

BUSINESS INTELLIGENCE: DATA WAREHOUSING, DATA ACQUISITION, DATA MINING, BUSINESS ANALYTICS, AND VISUALIZATION

LEARNING OBJECTIVES

- Describe issues in data collection, problems, and quality .
- Describe the characteristics and organization of database management systems.
- Explain the importance and use of a data warehouse and data mart.
- Describe business intelligence/business analytics and their importance to organizations .
- Describe how online analytical processing (OLAP), data mining, data visualization, multidimensionality, and real-time analytics can improve decision-making .
- Explain how the Web impacts database technologies and methods, and vice versa .
- Describe how database technologies and methods as part of business intelligence/business analytics improve decision-making .
- Describe Web intelligence/Web analytics and their importance to organizations.

Many organizations have amassed vast amounts of data that employees use to unlock valuable secrets to enable the organization to compete successfully. Some organizations do this extremely well, but others are quite ineffective. To use analytic tools to improve organizational decision-making, a foundational data architecture and enterprise architecture must be in place to facilitate effective decision analysis. Enabling decision analysis through access to all relevant information is known as *business intelligence*. Business intelligence includes data warehousing, online analytical processing, data mining, and visualization and multidimensionality. The outline of this chapter is as follows:

- 5.1 Opening Vignette: Information Sharing a Principal Component of the National Strategy for Homeland Security 5.2
- The Nature and Sources of Data
- 5.3 Data Collection, Problems, and Quality
- 5.4 The Web/Internet and Commercial Database Services
- 5.5 Database Management Systems in Decision-Support Systems/Business Intelligence
- 5.6 Database Organization and Structures 5.7
- Data Warehousing
- 5.8 Data Marts

5.9 Business Intelligence/Business Analytics 5.10

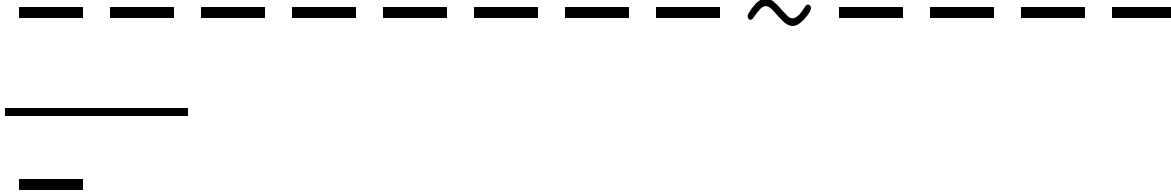
Online Analytical Processing (OLAP) 5.11 Data

Mining

5.U Data Visualization, Multidimensionality, and Real-Time Analytics

5.13 Geographic Information Systems

5.14 Business Intelligence and the Web: Web Intelligence/Web Analytics



5.1 OPENING VIGNETTE: INFORMATION SHARING A PRINCIPAL COMPONENT OF THE NATIONAL STRATEGY FOR HOMELAND SECURITY

Datawarehouses provide a strategic data architecture to enable decision support analysis. Data warehousing enables data mining, the ability to automatically synthesize vast amounts of information in order to discover *hidden truths* within the data. Data portals have emerged as the next generation in Web-enabled data warehouses. One of the most significant data portals has been developed in direct response to the terrorist attacks on the United States on September 11, 2001.

The National Strategy for Homeland Security of the United States includes a National Vision for the sharing of information related to the detection of terrorist activities. It states,

We will build a national environment that enables the sharing of essential homeland security information. We must build a *system of systems* that can provide the right information to the right people at all times. Information will be shared "horizontally" across each level of government and "vertically" among federal, state and local governments, private industry and citizens. With the proper use of people, processes, and technology, homeland security officials throughout the United States can have complete and common awareness of threats and vulnerabilities as well as knowledge of the personnel and resources available to address these threats. Officials will receive the information they need so they can anticipate threats and respond rapidly and effectively.

The goal of the project is to create a workable model for integrating knowledge that resides across many disparate data sources, while ensuring that privacy and civil liberties are adequately safeguarded. The five major initiatives that are identified within the strategy include:

1. To integrate information sharing across the federal government
2. To extend the integration of information sharing across state and local governments, private industry, and citizens
3. To adopt common metadata standards of electronic information relevant to homeland security
4. To improve public safety communication
5. To ensure reliable public health information.

Modified from the National Strategy for Homeland Security Web site,
www.whitehouse.gov/homeland/book/index.html.

These goals can only be accomplished if there is a means to facilitate the sharing of information among numerous agencies that currently maintain independent data silos. Border security alone engages eleven agencies. For the entire data warehouse project, approximately 80 percent of the architecture will be in place in 18 months, while the complete implementation will phase in over three to five years. Ultimately the data warehouse will lead to increased security for the United States. It will be a model for how all countries can interact to protect their borders and ensure the safety of their citizenry. This ambitious project is not without challenges. For example, data will need to be mined from immigration records, treasury records (dealing with the exchange of large sums of money), and FBI (criminal) records. The data exist in different formats and data types; a major effort is underway to establish a means to link and search through these data to identify potential threats and crimes.

•. QUESTIONS FOR THE OPENING VIGNETTE

1. Identify the challenges faced by the Office of Homeland Security in integrating disparate databases.
2. Identify the sources of information that will be required to make the information in this data portal useful.
3. What are the expected benefits?
4. Identify decisions supported by this data portal.
5. What decision support tools and techniques can be used to identify potential terrorist activities?
6. What would you recommend to the Office of Homeland Security to improve the capabilities of this data portal?



5.2 THE NATURE AND SOURCES OF DATA

In order to understand a situation, a decision-maker needs data, information, and knowledge. These must be integrated and organized in a manner that makes them useful. The decision-maker must be able to apply analysis tools (online analytical processing (OLAP), data mining, etc.) so that the data, information, and knowledge can be utilized to full benefit. These analysis tools fall under the general heading of business intelligence (BI) and business analytics (BA) (see Chapters 3 and 4). New tools allow decision-makers and analysts to readily identify relationships among data items that enable understanding and provide a competitive advantage. Foreexample, a customer-relationship (resource) management (CRM) system allows managers to better understand their customers. They can then determine a likely candidate for a particular product or service at a specific price (see Chapter 8). Marketing efforts are improved and sales are maximized. All enterprise information systems (e.g., CRM, executive information systems, content-management systems, revenue management systems, enterprise resource planning/enterprise resource management systems, supply chain management systems, knowledge management systems) utilize database management systems, data warehouses, OLAP, and data mining as their foundation (see Chapters 8 and 9). These business intelligence/business analytic (and Web intelligence/Web analytic) tools enable the modern enterprise to compete successfully. In the right hands, these tools provide great decision-makers with great capabilities. For example, see Case Application 5.2, which indicates how a firm developed and then utilized databases in an extremely competitive manner.

The Opening Vignette illustrates what can go wrong in the extreme when you do not gather data to track the activities of individuals and organizations that impact your organization (in a business environment, these are customers, potential customers and the competition). The critical issue for the U.S. Department of Homeland Security is to gather and analyze data from disparate sources. These data must be integrated in a data warehouse and analyzed automatically via data-mining tools or by analysts using OLAP tools. Of course, abuses can occur in the process of collecting and utilizing such a massive amount of data (see DSS in Focus 5.1).

The impact of tracking data and then exploiting them for competitive advantage can be enormous. Entire industries, such as travel, banking, and all successful e-commerce ventures, rely totally on their data and information content to flourish. Experian Automotive has developed a business opportunity from modern database, extraction and integration tools (see DSS in Action 5.2).

Songini (2002) provides an excellent description of databases, data, information, metadata, OLAP, repository, and data mining. Major database vendors include IBM, Oracle, Informix, Microsoft, and Sybase. Database vendors are reviewed on a regular basis by the trade press. For example, see Whiting (2000) and the "Annual Product Review" issue of *DM Review* (www.dmreview.com) every July.

All decision-support systems use data, information, and/or knowledge. These three terms are sometimes used interchangeably and may have several definitions. A common way of looking at them is as follows:

- **Data.** Items about things, events, activities, and transactions are recorded, classified, and stored but are not organized to convey any specific meaning. Data items can be numeric, alphanumeric, figures, sounds, or images.
- **Information.** Data that have been organized in a manner that gives them meaning for the recipient. They confirm something the recipient knows, or may have

DSS IN FOCUS 5.1

HOMELAND SECURITY PRIVACY AND COST CONCERNS~

The U.S. government plans to apply analytic technologies on a global scale in the war on terrorism, but will they prove an effective weapon? In the first year and a half after September 11, 2001, supermarket chains, home improvement stores, and others voluntarily handed over massive amounts of customer records to federal law enforcement agencies, almost always in violation of their stated privacy policies. Many others responded to court orders for information, as required by law. The government has a right to gather corporate data under legislation passed after September 11, 2001.

The FBI now mines enormous amounts of data looking for activity that could indicate a terrorist plot or crime. Transaction data are where law-enforcement agencies expect to find results. American businesses are stuck in the middle. Some have to create special systems to generate the data required by law-enforcement agencies. An average-size company will spend an average of \$5 million for a system. On the other hand, not comply-

ing can cost more. Western Union was fined \$8 million in December 2002 for not complying properly.

Privacy issues abound. Since the government is acquiring personal data to detect suspicious patterns of activity, there is the prospect of abuse and illegal use of the data. There may be significant privacy costs involved. There are major problems with violating people's freedoms and rights. There is a need for an oversight organization to "watch the watchers." The DHS must not mindlessly acquire data. It should only acquire pertinent data and information that can be mined to identify patterns that potentially could lead to stopping terrorist activities.

Source: Partly adapted from John Foley, "Data Debate," *Information Week*, May 19, 2003, pp. 22-24; S. Grimes, "Look Before You Leap," *Intelligent Enterprise*, June 2003; Ben Worthen, "What to Do When Uncle Sam Wants Your Data," *CIO*, April 15, 2003, pp. 56-66.

DSS IN ACTION 5.2**DATABASE TOOLS OPEN UP NEW REVENUE OPPORTUNITIES FOR EXPERIAN AUTOMOTIVE**

Experian Automotive has developed new business opportunities from data tools that manage, extract, and integrate. Experian has developed a system with a huge database (the world's 10th largest) to track automobile sales data. The acquired data are external and come from public records of automobile sales. Experian draws on these data to provide the ownership history of any vehicle bought or sold in the United States for an

inexpensive fee per query via the Web. There is a massive market for this service, especially from car dealerships. Experian also focuses on automobile parts companies to identify recalls and consider how to target automobile parts sales.

Source: Adapted from Pimm Fox, "Extracting Dollars from Data," *ComputerWorld*, April 15, 2002, p. 42.

"surprise" value by revealing something not known. An MSS application processes data items so that the results are meaningful for an intended action or decision.

- **Knowledge.** Knowledge consists of data items and/or information organized and processed to convey understanding, experience, accumulated learning, and expertise that are applicable to a current problem or activity. Knowledge can be the application of data and information in making a decision. (See Chapters 9 and 10.)

MSS data can include documents, pictures, maps, sound, video, and animation.

These data can be stored and organized in different ways before and after use. They also include concepts, thoughts, and opinions. Data can be raw or summarized. Many MSS applications use summary or extracted data that come from three primary sources: internal, external, and personal.

INTERNAL DATA

Internal data are stored in one or more places. These data are about people, products, services, and processes. For example, data about employees and their pay are usually stored in the corporate database. Data about equipment and machinery can be stored in the maintenance department database. Sales data can be stored in several places: aggregate sales data in the corporate database, and details at each region's database. An MSS can use raw data as well as processed data (e.g., reports and summaries). Internal data are available via an organization's intranet or other internal network.

EXTERNAL DATA

There are many sources of external data. They range from commercial databases to data collected by sensors and satellites. Data are available on CDs and DVDs, on the Internet, as films and photographs, and as music or voices. Government reports and files are a major source of external data, most of which are available on the Web today (e.g., see www.ftc.gov, the U.S. Federal Trade Commission). External data may also be available by using GIS (geographic information systems, see Section 5.13), from federal census bureaus, and other demographic sources that gather data either directly from customers or from data suppliers. Chambers of commerce, local banks, research institutions, and the like, flood the environment with data and information, resulting in *information overload* for the MSS user. Data can come from around the globe. Most external data are irrelevant to a specific MSS. Yet many external data must be monitored and captured to ensure that important items are not overlooked. Using intelli-

gent scanning and interpretation agents may alleviate this problem. For tips on how to manage external data, see Collett (2002).

PERSONAL DATA AND KNOWLEDGE

MSS users and other corporate employees have expertise and knowledge that can be stored for future use. These include subjective estimates of sales, opinions about what competitors are likely to do, and interpretations of news articles. What people really know and methodologies to capture, manage, and distribute it are the subject of *knowledge management* (Chapter 9).

5.3 DATA COLLECTION, PROBLEMS, AND QUALITY

The need to extract data from many internal and external sources complicates the task of MSS building. Sometimes it is necessary to collect raw data in the field. In other cases, it is necessary to elicit data from people or to find it on the Internet. Regardless of how they are collected, data must be validated and filtered. A classic expression that sums up the situation is "Garbage in, garbage out" (GIGO). Therefore, *data quality* (DQ) is an extremely important issue.

METHODS FOR COLLECTING RAW DATA

Raw data can be collected manually or by instruments and sensors. Representative data collection methods are time studies, surveys (using questionnaires), observations (e.g., using video cameras; see Exercise 9), and soliciting information from experts (e.g., using interviews; see Chapter 11). In addition, sensors and scanners are increasingly being used in data acquisition. Probably the most reliable method of data collection is from point-of-purchase inventory control. When you buy something, the register records sales information with your personal information collected from your credit card. This has enabled Wal-Mart, Sears, and other retailers to build complete, massive (petabyte-sized) data warehouses in which they collect and store business intelligence data about their customers. This information is then used to identify customer buying patterns to manage local store inventory and identify new merchandising opportunities. It also helps the retail organization manage its suppliers.

Ewalt (2003) describes how PDAs are utilized to collect and utilize data in the field. Logistics companies have been using PDAs for some time. Menlo Worldwide Forwarding, a global freight company, recently equipped over 800 drivers with PDAs. Radio links are used to dispatch drivers to pick up packages. The driver scans a bar code label on the package into the PDA, which then beams tracking data back to the home office.

The need for reliable, accurate data for any MSS is universally accepted. However, in real life, developers and users face ill-structured problems in "noisy" and difficult environments. There is a wide variety of hardware and software for data storage, communication, and presentation, but much less effort has gone into developing methods for MSS data capture in less tractable decision environments. Inadequate methods for dealing with these problems may limit the effectiveness of even sophisticated technologies in MSS development and use. Some methods involve physically capturing data via bar codes or by RFID (radio-frequency identification tag) technology. An RFID electronic button sends an identification signal with some data (several kilobytes when these devices were new) directly to a nearby receiver. A packing crate, or

even an individual consumer product, can readily be identified. In the early 2000s, manufacturers, airlines, and retailers were experimenting with utilizing RHD devices for security, speeding up processing in receiving, and customer checkout. Wal-Mart Stores Inc. announced in June 2003 that by January 2005 its 100 key suppliers must use RFID to track pallets of goods through its supply chain. See DSS in Action 5.3. Swatch incorporates the device into select watch models so that ski lift passes at ski resorts are automatically encoded into it. The resort can readily identify the types of slopes you like to ski and share the information with its other properties.

DSS IN ACTION 5.3

RFID TAGS HELP AUTOMATE DATA COLLECTION AND USE



In June 2003, Wal-Mart Stores Inc. announced that by 2005 its 100 key suppliers must use RFID to track pallets of goods through its supply chain. Wal-Mart considers this much more than a company-specific effort and urged all retailers and suppliers to embrace RFID and related standards. Wal-Mart's initiative should result in deploying about 1 billion RFID tags to track and identify items in the individual crates and pallets. Wal-Mart will first concentrate on using the technology to improve inventory management in its supply chain. Wal-Mart's decision to deploy the technology should legitimize it and push it into the mainstream. The Wal-Mart deadline will definitely speed adoption by the industry.

The RFID unit price must be 5 cents (United States) or less for the Wal-Mart initiative to be costeffective. In mid-2003, the RFID tags cost between 30 to 50 cents. Based on a 5 cent per tag cost, the outlay for the tags alone will total \$50 million. In 2003, the readers sold for \$1000 or more.

Wal-Mart is not the only retailer moving toward RFID. Marks & Spencer PLC, one of Britain's largest retailers, utilizes RFID technology in its food supply-chain operations. Each of 3.5 million plastic trays used to ship products has an RFID tag on it. Procter & Gamble Co. experimented with RFID for more than six months in 2003, running tests with several retailers.

In 2003, Delta Airlines started tests of using RFID to identify baggage while bags are loaded and unloaded on airport tarmacs. Delta will load data into the tags as the bar code is printed. Testing is critical because of potential interference from other airport wireless systems. Delta expected to see a higher level of accuracy than from the existing bar-code system. Even so, Delta delivers 99 percent of the 100 million or so bags it handles each year. But it still costs Delta a small fortune to find missing bags.

RFID tags have been utilized to track the movement of pharmaceuticals through Europe's "gray" (i.e., semi-legal) markets. At the time, medicines were generally much less expensive in southern Europe than in northern Europe, so unscrupulous wholesalers traveled south to buy them for resale in the north. RFID tags were installed inside the labels. When a vendor representative visited the dishonest wholesalers, he was able to identify the source of their stock once he got within 3 meters of the containers. All contracts with these wholesalers were immediately cancelled.

Others possible uses of RFID include embedding them in badges so that doors will automatically unlock for an authorized person, and providing access to movies and other events (through a watch-embedded or card-embedded RFID tag). They could be embedded in automobiles for automatic toll charges (as in the City of London, see Exercise 9), used in automobiles to store an entire maintenance and repair record (this is currently done for industrial fork lifts), or even under the skin for identification (by ATMs, museums, transit systems, admission to any facility, or law enforcement officials). Some pet owners have had these tags surgically embedded under their pet's skin for identification if lost or stolen. Eventually, consumer product packages and suitcases may be manufactured to contain RFID tags so that when you walk out of a store, readers detect what you have selected, and your account will automatically be charged for what you have, through an RFID tag either under your skin or in a credit card.

Source: Partly adapted from Bob Brewin, "Delta to Test RFID Tags on Luggage," *ComputerWorld*, Vol. 37, No. 25, June 23, 2003, p. 7; Chris Murphy and Mary Hayes, "Tag Line," *Information Week*, June 15, 2003, pp. 18-20; Jaikumar Vijayan and Bob Brewin, "Wal-Mart Backs RFID Technology," *ComputerWorld*, Vol. 37, No. 24, June 16, 2003, pp. 1, 14.

50.

Even biometric (scanning) devices are used to collect real-world data. Biometric systems detect various physical and behavioral features of individuals and assess them to authenticate the identities of visitors and immigrants entering the United States. Databases and data mining methods are also used. Some \$400 million was spent on biometrics for U.S. border control in 2003. See Verton (2003).

DATA PROBLEMS

All computer-based systems depend on data. The quality and integrity of the data are critical if the MSS is to avoid the GIGO syndrome. MSS depend on data because compiled data that make up information and knowledge are at the heart of any decisionmaking system.

The major DSS data problems are summarized in Table 5.1 along with some possible solutions. Data must be available to the system or the system must include a data acquisition subsystem. Data issues should be considered in the planning stage of system development. If too many problems are anticipated, the costs of solving them can be estimated, If they are excessive, the MSS project should not be undertaken or should be put on hold until costs and problems decrease.

DATA QUALITY

Data quality (DQ) is an extremely important issue because quality determines the usefulness of data as well as the quality of the decisions based on them. Data in organizational databases are frequently found to be inaccurate, incomplete, or ambiguous. The

TABLE 5.1 Data Problems		
Problem	Typical Cause	Possible Solutions
Data are not correct. Develop a systematic way to enter data.	Data were generated carelessly. Raw data were entered inaccurately. Data were tampered with.	Automate data entry. Introduce quality controls on data generation. Establish appropriate security programs. Modify the system for generating data.
Data are not timely. Use the Web to get fresh data. Develop a system for rescaling or recombining properly indexed data.	The method for generating data is not rapid enough to meet the need for data.	
Data are not measured or indexed properly.	Raw data are gathered inconsistently with the purposes of the analysis. Use of complex models.	Use a data warehouse. Use appropriate search engines. Develop simpler or more highly aggregated models. Predict what data may be needed in the future. Use a data warehouse. Generate new data or estimate them.
Needed data simply do not exist.	No one ever stored data needed in 1999. Required data never existed.	

Source: Based on Alter (1980), p.130. Alter (1980). Decision Support Systems: Current Practices and Continuing Challenges. Reading, MA: Addison-Wesley.

economic and social damage from poor-quality data costs billions of dollars (Redman, 1998),

The Data Warehousing Institute (TDWI) estimated in 2001 that poor-quality customer data caused U.S. businesses \$611 billion a year in postage, printing, and the staff overhead to deal with the mass of erroneous communications and marketing (from a TDWI report: Wayne Erickson, "Data Quality and the Bottom Line" www.dw-institute.com/dqreport/). Frighteningly, the real cost of poor-quality data is much higher. Organizations can frustrate and alienate loyal customers by incor-

rectly addressing letters or failing to recognize them when they call, or visit a store or Web site. Once a company loses its loyal customers, it loses its base of sales and referrals, as well as future revenue potential. See Eckerson (2002a). Some typical costs include those of rework, lost customers, late reporting, wrong decisions, wasted project activities, slow response to new needs (missed opportunities), and delays in implementing large projects that depend on existing databases (Olson, 2003a, 2003b).

Data quality is one of those topics that everyone knows is important but tends to neglect. Data quality often generates little enthusiasm and is typically viewed as a maintenance function. Firms have clearly been willing to accept poor data quality. Companies can even survive and flourish with poor data quality. It is not considered a life-and-death issue, but sometimes it can be. Data inaccuracies can be extremely costly (see Olson, 2003a, 2003b). Even so, most firms manage data quality in a casual manner (Eckerson, 2002a). According to Hatcher (2003), data quality is a major problem in data warehouse development and business intelligence/business analytics utilization. Data quality can delay the implementation of a warehouse or a data mart six months or more. Inaccurate data stored in a data warehouse and then reported to someone will instantly kill a user's trust in the new system.

A recent TDWI (The Data Warehouse Institute) survey uncovered the sources of dirty data. These are shown in Table 5.2. Unsurprisingly, respondents to TDWI's survey cite data-entry errors by employees as the primary cause of dirty data.

Data quality was often overlooked in the early days of data warehousing. Many of the original decisions about data quality now need to be revisited by data warehouse practitioners in order to keep pace with the demands of enterprise decision-making (see Canter, 2002). For an example of an organization that suffered because of data quality, see DSS in Action SA.

Strong et al. (1997) conducted extensive research on data quality problems and divided them into the following four categories and dimensions:

51.

TABLE 5.2 Source of Data Quality Problems

<i>Source of Data Quality Problem</i>	<i>Percent Response</i>
Data entry by employees	76
Changes to source systems;	5
Data migration or conversion projects	3
Mixed expectations by users	48
External data	4
Systems' errors	6
Data entry by customers	34
Other	2
<i>Source: Adapted from Wayne Eckerson, "Data Quality and the Bottom Line," Application Development Trends, May 2002, pp. 24--30.</i>	
	2
	5
	1
	2

DSS IN ACTION 5.4

DATA QUALITY IS THE CULPRIT IN
MONTANA PRISONS

Data quality held the Montana Department of Corrections prisoner for years. As IT systems aged, data entry errors in reports built up. Required forms that were submitted to state and federal authorities could not pass a lie detector test. Even though the department's IS group spent countless hours of manual effort in attempting to maintain some level of reporting integrity, overall confidence in data quality was low. The issue came to breakout proportions when, in 1997, the department lost a \$1 million federal grant. The guilty party was its information systems, which lacked business rules and a data dictionary. The systems could not accurately forecast how many of any type of offender would be incarcerated. Fortunately, no offenders were lost in the data shuffle, but there was no way to predict demand for prison "services" to "customers" over the next two to five years.

By mid-1999, a major effort focused on cleaning up the prison information systems through quality and accurate data was completed. By 2001, the department's information systems gatekeepers (everyone who entered and maintained data) had developed a culture of data quality. Though not unusual, it is important to note that some 15 to 20 percent of a company's operating revenue may be spent on workarounds or repairs of data-quality problems. And some organizations, like the Montana Department of Corrections, have created fulltime positions devoted to ensuring data quality.

Source: Adapted from Beth Stackpole, "Dirty Data Is the Dirty Little Secret That Can Jeopardize Your CRM Effort," C/O, February 15, 2001, pp.101-114.

Contextual DQ: Relevancy, value added, timeliness, completeness, amount of data

- *Intrinsic DQ:* accuracy, objectivity, believability, reputation
- *Accessibility DQ:* accessibility, access security
- *Representation DQ:* interpretability, ease of understanding, concise representation, consistent representation.

Strong et al. (1997) developed a framework that presents the major issues and barriers in each of the categories. They suggested that once the major variables and relationships in each category are identified, an attempt can be made to find out how to better manage the data. Some of the problems are technical ones, such as capacity, while others relate to potential computer crimes. For a comprehensive discussion, see Wang (1998).

Data quality is important, especially for CRM, ERP, and other enterprise information systems. The problem is that data warehousing, e-business, and CRM projects often expose poor-quality data because they require companies to extract and integrate data from multiple operational systems that are often peppered with errors, missing values, and integrity problems. These problems do not show up until someone tries to summarize or aggregate the data. See Dyche (2001).

Improved data quality is the result of a business improvement process designed to identify and eliminate the root causes of bad data. Data warehouse applications require data cleansing every time the warehouse is populated or updated. See King (2002). To improve data quality and maintain accuracy requires an active data quality assurance program. Berg and Heagele (1997) provide a management perspective and model for improving data quality. We describe their *data quality action plan*, which provides a framework, in DSS in Focus 5.5. Some specific major benefits from examples of improving data quality include integrating the information systems of two businesses that merged after an acquisition. Instead of a three-year effort, it was completed in one year. Another example is that of getting a CRM system completed and serving the sales and marketing organizations in one year instead of working on it for three years.

