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 is a database, and how does a relational database organize data?

6-2 What are the principles of a database management system?

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 data governance and data quality assurance essential for managing the firm's data resources?

6-5 How will MIS help my career?

CHAPTER CASES

- Astro: Leveraging Data for Customer-driven Service
- The Paradise Papers and Big Data Journalism
- DEWA: Evolving Utilities for a Smart City
- Does Big Data Provide the Answer?

VIDEO CASES

- Dubuque Uses Cloud Computing and Sensors to Build a Smarter City
- Brooks Brothers Closes in on Omnichannel Retail
- Maruti Suzuki Business Intelligence and Enterprise Databases

MyLab MIS

- Discussion Questions: 6-5, 6-6, 6-7
- Hands-On MIS Projects: 6-8, 6-9, 6-10, 6-11

ASTRO: LEVERAGING DATA FOR CUSTOMER-DRIVEN SERVICE

Astro is the leading Malaysian satellite television broadcaster, with an active subscriber list of nearly 5 million, or about 71 percent of Malaysian households. The growth and success of the company has been based on satisfying its customers' taste for Western, Indian, and Korean programming across a wide portfolio of services, including IPTV, broadband, and other streaming services.

However, a number of local and international firms are beginning to pose a threat to Astro's market share. Locally, the line between infrastructure and service providers has blurred as Astro ventures into the broadband market, while major telecommunications providers such as Maxis and TeleKom Malaysia have expanded their own portfolios to include television and on-demand services. Additionally, Astro is increasingly under pressure from regional and international media service providers such as iFlix and, more recently, the U.S.-based Netflix.

To address these challenges, Astro turned to its data to better know and understand its customers. The firm has multiple touchpoints from which it obtains customer data and can secure continuous feedback. These include customer interactions via self-service platforms, on-demand services and libraries, and pay-per-view content, all of which provide data that Astro can use to develop insights into customer behavior.

The challenge for Astro lay in its existing infrastructure and legacy practices. The organization's data were housed at multiple locations, across various sites and on the cloud. Structured data from its PayTV services were stored on a centralized enterprise data warehouse, while unstructured data from other digital products were spread across multiple cloud-based systems. This practice reduced the visibility of data across the organization, necessitating a change in its operational model.

As part of a RM237-million investment in technology infrastructure, Astro chose to shed the traditional enterprise data warehouse paradigm and fully adopt a cloud base data lake. The Astro Data Lake is built using Amazon Web Services (AWS) and aggregates several sources of data, including viewing information, transactions, and interactions across all touchpoints and platforms. This includes data from the firm's extensive library of over 23,000 on-demand titles hosted using Amazon Simple Storage Service (S3), with the processing and computation facilitated through Amazon Elastic Compute Cloud (EC2), delivering 2 petabytes of content every month. Amazon's AWS cloud services provides a scalable solution for data-intensive firms like Astro with a number of features: relatively easy deployment of a data lake, with little to no backend server administration through AWS Lambda; robust search functionality and user authentication via Amazon Elasticsearch and Amazon Cognito, respectively; and data transformation and analytics through the use of AWS Glue and Amazon Athena. These

© sauvignon/123rf

services are powered by Amazon's DynamoDB document database, which provides Astro with real-time performance with minimal latency.

The creation of the Astro Data Lake has enabled the organization to integrate the silos of analytics across the various divisions of the organization, thereby introducing enterprise-wide best practices and standards on data quality. In particular, the data lake architecture is helping to facilitate precise multi-channel advertising, individual-curated recommendations, and an in-depth understanding of individual media consumption preferences, behavior, and sentiment.

Sources: Astro, "Transforming Our Technology Landscape to Deliver Exceptional Customer Experiences," July 7 2017, Astro.com; Digital News Asia, "Astro Accelerates Digital, Business Transformation With Amazon Web Services," April 19, 2017; Goh Thean Eu, "5 Interesting Takeaways from Astro's Q1 2016 Results," Digital News Asia, June 15, 2016; "Astro Inks Deal with Amazon Web Services as Part of Its Digital Strategy," Marketing-Interactive.com, April 18, 2017; Srishti Deoras, "Now Amazon Is a Front Runner in the Lucrative Data Lake Market," *Analytics India Magazine.*

Case contributed by Imran Medi, Asia Pacific University of Technology and Innovation

Astro's centralization of its data illustrates the importance of data management, particularly when an organization is faced with multiple sources of data across several platforms. For most organizations, one of the key operating challenges of the twenty-first century is effectively leveraging information technology to capture, process, and analyze multivariate data in the wake of increased competition. This is exactly the situation Astro found itself in as a digital data-driven organization.

The chapter-opening diagram highlights the important considerations touched upon in the case and in particular draws attention to the socio-technical considerations involved in refactoring the technological infrastructure of Astro. This has involved integrating existing data from multiple sources and repositories with the continuous streams of data acquired across a variety of platforms into a single cloud-based data lake accessible to all the organization's divisions.

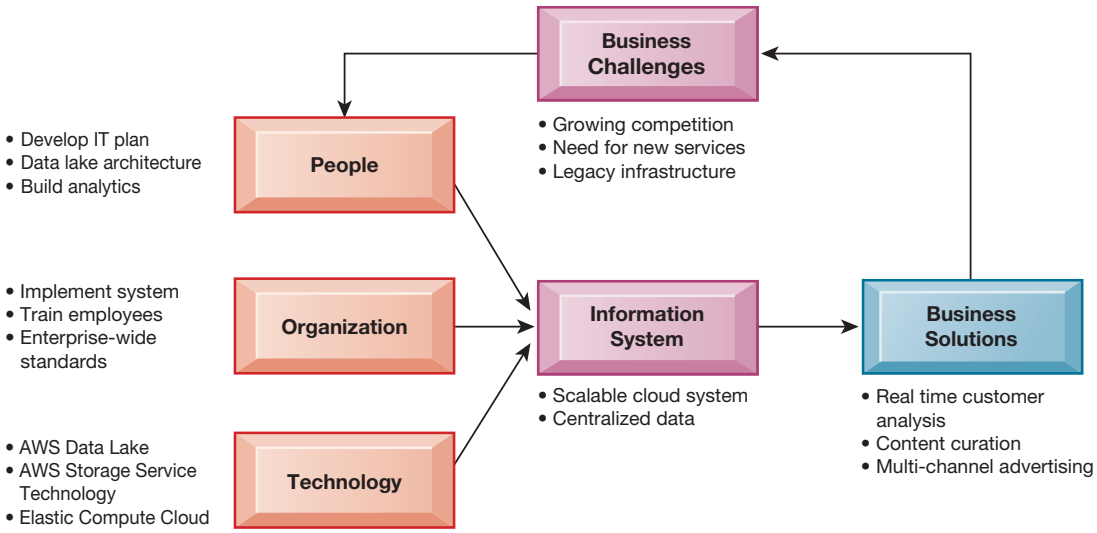
Part of this change involves training and upskilling of employees as well as reengineering business processes to align with the capabilities of a centralized repository. Employees need to be educated about cloud security and agile operations to fully exploit the automation potential of the cloud environment. This has required extensive organization-wide training in the use of various AWS provisions, data analytics, and security operations.

The ultimate aim of the organization is improved decision-making predicated upon a complete understanding of the customers. While various divisions had previously been engaged in some form of data analytics, this had been based upon siloed data separated from other organizational repositories.

Here are some questions to reflect upon: What considerations should a firm have when moving to a cloud based infrastructure such as Astro's? Why is it necessary to review current business processes? How will data analytics benefit Astro, and how will AI services such as Amazon's Athena facilitate this?

6-1 What is a database, and how does a relational database organize data?

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 a student's identification number (ID), the course taken, the date, and the grade, comprises a **record**; a group of records of the same type is called a **file**. 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. Databases are at the heart of all information systems because they keep track of the people, places, and things that a business must deal with on a continuing, often instant basis.

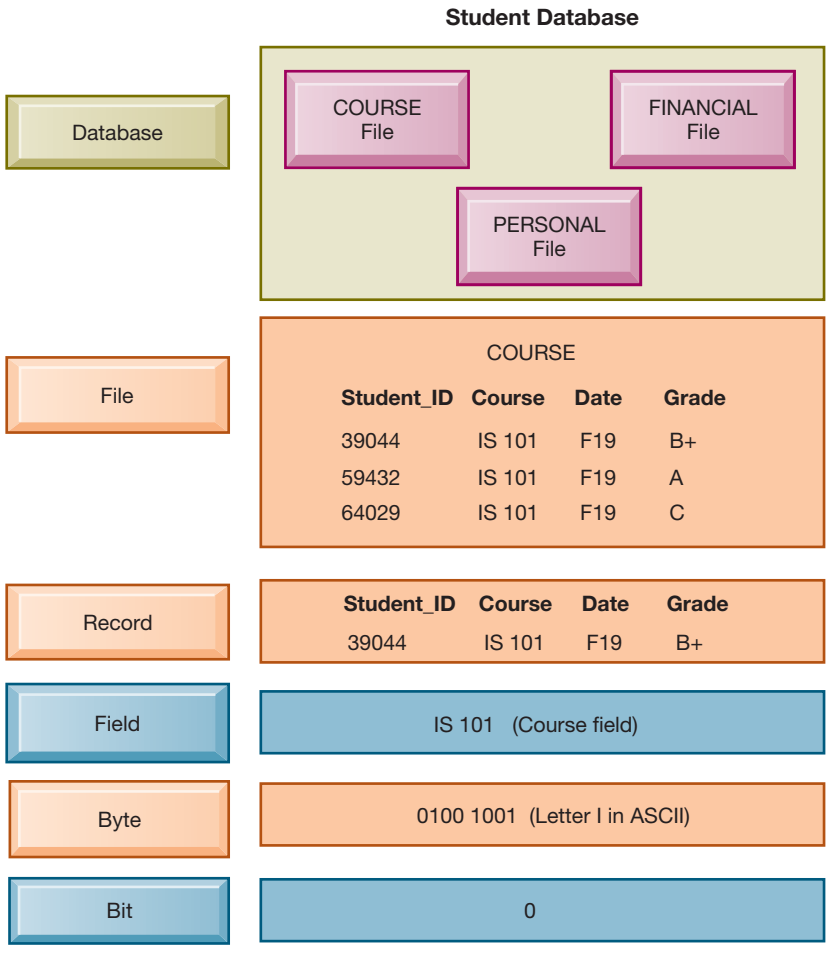


Figure 6.1
The Data Hierarchy.
A computer system organizes data in a hierarchy that starts with the bit, which represents either a 0 or a 1. Bits can be grouped to form a byte to represent one character, number, or symbol. Bytes can be grouped to form a field, and related fields can be grouped to form a record. Related records can be collected to form a file, and related files can be organized into a database.

