. Figure 6.8 illustrates how the same query as the SQL query to select parts and suppliers would be constructed using the Microsoft Access query-building tools.

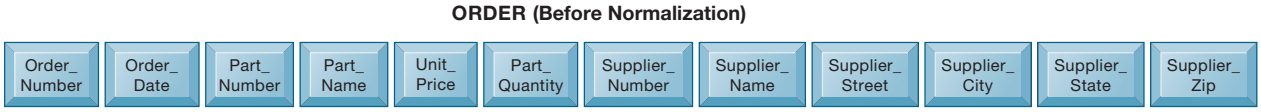
Microsoft Access and other DBMS include capabilities for report generation so that the data of interest can be displayed in a more structured and polished format than would be possible just by querying. Crystal Reports is a popular report generator for large corporate DBMS, although it can also be used with Access. Access also has capabilities for developing desktop system applications. These include tools for creating data entry screens, reports, and developing the logic for processing transactions.

Designing Databases

To create a database, you must understand the relationships among the data, the type of data that will be maintained in the database, how the data will be used, and how the organization will need to change to manage data from a companywide perspective. The database requires both a conceptual design and a physical design. The conceptual, or logical, design of a database is an abstract model of the database from a business perspective, whereas the physical design shows how the database is actually arranged on direct-access storage devices.

FIGURE 6.9 AN UNNORMALIZED RELATION FOR ORDER

An unnormalized relation contains repeating groups. For example, there can be many parts and suppliers for each order. There is only a one-to-one correspondence between Order_Number and Order_Date.



Normalization and Entity-Relationship Diagrams

The conceptual database design describes how the data elements in the database are to be grouped. The design process identifies relationships among data elements and the most efficient way of grouping data elements together to meet business information requirements. The process also identifies redundant data elements and the groupings of data elements required for specific application programs. Groups of data are organized, refined, and streamlined until an overall logical view of the relationships among all the data in the database emerges.

To use a relational database model effectively, complex groupings of data must be streamlined to minimize redundant data elements and awkward many-to-many relationships. The process of creating small, stable, yet flexible and adaptive data structures from complex groups of data is called **normalization**. Figures 6.9 and 6.10 illustrate this process.

In the particular business modeled here, an order can have more than one part, but each part is provided by only one supplier. If we build a relation called ORDER with all the fields included here, we would have to repeat the name and address of the supplier for every part on the order, even though the order is for parts from a single supplier. This relationship contains what are called repeating data groups because there can be many parts on a single order to a given supplier. A more efficient way to arrange the data is to break down ORDER into smaller relations, each of which describes a single entity. If we go step by step and normalize the relation ORDER, we emerge with the relations illustrated in Figure 6.10. You can find out more about normalization, entity-relationship diagramming, and database design in the Learning Tracks for this chapter.

FIGURE 6.10 NORMALIZED TABLES CREATED FROM ORDER

After normalization, the original relation ORDER has been broken down into four smaller relations. The relation ORDER is left with only two attributes, and the relation LINE_ITEM has a combined, or concatenated, key consisting of Order_ Number and Part_Number.

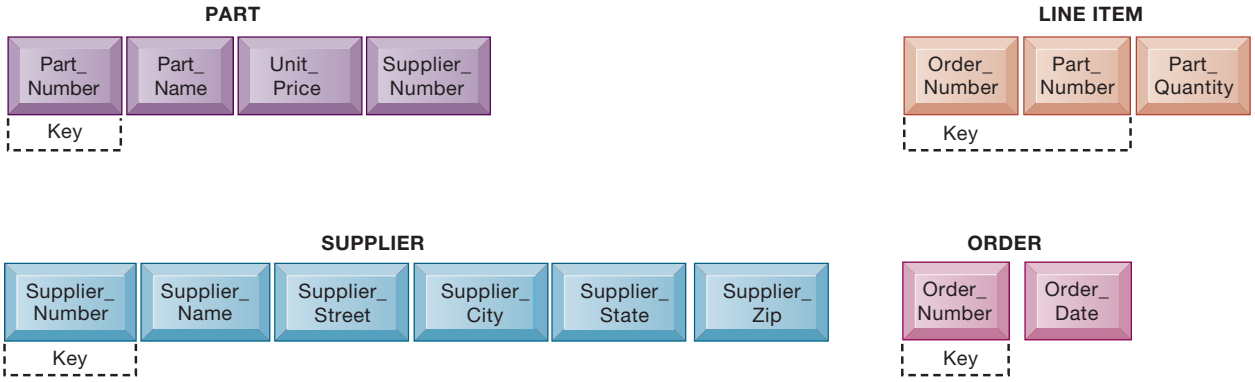
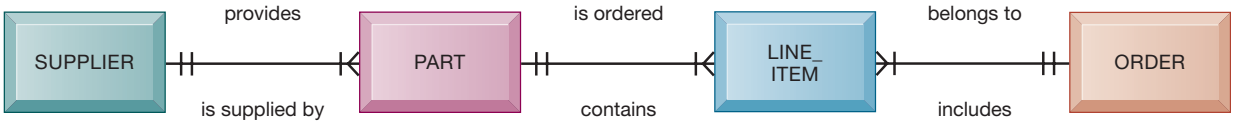


FIGURE 6.11 AN ENTITY-RELATIONSHIP DIAGRAM

This diagram shows the relationships between the entities SUPPLIER, PART, LINE_ITEM, and ORDER that might be used to model the database in Figure 6.10.



Relational database systems try to enforce **referential integrity** rules to ensure that relationships between coupled tables remain consistent. When one table has a foreign key that points to another table, you may not add a record to the table with the foreign key unless there is a corresponding record in the linked table. In the database we examined earlier in this chapter, the foreign key Supplier_Number links the PART table to the SUPPLIER table. We may not add a new record to the PART table for a part with Supplier_Number 8266 unless there is a corresponding record in the SUPPLIER table for Supplier_Number 8266. We must also delete the corresponding record in the PART table if we delete the record in the SUPPLIER table for Supplier_Number 8266. In other words, we shouldn't have parts from nonexistent suppliers!

Database designers document their data model with an **entity-relationship diagram**, illustrated in Figure 6.11. This diagram illustrates the relationship between the entities SUPPLIER, PART, LINE_ITEM, and ORDER. The boxes represent entities. The lines connecting the boxes represent relationships. A line connecting two entities that ends in two short marks designates a one-to-one relationship. A line connecting two entities that ends with a crow's foot topped by a short mark indicates a one-to-many relationship. Figure 6.11 shows that one ORDER can contain many LINE_ITEMs. (A PART can be ordered many times and appear many times as a line item in a single order.) Each PART can have only one SUPPLIER, but many PARTs can be provided by the same SUPPLIER.

It can't be emphasized enough: If the business doesn't get its data model right, the system won't be able to serve the business well. The company's systems will not be as effective as they could be because they'll have to work with data that may be inaccurate, incomplete, or difficult to retrieve. Understanding the organization's data and how they should be represented in a database is perhaps the most important lesson you can learn from this course.

For example, Famous Footwear, a shoe store chain with more than 800 locations in 49 states, could not achieve its goal of having "the right style of shoe in the right store for sale at the right price" because its database was not properly designed for rapidly adjusting store inventory. The company had an Oracle relational database running on a midrange computer, but the database was designed primarily for producing standard reports for management rather than for reacting to marketplace changes. Management could not obtain precise data on specific items in inventory in each of its stores. The company had to work around this problem by building a new database where the sales and inventory data could be better organized for analysis and inventory management.

Non-relational Databases, Cloud Databases, and Blockchain

For more than 30 years, relational database technology has been the gold standard. Cloud computing, unprecedented data volumes, massive workloads for web services, and the need to store new types of data require

database alternatives to the traditional relational model of organizing data in the form of tables, columns, and rows. Companies are turning to "NoSQL" non-relational database technologies for this purpose. **Non-relational database management systems** use a more flexible data model and are designed for managing large data sets across many distributed machines and for easily scaling up or down. They are useful for accelerating simple queries against large volumes of structured and unstructured data, including web, social media, graphics, and other forms of data that are difficult to analyze with traditional SQL-based tools.

There are several different kinds of NoSQL databases, each with its own technical features and behavior. Oracle NoSQL Database is one example, as is Amazon's SimpleDB, one of the Amazon Web Services that run in the cloud. SimpleDB provides a simple web services interface to create and store multiple data sets, query data easily, and return the results. There is no need to predefine a formal database structure or change that definition if new data are added later.

MetLife's MongoDB open source NoSQL database brings together data from more than 70 separate administrative systems, claims systems, and other data sources, including semi-structured and unstructured data, such as images of health records and death certificates. The NoSQL database can handle structured, semi-structured, and unstructured information without requiring tedious, expensive, and time-consuming database mapping to normalize all data to a rigid schema, as required by relational databases.

Cloud Databases and Distributed Databases

Among the services Amazon and other cloud computing vendors provide are relational database engines. Amazon Relational Database Service (Amazon RDS) offers MySQL, Microsoft SQL Server, Oracle Database, PostgreSQL, or Amazon Aurora as database engines. Pricing is based on usage. Oracle has its own Database Cloud Services using its relational Oracle Database, and Microsoft Azure SQL Database is a cloud-based relational database service based on the Microsoft SQL Server DBMS. Cloud-based data management services have special appeal for web-focused startups or small to medium-sized businesses seeking database capabilities at a lower cost than in-house database products.