DSS IN FOCUS 5.5**A DATA QUALITY ACTION PLAN**

A data quality action plan is a recommended framework for guiding data quality improvement. Here are the steps to follow:

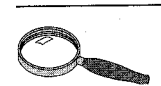
1. Determine the critical business functions to be considered.
2. Identify criteria for selecting critical data elements.
3. Designate the critical data elements.
4. Identify known data-quality concerns for the critical data elements, and their causes.
5. Determine the quality standards to be applied to each critical data element.
6. Design a measurement method for each standard.
7. Identify and implement quick-hit data quality improvement initiatives.
8. Implement measurement methods to obtain a data-quality baseline.
9. Assess measurements, data quality concerns, and their causes.
10. Plan and implement additional improvement initiatives.
11. Continue to measure quality levels and tune initiatives.
12. Expand process to include additional data elements.

Source: Adapted from Berg and Heagele (1997).

and then canceling it (see Olson, 2003a, 2003b). The Montana Department of Corrections situation described in DSS in Action 5.4 recovered from its low-quality data problem by developing a culture of quality through a data quality assurance plan.

We describe some best practices for data quality in DSS in Focus 5.6. Practitioners have identified these as important for an organization to maintain a high level of data quality and integrity.

Data-quality issues, methods, and solutions are discussed in great detail by Berson et al. (2000), Canter (2002), Dasu and Johnson (2003), Dravis (2002), Dyché (2001),

DSS IN FOCUS 5.6**BEST PRACTICES FOR DATA QUALITY**

Here are some best practices for ensuring data quality in practice.

- *Data scrubbing is not enough.* Data cleansing software only handles a few issues: inaccurate numbers, misspellings, incomplete fields. Comprehensive data-quality programs approach data standardization so that information can maintain its integrity.
- *Start at the top.* Top management must be aware of data quality issues and how they impact the organization. They must buy into any repair effort, because resources will be needed to address long-standing issues.
- *Know your data.* Understand what data you have, and what they are used for. Determine the appropriate level of precision necessary for each data item.
- *Make it a continuous process.* Develop a culture of data quality. Institutionalize a methodology and best practices for entering and checking information.
- *Measure results.* Regularly audit the results to ensure that standards are being enforced and to estimate impacts on the bottom line.

Source: Adapted from Beth Stackpole, "Dirty Data Is the Dirty Little Secret That Can Jeopardize Your CRM Effort," *CIO*, February 15, 2001, pp.101-114.

Eckerson (2002a), King (2002), Loshin (2001, 2003), Olson, (2003a, 2003b), Stackpole (2001), Stodder (2002), and Theodoratos and Bouzeghoul (2001).

DATA INTEGRITY

One of the major issues of DQ is data integrity. Older filing systems may lack integrity. That is, a change made in the file in one place may not be made in the file in another place or department. This results in conflicting data. Data quality specific issues and measures depend on the application of the data. This is an especially important issue in collaborative computing environments (Chapter 7), such as the one provided by Lotus Notes/Domino and Groove. In the area of the data warehouse, for example, Gray and Watson (1998) distinguish the following five issues:

- *Uniformity.* During data capture; uniformity checks ensure that the data are within specified limits.
- *Version.* Version checks are performed when the data are transformed through the use of metadata to ensure that the format of the original data has not been changed.
- *Completeness check.* A completeness check ensures that the summaries are correct and that all values needed to create the summary are included.
- *Conformity check.* A conformity check makes sure that the summarized data are "in the ballpark." That is, during data analysis and reporting, correlations are run between the value reported and previous values for the same number. Sudden changes can indicate a basic change in the business, analysis errors; or bad data.
- *Genealogy check or drill down.* A genealogy check or drill down is a trace back to the data source through its various transformations.

DATA ACCESS AND INTEGRATION

A decision-maker typically needs access to multiple sources of data that must be integrated (see the Opening Vignette and Case Applications 5.1 and 5.2). Before data warehouses, data marts, and business intelligence software, providing access to data sources was a major, laborious process. Even with modern Web-based data management tools, recognizing what data to access and providing it to the decision-maker is a nontrivial task that requires database specialists. As data warehouses grow in size, the issues of integrating data exasperate. This is especially important for the Department of Homeland Security. See Chabrow (2002) and DSS in Action 5.7 for how the DHS is working on a massive enterprise data and application integration project.

The needs of business analytics continue to evolve. In addition to historical, cleansed, consolidated, and point-in-time data, business users increasingly demand access to real-time, unstructured, and/or remote data. In addition, everything has to be integrated with the contents of their existing data warehouse. See Devlin, 2003. Moreover, access via PDAs and through speech recognition and synthesis is becoming more commonplace, further complicating integration issues (see Edwards, 2003).

Fox (2003) describes active information models for data transformation in developing an enterprise-wide system. These models take into consideration the necessary transformation logic to custom-developed high cost applications. Further, they must include the semantic and syntactic differences between schemas. This is especially important when corporate mergers occur and parallel applications must be integrated. Enterprise data resources can take many different forms: Relational Database (RDB) tables, XML documents, Electronic Data Interchange (EDI) messages, COBOL records, and so on. Independent Software Vendor (ISV) applications, such as enter-

DSS IN ACTION 5.7

HOMELAND SECURITY DATA INTEGRATION



Steve Cooper, special assistant to the president and CIA of the U.S. Department of Homeland Security (DHS), is responsible for determining which existing applications and types of data can help the organization meet its goal, migrating the data into a secure, usable, state-of-the-art framework, and integrating the disparate networks and data standards of 22 federal agencies, with 170,000 employees, that merged to form the DHS. This task is to be completed by mid-2005. The real problem is that federal agencies have historically operated autonomously, and their IT systems were not designed to interoperate with one another. Essentially, the DHS needs to link silos of data together.

The DHS has one of the most complex information-gathering and data migration projects under way in the federal government. The challenge of moving data from legacy systems (see Case Application 5.2), within or across agencies, is something all departments must address. Complicating the issue is the plethora of rapidly aging applications and databases throughout government. Data integration improvement is under way at the federal, local, and state levels. The government is utilizing tools from the corporate world.

Major problems have occurred because each agency has its own set of business rules that dictate how data are described, collected, and accessed. Some of the data are unstructured and not located in relational databases, and they cannot be easily manipulated and analyzed. Commercial applications will definitely be used in this major integration. Probably the bulk of the effort will be accomplished with data warehouse and datamart technologies. Informatica, among other software vendors, has developed data integration solutions that enable organizations to combine disparate systems to make information more widely accessible throughout an organization. Such software may be ideal for such a large-scale project.

The idea is to decide on and create an enterprise architecture (see Case Application 5.2) for federal and state agencies involved in homeland security. The architecture will help determine the success of homeland defense. The first step in migrating data is to identify all the applications and data in use. After identifying applications and databases, the next step is to determine which to use and which to discard. Once an organization knows which data and applications it wants to keep, the

difficult process of moving the data starts. First, it is necessary to identify and build on a common thread in the data. Another major challenge in the data-migration arena is security, especially when dealing with data and applications that are decades old.

Homeland Security will definitely have an information-analysis and infrastructure-protection component. This may be the single most difficult challenge for the DHS. Not only will Homeland Security have to make sense of a huge mountain of intelligence gathered from disparate sources, but then it will have to get that information to the people who can most effectively act on it. Many of them are outside the federal government.

Even the central government recognizes that data deficiencies may plague the DHS. Moving information to where it is needed, and doing so when it is needed, is critical and exceedingly difficult. Some 650,000 state and local law enforcement officials "operate in a virtual intelligence vacuum, without access to terrorist watch lists provided by the State Department to immigration and consular officials," according to the October 2002 Hart-Rudman report, "America Still Unprepared America Still in Danger," sponsored by the Council on Foreign Relations. The task force cited the lack of intelligence sharing as a critical problem deserving immediate attention. "When it comes to combating terrorism, the police officers on the beat are effectively operating deaf, dumb and blind," the report concluded.

DARPA, the Defense Advanced Research Projects Agency, spent \$240 million on combined projects on Total Information Awareness, to develop ways of treating worldwide, distributed legacy databases as if they were a single, centralized database.

Sources: Adapted from Eric Chabrow, "One Nation, Under I.T." *Information Week*, No. 914, November 11, 2002, pp. 47-50; Todd Datz, "Integrating America," *CIO*, December 2002, p. 44-51; John Foley, "Data Debate." *Informationweek*, May 19, 2003, pp. 22-24; Amy Rogers Nazarov, "Informatica Seeks Partners to Gain Traction in Fed Market." *CRN*, June 9, 2003, p. 39; Patrick Thibodeau, "DHS Sets Timeline for IT Integration," *Computer World*, June 16, 2003, p. 7; Katherine McIntire Peters, "5 Homeland Security Hurdles," *Government Executive*, Vol. 35, No. 2, pp. 18-21, February 2003; Amy Rogers, "Data Sharing Key to Homeland Security Efforts," *eRN*, No. 1019, November 4, 2002, pp. 39-40; and Karen D. Schwartz, "The Data Migration Challenge," *Government Executive*, Vol. 34, No. 16, December 2002, pp. 70-72.

prise resource planning, customer relationship management software, and in-house-developed software, define their own input and output schemas. Often, different schemas hold similar information structured differently. The information model is central in that it represents a neutral semantic view of the enterprise. See Fox (2003) for details. Case Application 5.2 describes how a firm developed an infrastructure for integrating data from disparate sources. DSS in Focus 5.8 describes the processes of extract, transform, and load (ETL), which are the basis for all data-integration efforts.

Many integration projects involve enterprise-wide systems. In DSS in Focus 5.9, we provide a checklist of what works and what does not work when attempting such a project. See Orovic (2003) for details and impacts. Also see Chapter 6 for details on DSS implementation.

Integrating data properly from various databases and other disparate sources is difficult. But when not done properly, it can lead to disaster in enterprise-wide systems like CRM, ERP, and supply chain projects (Nash, 2002). See DSS in Focus 5.10 for issues relating to data cleansing as a part of data integration. Also see Dasu and Johnson (2003). Madsen (2003) describes how a real-time delivery infrastructure (see Section 5.12) allows an enterprise to easily integrate applications on a repeatable basis and yet remain flexible enough to accommodate change.

The following authors discuss data integration issues, models, methods, and solutions: Balen (2000), Calvanese et al. (2001), Devlin (2003), Erickson (2003), Fox (2003), Holland (2000), McCright (2001), Meehan (2002), Nash (2002), Orovic (2003), Vaughan (2003), Pelletier, Pierre, and Hoang (2003), and Whiting (2002).

DATA INTEGRATION VIA XML

XML is quickly becoming the standard language for database integration and data, transfer (Balen, 2000). By 2004, some 40 percent of all e-commerce transactions occurred over XML servers. This was up from 16 percent in 2002 (see Savage, 2001). XML may revolutionize electronic data exchange by becoming the *universal data translator* (Savage, 2001). Systems developers must be extremely careful because XML *cannot overcome poor business logic*. If the business processes are bad, no data integration method will improve them.

Even though XML is an excellent way to exchange data among applications and organizations, a critical issue is whether it can function well as a native database format in practice. XML is a mismatch with relational databases: it works, but is hard to maintain. There are difficulties in performance, specifically in searching large databases.

DSS IN FOCUS 5.8

WHAT IS ETL?



Extract, transform, and load (ETL) programs periodically extract data from source systems, transform them into a common format, and then load them into the target data store, typically a data warehouse or data mart. ETL tools also typically transport data between sources and targets, document how data elements change as they move between source and target (e.g., metadata), exchange metadata with other applications as needed,

and administer all run-time processes and operations (e.g., scheduling, error management, audit logs, statistics). ETL is extremely important for data integration and data warehousing.

Source: Adapted from Wayne Erickson, "The Evolution of ETL," in *What Works: Best Practices in Business Intelligence and Data Warehousing*, Vol. 15, The Data Warehousing Institute, Chatsworth, CA, June, 2003.

DSS IN FOCUS 5.9**WHAT TO DO AND WHAT NOT TO DO WHEN
IMPLEMENTING AN ENTERPRISEWIDE
INTEGRATION PROJECT****WHAT TO DO:**

1. Think globally and act locally. Plan enterprise-wide; implement incrementally.
2. Define integration framework components.
3. Focus on business-driven goals with high cost and low technical complexity.
4. Treat the enterprise system as your strategic application.
5. Pursue reusable, template-based approaches to development.
6. Use prototyping as the project estimate generator.
7. Think of integration at different levels of abstraction.
8. Expect to build application logic into the enterprise infrastructure.
9. Assign project responsibility at the highest corporate level and negotiate, negotiate, negotiate.
10. Plan for message logging and warehouse to track audit and recovery.

2. Purchase more than you need for a given phase.
3. Substitute an enterprise application architecture for a data warehouse.
4. Force usage of near-real-time message-based integration unless it is absolutely mandatory.
5. Assume that existing process models will suffice for process integration; they are not the same.
6. Plan to change your business processes as part of the enterprise application implementation.
7. Assume that all relevant knowledge resides within the project team.
8. Be driven by centralizing any enterprise-level business objects as part of the enterprise application implementation.
9. Be intrusive into the existing applications.
10. Use ad hoc process and message modeling techniques.

WHAT NOT TO DO:

1. Critique business strategy through the enterprise architecture. Instead evaluate the impact of the business strategy on IT.

Source: Adapted from V. Orovic, "To Do & Not to Do," eAI Journal, June, 2003, pp. 37-43.

XML uses a lot of space. Even so, there are native XML database engines. See DeJesus (2000) for more on these.

DATA INTEGRATION SOFTWARE

Developers of document and data capture and management software are increasingly utilizing XML to transport data from sources to destinations. For example, Captiva Software Corp., RTSe USA Inc., Kofax Image Products Inc., and Tower Software all utilize XML to move and upload documents to the Web, intranets, and wireless applications. RosettaNet XML Solutions create standard B2B protocols that increase supply chain efficiency. BizTalk Server 2000 uses XML to help companies manage their data, conduct data exchanges with e-commerce partners more easily, and lower costs (Savage, 2001). The ADT (formerly InfoPump) data-transformation tools from Computer Associates track changes in data and applications. The software lets companies extract and transform data from up to 30 sources including relational databases, mainframe IMS and VSAM files, and applications, and load them into a database or data warehouse. Vaughan (2003) provides a list of software tools that use XML to extract and transform data.

DSS IN FOCUS 5.10

ENTERPRISE DATA HOUSE CLEANING



Every organization has redundant data, wrong data, missing data, and miscoded data, probably buried in systems that do not communicate much. This is the attic problem familiar to most homeowners: Throw in enough boxes of seasonal clothes, holiday trim, family history documents, and other important items, and soon the mess is too big to manage. It happens at companies, too. Multiple operating units, manufacturing plants, and other facilities may all run different vendors' applications for sales, human resources, and other tasks. The mix of disparate data makes for a pile of unsorted and unreconciled information. Integration becomes a major effort.

(LEANING HOUSE:

Before any data can be cleansed, your IT department must create a plan for finding and collecting all the data and then decide how to manage them. Practitioners offer this advice:

1. Decide what types of information must be captured. Set up a small data-mapping committee to do this.
2. Find mapping software that can harvest data from many sources, including legacy applications, PC files, HTML documents, unstructured sources, and enterprise systems. Several vendors have developed such software.
3. Start with a high-payoff project. The first integration project should be in a business unit that generates high revenue. This helps obtain upper-management buy-in.
4. Create and institutionalize a process for mapping, cleansing, and collating data. Companies must continually capture information from disparate sources.

Source: Adapted from Kim S. Nash, "Merging Data Silos," ComputerWorld, April 25, 2002, pp. 30-32.

5 .. 4 THE WEB/INTERNET AND COMMERCIAL DATABASE SERVICES

External data pour into organizations from many sources. Some of the data come on a regular basis from business partners through collaboration (e.g., collaborative supplychain management; see Chapters 7 and 8). The Internet is a major source of data.

- The **WebInternet**. Many thousands of databases all over the world are accessible through the Web/Internet. A decision-maker can access the home pages of vendors, clients, and competitors, view and download information, or conduct research. The Internet is the major supplier of external data for many decision situations.
- Commercial data banks. An **online (commercial) database** service sells access to specialized databases. Such a service can add external data to the MSS in a timely manner and at a reasonable cost. For example, GIS data must be accurate; regular updates are available. Several thousand services are currently available, many of which are accessible via the Internet. Table 5.3 lists several representative services.

The collection of data from multiple external sources may be complicated.

Products from leading companies, such as Oracle, IBM, and Sybase, can transfer information from external sources and put it where it is needed, when it is needed, in a usable form.

Since most sources of external data are on the Web, it makes sense to use intelligent agents to collect and possibly interpret external data. Pelletier, Pierre, and Hoang (2003) describe a multi-agent system designed for intelligent information retrieval from het-

52.

TABLE 5.3 Representative Commercial Database (Data Bank) Services

CompuServe (compuserve.com) and The Source. Personal computer networks providing statistical data banks (business and financial market statistics) as well as bibliographic data banks (news, reference, library, and electronic encyclopedias). CompuServe is the largest supplier of such services to personal computer users.

Compustat (compustat.com). Provides financial statistics about tens of thousands of corporations. Data Resources Inc. offers statistical data banks for agriculture, banking, commodities, demographics, economics, energy, finance, insurance, international business, and the steel and transportation industries. DRI economists maintain a number of these data banks. Standard & Poor's is also a source. It offers services under the U.S. Central Data Bank.

Dow-Jones Information Service. Provides statistical data banks on stock market and other financial markets and activities, and in-depth financial statistics on all corporations listed on the New York and American stock exchanges, plus thousands of other selected companies. Its Dow Jones News/Retrieval System provides bibliographic data banks on business, financial, and general news from the *Wall Street Journal*, *Barron's*, and the Dow Jones News Service.

Lockheed Information Systems. The largest bibliographic distributor. Its DIALOG system offers extracts and summaries of hundreds of different data banks in agriculture, business, economics, education, energy, engineering, environment, foundations, general news publications, government, international business, patents, pharmaceuticals, science, and social sciences. It relies on many economic research firms, trade associations, and government agencies for data.

Mead Data Central (www.mead.com). This data bank service offers two major bibliographic data banks. Lexis provides legal research information and legal articles. Nexis provides a full-text (not abstract) bibliographic database of hundreds of newspapers, magazines, and newsletters, news services, government documents, and so on. It includes full text and abstracts from the *New York Times* and the complete 29-volume *Encyclopedia Britannica*. Also provided are the Advertising & Marketing Intelligence (AMI) data bank and the National Automated Accounting Research System.

erogeneous distributed sources. The system uses software agents and is ideal for intelligent integration. For another example of how this is performed, see Liu et al. (2000).

THE WEB AND CORPORATE DATABASES AND SYSTEMS

Developments in document management systems (DMS) and content management systems (eMS) include the use of Web browsers by employees and customers to access vital information. Critical issues have become more critical in Web-based systems (see Gates, 2002; Rapoza, 2003). It is important to maintain accurate, up-to-date versions of documents, data, and other content, since otherwise the value of the information will diminish. Real-time computing, especially as it relates to DMS and CMS, has become a reality. Managers expect their DMS and CMS to produce up-to-the-minute accurate documents and information about the status of the organization as it relates to their work (see Gates, 2002; Raden, 2003a, 2003b). This real-time access to data introduces new complications in the design and development of data warehouses and the tools that access them. See Section 5.12 for details. Other Web developments include Pilot Software's Decision Support Suite (pilotsw.com) combined with BlueIsle Software's InTouch (blueisle.com) and group support systems deployed via Web browsers (e.g., Lotus Notes/Domino and Groove), and database management systems that provide data directly in a format that a Web browser can display with delivery over the Internet or an intranet. Pilot's Internet Publisher is a standalone Web product, as is DecisionWeb from Corns hare (comshare.com).

The "big three" vendors of relational database management systems—Oracle, Microsoft, and IBM—all have core database products to accommodate a world of

client/server architecture and Internet/intranet applications that incorporate nontraditional, or rich, multimedia data types. So do other firms in this area. Oracle's Developer/2000 is able to generate graphical client/server applications in PLISQL code, Oracle's implementation of structured query language (SQL), as well as in COBOL, C++, and HTML. Other tools provide Web browser capabilities, multimedia authoring and content scripting, object class libraries, and OLAP routines. Microsoft's .Net strategy supports Web-based business intelligence.

Among the suppliers of Web site and database integration are Spider Technologies (spidertech.com), Hart Software (hart.com), Next Software Inc. (next.com), NetObjects Inc. (netobjects.com), Oracle Corporation (oracle.com), and OneWave Inc. (onewave.com). These vendors link Web technology to database sources and to legacy database systems.

The use of the Web has had a far-reaching impact on collaborative computing in the form of groupware (Chapter 7), enterprise information systems (Chapter 8), knowledge-management systems (Chapter 9), document management systems, and the whole area of interface design, including the other enterprise information systems: ERP/ERM, CRM, PLM, and SCM.

5.5 DATABASE MANAGEMENT SYSTEMS IN DECISION SUPPORT SYSTEMS/BUSINESS INTELLIGENCE

The complexity of most corporate databases and large-scale independent MSS databases sometimes makes standard computer operating systems inadequate for an effective and efficient interface between the user and the database. A **database management system (DBMS)** supplements standard operating systems by allowing for greater integration of data, complex file structure, quick retrieval and changes, and better data security. Specifically, a DBMS is a software program for adding information to a database and updating, deleting, manipulating, storing, and retrieving information. A DBMS combined with a modeling language is a typical system-development pair used in constructing decision support systems and other management-support systems. DBMS are designed to handle large amounts of information. Often, data from the database are extracted and put in a statistical, mathematical, or financial model for further manipulation or analysis. Large, complex DSS often do this.

The major role of DBMS is to manage data. By *manage*, we mean to create, delete, change, and display the data. DBMS enable users to query data as well as to generate reports. For details, see Ramakrishnan and Gehrke (2002). Effective database management and retrieval can lead to immense benefits for organizations, as is evident in the situation of Aviall Inc., described in DSS in Action 5.11.

Unfortunately, there is some confusion about the appropriate role of DBMS and spreadsheets. This is because many DBMS offer capabilities similar to those available in an integrated spreadsheet such as Excel, and this enables the DBMS user to perform DSS spreadsheet work with a DBMS. Similarly, many spreadsheet programs offer a rudimentary set of DBMS capabilities. Although such a combination can be valuable in some cases, it may result in lengthy processing of information and inferior results. The add-in facilities are not robust enough and are often very cumbersome. Finally, the computer's available RAM may limit the size of the user's spreadsheet. For some applications, DBMS work with several databases and deal with many more data than a spreadsheet can.

DSS IN ACTION 5.11**AVIALL LANDS \$3 BILLION DEAL**

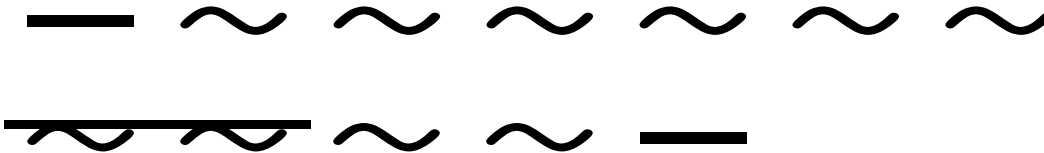
How important is effective data management and retrieval? Aviall Inc. attributes a \$3 billion spare parts distribution contract that it won to its IT infrastructure. The ten-year contract requires the company to distribute spare parts for Rolls-Royce aircraft engines. The ability to offer technology-driven services, such as sales forecasting, down to the line-item level was cited as one of the reasons why Aviall was successful. It recently linked information from its ERP, supply chain management, customer-relationship management, and

e-business applications to provide access to its marine and aviation parts inventory and distribution (at a cost of some \$30 to \$40 million). The system is expected to pay for itself by cutting costs associated with "lost" inventory. Timely access to information is proving to be a competitive resource that results in a *big payoff*

Source: Adapted from Marc L. Songini, "Distribution Deal Prods Tight IT Ties Between Aviall, Rolls-Royce," *ComputerWorld*, January 14, 2002.

For DSS applications, it is often necessary to work with both data and models. Therefore, it is tempting to use only one integrated tool, such as Excel. However, interfaces between DBMS and spreadsheets are fairly simple, facilitating the exchange of data between more powerful independent programs. Web-based modeling and database tools are designed to seamlessly interact (Fourer, 2001).

Small to medium DSS can be built either by enhanced DBMS or by integrated spreadsheets. Alternatively, they can be built with a DBMS program and a spreadsheet program. A third approach to the construction of DSS is to use a fully integrated DSS generator (Chapter 6).



5.6 DATABASE ORGANIZATION AND STRUCTURES

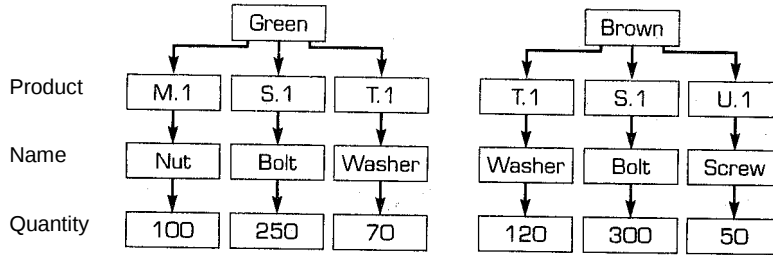
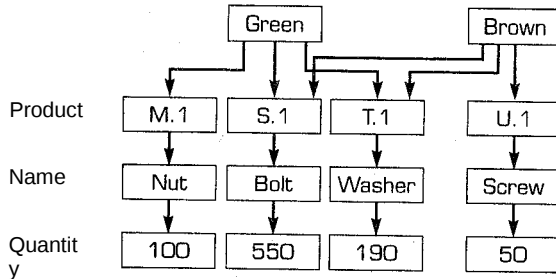
The relationships between the many individual records stored in a database can be expressed by several logical structures (see Korth, 2002; Mannino, 2001; McFadden et al., 2002; Post, 2002; and Riccardi, 2003). DBMS are designed to use these structures to perform their functions. The three conventional structures--relational, hierarchical, and network--are shown in Figure 5.1.

