CHAPTER

6 Foundations of Business Intelligence: Databases and Information Management

LEARNING OBJECTIVES

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

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

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

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

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

6-5 How will MIS help my career?

CHAPTER CASES

Astro: Leveraging Data for Customer-Driven Service

New Cloud Database Tools Help Vodafone Fiji Make Better Decisions

Big Data Baseball

Big Data in the Asia-Pacific Takes On the COVID-19 Pandemic

VIDEO CASES

Brooks Brothers Closes In on Omnichannel Retail

Maruti Suzuki Business Intelligence and Enterprise Databases

MyLab MIS

Discussion Questions: 6-5, 6-6, 6-7; **Hands-On MIS Projects:** 6-8, 6-9, 6-10, 6-11; **eText with Conceptual Animations**

Astro: Leveraging Data for Customer-Driven Service

Astro is the leading Malaysian satellite television broadcaster. The growth and success of the company has been based on satisfying its customers' taste for Western, Indian, and Korean programming across a wide portfolio of services, including IPTV, broadband, and other streaming services.

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

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

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

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

© sauvignon/123rf

via Amazon Elasticsearch and Amazon Cognito, respectively; and data transformation and analytics through the use of AWS Glue and Amazon Athena. These services are powered by Amazon's DynamoDB document database, which provides Astro with real-time performance with minimal latency.

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

Today, Astro has an active subscriber list of nearly 5.7 million, or about 75 percent of Malaysian households. It delivers content to around 24 million people and in 2019, it launched a 4K Ultra High Definition set-top box that offers cloud recording, a broader choice of content, and broadband bundles. Astro has since introduced three streaming services, Astro GO, HBO GO, and iQIYI, attracting over 2.6 million subscribers.

Sources: P. Prem Kumar, "Malaysia's Astro Targets Netflix and Amazon with Streaming Launch," Nikkei Asia, October 8, 2020; Astro, "Transforming Our Technology Landscape to Deliver Exceptional Customer Experiences," astro.com, July 7, 2017; Digital News Asia, "Astro Accelerates Digital Business Transformation with Amazon Web Services," April 19, 2017; "Astro Inks Deal with Amazon Web Services as Part of Its Digital Strategy," Marketing-Interactive.com, April 18, 2017.

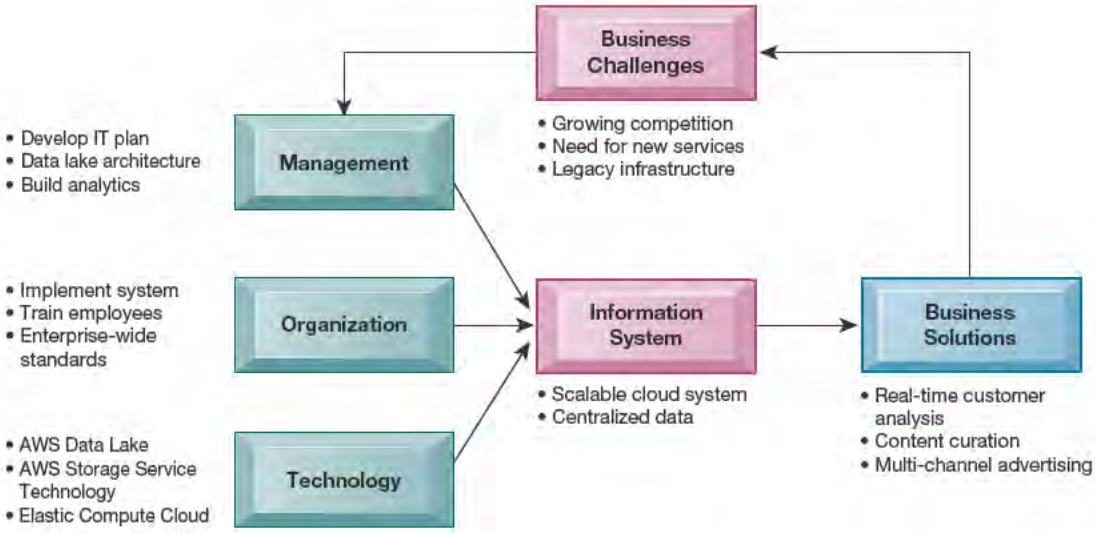
Case contributed by Imran Medi, Asia Pacific University of Technology and Innovation

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

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

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

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



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

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

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

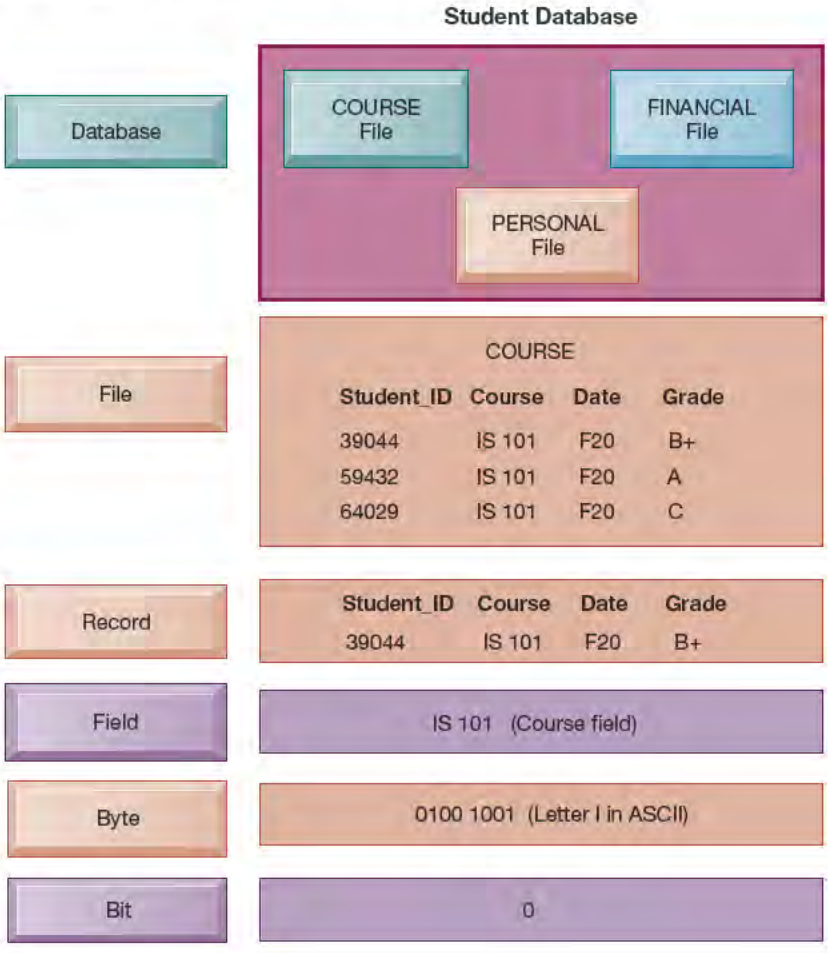
You might be surprised to learn that many businesses don't have timely, accurate, or relevant information because the data in their information systems have been poorly organized and maintained. That's why data management is so essential. To understand the problem, let's look at how information systems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