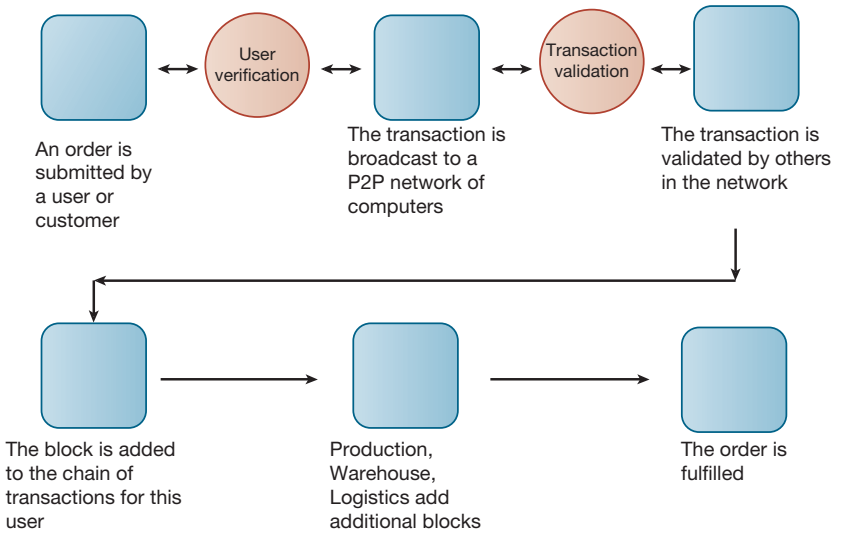
Google now offers its Spanner distributed database technology as a cloud service. A **distributed database** is one that is stored in multiple physical locations. Parts or copies of the database are physically stored in one location and other parts or copies are maintained in other locations. Spanner makes it possible to store information across millions of machines in hundreds of data centers around the globe, with special time-keeping tools to synchronize the data precisely in all of its locations and ensure the data are always consistent. Google uses Spanner to support its various cloud services, including Google Photos, AdWords (Google's online ad system), and Gmail, and is now making the technology available to other companies that might need such capabilities to run a global business.

Blockchain

Blockchain is a distributed database technology that enables firms and organizations to create and verify transactions on a network nearly instantaneously without a central authority. The system stores transactions as a distributed ledger among a network of computers The information held in the database is continually reconciled by the computers in the network.

FIGURE 6.12 HOW BLOCKCHAIN WORKS

A blockchain system is a distributed database that records transactions in a peer-to-peer network of computers



The blockchain maintains a continuously growing list of records called blocks. Each block contains a timestamp and link to a previous block. Once a block of data is recorded on the blockchain ledger, it cannot be altered retroactively. When someone wants to add a transaction, participants in the network (all of whom have copies of the existing blockchain) run algorithms to evaluate and verify the proposed transaction. Legitimate changes to the ledger are recorded across the blockchain in a matter of seconds or minutes and records are protected through cryptography. What makes a blockchain system possible and attractive to business firms is encryption and authentication of the actors and participating firms, which ensures that only legitimate actors can enter information, and only validated transactions are accepted. Once recorded, the transaction cannot be changed. Figure 6.12 illustrates how blockchain works for fulfilling an order.

There are many large benefits to firms using blockchain databases. Blockchain networks radically reduce the cost of verifying users, validating transactions, and the risks of storing and processing transaction information across thousands of firms. Instead of thousands of firms building their own private transaction systems, then integrating them with suppliers, shippers, and financial institution systems, blockchain can provide a single, simple, low-cost transaction system for participating firms. Standardization of recording transactions is aided through the use of *smart contracts*. Smart contracts are computer programs that implement the rules governing transactions between firms, e.g., what is the price of products, how will they be shipped, when will the transaction be completed, who will finance the transaction, what are financing terms, and the like.

The simplicity and security that blockchain offers has made it attractive for storing and securing financial transactions, supply chain transactions, medical records, and other types of data. Blockchain is a foundation technology for Bitcoin, Ethereum, and other cryptocurrencies. Chapter 8 provides more detail on securing transactions with blockchain.

6-3 What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?

Businesses use their databases to keep track of basic transactions, such as paying suppliers, processing orders, keeping track of customers, and paying employees. But they also need databases to provide information that will help the company run the business more efficiently and help managers and employees make better decisions. If a company wants to know which product is the most popular or who is its most profitable customer, the answer lies in the data.

The Challenge of Big Data

Most data collected by organizations used to be transaction data that could easily fit into rows and columns of relational database management systems. We are now witnessing an explosion of data from web traffic, email messages, and social media content (tweets, status messages), as well as machine-generated data from sensors (used in smart meters, manufacturing sensors, and electrical meters) or from electronic trading systems. These data may be unstructured or semi-structured and thus not suitable for relational database products that organize data in the form of columns and rows. We now use the term **big data** to describe these data sets with volumes so huge that they are beyond the ability of typical DBMS to capture, store, and analyze.

Big data is often characterized by the "3Vs": the extreme *volume* of data, the wide *variety* of data types and sources, and the *velocity* at which data must be processed. Big data doesn't designate any specific quantity but usually refers to data in the petabyte and exabyte range—in other words, billions to trillions of records, many from different sources. Big data are produced in much larger quantities and much more rapidly than traditional data. For example, a single jet engine is capable of generating 10 terabytes of data in just 30 minutes, and there are more than 25,000 airline flights each day. Twitter generates more than 8 terabytes of data daily. According to the International Data Center (IDC) technology research firm, data are more than doubling every two years, so the amount of data available to organizations is skyrocketing.

Businesses are interested in big data because they can reveal more patterns and interesting relationships than smaller data sets, with the potential to provide new insights into customer behavior, weather patterns, financial market activity, or other phenomena. For example, Shutterstock, the global online image marketplace, stores 24 million images, adding 10,000 more each day. To find ways to optimize the buying experience, Shutterstock analyzes its big data to find out where its website visitors place their cursors and how long they hover over an image before making a purchase. Big data is also finding many uses in the public sector, For example, city governments have been using big data to manage traffic flows and to fight crime.

However, to derive business value from these data, organizations need new technologies and tools capable of managing and analyzing nontraditional data along with their traditional enterprise data. They also need to know what questions to ask of the data and limitations of big data. Capturing, storing, and analyzing big data can be expensive, and information from big data may not necessarily help decision makers. It's important to have a clear understanding of the problem big data will solve for the business. The chapter-ending case explores these issues.

Business Intelligence Infrastructure

Suppose you wanted concise, reliable information about current operations, trends, and changes across the entire company. If you worked in a large company, the data you need might have to be pieced together from separate systems, such as sales, manufacturing, and accounting, and even from external sources, such as demographic or competitor data. Increasingly, you might need to use big data. A contemporary infrastructure for business intelligence has an array of tools for obtaining useful information from all the different types of data used by businesses today, including semi-structured and unstructured big data in vast quantities. These capabilities include data warehouses and data marts, Hadoop, in-memory computing, and analytical platforms. Some of these capabilities are available as cloud services.

Data Warehouses and Data Marts

The traditional tool for analyzing corporate data for the past two decades has been the data warehouse. A **data warehouse** is a database that stores current and historical data of potential interest to decision makers throughout the company. The data originate in many core operational transaction systems, such as systems for sales, customer accounts, and manufacturing, and may include data from website transactions. The data warehouse extracts current and historical data from multiple operational systems inside the organization. These data are combined with data from external sources and transformed by correcting inaccurate and incomplete data and restructuring the data for management reporting and analysis before being loaded into the data warehouse.

The data warehouse makes the data available for anyone to access as needed, but the data cannot be altered. A data warehouse system also provides a range of ad hoc and standardized query tools, analytical tools, and graphical reporting facilities.

Companies often build enterprise-wide data warehouses, where a central data warehouse serves the entire organization, or they create smaller, decentralized warehouses called data marts. A **data mart** is a subset of a data warehouse in which a summarized or highly focused portion of the organization's data is placed in a separate database for a specific population of users. For example, a company might develop marketing and sales data marts to deal with customer information. Bookseller Barnes & Noble used to maintain a series of data marts—one for point-of-sale data in retail stores, another for college bookstore sales, and a third for online sales.

Hadoop

Relational DBMS and data warehouse products are not well suited for organizing and analyzing big data or data that do not easily fit into columns and rows used in their data models. For handling unstructured and semi-structured data in vast quantities, as well as structured data, organizations are using **Hadoop**. Hadoop is an open source software framework managed by the Apache Software Foundation that enables distributed parallel processing of huge amounts of data across inexpensive computers. It breaks a big data problem down into subproblems, distributes them among up to thousands of inexpensive computer processing nodes, and then combines the result into a smaller data set that is easier to analyze. You've probably used Hadoop to find the best airfare on the Internet, get directions to a restaurant, do a search on Google, or connect with a friend on Facebook.

Hadoop consists of several key services, including the Hadoop Distributed File System (HDFS) for data storage and MapReduce for high-performance parallel data processing. HDFS links together the file systems on the numerous nodes in a Hadoop cluster to turn them into one big file system. Hadoop's MapReduce was inspired by Google's MapReduce system for breaking down processing of huge data sets and assigning work to the various nodes in a cluster. HBase, Hadoop's non-relational database, provides rapid access to the data stored on HDFS and a transactional platform for running high-scale real-time applications.

Hadoop can process large quantities of any kind of data, including structured transactional data, loosely structured data such as Facebook and Twitter feeds, complex data such as web server log files, and unstructured audio and video data. Hadoop runs on a cluster of inexpensive servers, and processors can be added or removed as needed. Companies use Hadoop for analyzing very large volumes of data as well as for a staging area for unstructured and semi-structured data before they are loaded into a data warehouse. Yahoo uses Hadoop to track users' behavior so it can modify its home page to fit their interests. Life sciences research firm NextBio uses Hadoop and HBase to process data for pharmaceutical companies conducting genomic research. Top database vendors such as IBM, Hewlett-Packard, Oracle, and Microsoft have their own Hadoop software distributions. Other vendors offer tools for moving data into and out of Hadoop or for analyzing data within Hadoop.