RELATIONAL DATABASES

The relational form of DSS database organization, described as tabular or flat, allows the user to think in terms of two-dimensional tables, which is the way many people see data reports. Relational DBMS allow multiple access queries. Thus, a data file consists of a number of columns proceeding down a page. Each column is considered a separate field. The rows on a page represent individual records made up of several fields, the same design that is used by spreadsheets. Several such data files can be related by means of a common data field found in two (or more) data files. The names of common fields must be spelled exactly alike, and the fields must be the same size (the same number of bytes) and type (e.g., alphanumeric or dollar). For example, in Figure 5.1 the data field Customer Name is found in both the customer and the usage file, and thus they are related. The data field Product Number is found in the product file and the

a Relational

I		I		t		~		Field s
Customer Number	Customer Name	Product Number	Product Name	Customer Name	Product Number	Quantity		
8	Green	M.1	Nut	Green	M.1	10	Records	
10	Brown	8.1	Bolt	Brown	8.1	300		
30	Black	11	Washer	Green	11	70		
45	White	U.1	8crew	White	8.1	30		
Customer Records		Product Records		Green	8.1	250		
				Brown	11	120		
				Brown	U.1	50		
				Usage Records				

b. Hierarchical**c. Network****FIGURE 5.1 DATABASE STRUCTURES**

usage file. It is through these common linkages that all three files are related and in combination form a **relational database**.

The advantage of this type of database is that it is simple for the user to learn, is easily expanded or altered, and can be accessed in a number of formats not anticipated at the time of the initial design and development of the database. It can support large amounts of data and efficient access. Many data warehouses are organized this way.

HIERARCHICAL DATABASES

A hierarchical model orders data items in a top-down fashion, creating logical links between related data items. It looks like a tree or an organization chart. It is used mainly in transaction processing, where processing efficiency is a critical element.

NETWORK DATABASES

The network database structure permits more complex links, including lateral connections between related items. This structure is also called the CODASYL model. It can save storage space through the sharing of some items. For example, in Figure 5.1, Green and Brown share Sol and T.1.

OBJECT-ORIENTED DATABASES

Comprehensive MSS applications, such as those involving computer-integrated manufacturing (CIM), require accessibility to complex data, which may include pictures and elaborate relationships. Such situations cannot be handled efficiently by hierarchical, network, or even relational database architectures, which mainly use an alphanumeric approach. Even the use of SQL to create and access relational databases may not be effective. For such applications, a graphical representation, such as the one used in object-oriented systems, may be useful.

Object-oriented data management is based on the principle of object-oriented programming (see details in the Web Chapter; also see Moore and Britt, 2001). Object-oriented database systems combine the characteristics of an object-oriented programming language, such as Veritos or UML, with a mechanism for data storage and access. The object-oriented tools focus directly on the databases. An **object-oriented database management system (OODBMS)** allows one to analyze data at a conceptual level that emphasizes the natural relationships between objects. Abstraction is used to establish inheritance hierarchies, and object encapsulation allows the database designer to store both conventional data and procedural code within the same objects.

An object-oriented data management system defines data as objects and encapsulates data along with their relevant structure and behavior. The system uses a hierarchy of classes and subclasses of objects. Structure, in terms of relationships, and behavior, in terms of methods and procedures, are contained within an object.

The worldwide relational and object-relational database management systems software market is expected to grow to almost \$20 billion by 2006, according to IDC (The Day Group, 2002). Object-oriented database managers are especially useful in distributed DSS for very complex applications. Object-oriented database systems have the power to handle the complex data used in MSS applications. For a descriptive example, see DSS in Action 5.12. Trident Systems Group Inc. (Fairfax, Virginia) has developed a large-scale object-oriented database system for the U.S. Navy (see Sgarioto, 1999).

MULTIMEDIA-BASED DATABASES

Multimedia database management systems (MMDBMS) manage data in a variety of formats, in addition to the standard text or numeric field. These formats include images, such as digitized photographs, and forms of bit-mapped graphics, such as maps or .PIC files, hypertext images, video clips, sound, and virtual reality (multidimensional images). Cataloguing such data is tricky. Accurate and known key words must be used. It is critical to develop effective ways to manage such data for GIS and for many other Web applications. Managing multimedia data continues to become more important for business intelligence (see D'Agostino, 2003).

Most corporate information resides outside the computer in documents, maps, photos, images, and videotapes. For companies to build applications that take advantage of such rich data types, a special database management system with the ability to manage and manipulate multiple data types must be used. Such systems store rich mul-

c. PIERCE WOOD MEMORIAL HOSPITAL OBJECTS



Glenn Palmier, data processing manager for G. Pierce Wood Memorial Hospital (GPW), was not happy that the vendor of his database-management systems, InterSystems Corp., was upgrading to an object-oriented architecture in its core product, CACHE. At the time, GPW had 45 different systems developed over 15 years at the state mental health facility in Arcadia, Florida. Smooth operations and fast data access were critical to GPW. The vendor moved quickly, reducing a five-year conversion plan to eight months. By then, GPW had converted all its systems to be object-oriented and Web-based. GPW focused on data usability in the conversion process. Databases were updated to

work better in the new object-oriented environment. After reengineering the databases and upgrading, the newsystems ran faster than ever before. For example, the old system required almost two hours to perform a certain query. The new system takes less than a minute. Personnel have been easily and quickly trained in the new systems, and the use of Web browsers to access data fits perfectly into the state's Internet strategy.

Source: Adapted from Jon William Toigo, "Objects Are Good for Your Mental Health," *Enterprise Systems*, June 2001, pp. 34-35.

multimedia data types as *binary large objects* (BLOBS). Database management systems are evolving to provide this capability (McFadden et al., 2002). It is critical to design the management capability upfront, with scalability in mind. For a lucky example of a situation that was not developed as such, but worked, Hurwicz (2002) describes NASA's experience when it endeavored to download and catalogue images from space for educational purposes, as envisioned by astronaut Sally Ride. Fortunately, there was time and volunteer effort enough to redesign the cataloguing mechanism on the Web-based, multimedia database system. See Hurwicz (2002) for details about the development issues, and the EarthKAM Web site (www.earthkam.ucsd.edu) for direct access to the online, running database system. Note that similar problems can occur in data warehouse design and development.

For Web-related applications of multimedia databases, see Maybury (1997), and multimedia demonstrations on the Web, including those of Macromedia's products and Visual Intelligence Corporation. Also see DSS in Action 5.13. In DSS in Action 5.14, we describe how an animated film production company utilized several multimedia databases to develop the *Jimmy Neutron: Boy Genius* film. The databases and managerial techniques have since led to lower overall production costs for the animated television series.

Some computer hardware (including the communication system with the database) may not be capable of playback in real-time. A delay with some buffering might be necessary (e.g., try any audio or video player in Windows). Intel Corporation's Pentium processor chips incorporate multimedia extension (MMX) technology for processing multimedia data for real-time graphics display. Since then, this and similar technologies have been embedded in many CPU and auxiliary processor chips.

DOCUMENT-BASED DATABASES

Document-based databases, also known as electronic document management (EDM) systems (Swift, 2001), were developed to alleviate paper storage and shuffling. They are used for information dissemination, form storage and management, shipment tracking, expert license processing, and workflow automation. Many content management systems (CMS) are based on EDM. In practice, most are implemented in Web-based sys-

DSS IN ACTION 5.13

MULTIMEDIA DATABASE MANAGEMENT SYSTEMS: A SAMPLER



IBM developed its DB2 Digital Library multimedia server architecture for storing, managing, and retrieving text, video, and digitized images over networks. Digital Library consists of several existing IBM software and hardware products combined with consulting and custom development (see ibm.com). Digital Library will compete head to head with multimedia storage and retrieval packages from other leading vendors.

MediaWay Inc. (mediaway.com) claims that its multimedia database management system can store, index, and retrieve multimedia data (sound, video, graphics) as easily as relational databases handle tabular data. The DBMS is aimed at companies that want to build what Media Way calls *multimedia cataloging applications* that manage images, sound, and video across

multiple back-end platforms. An advertising agency, for example, might want to use the product to build an application that accesses images of last year's advertisements stored on several servers. It is a client/server implementation. MediaWay is not the only vendor to target this niche, however. Relational database vendors, such as Oracle Corporation and Sybase Inc., have incorporated multimedia data features in their database servers.

In addition, several desktop software companies promote client databases for storing scanned images. Among the industries that use this technology are health care, real estate, retailing, and insurance.

Source: Condensed and adapted from the Web sites and publicly advertised information of various vendors.

JIMMY NEUTRON: THE "I CAN FIX THAT" DATABASE ~

Producers and animators working on the film *Jimmy Neutron: Boy Genius* tracked thousands of frames on four massive databases. DNA Productions (Irving, Texas), the animation services company that worked with Nickelodeon and screenwriter and director Steve Oedekerk to produce the film, addressed the problem of assembling the 1800 shots' that comprise the 82minute film by logging and tracking them in four FileMaker Pro databases. One tracked initial storyboards, another tracked the shots assigned to individual artists, the third tracked the progress of each frame throughout the production process, and the fourth tracked retakes (changes to completed shots). At the

film's completion, there were 20,000 entries. Each record tracked information about each shot dating back to the beginning of the project. The databases enabled the film to be completed in a mere eighteen months. The best part is that everyone had access to the shots instantly, instead of having to track down an individual or walk over to a large 4 by 8 foot (1.3 by 2.6 meter) board and look for it. Since making the film, the *Jimmy Neutron* TV series continues to utilize the database technology.

Source: Adapted from Stephanie Overby, "Animation Animation," *CIO*, 2002, May 15, 2002, pp. 22-24.

terns. See Bolles (2003), Gates (2002), and Rapoza (2003). Since EDM uses both object-oriented and multimedia databases, document-based databases were included in the preceding two sections. What is unique to EDM are the implementation and the applications. McDonnell Douglas Corporation distributes aircraft service bulletins to its customers around the world through the Internet. The company used to distribute a staggering volume of bulletins to over 200 airlines, using over 4 million pages of documentation every year. Now it is all on the Web, saving money and time both for the company and for its customers. Motorola uses DMS not only for document storage and retrieval but also for small-group collaboration and company-wide knowledge sharing. It has developed virtual communities where people can discuss and publish information, all with the Web-enabled DMS.

56.

Web-enabled document management systems have become an efficient and cost effective delivery system. American Express now offers its customers the option of receiving monthly billing statements online, including the ability to download statement detail, retrieve prior billing cycles, and view activity that has been posted but not yet billed. As this option grows in popularity, it will reduce production and mailing costs. Xerox Corporation developed its first knowledge management system on its EDM platform (see Chapter 9).

INTELLIGENT DATABASES

Artificial intelligence (AI) technologies, especially Web-based intelligent agents and artificial neural networks (ANN), simplify access to and manipulation of complex databases. Among other things, they can enhance the database management system by providing it with an inference capability, resulting in an **intelligent** database.

Difficulties in integrating ES into large databases have been a major problem even for major corporations. Several vendors, recognizing the importance of integration, have developed software products to support it. An example of such a product is the Oracle relational DBMS, which incorporates some ES functionality in the form of a query optimizer that selects the most efficient path for database queries to travel. In a distributed database, for example, a query optimizer recognizes that it is more efficient to transfer two records to a machine that holds 10,000 records than vice versa. (The optimization is important to users because with such a capability they need to know only a few rules and commands to use the database.) Another product is the INGRES II Intelligent Database.

Intelligent agents can enhance database searches, especially in large data warehouses. They can also maintain user preferences (e.g., amazon.com) and enhance search capability by anticipating user needs. These are important concepts that ultimately lead to ubiquitous computing. See DSS in Focus 5.15 for details of recent developments in intelligent agents.

DSS IN FOCUS 5.15

THE BOTS OF THE FUTURE



There are plenty of software agents in use today. They are found in help systems, search engines, and comparison-shopping tools. During the next few years, as technologies mature and agents radically increase their value by communicating with one another, they will significantly affect an organization's business processes. Training, decision support, and knowledge sharing will be affected, but experts see procurement as the killer application of business-to-business agents. Intelligent software agents (bots) feature triggers that allow them to execute without human intervention. Most agents also feature adaptive learning of users' tendencies and preferences and offer personalization based on what they learn about users.

One goal of software agent developers is to develop machines that perform tasks that people do not

want to do. Another is to delegate to machines tasks at which they are vastly superior to humans, such as comparing the price, quality, availability, and shipping cost of items.

BotKnowledge.com Agents can automatically perform intelligent searches, answer questions, tell you when an event occurs, individualize news delivery, tutor, and comparison shop.

Agents migrate from system to system, communicating and negotiating with each other. They are evolving from facilitators into decision-makers.

Source: Adapted from S. Ulfelder, "Undercover Agents," Computer World, June 5, 2000.

One of IBM's main initiatives in commercial AI provides a knowledge-processing subsystem that works with a database, enabling users to extract information from the database and pass it to an expert system's knowledge base in several different knowledge representation structures. Databases now store photographs, sophisticated graphics, audio, and other media. As a result, access to and management of databases are becoming more difficult, and so are the accessibility and retrieval of information. The use of intelligent systems in database access is also reflected in the use of natural language interfaces which can be used to help nonprogrammers retrieve and analyze data.

5.7 DATA WAREHOUSING

The Opening Vignette demonstrates a scenario in which a **data warehouse** can be utilized to support decision-making, analyzing large amounts of data from various sources to provide rapid results to support a critical process. The necessary data are scattered across many government agencies, and consolidating the data to make them available when needed will entail serious organizational and technical challenges.

Organizations, private and public, continuously collect data, information, and knowledge at an increasingly accelerated rate and store them in computerized systems. Updating, retrieving, using, and removing this information becomes more complicated as the amount increases. At the same time, the number of users that interact with the information continues to increase as a result of improved reliability and availability of network access, especially including the Internet. Working with multiple databases is becoming a difficult task that requires considerable expertise (see DSS inAction 5.16). Data for the data warehouse are brought in from various external and internal

DSS IN ACTION 5.16

DATA WAREHOUSING SUPPORTS FIRST AMERICAN ~ CORPORATION'S CORPORATE STRATEGY

First American Corporation changed its corporate strategy from a traditional banking approach to one that was centered on customer relationship management. This enabled First American to transform itself from a company that lost \$60 million in 1990 to an innovative financial services leader a decade later. The successful implementation of this strategy would not have been possible without a data warehouse called VISION that stored information about customer behaviors, such as products used, buying preferences, and client value positions. VISION provided:

- Identification of the top 20 percent of profitable customers
- Identification of the 40-50 percent of unprofitable customers
- Retention strategies

- Lower-cost distribution channels
- Strategies to expand customer relationships
- Redesigned information flows.

Access to information through a data warehouse can enable both evolutionary and revolutionary change. First American Corporation was able to achieve revolutionary change, transforming itself into the *Sweet 16* of financial services corporations.

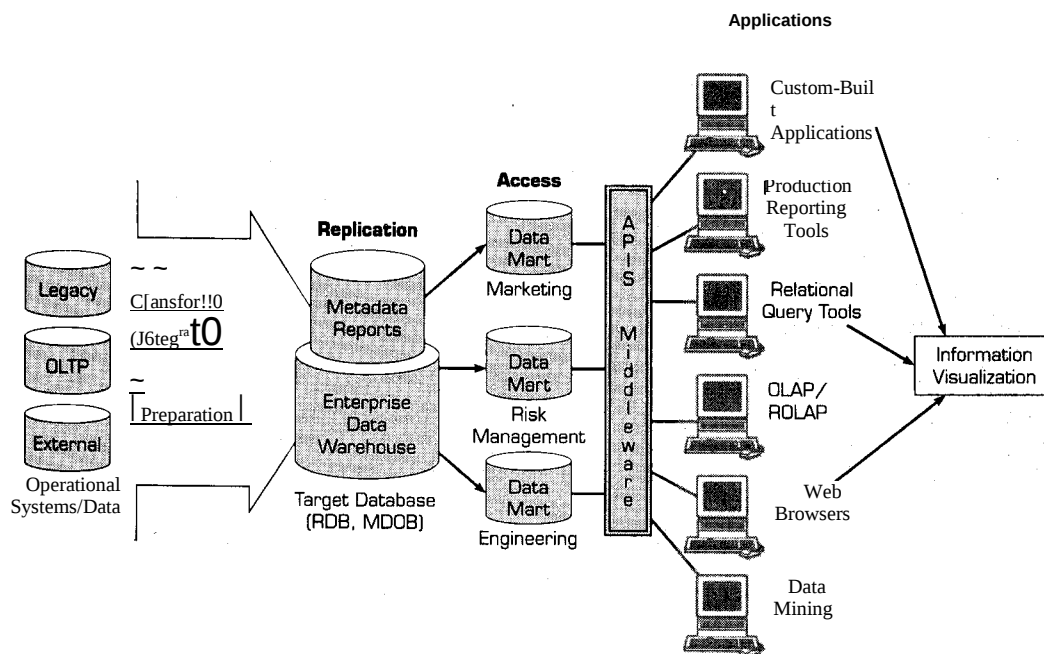
Source: Adapted from B. Cooper, H. I. Watson, B. H. Wixom, and D. Goodhue, "First American Tennessee Case Study," *MIS Quarterly*, 2004, forthcoming. Also presented as "Data Warehousing Supports Corporate Strategy at First American Corporation." SIM International's Best Paper Contest Recipients, 1999.

resources and are cleansed and organized in a manner consistent with the organization's needs. Once the data are populated in the data warehouse, data marts may be loaded for a specific area or department. Often, the data marts are bypassed, and business intelligence tools on client pes simply load and manipulate local data cubes. Data warehouses can be described as subject-oriented, integrated, time-variant, nonnormalized, non-volatile collections of data that support analytical decision-making. See Figure 5.2 for the data warehouse framework and views. Edelstein (1997) presents a good general introduction to data warehousing. Mannino (2001) discusses data Warehouse technology and management.

Since enterprise information management solutions aggregate or consolidate report information and electronic documents created by any application running on any platform, the enterprise information management solution extends the access to information and reports processed from the data warehouse (see Mullin, 2002). An enterprise data warehouse is a comprehensive database that supports all decision analysis required by an organization by providing summarized and detailed information. As implied in this definition, the data warehouse has access to all information *relevant* to the organization, which may come from many different sources, both internal and external. See Figure 5.2 for how data work their way into the data warehouse (on the left), for further analysis by tools (to the right).

A data warehouse begins with the physical separation of a company's operational and decision support environments. At the heart of many companies lies a store of *operational data*, usually derived from critical mainframe-based online transaction processing (OLTP) systems, such as order entry point of sales applications. Many legacy

FIGURE 5.2 DATA WAREHOUSE FRAMEWORK AND VIEWS



OLTP systems were implemented primarily in COBOL (especially banking systems), and still operate in a customer information control system (CICS) environment. OLTP systems for financial and inventory management and control, for example, also produce operational data. (Many firms are implementing Web front ends for such *legacy* systems. This could be a major and costly mistake. See Case Application 5.2 and Chapter 6.) In the operational environment, data access, application logic tasks, and data-presentation logic are tightly coupled together, usually in non-relational databases. OLTP data are usually detail data that control a specific event, such as the recording of a sales transaction, and are generally not summarized. These nonrelational data stores are not very conducive to data retrieval for decision support/business intelligence/business analytic applications. However, decision support information must be made accessible to management. *It is important to physically separate the data warehouse from the OLTP system.*

CHARACTERISTICS OF DATA WAREHOUSING

The major characteristics of data warehousing are as follows:

- *Subject-oriented.* Data are organized by detailed subject (e.g., by customer, policy type, and claim in an insurance company), containing only information relevant for decision support. Subject orientation enables users to determine not only how their business is performing, but why. A data warehouse differs from an operational database in that most operational databases have a product orientation and are tuned to handle transactions that update the database; subject orientation provides a more comprehensive view of the organization.
- *Integrated.* Data at different source locations may be encoded differently. For example, gender data may be encoded as 0 and 1 in one place and "m" and "f" in another. In the warehouse they are *scrubbed* (cleaned) into one format so that they are standardized and consistent. Many organizations use the same terms for data of different kinds. For example, "net sales" may mean net of commission to the marketing department but gross sales returns to the accounting department. Integrated data resolve inconsistent meanings and provide uniform terminology throughout the organization. Also, data and time formats vary around the world.
- *Time-variant (time series).* The data do not provide the current status. They are kept for five or ten years or more and are used for trends, forecasting, and comparisons. There is a *temporal* quality to a data warehouse. *Time is the one important dimension that all data warehouses must support.* Data for analysis from multiple sources contain multiple time points (e.g., daily, weekly, monthly views).
- *Nonvolatile.* Once entered into the warehouse, data are read-only, they cannot be changed or updated. Obsolete data are discarded, and changes are recorded as new data. This enables the data warehouse to be tuned almost exclusively for data access. For example, large amounts of free space (for data growth) typically are not needed, and database reorganizations can be scheduled in conjunction with the load operations of a data warehouse.
- *Summarized.* Operational data are aggregated, when needed, into summaries.
- *Not normalized.* Data in a data warehouse are generally not normalized and highly redundant.
- *Sources.* All data are present; both internal and external.
- *Metadata.* Metadata (defined as *data about data*) are included.

METADATA

We include a discussion of metadata in the data warehousing section because they have major impacts on how data warehouses function. As mentioned earlier, the term metadata refers to data about data. Metadata describe the structure of and some meaning about the data, thereby contributing to their effective or ineffective use.

Marco (2001) indicates that metadata hold the key to resolving the challenge of making users comfortable with technology. Executives realize that knowledge differentiates corporations in the information age. Metadata involve knowledge, and capturing and making them accessible throughout an organization have become important success factors. With metadata and a metadata repository, organizations can dramatically improve their use of both information and application development processes. Building a metadata repository should be mandatory for many organizations. Business metadata benefits include the reduction of IT-related problems, increased system value to the business, and improved business decision-making.

According to Kassam (2002), *business metadata* comprises information that increases our understanding of traditional (i.e., structured) data reported. The primary purpose of metadata should be to provide context to the data; that is, enriching information leading to knowledge. Business metadata, though difficult to provide efficiently, releases more of the potential of structured data. The context need not be the same for all users. In many ways, metadata assist in the conversion of data and information into knowledge (see Chapter 9). Metadata form a foundation for a *metabusiness* architecture (see Bell, 2001). Tannenbaum (2002) describes how to identify metadata requirements. Vaduva and Vetterli (2001) provide an overview of metadata management for data warehousing.

Semantic metadata are metadata that describe contextually relevant or domainspecific information about content, in the right context, based on an industry-specific or enterprise-specific custom metadata model or ontology. Basically, this involves putting a level of understanding into metadata. Text mining (Section 5.11) may be a viable way to capture semantic metadata. See Sheth (2003) for details. ADT Enterprise Metadata Edition from Computer Associates extends the capabilities of ADT (described in the *Data Access and Integration* subsection of Section 5.3) to include metadata management capabilities (see Whiting, 2002).

DATA WAREHOUSING ARCHITECTURE AND PROCESS

There are several basic architectures for data warehousing. Two-tier and three-tier architectures are quite common, but sometimes there is only one tier. McFadden, Hoffer, and Prescott (2003) distinguished among these by dividing the data warehouse into three parts:

1. The data warehouse itself, which contains the data and associated software
2. Data acquisition (back-end) software, which extracts data from legacy systems and external sources, consolidates and summarizes them, and loads them into the data warehouse
3. Client (front-end) software, which allows users to access and analyze data in the warehouse (e.g., a DSSIBI/BA engine)

In three-tier architecture, operational systems contain the data and the software for data acquisition in one tier (server), the data warehouse is another tier, and the third tier includes the decision support/business intelligence/business analytics engine (i.e., the application server) and the client. The advantage of this architecture is its sep-

58.

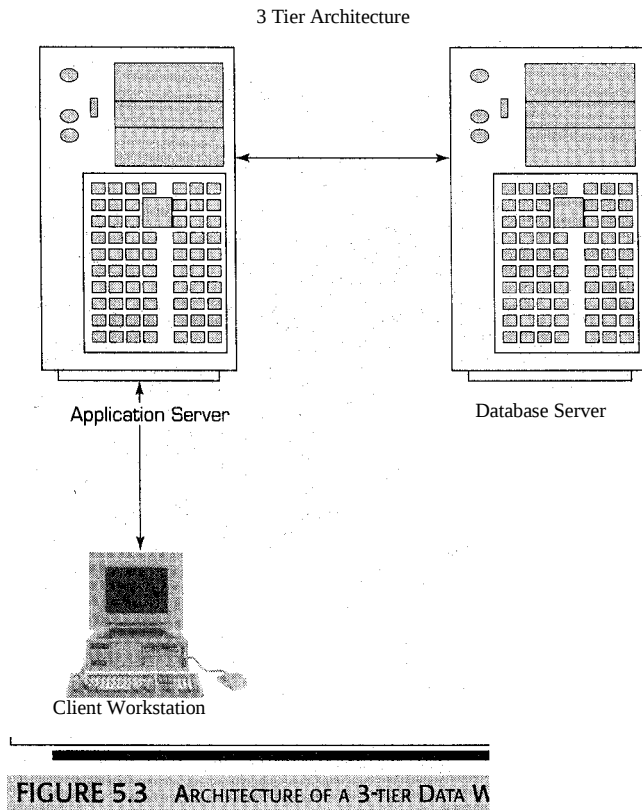


FIGURE 5.3 ARCHITECTURE OF A 3-TIER DATA W

aration of the functions of the data warehouse, which eliminates resource constraints and makes it possible to easily create data marts. See Figure 5.3.

The Vanguard Group moved to a Web-based three-tier architecture for its enterprise architecture to integrate all its data and provide customers with the same views of data as internal users (see Dragoon, 2003b). Likewise, Hilton migrated all of its independent client/server systems to a three-tier data warehouse using a Web design enterprise system. This change involved an investment of \$3.8 million (excluding labor) and affected 1500 users. It increased processing efficiency (speed) by a factor of 6. Hilton expects to save \$4.5 to \$5 million annually. Hilton plans to experiment with Dell's clustering technology next (see Anthes, 2003.)

In two-tier architecture, the DSS engine is on the same platform as the warehouse. Therefore, it is more economical than the three-tier structure. See Figure 5.4. See Mimno (1997) for more on data warehouse architectures.

Web architectures are similar in structure, requiring a design choice for housing the Web data warehouse with the transaction server or as a separate server(s). Page loading speed is an important consideration in designing Web-based applications; therefore server capacity must be carefully planned for.