File Organization Terms and Concepts

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

FIGURE 6.1 THE DATA HIERARCHY

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



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

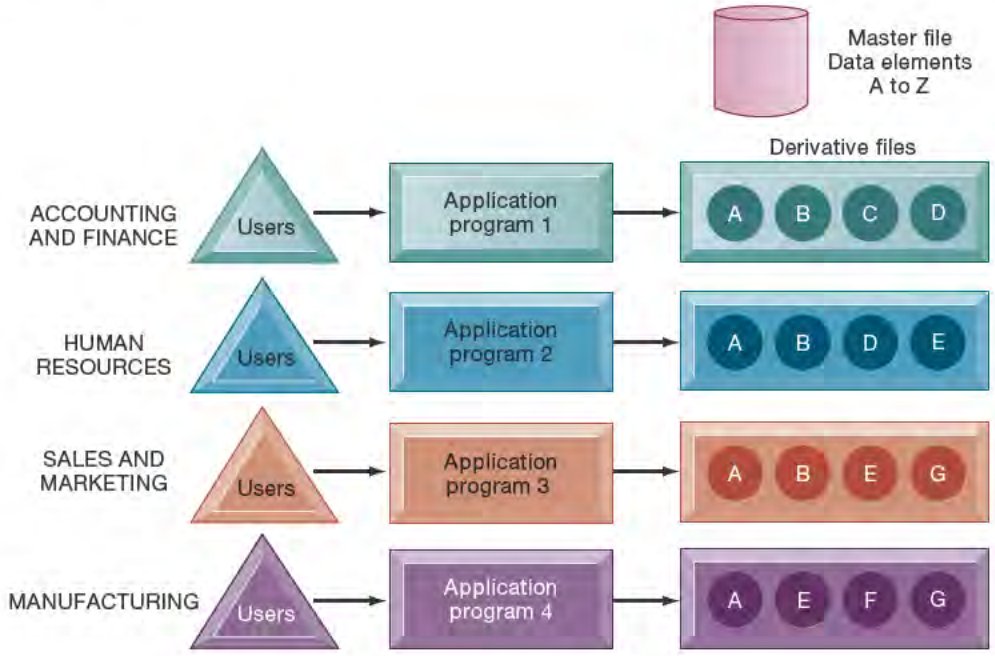
A record describes an entity. An **entity** is a person, place, thing, or event on which we store and maintain information. Each characteristic or quality describing a particular entity is called an **attribute**. For example, Student_ID, Course, Date, and Grade are attributes of the entity COURSE. The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

Problems With the Traditional File Environment

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

FIGURE 6.2 TRADITIONAL FILE PROCESSING

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



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

Data Redundancy and Inconsistency

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

Additional confusion can result from using different coding systems to represent values for an attribute. For instance, the sales, inventory, and manufacturing systems of a clothing retailer might use different codes to represent clothing size. One system might represent clothing size as "extra-large," whereas another might

use the code "XL" for the same purpose. The resulting confusion would make it difficult for companies to create customer relationship management, supply chain management, or enterprise systems that integrate data from different sources.

Program-Data Dependence

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

Lack of Flexibility

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

Poor Security

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

Lack of Data Sharing and Availability

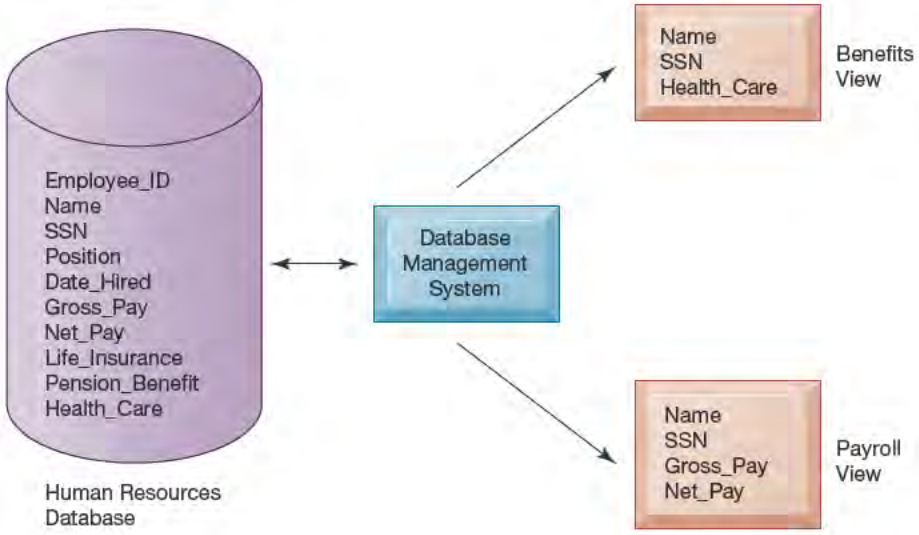
Because pieces of information in different files and different parts of the organization cannot be related to one another, it is virtually impossible for information to be shared or accessed in a timely manner. Information cannot flow freely across different functional areas or different parts of the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

