nearly 11 percent of the U.S. population was supposed to have had influenza at the flu season's peak in mid-January 2013. However, an article in the science journal Nature stated that Google's results were nearly twice the actual amount estimated by the CDC, which had 6 percent of the population coming down with the disease. Why did this happen? 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. Google's algorithm only looked at number of flu search requests, not the context of the searches.

Big data can also provide a distorted picture of the problem. Boston's Street Bump app uses a smartphone's accelerometer to detect potholes without the need for city workers to patrol the streets. Users of this mobile app collect road condition data while they drive and automatically provide city government with real-time information to fix problems and plan long-term investments. However, what Street Bump actually produces is a map of potholes that favors young, affluent areas where more people own smartphones. The capability to record every road bump or pothole from every enabled phone is not the same as recording every pothole. Data often contain systematic biases, and it takes careful thought to spot and correct for those biases.

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 or manipulated in the marketplace.

*Sources:* Nicole Laskowski and Niel Nikolaisen, "Seven Big Data Problems and How to Avoid Them," TechTarget Inc., 2016; "The Most Innovative Companies of 2016: Top Companies by Sector," www.fastcompany.com, accessed March 4. 2016; Ed Burns, "Big Data Analytics Not Just a Grab and Go Process," *Business Information*, February 2015; Elizabeth Dwoskin, "The Joys and Hype of Software Called Hadoop," *Wall Street Journal*, December 16, 2014; Tim Harford, "Big Data: Are We Making a Big Mistake?" *Financial Times Magazine*, March 28, 2014; Laura Kolodny, "How Consumers Can Use Big Data," *Wall Street Journal*, March 23, 2014; Joseph Stromberg, "Why Google Flu Trends Can't Track the Flu (Yet)," smithsonianmag.com, March 13, 2014; Gary Marcus and Ernest Davis, "Eight (No, Nine!) Problems with Big Data," *New York Times*, April 6, 2014; and Shira Ovide, "Big Data, Big Blunders," *Wall Street Journal*, March 11, 2013.

## CASE STUDY QUESTIONS

**6-13** What business benefits did the companies and services 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 were not improved by using big data.

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

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

MyMISLab

Go to the Assignments section of MyMISLab 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.

will achieve their goal or what questions they are trying to answer. Darian Shirzai, founder of Radius Intelligence Inc., likens this to haystacks without needles. Companies don't know what they're looking for because they think big data alone will solve their problem.

According to Michael Walker of Rose Business Technologies, which helps companies build big data systems, a significant majority of big data projects aren't producing any valuable, actionable results. A recent report from Gartner, Inc. stated that through 2017, 60 percent of big data projects will fail to go beyond piloting and experimentation and will eventually be abandoned. This is especially true for very large-scale big data projects. Companies are often better off starting with smaller projects with narrower goals.

Hadoop has emerged as a major technology for handling big data because it allows distributed processing of large unstructured as well as structured data sets across clusters of inexpensive computers. However, Hadoop is not easy to use, requires a considerable learning curve, and does not always work well for all corporate big data tasks. For example, when Bank of New York Mellon used Hadoop to locate glitches in a trading system, Hadoop worked well on a small scale, but it slowed to a crawl when many employees tried to access it at once. Very few of the company's 13,000 IT specialists had the expertise to troubleshoot this problem. David Gleason, the bank's chief data officer at the time, said he liked Hadoop but felt it still wasn't ready for prime time. According to Gartner, Inc. research director for information management Neil Heudecker, technology originally built to index the web may not be sufficient for corporate big data tasks.

It often takes a lot of work for a company to combine data stored in legacy systems with data stored in Hadoop. Although Hadoop can be much faster than traditional databases for some tasks, it often isn't fast enough to respond to queries immediately or to process incoming data in real time (such as using smartphone location data to generate just-in-time offers).

Hadoop vendors are responding with improvements and enhancements. For example, Hortonworks produced a tool that lets other applications run on top of Hadoop. Other companies are offering tools as Hadoop substitutes. Databricks developed Spark open source software that is more adept than Hadoop at handling real-time data, and the Google spinoff Metanautix is trying to supplant Hadoop entirely.

It is difficult to find enough technical IT specialists with expertise in big data analytical tools, including Hive, Pig, Cassandra, MongoDB, or Hadoop. On top of that, many business managers lack numerical and statistical skills required for finding, manipulating, managing, and interpreting data.

Even with big data expertise, data analysts need some business knowledge of the problem they are trying to solve with big data. For example, if a pharmaceutical company monitoring point-of-sale data in real time sees a spike in aspirin sales in January, it might think that the flu season is intensifying. However, before pouring sales resources into a big campaign and increasing flu medication production, the company would do well to compare sales patterns to past years. People might also be buying aspirin to nurse their hangovers following New Year's Eve parties. In other words, analysts need to know the business and the right questions to ask of the data.

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 U.S. 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.