In-Memory Computing

Another way of facilitating big data analysis is to use **in-memory computing**, which relies primarily on a computer's main memory (RAM) for data storage. (Conventional DBMS use disk storage systems.) Users access data stored in system primary memory, thereby eliminating bottlenecks from retrieving and reading data in a traditional, disk-based database and dramatically shortening query response times. In-memory processing makes it possible for very large sets of data, amounting to the size of a data mart or small data warehouse, to reside entirely in memory. Complex business calculations that used to take hours or days are able to be completed within seconds, and this can even be accomplished using handheld devices.

The previous chapter describes some of the advances in contemporary computer hardware technology that make in-memory processing possible, such as powerful high-speed processors, multicore processing, and falling computer memory prices. These technologies help companies optimize the use of memory and accelerate processing performance while lowering costs.

Leading in-memory database products include SAP HANA, Oracle Database In-Memory, and Teradata Intelligent Memory. The chapter-opening case on the Charlotte Hornets and the Interactive Session on the Kraft Company show how organizations are benefiting from in-memory technology.

Analytic Platforms

Commercial database vendors have developed specialized high-speed **analytic platforms** using both relational and non-relational technology that are optimized for analyzing large data sets. Analytic platforms feature preconfigured hardware-software systems that are specifically designed for query processing and analytics. For example, the IBM PureData System for Analytics features tightly integrated database, server, and storage components that handle complex analytic queries 10 to 100 times faster than traditional systems.

INTERACTIVE SESSION TECHNOLOGY

Kraft Heinz Finds a New Recipe for Analyzing Its Data

When the Kraft Foods Group and Heinz finalized their merger in July 2015, it was the marriage of two giants. The new Kraft Heinz Company became the fifth-largest consumer-packaged food and beverage organization in the world. The combined company has more than 200 global brands, $26.5 billion in revenue, and over 40,000 employees. Eight of the brands each have annual revenue exceeding $1 billion: Heinz, Maxwell House, Kraft Lunchables, Planters, Velveeta, Philadelphia, and Oscar Mayer. Running these companies required huge amounts of data from all of these brands. This is clearly the world of big data.

To remain profitable, enterprises in the fast-moving consumer goods industry require very lean operations. The uncertain global economy has dampened consumer spending, so companies such as Kraft Heinz must constantly identify opportunities for improving operational efficiencies to protect their profit margins. Kraft Heinz decided to deal with this challenge by focusing on optimizing its supply chain, manufacturing optimal quantities of each of its products, and delivering them to retailers at the best time and least cost to capitalize on consumer demand.

Managing a supply chain as large as that of Kraft Heinz requires timely and accurate data on sales forecasts, manufacturing plans, and logistics, often from multiple sources. To ensure that Kraft Heinz would be able to use all of its enterprise business data effectively, management decided to split the data among two large SAP enterprise resource planning (ERP) systems, one for North American business and the other for all other global business. The combined company also had to rethink its data warehouse.

Before the merger, the North America business had maintained nearly 18 terabytes of data in a SAP Business Warehouse and was using SAP Business Warehouse Accelerator to facilitate operational reporting. SAP Business Warehouse is SAP's data warehouse software for consolidating organizational data and supporting data analytics and reporting. The SAP Business Warehouse (BW) Accelerator is used to speed up database queries. Kraft Heinz management wanted decision makers to obtain more fine-grained views of the data that would reveal new opportunities for improving efficiency, self-service reporting, and real-time analytics.

SAP BW Accelerator was not suitable for these tasks. It could optimize query runtime (the period of time when a query program is running) only for a specific subset of data in the warehouse, and was limited to reporting on selected views of the data. It could not deal with data load and calculation performance and required replication of Business Warehouse data in a separate accelerator. With mushrooming data on the merged company's sales, logistics, and manufacturing, the warehouse was too overtaxed to generate timely reports for decision makers. Moreover, Kraft Heinz's complex data model made building new reports very time-consuming—it could take as much as six months to complete. Kraft Heinz needed a solution that would deliver more detailed reports more quickly without affecting the performance of underlying operational systems.

Kraft Heinz business users had been building some of their own reports using SAP BusinessObjects Analysis edition for Microsoft Office, which integrates with Microsoft Excel and PowerPoint. This tool allows ad hoc multidimensional analysis. What these users needed was to be able to build self-service reports from a single source of data and find an efficient way to collate data from multiple sources to obtain an enterprise-wide view of what was going on.

Kraft Heinz decided to migrate its data warehouse from its legacy database to SAP BW powered by SAP HANA, SAP's in-memory database platform, which dramatically improves the efficiency at which data can be loaded and processed, calculations can be computed, and queries and reports can be run. The new data warehouse would be able to integrate with existing SAP ERP applications driving day-to-day business operations. The company worked with IBM Global Services consultants to cleanse and streamline its existing databases. It archived and purged unwanted or unused data, with the IT department working closely with business professionals to jointly determine what was essential, what was still being used, and what data thought to be unused had been moved to a different functional area of the company.

Cleansing and streamlining data reduced the database size almost 50 percent, to 9 terabytes.

According to Sundar Dittakavi, Kraft Heinz Group Leader of Global Business Intelligence, in addition to providing better insights, the new data warehouse environment has achieved a 98 percent improvement in the production of standard reports. This is due to the 83 percent reduction in load time to execution time to make the data available, and reduction in execution time to complete the analysis. Global key performance indicators for the Kraft side of the business are built into SAP HANA.

Kraft Heinz can now accommodate exploding volumes of data and database queries easily, while maintaining enough processing power to handle unexpected issues. The company is also able to build new reports much faster and the flexibility of SAP HANA makes it much easier to change the company's data model. Now Kraft Heinz can produce new reports for business users in weeks instead of months and give decision makers the insights they need to boost efficiency and lower operating costs.

Sources: Ken Murphy, "The Kraft-Heinz Company Unlocks Recipe for Strategic Business Insight," *SAP Insider Profiles*, January 25, 2017; "The Kraft Heinz Company Migrates SAP Business Warehouse to the Lightning-Fast SAP HANA Database," IBM Corp. and SAP SE 2016; and www.kraftheinzcompany.com, accessed February 15, 2018.

CASE STUDY QUESTIONS

1. Identify the problem in this case study. To what extent was it a technology problem? Were any management and organizational factors involved?
2. How was information technology affecting business performance at Kraft Heinz?
3. How did new technology provide a solution to the problem? How effective was the solution?
4. Identify the management, organizational, and technology factors that had to be addressed in selecting and implementing Kraft-Heinz's new data warehouse solution.

Analytic platforms also include in-memory systems and NoSQL non-relational database management systems and are now available as cloud services.

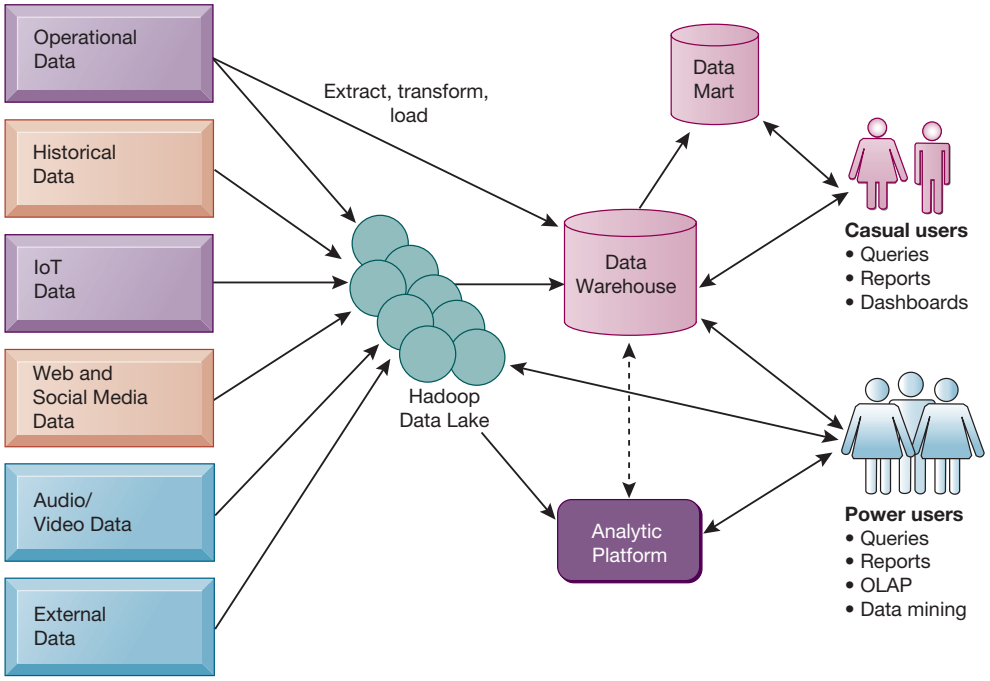
Figure 6.13 illustrates a contemporary business intelligence technology infrastructure using the technologies we have just described. Current and historical data are extracted from multiple operational systems along with web data, social media data, Internet of Things (IoT) machine-generated data, unstructured audio/visual data, and other data from external sources. Some companies are starting to pour all of these types of data into a data lake. A **data lake** is a repository for raw unstructured data or structured data that for the most part has not yet been analyzed, and the data can be accessed in many ways. The data lake stores these data in their native format until they are needed. The Hadoop Distributed File System (HDFS) is often used to store the data lake contents across a set of clustered computer nodes, and Hadoop clusters may be used to pre-process some of these data for use in the data warehouse, data marts, or an analytic platform, or for direct querying by power users. Outputs include reports and dashboards as well as query results. Chapter 12 discusses the various types of BI users and BI reporting in greater detail.