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

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

FIGURE 6.3 HUMAN RESOURCES DATABASE WITH MULTIPLE VIEWS

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



Database Management Systems

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

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

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

How a DBMS Solves the Problems of the Traditional File Environment

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

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

Relational DBMS

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

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

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

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

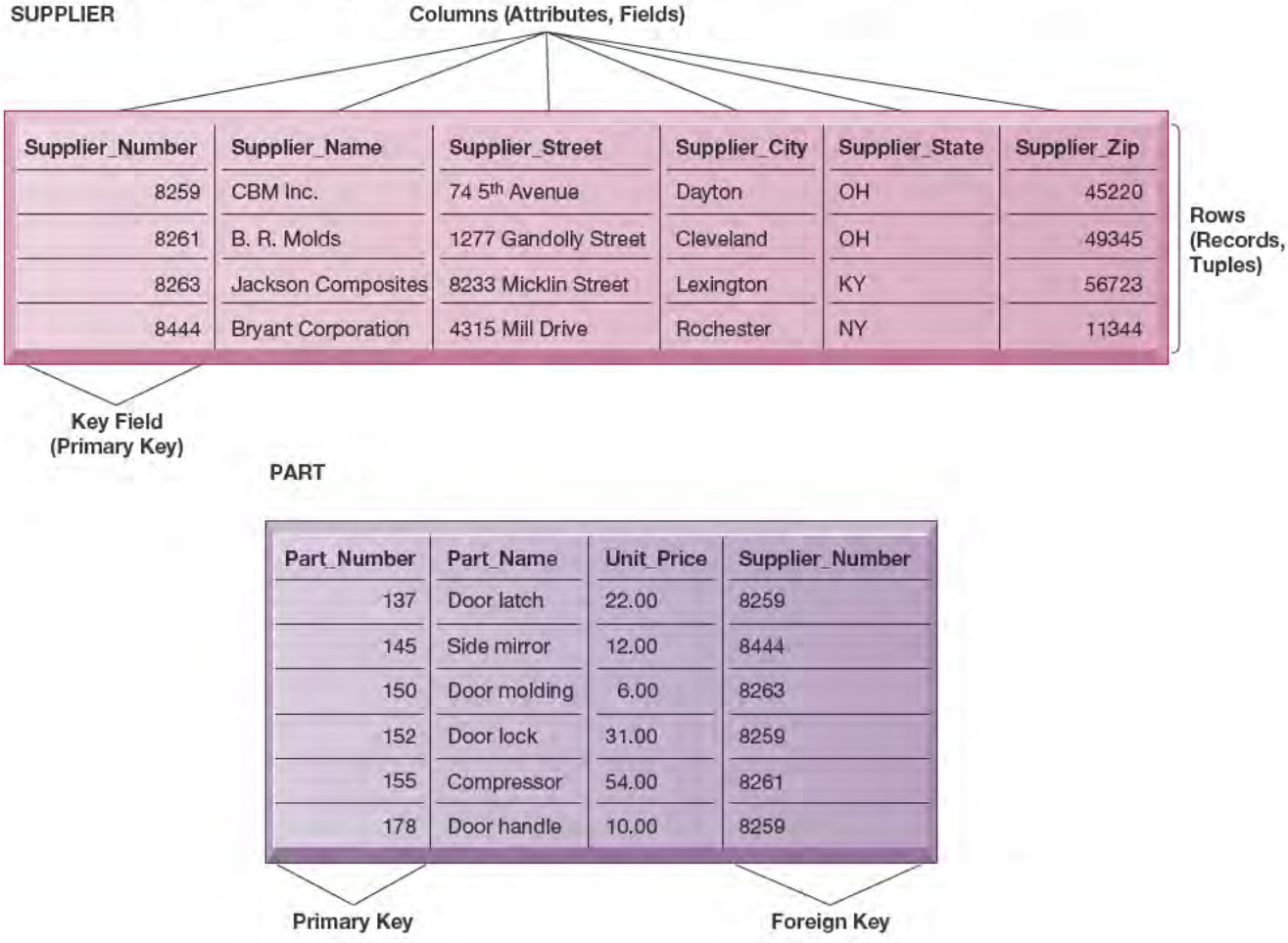
Operations of a Relational DBMS

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

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

FIGURE 6.4 RELATIONAL DATABASE TABLES

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



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

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

Capabilities of Database Management Systems

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

FIGURE 6.5 THE THREE BASIC OPERATIONS OF A RELATIONAL DBMS

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

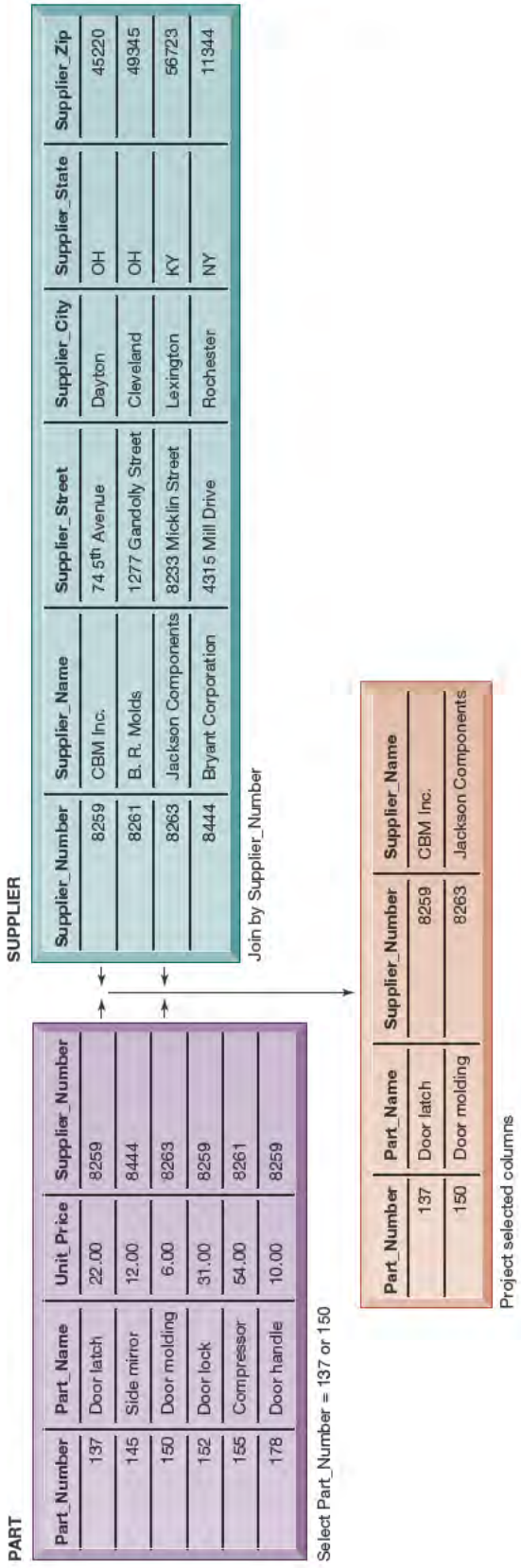
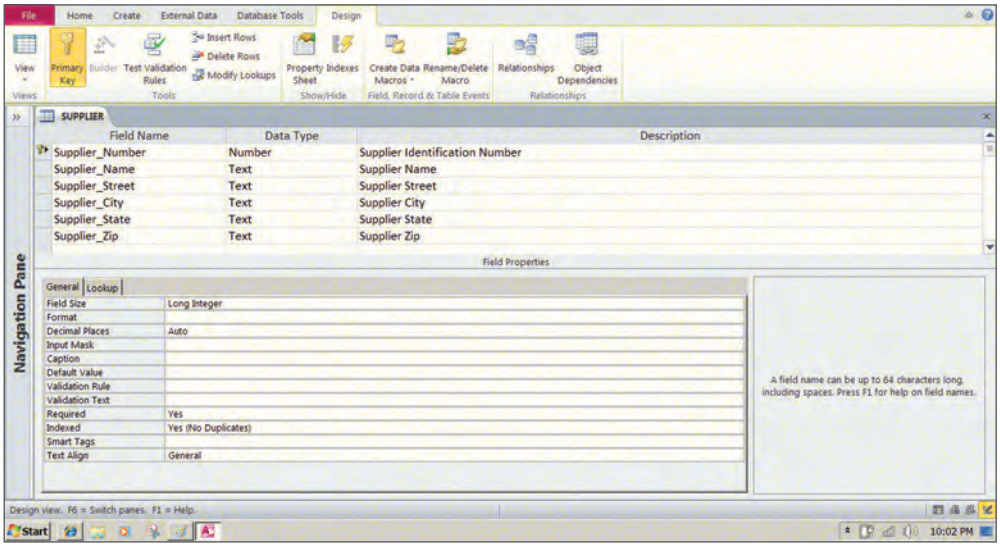


FIGURE 6.6 ACCESS DATA DICTIONARY FEATURES

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

Courtesy of Microsoft Corporation



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

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

Querying and Reporting

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

FIGURE 6.7 EXAMPLE OF AN SQL QUERY

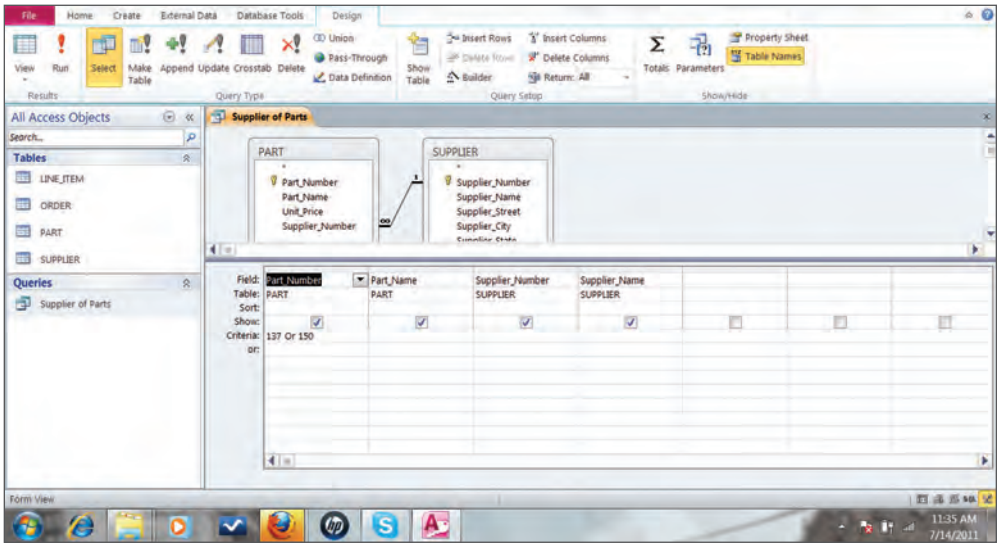
Illustrated here are the SQL statements for a query to select suppliers for parts 137 or 150. They produce a list with the same results as Figure 6.5.

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

FIGURE 6.8 AN ACCESS QUERY

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

Courtesy of Microsoft Corporation



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

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

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

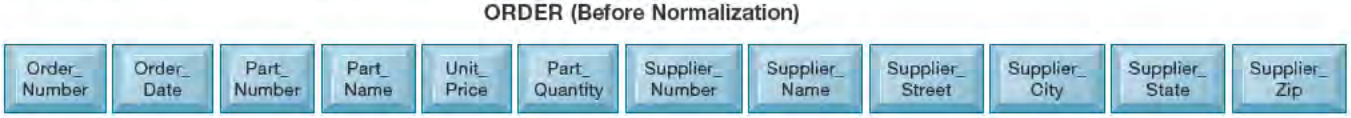
Microsoft Access and other DBMS include capabilities for report generation so that the data of interest can be displayed in a more structured and polished format than would be possible just by querying. Crystal Reports is a popular report generator for large corporate DBMS, although it can also be used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