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. The service has consistently overestimated flu rates when compared to conventional data collected afterward by the U.S. Centers for Disease Control and Prevention (CDC). According to Google Flu Trends,

## 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.

## Can We Trust 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.

Big data helps streaming music service Spotify create a service that feels personal to each of its 75 million global users. Spotify uses the big data it collects on user listening habits (more than 600 gigabytes daily) to design highly individualized products that captivate its users around a particular mood or moment in time rather than offering the same tired genres. Users can constantly enhance their listening experience with data-driven features such as the Discovery tool for new music, a Running tool that curates music timed to the beat of their workout, and Taste Rewind—which tells what they would have listened to in the past by analyzing what they listen to now. By constantly using big data to fine-tune its services, Spotify hopes to create the perfect user experience.

A number of services have emerged to analyze big data to help consumers. There are now online services to enable consumers to check thousands of different flight and hotel options and book their own reservations, tasks previously handled by travel agents. For instance, a mobile app from Skyscanner shows deals from all over the web in one list—sorted by price, duration, or airline—so travelers don't have to scour multiple sites to book within their budget. Skyscanner uses information from more than 300 airlines, travel agents, and timetables and shapes the data into at-a-glance formats with algorithms to keep pricing current and make predictions about who will have the best deal for a given market.

Big data is also providing benefits in law enforcement (see this chapter's Interactive Session on Organizations), sports, education, science, and health care. A recent McKinsey Global Institute report estimated that the U.S. healthcare system could save $300 billion each year—$1,000 per American—through better integration and analysis of the data produced by everything from clinical trials to health insurance transactions to "smart" running shoes. 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. Swimming in numbers and other data 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 and analyzing mountains of data without no clear objective or understanding of exactly how analyzing big data

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 MyMISLab, 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 MyMISLab, 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.

## 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** MyMISLab It has been said there is no bad data, just bad management. Discuss the implications of this statement.

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

**6-7** MyMISLab 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 MyMISLab's Multimedia Library 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

**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 webpages. 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, 231*
*Attribute, 215*
*Big data, 227*
*Bit, 214*
*Byte, 214*
*Data administration, 237*
*Data cleansing, 238*
*Data definition, 220*
*Data dictionary, 220*
*Data governance, 237*
*Data inconsistency, 216*
*Data manipulation language, 220*
*Data mart, 229*
*Data mining, 233*
*Data quality audit, 238*
*Data redundancy, 216*
*Data warehouse, 227*
*Database, 217*
*Database administration, 237*
*Database management system (DBMS), 217*
*Database server, 235*
*Entity, 215*
*Entity-relationship diagram, 224*
*Field, 214*
*File, 214*
*Foreign key, 219*
*Hadoop, 230*
*In-memory computing, 230*
*Information policy, 237*
*Key field, 219*
*Non-relational database management systems, 225*
*Normalization, 223*
*Online analytical processing (OLAP), 232*
*Primary key, 219*
*Program-data dependence, 216*
*Record, 214*
*Referential integrity, 224*
*Relational DBMS, 218*
*Sentiment analysis, 234*
*Structured Query Language (SQL), 220*
*Text mining, 234*
*Tuple, 219*
*Web mining, 234*

MyMISLab

To complete the problems marked with the MyMISLab, go to EOC Discussion Questions in MyMISLab.

business processes. Keurig believes its work building a fully automated data governance structure where the entire company is aligned to an enterprise-wide data strategy and standards won't be completed until 2020.

*Sources:* Ken Murphy, "Keurig Green Mountain Brews Up Data Governance," SAP Insider Profiles, January 2016; www.sap.com, accessed March 5, 2016; www.datumstrategy.com, accessed March 5, 2016; and Keurig Green Mountain Inc. Form 10-K, November 19, 2015.

## CASE STUDY QUESTIONS

1. Discuss the role of data governance at Keurig Green Mountain.
2. What management, organization, and technology issues had to be addressed in order to establish enterprise-wide data governance?
3. What were the business benefits of data governance for Keurig Green Mountain?
4. How did data governance improve operations and management decision making?

# 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 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.

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.

## INTERACTIVE SESSION: MANAGEMENT

### Keurig Green Mountain Embraces Data Governance

More than 25 percent of all coffee consumed in the United States today comes from Keurig Green Mountain single-serve K-Cups. Keurig Green Mountain, headquartered in Waterbury, Vermont, has expanded so rapidly over the past decade that it has 80 brands and nearly 600 product varieties of hot and cold coffees (including Green Mountain), teas, cocoas, dairy-based beverages, cider, and fruit-based drinks. It also partners with other vendors such as Dunkin Donuts, Newman's Own, and Starbucks to package and sell their products in K-Cup pods. The company has more than 6,600 employees and generated nearly $4.4 billion in revenue in 2015. A business this large and complex must maintain a vast amount of data.

