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:

1. What are the problems of managing data resources in a traditional file environment and how are they solved by a database management system?
2. What are the major capabilities of database management systems (DBMS) and why is a relational DBMS so powerful?
3. What are some important principles of database design?
4. What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?
5. Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?

Interactive Sessions:

What Can Businesses Learn from Text Mining?

Credit Bureau Errors—Big People Problems

CHAPTER OUTLINE

RR DONNELLEY TRIES TO MASTER ITS DATA

Right now you are most likely using an RR Donnelley product. Chicago-based RR Donnelley is a giant commercial printing and service company providing printing services, forms and labels, direct mail, and other services. This textbook probably came off its presses. The company's recent expansion has been fueled by a series of acquisitions, including commercial printer Moore Wallace in 2005 and printing and supply chain management company Banta in January 2007. RR Donnelley's revenue has jumped from $2.4 billion in 2003 to over $9.8 billion today.

However, all that growth created information management challenges. Each acquired company had its own systems and its own set of customer, vendor, and product data. Coming from so many different sources, the data were often inconsistent, duplicated, or incomplete. For example, different units of the business might each have a different meaning for the entity "customer." One might define "customer" as a specific billing location, while another might define "customer" as the legal parent entity of a company. Donnelley had to use time-consuming manual processes to reconcile the data stored in multiple systems in order to get a clear enterprise-wide picture of each of its customers, since they might be doing business with several different units of the company. These conditions heightened inefficiencies and costs.

RR Donnelley had become so big that it was impractical to store the information from all of its units in a single system. But Donnelley still needed a clear single set of data that was accurate and consistent for the entire enterprise. To solve this problem, RR Donnelley turned to master data management (MDM). MDM seeks to ensure that an organization does not use multiple versions of the same piece of data in different parts of its operations by merging disparate records into a single authenticated master file. Once the master file is in place, employees and applications access a single consolidated view of the company's data. It is especially useful for companies such as Donnelley that have data integration problems as a result of mergers and acquisitions.

Implementing MDM is a multi-step process that includes business process analysis, data cleansing, data consolidation and reconciliation, and data migration into a master file of all the company's data. Companies must identify what group in the company "owns" each piece of data and is responsible for resolving inconsistent definitions of data and other discrepancies. Donnelley launched its MDM program in late 2005 and began creating a single set of identifiers for its customer and vendor data. The company opted for a registry model using Purisma's Data Hub in which customer data continue to reside in the system where they originate but are registered in a master "hub" and cross-referenced so applications can find the data. The data in their source system are not touched.

Nearly a year later, Donnelley brought up its

Customer Master Data Store, which integrates the data from numerous systems from Donnelley acquisitions. Data that are outdated, incomplete, or incorrectly formatted are corrected or eliminated. A registry points to where the source data are stored. By having a single consistent enterprise-wide set of data with common definitions and standards, management is able to easily find out what kind of business and how much business it has with a particular customer to identify top customers and sales opportunities. And when Donelley acquires a company, it can quickly see a list of overlapping customers.

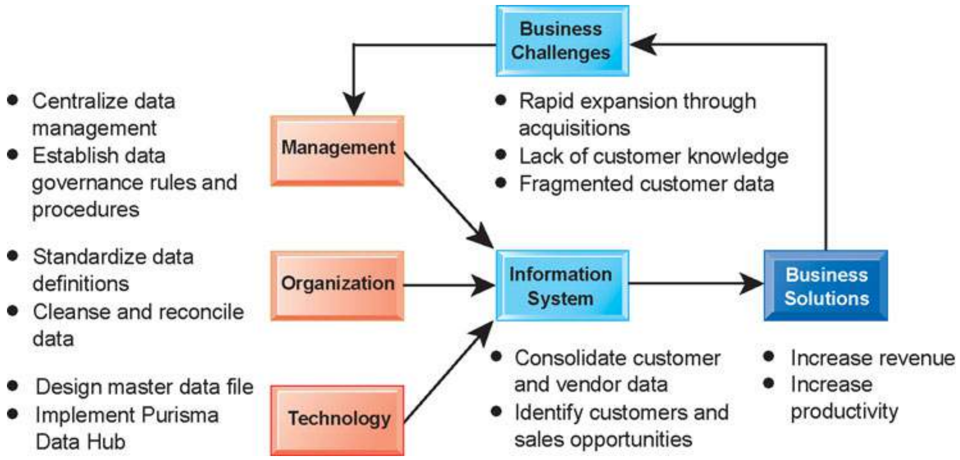
Sources: John McCormick, "Mastering Data at R.R. Donnelley," *Information Management Magazine*, March 2009; www.rrdonnelley.com, accessed June 10, 2010; and www.purisma.com, accessed June 10, 2010.

RR Donnelley's experience illustrates the importance of data management for businesses. Donnelley has experienced phenomenal growth, primarily through acquisitions. But its business performance depends on what it can or cannot do with its data. How businesses store, organize, and manage their data has a tremendous impact on organizational effectiveness.

The chapter-opening diagram calls attention to important points raised by this case and this chapter. Management decided that the company needed to centralize the management of the company's data. Data about customers, vendors, products, and other important entities had been stored in a number of different systems and files where they could not be easily retrieved and analyzed. They were often redundant and inconsistent, limiting their usefulness. Management was unable to obtain an enterprise-wide view of all of its customers at all of its acquisitions to market its products and services and provide better service and support.

In the past, RR Donnelley had used heavily manual paper processes to reconcile its inconsistent and redundant data and manage its information from an enterprise-wide perspective. This solution was no longer viable as the organization grew larger. A more appropriate solution was to identify, consolidate, cleanse, and standardize customer and other data in a single master data management registry. In addition to using appropriate technology, Donnelley had to correct and reorganize the data into a standard format and establish rules, responsibilities, and procedures for updating and using the data.

A master data management system helps RR Donnelley boost profitability by making it easier to identify customers and sales opportunities. It also improves operational efficiency and decision making by having more accurate and complete customer data available and reducing the time required to reconcile redundant and inconsistent data.


Business Challenges
Management
Organization
Technology
Information System
Business Solutions
• Centralize data management
• Establish data governance rules and procedures
• Standardize data definitions
• Cleanse and reconcile data
• Design master data file
• Implement Purisma Data Hub
• Rapid expansion through acquisitions
• Lack of customer knowledge
• Fragmented customer data
• Consolidate customer and vendor data
• Identify customers and sales opportunities
• Increase revenue
• Increase productivity

6.1 Organizing Data in a Traditional File Environment

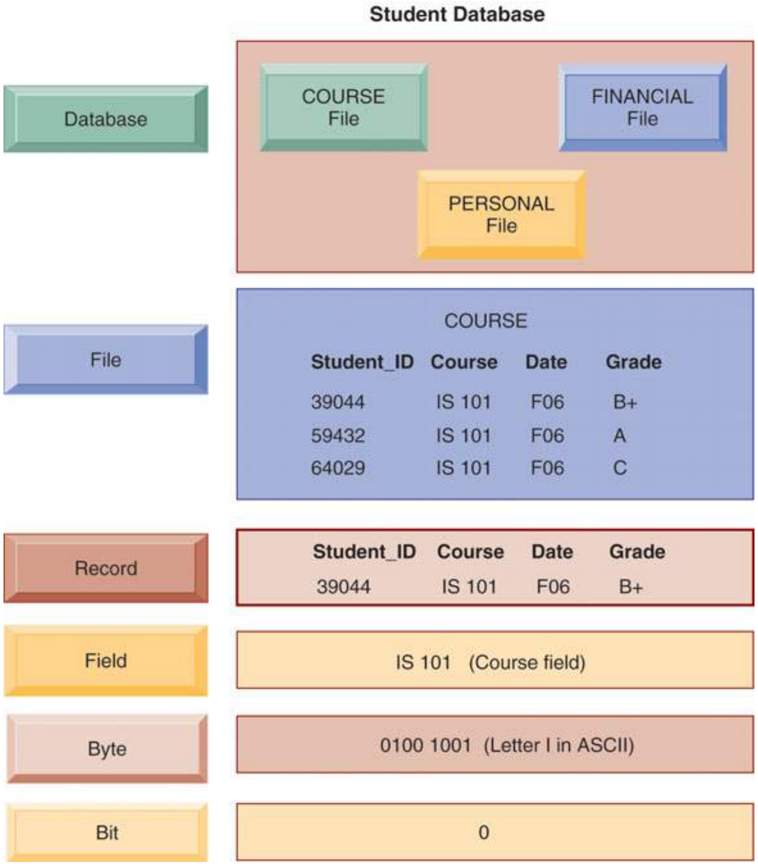
An effective information system provides users with accurate, timely, and relevant information. Accurate information is free of errors. Information is timely when it is available to decision makers when it is needed. Information is relevant when it is useful and appropriate for the types of work and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

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

FILE ORGANIZATION TERMS AND CONCEPTS

A computer system organizes data in a hierarchy that starts with bits and bytes and progresses to fields, records, files, and databases (see Figure 6-1). A bit represents the smallest unit of data a computer can handle. A group of bits, called a byte, represents a single character, which can be a letter, a

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.

number, or another symbol. A grouping of characters into a word, a group of words, or a complete number (such as a person's name or age) is called a **field**. A group of related fields, such as the student's name, the course taken, the date, and the grade, comprises a **record**; a group of records of the same type is called a **file**.

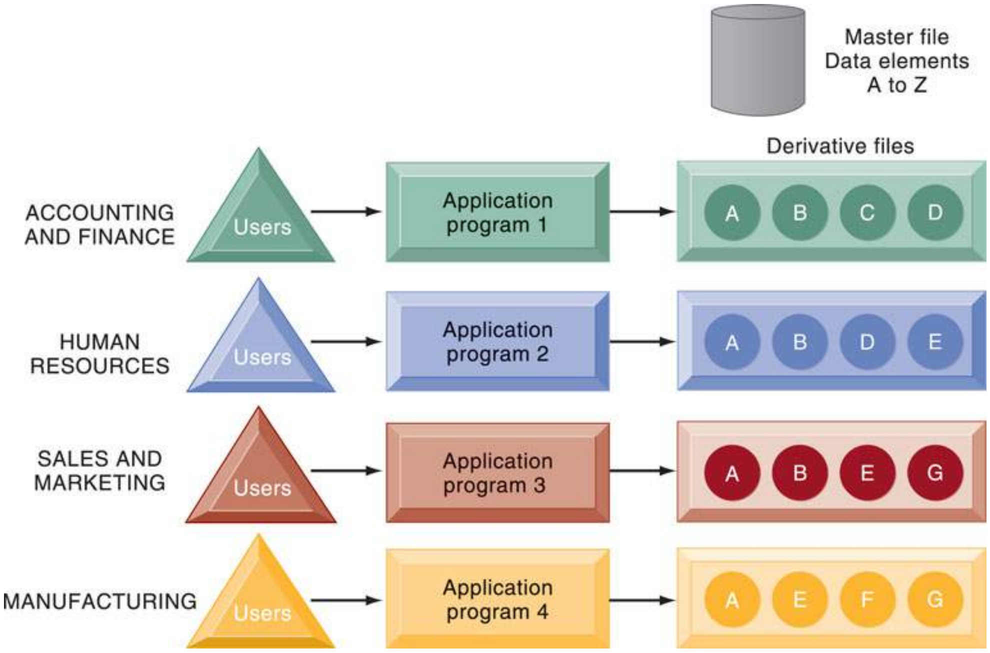
For example, the records in Figure 6-1 could constitute a student course file. A group of related files makes up a **database**. The student course file illustrated in Figure 6-1 could be grouped with files on students' personal histories and financial backgrounds to create a student database.