Designing Databases

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

FIGURE 6.9 AN UNNORMALIZED RELATION FOR ORDER

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



Normalization and Entity-Relationship Diagrams

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

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

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

FIGURE 6.10 NORMALIZED TABLES CREATED FROM ORDER

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

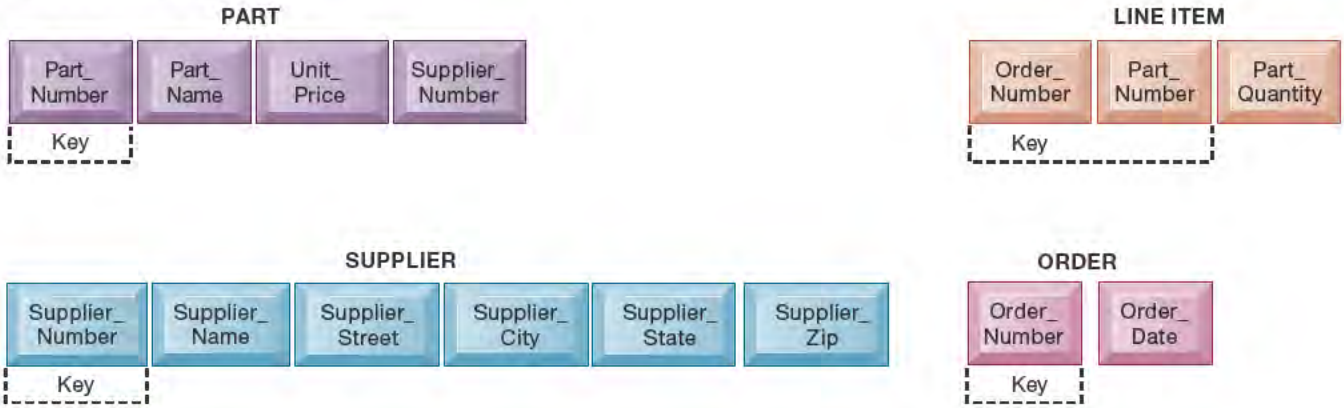
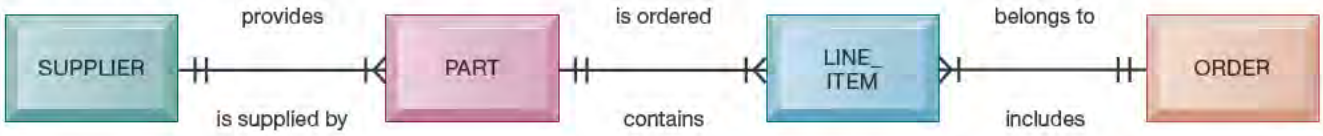


FIGURE 6.11 AN ENTITY-RELATIONSHIP DIAGRAM

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



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

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

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

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

Nonrelational Databases, Cloud Databases, and Blockchain

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

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

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

Phamaceutical giant AstraZeneca employed MongoDB NoSQL database in next-generation genome sequencing to develop drugs to fight disease. The technology creates a synthetic version of messenger RNA, which helps create proteins that can fight cancer and other diseases. Analyzing a single human genome can take four days. MongoDB's cross-platform document-oriented database can work with unstructured data pulled from disparate database systems to crunch numbers on a grand sale. With a greater command of the data, AstraZeneca can pursue links and patterns that it never noticed before.

Cloud Databases and Distributed Databases

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

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

Blockchain

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

INTERACTIVE SESSION TECHNOLOGY

New Cloud Database Tools Help Vodafone Fiji Make Better Decisions

Vodafone Fiji Limited is a 100 percent locally owned wireless telecommunications services provider, with more than 760,000 subscribers on its network, 260 employees, and revenue of U.S. $185 million. Vodafone Fiji works closely with the Vodafone Group, a British multinational telecommunications conglomerate, which predominantly services Asia, Africa, Europe, and Oceania, and other operating companies to deliver cutting-edge technology to the people of Fiji. It has 85 percent of that country's market share in telecommunications services.

Prepaid customers account for 96 percent of Fiji's mobile communications market. (Less than 25 percent of U.S. mobile customers use prepaid services.) Prepaid customers are not bound by mobile service contracts, and thus are always looking for better ways to save. Consequently, Vodafone Fiji has to come up with better deals for its customers to persuade them to continue with its services. Unfortunately Vodafone Fiji's systems were not able to deliver the information for the task.

The company lacked the computing power, storage, and data management tools to analyze data quickly to make informed decisions about which deals to offer to which customers and the right times to do so. The data were stored on premises in multiple databases. Making matters worse, Amalgamated Telecom Holdings (Vodafone Fiji's major shareholder) had recently acquired several other telecommunications companies serving nearby markets in Samoa, American Samoa, Vanuatu, Cook Islands, and Kiribati. The data to be managed and mined for insights increased threefold. All those businesses had different types of data in different systems and different formats.

Vodafone Fiji's chief commercial officer Ronald Prasad and his team calculated that it would cost the company about U.S. $2.5 million to upgrade the company's on-premises systems. They chose Oracle Autonomous Data Warehouse and Oracle Analytics Cloud services as a much more suitable and cost-effective solution. Oracle Autonomous Data Warehouse Cloud (ADWC) is a fully managed, high-performance elastic cloud service providing analytical capability over data stored in an Oracle database. The environment is optimized for data warehouse workloads and supports all standard SQL and business intelligence (BI) tools. The Oracle Autonomous Data Warehouse provides an easy-to-use, fully autonomous database that scales elastically as workloads increase, delivers fast query performance, and requires no database administration. Vodafone Fiji would be able to extract, move, and transform data from disparate sources in the cloud, where it could be analyzed much more rapidly.

By automating many of the routine tasks required to manage Oracle databases, Oracle Autonomous Database can free up database administrators (DBAs) to do higher-level and more strategic work. The warehouse system handles a great deal of tedious technical work on its own. Vodafone Fiji would not need to hire people with specialized database management skills, which would be challenging in a small job market such as Fiji's. An on-premises data warehouse would have taken the company two months to implement, whereas it set up Oracle Autonomous Data Warehouse Cloud within 30 minutes and reporting functionality within one week. Vodafone Fiji pays only for computing resources consumed.

Oracle Analytics Cloud is a cloud-based platform that can take data in almost any form or size from almost any source (desktop, enterprise, data center), with capabilities for collecting, consolidating, and transforming data, and creating transactional and analytical reports and dashboards. Oracle Analytics Cloud provides self-service capabilities for users to perform what-if modeling and analysis. Users are empowered to visualize and discover data, including working with Big Data.

Vodafone Fiji can now easily obtain insights from the data collected and adjust its promotions to changing market dynamics. It is able to target customers who are at risk of cancelling Vodafone services as well as those most likely to respond to a special offer. The warehouse system can produce reports on call patterns to make sure customers are signed up for the optimal set of services, or target customers with 4G plans with free trials to entice them into subscribing to a data plan. Promotional campaigns customized to specific cell sites can encourage customers to get online during low utilization periods.

More than 50 rich reports are available to decision makers within minutes. Queries can be completed 5 times faster than before. A data mining procedure that used to take 125 minutes takes only 25 minutes

with Oracle Autonomous Data Warehouse; an aggregation query that used to take 294 seconds with the old on-premises system takes only 5 seconds with Oracle Autonomous Data Warehouse. Examining customer service call data, warehouse analytic tools have been able to identify trivial, repetitive inquiries where automated responses are possible, reducing inbound calls by 8 to 10 percent and freeing up customer service agents to focus on up-selling and cross-selling based on specific customer profiles. Prasad's same three-member team has been able to handle quadruple the preacquisition workload very effectively.

Sources: www.vodafone.com.fj, accessed January 12, 2020; Tara Swords, "Call to Action," *Profit Magazine*, Fall 2019; and www.oracle.com, accessed January 12, 2020.