Keurig Green Mountain's meteoric growth called for a better approach to managing those data. The company had relied on what it called a "hero culture" for data governance. Different groups in charge of providing data would set up the data they were responsible for, such as customer records, vendor records, or material master data. The department receiving the data would correct any inaccuracies to make sure that the right products were produced, orders were placed, and items were shipped. The data providers were called "heroes" because their work was of such high value to the company. This way of working sufficed before Keurig's growth spurt. However, because the data corrections that were made downstream were not always conveyed to the data providers, the process was not repeatable and corrections might have to be made again the next time the data were used. This added to the time and cost to conduct business. Additionally, having different groups of "heroes" fix the data only for their specific business processes meant that managing data from a companywide standpoint was limited.

By 2013 Keurig Green Mountain had outgrown its legacy ERP system and switched to SAP ERP. This gave the entire company the opportunity to review how it was managing data and to take the necessary steps toward master data management, well-defined processes, standards for the maintenance of data across the organization, and comprehensive data cleansing. Master data management (MDM) is the organizational effort to create one single master reference source for all critical business data, leading to fewer errors and less redundancy in business processes. By providing one point of reference for critical information, MDM eliminates costly redundancies that occur when organizations rely upon multiple conflicting sources of data.

Keurig Green Mountain enlisted DATUM LLC to help it establish a strong data governance framework. This was necessary to ensure that as the company's volume of data increased, it wouldn't return to disparate data management practices that would negate the efficiencies and benefits of the SAP ERP software. DATUM LLC is an information management solutions company based in Annapolis, Maryland, that provides data governance software and consulting services. Its Information Value Management SaaS (software as a service) translates business objectives into functional designs that improve quality and processing speed.

DATUM supplies expertise on defining data-centric processes, information value, and accountability. By incorporating best practices, Information Value Management provides a framework of data standards, governance rules, business metrics, and business processes that is useful for analytics, business intelligence, data governance, process improvement, performance management, ERP implementation, and managing big data.

Information Value Management (IVM) can be integrated with SAP Information Steward software, which provides a single environment to discover, assess, define, monitor, and improve the quality of enterprise data assets. Information Steward's functionality includes modules for discovering data characteristics and relationships, creating and running data validation rules, identifying bad data and improving data quality, cataloging data, defining business terms for data and organizing the terms into categories, and data cleansing tools. Information Steward helps ensure companywide reporting consistency so that the company's data stewards can easily monitor hanging data and make sure these changes are reflected in the organization's master data management. IVM can also be used with other SAP solutions for enterprise information management (EIM), including data quality assurance, master data management, content management, and information lifecycle management.

Keurig Green Mountain has used Information Steward to implement data quality reports. As data are collected, the tool alerts users to missing required fields. This capability lessens the need for a "hero culture," repeated errors, and repeated fixes to data downstream. Better data quality leads to more informed business decisions, and users of company data will have more trust in the data behind their

*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 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.

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.

The Interactive Session on Management illustrates Keurig Green Mountain's experience with managing data as a resource. As you read this case, try to identify the policies, procedures, and technologies that were required to improve data management at this company.

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

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 6.14 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 700,000 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 1.6 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.

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 webpages, 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.

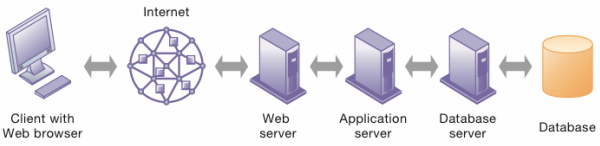
## Databases and the Web

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 6.14 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 working with the database. In a client/server environment, the DBMS resides on a dedicated computer called a **database server**. The DBMS receives the

**FIGURE 6.14 LINKING INTERNAL DATABASES TO THE WEB**




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

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. E-mail, 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 big data sets, 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 e-mail message, blog, social media conversation, or survey form to detect favorable and unfavorable opinions about specific subjects.

For example, the discount broker Charles Schwab uses Attensity Analyze software to analyze hundreds of thousands of its customer interactions each month. The software analyzes Schwab's customer service notes, e-mails, survey responses, and online discussions to discover signs of dissatisfaction that might cause a customer to stop using the company's services. Attensity is able to automatically identify the various "voices" customers use to express their feedback (such as a positive, negative, or conditional voice) to pinpoint a person's intent to buy, intent to leave, or reaction to a specific product or marketing message. Schwab uses this information to take corrective actions such as stepping up direct broker communication with the customer and trying to quickly resolve the problems that are making the customer unhappy.

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

Figure 6.13 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. 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 2016?" 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

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

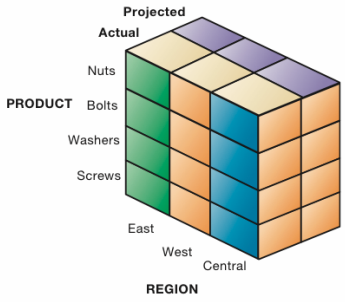
Once data have been captured and organized using the business intelligence technologies we have just described, they are available for further analysis using software for database querying and reporting, multidimensional data analysis (OLAP), and data mining. This section will introduce you to these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

### 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 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 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 6.13** 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.

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.