There are several issues to consider when deciding which architecture to use. Among them are:

1. Which database management system to use? Most data warehouses are built using relational database management systems. Oracle (Oracle Corporation),

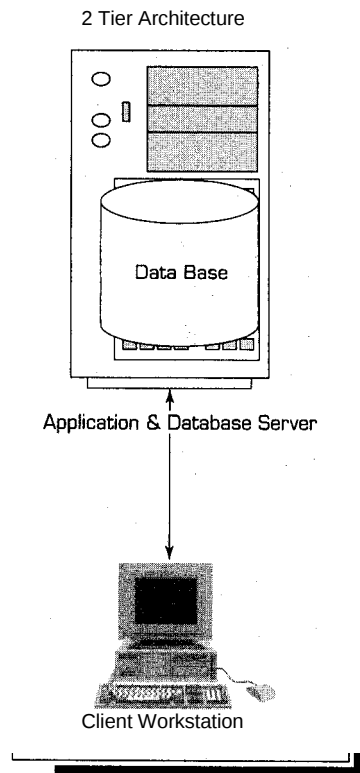


FIGURE 5.4 ARCHITECTURE OF A 2-TIER DATA WAREHOUSE

SQL Server (Microsoft), and DB2 (IBM) are most commonly used. Each of these products supports both client-server and Web-based architectures.

2. Will parallel processing and/or partitioning be utilized? Parallel processing enables multiple CPU's to process data warehouse query requests simultaneously and provides scalability. Data warehouse designers need to decide whether the database tables will be partitioned (split into smaller tables) for access efficiency and what the criteria will be. This is an important consideration that is necessitated by the large amounts of data contained in a typical data warehouse. Teradata has adopted this approach.
3. Will data migration tools be used to load the data warehouse?
4. What tools will be used to support data retrieval and analysis?

DATA WAREHOUSE DEVELOPMENT

A typical data warehouse structure is shown in Figure 5.2. The process of migrating data to a data warehouse involves the extraction of data from *all* relevant sources. Data sources may consist of files extracted from OLTP databases, spreadsheets, personal databases (e.g., Microsoft Access), or external files. Typically, all of the input files are written to a set of staging tables, which are designed to facilitate the load process. A data warehouse contains numerous business rules that define such things as how the data will be used, summarization rules, standardization of encoded attrib-

utes, and calculation rules. Any data quality issues pertaining to the source files need to be corrected before the data are loaded into the data warehouse. One of the benefits of a well-designed data warehouse is that these rules can be stored in a metadata repository and applied to the data warehouse centrally. This differs from an OUIP approach, which typically has data and business rules scattered throughout the system. The load process into a data warehouse can be performed either through data-transformation tools that provide a graphical user interface to aid in the development and maintenance business rule development or through more traditional methods by developing programs or utilities to load the data warehouse using programming languages such as PL/SQL, C++, or .Net. This decision does not come lightly for organizations. There are several issues that affect whether an organization will purchase data transformation tools or build the transformation process itself. These include:

1. Data transformation tools are expensive.
2. They may have a long learning curve.
3. It is difficult to measure how the IT organization is doing until it has learned to use the tools.

In the long run, a transformation-tool approach should simplify the maintenance of an organization's data warehouse. Transformation tools can also be effective in detecting and scrubbing; removing any anomalies in the data. OLAP and data-mining tools rely on how well the data are transformed.

STAR SCHEMAS

The data warehouse design is based upon the concept of dimensional modeling. Dimensional modeling is a retrieval-based model that supports high-volume query access. The star schema is the means by which dimensional modeling is implemented. A star schema contains a central *fact table*. A fact table contains the attributes needed to perform decision analysis, descriptive attributes used for query reporting, and foreign keys to link to dimension tables. The decision analysis attributes consist of performance measures, operational metrics, aggregated measures, and all other metrics needed to analyze the organization's performance. In other words, the fact table primarily addresses *what* the data warehouse supports for decision analysis. Surrounding the central fact tables (and linked via foreign keys) are dimension tables. Dimension tables contain attributes that describe the data contained within the fact table. Dimension tables address *how* data will be analyzed. Some examples of dimensions that would support a product fact table are location, time, and size. An example of a star schema is presented in Figure 5.5.

The grain of a data warehouse defines the highest level of detail that is supported. The grain will indicate whether the data warehouse is highly summarized or also includes detailed transaction data. If the grain is defined too high, then the warehouse may not support detail requests to drill down into the data. Drill down analysis is the process of probing beyond a summarized value to investigate each of the detail transactions that comprise the summary. A low level of granularity will result in more data being stored in the warehouse. Larger amounts of detail may impact the performance of queries by making the response times longer. Therefore, during the scoping of a data warehouse project, it is important to identify the right level of granularity that will be needed. See Tennant (2002) for a discussion of granularity issues in metadata.

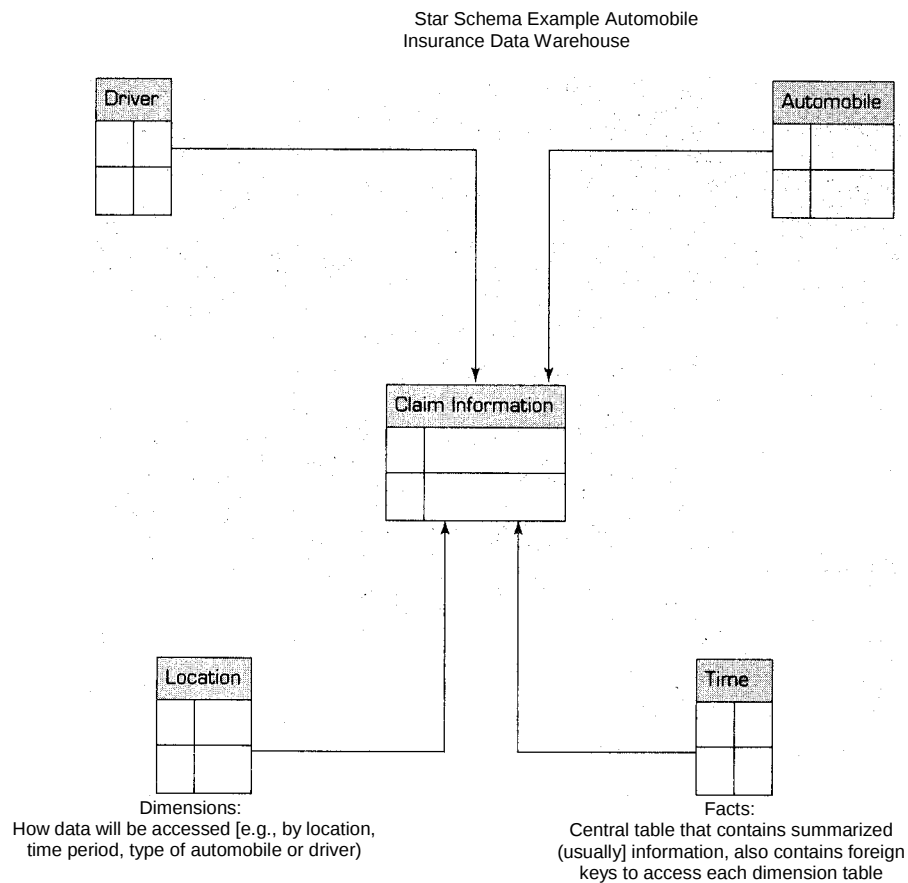


FIGURE 5.5 STAR SCHEMA

IMPLEMENTING DATA WAREHOUSING

Research by McKinsey and Co. indicates that much of the money invested in IT is wasted. IDC estimates that the world invested \$5.6 trillion in IT during the 1990s (\$2.6 trillion in the United States). IT investment had no impact on productivity in 53 of the 59 economic sectors of the McKinsey study. (We discuss IT effectiveness in Chapter 6.) However, McKinsey reports that IT investment can have an effective return on investment if applications are tied to specific business processes and linked to performance indicators (see Blair, 2003). This is critical in data warehouse and other large-scale database implementations. They must be useful, not just repositories of endless, useless data. They must drive business applications in ERP/ERM, "revenue management, SCM, CRM, and so on.

Implementing a data warehouse is generally a massive effort that must be planned and executed according to established methods. In Chapter 6, we discuss these methods in detail. Here we discuss specific ideas and issues as they relate to data warehousing. Eckerson (2002b, 2003) describes the four major ways to develop a data ware-

DSS IN FOCUS 5.17**THE FOUR MAJOR APPROACHES TO
BUILDING A DATA WAREHOUSE**

There are four major approaches to building a data warehousing environment: (1) top-down, (2) bottom-up, (3) hybrid, and (4) federated. Most organizations follow one or another of these approaches. In the top-down approach, the data warehouse is the center of the analytic environment. It is carefully designed and implemented. The design and implementation of all other aspects of business intelligence are based on it. This approach provides an integrated, flexible architecture to support later analytic data structures. In the bottom-up approach, the goal is to deliver business value by deploying multidimensional data marts quickly. Later these are organized into a data warehouse. The hybrid approach attempts to blend the first two

approaches. The federated approach is a concession to the natural forces that undermine the best plans for developing a perfect system. It uses all possible means to integrate analytical resources to meet changing needs or business conditions. Essentially, the federated approach involves integrating disparate systems (see the Opening Vignette and DSS in Action 5.7).

Sources: Adapted from Wayne Eckerson, "Four Ways to Build a Data Warehouse," *Application Development Trends*, May 2002, pp. 20-21; Wayne Eckerson, "Four Ways to Build a Data Warehouse," *What Works: Best Practices in Business Intelligence and Data Warehousing*, Vol. 15, The Data Warehousing Institute, Chatsworth, CA, June, 2003, pp. 46-49.

house. These include (1) top-down, (2) bottom-up, (3) hybrid, and (4) federated. We summarize these in DSS in Focus 5.17.

The federated approach is probably the least well known. Federation is often viewed as a form of information integration. It complements the traditional BTL and replication approaches by creating and maintaining a logical view of a single warehouse or mart, whereas the data reside in separate systems. See Devlin (2003) for details. One approach that is currently under development matches the notions that underlie peer-to-peer networks. Semantic Webs are used to wrap data into containers that reside in repositories in information space. This approach may be the solution to the massive data integration problem facing the Department of Homeland Security. See King (2003) for details.

Weir (2002) describes the best practices for implementing a data warehouse. We summarize these in DSS in Focus 5.18. Disaster may strike if one does not follow in the path of successful implementations. Adelman and Moss (2001) describe the risks confronting data warehouse projects. See DSS in Focus 5.19. Practitioners have unearthed a wealth of mistakes that have been made in the development of data warehouses. We summarize these in DSS in Focus 5.20. The three DSS in Focus boxes are, of course, interrelated. Watson et al. (1999) further discusses how such mistakes can lead to data warehouse failures.

Watson and Haley (1998) identified data warehouse projects as either data-centric or application-centric. A data-centric warehouse is based upon a data model that is independent of any application. It is designed to support a variety of user needs and applications. The methodological approach to designing a data-centric warehouse involves data modeling with a group of business experts who are familiar with the different information views needed to support the business. This consists of a top-down approach in producing specifications of information needs so as to not leave data behind. It is broad in scope and requires knowledge of current and anticipated data needs. A mapping approach should be used to provide a structured approach to classification of data. Data-centric warehouses should support flexibility because enterprise information constantly needs change based upon changes in the underlying business.

DSS IN FOCUS 5.18**BEST PRACTICES FOR DATA
WAREHOUSE IMPLEMENTATION**

Here is a list of best practices for implementing a data warehouse. They have been demonstrated in practice and constitute an excellent set of guidelines to follow.

- The project must fit with corporate strategy and business objectives.
- There must be complete buy-in to the project (executives, managers, users).
- Manage expectations.
- The data warehouse must be built incrementally.
- Build in adaptability.
- The project must be managed by both IT and business professionals.
- Develop a business/supplier relationship.
- Only load data that have been cleaned and are of a quality understood by the organization.
- Do not overlook training requirements.
- Be politically aware.

Source: Adapted from Robert Weir, "Best Practices for Implementing a Data Warehouse," *Journal of Data Warehousing*, Vol. 7, No.1, Winter, 2002, pp. 21-29.

The more dynamic the business, the greater the possibility that data needs will change during the development of the data warehouse. An application-centric warehouse is one initially designed to support a single initiative or small set of initiatives. This is a preferred approach for independent data mart development (see Section 5.8). The advantage of an application-centric approach is that it provides a more focused scope, and therefore increases the likelihood of successful data warehouse implementation. Its biggest disadvantage, however, is that critical data needs may be left out during the initial development, and therefore multiple iterations may be necessary.

DSS IN FOCUS 5.19**DATA WAREHOUSE RISKS**

There are many risks in data warehouse projects. Most of them are also found in other IT projects (see Chapter 6), but they are more serious here because data warehouses are large-scale, expensive projects. Each risk should be assessed at the inception of the project. See the source for information on details and how to mitigate the risks:

- No mission or objective
- Quality of source data is not known
- Skills are not in place
- Inadequate budget
- Lack of supporting software
- Source data are not understood
- Weak sponsor.
- Users are not computer literate
- Political problems, turf war
- Unrealistic user expectations
- Architectural and design risks
- Scope creep and changing requirements
- Vendors out of control
- Multiple platforms
- Key people may leave the project
- Loss of the sponsor
- Too much new technology
- Having to fix an operational system
- Geographically distributed environment
- Team geography, language culture

Source: Adapted from Sid Adelman and Larissa Moss, "Data Warehouse Risks," *Journal of Data Warehousing*, Vol. 6, No.1, Winter, 2001, pp. 9-15.

MISTAKES TO AVOID IN DEVELOPING A SUCCESSFUL DATA WAREHOUSE



When developing a successful data warehouse, watch out for these problems (see the explanations about each one):

1. *Starting with the wrong sponsorship chain.* You need an executive sponsor with influence over the necessary resources to support and invest in the data warehouse. You also need an executive *project driver*, someone who has earned the respect of other executives, has a healthy skepticism about technology, and is decisive but flexible. And you need an IS/IT manager to head up the project (the you in the project).
2. *Setting expectations that you cannot meet and frustrating executives at the moment of truth.* There are two phases in every data warehousing project: Phase 1 is the selling phase, where you internally market the project by selling the benefits to those who have access to needed resources. Phase 2 is the struggle to meet the expectations described in phase 1. For a mere \$1-7 million, you can hopefully deliver.
3. *Engaging in politically naive behavior.* Do not simply state that a data warehouse will help managers make better decisions. This may imply that you feel they have been making bad decisions until now. Sell the idea that they will be able to get the information they need to help in decision-making.
4. *Loading the warehouse with information just because it was available.* Do not let the data warehouse become a data landfill. This would unnecessarily slow down the use of the system. There is a trend toward real-time computing and analysis. Data warehouses must be shut down to load data in a timely way.
5. *Believing that data warehousing database design is the same as transactional database design.* In general, it is not. The goal of data warehousing is to access aggregates rather than a single or a few records, as in transaction-processing systems. Content is also different, as is evident in how data are organized. Database management systems tend to be nonredundant, normalized, and relational, whereas data warehouses are redundant, unnormalized, and multidimensional.
6. *Choosing a data warehouse manager who is technology-oriented rather than user-oriented.* One key to data warehouse success is to understand that the users must get what they need, not advanced technology for technology's sake.
7. *Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, and, perhaps, sound and video.* Data come in many formats and must be made accessible to the right people at the right time in the right format. They must be catalogued properly.
8. *Delivering data with overlapping and confusing definitions.* Data cleansing is a critical aspect of data warehousing. This includes reconciling conflicting data definitions and formats organizationwide. Politically, this may be difficult, because it involves change, typically at the executive level.
9. *Believing promises of performance, capacity, and scalability.* Data warehouses generally require more capacity and speed than is originally budgeted for. Plan ahead to scale up.
10. *Believing that your problems are over once the data warehouse is up and running.* DSS/business intelligence projects tend to evolve continually (see Chapter 6). Each deployment is an iteration of the prototyping process. There will always be a need to add more and different data sets to the data warehouse, as well as additional analytic tools for existing and additional groups of decision-makers. High energy and annual budgets must be planned for because success breeds success. Data warehousing never ends.
11. *Focusing on ad hoc data mining and periodic reporting instead of alerts.*

The natural progression of information in a data warehouse is

1. *Extract* the data from legacy systems, clean them, and feed them to the warehouse;
2. *Support* ad hoc reporting until you learn what people want; and then
3. *Convert* the ad hoc reports into regularly scheduled reports.

This may be natural, but it is not optimal or even practical. Managers are busy and need time to read reports. *Alert systems* are better and can make a data warehouse mission critical. Alert systems monitor the data flowing into the warehouse and inform all key people with a need to know as soon as a critical event occurs:

Source: Adapted from R. C. Barquin, A. Paller, and H. Edelstein, "Ten Mistakes to Avoid for Data Warehousing Managers," Chapter 7 in R. Barquin and H. Edelstein. (eds.). *Building, Using, and Managing the Data Warehouse*, Upper Saddle River, NJ: Prentice Hall PTR, 1997.

Wixom and Watson (2001) defined a research model for data warehouse success that identified seven important implementation factors that can be categorized into three criteria (organizational issues, project issues, and technical issues). The factors are:

1. Management support
2. Champion
3. Resources
4. User participation
5. Team skills
6. Source systems
7. Development technology

In many organizations, a data warehouse will only be successful if there is strong senior *management support* for its development and a *project champion* (see the best practices, risks, and mistakes described above). Although one might argue that this would be true for any information technology project, it is especially important for a data warehouse. The successful implementation of a data warehouse results in the establishment of an architectural framework that may allow for decision analysis throughout an organization and in some cases also provides comprehensive supplychain management by granting access to an organization's customers and suppliers. The implementation of Web-based data warehouses (*Webhousing*) has facilitated ease of access to vast amounts of data, but it is difficult to determine the *hard benefits* associated with a data warehouse. Hard benefits are defined as benefits to an organization that can be expressed in monetary terms. Many organizations have limited informationtechnology resources and must prioritize which projects will be worked on first. Management support and a strong project champion can help ensure that a data warehouse project will receive the resources necessary for successful implementation. Data warehouse *resources* can be a significant cost, in some cases requiring high-end processors and large increases in direct-access storage devices (DASD). Web-based warehouses may also have special security requirements to ensure that only authorized users have access to the data.

User participation in the development of data and access modeling is a critical success factor in data warehouse development. During data modeling, expertise is required to determine what data are needed, define business rules associated with the data, and decide what aggregations and other calculations may be necessary. Access modeling is needed to determine how data are to be retrieved from a data warehouse, and will assist in the physical definition of the warehouse by helping to define which data require indexing. It may also indicate whether dependent data marts are needed to facilitate information retrieval. The *team skills* needed to develop and implement a data warehouse require in-depth knowledge of the database technology and development tools utilized. **Source systems** and **development technology**, as mentioned previously, reference the many inputs and the process used to load and maintain a data warehouse.

MASSIVE DATA WAREHOUSES AND SCALABILITY

In addition to flexibility, a data warehouse needs to support scalability. The main issues pertaining to scalability are the amount of data in the warehouse, how quickly the warehouse is expected to grow, the number of concurrent users, and the complexity of user queries. A data warehouse must scale both horizontally and vertically. The warehouse will grow as a function of data growth and the need to expand the warehouse to support new business functionality. Data growth may be caused by the addition of current cycle data (e.g., this month's results) and/or historical data.

Hicks (2002) describes huge databases and data warehouses. In 2002, the Wal-Mart data warehouse was estimated to have a 200-terabyte capacity. The first petabyte-capacity data warehouse was made available in early 2004. Because of the storage required to archive its news footage, CNN plans to be one of the first organizations to install a petabyte-sized data warehouse (see Newman, 2002).

Given that the size of data warehouses is expanding at an exponential rate, *scalability* is an important issue. Good scalability means that queries and other data-access functions will grow (ideally) linearly with the size of the warehouse. In practice, specialized methods have been developed, to create scalable data warehouses. Nance (2001) describes scalability issues in data warehouse situations. Scalability is difficult in managing hundreds of terabytes or more. Terabytes of data have considerable inertia, occupy a lot of physical space, and require powerful computers. Some firms utilize parallel processing, others use clever indexing and search schemes to manage their data. Some spread their data across different physical data stores. As data warehouses approach the petabyte size, better and better solutions to scalability continue to be developed.

Deng (2003) describes the importance of effective indexing for data warehouses. Correct indexing can definitely lead to efficient searches through massive amounts of data. As a data warehouse is designed, it is important to consider correct indexing to help solve scalability problems. Hall (2002) also addresses scalability issues. Sears is an industry leader in deploying and utilizing massive data warehouses. See DSS in Action 5.21 for details.

USERS, CAPABILITIES, AND BENEFITS

Analysts, managers, executives, administrative assistants, and professionals are the major end-users of data warehouses. A data warehousing solution should provide ready access to critical data, insulate operation databases from ad hoc processing that can slow TPS systems, and provide high-level summary information as well as data drill-down capabilities. These properties can improve business knowledge, provide competitive advantage, enhance customer service and satisfaction, facilitate decision making, improve worker productivity, and help streamline business processes.

DATA WAREHOUSING APPLICATIONS

Allan (2001) provides an excellent example of a data warehouse. He addresses issues associated with the modeling of student record data for use in the student record data mart portion of a data warehouse for a college or university. Ryder uses its data warehouse for logistics. See DSS in Action 5.22.

DSS IN ACTION 5.21

THE SEARS DATA WAREHOUSE GROWS

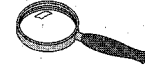


By April 2002, Sears, Roebuck and Co. had deployed 95 terabytes of new storage capacity, tripling its capacity. This allowed Sears to consolidate two key data warehouses and build a storage area network that handles its inventory and sales data warehouse with its customer information.

With the system, Sears can perform effective targeted promotional mailings. About 5,000 Sears employ-

ees use the data warehouse for analytical purposes. They can get daily product-sales information, analyze the purchases of individual customers, and correlate them with previous purchases.

Source' Adapted from Lucas Mearian, "Sears Triples Its Storage Capacity," *ComputerWorld*, February 28, 2002, pp. 1, 53.

DSS IN ACTION 5.22**RYDER RIDES INTO E-LOGISTICS**

With a new data warehouse, Ryder Systems Inc. has revamped its e-commerce strategy to match more than 1000 fleet customers and common carriers with freight that needs to be moved immediately. The effort is aimed at expanding Ryder's fleet-management supply-chain business. The system uses a transportation analytics package based on technology from NCR Corp.'s

Teradata data warehouse division and MicroStrategy Inc., a business analytics software vendor. The new system will let shippers place orders online and let carriers book orders in real-time. More is planned for the future.

Source: Adapted from Steve Konicki, "Ryder Trucks into New E-logistics Strategy," *Information Week*, June 11, 2001, p. 40.

DSS IN ACTION 5.23**WAL-MART IDENTIFIES AND MEETS UNEXPECTED CUSTOMER DEMAND THROUGH A DATA WAREHOUSE**

~

One instance of timely information being crucial to Wal-Mart took place after the attacks of September 11, 2001. Wal-Mart was able to quickly identify the buying patterns of its customers on the day of the attacks as the demand for weapons, bottled water, and survival gear increased, and then shifted to American flags the day afterwards. Wal-Mart was able to meet customer demand rapidly and could plan accordingly. It was able

to project that customers were delaying normal purchases for a few days, and expected and met the unusual higher demand afterwards.

Source: Adapted from C. Newman, "Teradata: Your Next Best Action with Your Customers," *Teradata Magazine*, Quarter 3, 2002.)

Wal-Mart is an undisputed leader in the data warehouse area. Westerman (2000) describes the effective Wal-Mart model. DSS in Action 5.23 is a small example of the effective use of Wal-Mart's data warehouse.

The major data warehouse vendors are Carleton, IBM, Informix, Microsoft, NCR, Oracle, Red Brick, and Sybase. For more on data warehousing, see Adelman and Moss (2001), Allan (2001), Barquin and Edelstein (1997a, 1997b), Barquin, Paller, and Edelstein (1997), Deng (2003), Eckerson (2002b, 2003), Edelstein (01997), Hall (2002), Konicki (2001), Mannino (2001), Mearian (2002), Mimno (1997), Mullin (2002), Nance (2001), Newman (2002), Watson and Haley (1998), Watson et al. (1999), Weir (2002), Westerman (2000), and Wixom and Watson (2001).

**5.8 DATA MARTS**

A **data mart** is a subset of the data warehouse, typically consisting of a single subject area (e.g., marketing, operations). A data mart can be either *dependent* or *independent*. A **dependent data mart** is a subset that is created directly from the data warehouse. It has the advantages of using a consistent data model and providing quality data. Dependent data marts support the concept of a single enterprise wide data model, but the data warehouse must be constructed first. A dependent data mart ensures that the end-user is viewing the same version of the data that is accessed by all other data warehouse users.

The high cost of data warehouses limits their use to large companies. As an alternative, many firms use a lower-cost, scaled-down version of a data warehouse referred to as an independent data mart. An independent data mart is a small warehouse designed for a strategic business unit (SBU) or a department, but its source is not an enterprise data warehouse.

The advantages of data marts include the following:

- The cost is low in comparison to an enterprise data warehouse (under \$100,000 vs. \$1 million or more).
- The lead time for implementation is significantly shorter, often less than 90 days.
- They are controlled locally rather than centrally, conferring power on the user.
- They contain less information than the data warehouse and hence have more rapid response and are more easily understood and navigated than an enterprise-wide data warehouse.
- They allow a business unit to build its own decision support systems without relying on a centralized IS department.
- An independent data mart can serve as a proof of concept prior to investing the resources needed to develop a comprehensive enterprise data warehouse. This will generate a quicker return on investment by realizing benefits sooner.