Analytical Tools: Relationships, Patterns, Trends

Once data have been captured and organized using the business intelligence technologies we have just described, they are available for further analysis using software for database querying and reporting, multidimensional data analysis (OLAP), and data mining. This section will introduce you to these tools, with more detail about business intelligence analytics and applications in Chapter 12.

FIGURE 6.13 CONTEMPORARY BUSINESS INTELLIGENCE INFRASTRUCTURE

A contemporary business intelligence infrastructure features capabilities and tools to manage and analyze large quantities and different types of data from multiple sources. Easy-to-use query and reporting tools for casual business users and more sophisticated analytical toolsets for power users are included.



Online Analytical Processing (OLAP)

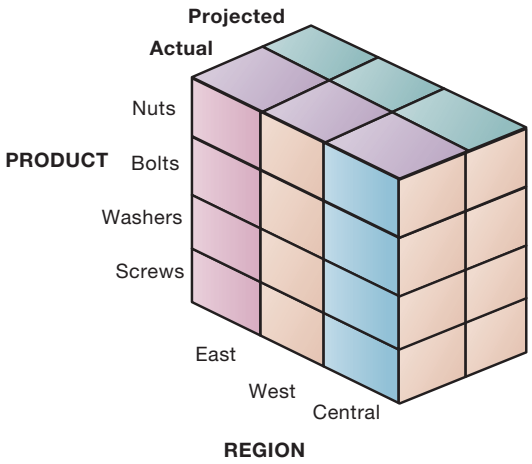
Suppose your company sells four different products—nuts, bolts, washers, and screws—in the East, West, and Central regions. If you wanted to ask a fairly straightforward question, such as how many washers were sold during the past quarter, you could easily find the answer by querying your sales database. But what if you wanted to know how many washers were sold in each of your sales regions and compare actual results with projected sales?

To obtain the answer, you would need **online analytical processing (OLAP)**. OLAP supports multidimensional data analysis, enabling users to view the same data in different ways using multiple dimensions. Each aspect of information—product, pricing, cost, region, or time period—represents a different dimension. So, a product manager could use a multidimensional data analysis tool to learn how many washers were sold in the East in June, how that compares with the previous month and the previous June, and how it compares with the sales forecast. OLAP enables users to obtain online answers to ad hoc questions such as these in a fairly rapid amount of time, even when the data are stored in very large databases, such as sales figures for multiple years.

Figure 6.14 shows a multidimensional model that could be created to represent products, regions, actual sales, and projected sales. A matrix of actual sales can be stacked on top of a matrix of projected sales to form a cube with six faces. If you rotate the cube 90 degrees one way, the face showing will be product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. If you rotate 180 degrees from the original view, you will see projected sales and product versus region.

FIGURE 6.14 MULTIDIMENSIONAL DATA MODEL

This view shows product versus region. If you rotate the cube 90 degrees, the face that will show is product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. Other views are possible.



Cubes can be nested within cubes to build complex views of data. A company would use either a specialized multidimensional database or a tool that creates multidimensional views of data in relational databases.

Data Mining

Traditional database queries answer such questions as "How many units of product number 403 were shipped in February 2018?" OLAP, or multidimensional analysis, supports much more complex requests for information, such as "Compare sales of product 403 relative to plan by quarter and sales region for the past two years." With OLAP and query-oriented data analysis, users need to have a good idea about the information for which they are looking.

Data mining is more discovery-driven. Data mining provides insights into corporate data that cannot be obtained with OLAP by finding hidden patterns and relationships in large databases and inferring rules from them to predict future behavior. The patterns and rules are used to guide decision making and forecast the effect of those decisions. The types of information obtainable from data mining include associations, sequences, classifications, clusters, and forecasts.

- *Associations* are occurrences linked to a single event. For instance, a study of supermarket purchasing patterns might reveal that, when corn chips are purchased, a cola drink is purchased 65 percent of the time, but when there is a promotion, cola is purchased 85 percent of the time. This information helps managers make better decisions because they have learned the profitability of a promotion.
- In *sequences*, events are linked over time. We might find, for example, that if a house is purchased, a new refrigerator will be purchased within two weeks 65 percent of the time, and an oven will be bought within one month of the home purchase 45 percent of the time.
- *Classification* recognizes patterns that describe the group to which an item belongs by examining existing items that have been classified and by inferring a set of rules. For example, businesses such as credit card or telephone companies worry about the loss of steady customers. Classification helps discover the characteristics of customers who are likely to leave and can

provide a model to help managers predict who those customers are so that the managers can devise special campaigns to retain such customers.

- *Clustering* works in a manner similar to classification when no groups have yet been defined. A data mining tool can discover different groupings within data, such as finding affinity groups for bank cards or partitioning a database into groups of customers based on demographics and types of personal investments.
- Although these applications involve predictions, *forecasting* uses predictions in a different way. It uses a series of existing values to forecast what other values will be. For example, forecasting might find patterns in data to help managers estimate the future value of continuous variables, such as sales figures.

These systems perform high-level analyses of patterns or trends, but they can also drill down to provide more detail when needed. There are data mining applications for all the functional areas of business and for government and scientific work. One popular use for data mining is to provide detailed analyses of patterns in customer data for one-to-one marketing campaigns or for identifying profitable customers.

Caesars Entertainment, formerly known as Harrah's Entertainment, is the largest gaming company in the world. It continually analyzes data about its customers gathered when people play its slot machines or use its casinos and hotels. The corporate marketing department uses this information to build a detailed gambling profile, based on a particular customer's ongoing value to the company. For instance, data mining lets Caesars know the favorite gaming experience of a regular customer at one of its riverboat casinos along with that person's preferences for room accommodations, restaurants, and entertainment. This information guides management decisions about how to cultivate the most profitable customers, encourage those customers to spend more, and attract more customers with high revenue-generating potential. Business intelligence improved Caesars's profits so much that it became the centerpiece of the firm's business strategy.

Text Mining and Web Mining

Unstructured data, most in the form of text files, is believed to account for more than 80 percent of useful organizational information and is one of the major sources of big data that firms want to analyze. Email, memos, call center transcripts, survey responses, legal cases, patent descriptions, and service reports are all valuable for finding patterns and trends that will help employees make better business decisions. **Text mining** tools are now available to help businesses analyze these data. These tools are able to extract key elements from unstructured natural language text, discover patterns and relationships, and summarize the information.

Businesses might turn to text mining to analyze transcripts of calls to customer service centers to identify major service and repair issues or to measure customer sentiment about their company. **Sentiment analysis** software is able to mine text comments in an email message, blog, social media conversation, or survey forms to detect favorable and unfavorable opinions about specific subjects. For example, Kraft Foods uses a Community Intelligence Portal and sentiment analysis to tune into consumer conversations about its products across numerous social networks, blogs, and other websites. Kraft tries to make sense of relevant comments rather than just track brand mentions and can identify customers' emotions and feelings when they talk about how they barbecue and what sauces and spices they use.

The web is another rich source of unstructured big data for revealing patterns, trends, and insights into customer behavior. The discovery and analysis of useful patterns and information from the World Wide Web are called **web mining**. Businesses might turn to web mining to help them understand customer behavior, evaluate the effectiveness of a particular website, or quantify the success of a marketing campaign. For instance, marketers use the Google Trends service, which tracks the popularity of various words and phrases used in Google search queries, to learn what people are interested in and what they are interested in buying.

Web mining looks for patterns in data through content mining, structure mining, and usage mining. Web content mining is the process of extracting knowledge from the content of web pages, which may include text, image, audio, and video data. Web structure mining examines data related to the structure of a particular website. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. Web usage mining examines user interaction data recorded by a web server whenever requests for a website's resources are received. The usage data records the user's behavior when the user browses or makes transactions on the website and collects the data in a server log. Analyzing such data can help companies determine the value of particular customers, cross-marketing strategies across products, and the effectiveness of promotional campaigns.

The chapter-ending case describes organizations' experiences as they use the analytical tools and business intelligence technologies we have described to grapple with "big data" challenges.

Databases and the Web

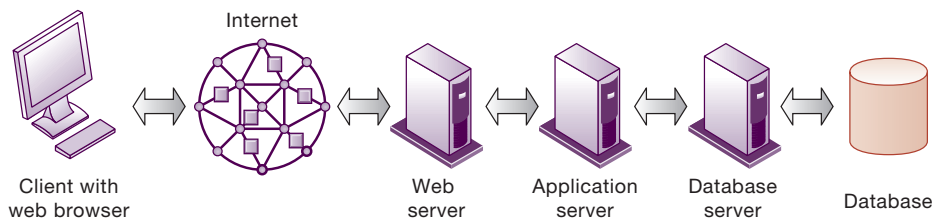
Have you ever tried to use the web to place an order or view a product catalog? If so, you were using a website linked to an internal corporate database. Many companies now use the web to make some of the information in their internal databases available to customers and business partners.

Suppose, for example, a customer with a web browser wants to search an online retailer's database for pricing information. Figure 6.15 illustrates how that customer might access the retailer's internal database over the web. The user accesses the retailer's website over the Internet using a web browser on his or her client PC or mobile device. The user's web browser software requests data from the organization's database, using HTML commands to communicate with the web server. Apps provide even faster access to corporate databases.

Because many back-end databases cannot interpret commands written in HTML, the web server passes these requests for data to software that translates HTML commands into SQL so the commands can be processed by the DBMS

FIGURE 6.15 LINKING INTERNAL DATABASES TO THE WEB

Users access an organization's internal database through the web using their desktop PC browsers or mobile apps.



working with the database. In a client/server environment, the DBMS resides on a dedicated computer called a **database server**. The DBMS receives the SQL requests and provides the required data. Middleware transfers information from the organization's internal database back to the web server for delivery in the form of a web page to the user.