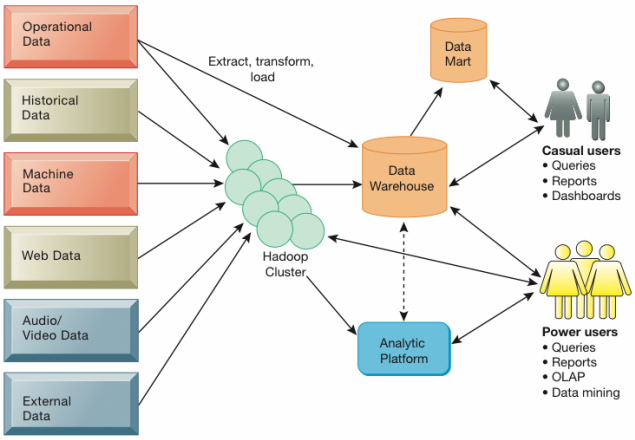
Leading commercial products for in-memory computing include SAP HANA and Oracle Exalytics. Each provides a set of integrated software components, including in-memory database software and specialized analytics software, that run on hardware optimized for in-memory computing work.

## 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 such as IBM PureData System for Analytics, feature preconfigured hardware-software systems that are specifically designed for query processing and analytics. For example, 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. Analytic platforms also include in-memory systems and NoSQL non-relational database management systems. Analytic platforms are now available as cloud services.

Figure 6.12 illustrates a contemporary business intelligence infrastructure using the technologies we have just described. Current and historical data are extracted from multiple operational systems along with web data, machine-generated data, unstructured audio/visual data, and data from external sources that have been restructured and reorganized for reporting and analysis. Hadoop clusters pre-process big data for use in the data warehouse, data marts, or an

**FIGURE 6.12 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.

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 sub-problems, 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

19 times but only for a misdemeanor charge and never served more than five months in jail. When flagged by CSU after his arrest in July 2010, he was convicted of felony robbery and sentenced to three and a half to seven years in prison.

Information developed by CSU helped Vance's Violent Criminal Enterprises Unit break up the most violent of Manhattan's 30 gangs. Since 2011, 17 gangs have been dismantled. According to New York's chief assistant district attorney Karen Friedman Agnifilo, murders dropped from 70 in 2010 to 29 in 2013 because the DA's office and police now had the information to identify the people driving crime in Manhattan and to take these people off the streets and put them behind bars.

There's another side to this, however. When prosecutors begin to compile databases for data-driven crime fighting, one needs to ask what people have been selected for inclusion in these databases, what are the selection criteria, and how harmful is this practice. Could the criminal justice databases include people who really shouldn't be there and nevertheless are targets for police scrutiny? According to Steven Zeidman, director of the criminal-defense clinic at the City University of New York (CUNY) School of Law, the answer is yes. More than 1,000 people are arrested in New York City each day. An overwhelming and disproportionate number are people of color arrested for minor offenses like riding a bicycle on the sidewalk or jaywalking. Zeidman recalled a time when he was in court with a teenager arrested for jaywalking. The arresting officer said he had stopped the young man because he was wearing a red shirt that was known to be a gang color. The young man was not a gang member, but he's probably in the database. People with arrest and conviction records find it next to impossible to find legitimate work on release, and this result lasts for as long as the records are retained.

*Sources:* Pervaiz Shallwani and Mark Morales, "NYC Officials Tout New Low in Crime, but Homicide, Rape, Robbery Rose," *Wall Street Journal*, January 4, 2016; "Prosecution Gets Smart" and "Intelligence-Driven Prosecution/Crime Strategies Unit," www.manhattanda.org, accessed March 4, 2016; and Chip Brown, "The Data D.A.", *New York Times Magazine*, December 7, 2014.

## CASE STUDY QUESTIONS

1. What are the benefits of intelligence-driven prosecution for crime fighters and the general public?
2. What problems does this approach to crime fighting pose?
3. What management, organization, and technology issues should be considered when setting up information systems for intelligence-driven prosecution?

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

## INTERACTIVE SESSION: ORGANIZATIONS

### New York City Moves To Data-Driven Crime Fighting

Nowhere have declining crime rates been as dramatic as in New York City. As reflected in the reported rates of the most serious types of crime, the city in 2015 was as safe as it had been since statistics have been kept. Crimes during the preceding few years have also been historically low.

Why is this happening? Experts point to a number of factors, including demographic trends, the proliferation of surveillance cameras, and increased incarceration rates. But New York City would also argue it is because of its proactive crime prevention program along with district attorney and police force willingness to aggressively deploy information technology.

Cyrus Vance Jr., New York County's district attorney, is vigorously trying to mine more crime-fighting information from the data collected by the city to drive crime rates even lower. He believes that New York could get crime rates to zero—if one looked harder at the data.

There has been a revolution in the use of big data for retailing and sports (think baseball and *Moneyball*) as well as for police work. New York City has been at the forefront in intensively using data for crime fighting, and its CompStat crime-mapping program has been replicated by other cities.

