Chapter 3

Data, Text, and Document Management

Learning Objectives

1. Describe data, text, and document management as well as their impacts on performance.
2. Understand file management systems.
3. Understand the functions of databases and database management systems.
4. Describe the tactical and strategic benefits of data warehouses, data marts, and data centers.
5. Explain how enterprise content management and electronic records management reduce cost, support business operations, and help companies meet their regulatory and legal requirements.

Integrating *IT*

Chapter 3 LINK LIBRARY

Advizor Solutions, data analytics and visualization *advizorsolutions.com/*
Clarabridge: How Text Mining Works *clarabridge.com/*
SAS Text Miner *sas.com/*
Tableau data visualization software *tableausoftware.com/data-visualization-software/*
EMC Corp., enterprise content management *emc.com/*
Oracle DBMS *oracle.com/*

QUICK LOOK at Chapter 3, Data, Text, and Document Management

This section introduces you to the business issues, challenges, and IT solutions in Chapter 3. Topics and issues mentioned in the Quick Look are explained in the chapter.

Results of the *Information, Unplugged* research study published in mid-2010 by Informatica Corporation revealed that enterprises are drowning in their own data (Silva, 2010). The IT department faces enormous challenges because of the surge in data flowing through today's enterprise applications and databases. Some enterprise applications and databases increase in size by as much as 50 percent per year. More than 87 percent of respondents blamed database and network performance issues on data growth. Having more data makes it harder to find the information you need, or the cost of managing that data may exceed the value of the information.

The data explosion is in part the result of the many more channels bringing in data, more types of data, and more complete data. In this chapter, we focus on data and text management, business records, and data infrastructure. Data infrastructure refers to the fundamental structure of an information system (IS), which determines how it functions and how flexible it is in meeting future data requirements. This chapter covers how effective data management improves the performance and productivity of enterprises, managers, and data workers. You will learn the importance of master data management, which improves data sharing, and the regulatory and legal requirements for managing electronic records that all types of organizations currently face.

Performance of every type of organization, including police departments (as you will read in the public sector case at the end of this chapter), depends on timely access to data that can be quickly analyzed and used to anticipate needs of customers, suppliers, or business partners.

Wendy's International Relies on Text Mining for Customer Experience Management

Collecting and analyzing customer feedback to quickly spot product or service problems is part of the new customer experience management (CEM) strategy of Wendy's International (*wendys.com/*), the well-known fast-food chain (Figure 3.1). Wendy's invested in text mining technology, also known as text analytics, from Clarabridge (*clarabridge.com/*) to enhance its CEM program. **Text** is unstructured data and an asset that can be managed. **Text mining** consists of powerful software tools to discover and extract knowledge from text documents; these tools can integrate text-based information with structured data and predictive analytics for better answers to complex questions.

Wendy's text mining software analyzes half a million text-based customer comments per year. Previously, the company's customer satisfaction team used manual methods to process text. The team had used a combination of spreadsheets and keyword searches to review comments manually, a process that was both slow and limited. To compete successfully against strong rivals McDonald's and Burger King, management wanted a solution that would speed analysis, detect

Figure 3.1 Wendy's International Inc. fast-food restaurant.

emerging issues, and pinpoint troubled business areas at the store, regional, and corporate levels.

Gaining Insight by Mining and Managing Text

Wendy's collects information from customers via multiple channels that include Web-based feedback forms, call center notes, e-mail messages, receipt-based surveys, and social media. Most feedback is in unstructured text format, not structured numeric format. For instance, comments like "better menu than McD & BK" or "messed up my order" cannot be put into a database and analyzed as can a customer rating of 5 for "menu options" and 1 for "quality of service" survey questions. A major benefit of text, however, is that it is more robust, that is, it's more complete and precise.

With text mining, managers receive reports on detailed customer experiences down to the store level within minutes. This knowledge enables corporate, regional, and store managers to spot and address problems related to the three customer experience factors impacting the company's profitability—cleanliness, speed of service, and quality of the meals.

Text Mining Combined with Other Analytics

Over the past decade, text analytics has evolved from a mysterious technology used almost exclusively by government intelligence agencies and huge financial firms to a toolkit that's being used by smaller organizations. Factors that have increased the use of text mining are the increasing amounts of text and decreasing costs of the software. Choice Hotels and Gaylord Hotels, for example, use text analytics to quickly make sense of thousands of customer satisfaction surveys gathered every day. Companies typically integrate text mining with their broader business intelligence (BI) and analytics solutions. The leading text mining and BI vendors include SAS, IBM SPSS, and SAP Business Objects.

Sources: Compiled from Henschen (2010), BusinessWire (2010), and Clarabridge (2010).

For Class Discussion and Debate

1. ***Scenario for Brainstorming and Discussion:*** In making critical decisions for your company or agency, there is a cost ("a price to pay") for errors and/or ignorance. For every industry, managing customer service is essential to productivity and growth. Select an industry, company, or public sector and identify some of the costs due to ignorance about customers' or constituents' experiences. Explain how your selection could benefit from text analytics that provided feedback within 24 hours. Compare and assess your answers with others in your class.

2. ***Debate:*** For service providers, the most accurate and meaningful feedback a customer can provide is communicated as close to the time of service as possible, which is a perfect fit for text messaging. Tapping into the immediacy of text messaging could provide timely and detailed data. Text analytics could then provide insights to improve CES. However, it's been argued that companies seldom fail for lack of insight or vision. Rather, they fail because of poor execution. Execution refers to action taken based on those insights. For this debate, consider the following. Customer service and marketing departments would likely be in favor of the investment in text mining to spot problems as soon as possible. In contrast, the finance and accounting departments might take the stand that giving customers an easy way to complain will increase complaints and that satisfied customers are less likely to comment. In addition, finance/accounting may claim that even if problems were spotted, the cost to fix them might not be worth it.

3. To do: Select one side of the argument, as just described. Debate whether investments in text message collection and mining should be made even if no clear positive ROI (return on investment) from better execution can be determined in advance. Provide convincing arguments either in favor of or against the investment.

3.1 Data, Text, and Document Management

Organizations' most strategic assets are their data, text, and documents. **Assets** are resources with recognized value that are under the control of an individual or organization.

Vast quantities of data, text, and documents are created or collected and then stockpiled through some type of storage method. Often data and documents get stored in multiple locations, perhaps five or more of them. Data, text, and document management helps companies improve productivity by ensuring that people can find what they need without having to conduct a long and difficult search. We use the term *data management* to refer to data, text, and document management, for simplicity, unless otherwise stated.

THE IMPORTANCE OF DATA MANAGEMENT

Why does data management matter? And how much does it matter? These are crucial questions because no enterprise can be effective without high-quality data that is accessible when needed. The goal of data management is to provide the infrastructure and tools to transform raw data into usable corporate information of the highest quality. Data is an organization's informational asset. Just as you study how to manage financial assets (i.e., identify, control, protect, analyze, and invest capital) to maximize their value in accounting and finance courses, here you learn how to manage informational assets. You will notice that the underlying concepts for managing financial and nonfinancial assets are similar. The basic rule is that to maximize earnings, companies invest in data management technologies that increase both of the following:

- opportunity to earn revenues (e.g., customer relationship management, or CRM)
- ability to cut expenses (e.g., inventory management)

When organizations analyze data, they need to consider all relevant expenses, including the expected costs of lost customers and earnings, penalties for noncompliance with regulations, and legal fines and losses stemming from the failure to protect confidential data from identity thieves. They calculate these costs by multiplying the probability of an event by the cost of the loss.

Managers and other decision makers need rapid access to correct, comprehensive, and consistent data across the enterprise if they are to improve their business processes and performance. They make decisions and service customers based on the data available to them. They rely on data retrieved from a data repository, such as a **database** or **data warehouse.** *Databases* store enterprise data that a company's business applications create or generate, such as sales, accounting, and employee data. Data entering the databases from POS terminals, online sales, and other sources is stored in an organized format so it can be managed and retrieved. A *data warehouse* is a specialized type of database that aggregates data from transaction databases so it can be analyzed. For example, management might scrutinize this data to identify and examine business trends in order to support planning and decision making, as you will read in Section 3.3.

Uncertainty: A Constraint on Managers. The viability of business decisions depends on access to high-quality data, and the quality of the data depends on effective approaches to data management. Too often managers and information workers are actually constrained by data that cannot be trusted because it is incomplete, out of context, outdated, inaccurate, inaccessible, or so overwhelming that it requires weeks to analyze. In those situations, the decision maker is facing too much uncertainty to make intelligent business decisions.

Data errors and inconsistencies lead to mistakes and lost opportunities, such as failed deliveries, invoicing blunders, and problems synchronizing data from multiple locations. In addition, data analysis errors that have resulted from the use of inaccurate formulas or untested models have harmed earnings and careers. Here are three examples of damages due to data analysis failures:

- TransAlta is a Canadian power generator company. A spreadsheet mistake led to TransAlta's buying more U.S. power transmission hedging contracts at higher prices than it would have if the decision had been based on accurate information. The data error cost the firm US$24 million (Wailgum, 2007).
- In the retail sector, the cost of errors due to unreliable and incorrect data alone is estimated to be as high as $40 billion annually (Snow, 2008).
- In the healthcare industry, one of the largest industries in the United States, data errors not only increase healthcare costs by billions of dollars but also cost thousands of lives, as discussed in *IT at Work 3.1*.

Well-designed data infrastructures provide employees with complete, timely, accurate, accessible, understandable, and relevant data; this is what data management is about. Data management decisions require tough trade-offs among many complex factors, especially in recessionary times when cost cutting is a powerful force. Cost-cutting efforts should not make it more difficult to generate revenues, but that is what happens in the business, healthcare, and government sectors. In recessionary times, the payoff from having effective IT strategy and planning becomes more evident. We focus now on data management and infrastructure.

IT at Work 3.1

Data Errors Cost Billions of Dollars and Put Lives at Risk

Every day, healthcare administrators and others throughout the healthcare supply chain waste 24 to 30 percent of their time correcting data errors. Each incorrect transaction costs $60 to $80 to correct. In addition, about 60 percent of all invoices among supply chain partners have errors, and each invoice error costs $40 to $400 to reconcile. Altogether, errors and conflicting data increase supply costs by 3 to 5 percent. In other words, each year billions of dollars are wasted in the healthcare supply chain because of supply chain data disconnects, which refer to one organization's IS not understanding data from another's IS. Unless the healthcare system developed a data synchronization tool to prevent data disconnects, any attempts to streamline supply chain costs by implementing new technologies, such as radio frequency identification (RFID) to automatically collect data, would be sabotaged by dirty data. RFID is data transmission using radio waves. **Dirty data**—that is, poor-quality data—lacks integrity and cannot be trusted. New regulatory requirements, such as the Florida Pedigree Act, mandate that important information accompany each drug throughout the supply chain. Using RFID, healthcare companies can capture required information such as drug name, dosage, container size, number of containers, lot/control numbers, and so on.

Consider the problems created by the lack of data consistency in the procurement (purchasing) process. Customers of the Defense Supply Center Philadelphia (DSCP), a healthcare facility operated by the Department of Defense (DoD), were receiving the wrong healthcare items, the wrong quantity of items, or an inferior item at a higher price. Numerous errors occurred whenever a supplier and DSCP or any other DoD healthcare facility referred to the same item (e.g., a surgical instrument) with different names or item numbers. These problems were due in large part to inaccurate or difficult-to-manage data.

For three years, efforts were made to synchronize DoD's medical/surgical data with data used by medical industry manufacturers and distributors. First, the healthcare industry had to develop a set of universal data standards or codes that uniquely identified each item. Those codes would enable organizations to accurately share data electronically because everyone would refer to each specific item the exact same way. A data synchronization tool provided data consistency starting with the cataloging process and proceeding through purchasing and billing operations. Results from this effort improved DSCP's operating profit margin and freed personnel to care for patients rather than spend their time searching through disparate product data. Other improvements and benefits of the data synchronization efforts are the following:

- Accurate and consistent item information enables easier and faster product sourcing. Product sourcing simply means finding products to buy.
- Matching of files ensures lowest contracted price for purchases for quicker, automatic new item entry. If the lowest contracted prices cannot be matched and verified automatically, then it must be done manually.
- Significant reduction in the amount of fraudulent or unauthorized purchasing and unnecessary inventories.
- Leveraged purchasing power to get lower prices because purchase volumes were now apparent.
- Better patient safety.
- Improved operating efficiency and fewer invoice errors

Sources: Compiled from Barlow (2007), Chisholm (2008), and Levine (2007).

Discussion Questions: How does dirty data create waste? Why is data synchronization across an enterprise a challenging problem? How can accurate data and verification systems deter and detect fraud?

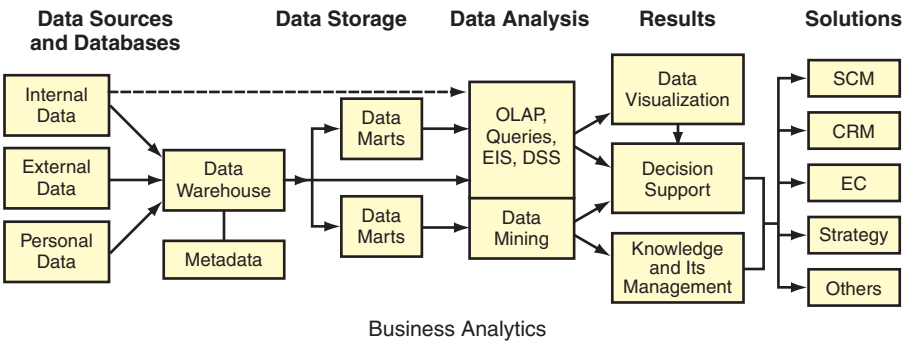

Figure 3.2 Data life cycle.

Data Management. **Data management** is a structured approach for capturing, storing, processing, integrating, distributing, securing, and archiving data effectively throughout its life cycle, as shown in Figure 3.2. The life cycle identifies the way data travels through an organization, from its capture or creation to its use in supporting data-driven solutions, such as supply chain management (SCM), CRM, and electronic commerce (EC). SCM, CRM, and EC are enterprise applications that require current and readily accessible data to function properly. One of the foundational structures of a business solution is the data warehouse. We discuss managing data and applying it to solving business problems throughout this chapter.

Three general data principles illustrate the importance of the data life cycle perspective and guide IT investment decisions.