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

PROBLEMS WITH THE TRADITIONAL FILE ENVIRONMENT

In most organizations, systems tended to grow independently without a company-wide plan. Accounting, finance, manufacturing, human resources, and sales and marketing all developed their own systems and data files. Figure 6-2 illustrates the traditional approach to information processing.

FIGURE 6-2 TRADITIONAL FILE PROCESSING



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

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

Data Redundancy and Inconsistency

Data redundancy is the presence of duplicate data in multiple data files so that the same data are stored in more than place or location. Data redundancy occurs when different groups in an organization independently collect the same piece of data and store it independently of each other. Data redundancy wastes storage resources and also leads to **data inconsistency**, where the same attribute may have different values. For example, in instances of the entity COURSE illustrated in Figure 6-1, the Date may be updated in some systems but not in others. The same attribute, Student_ID, may also have different names in different systems throughout the organization. Some systems might use Student_ID and others might use ID, for example.

Additional confusion might result from using different coding systems to represent values for an attribute. For instance, the sales, inventory, and manufacturing systems of a clothing retailer might use different codes to represent clothing size. One system might represent clothing size as "extra large," whereas another might use the code "XL" for the same purpose. The resulting confusion would make it difficult for companies to create customer relationship management, supply chain management, or enterprise systems that integrate data from different sources.

Program-Data Dependence

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

Lack of Flexibility

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

Poor Security

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

Lack of Data Sharing and Availability

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

6.2 The Database Approach to Data Management

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

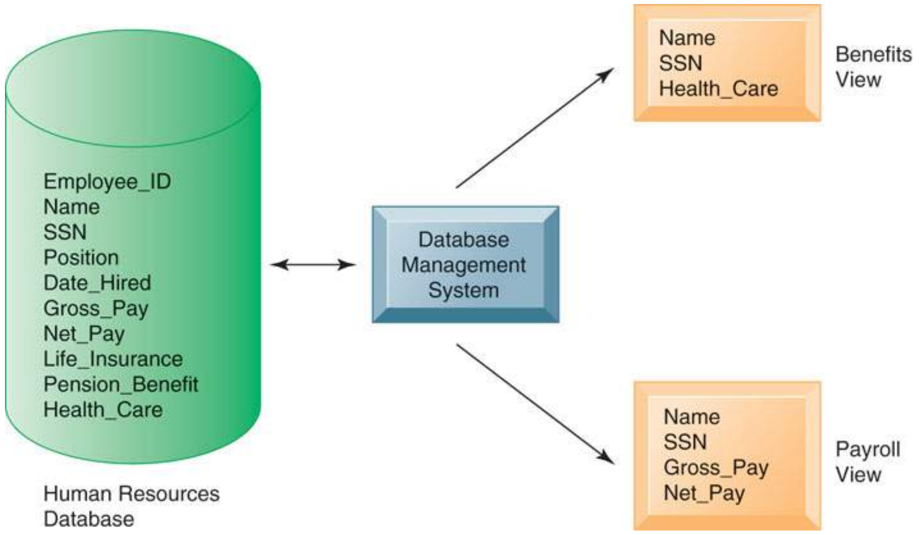
DATABASE MANAGEMENT SYSTEMS

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

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

The database management software makes the physical database available for different logical views required by users. For example, for the human resources database illustrated in Figure 6-3, a benefits specialist might require a view consisting of the employee's name, social security number, and health insurance coverage. A payroll department member might need data such as the employee's name, social security number, gross pay, and net pay. The data for

FIGURE 6-3 HUMAN RESOURCES DATABASE WITH MULTIPLE VIEWS



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

all these views are stored in a single database, where they can be more easily managed by the organization.

How a DBMS Solves the Problems of the Traditional File Environment

A DBMS reduces data redundancy and inconsistency by minimizing isolated files in which the same data are repeated. The DBMS may not enable the organization to eliminate data redundancy entirely, but it can help control redundancy. Even if the organization maintains some redundant data, using a DBMS eliminates data inconsistency because the DBMS can help the organization ensure that every occurrence of redundant data has the same values. The DBMS uncouples programs and data, enabling data to stand on their own. Access and availability of information will be increased and program development and maintenance costs reduced because users and programmers can perform ad hoc queries of data in the database. The DBMS enables the organization to centrally manage data, their use, and security.

Relational DBMS

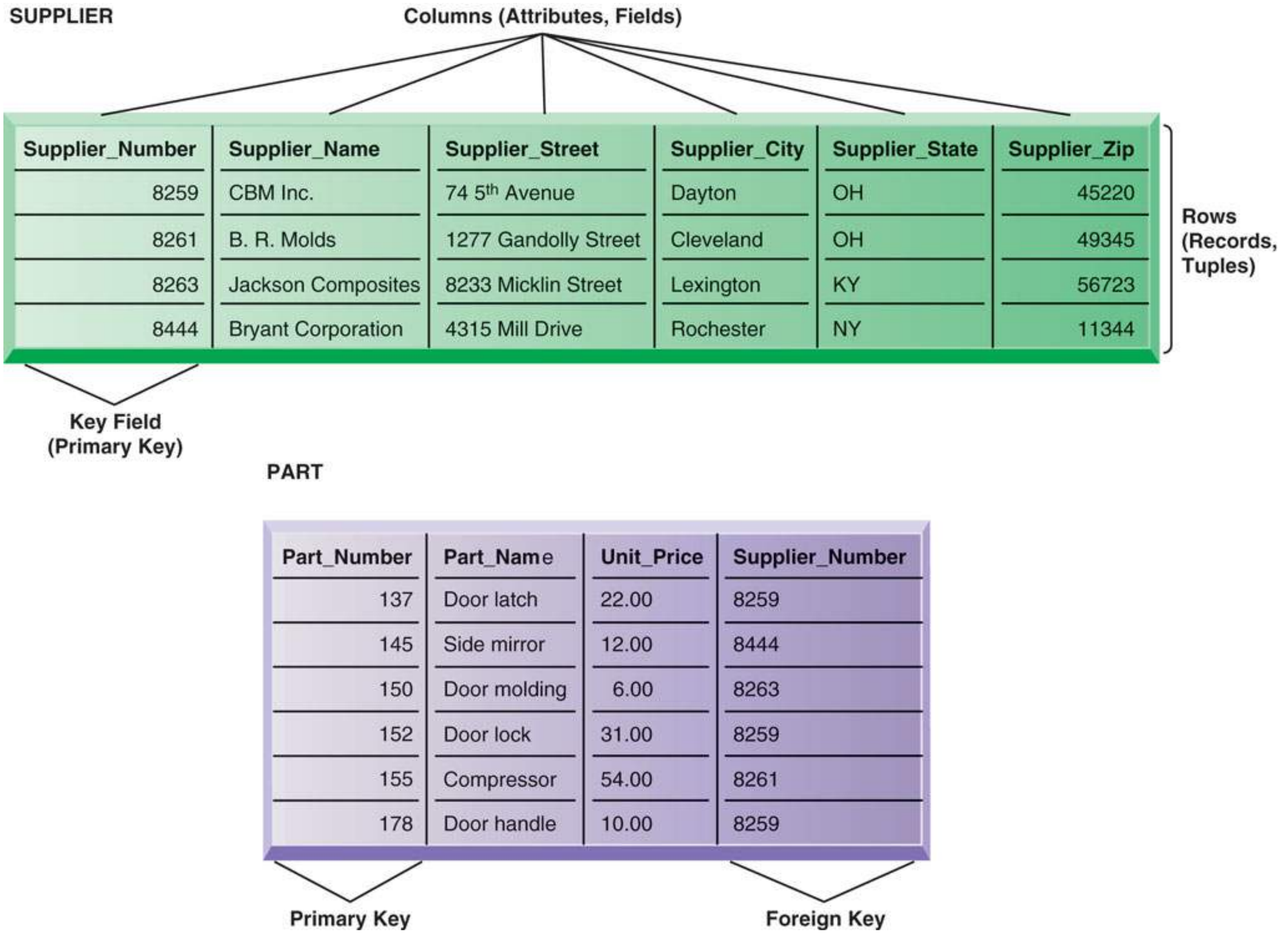
Contemporary DBMS use different database models to keep track of entities, attributes, and relationships. The most popular type of DBMS today for PCs as well as for larger computers and mainframes is the **relational DBMS**. Relational databases represent data as two-dimensional tables (called relations).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

Let's look at how a relational database organizes data about suppliers and parts (see Figure 6-4). The database has a separate table for the entity SUPPLIER and a table for the entity PART. Each table consists of a grid of columns and rows of data. Each individual element of data for each entity is stored as a separate field, and each field represents an attribute for that entity. Fields in a relational database are also called columns. For the entity SUPPLIER, the supplier identification number, name, street, city, state, and ZIP code are stored as separate fields within the SUPPLIER table and each field represents an attribute for the entity SUPPLIER.

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

The field for Supplier_Number in the SUPPLIER table uniquely identifies each record so that the record can be retrieved, updated, or sorted and it is called a **key field**. Each table in a relational database has one field that is designated as its **primary key**. This key field is the unique identifier for all the information in any row of the table and this primary key cannot be duplicated.

FIGURE 6-4 RELATIONAL DATABASE TABLES


SUPPLIER
Columns (Attributes, Fields)

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 Composites	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Rows (Records, Tuples)
Key Field (Primary Key)
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

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

Supplier_Number is the primary key for the SUPPLIER table and Part_Number is the primary key for the PART table. Note that Supplier_Number appears in both the SUPPLIER and PART tables. In the SUPPLIER table, Supplier_Number is the primary key. When the field Supplier_Number appears in the PART table it is called a **foreign key** and is essentially a lookup field to look up data about the supplier of a specific part.