CompStat features a comprehensive, citywide database that records all reported crimes or complaints, arrests, and summonses in each of the city's 76 precincts, including their time and location. The CompStat system analyzes the data and produces a weekly report 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 and NYPD's senior leadership quickly identify patterns and trends and develop a targeted strategy for fighting crime, such as dispatching more foot patrols to high-crime neighborhoods.

Vance and his team think there is much more that can be done with data to reduce crime. Dealing with more than 105,000 cases per year in Manhattan, New York's assistant district attorneys did not have enough information to make fine-grained decisions about charges, bail, pleas, or sentences. They couldn't quickly separate minor delinquents from serious offenders.

In 2010 Vance's team created a Crime Strategies Unit (CSU) to identify and address crime issues and target priority offenders for aggressive prosecution. Rather than information being left on thousands of legal pads in the offices of hundreds of assistant district attorneys, CSU gathers and maps crime data for Manhattan's 22 precincts to visually depict criminal activity based on multiple identifiers such as gang affiliation and type of crime. Police commanders supply a list of each precinct's 25 worst offenders, which is added to a searchable database that now includes more than 9,000 chronic offenders. A large percentage are recidivists who have been repeatedly convicted of grand larceny, active gang members, and other priority targets. These are the people law enforcement wants to know about if they are arrested.

This database is used for an arrest alert system. When someone considered a priority defendant is picked up (even on a minor charge or parole violation) or arrested in another borough of the city, any interested prosecutor, parole officer, or police intelligence officer is automatically sent a detailed e-mail. The system can use the database to send arrest alerts for a particular defendant, a particular gang, or a particular neighborhood or housing project, and the database can be sorted to highlight patterns of crime ranging from bicycle theft to homicide.

The alert system helps assistant district attorneys ensure that charging decisions, bail applications, and sentencing recommendations address that defendant's impact on criminal activity in the community. The information gathered by CSU and disseminated through the arrest alert system differentiates among those for whom incarceration is an imperative from a community-safety standpoint and those defendants for whom alternatives to incarceration are appropriate and will not negatively affect overall community safety. If someone leaves a gang, goes to prison for a long time, moves out of the city or New York state, or dies, the data in the arrest alert system are edited accordingly.

In speeches praising intelligence-driven prosecution, Vance often cites the example of a 270-pound scam artist who for more than a decade made a living by bumping into pedestrians in the Times Square area and demanding money, claiming they had broken his glasses. He had been convicted

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 doesn't refer to any specific quantity but usually refers to data in the petabyte and exabyte range—in other words, billions to trillions of records, all 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. Even though "tweets" are limited to 140 characters each, 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. The chapter-opening case on the U.S. Postal Service is one example, as are city governments using big data to manage traffic flows and fight crime. The Interactive Session on Organizations describes how New York City is using big data to lower its crime rate.

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

For example, MetLife used the MongoDB open source NoSQL database to quickly integrate disparate data on more than 100 million customers and deliver a consolidated view of each. MetLife's 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 is able to use structured, semi-structured, and unstructured information without requiring tedious, expensive, and time-consuming database mapping.

### Cloud Databases

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

In addition to public cloud-based data management services, companies now have the option of using databases in private clouds. For example, Sabre Holdings, the world's largest software as a service (SaaS) provider for the aviation industry, has a private database cloud that supports more than 100 projects and 700 users. A consolidated database spanning a pool of standardized servers running Oracle Database provides database services for multiple applications.

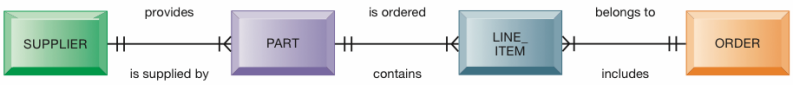
## 6-3 What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?

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, e-mail 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

**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.

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.

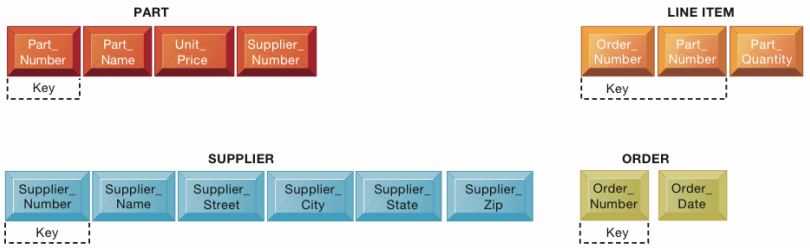
For example, Famous Footwear, a shoe store chain with more than 800 locations in 49 states, could not achieve its goal of having "the right style of shoe in the right store for sale at the right price" because its database was not properly designed for rapidly adjusting store inventory. The company had 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 and Databases in the Cloud

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.