1. **Principle of diminishing data value.** Viewing data in terms of a life cycle focuses attention on how the value of data diminishes as the data ages. The more recent the data, the more valuable it is. This is a simple, yet powerful, principle. Most organizations cannot operate at peak performance with blindspots (lack of data availability) of 30 days or longer.
2. **Principle of 90/90 data use.** Being able to act on real-time or near-real-time operational data can have significant advantages. According to the 90/90 data-use principle, a majority of stored data, as high as 90 percent, is seldom accessed after 90 days (except for auditing purposes). Put another way, data loses much of its value after three months.
3. **Principle of data in context.** The ability to capture, process, format, and distribute data in near real time or faster requires a huge investment in data management infrastructure to link remote POS systems to data storage, data analysis systems, and reporting applications. The investment can be justified on the principle that data must be integrated, processed, analyzed, and formatted into "actionable information." End users need to see data in a meaningful format and context if the data is to guide their decisions and plans.

Data Visualization. To format data into meaningful contexts for users, businesses employ **data visualization** and decision support tools. Data or information visualization, as the name suggests, refers to presenting data in ways that are faster and easier for users to understand. To better understand this process, examine the two data displays in Figure 3.3. The tabular and graphical displays both depict one-day changes in the Dow Jones Industrial Average (DJIA). The table provides more precise data, whereas the graph takes much less time and effort to understand. Data presentation and visualization tools offer both display options.

Data visualization tools and technology are becoming more popular and widely used as they become less expensive and easier to manipulate. As one example, Dartmouth University's Development Department, which is responsible for fundraising, realized that its efforts to target alumni for contributions to its capital campaign were not as effective as they could be. To reduce missed opportunities, they

Last: **11,893.690** Net Change: **–146.700%** Change: **–1.22%**

Open	12,039.090	52-Week High	14,198.100
High	12,094.210	52-Week Low	11,634.820
Low	11,819.690	Volume	07,579.268

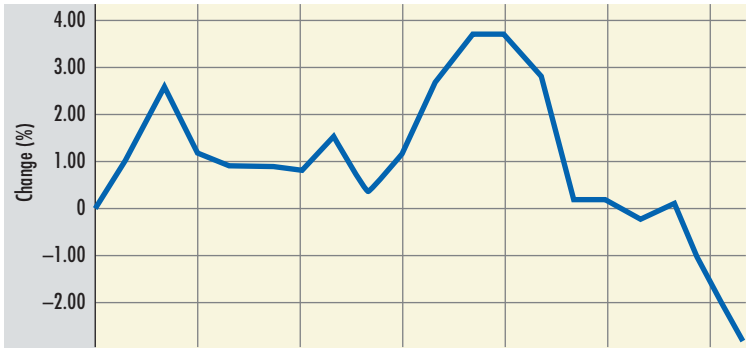

Figure 3.3 Dow Jones industrial average (DJIA) for a single day in tabular display and graphical display.

invested in data visualization tools that they were able to use themselves. As described in *IT at Work 3.2*, these tools enabled the Development Department to overcome data limitations. That is, they knew where and when to invest their time to maximize return on that time. The value of the data visualization tools can be measured by Dartmouth's hugely successful capital campaign.

Data Management: Problems and Challenges. In *IT at Work 3.1*, you read about the problems and costs associated with nonstandardized, unsynchronized data among organizations in the healthcare supply chain. One widespread problem is that people do not get data in the format they need to do their jobs. Therefore, even if the data is accurate, timely, and clean, it still might not be usable. According to the market intelligence firm IDC (*idc.com*), organizations with at least 1,000 knowledge workers (workers who rely on data to perform their jobs) lose $5.7 million annually in time wasted by employees reformatting data as they move among applications. Just as workers waste time tracking down and correcting invoicing and ordering

IT at Work 3.2

Finding Million-Dollar Donors in Three Minutes

The Development Department at Dartmouth University faced a common data problem. Their database stored millions of rows of alumni data, but it was totally dependent on the IT department for reports. Worse, these reports did not contain the types of information that development needed. Specifically, the data could not answer the basic questions that were critical to the success of the $1.3 billion capital campaign:

- Which alumni have the greatest donation potential?.
- Which alumni segments are most likely to donate, and in what ways?
- Which prospects are not donating to their potential?

In order for the managers to explore and navigate on their own and collect data that mattered to their campaign, the university invested in visual discovery software tools from Advizor Solutions (*advizorsolutions.com*). The Development Department used these tools to create a set of dashboards, which they made available over the Web. **Dashboards** are visual displays similar to the dashboard on an automobile (Figure 2.7 in Chapter 2 is an example). Once the dashboards were created, the development managers were able to answer the questions listed above, without help from the IT department. Managers now get answers within three minutes that used to take three weeks due to bottlenecks in the IT department. Most importantly, better-targeted prospect messages and trips have been critical to achieving the goal of the capital campaign.

Sources: Compiled from Advizor Solutions (*advizorsolutions.com*) and Teradata (2007).

Discussion Questions: Why were managers missing opportunities to obtain donations from prospective donors? How did end-user data visualization tools improve the managers' ability to perform their jobs?

errors among healthcare suppliers, they also spend significant amounts of time getting data into usable formats. In Chapter 4 you will read about how businesses resolve some of their data deficiencies using information or enterprise portals. **Enterprise portals** are a set of software applications that consolidate, manage, analyze, and transmit data to users through Web-based interface.

Managing, searching for, and retrieving data located throughout the enterprise is a major challenge, for various reasons:

- The volume of data increases exponentially with time. New data is added constantly and rapidly. Business records must be kept for a long time for auditing or legal reasons, even though the organization itself may no longer access them. Only a small percentage of an organization's data is relevant for any specific application or time.
- External data that needs to be considered in making organizational decisions is constantly increasing in volume.
- Data is scattered throughout organizations and is collected and created by many individuals using different methods, devices, and channels. Data is frequently stored in multiple servers and locations and also in different computing systems, databases, formats, and human and computer languages.
- Data security, quality, and integrity are critical yet easily jeopardized. In addition, legal requirements relating to data differ among countries, and they change frequently.
- Data is being created and used offline without going through quality control checks; hence, the validity of the data is questionable.
- Data throughout an organization may be redundant and out-of-date, creating a huge maintenance problem for data managers.

To deal with these difficulties, organizations invest in data management solutions. Historically, data management has been geared to support transaction processing by organizing the data in one location. This approach supports more secure and efficient high-volume processing. Because the amount of data being created and stored on end-user computers is increasing so dramatically, however, it is inefficient or even impossible for queries and other ad hoc applications to use traditional data management methods. Therefore, organizations have implemented relational databases, in which data is organized into rows and columns, to support end-user computing and decision making. Data organization is covered in Section 3.3.

With the prevalence of **client/server networks** (also called *client/server computing*) and Web technologies, numerous distinct databases are created and spread throughout the organization, creating problems in managing this data so that it's consistent in each location. Client/server networks consist of user PCs, called clients, linked to high-performance computers, called servers, which provide software, data, or computing services over a network. As businesses become more complex and their volumes of enterprise data explode, they increasingly are turning to master data management as a way to intelligently consolidate and manage these data.

MASTER DATA MANAGEMENT

Master Data Management. **Master data management (MDM)** is a process whereby companies integrate data from various sources or enterprise applications to provide a more unified view of the data. Although vendors may claim that their MDM solution creates "a single version of the truth," this claim is probably not true. In reality, MDM cannot create a single unified version of the data because constructing a completely unified view of all master data is simply not possible. Realistically, MDM consolidates data from various data sources into a **master reference file,** which then feeds data back to the applications, thereby creating accurate and consistent data across the enterprise. In *IT at Work 3.1*, participants in the healthcare supply chain were essentially developing a master reference file to obtain a more unified version of the data. A master data reference file is based on data entities. A **data entity** is anything real or abstract about which a company wants to collect and store data. Common data entities are a customer, a vendor, a product, and an employee.

Master Data Entities. **Master data entities** are the main entities of a company, such as customers, products, suppliers, employees, and assets. Each organizational department has distinct master data needs. Marketing, for example, is concerned with product pricing, brand, and product packaging, whereas production is concerned with product costs and schedules. A customer master reference file can feed data to all enterprise systems that have a customer relationship component, thereby providing a unified picture of the customers. Similarly, a product master reference file can feed data to all of the production systems within the enterprise. Three benefits of a unified view of customers are the following:

- Better, more accurate customer data to support marketing, sales, support, and service initiatives
- Better responsiveness to ensure that all employees who deal with customers have up-to-date, reliable information on the customers
- Better revenue management and more responsive business decisions

An MDM includes tools for cleaning and auditing the master data elements as well as tools for integrating and synchronizing data to make the data more accessible. MDM offers a solution for managers who are frustrated with how fragmented and dispersed their data sources are. According to Ventana Research (*ventanaresearch.com*), 4 percent of the 515 respondents to their survey said their organizations had implemented MDM, and 31 percent reported that their organizations had projects currently under way.

Transforming Data into Knowledge. Our discussion thus far has focused primarily on ways in which businesses accumulate and integrate data. Businesses do not run on raw data, however. They run on data that has been processed into information and knowledge, which managers apply to business problems and opportunities. As real-world examples throughout this chapter illustrate, knowledge learned from data fuels business solutions. Everything from innovative product designs to brilliant competitive moves relies on timely knowledge. However, because of the difficulties inherent in managing data, deriving knowledge from collected data is a complicated process.

Organizations transform data into knowledge in several ways. In general, this transformation process resembles the one shown in Figure 3.4. The desired data is

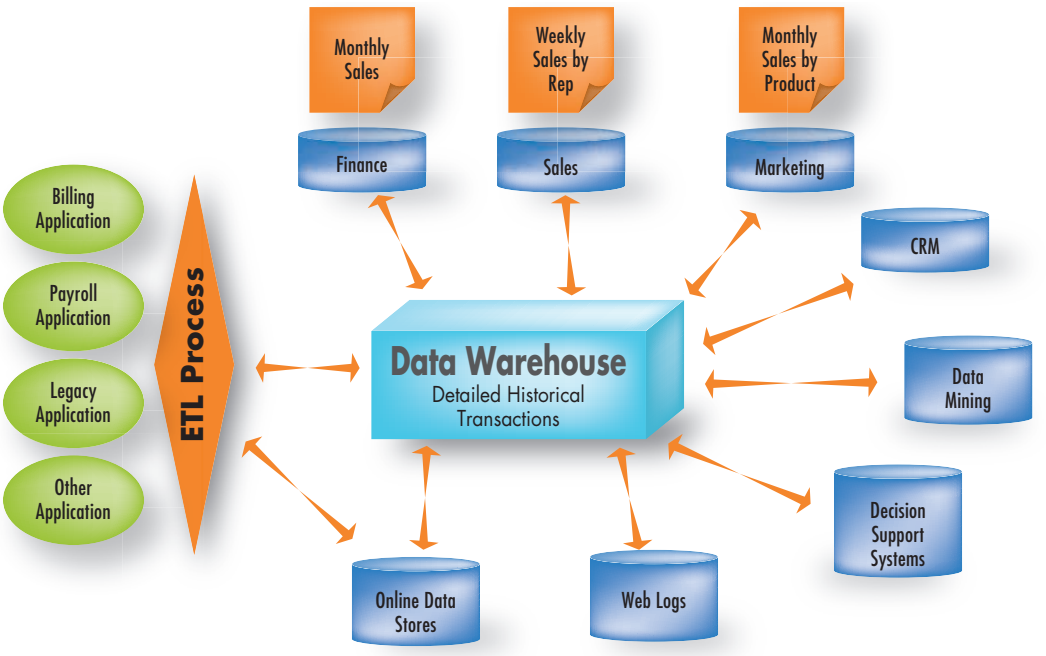

Figure 3.4 Model of an enterprise data warehouse. *(Source: From Syncsort, syncsort.com. Used with permission.)*

extracted from databases and preprocessed to fit the format of a data warehouse or a **data mart,** into which it is loaded. A data mart is a small data warehouse designed for a strategic business unit (SBU) or a single department. This series of processes is referred to as **ETL**, which stands for **extract, transform, and load.** Figure 3.4 shows the data transformation model of Syncsort Inc., which develops, markets, and services high-performance software for data management and data protection. ETL processes move data from multiple sources, reformat and cleanse them, and load them into another data warehouse or data mart for analysis or onto another operational system to support a business process.

Users then access the data warehouse or data mart and take a copy of the data needed for analysis. They scrutinize this material using data analysis and data mining tools. Data mining tools are specialized software used to analyze data to find patterns, correlations, trends, or other meaningful relationships. Data mining, which may also be called data discovery, is the process of analyzing data from different perspectives and summarizing it into information that can be used to increase revenue, decrease costs, or both. Data mining software allows users to analyze data from various dimensions or angles, categorize it, and find correlations or patterns among fields in the data warehouse. These activities ultimately generate valuable information and knowledge. Both the data (at various times during the process) and the knowledge (derived at the end of the process) may need to be sent and presented to users via visualization tools.

Data Quality and Integrity. Data collection is a highly complex process that can create problems concerning the quality of the data that is being collected. Therefore, regardless of how the data is collected, it needs to be validated so users know they can trust it. Classic expressions that sum up the situation are "garbage in, garbage out" (GIGO) and the potentially riskier "garbage in, gospel out." In the latter case, poor-quality data is trusted and used as the basis for planning. You have encountered data safeguards, such as integrity checks, to help improve data quality when you fill in an online form. For example, the form will not accept an e-mail address that is not formatted correctly.

Data quality is a measure of the data's usefulness as well as the quality of the decisions based on the data. It has the following five dimensions: accuracy, accessibility, relevance, timeliness, and completeness. As we have discussed, data frequently is inaccurate, incomplete, or ambiguous, particularly when it is stored in large, centralized databases. Examples of common data problems and possible solutions are listed in Table 3.1.

Although having high-quality data is essential for business success, numerous organizational and technical issues make it difficult to reach this objective. One problematic issue is data ownership. That is, who owns or is responsible for the data? Data ownership issues arise from the lack of policies defining responsibility and accountability in managing data. Inconsistent data quality requirements of various stand-alone applications create an additional set of problems as organizations try to combine individual applications into integrated enterprise systems.

TABLE 3.1 Data Problems and Solutions

Problems	Solutions
Data errors	Use automated data entry, Web forms for individuals entering data with data integrity checks and drop-down menus and radio buttons.
Duplicated data	Redesign the data model; normalize the relational database.
Compromised data	Implement a defense-in-depth approach to data security.
Missing data	Make fields mandatory on data entry forms.