Operations of a Relational DBMS

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

In a relational database, three basic operations, as shown in Figure 6-5, are used to develop useful sets of data: select, join, and project. The *select* operation creates a subset consisting of all records in the file that meet stated criteria. Select creates, in other words, a subset of rows that meet certain criteria. In our example, we want to select records (rows) from the PART table where the Part_Number equals 137 or 150. The *join* operation combines relational tables to provide the user with more information than is available in individual tables. In our example, we want to join the now-shortened PART table (only parts 137 or 150 will be presented) and the SUPPLIER table into a single new table.

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

Object-Oriented DBMS

Many applications today and in the future require databases that can store and retrieve not only structured numbers and characters but also drawings, images, photographs, voice, and full-motion video. DBMS designed for organizing structured data into rows and columns are not well suited to handling graphics-based or multimedia applications. Object-oriented databases are better suited for this purpose.

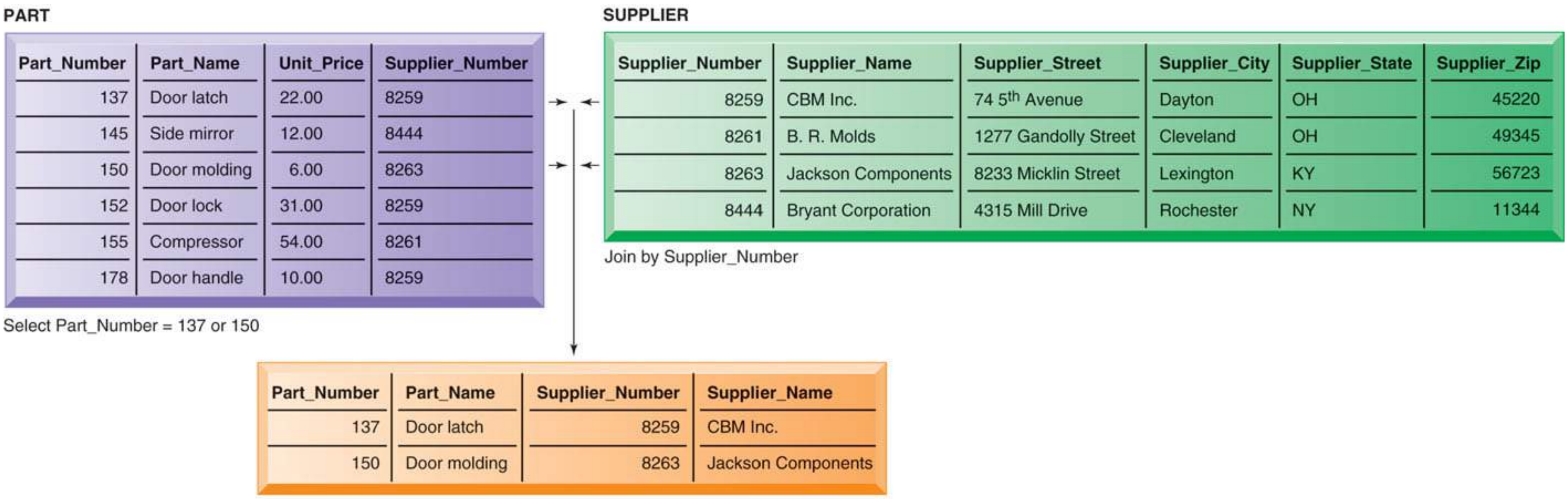
An **object-oriented DBMS** stores the data and procedures that act on those data as objects that can be automatically retrieved and shared. Object-oriented database management systems (OODBMS) are becoming popular because they can be used to manage the various multimedia components or Java applets used in Web applications, which typically integrate pieces of information from a variety of sources.

Although object-oriented databases can store more complex types of information than relational DBMS, they are relatively slow compared with relational DBMS for processing large numbers of transactions. Hybrid **object-relational DBMS** systems are now available to provide capabilities of both object-oriented and relational DBMS.

Databases in the Cloud

Suppose your company wants to use cloud computing services. Is there a way to manage data in the cloud? The answer is a qualified "Yes." Cloud computing providers offer database management services, but these services typically have less functionality than their on-premises counterparts. At the moment,

FIGURE 6-5 THE THREE BASIC OPERATIONS OF A RELATIONAL DBMS



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

Select Part_Number = 137 or 150

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

Join by Supplier_Number

Part_Number	Part_Name	Supplier_Number	Supplier_Name
137	Door latch	8259	CBM Inc.
150	Door molding	8263	Jackson Components

Project selected columns

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

the primary customer base for cloud-based data management consists of Web-focused start-ups or small to medium-sized businesses looking for database capabilities at a lower price than a standard relational DBMS.

Amazon Web Services has both a simple non-relational database called SimpleDB and a Relational Database Service, which is based on an online implementation of the MySQL open source DBMS. Amazon Relational Database Service (Amazon RDS) offers the full range of capabilities of MySQL. Pricing is based on usage. (Charges run from 11 cents per hour for a small database using 1.7 GB of server memory to $3.10 per hour for a large database using 68 GB of server memory.) There are also charges for the volume of data stored, the number of input-output requests, the amount of data written to the database, and the amount of data read from the database.

Amazon Web Services additionally offers Oracle customers the option to license Oracle Database 11g, Oracle Enterprise Manager, and Oracle Fusion Middleware to run on the Amazon EC2 (Elastic Cloud Compute) platform.

Microsoft SQL Azure Database is a cloud-based relational database service based on Microsoft's SQL Server DBMS. It provides a highly available, scalable database service hosted by Microsoft in the cloud. SQL Azure Database helps reduce costs by integrating with existing software tools and providing symmetry with on-premises and cloud databases.

TicketDirect, which sells tickets to concerts, sporting events, theater performances, and movies in Australia and New Zealand, adopted the SQL Azure Database cloud platform in order to improve management of peak system loads during major ticket sales. It migrated its data to the SQL Azure database. By moving to a cloud solution, TicketDirect is able to scale its computing resources in response to real-time demand while keeping costs low.

CAPABILITIES OF DATABASE MANAGEMENT SYSTEMS

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

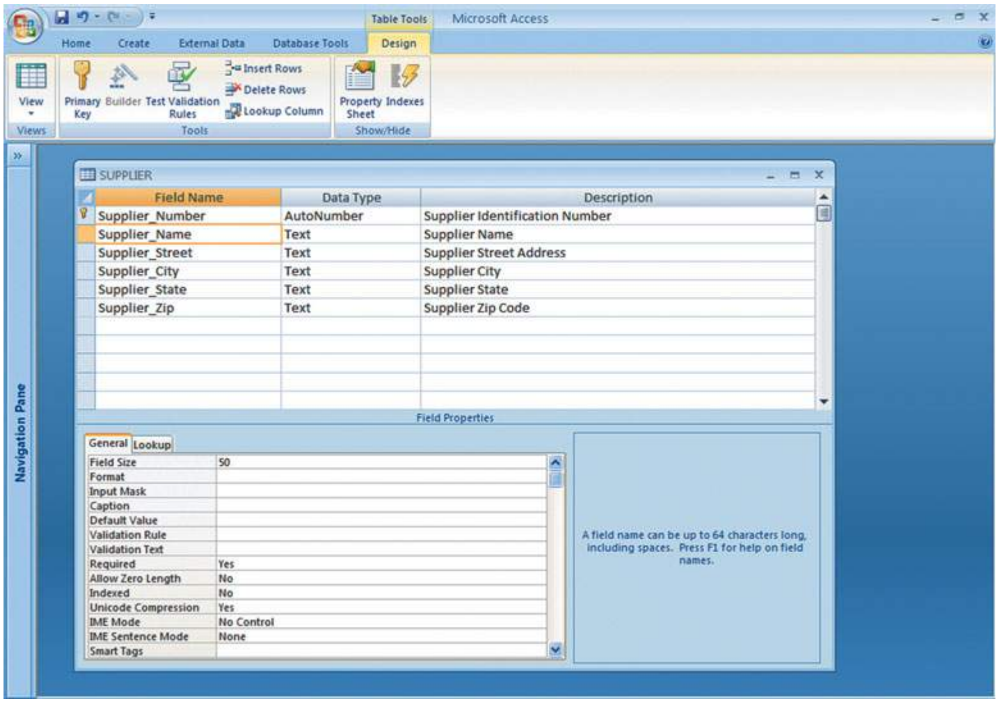
DBMS have a **data definition** capability to specify the structure of the content of the database. It would be used to create database tables and to define the characteristics of the fields in each table. This information about the database would be documented in a data dictionary. A **data dictionary**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

Microsoft Access has a rudimentary data dictionary capability that displays information about the name, description, size, type, format, and other properties of each field in a table (see Figure 6-6). Data dictionaries for large corporate databases may capture additional information, such as usage, ownership (who in the organization is responsible for maintaining the data), authorization; security, and the individuals, business functions, programs, and reports that use each data element.

Querying and Reporting

DBMS includes tools for accessing and manipulating information in databases. Most DBMS have a specialized language called a **data manipulation language** that is used to add, change, delete, and retrieve the data in the database. This language contains commands that permit end users and programming specialists to extract data from the database to satisfy information requests and develop applications. The most prominent data manipulation language today is **Structured Query Language**, or **SQL**. Figure 6-7 illustrates the SQL query that

FIGURE 6-6 MICROSOFT ACCESS DATA DICTIONARY FEATURES



Field Name	Data Type	Description
Supplier_Number	AutoNumber	Supplier Identification Number
Supplier_Name	Text	Supplier Name
Supplier_Street	Text	Supplier Street Address
Supplier_City	Text	Supplier City
Supplier_State	Text	Supplier State
Supplier_Zip	Text	Supplier Zip Code

Field Size	50
Format	
Input Mask	
Caption	
Default Value	
Validation Rule	
Validation Text	
Required	Yes
Allow Zero Length	No
Indexed	No
Unicode Compression	Yes
IME Mode	No Control
IME Sentence Mode	None
Smart Tags	

A field name can be up to 64 characters long, including spaces. Press F1 for help on field names.

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.

would produce the new resultant table in Figure 6-5. You can find out more about how to perform SQL queries in our Learning Tracks for this chapter.