ENTITIES AND ATTRIBUTES

To run a business, you most likely will be using data about categories of information such as customers, suppliers, employees, orders, products, shippers, and perhaps parts. Each of these generalized categories representing a person, place, or thing on which we store information is called an **entity**. Each entity has specific characteristics called **attributes**. For example, in Figure 6.1, COURSE would be an entity, and Student_ID, Course, Date, and Grade would be its attributes. If you were a business keeping track of parts you used and their suppliers, the entity SUPPLIER would have attributes such as the supplier's name and address, which would most likely include the street, city, state, and zip code. The entity PART would typically have attributes such as part description, price of each part (unit price), and the supplier who produced the part.

ORGANIZING DATA IN A RELATIONAL DATABASE

If you stored this information in paper files, you would probably have a file on each entity and its attributes. In an information system, a database organizes the data much the same way, grouping related pieces of data. The **relational database** is the most common type of database today. Relational databases organize data into two-dimensional tables (called *relations*) with columns and rows. Each table contains data about an entity and its attributes. For the most part, there is one table for each business entity, so, at the most basic level, you will have one table for customers and a table each for suppliers, parts in inventory, employees, and sales transactions.

Let's look at how a relational database would organize data about suppliers and parts. Look at the SUPPLIER table illustrated in Figure 6.2. It consists of a grid of columns and rows of data. Each element of data about a supplier, such as the supplier name, street, city, state, and zip code, is stored as a separate field within the SUPPLIER table. Each field represents an attribute for the entity SUPPLIER. Fields in a relational database are also called *columns*.

The actual information about a single supplier that resides in a table is called a *row*. Rows are commonly referred to as records.

Note that there is a field for Supplier_Number in this table. This field uniquely identifies each record so that the record can be retrieved, updated, or sorted, and it is called a **key field**. Each table in a relational database has one field 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.

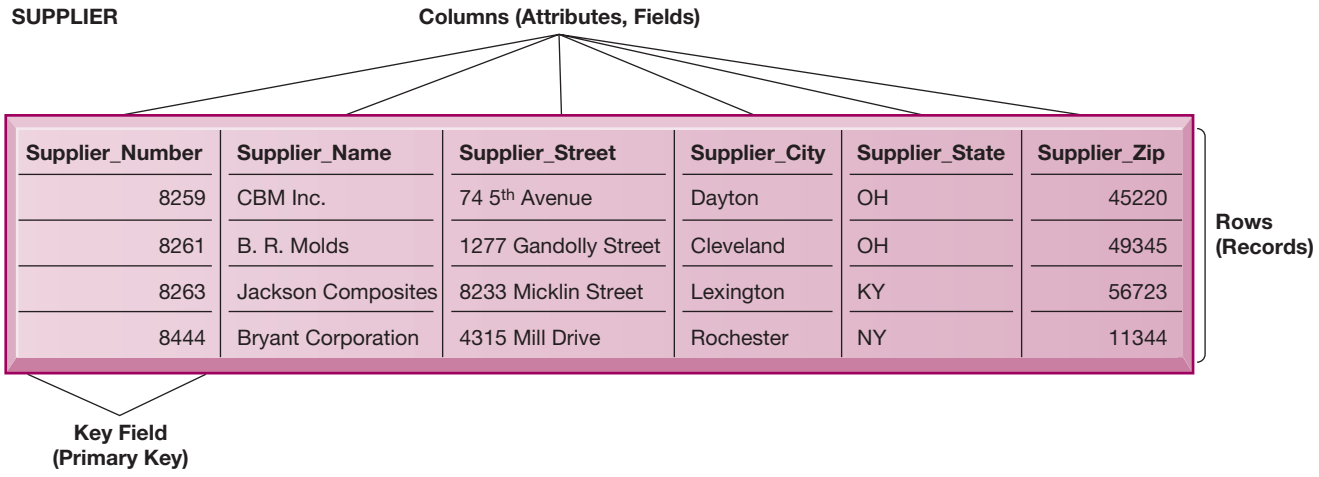


Figure 6.2
A Relational Database Table.
A relational database organizes data in the form of two-dimensional tables. Illustrated here is a table for the entity SUPPLIER showing how it represents the entity and its attributes. Supplier_Number is the key field.

We could use the supplier's name as a key field. However, if two suppliers had the same name (which does happen from time to time), supplier name would not uniquely identify each, so it is necessary to assign a special identifier field for this purpose. For example, if you had two suppliers, both named "CBM," but one was based in Dayton and the other in St. Louis, it would be easy to confuse them. However, if each has a unique supplier number, such confusion is prevented.

We also see that the address information has been separated into four fields: Supplier_Street, Supplier_City, Supplier_State, and Supplier_Zip. Data are separated into the smallest elements that one would want to access separately to make it easy to select only the rows in the table that match the contents of one field, such as all the suppliers in Ohio (OH). The rows of data can also be sorted by the contents of the Supplier_State field to get a list of suppliers by state regardless of their cities.

So far, the SUPPLIER table does not have any information about the parts that a particular supplier provides for your company. PART is a separate entity from SUPPLIER, and fields with information about parts should be stored in a separate PART table (see Figure 6.3).

Why not keep information on parts in the same table as suppliers? If we did that, each row of the table would contain the attributes of both PART and SUPPLIER. Because one supplier could supply more than one part, the table would need many extra rows for a single supplier to show all the parts that supplier provided. We would be maintaining a great deal of redundant data about suppliers, and it would be difficult to search for the information on any individual part because you would not know whether this part is the first or fiftieth part in this supplier's record. A separate table, PART, should be created to store these three fields and solve this problem.

The PART table would also have to contain another field, Supplier_Number, so that you would know the supplier for each part. It would not be necessary to keep repeating all the information about a supplier in each PART record because having a Supplier_ Number field in the PART table allows you to look up the data in the fields of the SUPPLIER table.

Notice 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 look-up field to find data about the supplier of a specific part. Note that the PART table would itself have its own primary key field, Part_Number, to identify each part uniquely. This key is not used to link PART with SUPPLIER but could be used to link PART with a different entity.

As we organize data into tables, it is important to make sure that all the attributes for a particular entity apply only to that entity. If you were to keep the supplier's address with the PART record, that information would not really relate only to PART; it would relate to both PART and SUPPLIER. If the supplier's address were to change,

PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

Primary Key

Foreign Key

Figure 6.3
The PART Table.
Data for the entity PART have their own separate table. Part_Number is the primary key and Supplier_ Number is the foreign key, enabling users to find related information from the SUPPLIER table about the supplier for each part.

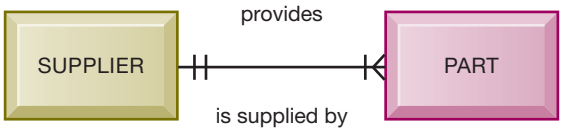


Figure 6.4
A Simple Entity-Relationship Diagram.
This diagram shows the relationship between the entities SUPPLIER and PART.

it would be necessary to alter the data in every PART record rather than only once in the SUPPLIER record.

ESTABLISHING RELATIONSHIPS

Now that we've broken down our data into a SUPPLIER table and a PART table, we must make sure we understand the relationship between them. A schematic called an **entity-relationship diagram** clarifies table relationships in a relational database. The most important piece of information an entity-relationship diagram provides is the manner in which two tables are related to each other. Tables in a relational database may have one-to-one, one-to-many, and many-to-many relationships.

An example of a one-to-one relationship is a human resources system that stores confidential data about employees. The system stores data, such as the employee name, date of birth, address, and job position, in one table and confidential data about that employee, such as salary or pension benefits, in another table. These two tables pertaining to a single employee would have a one-to-one relationship because each record in the EMPLOYEE table with basic employee data has only one related record in the table storing confidential data.

The relationship between the SUPPLIER and PART entities in our database is a one-to-many relationship. Each supplier can supply more than one part, but each part has only one supplier. For every record in the SUPPLIER table, many related records might be in the PART table.

Figure 6.4 illustrates how an entity-relationship diagram would depict this one-to-many relationship. 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 preceded by a short mark indicates a one-to-many relationship. Figure 6.4 shows that each part has only one supplier, but the same supplier can provide many parts.

We would also see a one-to-many relationship if we wanted to add a table about orders to our database because one supplier services many orders. The ORDER table would contain only the Order_Number and Order_Date fields. Figure 6.5 illustrates a

Figure 6.5
Sample Order Report.
The shaded areas show which data came from the ORDER, SUPPLIER, and LINE_ITEM tables. The database does not maintain data on extended price or order total because they can be derived from other data in the tables.

Order Number: 3502
Order Date: 1/15/2020

Supplier Number: 8259
Supplier Name: CBM Inc.
Supplier Address: 74 5th Avenue, Dayton, OH 45220

Order_Number	Part_Number	Part_Quantity	Part_Name	Unit_Price	Extended Price
3502	137	10	Door latch	22.00	$220.00
3502	152	20	Door lock	31.00	620.00
3502	178	5	Door handle	10.00	50.00
			Order Total:		$890.00

PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

LINE_ITEM

Order_Number	Part_Number	Part_Quantity
3502	137	10
3502	152	20
3502	178	5

ORDER

Order_Number	Order_Date
3502	1/15/2020
3503	1/16/2020
3504	1/17/2020

SUPPLIER

Supplier_Number	Supplier_Name	Supplier_Street	Supplier_City	Supplier_State	Supplier_Zip
8259	CBM Inc.	74 5th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Components	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Figure 6.6
The Final Database Design with Sample Records.
The final design of the database for suppliers, parts, and orders has four tables. The LINE_ITEM table is a join table that eliminates the many-to-many relationship between ORDER and PART.

report showing an order of parts from a supplier. If you look at the report, you can see that the information on the top-right portion of the report comes from the ORDER table. The actual line items ordered are listed in the lower portion of the report.