There are several types of data marts:

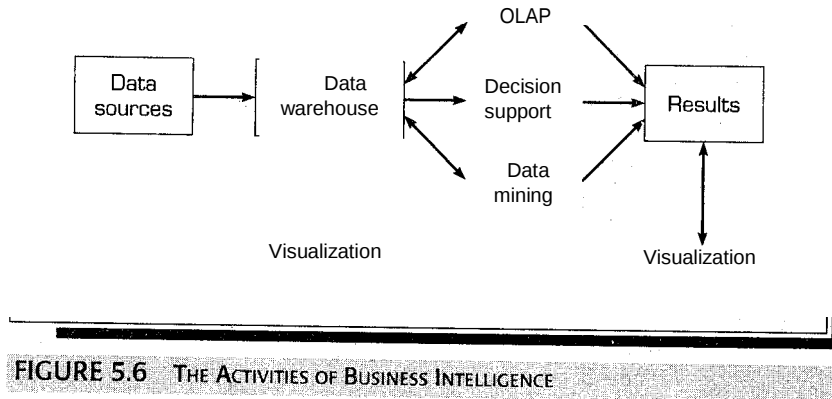
1. Replicated (dependent) data marts. Sometimes it is easier to work with smaller parts of the warehouse. In such cases one can replicate functional subsets of the data warehouse in smaller databases, each of which is dedicated to certain areas, as shown in Figure 5.2. In this case the data mart is an addition to the data warehouse.
2. Independent data marts. A company can have one or more independent data marts without having a data warehouse. In such cases there is a need to integrate the data marts. This is possible only if each data mart is assigned a specific set of information for which it is responsible. The IS department specifies the rules to the metadata so that the information kept by each mart is compatible with that provided by all the other marts. When this is not done, the data marts are difficult to integrate, creating potentially serious fragmentation problems for the organization.

5.9 BUSINESS INTELLIGENCE/BUSINESS ANALYTICS

Now that we know about databases, data warehouses, data marts, and the analytical decision-making methods discussed in Chapter 4, we are ready to discuss business intelligence/business analytics intelligently.

Business intelligence describes the basic architectural components of a business intelligence environment, ranging from traditional topics, such as business process modeling and data modeling, to more modern topics, such as business rule systems, data profiling, information compliance and data quality, data warehousing, and data mining (see Loshin, 2003).

Business intelligence involves acquiring data and information (and perhaps even knowledge, see Chapter 9) from a wide variety of sources and utilizing them in decision-making. Technically, business analytics adds an additional dimension to business intelligence: models and solution methods. These are often buried so deep within the tools, however, that the analyst need not get his or her hands "dirty." Typically, the



terms are used interchangeably. We show the activities of business intelligence in Figure 5.6. Business intelligence methods and tools are highly visual in nature. They provide charts and graphs of multidimensional data with the click of a mouse. These methods generally access data from data warehouses and deposit them into a local, multidimensional database system. *Online analytical processing (OLAP)* methods allow an analyst, or even (less typically) a manager to slice and dice the data, while observing graphs and tables that reflect the dimensions being observed. Models may be applied to the data for forecasting or to identify opportunities (for software examples, see Temtec Executive Viewer, Cognos Impromptu and PowerPlay, and IBM Cube Views). **Data mining** methods apply statistical and deterministic models, and artificial intelligence methods to data, perhaps guided by an analyst (or manager), to identify hidden relationships or induce/discover knowledge among the various data or text elements (for software examples, see IBM DB2 Intelligent Miner Scoring, Angoss KnowledgeSeeker, Megaputer Intelligence PolyAnalyst, and SAS Enterprise Miner). Data mining is also highly visual in the way results are displayed. Graphs and charts typically display results. Thus the key difference between OLAP and data mining is that data mining runs (mostly) automatically, while OLAP is driven. As tools improve in ease of use, more and more managers utilize them, resulting in a trend to move business intelligence from the analyst to the user (manager). This introduces a new problem: Managers sometimes do not fully understand business intelligence/business analytics methods. In consequence, their focus may be on visualization rather than application of appropriate and accurate analysis tools. With both tools, it is important to recognize that systems analysts are generally required to set up the access to the data to be analyzed. This involves dealing with data cleansing and integration, a task best left to IS specialists. See the Opening Vignette and DSS in Action 5.7.

All managers and executives should be using business intelligence systems, but some find the data irrelevant or the tools too complicated to use. Sometimes managers are not trained properly. Distributing information from analytics throughout a company is a major challenge; most businesses want a greater percentage of the enterprise to leverage analytics, but most of the challenges around technology involve culture, people, and processes (see Hatcher, 2003). A critical issue is to align BI systems to business needs. If the system does not provide useful information, it is considered useless. See DSS in Focus 5.24 for details of a recent study on how executives currently utilize business intelligence tools.

DSS IN FOCUS 5.24

ARE BUSINESS INTELLIGENCE SYSTEMS MAKING FIRMS SMARTER?

~|

More than 570 IT executives responded to *CIO Insight's* Business Intelligence Research Study. *CIO Insight* discovered some interesting facts about the current state of business intelligence.

- Most notably, the use of business intelligence technologies is high, and growing.
- Larger companies are somewhat more likely than smaller companies to use BI.
- In 2002, successful companies spent almost 50 percent more on BI technology than unsuccessful companies. BI seems to be necessary (but not sufficient) for success.
- The government utilizes virtually every market intelligence technology at significantly higher rates than any other sector of the economy.
- The technologies used to collect, aggregate, analyze, and report on competitive intelligence along with the percentage response in parentheses are: reporting tools (82.1), automated data/information

feeds (79), intra nets/portals (70.4), data warehousing (69.8), content management (63), data-visualization software (41.4), specialty search engines (41.4), work-flow software (41.4), and harvesting (e.g., intelligent agents) (38.9).

- Just 49 percent of less successful companies are happy with their competitive intelligence efforts.
- Some 88 percent of companies have confidence in the accuracy of the customer information they gather.
- Dissatisfaction with BI usually derives from difficulty in distributing the results.
- CIOs want to move firms to the real-time enterprise.

Source: Adapted from "The 2003 CIO Insight Business Intelligence Research Study: Are Your BI Systems Making You Smarter?" *CIO Insight*, No. 26, May 23, 2003.

In the first 50 years of computing history, computing systems have had a deep and comprehensive infusion into various business domains. Computing systems are now an indispensable infrastructure with which we run, manage, and coordinate business operations. In the first decade of the new millennium, we see a new era of ubiquitous computing systems. Analytics will interweave most, if not all, enterprise systems (Delic and Dayal, 2003). Decision-makers throughout every enterprise need an IT architecture that serves their needs, rather than the other way around. Delic and Dayal (2003) provide an impressive view of emerging enterprise analytic systems (see Chapter 8) that use business intelligence/business analytics requirements as their basis.

According to an IDC report issued in the fall of 2002, organizations that have successfully implemented and used analytic applications have realized returns ranging from 17 percent to more than 2000 percent, with a median ROI of 122 percent ("The Financial Impact of Business Analytics," IDC, October 2002; also see Kaliebe, 2003). Even so, more than half of all business intelligence projects fail. As with data warehousing, business intelligence activities should be regarded, not simply as another set of IT projects, but as a constantly evolving strategy, vision, and architecture that continuously seeks to align an organization's operations and direction with its strategic business goals. We discuss the notion that BI/DSS are never really complete in Chapter 6. They continue to evolve. Companies achieve success when they do the following (see Atre, 2003):

- Make better decisions with greater speed and confidence.
- Streamline operations.
- Shorten product development cycles.

- Maximize value from existing product lines and anticipate new opportunities.
- Create better, more focused marketing as well as improved relationships with customers and suppliers.

Organizations must understand and address many critical challenges for business intelligence success. We describe these in DSS in Focus 5.25.

Business intelligence tools (both data mining and OLAP) have been used to identify white-collar theft in organizations. They are able to identify inflated invoices, embezzlement, customer impersonation, and similar offenses. The estimate of total fraud in the United Kingdom is almost \$30 billion (U.S.). Frauds committed by employees cause median losses of \$60,000, while frauds committed by managers or executives cause median losses of \$250,000. When managers and employees conspire, the median loss rises to \$500,000. If all internal data systems are integrated with a data warehouse for fraud analysis so they can be compared to external fraud-related data. Patterns and anomalies become more readily identifiable. Suspicious activities can be isolated, measured, and tracked. See Dorrington (2003) for details.

Williams-Sonoma saves millions with targeted marketing, multichannel branding using the SAS data mining software, Enterprise Miner, along with a suite of CRM applications from SAS. The new marketing system models and explores customer data from more than 30 million households to help the retailer create a personalized, cohesive shopping experience across multiple channels and multiple brands. See Bolen (2003) for details. Callaghan (2003a) describes how SPSS Predictive Web Analytics and SAS Web can be utilized to predict customer Web behavior and develop customer segmentation models (clusters) that lead to better business performance. Retailers frequently use business intelligence tools, as we show in DSS in Action 5.26.

New forms of business intelligence continue to emerge. Performance management systems (PMS) are one of the new forms. These are business intelligence tools that provide scorecards and other relevant information with which decision-makers can determine their level of success in reaching their goals. Two tools include Business

DSS IN FOCUS 5.25

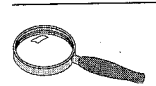
TEN CRITICAL CHALLENGES FOR BUSINESS INTELLIGENCE SUCCESS



There are 10 reasons why business intelligence projects fail. Organizations must understand and address these 10 critical challenges for success:

1. Failure to recognize BI projects as cross-organizational business initiatives, and to understand that as such they differ from typical standalone solutions.
2. Unengaged or weak business sponsors.
3. Unavailable or unwilling business representatives.
4. Lack of skilled and available staff, or suboptimal staff utilization.
5. No software release concept (no iterative development method).
6. No work breakdown structure (no methodology).
7. No business analysis or standardization activities.
8. No appreciation of the impact of dirty data on business profitability.
9. No understanding of the necessity for and the use of metadata.
10. Too much reliance on disparate methods and tools.

Source: Adapted from Shaku Atre, "The Top 10 Critical Challenges for Business Intelligence Success." *ComputerWorld, White Paper/Special Advertising Supplement*, Vol. 37, No. 26, June 30, 2003.

DSS IN ACTION 5.26**RETAIL MAKES STEADY BUSINESS INTELLIGENCE PROGRESS**

Hudson's Bay Co. turned 333 in May 2003. Despite its age, Hudson's Bay upgraded its information systems to give executives, store managers, and key suppliers methods to analyze reams of sales and customer data. The challenge the firm faces is to determine how to transform the data into useful information. The firm uses two data warehouses and business intelligence tools from the Teradata division of NCR Corp. to track and make decisions on product inventory and sales.

Most brick-and-mortar retailers lag other industries in business intelligence. Notable exceptions include Wal-Mart Stores Inc. and Sears. Other retailers continue to make impressive strides.

At Harry Rosen Inc., a chain of 17 men's clothing stores, executives use Cognos Inc.'s data analysis tools integrated into a merchandising system. There are more than a dozen sales and inventory reports for analyzing sales that help the firm identify sales trends, manage inventory, and improve gross profit margins.

Other retailers are looking for similar ways to obtain a competitive edge. Putting the right products in the right place at the right time at the right price (see revenue management in Chapter 4) is the goal of retail-

ers. Doing it right determines who succeeds, and who fails.

Using business intelligence and analysis tools from BusinessObjects SA, TruServ Corp. (the parent company of True Value Hardware and Taylor Rental) reduced its "red zone" inventory (products that have not sold in one-half year) by \$50 million over two years by analyzing product stockpiles. For about a year, the system has also identified products sitting in its 14 distribution centers that might sell better in other parts of the country.

Stores are learning from online retailers about how to perform analytic investigations of customer performance. For example, 1. Crew Group and Nordstrom Inc. use DigiMine to analyze online sales. Nordstrom had a situation where online shoppers were searching for navel rings just like the one that a model wore in an advertisement. Nordstrom was able to quickly obtain the rings for both its stores and online customers, even though it had not carried the product beforehand.

Source: Adapted from Rick Whiting, "Business-Intelligence Buy-In," Information Week, May 12, 2003, pp. 56-60.

Objects S.A. Performance Manager and the SAS Institute, Inc. Activity-Based Management 6.0. Both provide Web-enabled analysis. Business Objects S.A. Performance Manager software includes goal management, scorecards, and strategy maps, and enables customers to align goals and performance with enterprise strategy. The SAS Institute, Inc. Activity-Based Management 6.0 provides cost and productivity analysis of customers, products, services, or business processes. See Callaghan (2003c) for details.

C

DASHBOARDS

Dashboards provides managers with exactly the information they need in the correct format at the correct time. Business intelligence systems are the foundation of dashboards, which have evolved from executive information systems into enterprise information systems that access data warehouses via oLAP systems (see Chapter 8 and Leon, 2003). Dashboards can impact on communications and company politics. Dashboards and scorecards measure and display what is important. Each individual, ideally, can focus on what is important to him or her. Essentially a dashboard is a preset oLAP display. See Figure 5.7 for an example of Brio Performance Suite's "CEO Dashboard." MQSoftware, Inc.'s Q Pasa! Business Dashboard provides real-time views of data. Cognos Visualizer Series 7 is another example of a corporate dashboard that helps give managers insight to make better decisions. It is part of Cognos Business

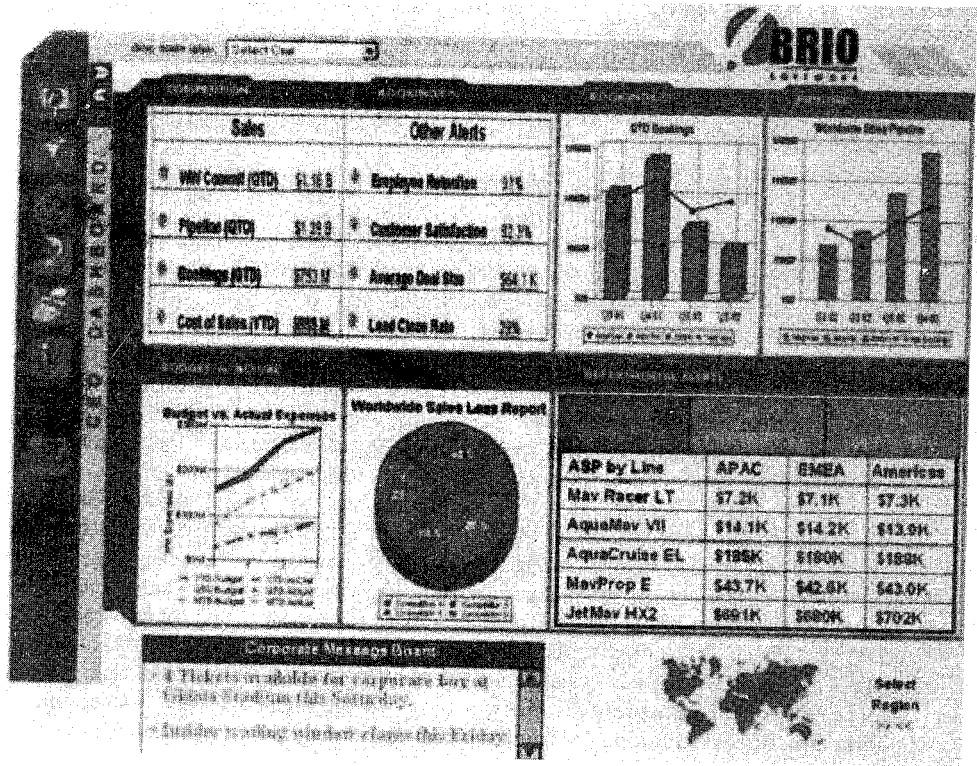


FIGURE 5.7 SCREEN SHOT FROM BIO PERFORMANCE SUITE DIGITAL DASHBOARD

Courtesy of Brio Software Inc.

Intelligence. Business intelligence dashboards have spread to various nonfinancial departments of firms, including sales and customer service. See Table 5.4 for details of how dashboards have spread through organizations.

At Southwest Airlines, they call digital dashboards *cockpits*. Individuals get customized views of the information they need for their work. At Honeywell Inc.'s Specialty Materials Division in Morristown, New Jersey, Cognos Inc. dashboards give everyone in sales a clear view into daily business performance. Sales representatives can see their own sales statistics, but they Can also see how others are doing, as can managers. This has led not only to a move from monthly and quarterly data views to daily views. Now the firm has a common definition and view of all information.

TABLE 5.4 epartments with Dashboards

Department	Percent
Sales	21.1
Finance	18.8
Customer service	14.3
Manufacturing/operations	12
Supply chain management	10.5
Human resources	8.3

Source: Adapted from M. Leon, "Dashboard Democracy," *Computer World*, June 16,2003.

DSS IN FOCUS 5.27**BUSINESS INTELLIGENCE ASSESSMENT**

A business intelligence assessment is a low-cost, actionable examination of the three areas critical to the implementation of any business intelligence initiative:

- *Business needs analysis:* Analyze the underlying strategic and tactical business goals and objectives that are driving the development of the BI solution, including whether executive sponsorship and funding are available.
- *Organizational analysis:* Analyze the existing business and technical organizational structures, including the level of IT/business partnering in place, the organization's culture and leadership style, its understanding of BI concepts, whether roles and responsibilities have been established, and whether people with the appropriate amount of time and skills are in place.

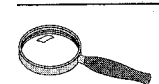
"Technical/methodology analysis: Analyze whether the appropriate technical infrastructure and development methodologies are in place, including all related hardware and software, the quality and quantity of the source data, and the methodology and change-control process.

The assessment forces an organization to examine strengths and weaknesses within these three areas and makes recommendations about how to fix potential problem areas. Ideally perform such an analysis before developing a costly set of systems, including data warehouses, OLAP, and data mining. The assessment itself helps build awareness and support for the initiative.

Source: Adapted from T. Burzinski, "The Case for Business Intelligence Assessments," *DM Review*, July 2002.

Burzinski (2002) recommends performing a business intelligence assessment before implanting any business intelligence initiative. See DSS in Focus 5.27 for details. The development of business intelligence and data warehousing initiatives over the last decade has led to many issues and their solution. We describe critical lessons learned in DSS in Focus 5.28.

The Web has had a profound impact on how these tools function and what they are utilized for. The visual nature of most business intelligence tools is often based on Web-browser interfaces. As Web use and e-commerce increase, there is more of a demand for gathering and analyzing data from the *clickstream*, to identify where cus-

DSS IN FOCUS 5.28**CRITICAL LESSONS IN BUSINESS INTELLIGENCE AND DATA WAREHOUSING**

The first 10 years of business intelligence and data warehousing initiatives have resulted in many successful, high-return applications of information technology. Here are some critical lessons that should be followed and examined to help ensure success:

- Create stability in the basic structures of data fundamental for providing business intelligence and running the business.
- Ensure that each data element stands on its own as a fact or attribute.
- Keep an enterprise-wide focus, not a departmental, regional, or other category focus.
- Make business intelligence not simply the analytical report, but the information a manager or executive needs to make informed decisions.
- Use several different business intelligence technologies that integrate well.

Source: Adapted from Richard Skrlitz, "New Directions for Business Intelligence," *DM Review*, April 2002, p.10.

59.

tomers go on a Web site, where they came from, where they go afterwards, and what they buy or don't buy. (These systems are often called Web intelligence/Web analytics; see Section 5.14.) Combining this with census data and geographic information systems, a firm can identify what to target market to existing and potential new customers. We indicate database and business intelligence technologies and Web impacts in Table 5.5.

Kurtyka (2003) discusses issues that relate to organizational learning and business intelligence. Smith (2001) describes a method for the strategic assessment of business intelligence tools. He provides an analysis of a large sample of tools. See Smith (2001) for details. Determining which tool to use has significant consequences on the decision analysis features that will be supported. We have purposely separated data mining tools from the OLAP discussion. Topics and issues pertaining to OLAP, data mining, and the Web are discussed in the remainder of the chapter.

For more on business intelligence, see Bolen (2003), Burzinski (2002), Callaghan (2003a), Delic and Dayal (2003), Dorrington (2003), Kaliebe (2003), Kurtyka (2003), Leon (2003), Loshin (2003), Pallatto (2002a,b), Smith (2001), Songini (2003), Ulfelder (2000a), Vitt, Luckevich, and Misner (2002), and Whiting (2003).

TABLE 5.5 Database and Business Intelligence Technologies, and Web Impacts

<i>Knowledge Management</i>	<i>Web Impacts</i>	<i>Impacts on the Web</i>
Databases	Consistent, friendly, graphical user interface Web database servers provide efficient and effective data storage and retrieval Convenient, fast and direct access to data on servers Multimedia data storage and retrieval expectations have become a reality Developments in search engines are directly applicable to database technologies Same as above	Data captured and shared are utilized in improving Web site design and performance Web servers are developed and sold specifically for database applications
Data warehouse and data mart	Distributed properties of Web servers have led to distributed data warehouses and data marts The distributed properties have led to improvements in data integration Improvements in technology help solve scalability problems Same as above	Same as above Led to the proliferation of Web technologies to provide massive communication for data warehouse use
OLAP	Here the Web-based graphics are critical to understanding results Access to analytical models and methods to solve business, engineering and other problems Same as above	Same as above Improvements in Web e-commerce and other sites Improvements in Web/Internet technologies
Data mining	Helps to automate the analytical methods	Same as above

ANALYTICAL PROCESSING (OLAP)-----5.10 ONLINE

For many years IT concentrated on building mission-critical systems that mainly supported corporate transaction processing. Such systems must be virtually fault-tolerant and provide rapid response. An effective solution was provided by online transaction processing (OLTP), which centers on a distributed relational database environment. The latest developments in this area are the utilization of ERP and SCM software for transaction processing tasks, CRM applications, and integration with Web-based technologies and intranets. Many tools were created for developing OLTP applications- the INFORMIX OnLine Dynamic Server (informix.com) is an-example of an effect;tiy~ tool.

Access to data is often needed by both OLTP and MSSlippliatidns.

Unfortunately, trying to serve both types of requests may be problematic (Gray and Watson, 1998). Therefore, some companies elect to separate databases into OLTP types and OLAP types. The OLAP type is based on the data warehouse.

Even so, Gonzales and Robinson (2003) indicate that for OLAP to work properly, the relational database management system must/be optimized to support OLAP instead of directly utilizing pure, multidimensionaf data cubes. The database must be integrated with the centralized, cohesive, and consistent control of.nfultidimensional data across the enterprise. To make the database aware of the higher-level data organization OLAP requires, the database catalogs need a set of higher-level objects that relate directly to OLAP and business models. In effect, these objects will take the existing atomic entities and compound them to make dimensional entities, such as attributes, facts, relationships, hierarchies, and dimensions. Once these high-level objects are defined, the new information can be stored and managed as part of the catalogs (see Gonzales and Robinson, 2003). In effect, managing metadata becomes part of a relational database management system in order to make it "OLAP aware."

The term online analytical processing (OLAP) refers to a variety of activities usually performed by end-users in online systems (see DSS in Action 5.29). There is no

60.

DSS IN ACTION 5.29

OUTOKUMPU COPPER PRODUCTS



Outokumpu Copper Products (Finland) processes millions of kilograms of base metals each year that are used in such products as belt buckles, drinking water tanks, and radiators. Its products serve industrial, electronic, mining and metal, transport, and construction companies worldwide. Outokumpu has four divisions with 13 business lines operating independently in Europe, America, and Asia. Determining the profit margin of a product, how to lower production costs, customer turnover, and profitability are critical business issues. Integrating and analyzing information from each business unit presented a tremendous challenge, because the markets in the different divisions operate with such specific requirements that a generic information system did not seem possible. The solution was to build a data warehouse and use a Web-based OLAP

server. The database utilized Hyperion's Essbase. Two models were quickly built to support customer profitability and products, and one to report on the delivery performance. They used Executive Viewer, an OLAP front-end from Temtec Corporation "(tty the online demo at www.temtec.com). Executive viewer is Web-based and integrates with numerous databases including Hyperion Essbase. It allowed the company to develop an application to quickly select information by market, product, or customer and supports drill-through analysis. The results have been the implementation of a flexible analytical tool that has met widespread acceptance throughout the organization.

. Source: Condensed from Case: "Outokumpu Copper Products-Finland," www.temtec.com.2003.

DSS IN ACTION 5.30 : ... ;, "' .. C.'

COGNOS OLAP TOOLS BENEFIT ALLIED BUILDING



Allied Building Products Corporation has grown its building materials distribution company by increasing market share and customers. Allied implemented *Cognos Finance*, a business intelligence (BI) solution from Cognos, in under 90 days. As a result, Allied was able to standardize data company-wide and automate processing to deliver a single coordinated view of financial performance. Cognos Finance allows Allied to reduce manual labor in producing reports, and at the

same time provides access to budgets, forecasts, and actuals across all of its branch operations. This enabled an integrated view of information and accelerated accurate financial reporting.

Source: Adapted from "Allied Building Products Corporation Monitors Corporate Performance with Cognos Finance," Cognos, Inc., cognos.com, 2003.

agreement on what activities are considered OLAP. Usually one includes such activities as generating queries (see DSS in Action 5.30), requesting ad hoc reports and graphs, conducting statistical analyses, and building DSS and multimedia applications. Some include executive and/or enterprise information systems and data mining. Essentially, OLAP provides modeling and visualization capabilities to large data sets, either to database management systems or, more often, data warehouse systems. OLAP is different from data mining in that users can ask specific, open-ended questions. Users, typically analysts, run OLAP systems. They drive OLAP, whereas data mining looks for relationships, with some direction from the analyst. OLAP is generally facilitated by working with the data warehouse (or with the data mart or a multidimensional warehouse) and with a set of OLAP tools. These tools can be **query tools**, spreadsheets, data mining tools, data visualization tools, and the like. For a list of OLAP tools, see Costanza (2000), Alexander (2003), Karpinski (1999), and periodic reviews in the software sections of *PCWeek*, *DM Review*, *InternetWeek*, *Intelligent Enterprise*, and *Software Review*. The major vendors of these tools include BusinessObjects S.A., Brio Software Inc., Computer Associates, Cognos Inc., Comshare, Crystal Decisions Inc., Hyperion Software Corporation, Informatica Corp., Information Builders, **IBM**, Intersolve, Microstrategy Corporation, Oracle, SAS Institute Inc., Software A&G, and Temtec. The multidimensional nature of almost all OLAP systems is described in Section 5.12.

ING Antai Insurance Co. (Taiwan) uses IBM's OLAP Server to analyze large amounts of data to detect fraudulent claims and speed up the processing of claims. It takes only a couple of days to analyze data that previously took several weeks. As the tools and hardware improve, claims can be analyzed instantaneously. The cost of processing claims is greatly reduced. See Raden (1997). TCF Bank utilizes OLAP to provide information directly to those who need it. The bank now understands its customer base more accurately for targeting its marketing efforts. See DSS in Action 5.31 for details. We cover more material in Section 5.11 and in Case Application 5.3 on how banks utilize business intelligence/business analytic tools for similar tasks.