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

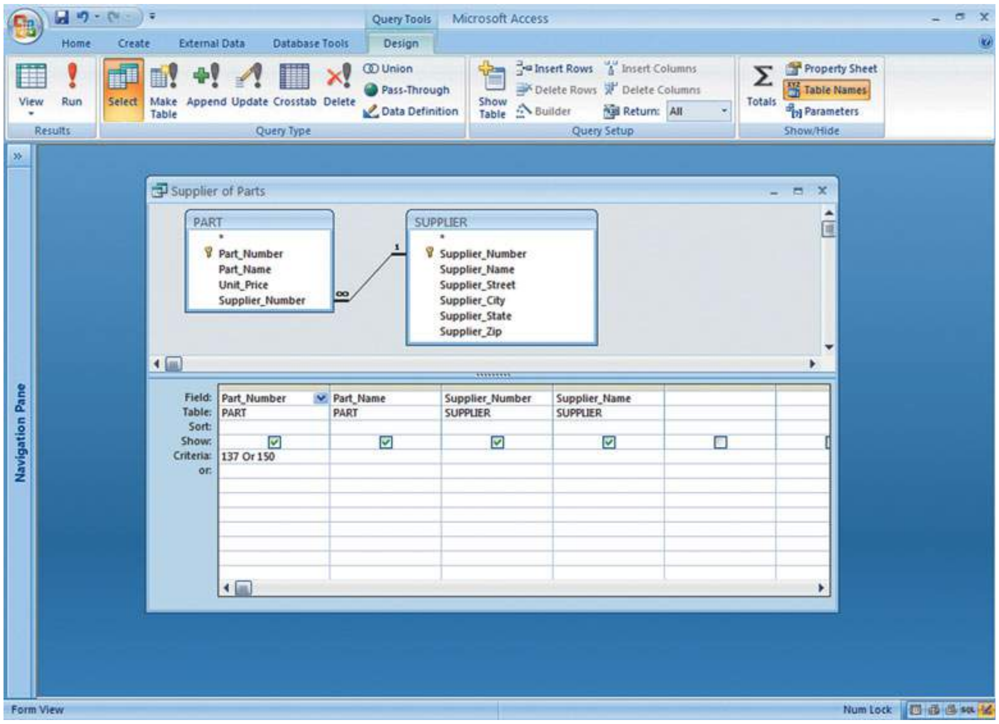
In Microsoft Access, you will find features that enable users to create queries by identifying the tables and fields they want and the results, and then selecting the rows from the database that meet particular criteria. These actions in turn are translated into SQL commands. Figure 6-8 illustrates how

FIGURE 6-7 EXAMPLE OF AN SQL QUERY

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

Illustrated here are the SQL statements for a query to select suppliers for parts 137 or 150. They produce a list with the same results as Figure 6-5.

FIGURE 6-8 AN ACCESS QUERY



Field:	Part_Number	Part_Name	Supplier_Number	Supplier_Name	
Table:	PART	PART	SUPPLIER	SUPPLIER	
Sort:					
Show:	☑	☑	☑	☑	☐
Criteria:	137 Or 150				
or:					

Illustrated here is how the query in Figure 6-7 would be constructed using Microsoft Access query-building tools. It shows the tables, fields, and selection criteria used for the query.

the same query as the SQL query to select parts and suppliers would be constructed using the Microsoft query-building tools.

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

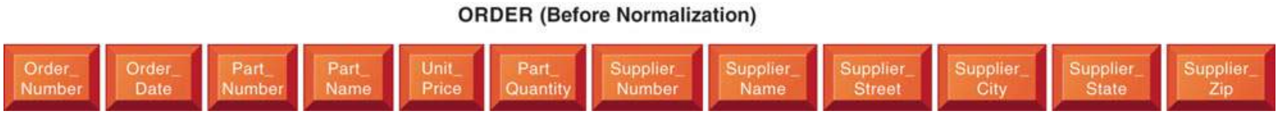
DESIGNING DATABASES

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

Normalization and Entity-Relationship Diagrams

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

FIGURE 6-9 AN UNNORMALIZED RELATION FOR ORDER



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

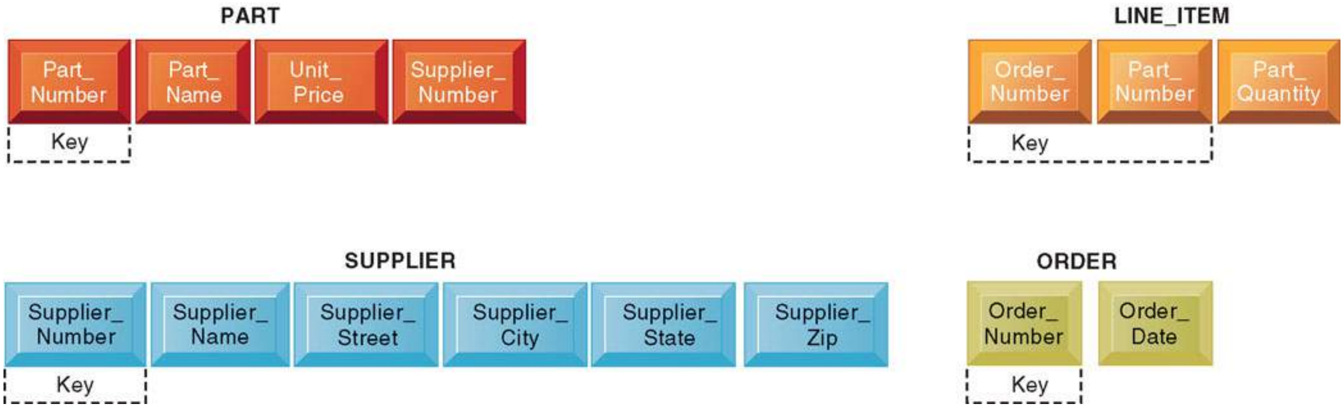
application programs. Groups of data are organized, refined, and streamlined until an overall logical view of the relationships among all the data in the database emerges.

To use a relational database model effectively, complex groupings of data must be streamlined to minimize redundant data elements and awkward many-to-many relationships. The process of creating small, stable, yet flexible and adaptive data structures from complex groups of data is called **normalization**. Figures 6-9 and 6-10 illustrate this process.

In the particular business modeled here, an order can have more than one part but each part is provided by only one supplier. If we build a relation called ORDER with all the fields included here, we would have to repeat the name and address of the supplier for every part on the order, even though the order is for parts from a single supplier. This relationship contains what are called repeating data groups because there can be many parts on a single order to a given supplier. A more efficient way to arrange the data is to break down ORDER into smaller relations, each of which describes a single entity. If we go step by step and normalize the relation ORDER, we emerge with the relations illustrated in Figure 6-10. You can find out more about normalization, entity-relationship diagramming, and database design in the Learning Tracks for this chapter.

Relational database systems try to enforce **referential integrity** rules to ensure that relationships between coupled tables remain consistent. When one table has a foreign key that points to another table, you may not add a record to the table with the foreign key unless there is a corresponding record in the linked table. In the database we examined earlier in this chapter, the foreign key

FIGURE 6-10 NORMALIZED TABLES CREATED FROM ORDER



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

Supplier_Number links the PART table to the SUPPLIER table. We may not add a new record to the PART table for a part with Supplier_Number 8266 unless there is a corresponding record in the SUPPLIER table for Supplier_Number 8266. We must also delete the corresponding record in the PART table if we delete the record in the SUPPLIER table for Supplier_Number 8266. In other words, we shouldn't have parts from nonexistent suppliers!

Database designers document their data model with an **entity-relationship diagram**, illustrated in Figure 6-11. This diagram illustrates the relationship between the entities SUPPLIER, PART, LINE_ITEM, and ORDER. The boxes represent entities. The lines connecting the boxes represent relationships. A line connecting two entities that ends in two short marks designates a one-to-one relationship. A line connecting two entities that ends with a crow's foot topped by a short mark indicates a one-to-many relationship. Figure 6-11 shows that one ORDER can contain many LINE_ITEMs. (A PART can be ordered many times and appear many times as a line item in a single order.) Each PART can have only one SUPPLIER, but many PARTs can be provided by the same SUPPLIER.

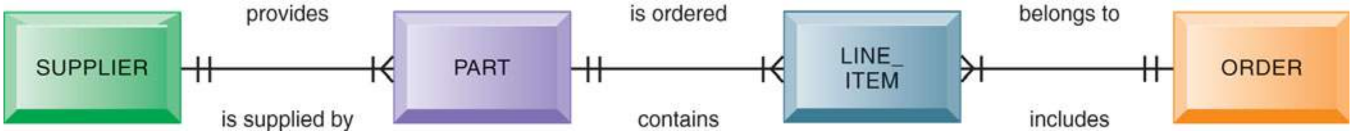
It can't be emphasized enough: If the business doesn't get its data model right, the system won't be able to serve the business well. The company's systems will not be as effective as they could be because they'll have to work with data that may be inaccurate, incomplete, or difficult to retrieve. Understanding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

For example, Famous Footwear, a shoe store chain with more than 800 locations in 49 states, could not achieve its goal of having "the right style of shoe in the right store for sale at the right price" because its database was not properly designed for rapidly adjusting store inventory. The company had an Oracle relational database running on an IBM AS/400 midrange computer, but the database was designed primarily for producing standard reports for management rather than for reacting to marketplace changes. Management could not obtain precise data on specific items in inventory in each of its stores. The company had to work around this problem by building a new database where the sales and inventory data could be better organized for analysis and inventory management.

6.3 Using Databases to Improve Business Performance and Decision Making

Businesses use their databases to keep track of basic transactions, such as paying suppliers, processing orders, keeping track of customers, and paying employees. But they also need databases to provide information that will help the company

FIGURE 6-11 AN ENTITY-RELATIONSHIP DIAGRAM



This diagram shows the relationships between the entities SUPPLIER, PART, LINE_ITEM, and ORDER that might be used to model the database in Figure 6-10.

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.

For example, by analyzing data from customer credit card purchases, Louise's Trattoria, a Los Angeles restaurant chain, learned that quality was more important than price for most of its customers, who were college-educated and liked fine wine. Acting on this information, the chain introduced vegetarian dishes, more seafood selections, and more expensive wines, raising sales by more than 10 percent.

In a large company, with large databases or large systems for separate functions, such as manufacturing, sales, and accounting, special capabilities and tools are required for analyzing vast quantities of data and for accessing data from multiple systems. These capabilities include data warehousing, data mining, and tools for accessing internal databases through the Web.

DATA WAREHOUSES

Suppose you want concise, reliable information about current operations, trends, and changes across the entire company If you worked in a large company, obtaining this might be difficult because data are often maintained in separate systems, such as sales, manufacturing, or accounting. Some of the data you need might be found in the sales system, and other pieces in the manufacturing system. Many of these systems are older legacy systems that use outdated data management technologies or file systems where information is difficult for users to access.

You might have to spend an inordinate amount of time locating and gathering the data you need, or you would be forced to make your decision based on incomplete knowledge. If you want information about trends, you might also have trouble finding data about past events because most firms only make their current data immediately available. Data warehousing addresses these problems.