Figure 6.15 shows that the middleware working between the web server and the DBMS is an application server running on its own dedicated computer (see Chapter 5). The application server software handles all application operations, including transaction processing and data access, between browser-based computers and a company's back-end business applications or databases. The application server takes requests from the web server, runs the business logic to process transactions based on those requests, and provides connectivity to the organization's back-end systems or databases. Alternatively, the software for handling these operations could be a custom program or a CGI script. A CGI script is a compact program using the *Common Gateway Interface (CGI)* specification for processing data on a web server.

There are a number of advantages to using the web to access an organization's internal databases. First, web browser software is much easier to use than proprietary query tools. Second, the web interface requires few or no changes to the internal database. It costs much less to add a web interface in front of a legacy system than to redesign and rebuild the system to improve user access.

Accessing corporate databases through the web is creating new efficiencies, opportunities, and business models. ThomasNet.com provides an up-to-date online directory of more than 500,000 suppliers of industrial products, such as chemicals, metals, plastics, rubber, and automotive equipment. Formerly called Thomas Register, the company used to send out huge paper catalogs with this information. Now it provides this information to users online via its website and has become a smaller, leaner company.

Other companies have created entirely new businesses based on access to large databases through the web. One is the social networking service Facebook, which helps users stay connected with each other and meet new people. Facebook features "profiles" with information on over 2.2 billion active users with information about themselves, including interests, friends, photos, and groups with which they are affiliated. Facebook maintains a very large database to house and manage all of this content. There are also many web-enabled databases in the public sector to help consumers and citizens access helpful information.

6-4 Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?

Setting up a database is only a start. In order to make sure that the data for your business remain accurate, reliable, and readily available to those who need them, your business will need special policies and procedures for data management.

Establishing an Information Policy

Every business, large and small, needs an information policy. Your firm's data are an important resource, and you don't want people doing whatever they want with them. You need to have rules on how the data are to be organized and maintained and who is allowed to view the data or change them.

An **information policy** specifies the organization's rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying information. Information policy lays out specific procedures and accountabilities, identifying which users and organizational units can share information, where information can be distributed, and who is responsible for updating and maintaining the information. For example, a typical information policy would specify that only selected members of the payroll and human resources department would have the right to change and view sensitive employee data, such as an employee's salary or social security number, and that these departments are responsible for making sure that such employee data are accurate.

If you are in a small business, the information policy would be established and implemented by the owners or managers. In a large organization, managing and planning for information as a corporate resource often require a formal data administration function. **Data administration** is responsible for the specific policies and procedures through which data can be managed as an organizational resource. These responsibilities include developing an information policy, planning for data, overseeing logical database design and data dictionary development, and monitoring how information systems specialists and end-user groups use data.

You may hear the term **data governance** used to describe many of these activities. Promoted by IBM, data governance deals with the policies and processes for managing the availability, usability, integrity, and security of the data employed in an enterprise with special emphasis on promoting privacy, security, data quality, and compliance with government regulations.

A large organization will also have a database design and management group within the corporate information systems division that is responsible for defining and organizing the structure and content of the database and maintaining the database. In close cooperation with users, the design group establishes the physical database, the logical relations among elements, and the access rules and security procedures. The functions it performs are called **database administration**.

Ensuring Data Quality

A well-designed database and information policy will go a long way toward ensuring that the business has the information it needs. However, additional steps must be taken to ensure that the data in organizational databases are accurate and remain reliable.

What would happen if a customer's telephone number or account balance were incorrect? What would be the impact if the database had the wrong price for the product you sold or your sales system and inventory system showed different prices for the same product? Data that are inaccurate, untimely, or inconsistent with other sources of information lead to incorrect decisions, product recalls, and financial losses. Gartner, Inc. reported that more than 25 percent of the critical data in large *Fortune* 1000 companies' databases is inaccurate or incomplete, including bad product codes and product descriptions, faulty inventory descriptions, erroneous financial data, incorrect supplier information, and incorrect employee data. Some of these data quality problems are caused by redundant and inconsistent data produced by multiple systems feeding a data warehouse. For example, the sales ordering system and the inventory management system might both maintain data on the

organization's products. However, the sales ordering system might use the term *Item Number* and the inventory system might call the same attribute *Product Number*. The sales, inventory, or manufacturing systems of a clothing retailer might use different codes to represent values for an attribute. One system might represent clothing size as "medium," whereas the other system might use the code "M" for the same purpose. During the design process for the warehouse database, data describing entities, such as a customer, product, or order, should be named and defined consistently for all business areas using the database.

Think of all the times you've received several pieces of the same direct mail advertising on the same day. This is very likely the result of having your name maintained multiple times in a database. Your name may have been misspelled or you used your middle initial on one occasion and not on another or the information was initially entered onto a paper form and not scanned properly into the system. Because of these inconsistencies, the database would treat you as different people! We often receive redundant mail addressed to Laudon, Lavdon, Lauden, or Landon.

If a database is properly designed and enterprise-wide data standards are established, duplicate or inconsistent data elements should be minimal. Most data quality problems, however, such as misspelled names, transposed numbers, or incorrect or missing codes, stem from errors during data input. The incidence of such errors is rising as companies move their businesses to the web and allow customers and suppliers to enter data into their websites that directly update internal systems.

Before a new database is in place, organizations need to identify and correct their faulty data and establish better routines for editing data once their database is in operation. Analysis of data quality often begins with a **data quality audit**, which is a structured survey of the accuracy and level of completeness of the data in an information system. Data quality audits can be performed by surveying entire data files, surveying samples from data files, or surveying end users for their perceptions of data quality.

Data cleansing, also known as *data scrubbing*, consists of activities for detecting and correcting data in a database that are incorrect, incomplete, improperly formatted, or redundant. Data cleansing not only corrects errors but also enforces consistency among different sets of data that originated in separate information systems. Specialized data-cleansing software is available to automatically survey data files, correct errors in the data, and integrate the data in a consistent companywide format.

Data quality problems are not just business problems. They also pose serious problems for individuals, affecting their financial condition and even their jobs. For example, inaccurate or outdated data about consumers' credit histories maintained by credit bureaus can prevent creditworthy individuals from obtaining loans or lower their chances of finding or keeping a job. And as the Interactive Session on Organization describes, incomplete or inaccurate databases also pose problems for criminal justice and public safety.

A small minority of companies allow individual departments to be in charge of maintaining the quality of their own data. However, best data administration practices call for centralizing data governance, standardization of organizational data, data quality maintenance, and accessibility to data assets.

INTERACTIVE SESSION ORGANIZATIONS

Databases Where the Data Aren't There

On November 5, 2017 Devin Patrick Kelley walked into the First Baptist Church in Sutherland Springs, Texas toting a Ruger AR-556 semi-automatic rifle and fired round after round into the congregation gathered for Sunday morning services. Within a few minutes, he had killed 26 people and injured 20 others. Kelley was later found dead in his SUV with a self-inflicted gunshot wound. The attack was the deadliest mass shooting by an individual in Texas, the fifth-deadliest mass shooting in the United States, as well as the deadliest shooting in an American place of worship in modern history.

This tragedy could have been avoided. Kelley was prohibited by law from purchasing or possessing firearms and ammunition due to a 2012 domestic violence conviction in a court martial while he was serving in the U.S. Air Force. The Air Force failed to record the conviction in the Federal Bureau of Investigation (FBI) National Crime Information Center (NCIC) database, which is used by the National Instant Check System (NICS) to flag prohibited gun purchases. This allowed Kelley to pass background checks and purchase four guns, one in each of the past four years.

Federally licensed firearm dealers are required to check the credentials of every potential buyer against the NICS system containing millions of criminal history records and protection orders. The system is supposed to flag any potential gun buyer who falls in various categories prohibiting a sale including fugitives, convicted felons, or those with dishonorable discharges from the military.

The Air Force acknowledged that it did not inform federal authorities about the domestic violence conviction, which should have prevented Kelley from buying firearms. Ann Stefanek, an Air Force spokesperson, stated that the Air Force would conduct a comprehensive review to ensure records in other cases have been reported correctly. The Defense Department plans to review how all U.S. military services report such cases into the background-check system. Members of the U.S. Senate have called for legislation to improve the completeness of NICS recordkeeping.

Individuals familiar with how NICS works observed that large gaps in information sharing between the military and the Justice Department have created a blind spot in background checks of veterans, allowing those barred from possessing weapons to get clearance. They believed that the failure to flag Kelley more likely reflected a systemic flaw rather than a one-time miss. Robert Belair, a Washington privacy lawyer and expert on the FBI's background-screening system, said the Air Force and other branches of the military seldom submit court-martial records to the FBI's screening database when the offense doesn't lead to a dishonorable discharge because this has never been a priority for the military.

According to a 2016 report by the U.S. Government Accountability Office (GAO), the FBI has struggled to collect domestic abuse records for background checks, in part because incomplete or missing criminal histories make it harder to determine if someone should be banned from obtaining a gun. The GAO focused on reporting by state and local authorities and reported that between 2006 and 2015 about 6,700 firearms were incorrectly transferred to individuals with domestic violence records.

Federal law requires federal departments, including the military branches, to notify the Justice Department at least quarterly about any records they have showing that someone is disqualified from buying a gun. At the state level, however, compliance is voluntary unless specified by state law or federal funding requirements. It isn't known how many court-martial records are submitted to the FBI, which said it couldn't provide the information.

Gaps in databases have also affected other aspects of law enforcement, such as sentencing and parole. The decision to parole O.J. Simpson in October 2017 is an example. Before voting to release O.J. Simpson from prison after nine years, the Nevada parole board discussed in detail the robbery that had put him behind bars and his conduct as an inmate. Members of the Nevada Board of Parole Commissioners stated that before Simpson's 2008 conviction for a Las Vegas hotel robbery, Simpson had no history of a criminal conviction. Although Simpson was acquitted in 1995 of the murders of his former wife Nicole Brown Simpson and Ronald Goldman, in 1989 he had pleaded no contest in Los Angeles to misdemeanor battery of Ms. Simpson, who was then his wife. The Nevada parole board did not have that information. The 1989 conviction was not considered when a four-member panel voted unanimously to release him in October 2017.