IT at Work 3.3

National Security Depends on Intelligence and Data Mining

Individuals and groups driven by ideological or political motives or the intent to do harm use the Internet to plan and coordinate their activities, as was the case for the September 11, 2001, attacks. Terrorists use various tactics and technologies to carry out their destructive plans—hacking, spamming, phishing, identity theft, and Web site propaganda and recruitment. Computers seized in Afghanistan reportedly revealed that al-Qaeda was collecting intelligence on targets and sending encrypted messages via the Internet.

National security depends on timely intelligence efforts to detect these activities as early as possible. Intelligence agencies, such as the FBI (*fbi.gov*) and CIA (*cia.gov*) in the United States and MI6 (*intelligence.gov.uk/agencies/mi6.asp*) and Defense Intelligence Staff (DIS) (*intelligence.gov.uk/agencies/dis.asp*) in the United Kingdom, mine enormous amounts of data to monitor potential threats to national security. Some data collection might infringe on citizens' privacy rights. The DIS, for example, conducts intelligence analysis from both overt and covert sources.

Data mining for intelligence purposes combines statistical models, powerful processors, and artificial intelligence (AI) to find and retrieve valuable information. There are two types of data mining systems: subject-based systems that retrieve data to follow a lead and pattern-based systems that look for suspicious behaviors. An example of a subject-based technique is link analysis, which uses data to make connections among seemingly unconnected people or events. Link analysis software identifies suspicious activities, such as a spike in the number of e-mail exchanges between two parties (one of whom is a suspect), checks written by different people to the same third party, or airline tickets bought to the same destination on the same departing date. Intelligence personnel then follow these "links" to uncover other people with whom a suspect is interacting.

Experts consider intelligence efforts such as these to be crucial to global security. Some military experts believe that war between major nations is becoming obsolete and that our future defense will rely far more on intelligence officers with databases than on tanks and artillery. A key lesson of September 11 is that America's intelligence agencies must work together and share information to act as a single, unified intelligence enterprise to detect risks.

Sources: Compiled from *Whitehouse.gov* (2004), Volonino et al. (2007), and Worthen (2006).

Discussion Questions: How does data mining provide intelligence to decision makers? What are the two types of data mining systems, and how do they provide value to defense organizations?

Interorganizational information systems add a new level of complexity to managing data quality. Companies must resolve the issues of administrative authority to ensure that each partner complies with the data quality standards. The tendency to delegate data quality responsibilities to the technical teams, who have no control over data quality, as opposed to business users, who do have such control, is another common barrier that stands in the way of accumulating high-quality data (Loshin, 2004).

Data Privacy and Ethics. Businesses that collect data about employees, customers, or anyone else have the duty to protect this data. Data should be accessible only to authorized people. Securing data from unauthorized access and from abuse by authorized parties is expensive and difficult. To motivate companies to invest in data security, the government has imposed enormous fines and penalties for data breaches, as you will read in Chapter 5.

Furthermore, providing information required by the government or regulators adds to the expense of data management. An example is the situation of homeland security described in *IT at Work 3.3*.

GAINING INSIGHT FROM TEXT AND DOCUMENTS

Managers who are committed to fact-based, data-driven decision making are recognizing the power hidden in text to yield insight into marketing, new product development, customer service, public relations, and competition. Techniques for analyzing text, documents, and other unstructured content are available from several vendors.

It's estimated that up to 75 percent of an organization's data is freeform or unstructured, consisting of word-processing documents, content of Web documents,

tweets and other social media, e-mail and text messages, audio, video, images and diagrams, fax and memos, call center or claims notes, and so on. Increasingly, text analytics software is being used to gain insights from freeform content. Gaining business insight is the value of business analytics in general, regardless of the source of the data—textual, numerical, or categorical. Text mining and analytics help organizations manage the information overload.

Two innovative applications of text analytics by organizations are described here:

- Agata, an Italian company, uses social networking tools to implement an online lending system that matches borrowers and investors without intervention from traditional institutions. Its new credit scoring process not only includes quantitative variables from past history and well-defined risk categories; it also integrates qualitative evaluations collected from written descriptions of the projects and business plans to facilitate better decisions on credit risk.
- A Hong Kong government office was faced with the challenge of processing large volumes of structured and unstructured text in traditional Chinese, simplified Chinese, and English. It implemented SAS software to decode and analyze messages in any of these languages, including information from call centers. The result is better public service and increased public satisfaction with the government.

Text Mining and Analytics. Text mining is a broad category that in general involves interpreting words and concepts in context. Then the text is organized, explored, and analyzed to provide actionable insights for managers. With text analytics, information is extracted out of large quantities of various types of textual information. It can be combined with structured data in an automated process.

Text analytics addresses two major business challenges. The first is information organization and the *findability* of the content in documents. The second is discovery of trends and patterns to allow foresight from textual information.

The process of performing analysis on text to discover insights is similar to analyzing traditional data types.

1. **Exploration.** First, documents are explored. This might mean doing simple word counts in a document collection or manually creating topic areas to categorize documents by reading a sample of them. For example, what are the major types of issues (brake or engine failure) that have been identified in recent automobile warranty claims? A challenge of the exploration effort is misspelled or abbreviated words, acronyms, or slang.
2. **Preprocessing.** Before analysis or the automated categorization of the content, the text may need to be preprocessed to standardize it to the extent possible. As in traditional analysis, up to 80 percent of the time can be spent preparing and standardizing the data. Misspelled words, abbreviations, and slang may need to be transformed into consistent terms. For instance, "BTW" would be standardized to "by the way" and "left voice message" could be tagged as "lvm."
3. **Categorizing and Modeling.** Content is then ready to be categorized. Categorizing messages or documents from information contained in them can be achieved using statistical models and business rules. As with traditional model development, sample documents are examined to train the models. Additional documents are then processed to validate the accuracy and precision of the model, and finally new documents are evaluated using the final model (scored). Models can then be put into production for automated processing of new documents as they arrive.

There is considerable overlap between text and document management, but document management has unique issues, which are discussed next.

Document Management. All companies create **business records,** which are documents that record business dealings such as contracts, research and development,

accounting source documents, memos, customer/client communications, and meeting minutes. **Document management** is the automated control of imaged and electronic documents, page images, spreadsheets, voice and e-mail messages, word-processing documents, and other documents through their life cycle in an organization, from initial creation to final archiving or destruction.

Document management systems (DMS) consist of hardware and software that manage and archive electronic documents and also convert paper documents into e-documents and then index and store them according to company policy. For example, companies may be required by law to retain financial documents for at least seven years, whereas e-mail messages about marketing promotions would be retained for a year and then discarded. DMS's have query and search capabilities so they can be identified and accessed like data in a database. These systems range from those designed to support a small work group to full-featured, Web-enabled enterprise-wide systems. DMS may be part of a newer integrated system called enterprise content management (ECM), which is discussed in Section 3.5.

Departments or companies whose employees spend most of the day filing or retrieving documents or warehouse paper records can reduce costs significantly with DMS. These systems minimize the inefficiencies and frustration associated with managing paper documents and paper workflows. Significantly, however, they do not create a paperless office, as had been predicted. Offices still use a lot of paper.

A DMS can help a business to become more efficient and productive by the following:

- Enabling the company to access and use the content contained in the documents
- Cutting labor costs by automating business processes
- Reducing the time and effort required to locate information the business needs to support decision making
- Improving the security of the content, thereby reducing the risk of intellectual property theft
- Minimizing the costs associated with printing, storage, and searching for content

The major document management tools are workflow software, authoring tools, scanners, and databases. When workflows are digital, productivity increases, costs decrease, compliance obligations are easier to verify, and **green computing** becomes possible. Green computing is an initiative to conserve our valuable natural resources by reducing the effects of our computer usage on the environment. Businesses also use a DMS for disaster recovery and business continuity, security, knowledge sharing and collaboration, and remote and controlled access to documents. Because DMS's have multilayered access capabilities, employees can access and change only the documents they are authorized to handle. Visit *altimate.ca/flash/viewer.html* to see how files can be opened directly within the Web browser without the file's native application being installed locally on the user's computer. When companies select a DMS, they ask the following questions:

1. Is the software available in a form that makes sense to your organization, whether you need the DMS installed on your network or will purchase the service?
2. Is the software easy to use and accessible from Web browsers, office applications and e-mail applications, and Windows Explorer? (If not, people won't use it.)
3. Does the software have lightweight, modern Web and graphical user interfaces that effectively support remote users via an intranet, a virtual private network (VPN, discussed in Chapter 4), or the Internet? A VPN allows a worker to connect to a company's network remotely through the Internet. A VPN is less expensive than having workers connect using a modem or dedicated line.

IT at Work 3.4 describes how several companies currently use DMS.

IT at Work 3.4

How Companies Use Document Management Systems

HRM FIN GOV LESSONS FROM IT FAILURES

Below are some examples of how real-life companies utilize DMS.

- The Surgery Center of Baltimore stores all medical records electronically, providing instant patient information to doctors and nurses anywhere and at any time. The system also routes charts to the billing department, which can then scan and e-mail any relevant information to insurance providers and patients. The DMS helps maintain the required audit trail, including providing records when they are needed for legal purposes. How valuable has the DMS been to the center? Since it was implemented, business processes have been expedited by more than 50 percent, the costs of these processes have been significantly reduced, and the morale of office employees in the center has improved noticeably.
- American Express (AMEX) uses TELEform, a DMS developed by Alchemy and Cardiff Software, to collect and process more than 1 million customer satisfaction surveys every year. The data is collected in templates that consist of more than 600 different survey forms in 12 languages and 11 countries. AMEX integrated TELEform with AMEX's legacy system, which enables it to distribute processed results to many managers. Because the survey forms are now so readily accessible, AMEX has been able to reduce the number of staff who process these forms from 17 to 1, thereby saving the company more than $500,000 each year (see *cardiff.com/customers/index.html*).
- In Toronto, Canada, the Department of Works and Emergency Services uses a Web-based document-retrieval solution. This DMS gives the department's employees immediate access to drawings and documents related to roads, buildings, utility lines, and other structures. The department has installed laptop computers loaded with maps, drawings, and historical repair data in each vehicle. Quick access to these documents enables emergency crews to solve problems and, more importantly, to save lives (see *laserfiche.com/newsroom/torontoworks.html)*.
- The University of Cincinnati provides authorized access to the personnel files of 12,000 active employees and tens of thousands of retirees. The university receives more than 75,000 queries about personnel records every year and then must search more than 3 million records to answer these queries. Using a microfilm system to find answers took days. The solution was a DMS that digitized all paper and microfilm documents, without help from the IT department, making them available via the Internet and the university's intranet. An authorized employee can now use a browser and access a document in seconds (see *captaris.com/alchemy*).
- Finally, the European Court of Human Rights has implemented a Web-based knowledge portal and document and case management systems that support more than 700 internal users and millions of external users worldwide. The DMS has streamlined case processing, which in turn has made internal operations more efficient and has significantly improved the court's services to the public. The Human Rights Documents project has had a significant return on investment (see *opentext.com/customers/ globalstar/awards/2007/winners-na-2007.html*).

Discussion Questions: What types of waste can DMS reduce? How? What is the value of providing access to documents via the Internet or a corporate intranet?

Review Questions

1. What is the goal of data management?
2. What constraints do managers face when they cannot trust data?
3. Why is it difficult to manage, search, and retrieve data located throughout the enterprise?
4. How can data visualization tools and technology improve decision making?
5. What is master data management?
6. What is text and document management?
7. What are three benefits of document management systems?

3.2 File Management Systems

The previous section discussed how businesses use computer systems, particularly DMS, to manipulate data much more efficiently and productively. In this section we explain how these systems actually work.

A computer system essentially organizes data into a hierarchy that begins with bits and proceeds to bytes, fields, records, files, and databases (see Figure 3.5). A **bit** represents the smallest unit of data a computer can process, which is either a 0 or a 1. A group of eight bits, called a **byte,** represents a single character, which can be a letter, a number, or a symbol. Characters that are combined to form a word, a group of words, or a complete number constitute a **field.** A key characteristic of a field is that all of the entries are related in some way. For example, a field titled "Cust_Name"

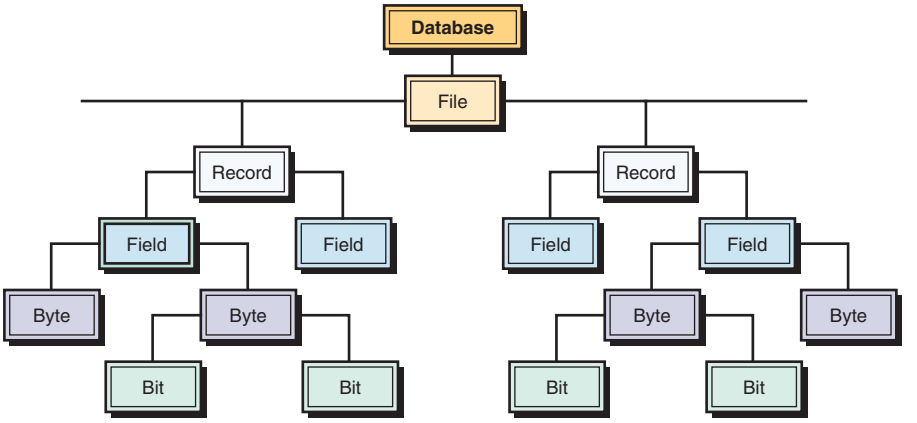

Figure 3.5 Hierarchy of data for a computer-based file.

might include the names of a company's customers. It would not, however, contain addresses or telephone numbers.

Just as related characters can be combined into a field, related fields—such as vendor name, address, and account data—can constitute a **record.** Moving up the hierarchy, a collection of related records is called a **file** or data file. For example, the records of all noncommercial customers who have a mortgage loan at a financial institution would constitute a data file. Finally, as we saw in our discussion of data management, a logical group of related files would constitute a database. All customer loan files, such as mortgages and auto, personal and home equity loans, could be grouped to create a noncommercial loan database.

Another way of thinking about database components is that a record describes an *entity*. Each characteristic describing an entity is called an **attribute.** An attribute corresponds to a field on a record. Examples of attributes are customer name, invoice number, and order date.