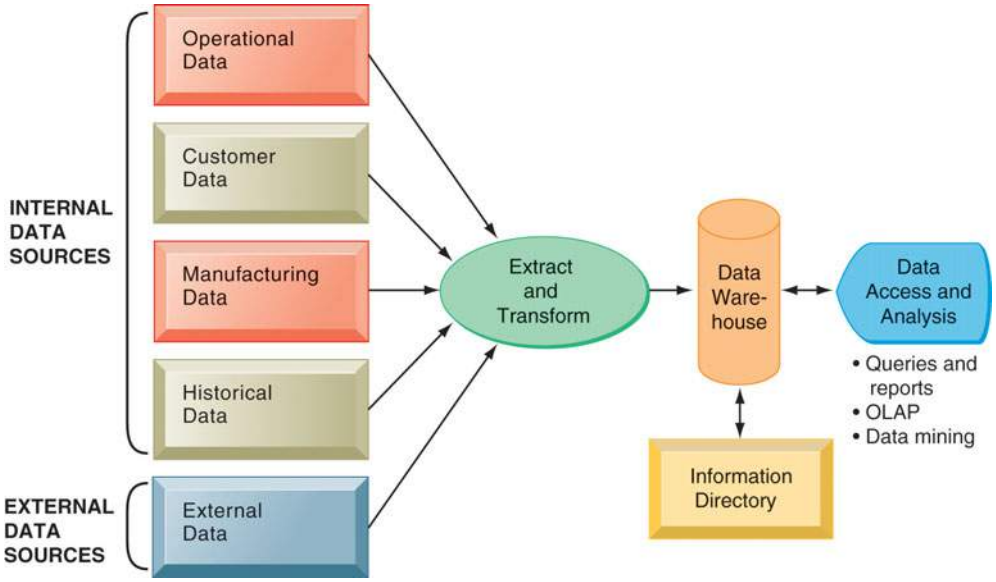
What Is a Data Warehouse?

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 Web site transactions. The data warehouse consolidates and standardizes information from different operational databases so that the information can be used across the enterprise for management analysis and decision making.

Figure 6-12 illustrates how a data warehouse works. 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. Many firms use intranet portals to make the data warehouse information widely available throughout the firm.

Catalina Marketing, a global marketing firm for major consumer packaged goods companies and retailers, operates a gigantic data warehouse that includes three years of purchase history for 195 million U.S. customer loyalty program members at supermarkets, pharmacies, and other retailers. It is the largest loyalty database in the world. Catalina's retail store customers analyze this database of customer purchase histories to determine individual customers' buying preferences. When a shopper checks out at the cash register of one of

FIGURE 6-12 **COMPONENTS OF A DATA WAREHOUSE**



The data warehouse extracts current and historical data from multiple operational systems inside the organization. These data are combined with data from external sources and reorganized into a central database designed for management reporting and analysis. The information directory provides users with information about the data available in the warehouse.

Catalina's retail customers, the purchase is instantly analyzed along with that customer's buying history in the data warehouse to determine what coupons that customer will receive at checkout along with a receipt.

The U.S. Internal Revenue Service (IRS) maintains a Compliance Data Warehouse that consolidates taxpayer data that had been fragmented among many different legacy systems, including personal information about taxpayers and archived tax returns. These systems had been designed to process tax return forms efficiently but their data were very difficult to query and analyze. The Compliance Data Warehouse integrates taxpayer data from many disparate sources into a relational structure, which makes querying and analysis much easier. With a complete and comprehensive picture of taxpayers, the warehouse helps IRS analysts and staff identify people who are most likely to cheat on their income tax payments and respond rapidly to taxpayer queries.

Data Marts

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. Before implementing an enterprise-wide data warehouse, bookseller Barnes & Noble maintained 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. A data mart typically focuses on a single subject area or line of business, so it usually can be constructed more rapidly and at lower cost than an enterprise-wide data warehouse.

TOOLS FOR BUSINESS INTELLIGENCE: MULTIDIMENSIONAL DATA ANALYSIS AND DATA MINING

Once data have been captured and organized in data warehouses and data marts, they are available for further analysis using tools for business intelligence, which we introduced briefly in Chapter 2. Business intelligence tools enable users to analyze data to see new patterns, relationships, and insights that are useful for guiding decision making.

Principal tools for business intelligence include software for database querying and reporting, tools for multidimensional data analysis (online analytical processing), and tools for data mining. This section will introduce you to these tools, with more detail about business intelligence analytics and applications in the Chapter 12 discussion of decision making.

Online Analytical Processing (OLAP)

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

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

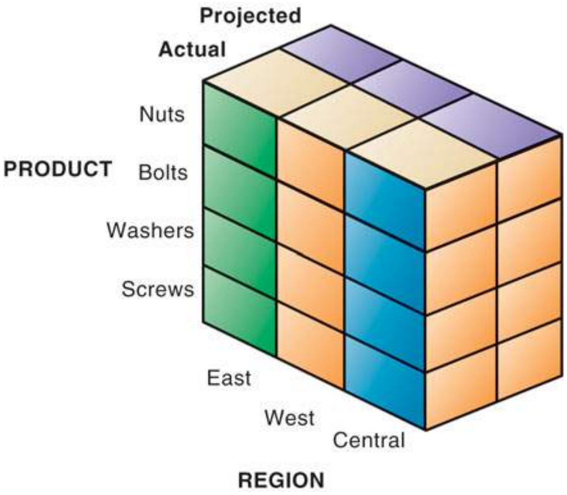
Figure 6-13 shows a multidimensional model that could be created to represent products, regions, actual sales, and projected sales. A matrix of actual sales can be stacked on top of a matrix of projected sales to form a cube with six faces. If you rotate the cube 90 degrees one way, the face showing will be product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. If you rotate 180 degrees from the original view, you will see projected sales and product versus region. Cubes can be nested within cubes to build complex views of data. A company would use either a specialized multidimensional database or a tool that creates multidimensional views of data in relational databases.

Data Mining

Traditional database queries answer such questions as, “How many units of product number 403 were shipped in February 2010?” 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

FIGURE 6-13 MULTIDIMENSIONAL DATA MODEL



The view that is showing is product versus region. If you rotate the cube 90 degrees, the face will show 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.

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.

For example, Harrah's Entertainment, the second-largest gambling company in its industry, uses data mining to identify its most profitable customers and generate more revenue from them. The company continually analyzes data about its customers gathered when people play its slot machines or use Harrah's casinos and hotels. Harrah's marketing department uses this information to build a detailed gambling profile, based on a particular customer's ongoing value to the company. For instance, data mining lets Harrah's know the favorite gaming experience of a regular customer at one of its Midwest riverboat casinos, along with that person's preferences for room accomodations, 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 has improved Harrah's profits so much that it has become the centerpiece of the firm's business strategy.

Predictive analytics use data mining techniques, historical data, and assumptions about future conditions to predict outcomes of events, such as the probability a customer will respond to an offer or purchase a specific product. For example, the U.S. division of The Body Shop International plc used predictive analytics with its database of catalog, Web, and retail store customers to identify customers who were more likely to make catalog purchases. That information helped the company build a more precise and targeted mailing list for its catalogs, improving the response rate for catalog mailings and catalog revenues.

Text Mining and Web Mining

Business intelligence tools deal primarily with data that have been structured in databases and files. However, unstructured data, most in the form of text files, is believed to account for over 80 percent of an organization's useful information. E-mail, memos, call center transcripts, survey responses, legal cases, patent descriptions, and service reports are all valuable for finding patterns and trends that will help employees make better business decisions. **Text mining** tools are now available to help businesses analyze these data. These tools are able to extract key elements from large unstructured 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.

Text mining is a relatively new technology, but what's really new are the myriad ways in which unstructured data are being generated by consumers and the business uses for these data. The Interactive Session on Technology explores some of these business applications of text mining.

The Web is another rich source of valuable information, some of which can now be mined for patterns, trends, and insights into customer behavior. The discovery and analysis of useful patterns and information from the World Wide Web is called **Web mining**. Businesses might turn to Web mining to help them understand customer behavior, evaluate the effectiveness of a particular Web site, or quantify the success of a marketing campaign. For instance, marketers use Google Trends and Google Insights for Search services, which track 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.

INTERACTIVE SESSION: TECHNOLOGY

WHAT CAN BUSINESSES LEARN FROM TEXT MINING?

Text mining is the discovery of patterns and relationships from large sets of unstructured data—the kind of data we generate in e-mails, phone conversations, blog postings, online customer surveys, and tweets. The mobile digital platform has amplified the explosion in digital information, with hundreds of millions of people calling, texting, searching, "apping" (using applications), buying goods, and writing billions of e-mails on the go.

Consumers today are more than just consumers: they have more ways to collaborate, share information, and influence the opinions of their friends and peers, and the data they create in doing so have significant value to businesses. Unlike structured data, which are generated from events such as completing a purchase transaction, unstructured data have no distinct form. Nevertheless, managers believe such data may offer unique insights into customer behavior and attitudes that were much more difficult to determine years ago.

For example, in 2007, JetBlue experienced unprecedented levels of customer discontent in the wake of a February ice storm that resulted in widespread flight cancellations and planes stranded on Kennedy Airport runways. The airline received 15,000 e-mails per day from customers during the storm and immediately afterwards, up from its usual daily volume of 400. The volume was so much larger than usual that JetBlue had no simple way to read everything its customers were saying.

Fortunately, the company had recently contracted with Attensity, a leading vendor of text analytics software, and was able to use the software to analyze all of the e-mail it had received within two days. According to JetBlue research analyst Bryan Jeppsen, Attensity Analyze for Voice of the Customer (VoC) enabled JetBlue to rapidly extract customer sentiments, preferences, and requests it couldn't find any other way. This tool uses a proprietary technology to automatically identify facts, opinions, requests, trends, and trouble spots from the unstructured text of survey responses, service notes, e-mail messages, Web forums, blog entries, news articles, and other customer communications. The technology is able to accurately and automatically identify the many different "voices" customers use to express their feedback (such as a negative voice, positive voice, or conditional voice), which helps organizations pinpoint key events and relationships, such as intent to buy, intent to leave, or customer "wish" events. It can reveal specific product and service issues, reactions to marketing and public relations efforts, and even buying signals.

Attensity's software integrated with JetBlue's other customer analysis tools, such as Satmetrix's Net Promoter metrics, which classifies customers into groups that are generating positive, negative, or no feedback about the company. Using Attensity's text analytics in tandem with these tools, JetBlue developed a customer bill of rights that addressed the major issues customers had with the company.