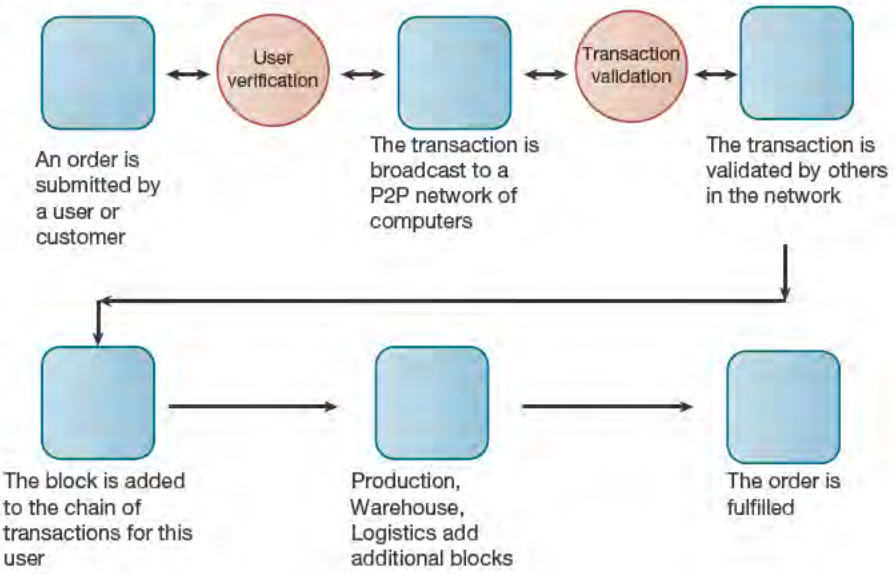
CASE STUDY QUESTIONS

1. Define the problem faced by Vodafone Fiji. What management, organization, and technology factors contributed to the problem?
2. Evaluate Oracle Autonomous Data Warehouse and Oracle Analytics Cloud as a solution for Vodafone Fiji.
3. How did the new Oracle tools change decision making at Vodafone Fiji?
4. Was using cloud services advantageous for Vodafone Fiji? Explain your answer.

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

FIGURE 6.12 HOW BLOCKCHAIN WORKS

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



information, and only validated transactions are accepted. Once recorded, the transaction cannot be changed. Figure 6.12 illustrates how blockchain works for fulfilling an order.

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

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

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

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

The Challenge of Big Data

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

Big data is often characterized by the "3Vs": the extreme *volume* of data, the wide *variety* of data types and sources, and the *velocity* at which data must be processed. Big data doesn't designate any specific quantity but usually refers to data in the petabyte and exabyte range—in other words, billions to trillions of records, many from different sources. Big data are produced in much larger quantities and much more rapidly than traditional data. For example, a single jet engine is capable of generating 10 terabytes of data in just 30 minutes, and

INTERACTIVE SESSION MANAGEMENT

Big Data Baseball

Big data and analytics are sweeping the business world, and the professional sports industry is no exception. Baseball, football, soccer, hockey, tennis, and even sailboat racing are finding ways to analyze data about players and competing teams in order to improve performance. The use of analytics and big data has revolutionized the game of baseball as we know it, including defensive shifts, swing path changes, and how teams acquire and develop players.

Given the huge disparities in Major League Baseball (MLB) team budgets, wealthier teams typically have the advantage in recruiting the best players. Michael Lewis's book *Moneyball*, published in 2003, describes how Oakland Athletics manager Billy Beane was able to turn the underdog A's into a winning team by using big data analytics to guide decisions about which players to recruit and cultivate. Rigorous statistical analysis had demonstrated that on-base percentage and slugging percentage were better indicators of offensive success (and cheaper to obtain on the open market) than more historically valued qualities such as speed and contact. These observations flew in the face of conventional baseball wisdom and the beliefs of many baseball talent scouts and coaches. Beane rebuilt the A's based on these findings, producing a consistently winning team for a number of years by using advanced analytics to gain insights into each player's value and contribution to team success that wealthier teams had overlooked.

Big data is credited with helping the Boston Red Sox win the World Series in 2004 and the St. Louis Cardinals win in 2006 and 2011. To varying degrees, every Major League Baseball team today uses big data and deep analytics to support decisions about many aspects of the game. However, some teams, such as the Pittsburgh Pirates, Chicago Cubs, and Houston Astros, were slower to do so than others, and suffered lackluster performance until they embraced big data more fully.

Findings from big data analytics have changed the importance baseball teams attach to specific skills of players. Skills that previously could not be quantified are now receiving more attention, including fielding, base running, and stealing. Skill in fielding is especially valued today. For example, Mike Trout, center fielder for the Los Angeles Angels, is highly regarded by team owners because he's an exceptional fielder and base runner and an exceptionally intelligent base ball player, even though he lacked stellar statistics in home runs. Today the biggest challenge is not whether to use big data in baseball but how to use it effectively. It is not always possible to interpret the data and separate out what is "noise" and what is actually actionable information. The amount of data players and pitchers must deal with can be overwhelming—pitch usage, swing planes, spin rates, etc. When a player steps into the batter's box, every hitter is different in terms of how much information that person can absorb before getting bogged down in it. Some want to know what a pitcher will do in certain situations—what pitches the pitcher will use and how often that person uses them—while some want to just step in with a clear head and look for the ball. There's only so much data a person can use without dissecting too much and getting too distracted from the task at hand.

Many baseball experts still believe that traditional methods of player evaluation, along with gut instinct, money, and luck, are still key ingredients for winning teams. For example, the San Francisco Giants use big data and statistics, but also base their player recruitment decisions on the opinions of scouts and coaches. According to Giants bench coach Ron Wotus, numbers really can't tell the whole story about the quality of the player; so the Giants integrate statistical data with scouting, coaching, and player experience, especially when dealing with opponents outside the National League that the Giants do not see regularly. Being able to exploit an individual player's strengths comes more from knowing the player and his ability as opposed to the statistics, Wotus believes. Shortstops with good arms can play farther from home plate than normal at times, while fast runners can play closer to home plate than usual. There are nuances to defending the opposition that are not statistically related, but statistics help when you don't know players well enough to know what to expect from them.

Sources: www.statsperform.com, accessed January 25, 2020; www.mlb.com, accessed January 25, 2020; "Sports Analytics Market Size, Shares 2020 By Top Key Players: IBM, SAP, SAS, Tableau Software, Oracle, STATS, Prozone," *MarketWatch*, January 30, 2020; "Nick San Miguel, 'San Francisco Giants: Analytics Are Not the Answer'," www.aroundthefoghorn.com, accessed February 4, 2020; Richard Justice, "MLB Clubs Stay Focused on Future of Analytics," www.mlb.com, accessed March 14, 2019; "Changing the Game: How Data Analytics Is Upending Baseball," *Knowledge @ Wharton*, February 21, 2019; and "A View from the Front Lines of Baseball's Data-Analytics Revolution," *McKinsey Quarterly*, July 2018.

CASE STUDY QUESTIONS

1. How did information technology change the game of baseball? Explain.
2. How did information technology affect decision making at MLB teams? What kinds of decisions changed as the result of using big data?
3. How much should baseball rely on big data and analytics? Explain your answer.

there are more than 100,000 airline flights each day. Twitter generates more than 12 terabytes of data daily. According to the International Data Center (IDC) technology research firm, data are more than doubling every two years, so the amount of data available to organizations is skyrocketing.

Businesses are interested in big data because they can reveal more patterns and interesting relationships than smaller data sets, with the potential to provide new insights into customer behavior, weather patterns, financial market activity, or other phenomena. For example, Shutterstock, the global online image marketplace, stores 24 million images, adding 10,000 more each day. To find ways to optimize the buying experience, Shutterstock analyzes its big data to find out where its website visitors place their cursors and how long they hover over an image before making a purchase. Big data is also finding many uses in the public sector. For example, the chapter-ending case explores how big data is being used to combat the COVID-19 pandemic. The Interactive Session on Management illustrates how Major League Baseball is using big data to improve player and team performance.

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

Business Intelligence Infrastructure

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

Data Warehouses and Data Marts

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

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

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

Hadoop

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

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

Hadoop can process large quantities of any kind of data, including structured transactional data, loosely structured data such as Facebook and Twitter feeds, complex data such as web server log files, and unstructured audio and video data. Hadoop runs on a cluster of inexpensive servers, and processors can be added or removed as needed. Companies use Hadoop for analyzing very large volumes of data as well as for a staging area for unstructured and

semistructured data before they are loaded into a data warehouse. Yahoo uses Hadoop to track users' behavior so it can modify its home page to fit their interests. Life sciences research firm NextBio uses Hadoop and HBase to process data for pharmaceutical companies conducting genomic research. Top database vendors such as IBM, Hewlett-Packard, Oracle, and Microsoft have their own Hadoop software distributions. Other vendors offer tools for moving data into and out of Hadoop or for analyzing data within Hadoop.