Because one order can be for many parts from a supplier, and a single part can be ordered many times on different orders, this creates a many-to-many relationship between the PART and ORDER tables. Whenever a many-to-many relationship exists between two tables, it is necessary to link these two tables in a table that joins this information. Creating a separate table for a line item in the order would serve this purpose. This table is often called a *join table* or an *intersection relation*. This join table contains only three fields: Order_Number and Part_Number, which are used only to link the ORDER and PART tables, and Part_Quantity. If you look at the bottom-left part of the report, this is the information coming from the LINE_ITEM table.

We would thus wind up with a total of four tables in our database. Figure 6.6 illustrates the final set of tables, and Figure 6.7 shows what the entity-relationship diagram for this set of tables would look like. Note that the ORDER table does not contain data on the extended price because that value can be calculated by

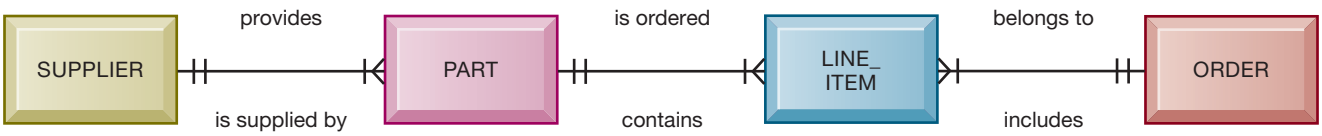


Figure 6.7
Entity-Relationship Diagram for the Database with Four Tables.
This diagram shows the relationship between the SUPPLIER, PART, LINE_ITEM, and ORDER entities.

multiplying Unit_Price by Part_Quantity. This data element can be derived when needed, using information that already exists in the PART and LINE_ITEM tables. Order_Total is another derived field, calculated by totaling the extended prices for items ordered.

The process of streamlining complex groups of data to minimize redundant data elements and awkward many-to-many relationships and increase stability and flexibility is called **normalization**. A properly designed and normalized database is easy to maintain and minimizes duplicate data. The Learning Tracks for this chapter direct you to more-detailed discussions of database design, normalization, and entity-relationship diagramming.

Relational database systems 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 have just created, 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!

The example provided here for parts, orders, and suppliers is a simple one. Even in a small business, you will have tables for other important entities such as customers, shippers, and employees. A large corporation typically has databases with thousands of entities (tables) to maintain. What is important for any business, large or small, is to have a good data model that includes all its entities and the relationships among them, one that is organized to minimize redundancy, maximize accuracy, and make data easily accessible for reporting and analysis.

It cannot be emphasized enough: If the business does not get its data model right, the system will not be able to serve the business properly. The company's systems will not be as effective as they could be because they will 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 a very important lesson you can learn from this course.

For example, Famous Footwear, a shoe store chain with more than 1,100 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 a rapidly adjusting store inventory. The company had a database that 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 that organized the sales and inventory data better for analysis and inventory management.

6-2 What are the principles of a database management system?

Now that you have started creating the files and identifying the data your business requires, you will need a database management system to help you manage and use the data. A **database management system (DBMS)** is a specific type of software for creating, storing, organizing, and accessing data from a database. Microsoft Access is a DBMS for desktop systems, whereas DB2, Oracle Database, and Microsoft SQL Server are DBMS for large mainframes and midrange computers. MySQL is a popular open-source DBMS. All these products are relational DBMS that support a relational database.

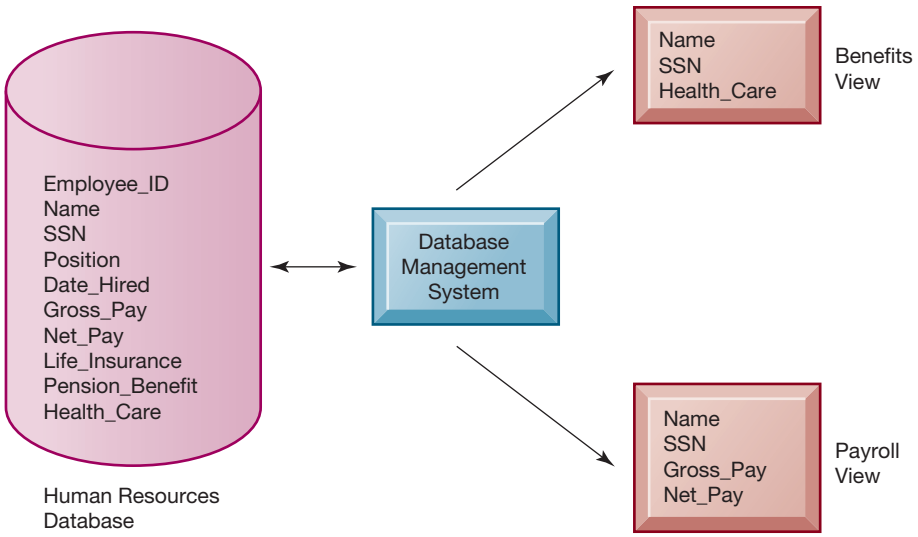


Figure 6.8
Human Resources Database with Multiple Views.
A single human resources database provides many views of data, depending on the information requirements of the user. Illustrated here are two possible views, one of interest to a benefits specialist and one of interest to a member of the company's payroll department.

The DBMS relieves the end user or programmer 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 end users or business specialists would perceive them, whereas the *physical view* shows how data are actually organized and structured on physical storage media, such as a hard disk.

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 6.8, a benefits specialist typically will require a view consisting of the employee's name, social security number, and health insurance coverage. A payroll department member will need data such as the employee's name, social security number, gross pay, and net pay. The data for all of these views is stored in a single database, where the organization can manage it more easily.

OPERATIONS OF A RELATIONAL DBMS

In a relational database, tables can be easily combined to deliver data that users require, provided that any two tables share a common data element. Let's return to the database we set up earlier with PART and SUPPLIER tables illustrated in Figures 6.2 and 6.3.

Suppose we wanted to find in this database the names of suppliers who could provide us with part number 137 or part number 150. We would need information from two tables: the SUPPLIER table and the PART table. Note that these two tables have a shared data element: Supplier_Number.

In a relational database, three basic operations, as shown in Figure 6.9, are used to develop useful sets of data: select, project, and join. 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 are 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 (see Figure 6.9).

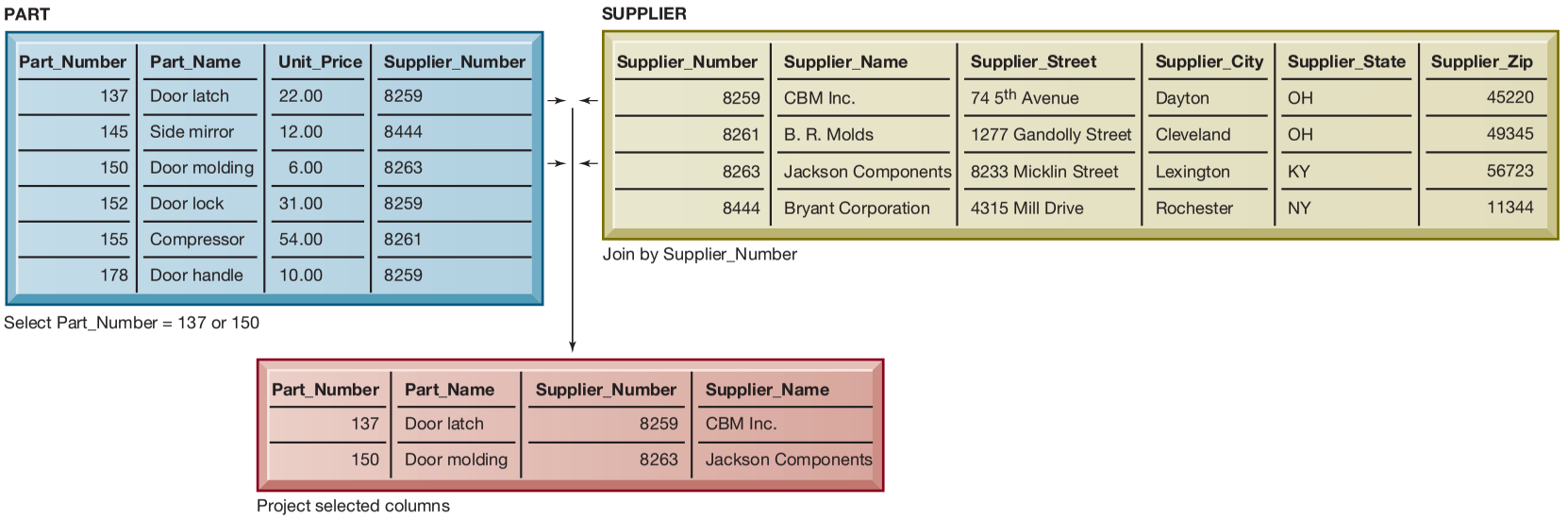


Figure 6.9
The Three Basic Operations of a Relational DBMS.
The select, join, and project operations enable data from two tables to be combined and only selected attributes to be displayed.