SQL FOR QUERYING

Structured query language (SQL) is a standard data language for data access and manipulation in relational database management systems. It is an English-like language consisting of several layers of increasing complexity and capability. SQL is used for online access to databases, DBMS operations from programs, and database admin-

DSS IN ACTION 5.31**TCF BANK'S OLAP AND DATA MINING**

One of the largest regional banks in the midwestern region of the United States, TCF Bank has more than 390 branches in six states and serves customers from all income groups. TCF Bank has \$12.2 billion in assets and operates the fourth-largest supermarket branch-banking system in the country. TCF focuses on being a convenient one-stop shop for customers; it is one of the few banks in the United States that is open 12 hours per day, seven days per week, including holidays.

Users in the bank's major groups (retail banking, consumer loans, mortgage banking, brokerage) found that the IT reports were not meeting their needs. Instead, they had to develop custom processes to download files from IT and then load the data into spreadsheets for further analysis. The time required to create a standard graph report was close to a month. It might take six weeks to generate a customer marketing list.

The information management department needed to come up with a new process to enable users to gain customer insight so as to uncover opportunities and

effectively offer new services to customers. TCF adopted Informatica PowerCenter and PowerAnalyzer in mid-2002. PowerAnalyzer's report-creating wizard, metrics-based reporting, and analysis-path drill-down features were important ease-of-use functions in the adoption decision. A number of key-indicator starter reports for user dashboards were developed. In a week, 550 loan officers and executives were using these and other reports on a daily basis.

With the new OLAP system, which includes a cross-sell application, TCF is able to identify classes of customers to approach with specific services and products. This is especially critical in identifying the needs of new customers. In addition, reports are generated immediately, so further analysis can be performed.

Source: Adapted from Ted Ledman, "TCF Bank," *What Works: Best Practices in Business Intelligence and Data Warehousing*, Vol. 15, The Data Warehousing Institute, Chatsworth, CA, June 2003, p. 10.

istration functions. It is also used for data access and manipulation functions of some leading DBMS software products (e.g., Oracle, IBM's DB2, Ingres).

Since SQL is nonprocedural and fairly user-friendly, many end-users can use it for their own queries and database operations. SQL can be employed for programs written in any standard programming language; thus, it facilitates software integration. Support of DSS/business intelligence is accomplished in the warehouse with products from vendors such as Brio, BusinessObjects, Cognos, Pilot Software, and SAS. SQL is a fairly simplistic OLAP tool. The real strength of OLAP is in its major analytic capabilities.

OLAP TOOLS

Using SQL and other conventional data access and analysis tools is helpful, but not sufficient, for OLAP. In OLAP a special class of tools is used, known as decision support/business intelligence/business analytic front ends, data-access front ends, database front ends, and visual information access systems. These methods go well beyond spreadsheets in power and results. They tools are intended to empower users.

OLAP tools have characteristics that distinguish them from reporting tools designed to support traditional OLTP reporting applications. The characteristics of OLAP tools were succinctly defined by E. F. Codd and associates (1993); Codd is considered to be the "inventor" of the relational database model. The twelve rUles for OLAP tools are summarized in Table 5.6 (see Raden, 1997). They defined four types of processing that are performed by analysts within an organization:

1. Categorical analysis is a static analysis based upon historical data. It is based upon the premise that past performance is an indicator of the future. This is the primary analysis supported by OLTP transaction-based databases.

TABLE 5.6 OLAP Product Evaluation Rules: Codd's Twelve Rules for OLAP

- Multidimensional Conceptual
- View Transparency
- Accessibility
- Consistent Reporting Performance
- Client-Server Architecture
- Generic Dimensionality
- Dynamic Sparse Matrix
- Handling Multi-User Support
- Unrestricted Cross-dimensional Operations
- Intuitive Data Manipulation
- Flexible Reporting
- Unlimited Dimensions and Aggregation Levels

Source: Adapted from "Providing OLAP to User-Analysts: An IT Mandate," Codd & Associates, White Paper, hyperion.com. Also see Radin (1997).

- 2. Exegetical analysis is also based upon historical data, adding the ability to perform drill down analysis. Drill down analysis is the ability to query further into the data to determine the detail data that were used to determine a derived value.
- 3. Contemplative analysis allows a user to change a single value to its impact.
- 4. Formulaic analysis permits changes to multiple variables.

Vendors in the BI arena are maneuvering to empower end users with the ability to customize analytic applications to meet evolving business needs. These include Spotfire (DecisionSite analytics platform), Business Objects (Enterprise BI Suite), and QlikTech (QlikTech). See Haverstein (2003b).

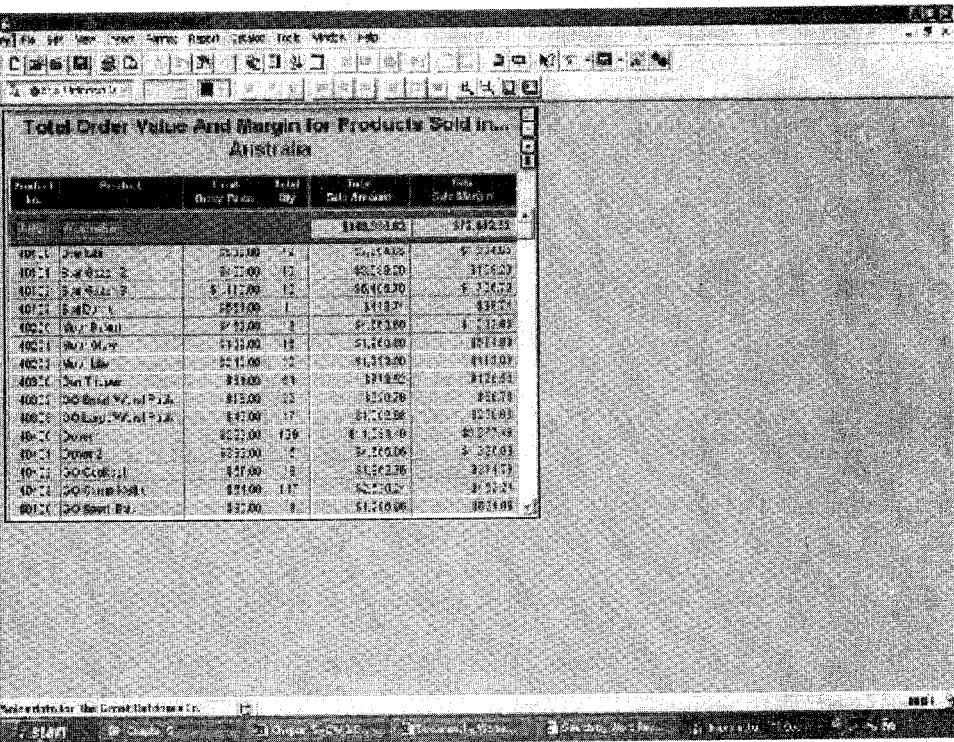
There are hundreds of OLAP tools available today. They share many features but also provide some distinct differences (see DSS in Action 5.30).An example of OLAP output is shown in Figure 5.8, the result of a Cognos Impromptu Version 7.0 query.

Temtec Executive Viewer provides all the expected OLAP features, including multidimensional views of data, dimension expansion and collapse, dynamic column selection,automatic calculations (sums, etc.), automatic charting and graphing, physical rnaps to display data (Figure 5.9), and instantaneous drilldown and rollup. Executive Viewer utilizes the idea of traffic lights in its display of data. We show an example of this in Figure 5.10. Note the shading in the last, variance column. The value of -2.57 is shaded in red, indicating "stop, there is a problem"; the 4.64 is shaded in yellow indicating "caution, a problem may be developing"; and the rest of the numbers are shaded in green, indicating that it is safe to proceed forward. Try Executive Viewer online (www.temtec.com) using the demo book (where data are integrated into a multidimensional framework).

We show screen shots of the Brio Performance Suite in Figure 5.11 and Figure 5.12. The Suite's Web-based drag and drop capabilities, which show how to construct a report by dragging rows/columns for analysis, are depicted in Figure 5.11. The Performance Suite reporting with OLAp capabilities is depicted in Figure 5.12. Note the mixture of graphs and tables.

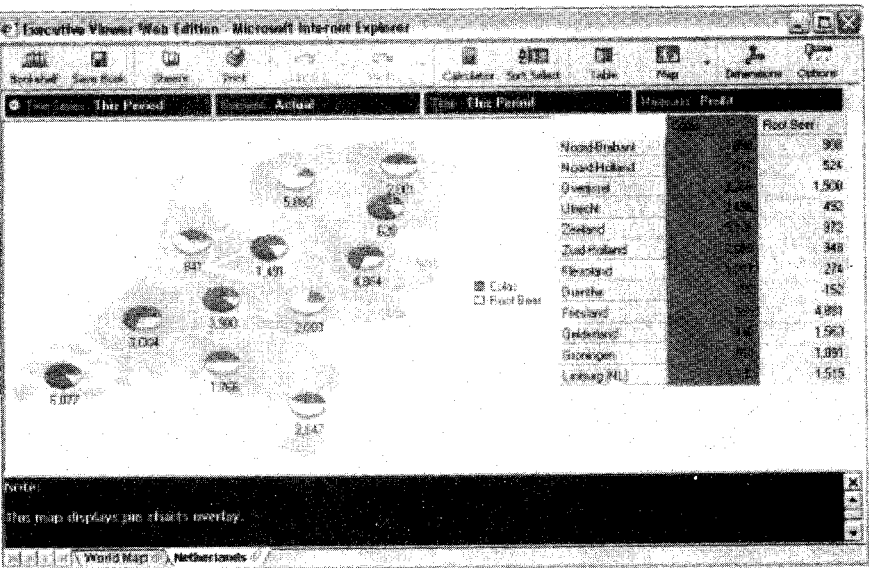
Raden (1997) discusses approaches that may be used in selecting the appropriate OLAP technology for an organization. Menninger (1997) discusses how an organization should go about developing an object-oriented OLAP application. For more on

FIGURE 5.8 COGNOS IMPROMPTU SAMPLE OUTPUT



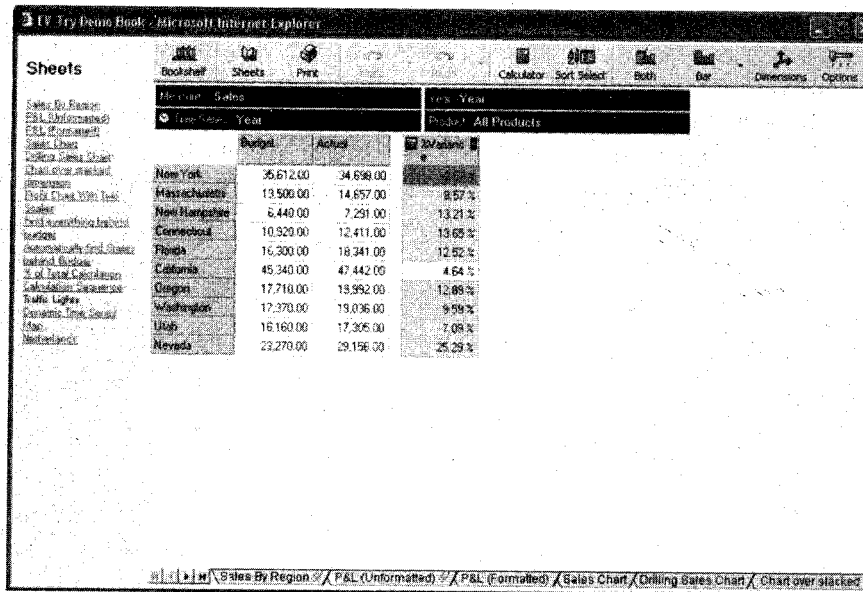
Courtesy of Cognos, Inc.

FIGURE 5.9 DATA DISPLAY



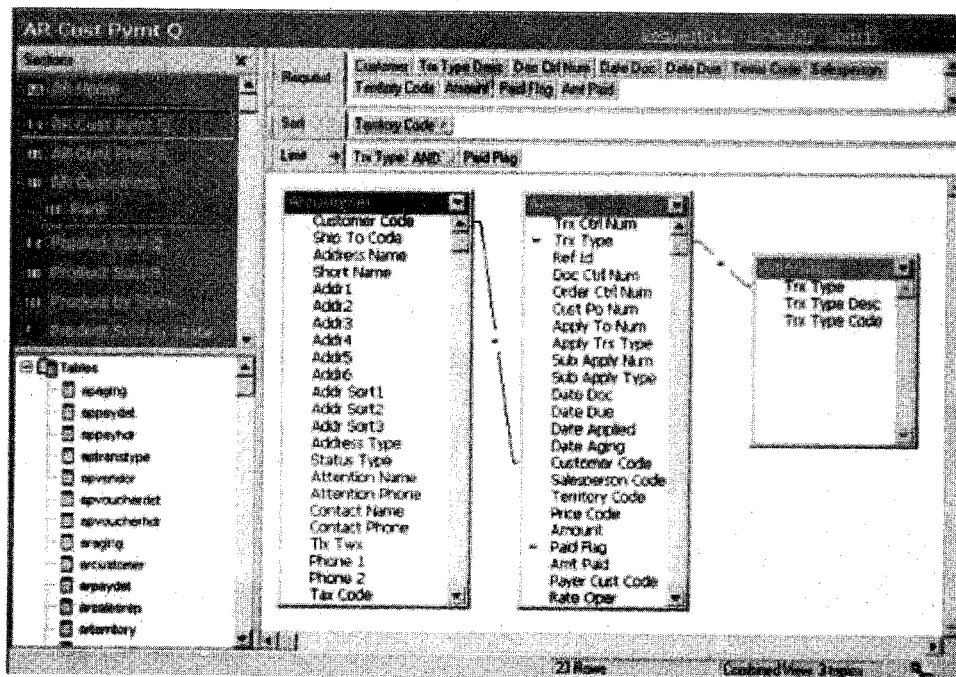
Courtesy of Temtec Inc.

FIGURE 5.10 TEMTEC EXECUTIVE VIEWER TRAFFIC LIGHT DISPLAY



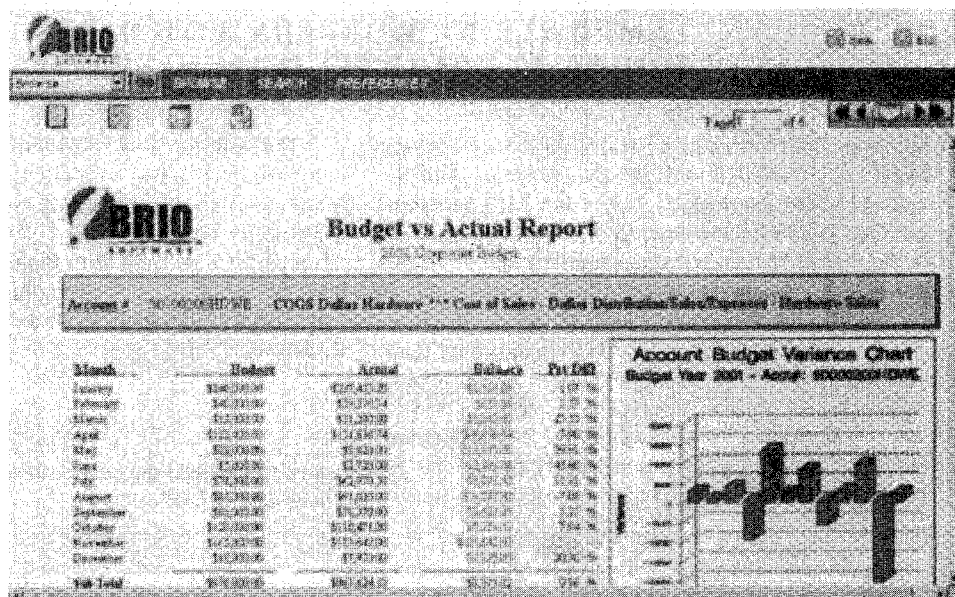
Courtesy of Temtec Inc.

FIGURE 5.11 BRIO PERFORMANCE SUITE SCREEN SHOT OF WEB-BASED DRAG-AND-DROP QUERY



Courtesy of Brio Software

FIGURE 5.12 BRIO PERFORMANCE SUITE SCREEN SHOT OF REPORTING WITH OLAP CAPABILITY



Courtesy of Brio Software Inc.

OLAP, see Barquin and Edelstein (1997a, 1997b), Dash (2001), Kudyba (2002), Ledman (2003), Menninger (1997), Gonzales and Robinson (2003), Havenstein (2003a), and Raden (1997).

5.11

Traditional data analysis is done by inserting data into standards or customized models. In either case, it is assumed that the relationships among various system variables are well known and can be expressed mathematically. However, in many cases, relationships may not be known. In such situations, modeling is not possible and a data mining approach may be attempted.

Data mining (DM) is a term used to describe knowledge discovery in databases.

Data mining is a process that uses statistical, mathematical, artificial intelligence and machine-learning techniques to extract and identify useful information and subsequent knowledge from large databases. Formerly the term was used to describe the process through which undiscovered patterns in data were identified. However, over time, the original definition has been modified to include most types of (automated) data analysis. According to the Gartner Group, data mining is the process of engineering mathematical patterns from usually large sets of data. These patterns can be rules, affinities, correlations, trends, or prediction models (see Nernati and Barko, 2001; Linden, 1999).

Data mining is on the interface of computer science and statistics, utilizing advances in both disciplines to make progress in extracting information from large databases. It is an emerging field that has attracted much attention in a very short time.

DSS IN ACTION 5.32**EMPOWERING WORKERS WITH DATA MINING
AT ROCKWELL INTERNATIONAL**

Rockwell International's air transport division (Cedar Rapids, Iowa) needed to access the corporate database frequently. For many years, only a few MIS personnel had the technical know-how to dig corporate data out of the mainframe. However, executives and managers increasingly demanded access to information stored in the mainframe. Frustration and delays in providing information were common. The MIS department operated under a heavy workload. Furthermore, because of the priority given to top management, other employees had to wait days or even months to get the information they needed. Today, managers can easily and quickly get most of the data they need by themselves.

The solution was provided by creating a special database on a server in a client/server environment. Managers can develop their own applications with LightShip (from Pilot Software, pilotsw.com). Managers can go right after the information they need without having to be programmers, resulting in less frustration and backlog, and happy employees at Rockwell.

Source: Condensed from public information provided by Pilot Software Inc., pilotsw.com.

Glymour et al. (1997) discuss statistical themes and lessons directly relevant to data mining, and some opportunities for synergy between the computational and statistical communities for further advances in data analysis.

Data mining includes tasks known as knowledge extraction, data archaeology, data exploration, data pattern processing, data dredging, and information harvesting. **All** these activities are conducted automatically and allow quick discovery even by nonprogrammers (see DSS in Action 5.32). The following are the major characteristics and objectives of data mining:

- Data are often buried deep within very large databases, which sometimes contain data from several years. In many cases, the data are cleaned and consolidated in a data warehouse.
- The data mining environment is usually a client/server architecture or a Web-based architecture.
- Sophisticated new tools, including advanced visualization tools, help to remove the information *ore* buried in corporate files or archival public records. Finding it involves massaging and synchronizing these data to get the right results. Cutting-edge data miners are also exploring the usefulness of *soft* data (unstructured text stored in such places as Lotus Notes databases, text files on the Internet, or a corporate-wide intranet).
- The miner is often an end-user, empowered by data drills and other power query tools to ask ad hoc questions and obtain answers quickly with little or no programming skill.
- *Striking it rich* often involves finding an unexpected result and requires end-users to think creatively.
- Data mining tools are readily combined with spreadsheets and other software development tools. Thus, the mined data can be analyzed and processed quickly and easily.
- Because of the large amounts of data and massive search efforts, it is sometimes necessary to use parallel processing for data mining.

Effectively leveraging data mining tools and technologies can lead to acquiring and maintaining a strategic competitive advantage. Data mining offers organizations an indispensable decision-enhancing environment to exploit new opportunities by transforming data into a strategic weapon. See Nemati and Barko (2001). Many authors have written on the practical and theoretical approaches of data mining, including Web, text, search, and organizational transformation issues. Among them are Anthes (2002), Barquin and Edelstein (1997a, 1997b), Barth (1997), Berry (2000, 2002, 2003a, 2003b), Bhandari et al. (1997), Bigus (1996), Bolen (2001), Boyd (2001), Buck (2000), Copeland (2001), Dunham (2003), Fayyad (2003), Finlay (2001), Gimes (2001), Glymour et al. (1997), Haskett (2000a, 2000b), Hudnell and Kitayama (2000), Lamont (2000), Linden (1999), Linoff and Berry (2002), Nemati and Barko (2001), Palshikar (2001), Saarevirta (2001), Swartz (2003), Skalak (2001), Small and Edelstein (1997), Swift (2001), Tillett (2000), Walker (2002), and Zaima (2003). Also see the Kdnuggets Web site (www.kdnuggets.com). See Groth (1998), Haskett (2000a, 2000b), and Nemati and Barko (2001) for excellent introductions to the topic.

HOW DATA MINING WORKS

Intelligent data mining, according to Dunham (2003), discovers information within data warehouses that queries and reports cannot effectively reveal. Data mining tools find patterns in data and may even infer rules from them. Three types of methods are used to identify patterns in data (Nemati and Barko, 2001):

- Simple models (SQL-based query, OLAP, human judgment)
- Intermediate models (regression, decision trees, clustering)
- Complex models (neural networks, other rule induction)

These patterns and rules can be used to guide decision-making and forecast the effect of decisions. Data mining can speed analysis by focusing attention on the most important variables. The dramatic drop in the cost/performance ratio of computer systems has enabled many organizations to start applying the complex algorithms of data mining techniques. Each *data mining application class* is supported by a set of algorithmic approaches to extract the relevant relationships in the data. These approaches differ in the classes of problems they are able to solve (see Haskett, 2000b). The classes are:

- **Classification:** infers the defining characteristics of a certain group (e.g., customers who have been lost to competitors). These methods involve seeding a set of data with a known set of classes (perhaps found by clustering), and mapping all other items (customers) into these sets. Decision trees and neural networks are useful techniques.
- **Clustering:** identifies groups of items that share a certain characteristic (clustering differs from classification in that no predefined characteristic is given). Clustering approaches address segmentation problems. Clustering algorithms can be used to identify classes of customers with certain needs to be met.
- **Association:** identifies relationships between events that occur at one time. Association approaches address a class of problems typified by market basket analysis. In retailing, there is an attempt to identify what products sell with what other ones, and to what degree. Statistical methods are typically used.

- **Sequencing:** similar to association, except that the relationship occurs over a period of time (e.g., repeat visits to a supermarket, use of a financial planning product). Purchases can be tracked because the purchaser can be identified by an account number or some other means.
- **Regression:** used to map data to a prediction value. Linear and nonlinear techniques are used. This is a form of *estimation*. It often involves identifying metrics and evaluating an item (customer) along the metrics by assigning scores. Sales predictions may be accomplished as well.
- **Forecasting:** estimates future values based on patterns within large sets of data (e.g., demand forecasting). This is another form of *estimation*. There is an attempt to utilize statistical time-series methods to predict future sales.
- **Other techniques:** these are typically based on advanced artificial intelligence methods. They include case-based reasoning, fuzzy logic, genetic algorithms, and fractal-based transforms.

In Table 5.7, we show these data mining functions along with representative algorithms and application examples. Also see Groth (1998).

Firms often effectively use their data mining systems to perform market segmentation by *cluster analysis*. Cluster analysis is a means of identifying classes of items so that items in a cluster have more in common with each other than with items in other clusters. We provide a detailed description of cluster analysis and an excellent example in the banking field in Case Application 5.3. Similarly, JCB Co., Ltd. (Japan) has effectively used cluster analysis as part of its data mining effort in segmenting customers and directing appropriate marketing products to the segments at the right time in the right format at the right price (see the *revenue management* subsection in Chapter 4). See DSS in Action 5.33 for a description. Tillett (2000) describes how another bank effectively mines customer data via cluster analysis using Web-based tools. In DSS in Action 5.31, we discussed how OLAP tools were used for a similar purpose.

Data mining can be either *hypothesis driven* or *discovery driven*. **Hypothesis-driven data mining** begins with a proposition by the user, who then seeks to validate the truthfulness of the proposition. For example, a marketing manager may begin with the proposition, "Are DVD players sales related to sales of television sets?" **Discovery-**

TABLE 5.7 Data Mining Functions, Algorithms, and Application Examples.

<u>Data Mining Function</u>	<u>Algorithm</u>	<u>Application Examples</u>
Associates	Statistics, set theory	Market basket analysis
Classification	Decision trees, neural networks	Target marketing, quality
Clustering	Neural networks, statistics, optimization, discriminant analysis	control, risk assessment
Modeling	Linear and nonlinear regression, curve fitting, neural networks	Market segmentation
Sequential patterns	Statistics, set theory	Sales forecasting, interest rate prediction, inventory control
		Market basket analysis over time, customer life cycle analysis

Source: Adapted from I. P. Bigus, *Data Mining with Neural Networks*, New York: McGraw-Hill, 1996.

DSS IN ACTION 5.33**DATA MINING AT JCB CO., LTD.**

As the largest credit card issuer in Japan, JCB Co., Ltd. has established itself as an international brand. The company, with 34 million cardholders, offers 200 services and 600 card types. In July 1999, JCB's sales department started to develop a data mining project to track increasingly diverse and complex customer needs. The system was completed in three months with SAS Enterprise Miner. (Enterprise Miner contains many integrated models and algorithms, including decision trees, neural networks, regression, memory-based reasoning, bagging and boosting ensembles, two-stage models, clustering, time series, and associations. See www.sas.com for details.) The JCB system includes customer profiling and customer relationship management.

JCB segments its customers (cluster analysis) to increase the response rate of its marketing campaigns, which in turn increases revenue. Then, through customer profiling technology, it devises customer-focused

sales strategies. "The system analyzes how members use their cards, helping JCB identify and retain its most profitable customers. JCB plans to serve its customers for several decades by offering different services for various life stages.