When states such as Nevada weigh the risk posed by an inmate, they routinely look through their own records, and also check with the NCIC. Mr. Simpson's 1989 conviction did not appear in the NCIC history when Nevada officials prepared a pre-sentencing report after his 2008 conviction. David M. Smith, hearings examiner for the Nevada parole board, said the parole commissioners relied in part on the information in that 2008 report in assessing whether Mr. Simpson should be released. Smith believed it was impossible to tell whether knowledge of Simpson's misdemeanor conviction would have made a difference in the Nevada parole board's decision.

Omission of Simpson's 1989 conviction in the federal system again highlights the problem of major gaps in federal criminal databases, which rely primarily on accurate and complete reporting by local and state agencies. The Justice Department has reported, for example, that states fail to transmit most of their active arrest warrants from their own databases into the federal system and often neglect to update records to show whether cases resulted in convictions. Some states still rely on paper files, making it likelier that they will not end up in the federal electronic records database, a problem that is more common with older records.

Sources: Kristina Peterson and Jacob Gershman, "Lapses in Gun Buyers' Records Come Under Scrutiny," *Wall Street Journal*, November 7, 2017; Melissa Jeltsen, "Air Force Failed to Enter Church Shooter's Domestic Violence Record In U.S. Database," *Huffington Post*, November 6, 2017; Richard Perez-Pena, "Nevada Parole Board Unaware of O.J. Simpson's Old Conviction," *New York Times*, August 11, 2017; and Eli Rosenberg, Mark Berman, and Wesley Lowery, "Texas Church Gunman Escaped Mental Health Facility in 2012 after Threatening Military Superiors," *Washington Post*, November 7, 2017.

CASE STUDY QUESTIONS

1. Define the problem described in this case. How serious a problem is it?
2. What management, organization, and technology factors contributed to this problem?
3. What is the political and social impact of incomplete recordkeeping in the FBI NCIC and NICS databases?

6-5 How will MIS help my career?

Here is how Chapter 6 and this book can help you find a job as an entry-level data analyst.

The Company

Mega Midwest Power, a large diversified energy company headquartered in Cleveland, Ohio, has an open position for an entry-level data analyst. The company is involved in the distribution, transmission, and generation of electricity as well as energy management and other energy-related services for 5 million customers in the Midwest and mid-Atlantic regions.

Position Description

Job responsibilities include:

- Maintaining the integrity of substation equipment and related data in multiple databases, including SAP.
- Querying databases in multiple systems.
- Modifying systems for proper data management and procedural controls.
- Recommending and implementing process changes based on data problems that are identified.
- Conducting business-specific research, gathering data, and compiling reports and summaries.
- Expanding knowledge of policies, practices, and procedures.

Job Requirements

- BA/BS degree in business, finance, accounting, economics, engineering, or related discipline
- 1–2 years professional work experience desirable
- Knowledge of Microsoft Office tools (Excel, PowerPoint, Access, and Word)
- Strong analytical capabilities, including attention to detail, problem solving, and decision making
- Strong oral and written communication and teamwork skills
- Familiarity with transmission substation equipment desirable

Interview Questions

1. What do you know about substation equipment? Have you ever worked with SAP for Utilities?
2. What do you know about data management and databases? Have you ever worked with data management software? If so, what exactly have you done with it?
3. Tell us what you can do with Access and Excel. What kinds of problems have you used these tools to solve? Did you take courses in Access or Excel?
4. What experience do you have analyzing problems and developing specific solutions? Can you give an example of a problem you helped solve?

Author Tips

1. Do some research on the electric utility industry equipment maintenance and software for electric utility asset management and predictive maintenance. Read blogs from IBM, Deloitte, and Intel about predictive maintenance and watch YouTube videos from GE and IBM on this topic.
2. Review Chapter 6 of this text on data management and databases, along with the Chapter 12 discussion of operational intelligence. Inquire what you would be expected to do with databases in this job position.
3. Do some research on the capabilities of SAP for Utilities and ask exactly how you would be using this software and what skills would be required. Watch SAP's YouTube video on SAP for Utilities.

REVIEW **SUMMARY**

6-1 What are the problems of managing data resources in a traditional file environment?

Traditional file management techniques make it difficult for organizations to keep track of all of the pieces of data they use in a systematic way and to organize these data so that they can be easily accessed. Different functional areas and groups were allowed to develop their own files independently. Over time, this traditional file management environment creates problems such as data redundancy and inconsistency, program-data dependence, inflexibility, poor security, and lack of data sharing and availability. A database management system (DBMS) solves these problems with software that permits centralization of data and data management so that businesses have a single consistent source for all their data needs. Using a DBMS minimizes redundant and inconsistent files.

6-2 What are the major capabilities of database management systens (DBMS), and why is a relational DBMS so powerful?

The principal capabilities of a DBMS include a data definition capability, a data dictionary capability, and a data manipulation language. The data definition capability specifies the structure and content of the

database. The data dictionary is an automated or manual file that stores information about the data in the database, including names, definitions, formats, and descriptions of data elements. The data manipulation language, such as SQL, is a specialized language for accessing and manipulating the data in the database.

The relational database has been the primary method for organizing and maintaining data in information systems because it is so flexible and accessible. It organizes data in two-dimensional tables called relations with rows and columns. Each table contains data about an entity and its attributes. Each row represents a record, and each column represents an attribute or field. Each table also contains a key field to uniquely identify each record for retrieval or manipulation. Relational database tables can be combined easily to deliver data required by users, provided that any two tables share a common data element. Non-relational databases are becoming popular for managing types of data that can't be handled easily by the relational data model. Both relational and non-relational database products are available as cloud computing services. A distributed database is one that is stored in multiple physical locations, including remote cloud computing centers.

Designing a database requires both a logical design and a physical design. The logical design models the database from a business perspective. The organization's data model should reflect its key business processes and decision-making requirements. The process of creating small, stable, flexible, and adaptive data structures from complex groups of data when designing a relational database is termed normalization. A well-designed relational database will not have many-to-many relationships, and all attributes for a specific entity will only apply to that entity. It will try to enforce referential integrity rules to ensure that relationships between coupled tables remain consistent. An entity-relationship diagram graphically depicts the relationship between entities (tables) in a relational database.

6-3 What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?

Contemporary data management technology has an array of tools for obtaining useful information from all the different types of data used by businesses today, including semi-structured and unstructured big data in vast quantities. These capabilities include data warehouses and data marts, Hadoop, in-memory computing, and analytical platforms. OLAP represents relationships among data as a multidimensional structure, which can be visualized as cubes of data and cubes within cubes of data, enabling more sophisticated data analysis. Data mining analyzes large pools of data, including the contents of data warehouses, to find patterns and rules that can be used to predict future behavior and guide decision making. Text mining tools help businesses analyze large unstructured data sets consisting of text. Web mining tools focus on analysis of useful patterns and information from the web, examining the structure of websites and activities of website users, as well as the contents of web pages. Conventional databases can be linked via middleware to the web or a web interface to facilitate user access to an organization's internal data.

6-4 Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?

Developing a database environment requires policies and procedures for managing organizational data as well as a good data model and database technology. A formal information policy governs the maintenance, distribution, and use of information in the organization. In large corporations, a formal data administration function is responsible for information policy as well as for data planning, data dictionary development, and monitoring data usage in the firm.

Data that are inaccurate, incomplete, or inconsistent create serious operational and financial problems for businesses because they may create inaccuracies in product pricing, customer accounts, and inventory data and lead to inaccurate decisions about the actions that should be taken by the firm. Firms must take special steps to make sure they have a high level of data quality. These include using enterprise-wide data standards, databases designed to minimize inconsistent and redundant data, data quality audits, and data cleansing software.

Key Terms

Analytic platform, 229
Attribute, 214
Big data, 227
Bit, 213
Blockchain, 225
Byte, 213
Data administration, 237
Data cleansing, 238

Data definition, 221
Data dictionary, 221
Data governance, 237
Data inconsistency, 215
Data lake, 231
Data manipulation language, 221
Data mart, 228
Data mining, 233
Data quality audit, 238
Data redundancy, 215
Data warehouse, 228
Database, 216
Database administration, 237
Database management system (DBMS), 217
Database server, 236
Distributed database, 225
Entity, 214
Entity-relationship diagram, 224
Field, 213
File, 213
Foreign key, 218
Hadoop, 228
In-memory computing, 229
Information policy, 237
Key field, 218
Non-relational database management systems, 225
Normalization, 223
Online analytical processing (OLAP), 232
Primary key, 218
Program-data dependence, 216
Record, 213
Referential integrity, 224
Relational DBMS, 218
Sentiment analysis, 234
Structured Query Language (SQL), 221
Text mining, 234
Tuple, 218
Web mining, 235

MyLab MIS

To complete the problems with MyLab MIS, go to the EOC Discussion Questions in MyLab MIS.

Review Questions

6-1 What are the problems of managing data resources in a traditional file environment?

- List and describe each of the components in the data hierarchy.
- Define and explain the significance of entities, attributes, and key fields.
- List and describe the problems of the traditional file environment.

6-2 What are the major capabilities of database management systems (DBMS), and why is a relational DBMS so powerful?

- Define a database and a database management system.
- Name and briefly describe the capabilities of a DBMS.
- Define a relational DBMS and explain how it organizes data.
- List and describe the three operations of a relational DBMS.
- Explain why non-relational databases are useful.
- Define and describe normalization and referential integrity and explain how they contribute to a well-designed relational database.
- Define and describe an entity-relationship diagram and explain its role in database design.