**FIGURE 6.10** NORMALIZED TABLES CREATED FROM ORDER



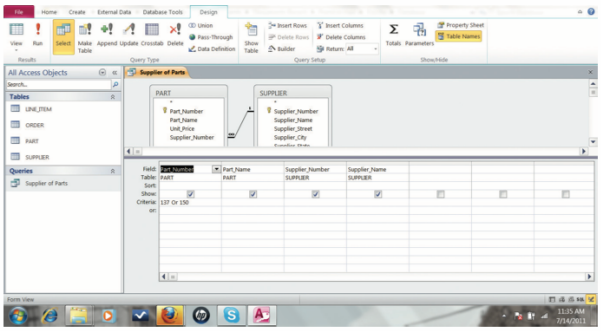

After normalization, the original relation ORDER has been broken down into four smaller relations. The relation ORDER is left with only two attributes, and the relation LINE_ITEM has a combined, or concatenated, key consisting of Order_Number and Part_Number.

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.

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

**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.

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.

## Normalization and Entity-Relationship Diagrams

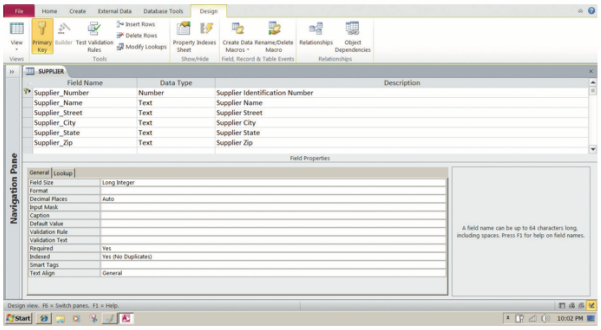
The conceptual database design describes how the data elements in the database are to be grouped. The design process identifies relationships among data elements and the most efficient way of grouping data elements together to meet business information requirements. The process also identifies redundant 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.

**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.

**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.

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

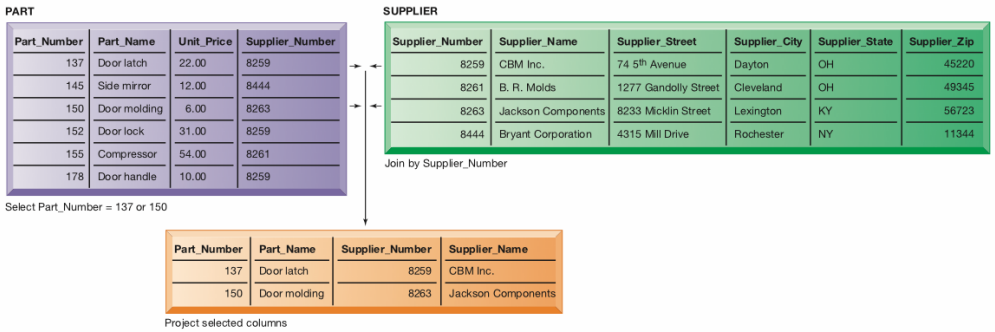
To create a database, you must understand the relationships among the data, the type of data that will be maintained in the database, how the data will be used, and how the organization will need to change to manage data from a

**FIGURE 6.7 EXAMPLE OF AN SQL QUERY**

```
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;
```

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.

**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.

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* 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.

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.

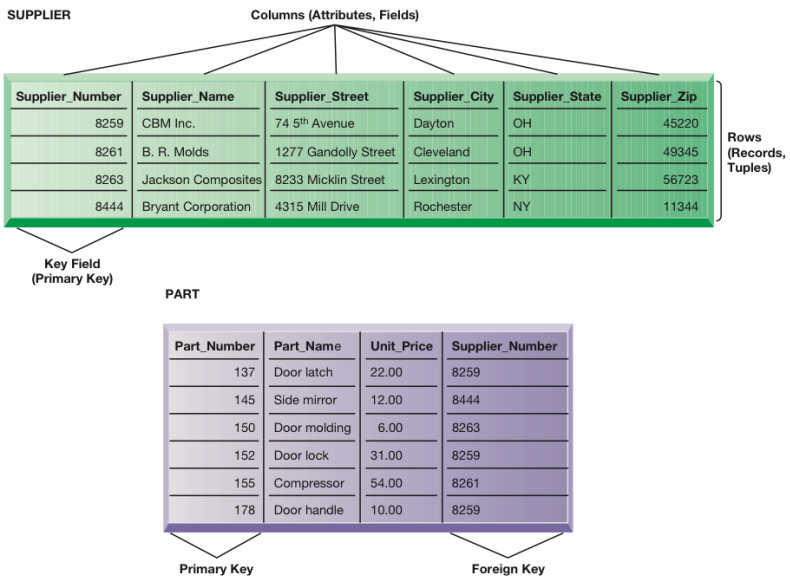
Microsoft Access has a rudimentary data dictionary capability that displays information about the name, description, size, type, format, and other properties of each field in a table (see Figure 6.6). Data dictionaries for large corporate databases may capture additional information, such as usage, ownership (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 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.