"By clustering and making associations with the data, we are trying to figure out what customers need. And we would like to polish our business models by repeating the process of planning, doing, checking and practicing," says Makoto Nakaoka, manager of JCB's business administration department. In a short time, JCB quadrupled the rate of customer responses to direct-mail solicitations and initiated a successful campaign to retain current cardholders, at a success rate six to ten times greater than ever before.

Source: Marla Hudnell and Hajime Kitayarna, "Transforming Business with Data Mining," *SAS com*, November/December 2000, pp. 22-23.

driven data mining finds patterns, associations, and relationships among the data. It can uncover facts that were previously unknown or not even contemplated by an organization.

Buck (2000) organized the classes of data mining tools and techniques as they relate to information and business intelligence technologies. Her taxonomy is

- Mathematical and statistical analysis packages
- Personalization tools for Web-based marketing
- Analytics built into marketing platforms
- Advanced CRM tools
- Analytics added to other vertical industry-specific platforms
- Analytics added to database tools (e.g., OLAP)
- Standalone data mining tools

In data mining (and OLAP), the scalability of the methods and of the data warehouse (or database) are critical issues. This is so because of the amount of data and searching required. See Small and Edelstein (1997) and Section 5.7 for more on these issues and how they can be handled.

Edelstein (2001) describes the seven steps necessary for successful data mining. See DSS in Focus 5.34. If these are followed, and business practices are right, then the data mining effort should succeed. Acting on the results is critical, because discovering relationships in the data has no impact unless the relationships are utilized.

A number of misconceptions have developed about data mining. We describe these in DSS in Focus 5.35. Many of these reflect the way that data mining is utilized in practice. For example, data mining methods are typically used mainly by IT staff and management, and by consultants/analysts because it is too hard for nontechnical per-

DSS IN FOCUS 5.34**THE SEVEN STEPS OF DATA MINING**

Data mining uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make accurate predictions. Data mining helps organizations develop the most accurate models of their customers and prospective customers. The seven steps of data mining are:

1. Define the business problem.
2. Build (find or acquire) the data-mining database.

3. Explore the data.
4. Prepare the data for modeling.
5. Build (or find) the models.
6. Evaluate the models.
7. Act on the results.

Source: Adapted from Herbert Edelstein, "Pan for Gold in the Clickstream," *InformationWeek*, March 12, 2001, pp. 77-91.

DSS IN FOCUS 5.35**DATA MINING MYTHS**

Data mining is a powerful analytic tool that enables business executives to advance from describing historical customer behavior to predicting the future. It finds patterns that unlock the mysteries of customer behavior. The results of data mining can be used to increase revenue, reduce expenses, identify fraud, and identify business opportunities, offering new competitive advantage. There are a number of myths about data mining, listed below. Data mining visionaries have gained enormous competitive advantage by understanding that these myths are just that—myths.

- *Data mining provides instant, crystal-ball predictions.* Data mining is a multi-step process that requires deliberate, proactive design and use.
- *Data mining is not yet viable for business applications.* The current state-of-the-art is ready to go for almost any business.

- *Data mining requires a separate, dedicated database.* Because of advances in database technology, a dedicated database is not required, even though it may be desirable.
- *Only Ph.D.s can do data mining.* Newer Webbased tools make data mining by managers possible.
- *Data mining is only for large firms with lots of customer data.* If the data accurately reflect the business or its customers, a company can utilize data mining.

Source: Adapted partly from Arlene Zaima, "The Five Myths of Data Mining," *What Works: Best Practices in Business Intelligence and Data Warehousing*, Vol. 15, The Data Warehousing Institute, Chatsworth, CA, June, 2003, pp. 42-43.

sonne 1 to interpret results. But this is changing as the tools become easier to use. In 2001, only around 35 percent of corporate management and staff directly used datamining tools (see Nemati and Barko, 2001). These results are comparable to those for data warehouses and OLAP.

Data mining is iterative because data miners make mistakes. Actually, it is the process of discovery that is iterative. Thomas A. Edison quipped that he *failed* to invent the light bulb 100 times before he succeeded. So, just like the famous inventor's work process, data mining is a process of discovery. It is an experimental process that requires sound experimental design. See DSS in Focus 5.36 for specific "errors" that data miners typically make in practice because they often do not understand the process but do understand the expected results.

DSS IN FOCUS 5.36

DATA MINING BLUNDERS



Here are ten data mining mistakes that are often made in practice. Try to avoid them:

- Select the wrong problem for data mining.
- Ignore what your sponsor thinks data mining is, and what it really can and cannot do.
- Leave insufficient time for data preparation. This takes more effort than is generally understood.
- Look only at aggregated results, never at individual records. IBM's DB2 Intelligent Miner Scoring can highlight individual records of interest.
- Be sloppy about 'keeping track of the mining procedure and results.
- Ignore suspicious findings and quickly move on.
- Run mining algorithms repeatedly and blindly. Don't think hard enough about the next stage of data analysis. Data mining is a very hands-on activity.
- Believe everything you are told about the data.'
- Believe everything you are told about your own data mining analysis.
- Measure your results differently from the way your sponsor measures them.

Source: Adapted from David Skalak, "Data Mining Blunders Exposed!" *DB2 Magazine*, Quarter 2, 2001, pp.10-13.

DATA MINING TOOLS AND TECHNIQUES

There are many methods for performing data mining. Data-mining software may utilize one or more of these techniques; this is one of the distinguishing characteristics of data-mining software. Data mining tools and techniques can be classified based upon the structure of the data and the algorithms used. The main ones are:

- *Statistical methods.* These include linear and nonlinear regression, point estimation, Bayes's theorem (probability distribution), correlations, and cluster analysis.
- *Decision trees.* Decision trees are used in classification and clustering methods. Decision trees break problems down into increasingly discrete subsets, by working from generalizations to increasingly more specific information. A decision tree can be defined as a root followed by internal nodes. Each node (including the root) is labeled with a question. The arcs associated with each node cover all possible responses. Each response represents a probable outcome (see Dunham, 2003).
- *Case-based reasoning.* Using historical cases, the case-based reasoning approach can be used to recognize patterns. For example, customers of Cognitive Systems Inc. use such an approach for help desk applications. One customer has a 50,000 query case library. New cases can be matched quickly against the 50,000 samples in the library, providing automatic answers to queries with more than 90 percent accuracy. For more on case-based reasoning, see Chapter 12.
- *Neural computing.* Neural networks utilize many connected nodes (which operate in a manner similar to how the neurons of the human brain function). This approach examines a massive amount of historical data for patterns. Thus, one can go through large databases and, for example, identify potential customers for a new product (see DSS in Action 5.37) or companies whose profiles suggest that they are heading for bankruptcy. Many applications are in financial services (Fadlalla and Lind, 2001) and in manufacturing. A comprehensive description of neural networks is covered later in the text.

DSS IN ACTION 5.37**DATA MINING AT MARRIOTT**

Marriott Club International (www.marriot.com), America's largest seller of vacation time-share condos, had a problem. The company has a database with millions of names. It used to send advertisements to all of them at great expense, but the response was minimal. The company decided to identify the customers on their list who were more likely to respond. Marriott uses neural computing technology in its data mining; the objective is to detect patterns by combing through the digitized customer lists.

Marriott started with names, mostly of hotel guests. Digging into a trove of motor vehicle records, property records, warranty cards, and lists of people who have bought by mail or on the Web, a computer program enriches the prospect list. It adds such facts as the cus-

tomers' ages, their children's ages, their estimated income, what cars they drive, and whether they play golf. The Marriott system then identifies who is most likely to respond to a mailed flier.

With these factors, Marriott is able to cast its net a little more narrowly and catch more fish. Data mining has increased the response rate to Marriott's direct mail time-share pitches by 33 percent. In addition, the company has reported significant savings on its mail costs. The same approach can be applied to Internet advertising, as was done by Site59.com Inc.

Source: Condensed from 1. Novack, "The Data Miners," *Forbes*, February 12, 1996, and www.marriotclub.com.

- *Intelligent agents.* One of the most promising approaches to retrieving information from databases, especially external ones, is the use of intelligent agents. With the availability of a vast and growing amount of information through the Internet, finding the right information is becoming more difficult. Web-based data mining applications are typically enabled by intelligent software agents. This topic is discussed in a later chapter.
- *Genetic algorithms.* Genetic algorithms work on the principle of expansion of possible outcomes. Given a fixed number of possible outcomes, genetic algorithms seek to define new and better solutions. Genetic algorithms are used for clustering and association rules.
- *Other tools.* Several other tools can be used. These include rule induction and data visualization. The best source of new tool development is vendor Web sites.

The case-based reasoning, neural computing, intelligent agent, and genetic algorithm methods have their foundations in artificial intelligence.

Data mining algorithms are important (see Dasu and Johnson, 2003). When dealing with customer behavioral data, which can encompass a hundred dimensions or more, algorithms should be capable of dealing effectively with high-dimensional data. These algorithms must also be able to work with business constraints and rules. Simple statistics do not work. Knowledge of the business constraints, of the relations between products, and of the various behavioral segments of customers is a must.

Since the terrorist attacks on September 11, 2001, there have been numerous advances in the utilization of data mining methods by law-enforcement agencies to track terrorism and crime in general. See DSS in Action 5.38 for details and an example.

TEXT MINING

Text mining is the application of data mining to nonstructured or less structured text files. Data mining takes advantage of the infrastructure of stored data to extract additional useful information. For example, by data mining a customer database, an analyst

DSS IN ACTION 5.38**DMS DATA MINING SPINOFFS
AND ADVANCES IN LAW ENFORCEMENT**

In late 2002, John Poindexter, former head of the National Security Council, caused a flap with his proposal for a new Information Awareness Office within the Pentagon. Critics blasted Poindexter's plans for data mining numerous credit, banking, and retail purchase records of U.S. citizens, in the name of detecting possible terrorist patterns of behavior.

In reality, agencies like the National Reconnaissance Office and the National Security Agency have been doing this for years, and in mid-20m the Northern Command did so as well. In fact, data-mining tools used by national intelligence agencies are already being utilized by domestic law-enforcement agencies in the United States. The tools transferred from the U.S. Space Command to Northern Command, and from there to the Department of Homeland Security, show both the common technology base for all environments and the possible civil-liberties concerns inherent in such tech transfers. All agencies are concerned about respect for civil liberties. Better intelligence coordination with state and local police forces is a chief goal.

The NRO and NSA use large-scale commercial database tools and specialized pattern-recognition tools. Defense contractors are responsible for integrating tools together in software suites that would prove useful to intelligence agencies. Many were working with the Department of Homeland Security's constituent agencies before DHS formed at the end of 2002. They deploy the tools for domestic drug enforcement and counterterrorism duties through the channels of the Northern Command and DHS.

For example, Northrop-Grumman's Web-enabled Temporal Analysis System (WebTAS) was developed

in conjunction with the Air Force Research Labs and used during the Iraq campaign. It is available to regional police intelligence coalitions through the DHS. WebTAS displays maps and shows links corresponding to relations among targets. Clicking on a link calls up related databases that can tell an analyst, for example, all of the calls that the target has made in the last few days. To pick up patterns that might be buried in the noise of too much information, an embedded behavioral-correlation engine predicts possible trends for developing situations and flags circumstances that identify problems for gathering further intelligence.

New cooperation among Homeland Security investigators, especially in mining data, is producing major breakthroughs in nonterrorist cases, including the deaths of 19 illegal aliens found stuffed into a trailer in Victoria, Texas, on May 14, 2003. Detected via data mining techniques, money transfers and phone calls made by victims and more than 50 witnesses who survived the incident led authorities to a legal U.S. resident, who they believe led a smuggling ring that took aliens across the Mexican border to U.S. cities for a substantial fee. Coconspirators around the United States were also identified. After the suspect fled the country, she was lured by a sting operation to Honduras, where she was arrested and extradited to the United States.

Sources: Adapted from Loring Wirbel, "Data Mining Comes Down to Earth," *Electronic Engineering Times*, No. 1270, May 19, 2003, pp. 18, 22; Mark Hosenball, "Crime Breakthrough," *NewsWeek*, June 20, 2003, p. 9.

might discover that everyone who buys product A also buys products Band C, but six months later. Text mining operates with less structured information. Documents rarely have a strong internal infrastructure, and when they do, it is frequently focused on document format rather than document content. Text mining helps organizations to

- Find the "hidden" content of documents, including additional useful relationships.
- Relate documents across previous unnoticed divisions; for example, discover that customers in two different product divisions have the same characteristics.
- Group documents by common themes; for example, all the customers of an insurance firm who have similar complaints and cancel their policies.

Ellingsworth and Sullivan (2003) describe the process of text mining (see DSS in Focus 5.39). They also describe how the Fireman's Fund Insurance Company utilizes

DSS IN ACTION 5.40**TEXT MINING**

Text mining is a very effective approach to automatically performing analysis on standard and Web documents. For example, an international pharmaceutical firm used text mining to evaluate 500 text-based responses from patients participating in a clinical study of a new allergy medication. Text mining software detected a cluster of 50 patients who used specific words that described negative side effects. Further

examination indicated that these patients all received a high dosage of the drug, and that women older than 40 were especially sensitive to the high dosage. Consequently, dosage levels are adjusted, and warnings to women over 40 are included with the medicine.

Source: Adapted from A. Bolen, "Data Mining for Text," SAS com, November/December 2001.

DSS IN ACTION 5.41**DATA MINING AT PFIZER**

Pfizer, a large pharmaceutical company, uses text mining to look for parallels in pharmaceutical testing in the extremely large database that the National Institutes of Health uses to catalog medical research. The text mining project targets biomedical documents extracted from various external sources, such as MedLine, a medical research literature service provided by the National Institutes of Health.

The Pfizer system searches the database of documents and extracts a set of documents characterized by simple search criteria based on a combination of keywords. Next, the set of documents is further segmented into topics. Topics are characterized by lists of keywords extracted from the free-format text contained in the documents. The scientists choose topics of interest by

examining keyword lists. Pfizer has realized several benefits. First, the company has discovered that text mining is not only a technology for the categorization of information. The results of text mining also permit the building of new applications for further navigation of data and decision support. These new applications can take a prototype to complete development much faster than ever before. It is now possible to rapidly assemble powerful, easy-to-use analytical applications that address the full gamut of requirements.

Source: Adapted from Lawrence Bell, "For Pfizer, AlphaBlox Is Just What the Doctor Ordered." *What Works: Best Practices in Business Intelligence and Data Warehousing*, Vol. 10, The Data Warehousing Institute, Chatsworth, CA, June, 2003, p. 31.

SAMPLER OF DATA MINING APPLICATIONS

Data mining can be very helpful, as shown by the following representative examples. Note that the intent of most of these examples is to identify a business opportunity to create a sustainable competitive advantage.

- **Marketing:** predicting which customers will respond to Internet banners or buy a particular product; segmenting customer demographics.
- **Banking:** forecasting levels of bad loans and fraudulent credit card usage, credit card spending by new customers, and which kinds of customers will best respond to new loan offers or other products and services.
- **Retailing and sales:** predicting sales and determining correct inventory levels and distribution schedules among outlets.
- **Manufacturing and production:** predicting when to expect machinery failures, finding key factors that control the optimization of manufacturing capacity.

- *Brokerage and securities trading*: predicting when bond prices will change, forecasting the range of stock fluctuation for particular issues and the overall market; determining when to trade stocks.
- *Insurance*: forecasting claim amounts and medical coverage costs, classifying the most important elements that affect medical coverage, predicting which customers will buy new policies with special features.
- *Computer hardware and software*: predicting disk drive failure, forecasting how long it will take to create new chips, predicting potential security violations.
- *Government and defense*: forecasting the cost of moving military equipment, testing strategies for military engagements, predicting resource consumption.
- *Airlines*: capturing data not only on where customers are flying but also the ultimate destination of passengers who change carriers in mid-flight. With this information airlines can identify popular locations they are not currently serving so as to add routes and capture lost business.
- *Health care*: correlating demographics of patients with critical illnesses; using data mining, doctors can develop better insights on symptoms and how to provide proper treatments.
- *Broadcasting*: predicting what programs are best shown during prime time and how to maximize returns by inserting advertisements.
- *Police*: tracking crime patterns, locations, criminal behavior, and attributes to help solve criminal cases (see DSS in Action 5.55 in Section 5.13).

Palshikar (2001) provides several examples of effective data mining in practice.

See DSS in Focus 5.42 for information about data-mining and analysis efforts at DRS. Census data can be combined with other market data when segmenting customers (see Gimes, 2001). For the capabilities of data mining and a comparison of data mining tools, see Dunham (2003), Roiger and Geatz (2003).

Wal-Mart continues to pioneer data mining efforts. In fact, Wal-Mart even notices blips in data due to ethnic holidays and plans for them. See DSS in Action 5.43. Data mining is critical when utilized in a customer relationship (resource) management (CRM) system, as is described in DSS in Focus 5.44. See Berry (2000, 2002, 2003a, 2003b), Fayyad (2003), Linoff and Berry (2000), and Swift (2001).

A less typical application of data mining was applied to improving the performance of National Basketball Association (NBA) teams in the United States. The NBA developed *Advanced Scout*, a PC-based data mining application used by coaching staffs to discover interesting patterns in basketball game data. The process of pattern interpretation is facilitated by allowing the user to relate patterns to videotape. See Bhandari et al. (1997) for details.

KDD AND DATA MINING

Data mining and knowledge discovery in databases (KDD) are frequently used as synonyms. Fayyad et al. (1996) define knowledge discovery in databases (KDD) as a process of using data mining methods to find useful information and patterns in the data, whereas data mining is the use of algorithms to identify patterns in data -derived by the KDD process. KDD is a comprehensive process that encompasses data mining. The input to the KDD process consists of organizational data. The enterprise data warehouse enables KDD to be implemented efficiently because it provides a single source for data to be mined. Dunham (2003) summarizes the KDD process as consisting of the following steps:

- *Selection*: Identification of the data that will be considered within the data mining process.

DSS IN FOCUS 5.42**HOMELAND SECURITY DATA MINING AND ANALYSIS ~**

The U.S. government's Total Information Awareness (TIA) project, spearheaded by the Defense Advanced Research Projects Agency (DARPA), has been called "the mother of all data-mining projects." The research and development program, headed by John Poindexter, aims to identify, track, and prevent individuals from planning and organizing terrorist activities. Much of the effort focuses on unifying and probing databases that carry information on financial transactions. The program will also create large databases that sift through the purchases, travel, immigration status, income, and other data of millions of Americans. There are three parts to the TIA project:

- *Voice recognition.* Sifting through electronically recorded transmissions and providing rapid translations of foreign languages
- *A tool* to find connections between transactions, such as passports, airline tickets, rental cars, gun or chemical purchases, arrests, and other suspicious activities
- *Collaboration.* A mechanism to enable information and analysis sharing among agencies

Experts say it may only take the government a year to get this technology up and running. The key to success will be facilitating information sharing among departments. The technology to mine these data sources is there today, but developing systems to talk to each other may be a challenge for some time to come.

Even the National Science Foundation (NSF) is getting involved in developing and promoting data mining methods for the DHS. In 2002, the government intelligence community provided \$6 million to supplement existing NSF research on data mining, with comparable funding likely for two more years. Many data-mining projects are summarized on the NSF Web site.

Sources: Adapted from Todd Datz, "Integrating America." *C/O*, December 1, 2002, p. 44-51; Nikki Swartz, "Data Mining Initiatives," *Information Management Journal*, Vol. 37, No.2, March/April 2003, p. 17; NSF, "Fact Sheet: Data Mining and Homeland Security Applications," Office of Legislative and Public Affairs, www.nsf.gov/od/lpa/news/03/fact030124.htm January 24, 2003.

- *Preprocessing:* Erroneous and missing data must be dealt with. This involves correction and/or utilizing predicted values.
- *Transformation:* The data must be converted into a single common format for processing; this may involve encoding data or reducing the number of variables with which to deal.
- *Data mining:* Algorithms are applied to the transformed data in order to produce output.
- *Interpretation/evaluation:* To be useful, the results must be presented in a manner that is meaningful to the user.

DSS IN ACTION 5.43**WAL-MART: A PIONEER OF DATA MINING**

Wal-Mart has pioneered massive data mining efforts to transform its supplier relationships. Wal-Mart captures point-of-sale transactions from over 2,900 stores in six countries and continuously transmits these data to its massive Teradata data warehouse. Wal-Mart allows more than 3,500 suppliers to access data on their products and perform data analyses. The suppliers use these data to identify customer buying patterns at the store-

display level. They use this information to manage local store inventory and identify new merchandising opportunities. In 1995, Wal-Mart computers processed over 1 million complex data queries.

Source: Adapted from P. Westerman, *Data Warehousing: Using the Wal-Mart Model*, San Francisco: Morgan Kaufmann, 2000; and public sources.

DSS IN FOCUS 5.44**DATA MINING TO IDENTIFY CUSTOMER BEHAVIOR ~**

Understanding customer behavior is important to adjusting business strategies, increasing revenues, and identifying new opportunities. Many organizations have a massive amount and impressive variety of data and information resources that promise to reveal much more about customer behavior than was ever thought possible. Many firms have reached a point of rich data and poor utilization. For most retail environments, three sources of customer data are most critical to data mining efforts toward better understanding of behavior:

- Demographic data
- Transaction data
- Online interaction data

Clickstream analytics can identify who did and did not buy your product, why, and when.

Retail uses of data mining evolve as:

- Step 1: *Web analytics*. Gather Web site statistics that track customers' online behavior: hits, pages, sales volume, etc. This helps adjust a Web site to meet customer needs.
- Step 2: *Customer analytics*. These add depth to understanding customer interactions. Firms gather

data from multiple sources, including Web site interactions, transaction data from offline purchases, and demographic data. This is critical in CRM and revenue management in that a better understanding allows an organization to cluster customers into groupings.

- Step 3: *Optimization*. This promises the largest payoff. Subtle patterns can be detected and utilized to optimize customer interactions. This is the goal of CRM (Chapter 8) and revenue management (Chapter 4).

Consider 1. Crew, a major online and catalog retailer of men's and women's apparel, shoes, and accessories. Crew has had immense success with optimization analytics. The company previously used a cumbersome manual procedure to recommend similar and complementary styles to online purchasers. In the fall of 2002,¹ Crew deployed optimization analytics. The analytic engine recommendations, done automatically, generate twice as many sales as the older, manual system.

Source: Adapted from Usama Fayyad, "Optimizing Customer Insight," *Intelligent Enterprise*, May 2003.

INTELLIGENT DATA MINING AND TEXT MINING

New intelligent data and text mining methods, often based on artificial intelligence methods like artificial neural networks and intelligent agents, continue to be developed and applied in practice. These methods often prove to be very effective on specific kinds of problems and sets of data and text. Many are applied to identifying information and knowledge on Web pages scattered around the world. We describe some new intelligent-based methods for data mining in DSS in Focus 5.45. Also see Anthes (2002). See the Kdnuggets Web site (www.kdnuggets.com) for some additional information on intelligent methods for data mining.

When organizations are plagued by fraud, especially in financial transactions, as in e-commerce, they turn to specialized data-mining tools to detect patterns in the data. Generally these methods use neural networks in addition to clustering and statistical methods. SAS Anti-Money Laundering software is one example of how this is implemented in practice. See DSS in Action 5.46 for an example.

A team of Norwegian biologists have developed intelligent methods to *search and mine* the Web for genetic studies that contain information relevant to their endeavors. Since every three years we double the amount of information that we generate and store on earth (see Pallatto, 2002a), methods like these become increasingly important for scientific researchers as well as for smoothly running businesses. See DSS in Action 5.47 and Copeland (2001) for how this is done. Other methods, such as intelligent agents, may also be utilized in intelligent mining. Lamont (2000) describes how intelligent agents can

DSS IN FOCUS 5.45**NEW INTELLIGENT METHODS TO MINE DATA**

Here are some new intelligence-based methods for searching, sifting, and analyzing huge data sets and Web documents:

Non-Obvious Relationship Awareness (NORA) (Systems Research & Development). NORA can take information from disparate sources about people and their activities and find obscure, nonobvious relationships. Useful for reaching further into the world of criminals and terrorists (see the Opening Vignette and DSS in Action 5.38).

Outbreak detection (Tom Mitchell at Carnegie Mellon University). This is distributed data mining. Tracks millions to trillions of items looking for disease outbreaks in real-time.

Upside Down (Streamlogic Inc.). Instead of archiving data and running search queries, Upside Down archives search queries and runs data through them. The focus is on identifying what people are looking for rather than what is found. This is some 6,000 times faster than the conventional approach.

What's the Answer? (Verity Inc.). The smart software puts human learning (rules) into the search

software, enabling it to learn through logistic-regression classification. For example, instead of responding with a list of Web sites, a search engine could simply scan through several of them and answer the question that is posed (e.g., "What is the population of the world?").

Web Fountain (IBM). This software is based on Andrew Tomkin's research results. In teaching computers to read for comprehension and recognize patterns in text documents, he set the software up to *read everything on the Web*. Web Fountain developed from this. Now trends in public opinion and popular culture can be identified as they emerge, and tracked as they migrate quickly around the world. If you ask the Web Fountain the right kinds of questions, market research results can almost instantaneously appear. Web Fountain went online in late 2003 with a few pilot customers.

Sources: Adapted from Gary H. Anthes, "The Search Is On," *Computer World*, April 15, 2002; Brent Schlender, "How Big Blue Is Turning Geeks into Gold," *Fortune*, June 9, 2003, pp. 133-140.

be used to identify knowledge on the Web. Boyd (2001) describes how the BizWorks software package provides intelligent agents for internal and Web searches.

DATA MINING SOFTWARE

Data-mining software features more complicated algorithms for neural networking, clustering, segmentation, and classifications that are generally more sophisticated than OLAP methods (see Finlay, 2001). Many software vendors provide powerful data-

DSS IN ACTION 5.46**UNITY TRUST BANK FIGHTS FRAUD WITH SAS ANTI-MONEY LAUNDERING**

Unity Trust Bank attempts to detect suspicious transactions, as is required by British law. However, small charitable organization transactions look very similar to those of money launderers. Using SAS Anti-Money Laundering software, the bank focuses on suspicious behavior instead of transactions. Using the software, Unity Trust is able to successfully identify most fraudu-

lent and suspicious transactions, which can be further examined to determine their validity and turned over to law-enforcement officials.