In-Memory Computing

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

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

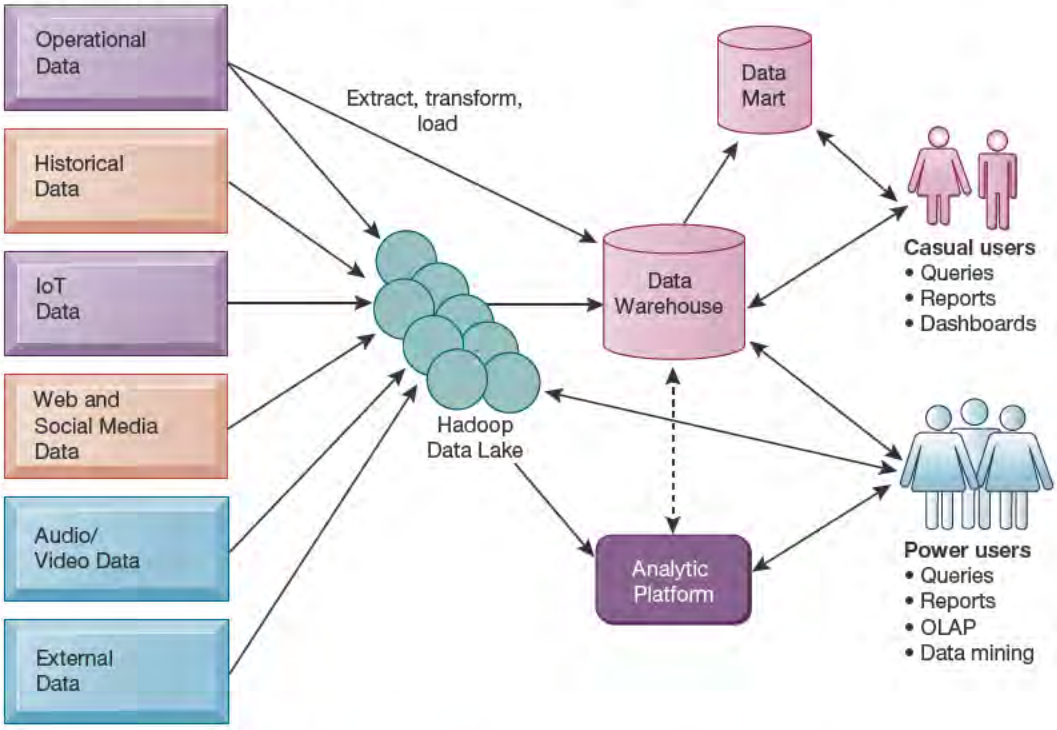
Analytic Platforms

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

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

FIGURE 6.13 CONTEMPORARY BUSINESS INTELLIGENCE INFRASTRUCTURE

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



Analytical Tools: Relationships, Patterns, Trends

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

Online Analytical Processing (OLAP)

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

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

FIGURE 6.14 MULTIDIMENSIONAL DATA MODEL

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

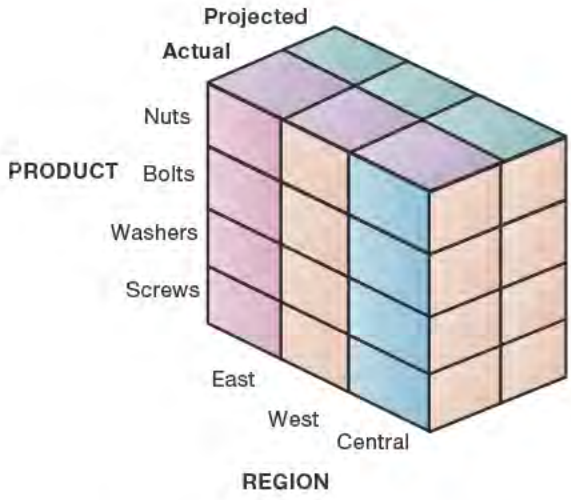


Figure 6.14 shows a multidimensional model that could be created to represent products, regions, actual sales, and projected sales. A matrix of actual sales can be stacked on top of a matrix of projected sales to form a cube with six faces. If you rotate the cube 90 degrees one way, the face showing will be product versus actual and projected sales. If you rotate the cube 90 degrees again, you will see region versus actual and projected sales. If you rotate 180 degrees from the original view, you will see projected sales and product versus region. 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 2020?" OLAP, or multidimensional analysis, supports much more complex requests for information, such as "Compare sales of product 403 relative to plan by quarter and sales region for the past two years." With OLAP and query-oriented data analysis, users need to have a good idea about the information for which they are looking.

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

- *Associations* are occurrences linked to a single event. For instance, a study of supermarket purchasing patterns might reveal that, when corn chips are purchased, a cola drink is purchased 65 percent of the time, but when there is a promotion, cola is purchased 85 percent of the time. This information helps managers make better decisions because they have learned the profitability of a promotion.
- In *sequences*, events are linked over time. We might find, for example, that if a house is purchased, a new refrigerator will be purchased within two weeks

65 percent of the time, and an oven will be bought within one month of the home purchase 45 percent of the time.

- *Classification* recognizes patterns that describe the group to which an item belongs by examining existing items that have been classified and by inferring a set of rules. For example, businesses such as credit card or telephone companies worry about the loss of steady customers. Classification helps discover the characteristics of customers who are likely to leave and can provide a model to help managers predict who those customers are so that the managers can devise special campaigns to retain such customers.
- *Clustering* works in a manner similar to classification when no groups have yet been defined. A data mining tool can discover different groupings within data, such as finding affinity groups for bank cards or partitioning a database into groups of customers based on demographics and types of personal investments.
- Although these applications involve predictions, *forecasting* uses predictions in a different way. It uses a series of existing values to forecast what other values will be. For example, forecasting might find patterns in data to help managers estimate the future value of continuous variables, such as sales figures.

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

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

Text Mining and Web Mining

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

Businesses might turn to text mining to analyze transcripts of calls to customer service centers to identify major service and repair issues or to measure customer sentiment about their company. **Sentiment analysis** software is able to mine text comments in an email message, blog, social media conversation, or survey forms to detect favorable and unfavorable opinions about specific

subjects. For example, Kraft Foods uses a Community Intelligence Portal and sentiment analysis to tune into consumer conversations about its products across numerous social networks, blogs, and other websites. Kraft tries to make sense of relevant comments rather than just track brand mentions and can identify customers' emotions and feelings when they talk about how they barbecue and what sauces and spices they use.

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

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

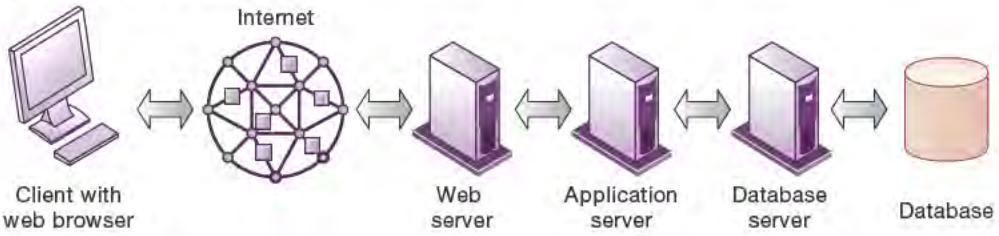
Databases and the Web

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

Suppose, for example, a customer with a web browser wants to search an online retailer's database for pricing information. Figure 6.15 illustrates how that customer might access the retailer's internal database over the web. The user accesses the retailer's website over the Internet using a web browser on

FIGURE 6.15 LINKING INTERNAL DATABASES TO THE WEB

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



his or her client PC or mobile device. The user's web browser software requests data from the organization's database, using HTML commands to communicate with the web server. Apps provide even faster access to corporate databases.

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

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

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

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

Other companies have created entirely new businesses based on access to large databases through the web. One is the social networking service Facebook, which helps users stay connected with each other and meet new people. Facebook features "profiles" with information on over 2.6 billion active users with information about themselves, including interests, friends, photos, and groups with which they are affiliated. Facebook maintains a very large database to house and manage all of this content. There are also many web-enabled databases in the public sector to help consumers and citizens access helpful information.