Hotel chains like Gaylord Hotels and Choice Hotels are using text mining software to glean insights from thousands of customer satisfaction surveys provided by their guests. Gaylord Hotels is using Clarabridge's text analytics solution delivered via the Internet as a hosted software service to gather and analyze customer feedback from surveys, e-mail, chat messaging, staffed call centers, and online forums associated with guests' and meeting planners' experiences at the company's convention resorts. The Clarabridge software sorts through the hotel chain's customer surveys and gathers positive and negative comments, organizing them into a variety of categories to reveal less obvious insights. For example, guests complained about many things more frequently than noisy rooms, but complaints of noisy rooms were most frequently correlated with surveys indicating an unwillingness to return to the hotel for another stay.

Analyzing customer surveys used to take weeks, but now takes only days, thanks to the Clarabridge software. Location managers and corporate executives have also used findings from text mining to influence decisions on building improvements.

Wendy's International adopted Clarabridge software to analyze nearly 500,000 messages it collects each year from its Web-based feedback forum, call center notes, e-mail messages, receipt-based surveys, and social media. The chain's customer satisfaction team had previously used spreadsheets and keyword searches to review customer comments, a very slow manual approach. Wendy's management was looking for a better tool to speed analysis, detect emerging issues, and pinpoint troubled areas of the business at the store, regional, or corporate level.

The Clarabridge technology enables Wendy's to track customer experiences down to the store level within minutes. This timely information helps store, regional, and corporate managers spot and address problems related to meal quality, cleanliness, and speed of service.

Text analytics software caught on first with government agencies and larger companies with information systems departments that had the means to properly use the complicated software, but Clarabridge is now offering a version of its product geared towards small businesses. The technology has already caught on with law enforcement, search tool interfaces, and "listening platforms" like Nielsen Online. Listening platforms are text mining tools that focus on brand management, allowing companies to determine how consumers feel about their brand and take steps to respond to negative sentiment.

Structured data analysis won't be rendered obsolete by text analytics, but companies that are able to use both methods to develop a clearer picture of their customers' attitudes will have an easier time establishing and building their brand and gleaning insights that will enhance profitability.

Sources: Doug Henschen, "Wendy's Taps Text Analytics to Mine Customer Feedback," *Information Week*, March 23, 2010; David Stodder," How Text Analytics Drive Customer Insight" *Information Week*, February 1, 2010; Nancy David Kho, "Customer Experience and Sentiment Analysis," *KMWorld*, February 1, 2010; Siobhan Gorman, "Details of Einstein Cyber-Shield Disclosed by White House," *The Wall Street Journal*, March 2, 2010; www.attensity.com, accessed June 16, 2010; and www.clarabridge.com, accessed June 17, 2010.

CASE STUDY QUESTIONS

1. What challenges does the increase in unstructured data present for businesses?

2. How does text-mining improve decision-making?

3. What kinds of companies are most likely to benefit from text mining software? Explain your answer.

4. In what ways could text mining potentially lead to the erosion of personal information privacy? Explain.

MIS IN ACTION

Visit a Web site such as QVC.com or TripAdvisor.com detailing products or services that have customer reviews. Pick a product, hotel, or other service with at least 15 customer reviews and read those reviews, both positive and negative. How could Web content mining help the offering company improve or better market this product or service? What pieces of information should highlighted?

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 extracts useful information from the links embedded in Web documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. Web usage mining examines user interaction data recorded by a Web server whenever requests for a Web site's resources are received. The usage data records the user's behavior when the user browses or makes transactions on the Web site 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

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

Suppose, for example, a customer with a Web browser wants to search an online retailer's database for pricing information. Figure 6-14 illustrates how that customer might access the retailer's internal database over the Web. The user accesses the retailer's Web site over the Internet using Web browser software on his or her client PC. The user's Web browser software requests 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 passes these requests for data to software that translates HTML commands into SQL so that they can be processed by the DBMS working with the database. In a client/server environment, the DBMS resides on a dedicated computer called a **database server**. The DBMS receives the SQL requests and provides the required data. The middleware transfers information from the organization's internal database back to the Web server for delivery in the form of a Web page to the user.

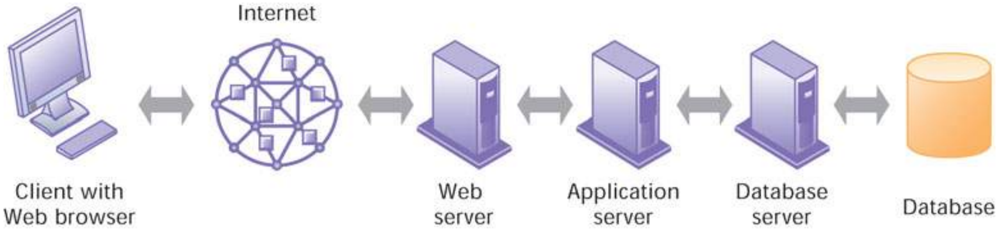
Figure 6-14 shows that the middleware working between the Web server and the DBMS is an application server running on its own dedicated computer (see Chapter 5). The application server software handles all application operations, including transaction processing and data access, between browser-based computers and a company's back-end business applications or databases. The application server takes requests from the Web server, runs the business logic to process transactions based on those requests, and provides connectivity to the organization's back-end systems or databases. Alternatively, the software for handling these operations could be a custom program or a CGI script. A CGI script is a compact program using the *Common Gateway Interface (CGI)* specification for processing data on a Web server.

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

Accessing corporate databases through the Web is creating new efficiencies, opportunities, and business models. ThomasNet.com provides an up-to-date online directory of more than 600,000 suppliers of industrial products, such as chemicals, metals, plastics, rubber, and automotive equipment. Formerly called Thomas Register, the company used to send out huge paper catalogs with this information. Now it provides this information to users online via its Web site and has become a smaller, leaner company.

Other companies have created entirely new businesses based on access to large databases through the Web. One is the social networking site MySpace, which helps users stay connected with each other or meet new people.

FIGURE 6-14 LINKING INTERNAL DATABASES TO THE WEB



Users access an organization's internal database through the Web using their desktop PCs and Web browser software.

MySpace features music, comedy, videos, and “profiles” with information supplied by 122 million users about their age, hometown, dating preferences, marital status, and interests. It maintains a massive database to house and manage all of this content. Facebook uses a similar database.

6.4 Managing Data Resources

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

ESTABLISHING AN INFORMATION POLICY

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

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

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

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

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

ENSURING DATA QUALITY

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

What would happen if a customer's telephone number or account balance were incorrect? What would be the impact if the database had the wrong price for the product you sold or your sales system and inventory system showed different prices for the same product? Data that are inaccurate, untimely, or inconsistent with other sources of information lead to incorrect decisions, product recalls, and financial losses. Inaccurate data in criminal justice and national security databases might even subject you to unnecessarily surveillance or detention, as described in the chapter-ending case study.

According to Forrester Research, 20 percent of U.S. mail and commercial package deliveries were returned because of incorrect names or addresses. 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. (Gartner, 2007).

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

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

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

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

Data quality problems are not just business problems. They also pose serious problems for individuals, affecting their financial condition and even their jobs. The Interactive Session on Organizations describes some of these impacts, as it details the data quality problems found in the companies that collect and report consumer credit data. As you read this case, look for the management, organization, and technology factors behind this problem, and whether existing solutions are adequate.

INTERACTIVE SESSION: ORGANIZATIONS

CREDIT BUREAU ERRORS—BIG PEOPLE PROBLEMS

You've found the car of your dreams. You have a good job and enough money for a down payment. All you need is an auto loan for $14,000. You have a few credit card bills, which you diligently pay off each month. But when you apply for the loan you're turned down. When you ask why, you're told you have an overdue loan from a bank you've never heard of. You've just become one of the millions of people who have been victimized by inaccurate or outdated data in credit bureaus' information systems.

Most data on U.S. consumers' credit histories are collected and maintained by three national credit reporting agencies: Experian, Equifax, and TransUnion. These organizations collect data from various sources to create a detailed dossier of an individual's borrowing and bill paying habits. This information helps lenders assess a person's credit worthiness, the ability to pay back a loan, and can affect the interest rate and other terms of a loan, including whether a loan will be granted in the first place. It can even affect the chances of finding or keeping a job: At least one-third of employers check credit reports when making hiring, firing, or promotion decisions.

U.S. credit bureaus collect personal information and financial data from a variety of sources, including creditors, lenders, utilities, debt collection agencies, and the courts. These data are aggregated and stored in massive databases maintained by the credit bureaus. The credit bureaus then sell this information to other companies to use for credit assessment.

The credit bureaus claim they know which credit cards are in each consumer's wallet, how much is due on the mortgage, and whether the electric bill is paid on time. But if the wrong information gets into their systems, whether through identity theft or errors transmitted by creditors, watch out! Untangling the mess can be almost impossible.

The bureaus understand the importance of providing accurate information to both lenders and consumers. But they also recognize that their own systems are responsible for many credit-report errors. Some mistakes occur because of the procedures for matching loans to individual credit reports.

The sheer volume of information being transmitted from creditors to credit bureaus increases the likelihood of mistakes. Experian, for example, updates 30 million credit reports each day and roughly 2 billion credit reports each month. It matches the identifying personal information in a credit application or credit account with the identifying personal information in a consumer credit file. Identifying personal information includes items such as name (first name, last name and middle initial), full current address and ZIP code, full previous address and ZIP code, and social security number. The new credit information goes into the consumer credit file that it best matches.

The credit bureaus rarely receive information that matches in all the fields in credit files, so they have to determine how much variation to allow and still call it a match. Imperfect data lead to imperfect matches. A consumer might provide incomplete or inaccurate information on a credit application. A creditor might submit incomplete or inaccurate information to the credit bureaus. If the wrong person matches better than anyone else, the data could unfortunately go into the wrong account.

Perhaps the consumer didn't write clearly on the account application. Name variations on different credit accounts can also result in less-than-perfect matches. Take the name Edward Jeffrey Johnson. One account may say Edward Johnson. Another may say Ed Johnson. Another might say Edward J. Johnson. Suppose the last two digits of Edward's social security number get transposed—more chance for mismatches.

If the name or social security number on another person's account partially matches the data in your file, the computer might attach that person's data to your record. Your record might likewise be corrupted if workers in companies supplying tax and bankruptcy data from court and government records accidentally transpose a digit or misread a document.

The credit bureaus claim it is impossible for them to monitor the accuracy of the 3.5 billion pieces of credit account information they receive each month. They must continually contend with bogus claims from consumers who falsify lender information or use shady credit-repair companies that challenge all the negative information on a credit report regardless of its validity. To separate the good from the bad, the credit bureaus use an automated e-OSCAR (Electronic Online Solution for Complete and Accurate Reporting) system to forward consumer disputes to lenders for verification.