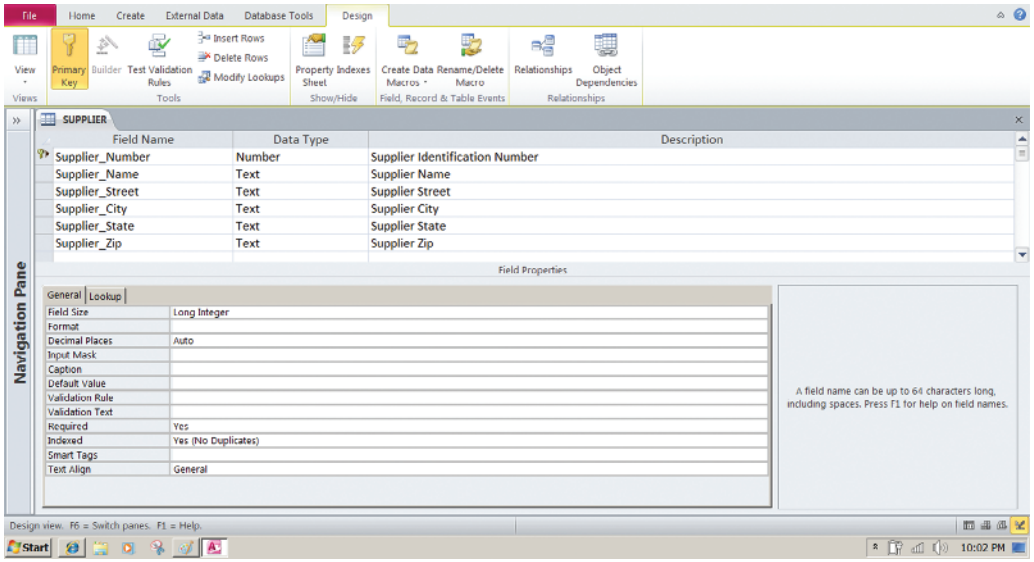


Figure 6.10
Access Data Dictionary Features.
Microsoft Access has a rudimentary data dictionary capability that displays information about the size, format, and other characteristics of each field in a database. Displayed here is the information maintained in the SUPPLIER table. The small key icon to the left of Supplier_Number indicates that it is a key field.

Source: Courtesy of Microsoft Corporation

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 capability, 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 6.10). 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 include 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 6.11 illustrates the SQL **query** that would produce the new resultant table in Figure 6.9. A query is a request for data from a database. You can find out more about how to perform SQL queries in our Learning Tracks for this chapter.

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

Figure 6.11
Example of a SQL Query.
Illustrated here are the SQL statements for a query to select suppliers for parts 137 or 150. They produce a list with the same results as Figure 6.9.

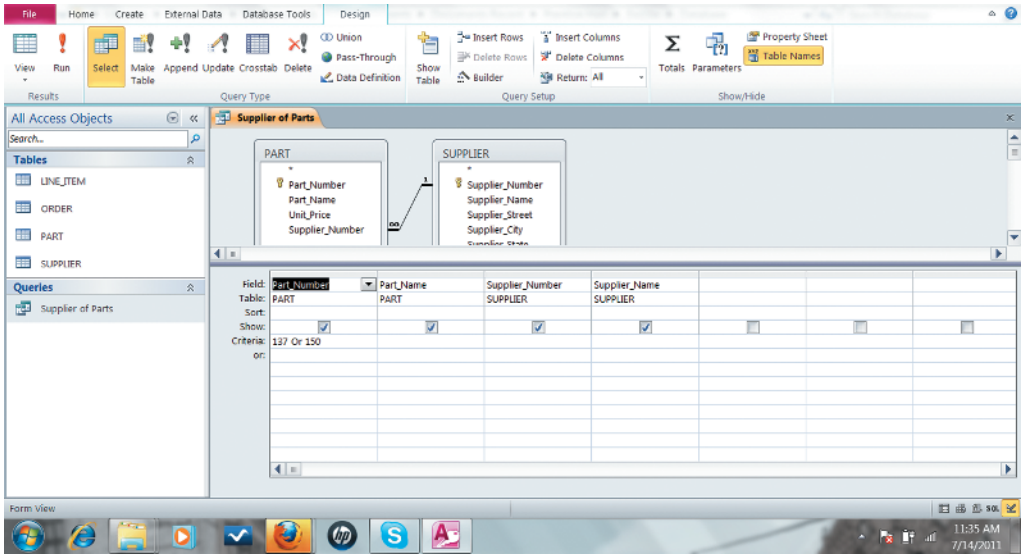


Figure 6.12
An Access Query.
Illustrated here is how the query in Figure 6.11 would be constructed using Microsoft Access query-building tools. It shows the tables, fields, and selection criteria used for the query.

Source: Courtesy of Microsoft Corporation

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.

Microsoft Access has capabilities to help users 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 6.12 illustrates how the SQL query to select parts and suppliers in Figure 6.11 would be constructed using Microsoft Access.

DBMS typically 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 Microsoft Access.

Microsoft Access also has capabilities for developing desktop system applications. These include tools for creating data entry screens and reports and developing the logic for processing transactions. Information systems specialists primarily use these capabilities.

NONRELATIONAL DATABASES, CLOUD DATABASES, AND BLOCKCHAIN

For more than 30 years, relational database technology has been the gold standard. Cloud computing, unprecedented data volumes, massive workloads for web services, and the need to store new types of data require database alternatives to the traditional relational model of organizing data in the form of tables, columns, and rows. Companies are turning to "NoSQL" nonrelational database technologies for this purpose. **Nonrelational database management systems** use a more flexible data model and are designed for managing large data sets across many distributed machines and for easily scaling up or down. They are useful for accelerating simple queries against large volumes of structured and unstructured data, including web, social media, graphics, and other forms of data that are difficult to analyze with traditional SQL-based tools.

There are several different kinds of NoSQL databases, each with its own technical features and behavior. Oracle NoSQL Database is one example, as is Amazon's SimpleDB, one of the Amazon Web Services that run in the cloud. SimpleDB provides a simple web services interface to create and store multiple data sets, query data

easily, and return the results. There is no need to predefine a formal database structure or change that definition if new data are added later.

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

Cloud Databases and Distributed Databases

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

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

Blockchain

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

The blockchain maintains a continuously growing list of records called blocks. Each block contains a timestamp and link to a previous block. Once a block of data is recorded on the blockchain ledger, it cannot be altered retroactively. When someone wants to add a transaction, participants in the network (all of whom have copies of the existing blockchain) run algorithms to evaluate and verify the proposed transaction. Legitimate changes to the ledger are recorded across the blockchain in a matter of seconds or minutes and records are protected through cryptography. What makes a blockchain system possible and attractive to business firms is encryption and authentication of the actors and participating firms which ensure only legitimate actors can enter information, and only validated transactions are accepted. Once recorded, the transaction cannot be changed. Figure 6.13 illustrates how blockchain works for fulfilling an order.

There are many benefits to firms using blockchain databases. Blockchain networks radically reduce the cost of verifying users, validating transactions, and the risks of storing and processing transaction information across thousands of firms. Instead of thousands of firms building their own private transaction systems, then integrating them with suppliers, shippers, and financial institution systems, blockchain can provide a single, simple, low-cost transaction system for participating firms.

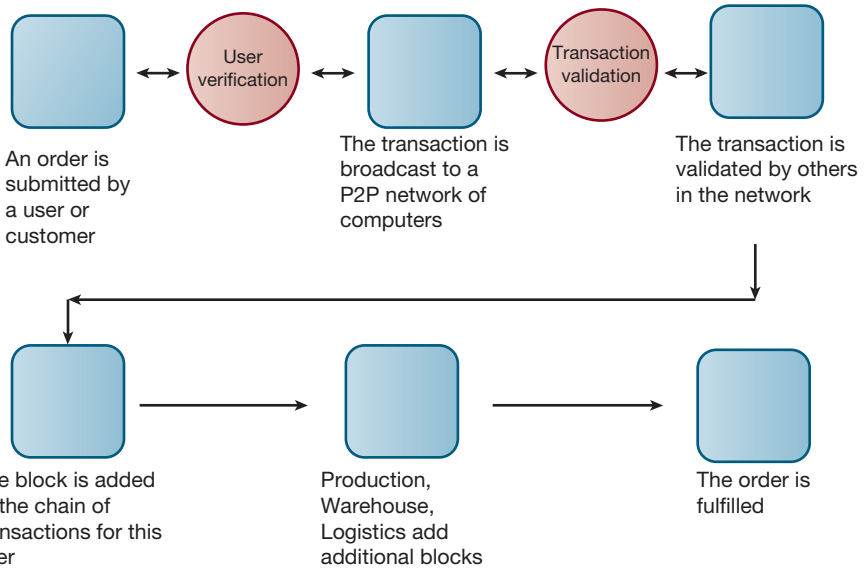


Figure 6.13 How Blockchain Works.
A blockchain system is a distributed database that records transactions in a peer-to-peer network of computers.

Standardization of recording transactions is aided through the use of *smart contracts*. Smart contracts are computer programs that implement the rules governing transactions between firms, e.g., what is the price of products, how will they be shipped, when will the transaction be completed, who will finance the transaction, what are financing terms, and the like.

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

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

Businesses use their databases to keep track of basic transactions, such as paying suppliers, processing orders, serving 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 of the data that organizations collected was transaction data that could easily fit into rows and columns of relational database management systems. There has been an explosion of data from many different sources, including web traffic, email messages, and social media content (tweets, status messages) as well as machine-generated data from sensors. These data may be unstructured or semistructured and thus not suitable for relational database products that organize data in the form of columns and rows. We now use the term **big data** to describe these data sets with volumes so huge that they are beyond the ability of typical DBMS to capture, store, and analyze.

The Paradise Papers and Big Data Journalism

In the last months of 2017, the news cycle around the world was dominated by reports of tax neutralization by the rich and powerful. This was the result of a thorough analysis by news outlets of 13.4 million leaked documents that detailed the tax-avoidance strategies of wealthy individuals and companies. The leaked documents, in large part originating from tax consultancy firm Appleby, were dubbed the "Paradise Papers" after the idyllic islands that had served as tax havens, such as Bermuda, where Appleby's headquarters are based.