Each record in a database needs an attribute (field) to uniquely identify it so that the record can be retrieved, updated, and sorted. This unique identifier field is called the **primary key.** Primary keys are typically numeric because they are easier to create. For example, the primary key of a product record would be the product ID. To find a group of records based on some common value (locating all products manufactured in Mexico) requires the use of secondary keys. **Secondary keys** are nonunique fields that have some identifying information (e.g., country of manufacture). **Foreign keys** are keys whose purpose is to link two or more tables together. Figure 3.6 illustrates primary and foreign keys.

Level 3 Customer Data Tables

Product Preferences Table

Field Name	Type	Length
Product SKU *	Number	12
Product name	Text	20
Product type	Text	12
Product characteristics	Number	30
Customer ID **	Number	8

Shop Locations Table

Field Name	Type	Length
Store number *	Number	8
Store neighborhood	Text	20
Store sales	Currency	8
Store customers	Number	5
Product SKU **	Number	12

Customer Contact Table

Field Name	Type	Length
Customer ID *	Number	8
Customer name	Text	20
Customer address	Text	30
Customer phone	Number	12

Product Sales Table

Field Name	Type	Length
Product SKU *	Number	12
Product name	Text	20
Product price	Currency	8
Product cost	Currency	8
Customer ID **	Number	8

** Primary Key ** Foreign Key*

Figure 3.6 Example of primary and foreign keys.

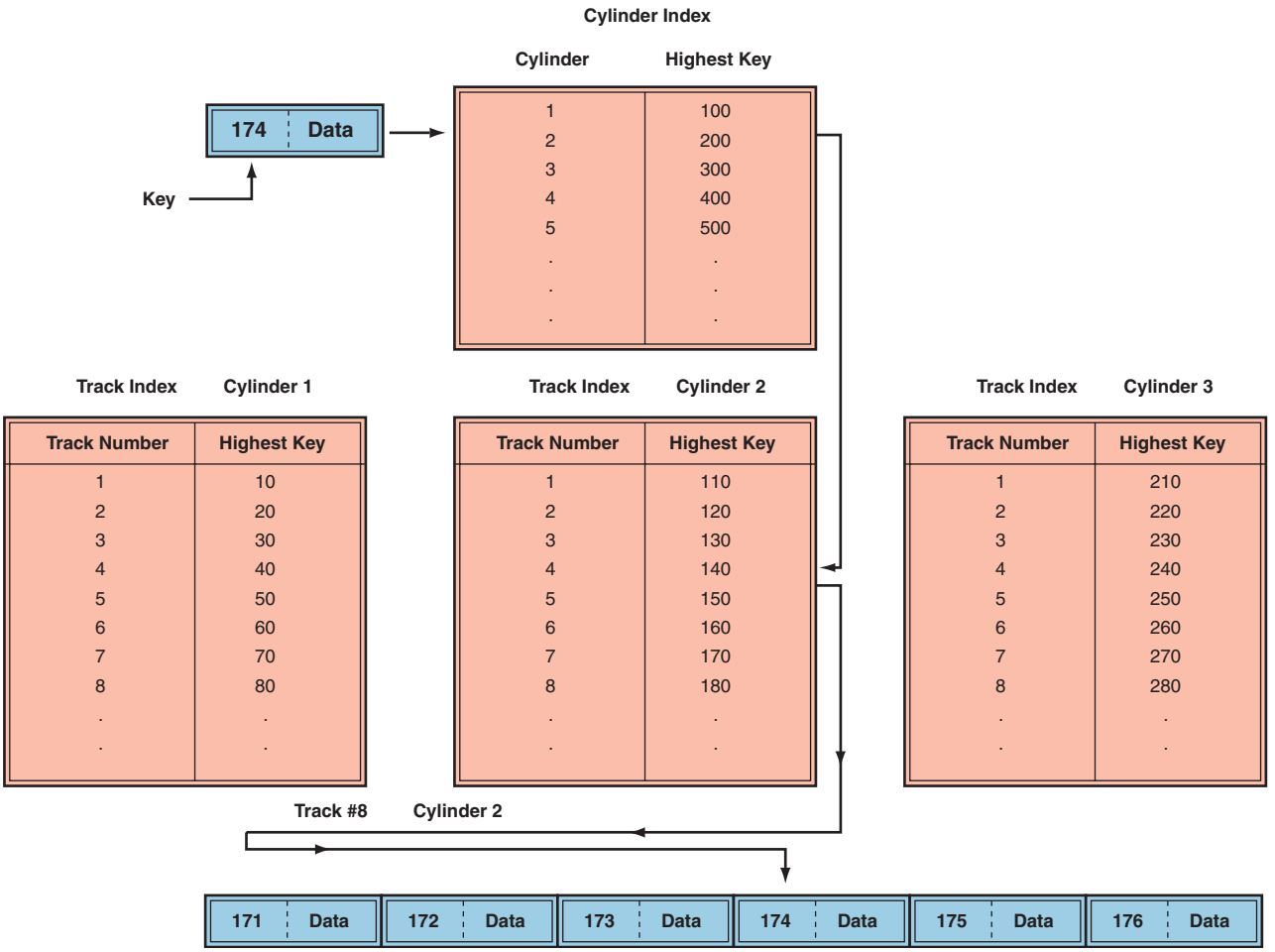

Figure 3.7 Indexed sequential access method.

Accessing Records from Computer Files. Records can be arranged in several ways on a storage medium. The arrangement determines how individual records can be accessed and how long it takes to access them. In **sequential file organization,** which is the way files are organized on tape, data records must be retrieved in the same physical sequence in which they are stored. The operation is like a tape recorder. In **direct file organization** or **random file organization,** records can be accessed directly regardless of their location on the storage medium. The operation is like a DVD drive. Magnetic tape uses sequential file organization, whereas magnetic disks use direct file organization.

The **indexed sequential access method (ISAM)** uses an index of key fields to locate individual records (see Figure 3.7). An *index* to a file lists the key field of each record and where that record is physically located on the storage media. Records are stored on disks in their key sequence. To locate a specific record, the system looks at the index (called track index) to locate the general location (identified by the cylinder and track numbers) containing the record. It then points to the beginning of that track and reads the records sequentially until it finds the correct record.

Limitations of the File Environment. When organizations began using computers to automate processes, they started with one application at a time, usually accounting, billing, or payroll. Each application was designed to be a stand-alone system that worked independently of other applications. For example, for each pay period, the payroll application would use its own employee and wage data to calculate and process the payroll. No other application would use this data without some manual

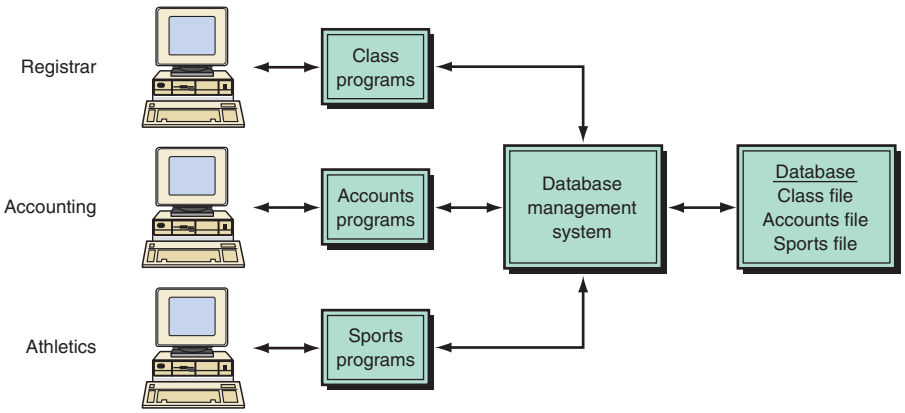

Figure 3.8 Database management system provides access to all data in the database.

intervention because, as just stated, the applications functioned independently of one another. This data file approach led to redundancy, inconsistency, data isolation, and other problems. We examine these problems below, and we illustrate them by showing a university file environment in Figure 3.8.

- **Data redundancy.** Because different programmers create different data-manipulating applications over long periods of time, the same data could be duplicated in several files. For example, in a loan application, each data file contains records about customers' loans. Many of these customers will be represented in other data files. This redundancy wastes physical storage media, makes it difficult to obtain a comprehensive view of customers, and increases the costs of entering and maintaining the data.
- **Data inconsistency.** Data inconsistency means that the actual data values are not synchronized across various copies of the data. Recall that unsynched data caused the problems faced by Dartmouth's development office, discussed earlier in this chapter. For example, if a financial institution has customers with several loans, and for each loan there is a file containing customer fields (e.g., name, address, e-mail, and telephone number), then a change to a customer's address in only one file creates inconsistencies with the address field in other files.
- **Data isolation.** File organization creates silos of data that make it extremely difficult to access data from different applications. For example, a manager who wants to know which customers owe more than $1,000 would probably not be able to obtain the answer from a data file system. To get the results, he would have to filter and integrate the data manually from multiple files.
- **Data security.** Securing data is difficult in the file environment because new applications are added to the system on an ad hoc basis. As the number of applications increases, so does the number of people who can access the data.

Data management problems arising from the file environment approach led to the development of better data management systems.

Review Questions

1. What are three limitations of the file management approach?
2. Why does each record in a database need a unique identifier (primary key)?
3. How do the data access methods of sequential file organization and direct file access methods differ?

3.3 Databases and Database Management Systems

Data flows into companies continuously and from many sources: clickstream data from Web and e-commerce applications, detailed data from POS terminals, and filtered data from CRM, supply chain, and enterprise resource planning applications. Databases are the optimal way to store and access organizational data.

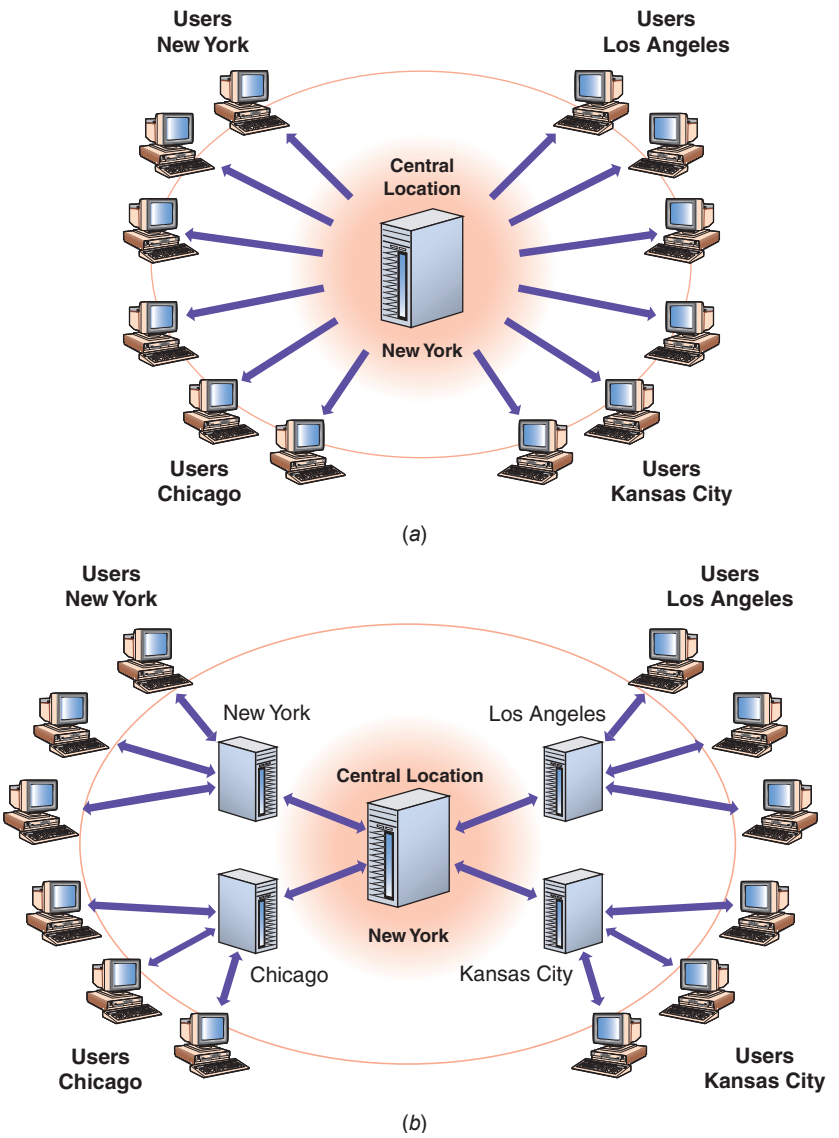

Figure 3.9 (a) Centralized database. (b) Distributed database with complete or partial copies of the central database in more than one location.

DATABASES

Database management programs can provide access to all of the data, alleviating many of the problems associated with data file environments. Therefore, data redundancy, data isolation, and data inconsistency are minimized, and data can be shared among data users. In addition, security and data integrity are easier to control, and applications are independent of the data they process. There are two basic types of databases: centralized and distributed.

Centralized Databases. A centralized database stores all related files in one physical location (see Figure 3.9). For decades the main database platform consisted of centralized database files on large, mainframe computers, primarily because of the enormous capital and operating costs associated with the alternative systems. Centralized databases offer many benefits to organizations. Files can generally be made more consistent with one another when they are physically kept in one location because file changes can be made in a supervised and orderly fashion. Also, files are not accessible except via the centralized host computer, where they can be protected more easily from unauthorized access or modification.

At the same time, however, centralized databases, like all centralized systems, are vulnerable to a single point of failure. That is, when the centralized database

computer fails to function properly, all users are affected. Additionally, when users are widely dispersed and must perform data manipulations from great distances, they often experience transmission delays.

Distributed Databases. A *distributed database* has complete copies of a database or portions of a database (see Figure 3.9). There are two types of distributed databases: replicated and partitioned.

A *replicated database* stores complete copies (replicas) of the entire database in multiple locations. This arrangement provides a backup in case of a failure or problems with the centralized database. It also improves the response time because it is local (closer) to users. On the negative side, it is much more expensive to set up and maintain because each replica must be updated as records are added to, modified in, and deleted from any of the databases. The updates may be done at the end of a day or some other schedule as determined by business needs. Otherwise, the various databases will contain conflicting data.

In contrast, a *partitioned database* is divided up so that each location has a portion of the entire database—usually the portion that meets users' local needs. Partitioned databases provide the response speed of localized files without the need to replicate all changes in multiple locations. One significant advantage of a partitioned database is that data in the files can be entered more quickly and kept more accurate by the users immediately responsible for the data.