**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.

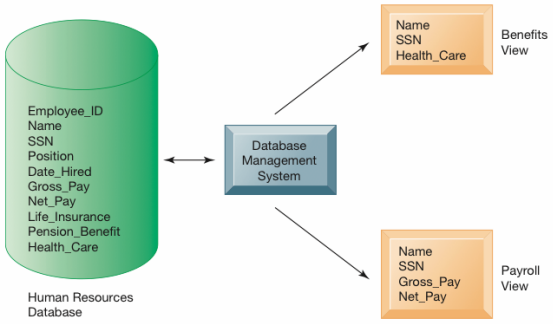
The actual information about a single supplier that resides in a table is called a row. Rows are commonly referred to as records, or in very technical terms, as **tuples**. Data for the entity PART have their own separate table.

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

**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.

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.

freely across different functional areas or different parts of the organization. If users find different values of 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.

### Database Management Systems

A **database management system (DBMS)** is software that permits 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

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, may 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 might 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 may 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

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.

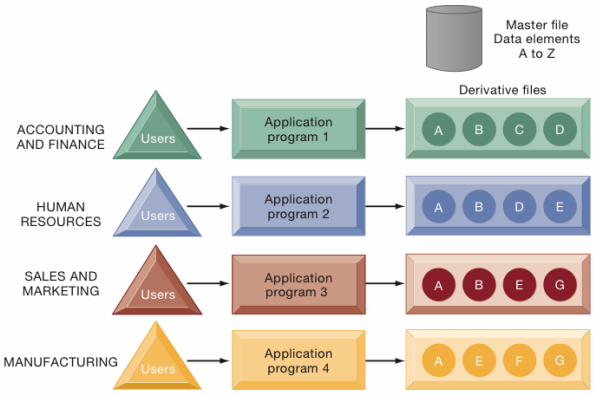
A record describes an entity. An **entity** is a person, place, thing, or event on which we store and maintain information. Each characteristic or quality describing a particular entity is called an **attribute**. For example, Student_ID, Course, Date, and Grade are attributes of the entity COURSE. The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

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

**FIGURE 6.2 TRADITIONAL FILE PROCESSING**



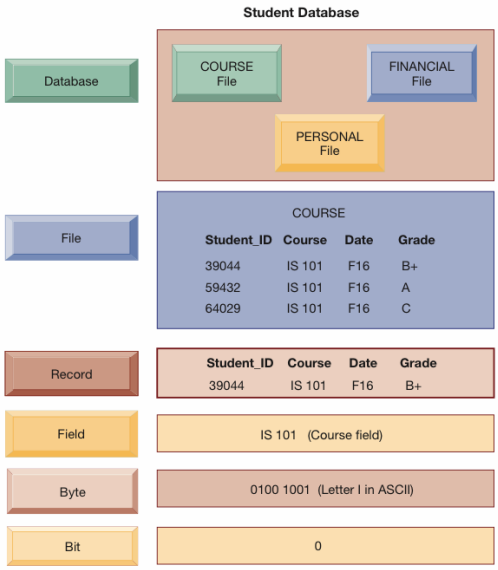

The use of a traditional approach to file processing encourages each functional area in a corporation to develop specialized applications. Each application requires a unique data file that is likely to be a subset of the master file. These subsets of the master file lead to data redundancy and inconsistency,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

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.

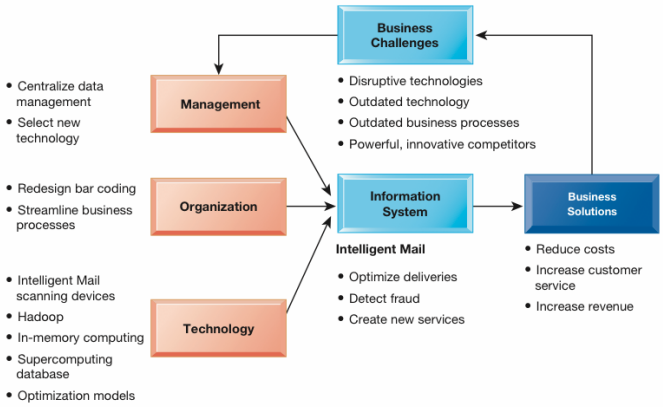
## File Organization Terms and Concepts

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.




The chapter-opening diagram calls attention to important points raised by this case. The USPS is struggling to stay alive and relevant in the face of disruptive technologies such as e-mail and messaging and powerful, innovative competitors. USPS management believes the solution is to use better information to drive its operational activities and decisions. USPS now consolidates data collected from packages and letters scanned during the delivery process into a large central database where the data can be more easily accessed and analyzed to improve operations as well as management reporting and decision making. Technologies such as in-memory computing tools and Apache Hadoop for distributed processing and storage of large data sets made it much easier for the USPS to rapidly access and analyze its vast stores of data and turn the data into valuable information. In order to achieve more value from automated scanning and processing technologies as well as information from the database, the USPS had to redesign jobs, business processes, and workflows The solution increased efficiency, improved customer service, reduced costs, and enabled the USPS to offer new services such as Sunday package delivery for Amazon. However, other commercial possibilities for wringing even more value out of USPS data by sharing the data with retailers may be limited by privacy restrictions.

Here are some questions to think about: What was the business impact of not making maximum use of data at the USPS? How did better use of data collected by the USPS improve operational efficiency and management decision making? How much will analyzing the data it collects help the USPS survive?