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

Setting up a database is only a start. To make sure that the data for your business remain accurate, reliable, and readily available to those who need it, your business will need special policies and procedures for data governance. **Data governance** encompasses policies and procedures through which data can be

managed as an organizational resource. It establishes the organization's rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying information. These include identifying which users and organizational units can share information, where information can be distributed, who is responsible for updating and maintaining the information, and how data resources should be secured (see Chapter 8). A firm's information policy might specify, for example, that only selected members of the payroll and human resources department would have the right to change or view sensitive employee data, such as an employee's salary or social security number, and that these departments are responsible for making sure that such employee data are accurate.

Assuring Data Quality

With today's organizations relying so heavily on data to drive operations and decision making, data quality assurance is especially important. What would happen if a customer's telephone number or account balance were incorrect? What would be the impact if the database had the wrong price for the product you sold? Data that are inaccurate, untimely, or inconsistent with other sources of information create serious operational and financial problems for businesses, even with a well-designed database and information policy. When faulty data go unnoticed, they often lead to incorrect decisions, product recalls, and even financial losses.

Gartner Inc. reported that more than 25 percent of the critical data in large *Fortune* 1000 companies' databases is inaccurate or incomplete, including bad product codes and product descriptions, faulty inventory descriptions, erroneous financial data, incorrect supplier information, and incorrect employee data. Some of these data quality problems are caused by redundant and inconsistent data produced by multiple systems. For example, the sales ordering system and the inventory management system might both maintain data on the organization's products. However, the sales ordering system might use the term *Item Number*, and the inventory system might call the same attribute *Product Number*. The sales, inventory, or manufacturing systems of a clothing retailer might use different codes to represent values for an attribute. One system might represent clothing size as extra large, whereas the other system might use the code XL for the same purpose. During the design process for a database, data describing entities, such as a customer, product, or order, should be named and defined consistently for all business areas using the database.

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

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

Before a new database is in place, organizations need to identify and correct their faulty data and establish better routines for editing data once their

database is in operation. Analysis of data quality often begins with a **data quality audit**, which is a structured survey of the accuracy and level of completeness of the data in an information system. Data quality audits can be performed by surveying entire data files, surveying samples from data files, or surveying end users for their perceptions of data quality.

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

Data quality problems are not just business problems. They also pose serious problems for individuals, affecting their financial condition and even their jobs. For example, inaccurate or outdated data about consumers' credit histories maintained by credit bureaus can prevent creditworthy individuals from obtaining loans or lower their chances of finding or keeping a job.

6-5 How will MIS help my career?

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

The Company

South Africa Power, a large diversified energy company headquartered in Johannesburg, has an open position for an entry-level data analyst. The company is involved in the distribution, transmission, and generation of electricity as well as energy management and other energy-related services for 5 million customers.

Position Description

Job responsibilities include:

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

Job Requirements

- BA/BS degree in business, finance, accounting, economics, engineering, or related discipline
- One to two years of professional work experience desirable

- Knowledge of Microsoft Office tools (Excel, PowerPoint, Access, and Word)
- Strong analytical capabilities, including attention to detail, problem solving, and decision making
- Strong oral and written communication and teamwork skills
- Familiarity with transmission substation equipment desirable

Interview Questions

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

Author Tips

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

REVIEW **SUMMARY**

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

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

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

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

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

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

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

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

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

Developing a database environment requires policies and procedures for managing organizational data as well as a good data model and database technology. Data governance encompasses organizational policies and procedures for the maintenance, distribution, and use of information in the organization. Data that are inaccurate, incomplete, or inconsistent create serious operational and financial problems for businesses because they may create inaccuracies in product pricing, customer accounts, and inventory data and lead to inaccurate decisions about the actions that should be taken by the firm. Firms must take special steps to make sure they have a high level of data quality. These include using enterprise-wide data standards, databases designed to minimize inconsistent and redundant data, data quality audits, and data cleansing software.

Key Terms

Analytic platform, 264
Attribute, 246
Big data, 260
Bit, 245
Blockchain, 257
Byte, 245
Data cleansing, 271
Data definition, 253
Data dictionary, 253
Data governance, 269
Data inconsistency, 247
Data lake, 264
Data manipulation language, 253
Data mart, 263
Data mining, 266
Data quality audit, 271
Data redundancy, 247
Data warehouse, 263
Database, 248
Database management system (DBMS), 249
Database server, 269
Distributed database, 257
Entity, 246
Entity-relationship diagram, 256
Field, 245
File, 245
Foreign key, 250
Hadoop, 263
In-memory computing, 264
Key field, 250
Nonrelational database management systems, 257
Normalization, 255
Online analytical processing (OLAP), 265
Primary key, 250
Program-data dependence, 248
Query, 253
Record, 245
Referential integrity, 256
Relational DBMS, 250
Sentiment analysis, 267
Structured Query Language (SQL), 253
Text mining, 267
Tuple, 250
Web mining, 268

MyLab MIS

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

Review Questions

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

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

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

- Define a database and a database management system.
- Name and briefly describe the capabilities of a DBMS.
- Define a relational DBMS and explain how it organizes data.

- List and describe the three operations of a relational DBMS.
- Explain why nonrelational databases are useful.
- Define and describe normalization and referential integrity and explain how they contribute to a well-designed relational database.
- Define and describe an entity-relationship diagram and explain its role in database design.

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

- Explain why businesses may be more interested in big data than in smaller data sets.
- Describe how a business might use a data mart.
- Describe the capabilities of Hadoop Distributed File System for data storage.
- Explain how in-memory computing can be used to facilitate data analysis.
- Explain what is meant by clustering in the context of data mining.
- Describe sentiment analysis and suggest why a business might use it.

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

- Define data governance and explain how it helps organizations manage their data.
- List and describe the most important data quality problems.
- List and describe the most important tools and techniques for assuring data quality.

Discussion Questions

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

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

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

Hands-On MIS Projects

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

Management Decision Problems

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

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

- What business problems are created by not having these data in a single standard format?
- How easy would it be to create a database with a single standard format that could store the data from both systems? Identify the problems that would have to be addressed.

- Should the problems be solved by database specialists or general business managers? Explain.
- Who should have the authority to finalize a single companywide format for this information in the data warehouse?

Achieving Operational Excellence: Building a Relational Database for Inventory Management

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

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

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

Improving Decision Making: Searching Online Databases for Overseas Business Resources

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

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

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

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

Collaboration and Teamwork Project

Identifying Entities and Attributes in an Online Database

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

Big Data in the Asia-Pacific Takes On the COVID-19 Pandemic

CASE STUDY

In December 2019, an epidemic of a novel coronavirus broke out in Wuhan, China, and quickly spread across the country to the rest of the world. As of October 28, 2020, there were more than 44 million confirmed cases of COVID-19 in 215 countries, with approximately 1.1 million deaths reported worldwide. Teams around the world raced to develop a vaccine. For instance, Pfizer, a global pharmaceutical company, partnered with BioNTech, a German biotechnology company, and at the conclusion of its Phase-3 trial in December 2020, it became the first to be authorized for emergency use in both the United Kingdom and the United States. The National Research Centre for Epidemiology and Microbiology in Russia has also developed a vaccine called Sputnik V, which it approved for release in August 2020.

During the months before an efficacious vaccine was developed and made available to the global community, various community containment measures and lockdowns were extensively implemented across countries to flatten the pandemic curve by slowing down the spread of the virus. Among the Asian countries, China went into a strict lockdown for one month but managed to return to some degree of normalcy. Similarly, Taiwan and South Korea managed to flatten the epidemic curve swiftly after monitoring the devastating outbreak in Wuhan. Their proactive responses to contain the outbreak included effective and vital interventions to promote social distancing, prompt isolation strategies, detection of active cases, active involvement of various policymakers to fund the cost associated with the pandemic, and the extensive use of big data for contact tracing to track individuals who may have come into contact with those who were tested positive.