These documents came into the possession of the German newspaper *Süddeutsche Zeitung*, whose managers quickly realized that they would be unable to analyze all the data by themselves. They reached out to other news organizations, including *The Guardian* and the BBC in the United Kingdom. No one knows for certain how many journalists and data analysts combed through the emails, reports, and accounts weighing in at over 1,400 gigabytes—there were likely hundreds of them. But all their efforts proved worthwhile, yielding a long list of scoops.

The star of the Irish band U2, Bono, turned out to be part owner of a shopping center in Lithuania that was being investigated for dodging taxes. A member of the British House of Lords, Baron Ashcroft, proved to have retained residency status outside the United Kingdom to evade taxes despite his protestations that he was a UK resident. The list of revelations went on. Most of those named in the Paradise Papers denied the accusations or stated that the tax neutralization methods they used were perfectly legal. Bono claimed ignorance of the shopping mall in Lithuania and said that he welcomed the insights provided by the analysis of the Paradise Papers.

The Paradise Papers reports were a triumph of what has come to be known as big data journalism. Few readers recognized what a major feat it had been, not least because the methods used were closely guarded. The International Consortium of Investigative Journalists, which played a major role in coordinating the analysis of the data, recognized that computer scientists would have to play a vital role in this enterprise. A chief technology officer was appointed to supervise the efforts along with six software developers. Access to the files was restricted; for one thing, the identity of the whistleblower had to be protected. At the instigation of the Consortium, all systems used by the journalists while studying the files were encrypted and a two-factor authentication system was applied. Many journalists were not well versed in security issues and had to be taught before they could work on the documents.

Not only did the leaked documents number in the millions, but they came in various formats too—some of these files were even in PSP (Paint Shop Pro). One reason for this was that Appleby was not the only source of the documents; in total, there were at least 19 data sources. A big challenge, thus, was devising a system that allowed easy access to all the files. However, the documents included emails, handwritten notes, photos, etc.—files that could not be read by machines, an absolutely necessary precondition for establishing a database.

A software company named Nuix stepped in and assisted in transforming all of these documents into a readable format through advanced optical character recognition software that could recognize text based on combinations of words that often occur together. Ingenious solutions like this allowed the journalists to finally put all the files in one database. Data analysts then devised algorithms that could cut across the many coding systems used in the 13.4 million documents and create links between companies and individuals and the data relevant to them.

This was perhaps the most difficult part of the enterprise, as it required close cooperation between the computer scientists and the journalists. To create a successful and efficient algorithm, the journalists needed to provide the data specialists with lists of the terms that were used in the Paradise Papers to refer to individuals and companies. As there are many of these—and they sometimes appear only as numerical codes—there was a strong need for cooperation between the journalists on the one hand and the data analysts on the other. Generally speaking, the success of big data journalism is contingent on close cooperation between data analysts and journalists; the challenge for journalists is to provide the data experts with clear information, while the challenge for the data experts is to create a knowledge center that journalists, many of whom are decidedly not computer wizards, can easily use.

The reporting on the Paradise Papers has been universally acclaimed as an outstanding

example of how new technology and techniques can be used to journalism's advantage. The news outlets won many awards for their work, including a prestigious Investigative Reporters and Editors award recognizing the innovative use of big data.

Sources: "Paradise Papers," BBC News, https://www.bbc.com/news/paradisepapers; "Paradise Papers: Secrets of the Global Elite," ICIJ, https://www.icij.org/investigations/paradise-papers/; "Paradise Papers: Das ist das Leak," Süddeutsche Zeitung, https://projekte.sueddeutsche.de/paradisepapers/politik/das-ist-das-leak-e229478/, all accessed December 1, 2019.

CASE STUDY QUESTIONS

1. Why was it a challenge to place all the documents from the Paradise Papers in one database?
2. Protecting the identity of a whistleblower or whistleblowers is of vital importance to journalists. Give at least one reason why this is so important.
3. Explain why cooperation between data experts and journalists was vital to the efficient analysis of data.
4. News outlets have been experiencing a severe crisis of profitability. What do you think are the causes of this crisis? What role can big data analytics play in countering it?

Case contributed by Bernard Bouwman, Avans University of Applied Sciences

Big data is often characterized by the "3Vs": the extreme *volume* of data, the wide *variety* of data types and sources, and the *velocity* at which the data must be processed. Big data doesn't designate any specific quantity but usually refers to data in the petabyte and exabyte range—in other words, billions to trillions of records, respectively, from different sources. Big data are produced in much larger quantities and much more rapidly than traditional data. For example, a single jet engine is capable of generating 10 terabytes of data in just 30 minutes, and there are more than 25,000 airline flights each day. Twitter generates more than 8 terabytes of data daily. Digital information is growing exponentially, to an expected 35 zettabytes in 2020. According to the International Data Center (IDC) technology research firm, the world's data are more than doubling every two years.

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 and adds 10,000 more each day. To find ways to optimize the Shutterstock experience, it analyzes its big data to find out where its website visitors place their cursors and how long they hover over an image before making a purchase. Big data is also finding many uses in the public sector. For example, city governments are using big data to manage traffic flows and fight crime. The Interactive Session on People shows how journalism uses big data to investigate stories of enormous scope.

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 the 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 problems big data will solve for the business. The chapter-ending case explores these issues.

BUSINESS INTELLIGENCE TECHNOLOGY 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 technology 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 semistructured 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 three decades has been the data warehouse. A **data warehouse** is a database that stores current and historical data of potential interest to decision makers throughout the company. The data originate in many core operational transaction systems, such as systems for sales, customer accounts, and manufacturing, and may include data from website transactions. The data warehouse extracts current and historical data from multiple systems inside the organization. These data are combined with data from external sources and transformed by correcting inaccurate and incomplete data and structuring the data in a common repository for management reporting and analysis.

The data warehouse makes the data available for anyone to access as needed, but it cannot be altered. A data warehouse system also provides a range of ad hoc and standardized query tools, analytical tools, and graphical reporting facilities. A data warehouse can be deployed on premises, in the cloud, or in a hybrid cloud environment (review Chapter 5).

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

Hadoop

Relational DBMS and data warehouse products are not well suited for organizing and analyzing big data or data that do not easily fit into columns and rows used in their data models. For handling unstructured and semistructured 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 very large amounts of data across inexpensive computers. It breaks a big data problem down into subproblems, distributes them among up to thousands of inexpensive computer processing nodes, and then combines the result into a smaller data set that is easier to analyze. You've probably used Hadoop to find the best airfare on the Internet, do a search on Google, or connect with a friend on Facebook.

Hadoop consists of several key services: 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 nonrelational 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 semistructured data before they are loaded into a data warehouse. Yahoo uses Hadoop to track user 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's 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 be accomplished even on handheld devices.

The previous chapter describes some of the advances in contemporary computer hardware technology that make in-memory processing possible, such as powerful high-speed processors, multicore processing, and falling computer memory prices. These technologies help companies optimize the use of memory and accelerate processing performance while lowering costs. Leading in-memory database products include SAP HANA, Oracle Database In-Memory, Microsoft SQL Server, and Teradata Intelligent Memory.

Analytic Platforms

Commercial database vendors have developed specialized high-speed **analytic platforms** using both relational and nonrelational technology that are optimized for analyzing large data sets. Analytic platforms feature preconfigured hardware–software systems that are specifically designed for query processing and analytics. For example, 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 nonrelational database management systems. Analytic platforms are now available as cloud services.

Figure 6.14 illustrates a contemporary business intelligence technology infrastructure using the technologies we have just described. Current and historical data are extracted from multiple operational systems along with web data, social media data, Internet of Things (IoT) machine-generated data, unstructured audio/visual data, and other data from external sources. Some companies are starting to pour all of these types of data into a data lake. A **data lake** is a repository for raw unstructured data or structured data that for the most part have not yet been analyzed, and the data can be accessed in many ways. The data lake stores these data in their native format until they are needed. The Hadoop Distributed File System (HDFS) is often used to store the data lake contents across a set of clustered computer nodes, and Hadoop clusters may be used to preprocess some of these data for use in the data warehouse, data marts, or an analytic platform or for direct querying by power users. Outputs include reports and dashboards as well as query results. Chapter 11 discusses the various types of BI users and BI reporting in greater detail.

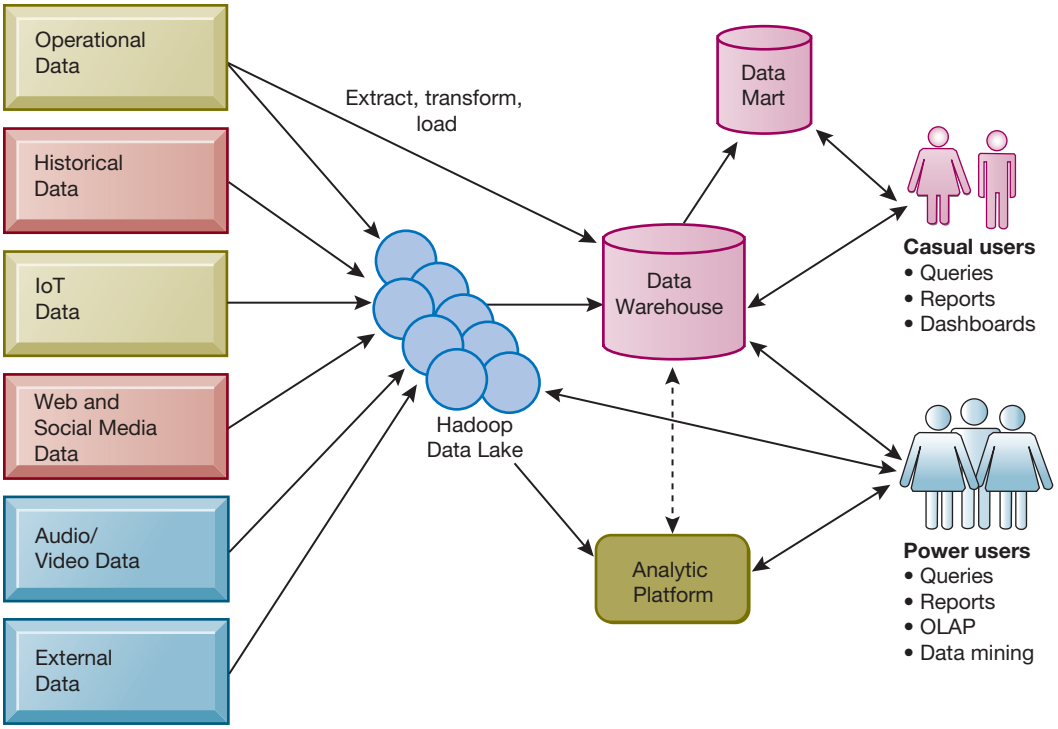


Figure 6.14
Business Intelligence Technology Infrastructure.
A contemporary business intelligence technology infrastructure features capabilities and tools to manage and analyze large quantities and different types of data from multiple sources. Easy-to-use query and reporting tools for casual business users and more sophisticated analytical toolsets for power users are included.