DATABASE MANAGEMENT SYSTEMS (DBMS)

A program that provides access to databases is known as a **database management system (DBMS).** The DBMS permits an organization to centralize data, manage it efficiently, and provide access to the stored data by application programs. DBMSs range in size and capabilities from the simple Microsoft Access to full-featured Oracle and DB2 solutions. Table 3.2 lists the major capabilities and advantages of DBMSs.

The DBMS acts as an interface between application programs and physical data files (see Figure 3.10). It provides users with tools to add, delete, maintain, display, print, search, select, sort, and update data. These tools range from easy-to-use natural

TABLE 3.2 Advantages and Capabilities of a DBMS

- **Permanence.** Data is permanently stored on a hard drive or other fast, reliable medium until explicitly removed or changed.
- **Querying.** Querying is the process of requesting data from various perspectives. Example: "How many trucks in Texas are green?"
- **Concurrency.** Many people may attempt to change or read the same data at the same time. Without rules for sharing changes, the data may become inconsistent or misleading. For example, if you change the color attribute of car 7 to be "blue" at the very same time somebody is changing it to "red," results are unpredictable.
- **Backup and replication.** Backup copies need to be made in case of equipment failure.
- **Rule enforcement.** Rules are applied to keep data clean and trustworthy. For example, a rule can state that each car can have only one engine associated with it (identified by engine number). If somebody tries to associate a second engine, the DBMS stops it and displays an error message. However, with new hybrid gas-electric cars, such rules may need to be relaxed. Rules can be added and removed as needed without significant redesign.
- **Security.** Limits on who can see or change attributes are necessary.
- **Computation.** Rather than have each computer application perform calculations, the DBMS performs them.
- **Change and access logging.** The DBMS creates a record and audit trail of who accessed what attributes, what was changed, and when it was changed.
- **Automated optimization.** If there are frequent usage patterns or requests, many DBMSs can adjust to improve the response time.

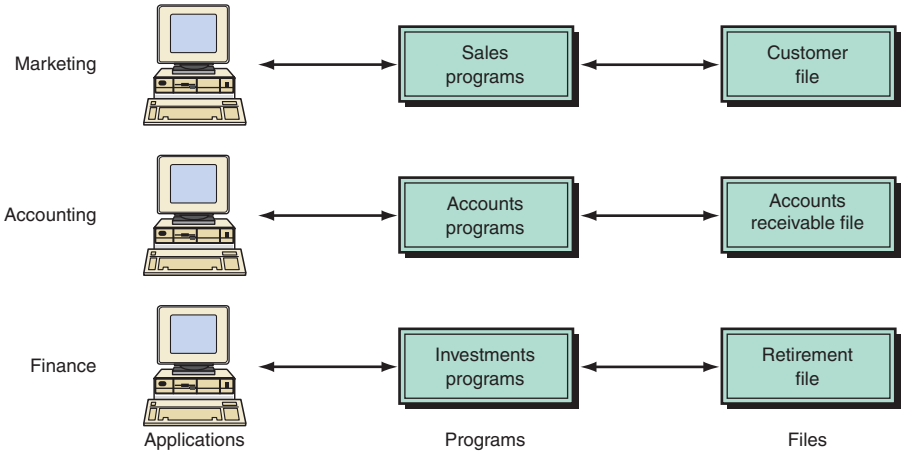

Figure 3.10 Computer-based files of this type cause problems such as redundancy, inconsistency, and data isolation.

language interfaces to complex programming languages used for developing sophisticated database applications. The major data functions performed by a DBMS are listed below.

- **Data filtering and profiling:** Inspecting the data for errors, inconsistencies, redundancies, and incomplete information.
- **Data quality:** Correcting, standardizing, and verifying the integrity of the data.
- **Data synchronization:** Integrating, matching, or linking data from disparate sources.
- **Data enrichment:** Enhancing data using information from internal and external data sources.
- **Data maintenance:** Checking and controlling data integrity over time.

Companies use DBMSs in a broad range of information systems. Some DBMSs, such as Microsoft Access, can be loaded onto a single user's computer and accessed in an ad hoc manner to support individual decision making. Others, such as IBM's DB2, are located on interconnected computers to support large-scale transaction processing systems, such as order entry and inventory control systems. DBMSs such as Oracle 11g, are interconnected throughout an organization's local area networks (LANs), giving departments access to corporate data. LANs are private networks owned and managed by the organization; they are discussed in detail in Chapter 4.

A DBMS enables many different users to share data and process resources. How can a single, unified database meet the differing requirements of so many users? For example, how can a single database be structured so that sales personnel can view customer, inventory, and production maintenance data while the human resources department maintains restricted access to private personnel data?

The answer is that a DBMS provides two views of the data: a physical view and a logical view. The physical view deals with the actual, physical arrangement and location of data in the direct access storage devices (DASDs). Database specialists use the physical view to configure storage and processing resources.

Users, however, need to see data differently from how they are stored, and they do not want to know all of the technical details of physical storage. After all, a business user is primarily interested in using the information, not in how it is stored. The logical view, or user's view, of data is meaningful to the user. What is important is that a DBMS provides endless logical views of the data. This feature allows users to see data from a business-related perspective rather than from a technical viewpoint. Clearly, users must adapt to the technical requirements of database information systems to some degree, but the logical views allow the system to adapt to the business needs of the users. The way in which you see data (the *logical view* or user's view) can vary; but the physical storage of data (*physical view*) is fixed.

Review Questions

1. What is a database? A database management system (DBMS)?
2. What are three data functions of a DBMS?
3. What is the difference between the physical view of and the logical view of data?

3.4 Data Warehouses, Data Marts, and Data Centers

It's not necessarily the biggest companies that are the most successful, but the smartest ones. Being a smart company means having on-demand access to relevant data, understanding it (usually with the help of data visualization tools), and using what you learn from it to increase productivity and/or profitability. Having complete information is critical to this process. Data warehouses enable managers and knowledge workers to leverage data for advantage from across the enterprise, thereby helping them make the smartest decisions.

Recall from our discussion of data management that a data warehouse is a repository in which data is organized so that it can be readily analyzed using methods such as data mining, decision support, querying, and other applications. Examples of uses of a data warehouse are revenue management, customer-relationship management, fraud detection, and payroll-management applications. To better understand data warehouses, it helps to compare them to databases.

COMPARING DATABASES TO DATA WAREHOUSES

Data warehouses and regular databases both consist of data tables (files), primary and other keys, and query capabilities. The main difference is that databases are designed and optimized to store data, whereas data warehouses are designed and optimized to respond to analysis questions that are critical for a business.

Databases are **online transaction processing (OLTP) systems** in which every transaction has to be recorded quickly. Consider, for example, financial transactions, such as withdrawals from a bank ATM or a debit account. These transactions must be recorded and processed as they occur, that is, in real time. Consequently, database systems for banking and debit cards are designed to ensure that every transaction gets recorded immediately.

Databases are volatile because data is constantly being added, edited, or updated. Consider a database of a bank. Every deposit, withdrawal, loan payment, or other transaction adds or changes data. The volatility caused by the transaction processing makes data analysis too difficult. To overcome this problem, data are extracted from designated databases, transformed, and loaded into a data warehouse. Significantly, these data are read-only data; that is, it cannot be updated. Rather, it remains the same until the next scheduled data extraction, transformation, and load (ETL). Unlike databases, then, warehouse data are not volatile. Thus, data warehouses are designed as **online analytical processing (OLAP) systems,** meaning that the data can be queried and analyzed much more efficiently than OLTP application databases.

REAL-TIME SUPPORT FROM A DATA WAREHOUSE

The modern business world is experiencing a growing trend toward real-time data warehousing and analytics. In the past, data warehouses primarily supported strategic applications, which did not require instant response time, direct customer interaction, or integration with operational systems. Today, businesses increasingly use information in the moment to support real-time customer interaction. Companies with an active data warehouse will be able to interact appropriately with a customer to provide superior customer service, which in turn improves revenues.

Companies, such as the credit card company Capital One, track each customer's profitability and use that score to determine the level of customer service. For example, when a customer calls Capital One, that customer is asked to enter the credit card number, which is linked to a profitability score. Low-profit customers get a voice

response unit only; high-profit customers get a live person—a customer service representative (CSR).

Consider, for example, the case of Charles, who is calling the customer service center because of frequent dropped cell calls. Through the call center application (attached to the active data warehouse), the CSR accesses not only the complete history of Charles's calls to the company but also a full view of all the services to which he subscribes—DSL, Internet, and cellular—along with his customer profitability score, which lets the CSR know how profitable (valuable) he is to the company. Intelligent and selective customer service is possible because all service lines and calls to customer service are stored in the active data warehouse. The CSR uses the data and company information to determine the best action or offer to resolve this issue to Charles's satisfaction. Additionally, the CSR will have the insight to cross-sell or up-sell additional services based on the details in Charles' profile and company interaction information. In this example, the active data warehouse provided a view of the customer that indicated what intervention to take based on the customer's profitability to the company. Because Charles subscribes to the company's high-profit-margin services, the company wants to minimize the risk of losing him as a customer.

THE NEED FOR DATA WAREHOUSING

Many organizations built data warehouses because they were frustrated with inconsistent decision support data or they needed to improve reporting applications or better understand the business. Viewed from this perspective, data warehouses are infrastructure investments that companies make to support current and future decision making.

The most successful companies are those that can respond quickly and flexibly to market changes and opportunities, and the key to this response is to use data and information effectively and efficiently. Companies perform this task not only via transaction processing but also through analytical processing, in which company employees—frequently end users—analyze the accumulated data. Analytical processing, also referred to as business intelligence (BI), includes data mining, decision support systems (DSSs), enterprise systems, Web applications, querying, and other end-user activities.

BENEFITS OF DATA WAREHOUSING

According to Teradata Corp., the benefits of a data warehouse are both business- and IT-related. From the business perspective, companies can make better decisions because they have access to better information. From an IT perspective, DWs deliver information more effectively and efficiently. Several areas of an organization that benefit from a DW are the following:

- **Marketing and sales.** Use a DW for product introductions, product information access, marketing program effectiveness, and product line profitability. Use the data to maximize per-customer profitability.
- **Pricing and contracts.** Use the data to calculate costs accurately to optimize pricing of a contract. Without accurate cost data, prices may be below or too near to cost or prices may be uncompetitive because they are too high.
- **Forecasting.** The DW assists in the timely visibility of end-customer demand.
- **Sales performance.** Use the data to determine sales profitability and productivity for all territories and regions; can obtain and analyze results by geography, product, sales group, or individual.
- **Financial.** Use daily, weekly, or monthly results for improved financial management.

Figure 3.11 diagrams the process of building and using a data warehouse. The organization's data are stored in operational systems (left side of the figure). Not all data are necessarily transferred to the data warehouse. Frequently, only a summary of the data is transferred. The data that is transferred is organized within the warehouse in

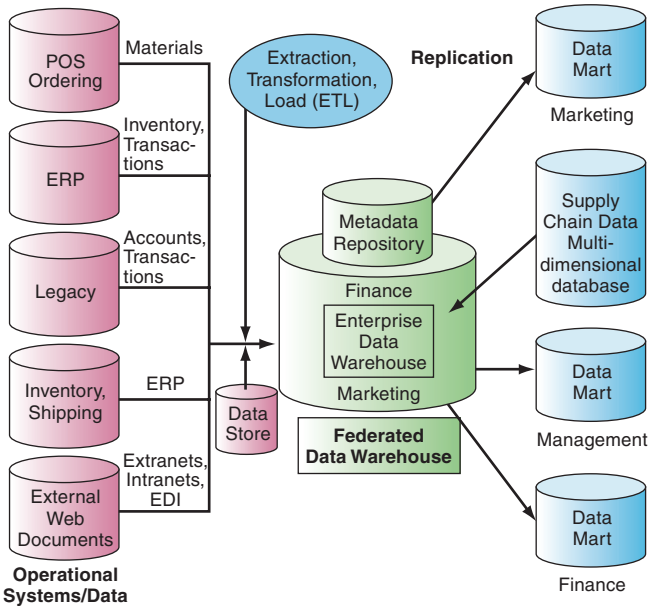

Figure 3.11 Data warehouse framework and views.

a form that is easy for end users to access and locate. The data is also standardized. Then the data is organized by subject, such as by functional area, vendor, or product.

CHARACTERISTICS OF A DATA WAREHOUSE

All types of data warehousing share nine major characteristics:

1. **Organization.** Data is organized by subject (e.g., by customer, vendor, product, price level, and region) and contains information relevant for decision support only.
2. **Consistency.** Data in different databases may be encoded differently. For example, gender data may be encoded 0 and 1 in one operational system, and "m" and "f" in another. In the warehouse it will be coded in a consistent manner.
3. **Time variant.** The data is kept for many years so it can be used for identifying trends, forecasting, and making comparisons over time.
4. **Nonvolatile.** Once the data are entered into the warehouse, they are not updated.
5. **Relational.** Typically the data warehouse uses a relational structure.
6. **Client/server.** The data warehouse uses the client/server architecture mainly to provide the end user an easy access to its data.
7. **Web-based.** Today's data warehouses are designed to provide an efficient computing environment for Web-based applications.
8. **Integration.** Data from various sources is integrated. Web Services are used to support integration.
9. **Real time.** Although most applications of data warehousing are not in real time, it is possible to arrange for real-time capabilities.

BUILDING A DATA WAREHOUSE

Building and implementing a data warehouse can present problems. Because a warehouse is very large and expensive to build, it's important to understand the key success factors in implementing one. Specifically, a company that is considering building a data warehouse first needs to address a series of basic questions:

- Does top management support the data warehouse?
- Do users support the data warehouse?
- Do users want access to a broad range of data? If they do, which is preferable: a single repository or a set of stand-alone data marts?
- Do users want data access and analysis tools?