Big data gives healthcare providers a competitive edge with the benefits of the "3Vs": volume, variety, and velocity. Some of the unique considerations with respect to big data in healthcare include timeliness in the synchronization of health information updates, ownership of health data that legally belongs to individuals with medical needs, data privacy and security, heterogeneity of data (structured, semistructured, and unstructured), and incompleteness of the data due to the continuous stream of it being fed into the systems. Nevertheless, harnessing such data provides new opportunities and life-saving outcomes. During the pandemic, big data was used to track, control, analyze, and eventually flatten the curve. Valuable information gleaned from big data helped in expediting the development of vaccines by providing more accurate and dynamic insights. This information, in conjunction with the Protein Data Bank—the sole global repository of three-dimensional structured data on large biomolecules such as proteins and nucleic acids—paved the way for medical researchers to understand the nature of the coronavirus and develop vaccination for it.

Big data also provided information to detect suspected COVID-19 infections, enabling an innovative epistemological approach to controlling the outbreak. The high volume of data helped in studying the pattern and rate of spread of the infection, and government bodies were able to optimize resource allocation and execute timely action.

In combating the COVID-19 crisis, Taiwan set a good example through its application of big data, artificial intelligence (AI), and frequent digital communication. The use of these technologies helped slow the spread of the COVID-19 virus through systematic tracking of the proliferation of the virus while providing citizens with real-time data on ways to minimize the risk of contracting it. AI was used extensively by the Taiwanese government to create real-time digital updates to caution its people about where infections had been detected.

The government conducted detailed mapping of the spread and systemic contact tracing across the country. The mapping was initiated during the outbreak's infancy and was able to curb the transmission of the virus early on. In addition, a team of programmers developed map-based code using Google Maps to detect and track the travel history of an individual over a specific period, and analytics highlighted whether the individual had come in contact with "red areas," or hotspots of coronavirus infection.

As most Taiwanese have signed up for the National Health Insurance (NHI), their health data is instantly available to the government. The NHI cards have many embedded security features to

prevent counterfeiting, and the data is protected through strict authorization features. A mutual recognition system allows medical personnel to retrieve health information of the NHI cardholders from the database using their own health professional cards. During the pandemic, Taiwan integrated its NHI database with its comprehensive immigration database and collection of travel histories. Foreigners were asked to scan a QR code and then fill out an online health declaration form, after which mobile technologies generated datasets for them. Through this integration, Taiwan generated high-volume datasets for analysis, which allowed researchers to create a management system that provided greater access to health and travel data for tracking purposes. This voluminous data was made accessible to healthcare providers to track 14-day travel histories and early symptoms of COVID-19 infection. This integration and analysis of big data allowed officials to distinguish between high-risk and low-risk individuals and decide between quarantining and issuing medical authorization.

Taiwan's success in flattening the curve may be attributed to its past experiences in combating the SARS epidemic in 2003. During this period, infrastructures such as active surveillance and screening systems using infrared thermal imaging scanning were implemented in most airports. These now proved beneficial in detecting COVID-19 symptoms and tracking the infected. However, Taiwan clearly owes much of its success to the emerging technologies that deploy big data analytics.

Big data is in the process of transforming Malaysia's digital ecosystem, and the pandemic has played a significant role in accelerating this digitization process. The government was already committed to adopting measures using big data analytics, which had gained wide popularity in the digitally networked nation. Data from various divisions of the healthcare sector were fed into big data analytics tools and enabled communal screenings, systemic contact tracing, and close monitoring of the impact of the virus. Among the measures implemented by the Malaysian government were the implementation of the Movement Control Order (MCO), the Enhanced Movement Control Order (EMCO), the Restricted Movement Control Order (RMCO), and the Controlled Movement Control Order (CMCO) based on the number of infected cases and the risk category obtained from the Ministry of Health database. This proved effective in controlling the spread of the virus, particularly in Selangor state, which had recorded the highest number of COVID-19 cases in Malaysia.

As a way forward, the government also introduced the MySejahtera application to assist in monitoring the outbreak. This mobile application captured the users' personal information, such as their name, identity card number, contact number, email address, residential address, and age, and fed it into the Ministry of Health's database for epidemiological analysis. Identity card numbers were verified with the National Registration Department to weed out bad data. Data from MySejahtera was extracted for contact tracing when a positive case was reported. There was a strict data process flow; the data was accessible to just seven people across the Crisis Preparedness and Response Centre (CPRC), the disease control center of MOH, and the Ministry of Health and National Cyber Security Agency (NACSA), which is responsible for data security and governance.

MySejahtera was equipped with a customized dashboard to track locations using a check-in feature that enabled visitor management for private and government premises. Once a COVID-19 case was confirmed, the CPRC acted immediately to retrieve data—using check-in timestamps at particular venues—on close contacts and other individuals who might have been in proximity to the infected person. Tracking of close contacts was prioritized for venues that carried a high probability of contracting COVID-19, such as enclosed and crowded venues like shopping malls and restaurants. Low-risk transmission venues were more likely to be omitted from the analysis.

Big data has thus proven to be a key player in flattening the COVID-19 curve in many of the Asia-Pacific countries. It has reduced the number of cases and deaths through early detection, provided pandemic prediction with high accuracy, assisted in effective disease management, and even improved patient experience. The use of analytics tools was particularly robust in Malaysia, where the Sungai Buloh Hospital's Infectious Disease Department team was awarded the Global Health Award in recognition of its determined fight against the COVID-19 virus. In addition, the World Brands Foundation has acknowledged the health director-general's leadership and his team's hard work in their efficient management of the coronavirus outbreak in the country.

The COVID-19 pandemic has left an indelible impact on human life and culture, but these victories in the Asia-Pacific countries have shown that success depends on governments' and organizations' willingness to act quickly and their readiness to harness the potential benefits of current and new technology.

Sources: "Pfizer and BioNTech Celebrate Historic First Authorization in the U.S. of Vaccine to Prevent COVID-19," Press Release, www.pfizer.com, December 11, 2020; "About Vaccine," www.sputnikvaccine.com, accessed December 19, 2020; "Protecting Personal Information on NHI Card," www.nhi.gov.tw, accessed December 19, 2020; Boo Su-Lyn, "How MySejahtera Protects Your Data and Does More Than Contact Tracing," www.codeblue.galencentre.org, August 12, 2020; "COVID-19 Outbreak in Malaysia: Actions Taken by the Malaysian Government," *International Journal of Infectious Diseases* 97 (August 2020): 108–116, www.sciencedirect.com; techUK, "How Taiwan Used Tech to Fight COVID-19," Press Release, www.techuk.org, March 31, 2020; Emily Waltz, "Big Data Helps Taiwan Fight Coronavirus," *IEEE Spectrum*, www.spectrum.ieee.org, March 12, 2020; C. Jason Wang, Chun Y. Ng, and Robert H. Brook, "Response to COVID-19 in Taiwan: Big Data Analytics, New Technology, and Proactive Testing," *JAMA* 323(14) (2020):1341–1342. doi:10.1001/jama.2020.3151, www.jamanetwork.com, March 3, 2020.

CASE STUDY QUESTIONS

6-13 Describe the role of big data in flattening the COVID-19 pandemic curve in Taiwan and Malaysia.

6-14 What features should be included in a mobile application intended to combat COVID-19, such as MySejahtera?

6-15 What are the challenges of big data in healthcare?

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

Chapter 6 References

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

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

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

Buff, Anne. "The Conundrum of Data Governance." *Big Data Quarterly* (Fall 2019).

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

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

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

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

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

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

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

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

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

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

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

Madnick, Stuart. "Blockchain Isn't as Unbreakable as You Think." *MIT Sloan Management Review* (Winter 2020).

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

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

_____"Data Governance in the Era of Heightened Regulation." *Big Data Quarterly* (Summer 2019).

_____"Seven Trends Shaping Big Data Into 'All Data'," *Big Data Quarterly* (Fall 2019).

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

Qu, Xinxue (Shawn), and Zhengrui Jiang. "A Time-Based Dynamic Synchronization Policy for Consolidated Database Systems." *MIS Quarterly* 43 No. 4 (December 2019).

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

Richardson, John. "This Is the Decade of Clean Data Management." *Supply Chain Brain* (February 7, 2020).

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

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