ANALYTICAL TOOLS: RELATIONSHIPS, PATTERNS, TRENDS

When data have been captured and organized using the business intelligence technologies we have just described, they are available for further analysis by 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 11.

Online Analytical Processing (OLAP)

Suppose your company sells four products—nuts, bolts, washers, and screws—in the East, West, and Central regions. If you wanted to ask a straightforward question, such as how many washers sold during the past quarter, you could easily find the answer by querying your sales database. However, 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. 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 rapid time, even when the data are stored in very large databases, such as sales figures for multiple years.

Figure 6.15 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

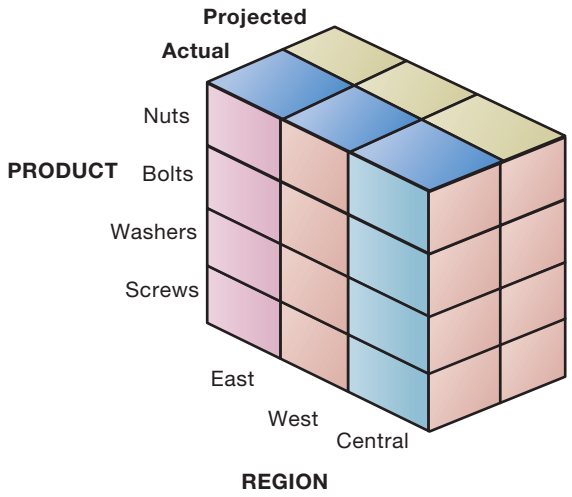


Figure 6.15
Multidimensional Data Model.
This view shows product versus region. If you rotate the cube 90 degrees, the face that will show is product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. Other views are possible.

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 2019?" OLAP, or multidimensional analysis, supports much more complex requests for information, such as, "Compare sales of product 403 relative to plan by quarter and sales region for the past two years." With OLAP and query-oriented data analysis, users need to have a good idea about the information for which they are looking.

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

- *Associations* are occurrences linked to a single event. For instance, a study of supermarket purchasing patterns might reveal that, when corn chips are purchased, a cola drink is purchased 65 percent of the time, but when there is a promotion, cola is purchased 85 percent of the time. This information helps managers make better decisions because they have learned the profitability of a promotion.
- In *sequences*, events are linked over time. We might find, for example, that if a house is purchased, a new refrigerator will be purchased within two weeks 65 percent of the time, and an oven will be bought within one month of the home purchase 45 percent of the time.
- *Classification* recognizes patterns that describe the group to which an item belongs by examining existing items that have been classified and by inferring a set of rules. For example, businesses such as credit card or telephone companies worry about the loss of steady customers. Classification helps discover the characteristics of customers who are likely to leave and can provide a model to help managers predict who those customers are so that the managers can devise special campaigns to retain such customers.
- *Clustering* works in a manner similar to classification when no groups have yet been defined. A data-mining tool can discover different groupings within data, such as finding affinity groups for bank cards or partitioning a database into groups of customers based on demographics and types of personal investments.

- Although these applications involve predictions, *forecasting* uses predictions in a different way. It uses a series of existing values to forecast what other values will be. For example, forecasting might find patterns in data to help managers estimate the future value of continuous variables, such as sales figures.

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

Caesars Entertainment, formerly known as Harrah's Entertainment, is the largest gaming company in the world. It continually analyzes data about its customers gathered when people play its slot machines or use its casinos and hotels. The corporate marketing department uses this information to build a detailed gambling profile, based on a particular customer's ongoing value to the company. For instance, data mining tells Caesars 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' profits so much that it became the centerpiece of the firm's business strategy, and customer data are Caesars' most valuable asset.

Text Mining and Web Mining

Unstructured data, most in the form of text files, is believed to account for more than 80 percent of useful organizational information and is one of the major sources of big data that firms want to analyze. Email, memos, call center transcripts, survey responses, legal cases, patent descriptions, and service reports are all valuable for finding patterns and trends that will help employees make better business decisions. **Text mining** tools are now available to help businesses analyze these data. These tools can 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 can mine text comments in an email message, blog, social media conversation, or survey form to detect favorable and unfavorable opinions about specific subjects. For example, Kraft Foods uses a Community Intelligence Portal and sentiment analysis to tune into consumer conversations about its products across numerous social networks, blogs, and other websites. Kraft tries to make sense of relevant comments rather than just track brand mentions and can identify customers' emotions and feelings when they talk about how they barbecue and what sauces and spices they use.

The web is another rich source of unstructured big data for revealing patterns, trends, and insights into customer behavior. The discovery and analysis of useful patterns and information from the web is 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 Google Trends, which tracks the popularity of various words and phrases used in Google search queries to learn what people are interested in and what they are interested in buying.

Web mining looks for patterns in data through content mining, structure mining, and usage mining. Web content mining is the process of extracting knowledge from the content of web pages, which may include text, image, audio, and video data. Web structure mining examines data related to the structure of a particular website. For example, links pointing to a document indicate the popularity of the document;

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 a web server records 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.

DATABASES AND THE WEB

Many companies are using the web to make some of the information in their internal databases available to customers and business partners. Prospective customers might use a company's website to view the company's product catalog or to place an order. The company in turn might use the web to check inventory availability for that product from its supplier.

These actions involve accessing and (in the case of ordering) updating corporate databases through the web. Suppose, for example, a customer with a web browser wants to search an online retailer's database for pricing information. Figure 6.16 illustrates how that customer might access the retailer's internal database over the web. The user would access the retailer's website over the Internet using web browser software on his or her client PC or mobile device. The user's web browser software would request data from the organization's database, using HTML commands to communicate with the web server.

Because many back-end databases cannot interpret commands written in HTML, the web server would pass these requests for data to software that translates HTML commands into SQL so the DBMS working with the database can process them. In a client/server environment, the DBMS often resides on a dedicated computer called a **database server**. The DBMS receives the SQL requests and provides the required data. The information is transferred from the organization's internal database back to the web server for delivery in the form of a web page to the user.

Figure 6.16 shows that the software working between the web server and the DBMS could be on an application server running on its own dedicated computer (see Chapter 5). 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.

There are a number of advantages to using the web to access an organization's internal databases. First, everyone knows how to use web browser software, and employees require much less training than if they used proprietary query tools. Second, the web interface requires few or no changes to the internal database. Companies leverage their investments in older systems because 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 and opportunities, and, in some cases, it is even changing the way business is being done. ThomasNet.com provides an up-to-date directory of information from more than 500,000 suppliers of industrial products such as chemicals, metals, plastics, rubber,

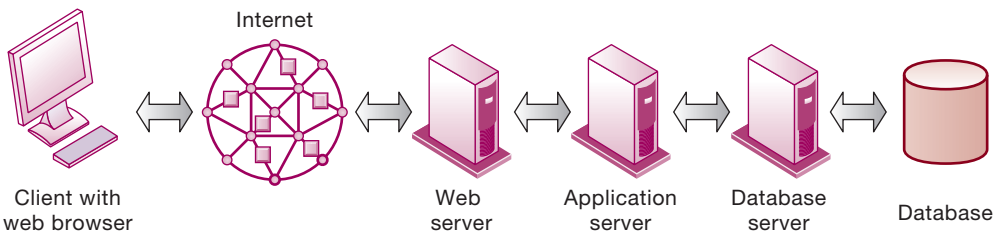


Figure 6.16
Linking Internal Databases to the Web.
Users access an organization's internal database through the web, using their desktop PCs or mobile devices and web browser software.

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 on its website and has become a smaller, leaner company. Facebook is an example of an entirely new business based on access to large databases through the web. Facebook maintains a massive database to house and manage all the data it collects about its 2.2 billion active users and their interests, friends, and photos.

6-4 Why are data governance and data quality assurance essential for managing the firm's data resources?

Setting up a database is only a start. To make sure that the data for your business remain accurate, reliable, and readily available to those who need it, your business will need special policies and procedures for data goverance. **Data governance** encompasses policies and procedures through which data can be managed as an organizational resource. It establishes the organization's rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying information. These include identifying which users and organizational units can share information, where information can be distributed, who is responsible for updating and maintaining the information, and how data resources should be secured (see Chapter 8). A firm's information policy might specify, for example, that only selected members of the payroll and human resources department would have the right to change or 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.

ASSURING DATA QUALITY

With today's organizations relying so heavily on data to drive operations and decision making, data quality assurance is especially important. 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? Data that are inaccurate, untimely, or inconsistent with other sources of information create serious operational and financial problems for businesses, even with a well-designed database and information policy. When faulty data go unnoticed, they often lead to incorrect decisions, product recalls, and even 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. For example, the sales ordering system and the inventory management system might both maintain data on the organization's products. However, the sales ordering system might use the term *Item Number*, and the inventory system might call the same attribute *Product Number*. The sales, inventory, or manufacturing systems of a clothing retailer might use different codes to represent values for an attribute. One system might represent clothing size as extra large, whereas the other system might use the code XL for the same purpose. During the design process for a 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 have received several pieces of the same direct mail advertising on the same day. This is very likely the result of your name being maintained multiple times in a database. Your name may have been misspelled, or you used your

DEWA: Evolving Utilities for a Smart City

The Dubai Electricity and Water Authority (DEWA) has been the sole provider of electricity and water services in Dubai since 1992, when the Dubai Electricity Company merged with the Dubai Water Department. Across the emirate, its 9,700 employees manage the generation and distribution of electricity to 652,200 customers and water to 580,678 customers. DEWA's mission is to provide its users with electricity and water in a sustainable way while maintaining high efficiency standards.