If your credit report showed an error, the bureaus usually do not contact the lender directly to correct the information. To save money, the bureaus send consumer protests and evidence to a data processing center run by a third-party contractor. These contractors rapidly summarize every complaint with a short comment and 2-digit code from a menu of 26 options. For example, the code A3 designates "belongs to another individual with a similar name." These summaries are often too brief to include the background banks need to understand a complaint.

Although this system fixes large numbers of errors (data are updated or corrected for 72 percent of disputes), consumers have few options if the system fails. Consumers who file a second dispute without providing new information might have their dispute dismissed as "frivolous." If the consumer tries to contact the lender that made the error on their own, banks have no obligation to investigate the dispute—unless it's sent by a credit bureau.

Sources: Dennis McCafferty, "Bad Credit Could Cost You a Job," *Baseline*, June 7, 2010; Kristen McNamara, "Bad Credit Derails Job Seekers," *The Wall Street Journal*, March 16, 2010; Anne Kadet, Lucy Lazarony, "Your Name Can Mess Up Your Credit Report, Bankrate.com, accessed July 1, 2009; "Credit Report Fix a Headache," *Atlanta Journal-Constitution*, June 14, 2009; and "Why Credit Bureaus Can't Get It Right," *Smart Money*, March 2009.

CASE STUDY QUESTIONS

1. Assess the business impact of credit bureaus' data quality problems for the credit bureaus, for lenders, for individuals.
2. Are any ethical issues raised by credit bureaus' data quality problems? Explain your answer.
3. Analyze the management, organization, and technology factors responsible for credit bureaus' data quality problems.
4. What can be done to solve these problems?

MIS IN ACTION

Go to the Experian Web site (www.experian.com) and explore the site, with special attention to its services for businesses and small businesses. Then answer the following questions:

1. List and describe five services for businesses and explain how each uses consumer data. Describe the kinds of businesses that would use these services.
2. Explain how each of these services is affected by inaccurate consumer data.

6.5 HANDS-ON MIS PROJECTS

The projects in this section give you hands-on experience in analyzing data quality problems, establishing company-wide data standards, creating a database for inventory management, and using the Web to search online databases for overseas business resources.

Management Decision Problems

1. 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 that 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, regardless of their location. In fact, cultural differences combined with complications from absorbing companies that Emerson had acquired led to multiple ways of entering quotes, 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?
2. Your industrial supply company wants to create a data warehouse where management can obtain a single corporate-wide view of critical sales information to identify best-selling products in specific geographic areas, key customers, and sales trends. Your sales and product information are stored in several different systems: a divisional sales system running on a Unix server and a corporate sales system running on an IBM mainframe. You would like to create a single standard format that consolidates these data from both systems. The following format has been proposed.

PRODUCT _ID	PRODUCT_ DESCRIPTION	COST_PER_ UNIT	UNITS_ SOLD	SALES_REGION	DIVISION	CUSTOMER_ID

The following are sample files from the two systems that would supply the data for the data warehouse:

CORPORATE SALES SYSTEM

PRODUCT_ID	PRODUCT_ DESCRIPTION	UNIT_COST	UNITS_SOLD	SALES_TERRITORY	DIVISION
60231	Bearing, 4”	5.28	900,245	Northeast	Parts
85773	SS assembly unit	12.45	992,111	Midwest	Parts

MECHANICAL PARTS DIVISION SALES SYSTEM

PROD_NO	PRODUCT_ DESCRIPTION	COST_PER_UNIT	UNITS_SOLD	SALES_REGION	CUSTOMER_ID
60231	4” Steel bearing	5.28	900,245	N.E.	Anderson
85773	SS assembly unit	12.45	992,111	M.W.	Kelly Industries

- 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 company-wide 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

Businesses today depend on databases to provide reliable information about items in inventory, items that need restocking, and inventory costs. In this exercise, you'll 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. This rapidly growing business needs a database system to manage this information.

Initially, the database should house information about suppliers and products. The database will contain two tables: a supplier table and a product table. The reorder level refers to the number of items in inventory that triggers a decision to order more items to prevent a stockout. (In other words, if the number of units of a particular item in inventory falls below the reorder level, the item should be reordered.) The user should be able to perform several queries and produce several managerial reports based on the data contained in the two tables.

Using the information found in the tables in MyMISLab, build a simple relational database for Sylvester's. Once you have built the database, perform the following activities:

- Prepare a report that identifies the five most expensive bicycles. The report should list the bicycles in descending order from most expensive to least expensive, the quantity on hand for each, and the markup percentage for each.
- Prepare a report that lists each supplier, its products, the quantities on hand, and associated reorder levels. The report should be sorted alphabetically by supplier. 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 items identified.
- Write a brief description of how the database could be enhanced to further improve management of the business. What tables or fields should be added? What additional reports would be useful?

Improving Decision Making: Searching Online Databases for Overseas Business Resources

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

Internet users have access to many thousands of Web-enabled databases with information on services and products in faraway locations. This project develops skills in searching these online databases.

Your company is located in Greensboro, North Carolina, and manufactures office furniture of various types. You have recently acquired several new customers in Australia, and a study you commissioned indicates that, with a presence there, you could greatly increase your sales. Moreover, your study indicates that you could do even better if you actually manufactured many of your products locally (in Australia). First, you need to set up an office in Melbourne to establish a presence, and then you need to begin importing from the United States. You then can plan to start producing locally.

You will soon be traveling to the area to make plans to actually set up an office, and you want to meet with organizations that can help you with your operation. You will need to engage people or organizations that offer many services necessary for you to open your office, including lawyers, accountants, import-export experts, telecommunications equipment and support, and even trainers who can help you to prepare your future employees to work for you. Start by searching for U.S. Department of Commerce advice on doing business in Australia. Then try the following online databases to locate companies that you would like to meet with during your coming trip: Australian Business Register (abr.business.gov.au/), Australia Trade Now (australiatradenow.com/), and the Nationwide Business Directory of Australia (www.nationwide.com.au). If necessary, you could also try search engines such as Yahoo and Google. Then perform the following activities:

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

Learning Track Modules

The following Learning Tracks provide content relevant to topics covered in this chapter:

1. Database Design, Normalization, and Entity-Relationship Diagramming
2. Introduction to SQL
3. Hierarchical and Network Data Models

Review Summary

1. ***What are the problems of managing data resources in a traditional file environment and how are they solved by a database management system?***

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

2. ***What are the major capabilities of DBMS and why is a relational DBMS so powerful?***

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

The relational database is the primary method for organizing and maintaining data today in information systems because it is so flexible and accessible. It organizes data in two-dimensional tables called relations with rows and columns. Each table contains data about an entity and its attributes. Each row represents a record and each column represents an attribute or field. Each table also contains a key field to uniquely identify each record for retrieval or manipulation. Relational database tables can be combined easily to deliver data required by users, provided that any two tables share a common data element.

3. ***What are some important database design principles?***

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

4. ***What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?***

Powerful tools are available to analyze and access the information in databases. A data warehouse consolidates current and historical data from many different operational systems in a central database designed for reporting and analysis. Data warehouses support multidimensional data analysis, also known as online analytical processing (OLAP). OLAP represents relationships among data as a multidimensional structure, which can be visualized as cubes of data and cubes within cubes of data, enabling more sophisticated data analysis. Data mining analyzes large pools of data, including the contents of data warehouses, to find patterns and rules that can be used to predict future behavior and guide decision making. Text mining tools help businesses analyze large unstructured data sets consisting of text. Web mining tools focus on analysis of useful patterns and information from the World Wide Web, examining the structure of Web sites and activities of Web site users as well as the contents of Web pages. Conventional databases can be linked via middleware to the Web or a Web interface to facilitate user access to an organization's internal data.

5. ***Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?***

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

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

Key Terms

Attribute, 210
Data administration, 230
Data cleansing, 231
Data definition, 217
Data dictionary, 217
Data governance, 230
Data inconsistency, 211
Data manipulation language, 217
Data mart, 223
Data mining, 224
Data quality audit, 231
Data redundancy, 211
Data warehouse, 222
Database, 210
Database (rigorous definition), 212
Database administration, 230
Database management system (DBMS), 212
Database server, 229
Entity, 210
Entity-relationship diagram, 221
Field, 210
File, 210
Foreign key, 215
Information policy, 230
Key field, 214
Normalization, 219
Object-oriented DBMS, 215
Object-relational DBMS, 215
Online analytical processing (OLAP), 224
Predictive analytics, 226
Primary key, 210
Program-data dependence, 211
Record, 214
Referential integrity, 220
Relational DBMS, 213
Structured Query Language (SQL), 217
Text mining, 226
Tuple, 214
Web mining, 226

Review Questions

1. What are the problems of managing data resources in a traditional file environment and how are they solved by a database management system?
   - List and describe each of the components in the data hierarchy.
   - Define and explain the significance of entities, attributes, and key fields.
   - List and describe the problems of the traditional file environment.
   - Define a database and a database management system and describe how it solves the problems of a traditional file environment.
2. What are the major capabilities of DBMS and why is a relational DBMS so powerful?
   - Name and briefly describe the capabilities of a DBMS.
   - Define a relational DBMS and explain how it organizes data.
   - List and describe the three operations of a relational DBMS.
3. What are some important database design principles?
   - Define and describe normalization and referential integrity and explain how they contribute to a well-designed relational database.
   - Define and describe an entity-relationship diagram and explain its role in database design.
4. What are the principal tools and technologies for accessing information from databases to improve business performance and decision making?
   - Define a data warehouse, explaining how it works and how it benefits organizations.
   - Define business intelligence and explain how it is related to database technology.
   - Describe the capabilities of online analytical processing (OLAP).
   - Define data mining, describing how it differs from OLAP and the types of information it provides.
   - Explain how text mining and Web mining differ from conventional data mining.
   - Describe how users can access information from a company's internal databases through the Web.
5. Why are information policy, data administration, and data quality assurance essential for managing the firm's data resources?
   - Describe the roles of information policy and data administration in information management.
   - Explain why data quality audits and data cleansing are essential.

Discussion Questions

1. It has been said that you do not need database management software to create a database environment. Discuss.
2. To what extent should end users be involved in the selection of a database management system and database design?
3. What are the consequences of an organization not having an information policy?

Video Cases

Video Cases and Instructional Videos illustrating some of the concepts in this chapter are available. Contact your instructor to access these videos.

Collaboration and Teamwork: Identifying Entities and Attributes in an Online Database

With your team of three or four students, select an online database to explore, such as AOL Music, iGo.com, or the Internet Movie Database (IMDb). Explore one of these Web sites to see what information it provides. Then list the entities and attributes that the company running the Web site must keep track of in its databases. Diagram the relationship between the entities you have identified. If possible, use Google Sites to post links to Web pages, team communication announcements, and work assignments; to brainstorm; and to work collaboratively on project documents. Try to use Google Docs to develop a presentation of your findings for the class.