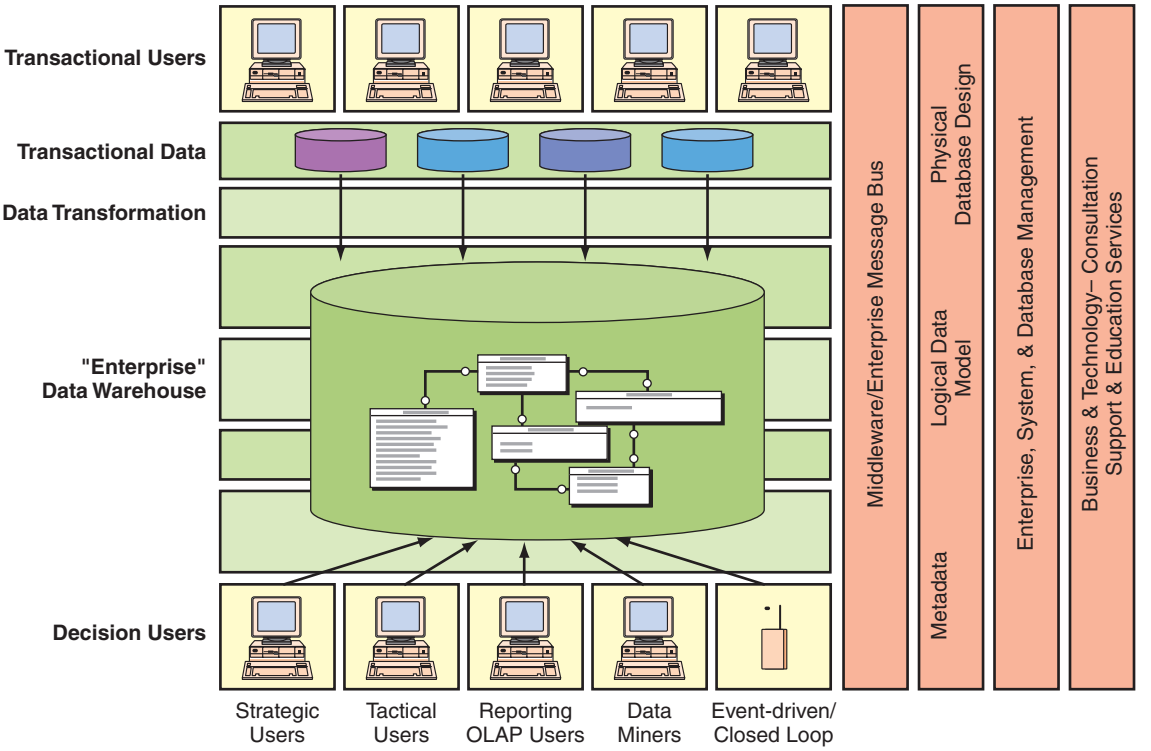

Figure 3.12 Teradata Corp.'s enterprise data warehouse. *(Source: Teradata Corporation [teradata.com], with permission.)*

- Do users understand how to use the data warehouse to solve business problems?
- Does the unit have one or more power users who can understand data warehouse technologies?

Architecture and Tools. There are several basic architectures for data warehousing. Two common ones are two-tier and three-tier architectures. In three-tier architecture, data from the warehouse is processed twice and deposited in an additional multidimensional database, in which it can be organized for easy multidimensional analysis and presentation or replicated in data marts.

There are two main reasons for creating a data warehouse as a separate data store. First, the performance of a separate data store is better because it is not competing (or waiting) for processing time. Second, modeling a database that can be used for both operational and analytical purposes can be difficult. Figure 3.12 represents an EDW developed by Teradata Corp. This centralized approach reduces the amount of data the technical team has to transfer, thereby simplifying data management and administration. Users are also provided with access to all of the data in the data warehouse instead of being limited to individual data marts.

Putting the Warehouse on the Intranet. Data warehouse content can be delivered to decision makers throughout the enterprise via an intranet. Users can view, query, and analyze the data and produce reports using Web browsers. This is an extremely economical and effective method of delivering data.

Suitability. Data warehousing is most appropriate for organizations that have some of the following characteristics:

- End users need to access large amounts of data.
- The operational data is stored in different systems.
- The organization employs an information-based approach to management.

TABLE 3.3 Strategic Uses of Data Warehousing

Industry	Functional Areas of Use	Strategic Use
Airline	Operations and Marketing	Crew assignment, aircraft deployment, mix of fares, analysis of route profitability, frequent-flyer program promotions
Apparel	Distribution and Marketing	Merchandising, inventory replenishment
Banking	Product Development, Operations, and Marketing	Customer service, trend analysis, product and service promotions, reduction of IS expenses
Credit card	Product Development and Marketing	Customer service, new information service for a fee, fraud detection
Defense contracts	Product Development	Technology transfer, production of military applications
E-Business	Distribution and Marketing	Data warehouses with personalization capabilities, marketing/shopping preferences allowing for up-selling and cross-selling
Government	Operations	Reporting on crime areas, homeland security
Healthcare	Operations	Reduction of operational expenses
Investment and insurance	Product Development, Operations, and Marketing	Risk management, market movements analysis, customer tendencies analysis, portfolio management
Retail chain	Distribution and Marketing	Trend analysis, buying pattern analysis, pricing policy, inventory control, sales promotions, optimal distribution channel decisions
Telecommunications	Product Development, Operations, and Marketing	New product and service promotions, reduction of IS budget, profitability analysis

- The organization serves a large, diverse customer base (such as in a utility company or a bank; for example, AT&T's 26-terabyte data warehouse is used for marketing analysis by 3,000 employees).
- The same data is represented differently in different systems.
- Data is stored in highly technical formats that are difficult to decipher.
- Extensive end-user computing is performed (many end users performing many activities).

Table 3.3 summarizes some of the successful applications of data warehouses. Hundreds of other successful applications have been reported (e.g., see client success stories and case studies at Web sites of vendors such as Hyperion Inc., Business Objects, Cognos Corp., Information Builders, NCR Corp., Oracle, Computer Associates, and Software A&G). For further discussion of this topic, visit the Data Warehouse Institute (*tdwi.org/*).

Many organizations, buoyed by the success of their data warehouse efforts, are taking data warehousing public. One example is Wells Fargo. Its development effort uses the resources of a Teradata warehouse to provide an online tool that collects and summarizes transactions for consumers—credit card, debit card, online bill payments, checking account—and generates an analysis of online banking sessions. Consumers are better able to understand their spending patterns, and they have reported a higher level of customer satisfaction.

Another company that continues to grow its enterprise using a public data warehouse is Travelocity. Part of the company's success lies in its innovative use of its EDW for marketing and CRM.

DATA MARTS, OPERATIONAL DATA STORES, AND MULTIDIMENSIONAL DATABASES

Organizations frequently implement data marts, operational data stores, and multidimensional databases either as supplements or substitutes for data warehouses. In this section we take a closer look at these systems, beginning with data marts.

Data Marts. The high costs of data warehouses can make them too expensive for a company to implement. As an alternative, many firms create a lower-cost,

TABLE 3.4 Reasons Data Warehouses Fail

Data warehousing design:
- Unrealistic expectations
- Inappropriate architecture
- Vendors overselling capabilities
- Lack of development expertise
- Lack of effective project sponsorship

Data warehousing implementation:
- Poor user training
- Failure to align data warehouses and data marts
- Lack of attention to cultural issues
- Corporate policies not updated

Data warehousing operation:
- Poor upkeep of technology
- Failure to upgrade modules
- Lack of integration
- Poor data quality

scaled-down version of a data warehouse called a data mart. Data marts are designed for a strategic business unit (SBU) or a single department.

In addition to lower costs (less than $100,000 versus $1 million or more for data warehouses), data marts require significantly shorter lead times for implementation, often less than 90 days. In addition, because they allow for local rather than central control, they confer power on the using group. They also contain less information than the data warehouse. Thus, they respond more quickly, and they are easier to understand and navigate. Finally, they allow a business unit to build its own decision support systems without relying on a centralized IS department.

Operational Data Stores. An **operational data store** is a database for transaction processing systems that use data warehouse concepts to provide clean data. It brings the concepts and benefits of the data warehouse to the operational portions of the business at a lower cost. Thus, it can be viewed as being situated between the operational data (in legacy systems) and the data warehouse. An operational data store is used for short-term decisions involving mission-critical applications rather than for the medium- and long-term decisions associated with the regular data warehouse. These decisions require access to much more current information. For example, a bank needs to know about all the accounts for a customer who is calling on the phone.

System Failures. Unfortunately, despite their potential benefits, implementations of large information systems often fail. Examples and reasons for failures are summarized in Table 3.4. Suggestions on how to avoid data warehouse failure are provided at *datawarehouse.com*, at *bitpipe.com*, and at *teradatauniversitynetwork.com*.

DATA CENTERS

Data center is the name given to facilities containing mission-critical ISs and components that deliver data and IT services to the enterprise. Data centers store and integrate networks, computer systems, and storage devices. Data centers need to ensure the availability of power and provide physical and data security. The newest data centers are huge and include temperature and fire controls, physical and digital security, redundant power supplies such as uninterruptible power sources (UPS), and redundant data communications connections. For example, in 2008, Christus Health Medical Center in Texas built a $23 million data center to house its patient insurance records, CT scans, and other data and documents. The size of the data

center is 48,000 square feet, which is over ten times the size of the hospital's former 4,000-square-foot data center. Demands for imaging data are growing quickly as more types of health reports and records get digitized, stored, and archived for decades.

Many companies are building or reconfiguring their data centers to save money. Some cannot afford the electricity and cooling costs. Others need more computing, storage, or network capacity to handle new applications or to cope with acquisitions. Still others need to improve their disaster recovery capabilities. Creating—or reducing the cost of—a disaster recovery site is often part of a data center upgrade plan.

Next-generation data centers will be more efficient in lowering operating expenses and energy consumption. They will have greater availability (uptime) to meet business needs and will be easier to manage. A networking company, Cisco (*cisco.com*), offers several podcasts and video demos of data centers.

Review Questions

1. What is the main difference in the designs of databases and data warehouses?
2. Compare databases and data warehouses in terms of data volatility and decision support.
3. What is an advantage of an active data warehouse?
4. What are the data functions performed by a data warehouse?
5. How can a data warehouse support a company's compliance requirements and green initiatives?
6. Why are data centers important to performance?

3.5 Enterprise Content Management

Enterprise content management (ECM) has become an important data management technology, particularly for large and medium-sized organizations. ECM includes electronic document management, Web content management, digital asset management, and **electronic records management (ERM)**. ERM infrastructures help reduce costs, easily share content across the enterprise, minimize risk, automate expensive time-intensive and manual processes, and consolidate multiple Web sites onto a single platform.

Four key forces are driving organizations to adopt a strategic, enterprise-level approach to planning and deploying content systems:

- Compounding growth of content generated by organizations
- The need to integrate that content within business processes
- The need to support increasing sophistication for business-user content access and interaction
- The need to maintain governance and control over content to ensure regulatory compliance and preparedness for legal discovery

Modern businesses generate volumes of documents, messages, and memos that, by their nature, contain unstructured content (data or information). Therefore, the contents of e-mail and instant messages, spreadsheets, faxes, reports, case notes, Web pages, voice mails, contracts, and presentations cannot be put into a database. However, many of these materials are business records (as discussed in Section 3.1) that need to be retained. As materials are not needed for current operations or decisions, they are archived—moved into longer-term storage. Because these materials constitute business records, they must be retained and made available when requested by auditors, investigators, the SEC, the IRS, or other authorities. To be retrievable, the records must be organized and indexed like structured data in a database.

Records are different from documents in that they cannot be modified or deleted except in controlled circumstances. In contrast, documents generally are subject to revision. Figure 3.13 shows the differences between documents and records as well as the relationship between document management and records management.

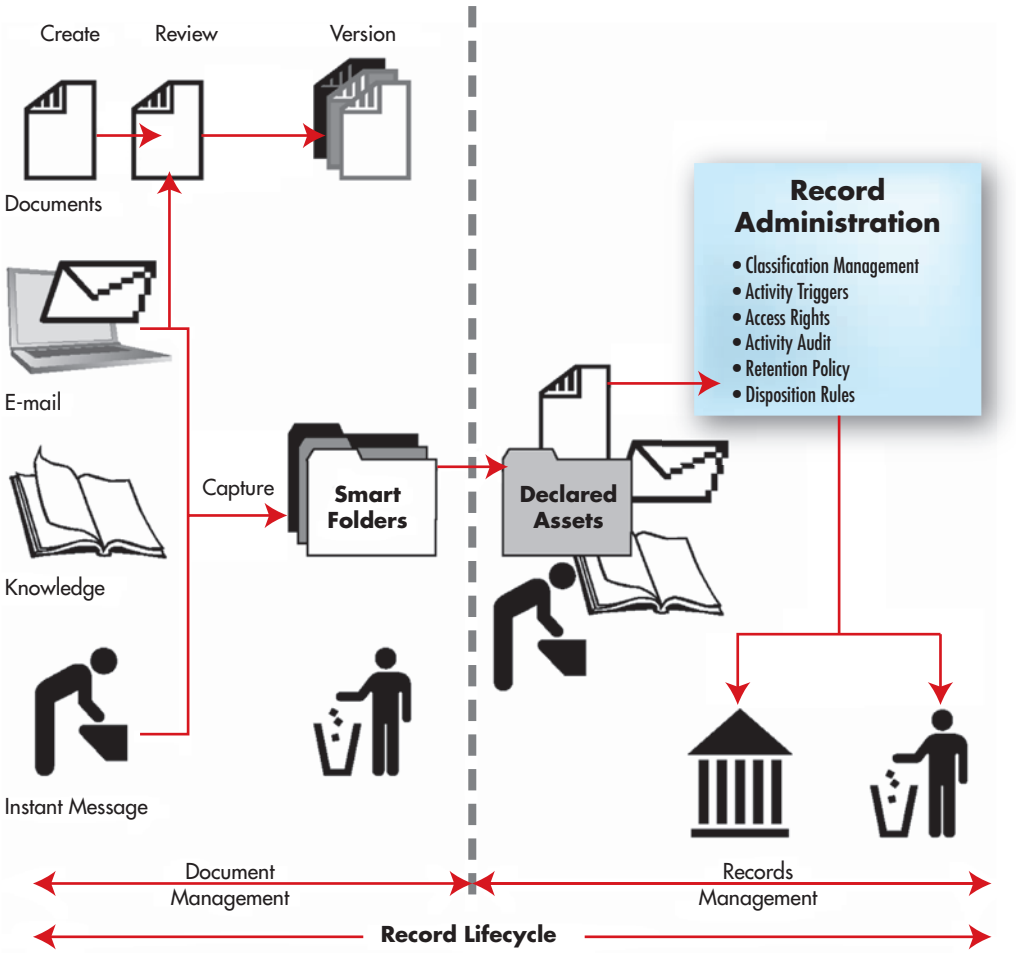

Figure 3.13 Electronic records management from creation to retention or destruction.

Electronic Records and Document Management. In Section 3.1 you read that document management systems organize and store e-mail, instant messages, and other types of unstructured content. In this section, we examine the related topic of ERM.