DEWA has a substantial generation capacity of around 10,000MW and water production of around 500 million imperial gallons per day. Utility companies measure their efficiency in customer minutes lost (CML), which stands for the average number of minutes a customer's supply has been interrupted. In 2019, the CML at DEWA was an impressive 2.39 minutes compared to the global standard of 15 minutes, and energy loss during transmission and distribution of electricity was 3.3 percent compared to 6–7 percent in Europe and the United States. Interruption of water supply, which indicates either low water quality or a situation where the customer has either not received water, was slightly higher, at 6.5 percent, compared to 15 percent in North America in 2018.

Dubai sits in the middle of a hot and sandy desert with a harsh climate and very few natural water sources. Its phenomenal growth over the last two decades has come with several costs, not least of which is the task of providing uninterrupted power and water to nearly 2 million people. In 2014, water connections increased by nearly 30 percent; in 2015, the demand for power shot up by 5–6 percent.

DEWA recognized that it had to deal with a range of problems. In addition to broader issues of inefficient processes, delays in upstream and downstream activities, and frequent mismatches between system inventory and physical stock, data management needed a complete overhaul. Information was still widely being stored on paper, which caused major delays in decision making. Employees had little access to information in the field via mobile; in fact, finding the information they needed sometimes required several trips to the field or the respective office. Furthermore, there was a lack of variety in payment options for customers and no scope for "smart" monitoring of electricity and water.

DEWA identified two main challenges: inventory-related matters and smart monitoring. The authority began working with SAP in 2012 for the former and with Honeywell in 2018 for the latter. To implement SAP Inventory Manager, DEWA collaborated with Wavelogix and On Device Solutions.

SAP Inventory Manager is a mobile application for wireless inventory management, including tracking of assets, tools, and parts. It automates stock functions like issues, transfers, returns, and audits, enabling an organization to respond appropriately to any fluctuations in demand. It does so by checking the availability of materials remotely using mobile devices, accepting and distributing incoming materials by purchase order, and prepicking materials based on these orders.

SAP Inventory Manager's implementation led to several changes at DEWA. Goods receipts were made against reservations and inbound delivery of goods to the warehouse; likewise, issuance of goods was done against reservations. This enabled goods to be seamlessly transferred internally between plants within the organization. Employees were able to efficiently pick, store, and track inventory in the stores through the SAP Material Master database, an application that contains information on all the materials that DEWA procures, produces, stores, and sells, information that can also be retrieved and edited through handheld devices.

The time between the purchase request and the purchase order cycle was reduced from 60 to 15 days, which led to savings of $3.4 million and 160,000 fewer sheets of paper wasted each year. The CML was brought down to 2.39 minutes from 5.62 minutes; water line losses, to 6.5 percent from 10.4 percent; and electricity losses, to 3.3 percent from 3.46 percent.

This and other, similar initiatives improved DEWA's cumulative efficiency by 29.68 percent over the course of a decade. In November 2012, DEWA won the bronze Quality Award in the Large Implementation category at the SAP EMEA Quality Awards, where it competed against organizations in Europe, the Middle East, and Africa; soon after, Dubai was ranked seventh in the world in the World Bank's Doing Business 2013 report for ease of access to electricity.

Dubai has invested heavily in its Smart Dubai initiative to transform into a pioneering smart city. Supporting this is DEWA's Smart Applications via

Smart Grid and Meters initiative, now collaborating with International conglomerate Honeywell in its third phase with the aim of installing a million smart meters by 2020. A smart infrastructure for electricity transmission and distribution networks would improve the speed with which power reconnects if any disruption occurs.

Honeywell's smart meters, equipped with wireless technology, sends detailed data from automatic readings to DEWA, which can adjust its supply as necessary, and directly to consumers as well, allowing them to make more informed decisions about their power usage. By using smart solutions to closely monitor electricity consumption and increase energy efficiency, DEWA also fulfills the Dubai government's twin objectives of true sustainability and satisfaction of its citizens and residents.

Sources: Navdeep Singla, "SAP Inventory Manager Case Study—DEWA," On Device Solutions, August 23, 2019; Dubai Electricity & Water Authority, "DEWA Sustainability Report 2018," www.dewa.gov.ae; Project Management Institute, "How DEWA Increased Efficiencies and Value Using Streamlined Project Management Processes" (2018), www.pmi.org; Jason Saundalkar, "DEWA Appoints Honeywell to Install 270,000 Smart Meters," *Middle East Construction News*, October 25, 2018; Edward Banda, "Dubai Electricity and Water Authority Continues Smart Transformation and Tests Blockchain," *Computer Weekly*, November 8, 2017.

CASE STUDY QUESTIONS

1. Identify the main challenge faced by DEWA in this case study.
2. Explain how people-based, organizational, and technological factors were involved in this challenge.
3. Why is accurate and consistent data particularly important for DEWA?

Case contributed by Saadat Alhashmi, University of Sharjah

middle initial on one occasion and not on another, or the information was initially entered on a paper form and not scanned properly into the system. Because of these inconsistencies, the database would treat you as different people! We often receive redundant mail addressed to Laudon, Lavdon, Lauden, or Landon.

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

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

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

Data quality problems are not just business problems. They also pose serious problems for individuals, affecting their financial condition and even their jobs. For example, inaccurate or outdated data about consumers' credit histories maintained by credit bureaus can prevent creditworthy individuals from obtaining loans or lower their chances of finding or keeping a job. And as the Interactive Session on Organizations shows, incomplete or inaccessible databases can cripple public authorities' ability to provide basic utilities to their citizens.

6-5 How will MIS help my career?

Here is how Chapter 6 and this book can help you find a job as an entry-level sales and marketing assistant for a global data services company.

THE COMPANY

Global Online Stats, a leading global provider of quantitative data, statistics, and market research products, has an open position for an entry-level sales and marketing assistant. The company has more than 500 employees and offices in Boston, London, and Paris. The company provides tools and services for accessing an online quantitative database aimed at business firms of all sizes, including consulting firms, media agencies, and marketing departments in large corporations from a variety of industries and countries.

POSITION DESCRIPTION

This position works closely with the Managing Director and Head of Global Sales to develop and maintain sales leads and new accounts. Job responsibilities include:

- Developing new accounts with leads generated by existing customers and relationships with media and industry associations as well as through cold calling, emailing, and online prospecting.
- Developing account relationships to turn sporadic customers into long-term business accounts.
- Developing sales opportunities for various categories of products and lines of business.
- Finding and scheduling appointments with new prospective clients.
- Updating customer and client profiles.

JOB REQUIREMENTS

- Four-year college degree
- Very strong verbal and written communication skills
- Microsoft Office skills
- Experience at a sales or marketing internship or in cold calling desirable
- Outgoing, competitive, proactive sales personality

INTERVIEW QUESTIONS

1. Have you ever worked with online databases or database software? Exactly what did you do with these databases? Did you ever take a database course?
2. Did you work with quantitative data in your college courses or at a prior job? What did you do with the data?
3. What is your level of expertise with Microsoft Office tools—Word, Excel, PowerPoint, Access?
4. What sales and marketing experience have you had?
5. Do you have any foreign language proficiency?
6. What challenges would you anticipate trying to sell our products and services to non-US organizations?

AUTHOR TIPS

1. Review this chapter and the discussion of business intelligence and analytics in Chapter 11.
2. Use the web to research the company, its products, services, and customers, and the way it operates. Try to find out more about the company's online quantitative database.
3. Ask exactly how you would be using Microsoft Office tools in your job.
4. Ask about how much training you would receive in how to use the company's data products.

Review Summary

6-1 What is a database, and how does a relational database organize data? A database is a group of related files that keeps track of people, places, and things (entities) about which organizations maintain information. The relational database is the primary method for organizing and maintaining data today in information systems. It organizes data in two-dimensional tables with rows and columns called relations. 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 identify each record uniquely for retrieval or manipulation. An entity-relationship diagram graphically depicts the relationship between entities (tables) in a relational database. The process of breaking down complex groupings of data and streamlining them to minimize redundancy and awkward many-to-many relationships is called normalization.

6-2 What are the principles of a database management system? A DBMS consists of software that permits centralization of data and data management so that businesses have a single, consistent source for all their data needs. A single database services multiple applications. The DBMS separates the logical and physical views of data so that the user does not have to be concerned with the data's physical location. The principal capabilities of a DBMS include a data definition capability, a data dictionary capability, and a data manipulation language. Nonrelational databases are becoming popular for managing types of data that can't be handled easily by the relational data model.

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 types of data businesses use today, including semistructured and unstructured big data in very large quantities from many different sources. 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. 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 analyzing useful patterns and information from the World Wide Web, examining the structure of websites, activities of website users, and the contents of web pages. Conventional databases can be linked to the web or a web interface to facilitate user access to an organization's internal data.

6-4 Why are data governance 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. Data governance encompasses organizatonal policies and procedures for the maintenance, distribution, and use of information in the organization. Data that are inaccurate, incomplete, or inconsistent create serious operational and financial problems for businesses if they lead to bad decisions about the actions the firm should take. Assuring data quality involves using enterprise-wide data standards, databases designed to minimize inconsistent and redundant data, data quality audits, and data cleansing software.