The Terror Watch List Database's Troubles Continue

CASE STUDY

In the aftermath of the 9-11 attacks, the FBI's Terrorist Screening Center, or TSC, was established to consolidate information about suspected terrorists from multiple government agencies into a single list to enhance inter-agency communication. A database of suspected terrorists known as the terrorist watch list was created. Multiple U.S. government agencies had been maintaining separate lists and these agencies lacked a consistent process to share relevant information.

Records in the TSC database contain sensitive but unclassified information on terrorist identities, such as name and date of birth, that can be shared with other screening agencies. Classified information about the people in the watch list is maintained in other law enforcement and intelligence agency databases. Records for the watchlist database are provided by two sources: The National Counterterrorism Center (NCTC) managed by the Office of the Director of National Intelligence provides identifying information on individuals with ties to international terrorism. The FBI provides identifying information on individuals with ties to purely domestic terrorism.

These agencies collect and maintain terrorist information and nominate individuals for inclusion in the TSC's consolidated watch list. They are required to follow strict procedures established by the head of the agency concerned and approved by the U.S. Attorney General. TSC staff must review each record submitted before it is added to the database. An individual will remain on the watch list until the respective department or agency that nominated that person to the list determines that the person should be removed from the list and deleted from the database

The TSC watch list database is updated daily with new nominations, modifications to existing records, and deletions. Since its creation, the list has ballooned to 400,000 people, recorded as 1.1 million names and aliases, and is continuing to grow at a rate of 200,000 records each year. Information on the list is distributed to a wide range of government agency systems for use in efforts to deter or detect the movements of known or suspected terrorists.

Recipient agencies include the FBI, CIA, National Security Agency (NSA), Transportation Security Administration (TSA), Department of Homeland Security, State Department, Customs and Border Protection, Secret Service, U.S. Marshals Service, and the White House. Airlines use data supplied by the TSA system in their NoFly and Selectee lists for prescreening passengers, while the U.S. Customs and Border Protection system uses the watchlist data to help screen travelers entering the United States. The State Department system screens applicants for visas to enter the United States and U.S. residents applying for passports, while state and local law enforcement agencies use the FBI system to help with arrests, detentions, and other criminal justice activities. Each of these agencies receives the subset of data in the watch list that pertains to its specific mission.

When an individual makes an airline reservation, arrives at a U.S. port of entry, applies for a U.S. visa, or is stopped by state or local police within the United States, the frontline screening agency or airline conducts a name-based search of the individual against the records from the terrorist watch list database. When the computerized name-matching system generates a "hit" (a potential name match) against a watch list record, the airline or agency will review each potential match. Matches that are clearly positive or exact matches that are inconclusive (uncertain or difficult to verify) are referred to the applicable screening agency's intelligence or operations center and to the TSC for closer examination. In turn, TSC checks its databases and other sources, including classified databases maintained by the NCTC and FBI to confirm whether the individual is a positive, negative, or inconclusive match to the watch list record. TSC creates a daily report summarizing all positive matches to the watch list and distributes them to numerous federal agencies.

The process of consolidating information from disparate agencies has been a slow and painstaking one, requiring the integration of at least 12 different databases. Two years after the process of integration took place, 10 of the 12 databases had been processed. The remaining two databases (the U.S. Immigration and Customs Enforcement's Automatic Biometric Identification System and the FBI's Integrated Automated Fingerprint Identification System) are both fingerprint databases. There is still more work to be done to optimize the list's usefulness.

Reports from both the Government Accountability Office and the Office of the Inspector General assert that the list contains inaccuracies and that government departmental policies for nomination and removal from the lists are not uniform. There has also been public outcry resulting from the size of the list and well-publicized incidents of obvious non-terrorists finding that they are included on the list.

Information about the process for inclusion on the list must necessarily be carefully protected if the list is to be effective against terrorists. The specific criteria for inclusion are not public knowledge. We do know, however, that government agencies populate their watch lists by performing wide sweeps of information gathered on travelers, using many misspellings and alternate variations of the names of suspected terrorists. This often leads to the inclusion of people who do not belong on watch lists, known as "false positives." It also results in some people being listed multiple times under different spellings of their names.

While these selection criteria may be effective for tracking as many potential terrorists as possible, they also lead to many more erroneous entries on the list than if the process required more finely tuned information to add new entries. Notable examples of 'false positives' include Michael Hicks, an 8-year-old New Jersey Cub Scout who is continually stopped at the airport for additional screening and the late senator Ted Kennedy, who had been repeatedly delayed in the past because his name resembles an alias once used by a suspected terrorist. Like Kennedy, Hicks may have been added because his name is the same or similar to a different suspected terrorist.

These incidents call attention to the quality and accuracy of the data in the TSC consolidated terrorist watch list. In June 2005, a report by the Department of Justice's Office of the Inspector General found inconsistent record counts, duplicate records, and records that lacked data fields or had unclear sources for their data. Although TSC subsequently enhanced its efforts to identify and correct incomplete or inaccurate watch list records, the Inspector General noted in September 2007 that TSC management of the watch list still showed some weaknesses.

Given the option between a list that tracks every potential terrorist at the cost of unnecessarily tracking some innocents, and a list that fails to track many terrorists in an effort to avoid tracking innocents, many would choose the list that tracked every terrorist despite the drawbacks. But to make matters worse for those already inconvenienced by wrongful inclusion on the list, there is currently no simple and quick redress process for innocents that hope to remove themselves from it.

The number of requests for removal from the watch list continues to mount, with over 24,000 requests recorded (about 2,000 each month) and only 54 percent of them resolved. The average time to process a request in 2008 was 40 days, which was not (and still is not) fast enough to keep pace with the number of requests for removal coming in. As a result, law-abiding travelers that inexplicably find themselves on the watch list are left with no easy way to remove themselves from it.

In February 2007, the Department of Homeland Security instituted its Traveler Redress Inquiry Program (TRIP) to help people that have been erroneously added to terrorist watch lists remove themselves and avoid extra screening and questioning. John Anderson's mother claimed that despite her best efforts, she was unable to remove her son from the watch lists. Senator Kennedy reportedly was only able to remove himself from the list by personally bringing up the matter to Tom Ridge, then the Director of the Department of Homeland Security.

Security officials say that mistakes such as the one that led to Anderson and Kennedy's inclusion on no-fly and consolidated watch lists occur due to the matching of imperfect data in airline reservation systems with imperfect data on the watch lists. Many airlines don't include gender, middle name, or date of birth in their reservations records, which increases the likelihood of false matches.

One way to improve screening and help reduce the number of people erroneously marked for additional investigation would be to use a more sophisticated system involving more personal data about individuals on the list. The TSA is developing just such a system, called "Secure Flight," but it has been continually delayed due to privacy concerns regarding the sensitivity and safety of the data it would collect. Other similar surveillance programs and watch lists, such as the NSA's attempts to gather information about suspected terrorists, have drawn criticism for potential privacy violations.

Additionally, the watch list has drawn criticism because of its potential to promote racial profiling and discrimination. Some allege that they were included by virtue of their race and ethnic descent, such as David Fathi, an attorney for the ACLU of Iranian descent, and Asif Iqbal, a U.S. citizen of Pakistani decent with the same name as a Guantanamo detainee. Outspoken critics of U.S. foreign policy, such as some elected officials and university professors, have also found themselves on the list.

A report released in May 2009 by Department of Justice Inspector General Glenn A. Fine found that the FBI had incorrectly kept nearly 24,000 people on its own watch list that supplies data to the terrorist watch list on the basis of outdated or irrelevant information. Examining nearly 69,000 referrals to the FBI list, the report found that 35 percent of those people remained on the list despite inadequate justification. Even more worrisome, the list did not contain the names of people who should have been listed because of their terrorist ties.

FBI officials claim that the bureau has made improvements, including better training, faster processing of referrals, and requiring field office supervisors to review watch-list nominations for accuracy and completeness. But this watch list and the others remain imperfect tools. In early 2008, it was revealed that 20 known terrorists were not correctly listed on the consolidated watch list. (Whether these individuals were able to enter the U.S. as a result is unclear.)

Umar Farouk Abdulmutallab, the Nigerian who unsuccessfully tried to detonate plastic explosives on the Northwest Airlines flight from Amsterdam to Detroit on Christmas Day 2009, had not made it onto the no-fly list. Although Abdulmutallab's father had reported concern over his son's radicalization to the U.S. State Department, the Department did not revoke Adbulmutallab's visa because his name was misspelled in the visa database, so he was allowed to enter the United States. Faisal Shahzad, the Times Square car bomber, was apprehended on May 3, 2010, only moments before his Emirates airline flight to Dubai and Pakistan was about to take off. The airline had failed to check a last-minute update to the no-fly list that had added Shahzad's name.

Sources: Scott Shane, "Lapses Allowed Suspect to Board Plane," *The New York Times*, May 4, 2010; Mike McIntire, "Ensnared by Error on Growing U.S. Watch List," *The New York Times*, April 6, 2010; Eric Lipton, Eric Schmitt, and Mark Mazzetti, "Review of Jet Bomb Plot Shows More Missed Clues," *The New York Times*, January 18, 2010; Lizette Alvarez, "Meet Mikey, 8: U.S. Has Him on Watch List," *The New York Times*, January 14, 2010; Eric Lichtblau, "Justice Dept. Finds Flaws in F.B.I. Terror List," *The New York Times*, May 7, 2009; Bob Egelko, "Watch-list Name Confusion Causes Hardship," *San Francisco Chronicle*, March 20, 2008; "Reports Cite Lack of Uniform Policy for Terrorist Watch List," *The Washington Post*, March 18, 2008; Siobhan Gorman, "NSA's Domestic Spying Grows as Agency Sweeps Up Data," *The Wall Street Journal*, March 10, 2008; Ellen Nakashima, and Scott McCartney, "When Your Name is Mud at the Airport," *The Wall Street Journal*, January 29, 2008.

CASE STUDY QUESTIONS

1. What concepts in this chapter are illustrated in this case?
2. Why was the consolidated terror watch list created? What are the benefits of the list?
3. Describe some of the weaknesses of the watch list. What management, organization, and technology factors are responsible for these weaknesses?
4. How effective is the system of watch lists described in this case study? Explain your answer.
5. If you were responsible for the management of the TSC watch list database, what steps would you take to correct some of these weaknesses?
6. Do you believe that the terror watch list represents a significant threat to individuals' privacy or Constitutional rights? Why or why not?

This page intentionally left blank