Simply creating backups of records is not a form of ERM, because the content is not organized so that it can be accurately and easily retrieved. ERM requires the involvement of not only key players in recordkeeping, such as records managers or record librarians, but also IT personnel and administrators under a shared responsibility to establish ERM policies. Those policies include schedules for retaining and destroying records, which must comply with state and federal regulations.

The requirement to manage records—regardless of whether they are paper or electronic—is not new. What is new is the volume of electronic records that must be reviewed to determine whether they should be retained or destroyed. Properly managed, electronic records are strategic assets. Improperly managed or destroyed, they are liabilities.

THE BUSINESS VALUE OF E-RECORDS MANAGEMENT

Companies need to be prepared to respond to an audit, federal investigation, lawsuit, or any other legal action against it. Types of lawsuits against companies include patent violations, product safety negligence, theft of intellectual property, breach of contract, wrongful termination, harassment, discrimination, and many more.

Nearly 90 percent of U.S. corporations become engaged in lawsuits; at any one time, the average $1 billion company in the United States faces 147 lawsuits (Kish, 2006). Each lawsuit will involve discovery, or the request for information (which almost always involves the request for e-mail and other electronic communications).

Discovery is the process of gathering information in preparation for trial, legal or regulatory investigation, or administrative action as required by law. When electronic information is involved, the process is called electronic discovery, or e-discovery. When a company receives an e-discovery request, the company must produce what is requested—or face charges of obstructing justice or being in contempt of court.

Several cases where a company incurred huge costs for not responding to e-discovery are the following:

- Failure to save e-mails resulted in a $2.75 million fine for Phillip Morris.
- Failure to respond to e-discovery requests cost Bank of America $10 million in fines.
- Failure to produce backup tapes and deleted e-mails resulted in a $29.3 million jury verdict against USB Warburg in what became a landmark case, *Zubulake v. UBS Warburg.*

ECM AND ERM GROWTH

At the 2008 EMC World Conference in Las Vegas, 9,000 attendees heard about the future of EMC for content management, virtualization, and Web 2.0 storage—indicating the importance of these technologies. Stored data is expected to reach 2 trillion gigabytes by 2011. Keeping corporate data safe will be an incredible challenge, as you will read in Chapter 5.

Unlike any other aspect of business, planning and managing electronic documents and records are elements of every business process. This situation has increased both the number of ECM and ERM vendors and the capabilities they provide. Vendors sell suites of products, including document management, collaboration, portals, and business intelligence.

Numerous major IT companies have become ECM and ERM vendors. ECM vendors include IBM (*ibm.com*), Oracle (*oracle.com*), and EMC (*emc.com*).

Major ERM vendors include Hummingbird (*hummingbird.com*), Iron Mountain (*ironmountain.com*), Oracle (*oracle.com*), and AccuTrac (*accutrac.com*). Visit *ironmountain.com/services/tours/records.asp* or *ironmountain.com/services/tours/dms.asp* to view videos of electronic records management and a records management center.

Review Questions

1. Define *ECM.*
2. What is the difference between a document and a record?
3. Why is ERM important to an organization?
4. Define discovery and e-discovery.
5. How does creating backups of electronic records differ from ERM?

Key Terms

assets *59*
bit *69*
business record *67*
byte *69*
client/server network *63*
dashboard *62*
data center *81*
data entity *63*
data management *61*
data mart *65*
data quality *65*
data synchronization *75*
data visualization *61*
data warehouse *59*
database *59*
database management system (DBMS) *74*
direct file organization *71*
dirty data *60*
discovery *84*
document management *68*
document management systems (DMS) *68*
electronic records management (ERM) *82*
enterprise content management (ECM) *82*
enterprise portal *63*
extract, transform, and load (ETL) *65*
field *69*
file *70*
indexed sequential access method (ISAM) *71*
master data entity *64*
master data management (MDM) *63*
master reference file *63*
online analytical processing (OLAP) systems *76*
online transaction processing (OLTP) systems *76*
operational data store *81*
primary key *70*
random file organization *71*
record *70*
secondary key *70*
sequential file organization *71*
text *57*
text mining *57*

Chapter Highlights and Insights

(Numbers refer to Learning Objectives)

1 Data is the foundation of any information system and needs to be managed throughout its useful life cycle, which converts data to useful information, knowledge, and a basis for decision support. Viewed from the basic profitability or net income model (profit = revenues − expenses), profit increases when employees learn from and use the data to increase revenues, reduce expenses, or both.

2 Managers and information workers may be constrained by data that cannot be trusted because it is incomplete, out of context, outdated, inaccurate, inaccessible, or so overwhelming that it requires too much time to analyze.

2 Data errors and inconsistencies lead to mistakes and lost opportunities, such as failed deliveries, invoicing blunders, and problems in synchronizing data from multiple locations.

2 Many factors that impact the quality of data must be recognized and controlled.

3 Programs that manage data and provide access to the database are called database management systems (DBMSs).

3 Data and documents are managed electronically. They are digitized, stored, and used in electronic management systems.

3 The benefits of using a DBMS include improved strategic use of corporate data, reduced complexity of the data environment, reduced data redundancy and enhanced data integrity, improved security, reduced data maintenance costs, and better access to data.

4 The logical view, or users' view, of data is meaningful to the user. Restricting access to data based on the users' job responsibilities increases data security.

4 The logical model is a detailed view of the data in which high-level entities are broken down into manageable data entities (e.g., customer data into product preferences, customer contact, shop locations, and product sales).

5 Data warehouses and data marts support the demanding data needs of decision makers. Relevant data is indexed and organized for easy access by end users.

5 Electronic document management, the automated control of documents, is a key to greater efficiency in handling documents in order to gain an edge on the competition.

5 How an organization manages its electronic records can directly affect its ability to compete intelligently, to comply with laws and regulations, to respond to litigation, and to recover from disaster. Regulatory requirements such as Sarbanes-Oxley, privacy requirements, and anti-fraud legislation have made managing information both a business priority and a legal obligation.

Questions for Discussion

1. What is the purpose of text mining?
2. Explain how having detailed real-time or near real-time data can improve productivity and decision quality.
3. Why does data and text management matter?
4. List three types of waste or damages that data errors can cause.
5. Explain the *principle of 90/90 data use*.
6. How does data visualization improve decision making?
7. Discuss the major drivers and benefits of data warehousing.
8. Why is master data management (MDM) important in companies with multiple data sources?
9. A data mart can substitute for a data warehouse or supplement it. Compare and discuss these options.
10. What ethical duties does the collection of data about customers impose on companies?
11. How are organizations using their data warehouses to improve consumer satisfaction and the company's profitability?
12. Relate document management to imaging systems.
13. Discuss the factors that make document management so valuable. What capabilities are particularly valuable?
14. Distinguish among operational databases, data warehouses, and data marts.

Exercises and Projects

1. Read *IT at Work 3.1,* "Data Errors Cost Billions of Dollars and Put Lives at Risk." Answer the questions at the end. Then visit the SAS Web site at *sas.com* and search for their data synchronization or data integration solution. List the key benefits of the SAS solution.
2. Interview a manager or other knowledge worker in a company you work for or to which you have access. Find the data problems the person has encountered and the measures he or she has taken to solve them.
3. Read *IT at Work 3.2,* "Finding Million-Dollar Donors in Three Minutes." Answer the questions at the end. Then visit the Business Objects Web site at *businessobjects.com* and search for "Xcelsius 2008 Demos and Sample Downloads." Click on one of the images of a dashboard or model to launch an interactive demo. Use the simulated controls in the demo to see *Xcelsius 2008 in action* (or visit *businessobjects. com/product/catalog/xcelsius/demos.asp*). Identify the model or dashboard whose interactive demo you viewed. Explain the benefits to decision makers of that dashboard or model.
4. Visit Analysis Factory at *analysisfactory.com*. Click to view the Interactive Business Solution Dashboards. Select one type of dashboard and explain its value or features.

5. Read *IT at Work 3.3,* "National Security Depends on Intelligence and Data Mining." Answer the questions at the end. Visit Oracle at *oracle.com* and do a search for Oracle Data Mining (ODM). Identify three functionalities of ODM.
6. At *teradatastudentnetwork.com*, read and answer the questions to the case "Harrah's High Payoff from Customer Information." Relate results from Harrah's to how other casinos use their customer data.
7. Go to *Teradata Magazine*, Volume 6, Number 2, and read "The Big Payoff." Then go to *teradatastudentnetwork.com* and read the case study "Harrah's High Payoff from Customer Information." What kind of payoff are they having from this investment in data warehousing?
8. At *teradatastudentnetwork.com*, read and answer the questions of the assignment entitled "Data Warehouse Failures." Choose one case and discuss the failure and the potential remedy.

Group Assignments and Projects

1. Prepare a report on the topic of data management and the intranet. Specifically, pay attention to the role of the data warehouse, the use of browsers for query, and data mining. Each group will visit one or two vendors' sites, read the white papers, and examine products (Oracle, Red Bricks, Brio, Siemens Mixdorf IS, NCR, SAS, and Information Advantage). Also, visit the Web site of the Data Warehouse Institute (*tdwi.org*).
2. Using data mining, it is possible not only to capture information that has been buried in distant courthouses but also to manipulate and cross-index it. This ability can benefit law enforcement but invade privacy. In 1996, Lexis-Nexis, the online information service, was accused of permitting access to sensitive information on individuals. The company argued that the firm was targeted unfairly because it provided only basic residential data for lawyers and law enforcement personnel. Should Lexis-Nexis be prohibited from allowing access to such information? Debate the issue.
3. Ocean Spray Cranberries, Inc. is a large cooperative of fruit growers and processors. Ocean Spray needed data to determine the effectiveness of its promotions and its advertising and to respond strategically to its competitors' promotions. The company also wanted to identify trends in consumer preferences for new products and to pinpoint marketing factors that might be causing changes in the selling levels of certain brands and markets.

   Ocean Spray buys marketing data from InfoScan (*us.infores.com*), a company that collects data using barcode scanners in a sample of 2,500 stores nationwide and from A.C. Nielsen. The data for each product includes sales volume, market share, distribution, price information, and information about promotions (sales, advertisements).

   The amount of data provided to Ocean Spray on a daily basis is overwhelming (about 100 to 1,000 times more data items than Ocean Spray used to collect on its own). All of the data is deposited in the corporate marketing data mart. To analyze this vast amount of data, the company developed a decision support system (DSS). To give end users easy access to the data, the company uses a data mining process called CoverStory, which summarizes information in accordance with user preferences. CoverStory interprets data processed by the DSS, identifies trends, discovers cause-and-effect relationships, presents hundreds of displays, and provides any information required by the decision makers. This system alerts managers to key problems and opportunities.
   a. Find information about Ocean Spray by entering Ocean Spray's Web site (*oceanspray.com*).
   b. Ocean Spray has said that it cannot run the business without the system. Why?
   c. What data from the data mart is used by the DSS?
   d. Enter *scanmar.nl* and click the Marketing Dashboard. How does the dashboard provide marketing and sales intelligence?

Internet Exercises

1. Conduct a survey on document management tools and applications.
2. Access the Web sites of one or two of the major data management vendors, such as Oracle, IBM, and Sybase, and trace the capabilities of their latest BI products.
3. Access the Web sites of one or two of the major data warehouse vendors, such as NCR or SAS; find how their products are related to the Web.
4. Access the Web site of the GartnerGroup (*gartnergroup .com*). Examine some of their research notes pertaining to marketing databases, data warehousing, and data management. Prepare a report regarding the state of the art.
5. Explore a Web site for multimedia database applications. Review some of the demonstrations, and prepare a concluding report.
6. *Enter microsoft.com/solutions/BI/customer/biwithinreach_ demo.asp* and see how BI is supported by Microsoft's tools. Write a report.
7. Visit *www-306.ibm.com/*. Find services related to dynamic warehouse and explain what it does.

BUSINESS CASE

Applebee's International Learns and Earns from Its Data

Over the past decades, businesses have invested heavily in IT infrastructures (e.g., ISs) to capture, store, analyze, and communicate data. However, the creation of ISs to manage and process data and the deployment of communication networks by themselves do not generate value, as measured by an increase in *profitability*. Viewed from the basic profitability or net income model (profit = revenues − expenses), profit increases when employees learn from and use the data to increase revenues, reduce expenses, or both. In this learn and earn model, managers learn—that is, gain insights—from their data to predict what actions will lead to the greatest increase in net earnings. *Net earnings* are also referred to as *net income* or the *bottom line*. The pursuit of earnings is the primary reason companies exist. Reducing uncertainty can improve the bottom line, as the examples in Table 3.5 show.

Applebee's International, Inc. (*applebees.com*), headquartered in Kansas, had faced these and other common business uncertainties and questions, but the company lacked the data infrastructure to answer them. Applebee's International develops, franchises, and operates restaurants under the Applebee's Neighborhood Grill & Bar brand, the largest casual dining enterprise in the world. As of 2008, there were nearly 2,000 Applebee's restaurants operating in 49 states and 17 countries, of which 510 were company owned. Despite its impressive size, however, Applebee's faced fierce competition.

To differentiate Applebee's from other restaurant chains and to build customer loyalty (defined as return visits), management wanted guests to experience a good time while having a great meal at attractive prices. To achieve their strategic objectives, management had to be able to forecast demand accurately and to become familiar with customers' experiences and regional food preferences. For example, knowing which new items to add to the menu based on past food preferences helps motivate return visits. However, identifying regional preferences, such as a strong demand for steaks in Texas but not in New England, by analyzing the relevant data was too time-consuming when it was done with the company's spreadsheet software.

The problem for many companies such as Applebee's International is that it is very difficult to bring together huge quantities of data located in different databases in a way that creates value. Without efficient processes for managing vast amounts of customer data and turning this data into usable knowledge, companies can miss critical opportunities to find insights hidden in the data.

Enterprise Data Warehousing Solution

Applebee's International implemented an **enterprise data warehouse (EDW)** from Teradata with data analysis capabilities that helped management acquire an accurate understanding of sales, demand, and costs. An EDW is a data repository whose data is analyzed and used throughout the organization to improve responsiveness and ultimately net earnings. Each day, Applebee's collects data concerning the previous day's sales from hundreds of point-of-sale (POS) systems located at every company-owned restaurant. The company then organizes this data to report every ticket item sold in 15-minute intervals. By reducing the amount of time required to collect POS data from two weeks to one day, the EDW has enabled management to respond quickly to guests' needs and to changes in guests' preferences. With greater knowledge about their customers, the company is better equipped to market and provide services that attract customers and build loyalty.