6-3 What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?

- Define big data and describe the technologies for managing and analyzing it.
- List and describe the components of a contemporary business intelligence infrastructure.
- Describe the capabilities of online analytical processing (OLAP).
- Define data mining, describing how it differs from OLAP and the types of information it provides.
- Explain how text mining and web mining differ from conventional data mining.
- Describe how users can access information from a company's internal databases through the web.

6-4 Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?

- Describe the roles of information policy and data administration in information management.
- Explain why data quality audits and data cleansing are essential.

Discussion Questions

6-5 MyLab MIS It has been said there is no bad data, just bad management. Discuss the implications of this statement.

6-6 MyLab MIS To what extent should end users be involved in the selection of a database management system and database design?

6-7 MyLab MIS What are the consequences of an organization not having an information policy?

Hands-On MIS Projects

The projects in this section give you hands-on experience in analyzing data quality problems, establishing companywide data standards, creating a database for inventory management, and using the web to search online databases for overseas business resources. Visit MyLab MIS to access this chapter's Hands-On MIS Projects.

Management Decision Problems

6-8 Emerson Process Management, a global supplier of measurement, analytical, and monitoring instruments and services based in Austin, Texas, had a new data warehouse designed for analyzing customer activity to improve service and marketing. However, the data warehouse was full of inaccurate and redundant data. The data in the warehouse came from numerous transaction processing systems in Europe, Asia, and other locations around the world. The team that designed the warehouse had assumed that sales groups in all these areas would enter customer names and addresses the same way. In fact, companies in different countries were using multiple ways of entering quote, billing, shipping, and other data. Assess the potential business impact of these data quality problems. What decisions have to be made and steps taken to reach a solution?

6-9 Your industrial supply company wants to create a data warehouse where management can obtain a single corporate-wide view of critical sales information to identify bestselling products, key customers, and sales trends. Your sales and product information are stored in two different systems: a divisional sales system running on a Unix server and a corporate sales system running on an IBM mainframe. You would like to create a single standard format that consolidates these data from both systems. In MyLab MIS, you can review the proposed format along with sample files from the two systems that would supply the data for the data warehouse. Then answer the following questions:

- What business problems are created by not having these data in a single standard format?
- How easy would it be to create a database with a single standard format that could store the data from both systems? Identify the problems that would have to be addressed.
- Should the problems be solved by database specialists or general business managers? Explain.
- Who should have the authority to finalize a single companywide format for this information in the data warehouse?

Achieving Operational Excellence: Building a Relational Database for Inventory Management

Software skills: Database design, querying, and reporting
Business skills: Inventory management

6-10 In this exercise, you will use database software to design a database for managing inventory for a small business. Sylvester's Bike Shop, located in San Francisco, California, sells road, mountain, hybrid, leisure, and children's bicycles. Currently, Sylvester's purchases bikes from three suppliers but plans to add new suppliers in the near future. Using the information found in the tables in MyLab MIS, build a simple relational database to manage information about Sylvester's suppliers and products. Once you have built the database, perform the following activities.

- Prepare a report that identifies the five most expensive bicycles. The report should list the bicycles in descending order from most expensive to least expensive, the quantity on hand for each, and the markup percentage for each.
- Prepare a report that lists each supplier, its products, the quantities on hand, and associated reorder levels. The report should be sorted alphabetically by supplier. For each supplier, the products should be sorted alphabetically.

- Prepare a report listing only the bicycles that are low in stock and need to be reordered. The report should provide supplier information for the items identified.
- Write a brief description of how the database could be enhanced to further improve management of the business. What tables or fields should be added? What additional reports would be useful?

Improving Decision Making: Searching Online Databases for Overseas Business Resources

Software skills: Online databases
Business skills: Researching services for overseas operations

6-11 This project develops skills in searching web-enabled databases with information about products and services in faraway locations.

Your company is located in Greensboro, North Carolina, and manufactures office furniture of various types. You are considering opening a facility to manufacture and sell your products in Australia. You would like to contact organizations that offer many services necessary for you to open your Australian office and manufacturing facility, including lawyers, accountants, import-export experts, and telecommunications equipment and support firms. Access the following online databases to locate companies that you would like to meet with during your upcoming trip: Australian Business Directory Online, AustraliaTrade Now, and the Nationwide Business Directory of Australia. If necessary, use search engines such as Yahoo and Google.

- List the companies you would contact on your trip to determine whether they can help you with these and any other functions you think are vital to establishing your office.
- Rate the databases you used for accuracy of name, completeness, ease of use, and general helpfulness.

Collaboration and Teamwork Project

Identifying Entities and Attributes in an Online Database

6-12 With your team of three or four other students, select an online database to explore, such as AOL Music, iGo.com, or the Internet Movie Database. Explore one of these websites to see what information it provides. Then list the entities and attributes that the company running the website must keep track of in its databases. Diagram the relationship between the entities you have identified. If possible, use Google Docs and Google Drive or Google Sites to brainstorm, organize, and develop a presentation of your findings for the class.

How Reliable Is Big Data?

CASE STUDY

Today's companies are dealing with an avalanche of data from social media, search, and sensors, as well as from traditional sources. According to one estimate, 2.5 quintillion bytes of data per day are generated around the world. Making sense of "big data" to improve decision making and business performance has become one of the primary opportunities for organizations of all shapes and sizes, but it also represents big challenges.

Businesses such as Amazon, YouTube, and Spotify have flourished by analyzing the big data they collect about customer interests and purchases to create millions of personalized recommendations. A number of online services analyze big data to help consumers, including services for finding the lowest price on autos, computers, mobile phone plans, clothing, airfare, hotel rooms, and many other types of goods and services. Big data is also providing benefits in sports (see the chapter-opening case), education, science, health care, and law enforcement.

Analyzing billions of data points collected on patients, healthcare providers, and the effectiveness of prescriptions and treatments has helped the UK National Health Service (NHS) save about 581 million pounds (U.S. $784 million). The data are housed in an Oracle Exadata Database Machine, which can quickly analyze very large volumes of data (review this chapter's discussion of analytic platforms). NHS has used its findings from big data analysis to create dashboards identifying patients taking 10 or more medications at once, and which patients are taking too many antibiotics. Compiling very large amounts of data about drugs and treatments given to cancer patients and correlating that information with patient outcomes has helped NHS identify more effective treatment protocols.

New York City analyzes all the crime-related data it collects to lower the crime rate. Its CompStat crime-mapping program uses a comprehensive citywide database of all reported crimes or complaints, arrests, and summonses in each of the city's 76 precincts to report weekly on crime complaint and arrest activity at the precinct, patrol borough, and citywide levels. CompStat data can be displayed on maps showing crime and arrest locations, crime hot spots, and other relevant information to help precinct commanders quickly identify patterns and trends and deploy police personnel where they are most needed. Big data on criminal activity also powers New York City's Crime Strategies Unit, which targets the worst offenders for aggressive prosecution. Healthcare companies are currently analyzing big data to determine the most effective and economical treatments for chronic illnesses and common diseases and provide personalized care recommendations to patients.

There are limits to using big data. A number of companies have rushed to start big data projects without first establishing a business goal for this new information or key performance metrics to measure success. Swimming in numbers doesn't necessarily mean that the right information is being collected or that people will make smarter decisions. Experts in big data analysis believe too many companies, seduced by the promise of big data, jump into big data projects with nothing to show for their efforts. They start amassing mountains of data with no clear objective or understanding of exactly how analyzing big data will achieve their goal or what questions they are trying to answer. Organizations also won't benefit from big data that has not been properly cleansed, organized, and managed—think data quality.

Just because something can be measured doesn't mean it should be measured. Suppose, for instance, that a large company wants to measure its website traffic in relation to the number of mentions on Twitter. It builds a digital dashboard to display the results continuously. In the past, the company had generated most of its sales leads and eventual sales from trade shows and conferences. Switching to Twitter mentions as the key metric to measure changes the sales department's focus. The department pours its energy and resources into monitoring website clicks and social media traffic, which produce many unqualified leads that never lead to sales.

Although big data is very good at detecting correlations, especially subtle correlations that an analysis of smaller data sets might miss, big data analysis doesn't necessarily show causation or which correlations are meaningful. For example, examining big data might show that from 2006 to 2011 the United States murder rate was highly correlated with the

market share of Internet Explorer, since both declined sharply. But that doesn't necessarily mean there is any meaningful connection between the two phenomena. Data analysts need some business knowledge of the problem they are trying to solve with big data.

Big data predictive models don't necessarily give you a better idea of what will happen in the future. Meridian Energy Ltd., an electricity generator and distributor operating in New Zealand and Australia, moved away from using an aging predictive equipment maintenance system. The software was supposed to predict the maintenance needs of all the large equipment the company owns and operates, including generators, wind turbines, transformers, circuit breakers, and industrial batteries. However, the system used outdated modeling techniques and could not actually predict equipment failures. It ran simulations of different scenarios and predicted when assets would fail the simulated tests. The recommendations of the software were useless because they did not accurately predict which pieces of equipment actually failed in the real world. Meridian eventually replaced the old system with IBM's Predictive Maintenance and Quality software, which bases predictions on more real-time data from equipment.

All data sets and data-driven forecasting models reflect the biases of the people selecting the data and performing the analysis. Several years ago, Google developed what it thought was a leading-edge algorithm using data it collected from web searches to determine exactly how many people had influenza and how the disease was spreading. It tried to calculate the number of people with flu in the United States by relating people's location to flu-related search queries on Google. Google consistently overestimated flu rates, when compared to conventional data collected afterward by the U.S. Centers for Disease Control (CDC). Several scientists suggested that Google was "tricked" by widespread media coverage of that year's severe flu season in the United States, which was further amplified by social media coverage. The model developed for forecasting flu trends was based on a flawed assumption—that the incidence of flu-related searches on Googles was a precise indicator of the number of people who actually came down with the flu. Google's algorithm only looked at numbers, not the context of the search results.