Key Terms

Analytic platform, 236
Attributes, 222
Big data, 232
Bit, 220
Blockchain, 231
Byte, 220
Data cleansing, 243
Data definition, 229
Data dictionary, 229
Data governance, 241
Data lake, 236
Data manipulation language, 229
Data mart, 235
Data mining, 238
Data quality audit, 243
Data warehouse, 235
Database, 221
Database management system (DBMS), 226
Database server, 240
Distributed database, 231
Entity, 222
Entity-relationship diagram, 224
Field, 220
File, 221
Foreign key, 223
Hadoop, 235
In-memory computing, 236
Key field, 222
Nonrelational database management systems, 230
Normalization, 226
Online analytical processing (OLAP), 237
Primary key, 222
Query, 229
Record, 221
Referential integrity, 226
Relational database, 222
Report generator, 230
Sentiment analysis, 239
Structured Query Language (SQL), 229
Text mining, 239
Web mining, 239

Review Questions

6-1 What is a database, and how does a relational database organize data?
- Define a database.
- Define and explain the significance of entities, attributes, and key fields.
- Define a relational database and explain how it organizes and stores information.
- Explain the role of entity-relationship diagrams and normalization in database design.

6-2 What are the principles of a database management system?
- Define a database management system (DBMS), describe how it works, and explain how it benefits organizations.
- Define and compare the logical and physical views of data.
- Define and describe the three operations of a relational database management system.
- Name and describe the three major capabilities of a DBMS.
- Define a nonrelational database management system and explain how it differs from a relational DBMS.

6-3 What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?
- Explain why businesses may be more interested in big data than in smaller data sets.
- Describe how a business might use a data mart.
- Describe the capabilities of Hadoop Distributed File System for data storage.

- Explain how in-memory computing can be used to facilitate data analysis.
- Explain what is meant by clustering in the context of data mining.
- Describe sentiment analysis and suggest why a business might use it.

6-4 Why are data governance and data quality assurance essential for managing the firm's data resources?

- Why is ensuring data quality important, and what might be involved in this?
- What is involved in a data quality audit?
- Identify individuals or organizations that might be affected by poor data quality.

MyLab MIS™

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

Discussion Questions

6-5 It has been said that you do not need database management software to create a database environment. Discuss. MyLab MIS

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

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

Hands-On MIS Projects

MANAGEMENT DECISION PROBLEMS

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

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

6-9 Your industrial supply company wants to create a data warehouse from which 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 several systems: a divisional sales system running on a UNIX server and a corporate sales system running on an IBM mainframe. You would like to create a single standard format that consolidates

these data from both systems. In MyLab MIS, you can review the proposed format along with sample files from the two systems that would supply the data for the data warehouse. Then answer the following questions:

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

ACHIEVING OPERATIONAL EXCELLENCE: BUILDING A RELATIONAL DATABASE FOR INVENTORY MANAGEMENT

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

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

After 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. Within each supplier category, 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 identified items.
- Write a brief description of how the database could be enhanced to improve management of the business further. 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 online 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 attorneys, 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, Australiatradenow, and the Nationwide Business Directory of Australia. If necessary, use search engines such as Yahoo and Google.

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

COLLABORATION AND TEAMWORK PROJECT

Identifying Entities and Attributes in an Online Database

6-12 With your team of three or four students, select an online database to explore, such as Best Buy or the Internet Movie Database. Explore one of these websites to see what information it provides. List the entities and attributes that the company running the website must keep track of in its databases. Diagram the relationships 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.

BUSINESS PROBLEM-SOLVING CASE

DOES BIG DATA PROVIDE THE ANSWER?

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

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

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

New York City analyzes all the crime-related data it collects to lower the crime rate. Its CompStat crime-mapping program uses a comprehensive citywide database of all reported crimes or complaints, arrests, and summonses in each of the city's 76 precincts to report weekly on crime complaint and arrest activity at the precinct, patrol borough, and citywide levels. CompStat data can be displayed on maps showing crime and arrest locations, crime hot spots, and other relevant information to help precinct commanders quickly identify patterns and trends and deploy police personnel where they are most needed.

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

Big Data does not always reflect emotions or intuitive feelings. For example, when LEGO faced bankruptcy in 2002–2003 the company used big data to determine that Millennials have short attention spans and get easily bored. The message from the data led LEGO to deemphasize their small iconic bricks in favor of large simplistic building blocks. This change only accelerated LEGO's decline, so the company decided to go into consumers' homes to try and reconnect with once-loyal customers. After meeting with an 11-year-old German boy, LEGO discovered that for children, playing and showing mastery in something were more valuable than receiving instant gratification. LEGO then pivoted again to emerge after its successful 2014 movie into the world's largest toy maker. Patterns and trends can sometimes be misleading.

Huge volumes of data do not necessarily provide more reliable insights. Sometimes the data being analyzed are not a truly representative sample of the data required. For example, election pollsters in the United States have struggled to obtain representative samples of the population because a majority of people do not have landline phones. It is more time-consuming and expensive for pollsters to contact mobile phone users, which now constitute 75 percent of some samples. U.S. law bans cell phone autodialing, so pollsters have to dial numbers by hand individually and make more calls, since mobile users tend to screen out unknown callers. Opinions on Twitter do not reflect the opinions

of the US population as a whole. The elderly, poor people or introverts, who tend not to use social media—or even computers—often get excluded.

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 the decline in United States crime rate was highly correlated with the decline in the market share of video rental stores such as Blockbuster. But that doesn't necessarily mean there is any meaningful connection between the two phenomena. Data analysts need some business knowledge of the problem they are trying to solve with big data.

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.

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

The New York Police Department (NYPD) recently developed a tool called Patternizr, which uses pattern recognition to identify potential criminals. The software searches through hundreds of thousands of crime records across 77 precincts in the NYPD database to find a series of crimes likely to have been committed by the same individual or individuals, based on a set of identifying characteristics. In the past, analysts had to manually review reports to identify patterns, a very time-consuming and inefficient process. Some experts worry that Patternizr inadvertently perpetuates bias. The NYPD used 10 years of manually identified pattern data to train Patternizr, removing attributes such as gender, race, and specific location from the data. Nevertheless such efforts may not eliminate racial and gender bias in Patternizr if race and gender played any role in past police actions used to model predictions. According to Gartner Inc. analyst Darin Stewart, Patternizr will sweep up individuals who fit a profile inferred by the system. At best, Stewart says, some people identified by Patternizr will be inconvenienced and insulted. At worst, innocent people will be incarcerated.

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. 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.

Insurance companies such as Progressive offer a small device to install in your car to analyze your driving habits, ostensibly to give you a better insurance rate. However, some of the criteria for lower auto insurance rates are considered discriminatory. For example, insurance companies like people who don't drive late at night and don't spend much time in their cars. However, poorer people are more likely to work a late shift and to have longer commutes to work, which might increase their auto insurance rates.

More and more companies are turning to computerized systems to filter and hire job applicants, especially for lower-wage, service-sector jobs. The algorithms these systems use to evaluate job candidates may be preventing qualified applicants from obtaining these jobs. For example, some of these algorithms have determined that, statistically, people with shorter commutes are more likely to stay in a job longer than those with longer commutes or less reliable transportation or those who haven't been at their address for very long. If asked, "How long is your commute?" applicants with long commuting times will be scored lower for the job. Although such considerations may be statistically accurate, is it fair to screen job applicants this way?

Sources: Brian Holak, "NYPD's Patternizr Crime Analysis Tool Raises AI Bias Concerns," searchbusinessanalytics.com, March 14, 2019; Linda Currey Post, "Big Data Helps UK National Health Service Lower Costs, Improve Treatments," *Forbes*, February 7, 2018; Michael Jude, Rajkumar Venkatesan and Christina Black, "Using Big Data: 3 Reasons It Fails and 4 Ways to Make It Work," University of Virginia Darden School of Business Press Release, February 2018; Alex Bekker, "Big Data: A Highway to Hell or a Stairway to Heaven? Exploring Big Data Problems," *ScienceSoft*, May 19, 2018; Ernest Davis, "The Problems of Big Data, and What to Do About Them," World Economic Forum, February 15, 2017; and Gary Marcus and Ernest Davis, "Eight (No, Nine!) Problems With Big Data," *New York Times*, April 6, 2014.

CASE STUDY QUESTIONS

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

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

6-15 Describe the limitations to using big data.

6-16 Should all organizations try to collect and 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?

Chapter 6 References

Barton, Dominic, and David Court. "Making Advanced Analytics Work for You." *Harvard Business Review* (October 2012).

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

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

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

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

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

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

DalleMule, Leandro, and Thomas H. Davenport. "What's Your Data Strategy?" *Harvard Business Review* (May–June 2017).

DataInformed. "The Database Decision: Key Considerations to Keep in Mind." *Wellesley Information Services* (2015).

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

Duncan, Alan D., Mei Yang Selvage, and Saul Judah. "How a Chief Data Officer Should Drive a Data Quality Program." Gartner Inc. (October 14, 2016).

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

Experian Data Quality. "Connecting Data Quality Initiatives with Business Drivers." *Experian Data Solutions* (2016).

Experian Information Solutions. "The 2018 Global Data Management Benchmark Report." (2018).

Felin, Teppo and Karim Lakhani. "What Problems Will You Solve with Blockchain?" *MIT Sloan Management Review* 60, No. 1 (Fall 2018).

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

Hoffer, Jeffrey A., Ramesh Venkataraman, and Heikki Toppi. *Modern Database Management*, 13th ed. (Upper Saddle River, NJ: Prentice-Hall, 2019).

Horst, Peter and Robert Dubroff. "Don't Let Big Data Bury Your Brand." *Harvard Business Review* (November 2015).

Kroenke, David M, David J. Auer., Robert C. Yoder, and Scott L. Vandenberg. *Database Processing: Fundamentals, Design, and Implementation*, 15th ed. (Upper Saddle River, NJ: Prentice-Hall, 2019).

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

Lukyanenko, Roman, Jeffrey Parsons, Yolanda F. Wiersma, and Mahed Maddah. "Expecting the Unexpected: Effects of Data Collection Design Choices on the Quality of Crowdsourced User-Generated Content," *MIS Quarterly* 43, No. 2 (June 2019).

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

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

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

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

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

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

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

Redman, Thomas C. "Data's Credibility Problem." *Harvard Business Review* (December 2013).

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

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

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

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