## 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.

By carefully tracking how mail moves around the country, from the moment a delivery vehicle arrives at a dock to the second a letter is delivered to its recipient, the Postal Service has built sophisticated computer models that map the most efficient and cost-effective mail delivery routes. The USPS is able to access and analyze its vast quantity of data very quickly because it uses in-memory computing tools that store the data in computer memory and the Apache Hadoop software framework to store and process large data sets in a distributed computing environment.

Another use for USPS big databases is to detect fraud in more than 528 million mail pieces each day. When a piece of mail is scanned at a post office facility, relevant data about the carrier, route, weight, and size are transmitted to the USPS supercomputing database in Eagan, Minnesota. The data from each mail piece are then compared to some 400 billion records in the database. Complex algorithms perform fraud detection and other tests on the data before transmission back to the delivery center. All of this happens in an average of 50 to 100 thousandths of a second. If something is not right, such as a package with insufficient, duplicated, or fraudulent postage, the problem is detected in near real time. Errors are tracked down, and fraud attempts are reported to the U.S. Postal Inspection Service for further investigation. With annual revenue of $65 billion, the USPS saves many millions per year by analyzing the data.

That's not all. Cochrane believes the USPS could use IMB data to help retailers and catalog companies drive more sales. For instance, the USPS could send a clothing retailer an e-mail or text message that a particular customer in Omaha, Nebraska, has just received the company's holiday catalog. Upon receiving this real-time alert, the retailer could immediately e-mail the customer a digital coupon or promotional offer. A great idea, but the USPS must adhere to privacy statutes binding U.S. government agencies. Although the service has data on every single piece of mail exchanged among hundreds of millions of Americans and the companies that sell to them, it is not supposed to use data for any purpose other than delivering the mail effectively. The USPS may or may not be able to use its data for other commercial purposes.

In November 2013, the USPS signed a deal with Amazon to deliver packages on Sundays in select cities, increasing its share of the profitable package-delivery market. USPS's package revenue increased 8 percent to $12.5 billion from 2012 to 2013. UPS and FedEx, watch out!

**Sources:** "Intelligent Mail" and "U.S. Postal Service 2015 Annual Report to Congress," www.usps.gov, accessed March 3, 2016; Derek Major, "USPS Plans for the Internet of Mailed Things," *Government Computer News*, June 15, 2015; Cindy Waxer, "Modernizing the Mail," *Computerworld*, December 2014; Federico Guerrini, "How Big Data and the Internet of Things Will Change the Postal Service," *Forbes*, July 3, 2014.

The experience of the U.S. Postal Service illustrates the importance of data management. Business performance depends on what an organization can or cannot do with its data. The U.S. Postal Service is a huge, sprawling government agency where insufficient data, along with antiquated equipment and business processes, affected both operational efficiency and management decision making. How businesses store, organize, and manage their data has an enormous impact on organizational effectiveness.

## Better Data Management Helps the U.S. Postal Service Rebound

The U.S. Postal Service (USPS) delivers 154 billion pieces of mail per year. It is one of the oldest government agencies in the United States, but now its future is uncertain. With more people using e-mail, messaging systems, and social networking instead of "snail mail," the volume of first-class mail has plummeted. Private-sector rivals like UPS and FedEx have been luring customers away from the USPS's package delivery services, while tech giants such as Amazon and Google plan innovative new services such as 24/7 delivery lockers, weekend pickup points, and future drone deliveries. The USPS must compete using 20-year-old delivery vehicles and aging parcel- and letter-sorting systems.

© Anton Novik/Shutterstock

It is no surprise, then, that the Postal Service handled nearly 1.4 billion fewer pieces of mail in 2015 than the year before and lost $5 billion. Despite aggressive cost-cutting measures like closing processing centers and slashing employee hours, 2015 marked the Postal Service's ninth consecutive year of losses. Can the ailing delivery service overcome its challenges?

James Cochrane, the USPS CIO, thinks it can, if it pays more attention to its data. The USPS collects a wealth of data as part of its daily operations. Cochrane wants to redesign the USPS mail tracking system to provide more useful information for operations and decision making. The USPS Intelligent Mail bar code (IMB) system uses Intelligent Mail scanning devices and more than 8,500 pieces of automated processing and sorting equipment to scan bar codes for data that are then transmitted to a central database. Bar codes for USPS letters and parcels contain as much data as possible, ranging from the type of mail being delivered to a parcel's final destination. The system is able to gather data from more than a billion tracking events each day. The IMB system provides more detailed and precise information about the mail stream, enabling the USPS to better manage cycle times, predict mail volume, and increase efficiencies at postal processing facilities and delivery routes.

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?

MyMISLab™

Visit **mymislab.com** for simulations, tutorials, and end-of-chapter problems.

## CHAPTER CASES

Better Data Management Helps the U.S. Postal Service Rebound
New York City Moves to Data-Driven Crime Fighting
Keurig Green Mountain Embraces Data Governance
Can We Trust 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