Business Improvements

Applebee's management gained clearer business insight by collecting and analyzing detailed data in near real time using an enterprise data warehouse. Regional managers can now select the best menu offerings and operate more efficiently. The company uses detailed sales data and data from customer satisfaction surveys to identify regional preferences, predict product demand, and build financial models that indicate which products are strong performers on the menu and which are not. By linking customer satisfaction ratings to specific menu items, Applebee's can determine which items are doing well, which ones taste good, and which food arrangements on the plates look most appetizing.

With detailed, near real-time data, Applebee's International improved its customers' experience, satisfaction, and loyalty—and increased the company's earnings. For the third quarter of 2007, total system-wide sales increased by 3.9 percent over the prior year, and Applebee's opened 16 new restaurants.

Lessons Learned from this Case

This case illustrates the importance of timely and detailed data collection, data analysis, and execution based on insights from

TABLE 3.5 How Data Can Reduce Uncertainty and Improve Accuracy and Performance

Business Uncertainty	Business Impact and Value
What will be monthly demand for Product X over each of the next three months?	Knowing demand for Product X means knowing how much to order. Sales quantity and sales revenues are maximized because there are no inventory shortages or lost sales. Expenses are minimized because there is no unsold inventory.
Which marketing promotions for Product Y are customers most likely to respond to?	Knowing which marketing promotion will get the highest response rate maximizes sales revenues while avoiding the huge expense of a useless promotion.

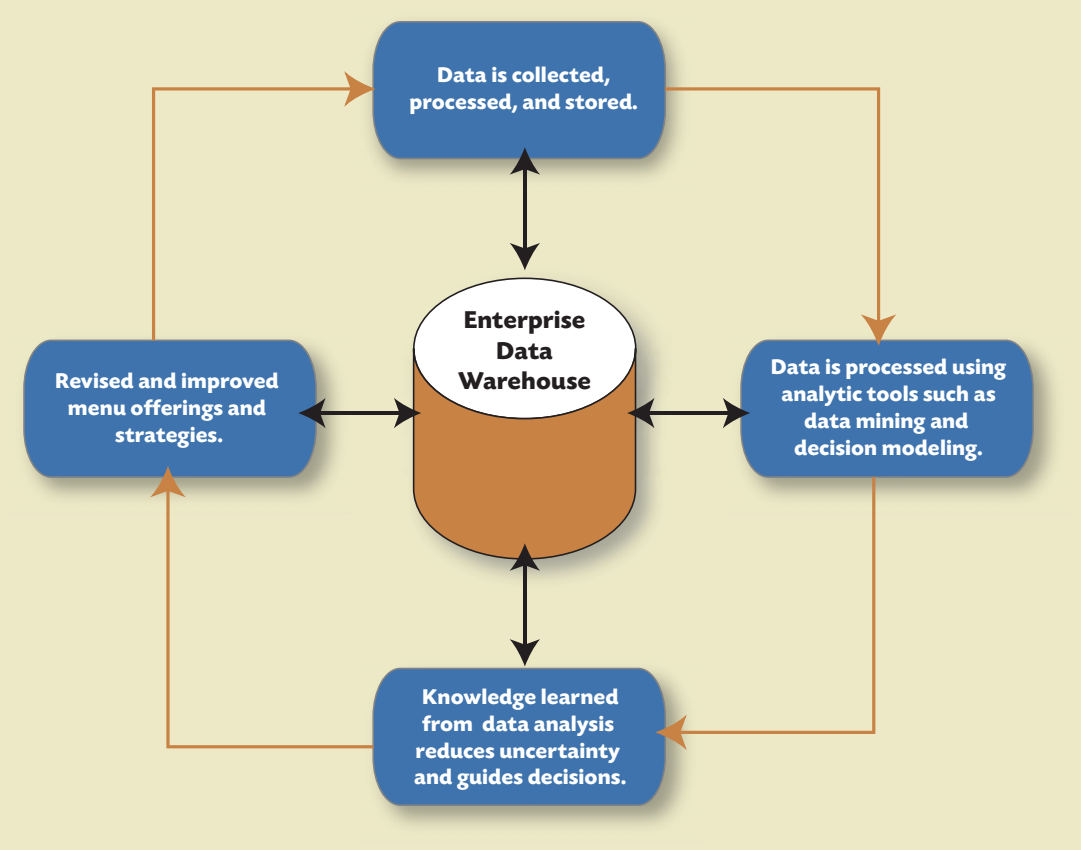

Figure 3.14 Applebee's enterprise data warehouse and feedback loop.

that data. It demonstrates that it is necessary to collect vast amounts of data, organize and store it properly in one place, analyze it, and then use the results of the analysis to make better marketing and strategic decisions. Companies seldom fail for lack of talent or strategic vision. Rather, they fail because of poor execution.

The case also illustrates data stages, as shown in Figure 3.14. First, data is collected, processed, and stored in a data warehouse. It is then processed by analytical tools such as data mining and decision modeling. Knowledge acquired from this data analysis directs promotional and other decisions. Finally, by continuously collecting and analyzing fresh data, management can receive feedback regarding the success of management strategies.

Sources: Compiled from *Applebees.com*, Business Wire (2007), and Teradata (2007).

Questions

1. Why is learning important to managers?
2. How does learning influence net earnings?
3. What is the value of the feedback loop at Applebee's?
4. How necessary is near real-time data?
5. Is it easier for IT to support planning or execution? Why?

PUBLIC SECTOR CASE

British Police Invest in Mobile IT to Improve Performance and Cut Costs

In England and Wales, the National Policing Improvement Agency (NPIA) is responsible for bringing high-tech equipment to the police service. NPIA published its *Science and Innovation in the Police Service 2010–2013*, a strategy for police service for years 2010–2013. Its three-year strategy includes a number of new ITs for police, including mobile fingerprinting and checking, wearable video devices, and digital forensics.

Mobile Biometric Devices Cost Savings

NPIA signed a contract with Cogent Systems, a biometrics firm, for mobile devices to check prints against those on the National Fingerprint Database at a cost of £9 million over three years. Deputy chief constable Peter Goodman, who leads mobile identification for the Association of Chief Police Officers (ACPO), said the MobileID devices would provide

cost savings equivalent to releasing some 360 officers back to front-line policing each year. He stated: "At a time when funding is likely to be constrained or reduced, technical innovation has promise in saving the police service time and money as well as in aiding bringing criminals to justice more quickly."

MobileID is a device about the size of a BlackBerry that lets police officers take suspects' fingerprints on the street and check the prints against the National Fingerprint Database without having to return to the police station. The process takes about two minutes. MobileID is part of a larger project called Mobile Identification at Scene (Midas). The devices work by scanning both of the suspect's index fingers and communicating over an encrypted wireless link with the database.

The NPIA will also deliver the first elements of the *Police National Database* and make images of suspects available on officers' mobile devices. In 2012, the agency plans to deploy noncooperative facial recognition systems.

Being Technologically Innovative and Responsive to the Public

NPIA chief executive Peter Neyroud said in a statement: "By applying modern science on the front line, police officers are detecting criminals faster, staying on the beat for longer and making decisions based on better evidence about what works." He also explained that: "Identification is crucial to police investigations, and giving officers the ability to do this on-the-spot within minutes is giving them more time to spend working in their communities, helping to fight crime, bringing more offenders to justice and better protecting the public.

An overall goal is to help ensure that the police service in England and Wales continues to harness science effectively and remains one of the most innovative of its kind in the world. The NPIA recognizes the need to do more than improve the *toolkit* available to police officers. The success of its strategy requires putting public confidence first, for example, by meeting the public's concerns about personal privacy.

Sources: Compiled from NPIA (2010), Kable (2010), and Thurston (2010).

Questions

1. What are some of the ways the NPIA has cost-justified significant investments in innovative IT for police service?
2. How will the new ITs improve policing services in England and Wales?
3. In your opinion, why might the success of NPIA's strategy require putting public confidence first—for example, by meeting the public's concerns about personal privacy—rather than putting public safety first?
4. What are some potentials risks to privacy that MobileID might cause? Does encryption eliminate those risks?
5. Download the NPIA's publication at the textbook's Web site or from *npia.police.uk/en/docs/science_and_innovation.pdf*. What are the primary objectives of its three-year strategy? What ITs are needed to meet those objectives?

ANALYSIS USING SPREADSHEETS

Calculating the Cost of Poor Document Management

Spring Street Company (SSC, a fictitious company) faced rising costs not only from sky-high energy prices, but also from what it considered "hidden costs" associated with its paper-intensive processes. The employees jokingly predicted that if the windows in the offices blew open on a very windy day, there would be total chaos as the papers started flying. The financial implications were that, if such a disaster occurred, the business would grind to a halt.

The company's accountant, Sam Spring, decided to calculate the costs of the paper-driven processes to identify their impact on the bottom line. He recognized that several employees spent most of their day filing or retrieving documents. In addition, there were the monthly costs to warehouse old paper records. Sam observed and measured the activities related to the handling of printed reports and paper files. His average estimates are as follows:

- It takes an employee five minutes to walk to the records room, locate a file, act on it, refile it, and return to his or her desk.
- Employees need to locate a file, act on it, and so on five times per day.
- There are 12 full-time employees who perform these functions.
- Once per day a document gets "lost" (destroyed, misplaced, or covered with massive coffee stains) and must be re-created. The total cost of replacing each lost document is $220.
- Warehousing costs as of the present time for the current volume of stored documents are $75 per month.

Sam would prefer a system that lets employees find and work with business documents without leaving their desks. He's most concerned about the human resources and accounting departments. These personnel are traditionally heavy users of paper files and would greatly benefit from a modern document management system. At the same time, however, Sam is also risk averse. He would rather invest in solutions that would reduce the risk of higher costs in the future. He recognizes that the U.S. Patriot Act's requirements that organizations provide immediate government access to records apply to SSC. He has read that manufacturing and government organizations rely on efficient document management to meet these broader regulatory imperatives. Finally, Sam wants to implement a disaster recovery system.

Your Mission

Prepare a report that provides Sam with the data and information he needs to select and implement a cost-effective alternative to the company's costly paper-intensive approach to managing documents. You will need to conduct research to provide data to prepare this report. Your report should include the following information:

1. Explain the similarities and differences between document imaging systems and document management systems (DMS). List the benefits and the basic hardware and software requirements for each system. Put this information into a table to help Sam readily understand the comparison.
2. Discuss why a DMS transforms the way a business operates. How should SSC prepare for a DMS if it decides to implement one?
3. Collect estimates for the costs of buying or implementing a DMS at SSC.
4. Using the data collected by Sam, create a spreadsheet that calculates the costs of handling paper at SSC based on hourly rates per employee of $16, $22, and $28. Add the cost of lost documents to this. Then, add the costs of warehousing the paper, which increases by 10% every month due to increases in volume. Present the results, showing both monthly totals and a yearly total. Prepare graphs as visualization tools so that Sam can easily perceive the projected growth in warehousing costs over the next three years. Download the spreadsheet to help you get started from the textbook's Web site.
5. Identify at least one additional cost factor (other than better security) that might be reduced or eliminated with the DMS.
6. How can DMS also serve as a disaster recovery system in case of fire, flood, or break-in?
7. Submit your recommendation for a DMS solution. Identify two vendors in your recommendation.

Resources on the Book's Web Site

More resources and study tools are located on the Student Web Site and on WileyPlus. You'll find additional chapter materials and useful Web links. In addition, self-quizzes that provide individualized feedback are available for each chapter.

References

Barlow, R. D., "Sync or Swim: Who Should Blink First and Why?" *Healthcare Purchasing News*, April 2007. *hpnonline.com/inside/2007-04/0704-DataSynch1.html*.

Business Wire, "Applebee's International Reports Third Quarter 2007 Results," October 29, 2007.

BusinessWire, "Wendy's Selects Clarabridge's Text Analytics Solution to Enhance Customer Feedback Program," March 22, 2010.

Chisholm, P., "Synchronizing the Supply Chain," *Military Medical Technology*, 12(2), March 7, 2008. *military-medical-technology.com/article.cfm?DocID=1218*

Clarabridge, Press Release, March 22, 2010. *clarabridge.com/*

Henschen, D., "Restaurant Chain Will Analyze Customer Comments to Quickly Spot Product and Service Problems," *Intelligent Enterprise,* March 2010. *intelligententerprise.com/showArticle.jhtml?articleID=224000398*

Kable, B., "Police to Be Issued with Mobile Fingerprinting Devices," *ZDNet UK*, March 5, 2010. *zdnet.co.uk/news/security-management/2010/03/05/police-to-be-issued-with-mobile-fingerprinting-devices-40072975/*

Kish, L., "Most U.S. Companies Engaged in Lawsuits," *E-Discovery Advisor*, May 2006.

Levine, L., "In Sync: Getting the Supply Chain Act Together," *Healthcare Purchasing News*, April 2007. *hpnonline.com/inside/2007-04/0704-DataSynch2.html*.

Loshin, D., "Issues and Opportunities in Data Quality Management Coordination," *DM Review*, April 2004.

NPIA (National Improvement Policing Agency), "Science and Innovation in the Police Service, 2010-2013." 2010. *npia.police.uk/en/docs/science_and_innovation.pdf*.

Silva, V., "Focus on Application Deployments Urgent: Informatica," *NetworkWorld*, April 8, 2010. *networkworld.com/news/2010/040810-focus-on-application-deployments-urgent.html*

Snow, C., "Embrace the Role and Value of Master Data Management," *Manufacturing Business Technology*, 26(2), February 2008.

Teradata, "Applebee's International," Video, 2007. *Teradata.com*

Thurston, T., "Police to Get Mobile Fingerprint-Checking Tech," *ZDNet UK*, March 25, 2010. *zdnet.co.uk/news/mobile-working/2010/03/25/police-to-get-mobile-fingerprint-checking-tech-40088449/*

Volonino, L., R. Anzaldua, and J. Godwin, *Computer Forensics: Principles and Practice*, Chapter 10. Upper Saddle River. NJ: Prentice-Hall, 2007.

Wailgum, T., "Eight of the Worst Spreadsheet Blunders," *CIO.com*, August 17, 2007. *cio.com/article/131500/Eight_of_the_Worst_Spreadsheet_Blunders*.

Whitehouse.gov, 2010.

Worthen, B., "IT versus Terror," *CIO*, August 1, 2006.