In addition to election tampering by hostile nations, insufficient attention to context and flawed assumptions may have played a role in the failure of most political experts to predict Donald Trump's victory over Hillary Clinton in the 2016 presidential election. Trump's victory ran counter to almost every major forecast, which had predicted Clinton's chances of winning to be between 70 to 99 percent.

Tons of data had been analyzed by political experts and the candidates' campaign teams. Clinton ran an overwhelmingly data-driven campaign, and big data had played a large role in Obama's victories in 2008 and 2012. Clinton's team added to the database the Obama campaigns had built, which connected personal data from traditional sources, such as reports from pollsters and field workers, with other data from social media posts and other online behavior as well as data used to predict consumer behavior. The Clinton team assumed that the same voters who supported Obama would turn out for their candidate, and focused on identifying voters in areas with a likelihood of high voter turnout. However, turnout for Clinton among the key groups who had supported Obama—women, minorities, college graduates, and blue-collar workers—fell short of expectations. (Trump had turned to big data as well, but put more emphasis on tailoring campaign messages to targeted voter groups.)

Political experts were misled into thinking Clinton's victory was assured because some predictive models lacked context in explaining potentially wide margins of error. There were shortcomings in polling, analysis, and interpretation, and analysts did not spend enough time examining how the data used in the predictive models were created. Many polls used in election forecasts underestimated the strength of Trump's support. State polls were inaccurate, perhaps failing to capture Republicans who initially refused to vote for Trump and then changed their minds at the last moment. Polls from Wisconsin shortly before the election had put Clinton well ahead of Trump. Polls are important for election predictions, but they are only one of many sources of data that should be consulted. Predictive models were unable to fully determine who would actually turn out to vote as opposed to how people thought they would vote. Analysts overlooked signs that Trump was forging ahead in the battleground states. Britain had a similar surprise when polls mistakenly predicted the nation would vote in June 2016 to stay in the European Union.

And let's not forget that big data poses some challenges to information security and privacy.

As Chapter 4 pointed out, companies are now aggressively collecting and mining massive data sets on people's shopping habits, incomes, hobbies, residences, and (via mobile devices) movements from place to place. They are using such big data to discover new facts about people, to classify them based on subtle patterns, to flag them as "risks" (for example, loan default risks or health risks), to predict their behavior, and to manipulate them for maximum profit.

When you combine someone's personal information with pieces of data from many different sources, you can infer new facts about that person (such as the fact that they are showing early signs of Parkinson's disease, or are unconsciously drawn toward products that are colored blue or green). If asked, most people might not want to disclose such information, but they might not even know such information about them exists. Privacy experts worry that people will be tagged and suffer adverse consequences without due process, the ability to fight back, or even knowledge that they have been discriminated against.

Sources: Linda Currey Post, "Big Data Helps UK National Health Service Lower Costs, Improve Treatments," *Forbes*, February 7, 2018; Michael Jude, "Data Preparation Is the Key to Big Data Success," *InfoWorld*, February 8, 2018; Rajkumar Venkatesan and Christina Black, "Using Big Data: 3 Reasons It Fails and 4 Ways to Make It Work," University of Virginia Darden School of Business Press Release, February 8, 3018; Ed Burns, "When Predictive Models Are Less Than Presidential," *Business Information*, February 2017; Aaron Timms, "Is Donald Trump's Surprise Win a Failure of Big Data? Not Really," *Fortune*, November 14, 2016; Steve Lohr and Natasha Singer, "The Data Said Clinton Would Win. Why You Shouldn't Have Believed It," *New York Times*, November 10, 2016; Nicole Laskowski and Niel Nikolaisen: "Seven Big Data Problems and How to Avoid Them," *TechTarget Inc.*, 2016; Joseph Stromberg, "Why Google Flu Trends Can't Track the Flu (Yet)," smithsonianmag.com, March 13, 2014; and Gary Marcus and Ernest Davis, "Eight (No, Nine!) Problems With Big Data," *New York Times*, April 6, 2014.

CASE STUDY QUESTIONS

6-13 What business benefits did the organizations described in this case achieve by analyzing and using big data?

6-14 Identify two decisions at the organizations described in this case that were improved by using big data and two decisions that big data did not improve.

6-15 List and describe the limitations to using big data.

6-16 Should all organizations try to collect and analyze big data? Why or why not? What management, organization, and technology issues should be addressed before a company decides to work with big data?

MyLab MIS

Go to the Assignments section of MyLab MIS to complete these writing exercises.

6-17 Identify the five problems of a traditional file environment and explain how a database management system solves them.

6-18 Discuss how the following facilitate the management of big data: Hadoop, in-memory computing, analytic platforms.

Chapter 6 References

Aiken, Peter, Mark Gillenson, Xihui Zhang, and David Rafner. "Data Management and Data Administration: Assessing 25 Years of Practice." *Journal of Database Management* (July-September 2011).

Beath, Cynthia, Irma Becerra-Fernandez, Jeanne Ross, and James Short. "Finding Value in the Information Explosion." *MIT Sloan Management Review* 53, No. 4 (Summer 2012).

Bessens, Bart. "Improving Data Quality Using Data Governance." *Big Data Quarterly* (Spring 2018).

Buff, Anne. "Adapting Governance to the Changing Data Landscape." *Big Data Quarterly* 3, No. 4 (Winter 2017).

Bughin, Jacques, John Livingston, and Sam Marwaha. "Seizing the Potential for Big Data." *McKinsey Quarterly* (October 2011).

Caserta, Joe, and Elliott Cordo. "Data Warehousing in the Era of Big Data." *Big Data Quarterly* (January 19, 2016).

Chai, Sen, and Willy Shih. "Why Big Data Isn't Enough." *MIT Sloan Management Review* (Winter 2017).

Clifford, James, Albert Croker, and Alex Tuzhilin. "On Data Representation and Use in a Temporal Relational DBMS." *Information Systems Research* 7, No. 3 (September 1996).

DalleMule, Landro, and Thomas H. Davenport. "What's Your Data Strategy?" *MIT Sloan Management Review* (Winter 2017).

Davenport, Thomas H. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Boston, MA: Harvard Business School Press (2014).

Devlin, Barry. "The EDW Lives On: The Beating Heart of the Data Lake." *9Sight Consulting* (April 2017).

Eckerson, Wayne W. "Analytics in the Era of Big Data: Exploring a Vast New Ecosystem." *TechTarget* (2012).

Experian Information Solutions. "The 2017 Global Data Management Benchmark Report." (2017).

Henschen, Doug. "MetLife Uses NoSQL for Customer Service Breakthrough." *Information Week* (May 13, 2013).

Hoffer, Jeffrey A., Ramesh Venkataraman, and Heikki Toppi. *Modern Database Management* (12th ed.). Upper Saddle River, NJ: Pearson (2016).

Imhoff, Claudia. "Data Warehouse Appliances and the New World Order of Analytics." *Intelligent Solutions Inc*. (August 2017).

King, Elliot. "Has Data Quality Reached a Turning Point?" *Big Data Quarterly* 3 No. 4 (Winter 2017).

Kroenke, David M., and David Auer. *Database Processing: Fundamentals, Design, and Implementation* (14th ed.). Upper Saddle River, NJ: Pearson (2016).

Lee, Yang W., and Diane M. Strong. "Knowing-Why About Data Processes and Data Quality." *Journal of Management Information Systems* 20, No. 3 (Winter 2004).

Loveman, Gary. "Diamonds in the Datamine." *Harvard Business Review* (May 2003).

Marcus, Gary, and Ernest Davis. "Eight (No, Nine!) Problems with Big Data." *New York Times* (April 6, 2014).

Martens, David, and Foster Provost. "Explaining Data-Driven Document Classifications." *MIS Quarterly* 38, No. 1 (March 2014).

McAfee, Andrew, and Erik Brynjolfsson. "Big Data: The Management Revolution." *Harvard Business Review* (October 2012).

McKendrick, Joe. "Building a Data Lake for the Enterprise." *Big Data Quarterly* (Spring 2018).

McKinsey Global Institute. "Big Data: The Next Frontier for Innovation, Competition, and Productivity." *McKinsey & Company* (2011).

Morrow, Rich. "Apache Hadoop: The Swiss Army Knife of IT." *Global Knowledge* (2013).

Mulani, Narendra. "In-Memory Technology: Keeping Pace with Your Data." *Information Management* (February 27, 2013).

O'Keefe, Kate. "Real Prize in Caesars Fight: Data on Players." *Wall Street Journal* (March 19, 2015).

Redman, Thomas. *Data Driven: Profiting from Your Most Important Business Asset*. Boston: Harvard Business Press (2008).

_____. "Data's Credibility Problem." *Harvard Business Review* (December 2013).

Ross, Jeanne W., Cynthia M. Beath, and Anne Quaadgras. "You May Not Need Big Data After All." *Harvard Business Review* (December 2013).

SAP. "Data Warehousing and the Future." (February 2017).

Shi, Donghui, Jian Guan, Josef Zurada, and Andrew Manikas. "A Data-Mining Approach to Identification of Risk Factors in Safety Management Systems." *Journal of Management Information Systems* 34 No. 4 (2017).

Wallace, David J. "How Caesar's Entertainment Sustains a Data-Driven Culture." *DataInformed* (December 14, 2012).

Zoumpoulis, Spyros, Duncan Simester, and Theos Evgeniou, "Run Field Experiments to Make Sense of Your Big Data." *Harvard Business Review* (November 12, 2015).