Source: Adapted partly from E. Walker, "Innovation and Social Responsibility in Financial Services," *SAS com*, November/December 2002.

three-dimensional presentations, and animation. Visual tools can help identify relationships directly. The ability to identify important trends in corporate and market data provides enormous advantages. More accurate predictive models provide significant business advantages in applications that drive content, transactions, or processes. Confident action, based on superior methods of visual data analysis, helps companies improve income and avoid costly mistakes (see Hallett, 1991). For example, network monitoring systems continue to become increasingly complicated and sophisticated. Visualization simplifies the reporting of test results. Consonus (Salt Lake City, Utah) designs, builds, and operates data centers, IT networks, and Web-enabled application delivery systems. Consonus uses the HP Open View Management Suite to help manage these data centers. OpenView helps manage customers' Web systems and provides them with an understanding of how customers view their sites' performance and availability. See McCarthy (2002) for details.

Data visualization enables OLAP and data mining, especially utilizing Web-based tools. Rather than having to wait for reports or compare sterile columns of numbers, a manager can utilize a browser interface in real-time to look at vital organizational performance data. By using visual analysis technologies, managers, engineers, and other professionals have spotted problems that for years went undetected by standard analysis methods.

Visualization software packages offer users capabilities for self-guided exploration and visual analysis of large amounts of data. For example, see the ILOG Visualization Suite (www.ilog.com). Some examples of OLAP systems with excellent visualization include Visual Insights (Cognos) and nVizn (SPSS Inc.). Cognos Visualizer, among other features, utilizes traffic light displays in tables and graphs. Numerical results are displayed in red/yellow/green indicating their status. SPSS's nVizn is a Java-based developer's tool kit for creating visualization applications. See Ulfelder (2000b) for details. See DSS in Focus 5.48 for ideas on visualization in Finance, and DSS in Action 5.49 for how it is quickly evolving. Visualization technologies can be integrated to ere-

DSS IN FOCUS 5.48

FINANCIAL DATA VISUALIZATION



To prevent systems from automatically identifying meaningless patterns in data, chief financial officers (CFOs) want to make sure that the processing power of a computer is always tempered with the insight of a human being. One way to do this is through data visualization, which uses color, form, motion, and depth to present masses of data in a comprehensible way. Andrew W. La, director of the Laboratory for Financial Engineering at Massachusetts Institute of Technology's Sloan School of Management, developed a program in which a CFO can use a mouse to "fly" over a three-dimensional landscape representing the risk, return, and liquidity of a company's assets. With practice, the CFO can begin to zero in on the choicest spot on the three-dimensional landscape: the one where the trade-

off among risk, return, and liquidity is most beneficial. La says, "The video-game generation just loves these 3D tools."

So far, very few CFOs have cruised three-dimensional cyberspace, but this continues to change. Most still spend the bulk of their time on routine matters, such as generating reports for the Securities and Exchange Commission. In 1996, Glassco Park president Robert I. Park said, "What we have in financial risk management today is like what we had in computer typesetting in 1981, before desktop publishing." See DSS in Action 5.49 for how this change is occurring.

Source: Condensed from P. Coy, "Higher Math and Savvy Software are Crucial," *Business Week*, October 28, 1996.

DSS IN ACTION 5.49**FINANCIAL DATA VISUALIZATION: PART 2**

Visualization, if done properly, is an incredibly powerful paradigm. SmartMoney.com's Web-based Maps gives Merrill Lynch analysts an easy-to-examine visual representation of financial information and customized three-dimensional pictures of the ever-changing financial markets. Beyond the cutting-edge financial services providers like Merrill Lynch, visualization is becoming ubiquitous in other enterprises as well. And developers are extending its use to serve increasingly diverse audiences. From the financial services markets to highly technical quality engineering companies that create testing tools to service providers using powerful network monitoring tools, visualization is becoming an indemand, value-added tool. Future uses are on the drawing boards as research scientists discover applications for tomorrow's business.

The financial services industry is a robust adopter, given the increasing sophistication of the market and

the high level of competition. To respond to some of this data complexity, Merrill Lynch uses Maps for its internal, proprietary data products to access real-time updates on stocks and mutual funds.

Merrill Lynch brokers can track client holdings individually or in aggregate, giving them an investment snapshot into their portfolio and their client's portfolio. The mapping technology can present data with size, value (through colors), and hierarchy, and allows users to click on and call up specific data sets from the map. Users can build dynamic, interactive, three-dimensional treetops from hierarchical company financial reports, share them with users via the Internet, and integrate them into existing applications.

Source: Adapted from Jack McCarthy, "Envisioning Enterprise Data," *Info World*, Vol. 24, No. 46, November 18, 2002, pp. 53-54.

ate different information presentations, especially with virtual reality (VR) methods. In Chapter 4, we discussed visual spreadsheets and visual interactive simulation. Data visualization enables these problem-solving methods in addition to providing graphic features to OLAP and data-mining tools.

Data visualization is easier to implement when the necessary data are in a data warehouse, or better yet in a multidimensional server. Harrah's does exactly this in managing its casinos. See DSS in Action 5.50. Our discussion focuses mainly on the concept of multidimensionality. In the next section, we present geographic information systems, a topic closely related to data visualization.

DSS IN ACTION 5.50**HARRAH'S VISUAL HURRAH!**

In 2002, Harrah's Entertainment, which runs 21 casinos in the United States, installed Compudigm International's visualization technology at its Las Vegas headquarters with plans to expand to more casinos. The visualization environment presents data in a form that allows decision-makers to see depth and worth in realtime and effect performance. Harrah's decision-makers can now view the flow of traffic across the casino floor in real-time. They can identify which slot machines are popular with the customers and which are most profitable by the minute. They can install more of the better

ones when needed. The data-visualization software also enables managers to determine casino layout on-the-fly. They can examine their *Rewards* program visually. Compudigm initially developed this product for the gaming industry, but since has extended its technology into financial services and telecommunications. See the Chapter 1 Opening Vignette for more details on how Harrah's utilizes business intelligence tools.

Source: Adapted from Brian Ploskina, "The Gestalt of Data," *Interactive Week*, June 18, 2001.

NEW DIRECTIONS IN DATA VISUALIZATION

Since the late 1990s data visualization has moved both into mainstream computing, where it is integrated with decision support tools and applications, and into intelligent visualization which includes data (information) interpretation. The following are some interesting areas:

- Interactive graphs and models that let users drill down into the underlying data to reorganize and compare data so that their meaning is clearer. Visualization tools can be useful in three areas: (1) statistical analysis, (2) graphical presentation tools, and (3) analytic applications.
- WatchMark Corporation, a subsidiary of Lucent Technologies uses a sophisticated data-visualization tool for wireless network operators. WatchMark Pilot Release 1.3 incorporates an innovative video replay engine with VCR-like controls, which enables network operators to quickly review the events that preceded a network problem, much like viewing an instant replay of a televised sporting event.
- Comshare Inc. provides Open Viz so that users can interact with images and data in meaningful ways. This reaffirms the notion that sophisticated visualization solutions now belong on the desktops of business professionals. Open Viz is a suite of components-supporting both Microsoft Common Object Model (COM) and JavaBean models-that enable IT developers to extend commercial and customdeveloped business intelligence solutions to encompass business-class data visualization.
- Identitech Inc. has developed Graphical Interface for Information Cognition, a data-visualization tool designed to support business decision-making. This software can be programmed to map data to sets of rectangles whose colors symbolize different levels of conditions, such as normal, high, and low.
- Analogous to a *visual spreadsheet* (Chapter 4), Visual Insights ADVIZOR allows users to find and understand patterns and trends hidden in complex data. It combines ease of use, industry-standard data access, and the power of interactive data visualization to create the next-generation user interface for business decision-making.
- There is an emerging new category of enterprise data visualization applications, termed on-line visualization for an enterprise (OLIVE). OLIVE systems are chart-centric applications that deliver visual business intelligence to the enterprise. There are 12 attributes that an enterprise charting application tool should have to qualify as an OLIVE tool, including (1) chart definition language and (2) a lifecycle process (see Craig, 1998).
- Visual software to reduce fraud and mitigate risk, especially in law enforcement, is an area of major developments. ChoicePoint (Alpharetta, Georgia, www.choicepoint.net) provides such software and services. Many organizations developing tools for the Department of Homeland Security are producing business intelligence spin-offs in law enforcement. See DSS in Action 5.38.
- Developments in virtual reality (VR) have wide-ranging impacts in business as well as other fields. See DSS in Action 5.51 for some applications and a Web software sampler. In addition, ChoicePoint's age-progression software, a form of predictive visualization, helps find missing children. By early 2003, ChoicePoint had assisted in recovering 782 missing children. See its Web site to see the results of age progression software.
- On the hardware side, there are continually new developments in visualization. Some involve special headgear or eyeglasses, others utilize holographic projec-

DSS IN ACTION 5.51

VIRTUAL REALITY VISUALIZATION



Visual applications include the latest developments in virtual reality (VR), which of course include more than just seeing images. Virtual reality representations have enabled advances in medicine, especially in training.

VR-simulations provide a way to train doctors and dentists on the look-and-feel of real surgical procedures. Three-dimensional images of organs (gall bladders, hearts, etc.) have enabled robotic surgery. VR can be used to treat phobias (fear of flying, thunderstorms, etc.). Haptics (virtual-touch technology) in conjunction with VR is accelerating applications. The Harvard School of Dental Medicine is working on a facility to enable dental training via haptics. Surgical applications are under development to accurately provide the texture, weight, and fragility of real body parts. Ortho Biotech Inc. has developed a mobile virtual reality simulator to help doctors understand how chemotherapy patients feel physically. Most doctors who try the simulator change how they talk about and treat cancer-related fatigue.

In Calgary, Alberta, the Cave Automated Virtual Environment is housed in a 3 by 3 meter room (possibly a forerunner of the holodeck of *Star Trek* fame). The Cave Project runs single cell simulations, cancer cell simulations, and a model human in a Java-based 3-D language. The project is at the forefront of bioinformatics. Eventually, Cave plans to develop 3-D models of diseases progressing through the human body.

VR has been used for flight training for years.

Pilots can learn their manual and technical skills through virtual reality-based simulations before assuming real flight responsibilities. Automobile manufacturers use virtual reality with simulations to help solve design problems and reduce costs. Math Works, Inc. provides a Virtual Reality Toolbox as part of its MATLAB and Simulink products. The Toolbox gives engineers an in-depth animated look at dynamic models.

Finally, VR is making major headway through Web applications, especially through the Virtual Reality Markup Language (VRML). For example, Land's End uses Web 3-D technology (My Virtual Model by Public Technologies Multimedia, Montreal, Quebec) to help shoppers evaluate garments on lifelike models. Three-D Web technology includes products like MetaCreations' MetaStream 3-D streaming format, Flatland Online's 3DML (markup language), Play Inc.'s Amorphium graphics engine, Oz.Com's Fluid3D plug-in for RealNetworks' RealPlayer G2; and Cycore's Cult3D modeling application.

Sources: Adapted from c.T. Heun, "Virtual Landscape: A Fantastic Human Voyage," *Information Week*, April 15, 2002; K. McMasters, "Almost There," *PC Magazine*, November 13, 2001; Business Wire, "Virtual Reality Doctors Empathize with Cancer Patients," *Business Wire*, November 1998; I. Edwards, "3-D Finally Gets Serious," *CIO*, March 1, 2000.

tion. Typically, these involve virtual reality representations of data as landscapes. Others are television-based. For example, Actuality Systems, Burlington, Massachusetts (www.actuality-systems.com), has developed a three-dimensional display for television. The Perspecta display is a 20-inch (51-cm) sphere in which 3-D images float. By mid 2003, ten displays had been sold (at \$40,000 each). NASA uses it to view star clusters, while the U.S. Army displays battlefield simulations. Experts are experimenting with the display in other areas, including medicine, energy, and software research. Eventually the company plans to develop a display system that eliminates the globe and is reasonably priced for general consumers. See the firm's Web site for sample images. Also see Copeland (2003).

Major OLAP vendors provide three-dimensional visualization tools with their decision support tools. For example, Forest Tree 6.0 is a Web-enabled development tool with a three-dimensional visualization version that enables users to visualize and easily manage multiple dimensions of data in a single view. New visual tools are continually being developed to analyze Web site performance. EBizinsights XL is one such tool. See Section 5.14 for more.

MULTIDIMENSIONALITY

Spreadsheet tables have two dimensions. Information with three or more dimensions can be presented by using a set of two-dimensional tables or a fairly complex table. In decision support, an attempt is made to simplify information presentation and allow the user to easily and quickly change the structure of tables to make them more meaningful (e.g., by flipping columns and rows, aggregating several rows and columns, or disaggregating a set of rows or columns—drill down).

MULTIDIMENSIONAL PRESENTATION

Summary data can be organized in different ways for analysis and presentation. An efficient way to do this is called multidimensionality. The major advantage of multidimensionality is that data can be organized the way managers rather than system analysts like to see them. Different presentations of the same data can be arranged easily and quickly.

Underlying every OLAP (and data mining) system is a conceptual data model often referred to as the multidimensional data model or multidimensional modeling (MDM). This technique helps conceptualize business models as a set of measures described by ordinary facets of business. The method is particularly useful for sifting, summarizing, and arranging data to facilitate analysis. In contrast to the techniques for designing online transaction processing systems, which rely on entities, relationships, functional decomposition, and state transition analysis, MDM utilizes the constructs of facts, dimensions, hierarchies, and sparsity. Choosing an appropriate tool requires examining the criteria of functionality, fit, performance, scalability, and future use. See Raden (1997).

Three factors are considered in multidimensionality: dimensions, measures, and time. Here are some examples:

- *Dimensions:* products, salespeople, market segments, business units, geographic locations, distribution channels, countries, industries
- *Measures:* money, sales volume, head count, inventory profit, actual vs. forecasted
- *Time:* daily, weekly, monthly, quarterly, yearly.

A manager may want to know the sales of a product in a certain geographic area, by a specific salesperson, during a specified month, or in terms of units. The answer to such a question can be provided regardless of the database structure, but it can be provided much faster, and by the user, if the data are organized in multidimensional databases or if the query or related software products are designed for multidimensionality. In either case, users can navigate through the many dimensions and levels of data via tables or graphs and are able to make quick interpretations, such as uncovering significant deviations or important trends.

Multidimensionality has some limitations, according to a Gartner Group research report (Gray and Watson, 1998):

- The multidimensional database can take up significantly more computer storage room than a summarized relational database.
- Multidimensional products cost significantly more, percentage-wise, than standard relational products.
- Database loading consumes system resources and time, depending on data volume and number of dimensions.
- Interfaces and maintenance are more complex than in relational databases.

Multidimensionality is available in different degrees of sophistication. Thus there are several types of software from which multidimensional systems can be constructed

at different price levels. Multidimensionality is especially important in *DSS/BI/BA* systems, including *enterprise information systems* (e.g., Decision Web from Cornshare Inc. . . , www.cornshare.com, and Pilot Analysis Server from Pilot Software Inc., www.pilotsw.com).

Tools with multidimensional capabilities often work in conjunction with database query systems and other OLAP tools. For example, IBM's Cube Views automates the creation of OLAP metadata at the database level so that metadata can be shared among applications that access the database. Cube Views aggregates data into multidimensional charts, allowing users to access the data from different perspectives, and returns answers to queries as XML-based Web services. Cube Views is supported by many business intelligence vendors, including Brio Software Inc., Crystal Decisions Inc., Cognos Inc., MicroStrategy Inc., Informatica Corp., InterNetivity, and BusinessObjects S.A. See Callaghan (2003b). Seagate Software's (part of Seagate Technology LLC) Crystal Reports builds reports that extract and analyze data from relational databases. This is part of the Crystal Enterprise software for distributing reports based on that information. Crystal Analysis Professional builds reports that extract and analyze multidimensional data from online analytical processing systems, such as Hyperion Essbase and Seagate Holos, as well as from mainstream databases, such as Microsoft SQL Server 2000, and IBM DB2 with built-in OLAP technology. See Whiting (2001). Other tools include Brio Enterprise (www.brio.com), PowerPlay (www.cognos.com), and InterNetivity Databeacon (www.internetivity.com), and Business Objects (www.businessobjects.com).

For examples of business intelligence software that readily handles multidimensionality, see Callaghan (2003b), Whiting (2001), and the "Annual Product Review" issue of *DM Review* every July (www.dmreview.com).

REAL-TIME ANALYTICS

A recent research study indicates that humans will record more information in the next three years than since the dawn of civilization. We need specialized methods to store our information in many formats, and to quickly retrieve and exploit it (see Pallatto, 2002a). Business users increasingly demand access to real-time, unstructured, or remote data, integrated with the contents of their data warehouse (see Devlin, 2003). For example, the buses in Houston, Texas, have been more reliable and efficient ever since they were equipped with instantaneous data gathering devices giving the drivers the ability to access information and modify traffic light changes (see "Houston Buses Due for 'Intellectual Overhaul,'" *ORMS Today*, June 2003, p.19). In many cases, real-time data updates and access are critical for the organization's success. See DSS in Action 5.52 for an example of real-time data collection and analysis, where they must be performed as a matter of life and death.

Data warehousing and business intelligence tools traditionally focus on assisting managers in making strategic and tactical decisions. In 2003, with the advent of realtime data warehousing, there was the start of a shift toward utilizing these technologies for operational decisions. This "active" use of data warehouses is just beginning to change the focus of these tools (see Coffee, 2003). See DSS in Focus 5.53 for some details of how the real-time concept evolved. Hewlett-Packard is moving toward an *Adaptive Enterprise* strategy for delivering on-demand computing (see Follett, 2003).

The trend to business intelligence software producing real-time data updates for real-time analysis and real-time decision-making is growing rapidly (see Baer, 2002; CIO Insight, 2003; Coffee, 2003; Devlin, 2003; Gates, 2002; Langseth and Vivatrat, 2002; Madsen, 2003; Pallatto, 2002a; Peterson, 2003; Raden, 2003a, 2003b; Barquin, Paller,

DSS IN ACTION 5.52**WHEN REAL-TIME DATA COLLECTION AND ANALYSIS MAKE RESPONSE TIMES REAL**

The City of Richmond, British Columbia, uses real-time data collection and analysis. Richmond is on a coastal island and has an average elevation of only 3 feet (1 meter) above sea level. It is important for city officials to know instantly whether its network of flood-control pumps is operating, how well, and why. Clearly, this is important in other parts of the world, as in The

Netherlands. Other excellent examples of the need for real-time data collection and analysis are in the source reference.

Source: Adapted from Tony Baer, "Analyzing Data in Real Time," *Application Development Trends*, April 2002, pp. 50-52.

DSS IN FOCUS 5.53**ACTIVE DATA WAREHOUSING: REAL-TIME REALITIES ~**

In 2003, an expansion of the role of data warehousing in practice was under way. Real-time systems were abuzz, along with all the usual complications of making data and information instantaneously available to those who need them. Peter Coffee believes that real-time systems must feed a real-time, decision-making process. Consequently, he was extremely interested in the remarks of Stephen Brobst, CTO of the Teradata division of NCR, speaking at a Chicago conference, "Creating the Real-Time Enterprise." Here is a summary of some comments:

Active data warehousing is a process of evolution in how an enterprise uses data. "Active" implies that the data warehouse becomes much more of an operational tool and represents much more of an opportunity to change the way the enterprise makes tactical decisions that dramatically increase its value to all its partners. Brobst provides a five-stage model that fits Coffee's experience of how organizations "grow" in their data utilization. The stages are reporting, analysis, prediction, operationalizing, and active warehousing:

Reporting. What happened?

Analysis. Why did it happen? *Prediction.*

What will happen? *Operationalizing.* What is happening?

Active warehousing. What do I want to happen?

"As the trend toward externalization escalates, the demand for near real-time decision support on organizationally consistent data is forcing IT groups to evaluate data management infrastructure agility. Organizations are enhancing centralized data ware-

houses to serve both operational and strategic decision-making." Anthony Bradley, META Group. Here is a comparison between traditional and active data warehousing environments.

Traditional Data Warehouse Environment

Strategic decisions only

Results sometimes hard to measure

Daily, weekly, monthly data currency acceptable; summaries often appropriate

Moderate user concurrency

Highly restrictive reporting used to confirm or check existing processes and patterns. Often uses pre-developed summary tables or data marts

Power users, knowledge workers, internal users

Active Data Warehouse Environment

Strategic and tactical decisions

Results measured with operations

Only comprehensive detailed data available within minutes is acceptable

High number (1000 or more) users accessing and querying the system simultaneously

Flexible ad hoc reporting, as well as machine assisted modeling (e.g., data mining) to discover new hypotheses and relationships

Operational staffs, call centers, external users

Source: Adapted from P. Coffee, " 'Active' Warehousing," *eWeek*, June 23, 2003; "Active Data Warehousing," Teradata Corp., www.teradata.com.

and Edelstein, 1997). Part of this push involves getting the right information to operational and tactical personnel so that they can utilize new business intelligence tools and up-to-the-minute results on which to base their decisions, since these employees generally deal with the short-term aspects of running an organization (see Chapter 2 for a discussion of Anthony's model).

Analytic systems continue to get faster, and many customers demand very current data. More and more IT managers are facing the expensive question of whether to take analytic systems real-time (see Baer, 2002). More and more *real-time* data warehousing/ analytics projects are under development and being deployed. The demand for real-time applications continues to grow. The proliferation of rules engines (business rules management), for example, creates pressure to implement more automated business processes that can best be implemented in a real-time data warehouse. When processes that require instantaneous updates are necessary for answering analytical questions, a real-time response is necessary. Query, OLAP, and data mining response times must be close to zero (see Raden, 2003a).

Real-time data warehouses are updated on a regular basis, not just weekly or monthly. In 2003, daily updating was expected; and the interval continued to shrink. In addition to real-time queries, business analytic applications are being deployed. The latter can instantaneously identify customer buying patterns based on store displays, and recommend immediate changes to placement or the display itself. Other applications include call-center support, fraud detection, revenue management, transportation, and many financial transactions. Obviously airlines, hotel chains, auto rental agencies, and even retail organizations in their *revenue management* efforts can update supply-and-demand elasticity curves to dynamically price their products and services (see Chapters 2 and 4).

On the other hand, an important issue in real-time computing is that not all data should be updated continuously. This may certainly cause problems when reports are generated in real-time, because one person's results may not match another person's. A company using BusinessObjects WebIntelligence noticed a significant problem with real-time intelligence. Real-time reports are all different when produced at slightly different times (see Peterson, 2003). Also, it may not be necessary to update certain data continuously, like course grades from three or more years ago.

Real-time requirements change the way we view the design of databases, data warehouses, OLAP, and data mining tools, since they are literally updated concurrently

DSS IN FOCUS 5.54

DATA WAREHOUSE ARCHITECTURES FOR REAL-TIME ANALYTICS



The primary schema for a data warehouse is either a star schema or a "normalized" schema. The latter is a term so loosely defined that it is hard to describe, but a normalized schema typically resembles a third (or higher) normal form (3NF) schema that is not dimensional. These 3NF designs do not support query and analysis. Their only purpose is to act as a staging area, an upstream data repository for a series of star schemas, OLAP cubes, and other structures that are directly

queried by analysts. Teradata implementations are the exception to the rule. Because of the unique characteristics of the massively parallel architecture and database optimizer, Teradata can process analytical SQL against a 3NF schema with acceptable performance.

Source: Adapted from Neil Raden, "Real Time: Get Real, Part II," *Intelligent Enterprise*, June 30, 2003, p. 16.

while queries are active. On the other hand, the substantial business value in doing so has been demonstrated, so it is crucial that organizations adopt these methods in their business processes. See DSS in Focus 5.54.

Examples of Web-based, real-time business intelligence software include BusinessObjects WebIntelligence, Cognos Supply Chain Analytics and BI Series 7, DataMirror Livebusiness, IBM DB2 Intelligent Miner Scoring (IMS), Informatica Analytics Delivery Platform, Informatica PowerAnalyzer, InterNetivity Databeacon, KnowNow LiveSheet for Excel, NetIQ Corp. WebTrends, PeopleSoft Enterprise Performance Management, SAS Supply Chain Intelligence Suite (SAS), and Sonic Software SonicMQ. For reviews, see Havenstein (2003b), Lindquist (2003), and Wallace (2000).

5.13 GEOGRAPHIC INFORMATION SYSTEMS

A geographic information system (GIS) is a computer-based system for capturing, storing, checking, integrating, manipulating, and displaying data with digitized maps. Its most distinguishing characteristic is that every record or digital object has an identified geographic location. By integrating maps with spatially oriented (geographic location) databases (called *geocoding*) and other databases, users can generate information for planning, problem-solving, and decision-making, thereby increasing their productivity and the quality of their decisions, as many banks and large retailers have done. Areas as diverse as retailing, banking, grocery, agriculture, natural resource management, public administration, NASA, the military, emergency preparedness, and urban planning have all successfully used GIS since the beginning of the 1970s.

Spatial data have become very important to many organizations. They are a new basis on which to manage infrastructures. As GIS tools and data sources become increasingly sophisticated and affordable, they help more companies and governments to understand precisely where their trucks, workers, and resources are, where they need to go to service a customer, and the best way to get from here to there. The areas of targeted marketing are growing rapidly. Organizations can easily segment a population. For example, the Credit Union of Texas (Dallas, Texas) utilizes a GIS to help decide where to place billboard and ATMs, and to help identify the areas most responsive to direct mailing. The typical response rates for the credit union is from 5 to 10 percent, much better than the average of 1 to 2 percent. Customers also enjoy receiving less mail from the credit union. They receive only relevant mailings. See Franklin (2002) for details. See DSS in Action 5.55 for some important examples.

Banks use GIS for displays that support

- Determining branch and ATM locations
- Analyzing customer demographics (e.g., residence, age, income level) for each of the bank's products
- Analyzing volume and traffic patterns of business activities
- Analyzing the geographic area served by each branch
- Finding the market potential for banking activities
- Evaluating strengths and weaknesses against those of the competition
- Evaluating branch performance.

A GIS is used as a geographic spreadsheet that allows managers to model business activities and perform what-if analyses (e.g., What if we close a branch or merge

DSS IN ACTION 5.55

CIS AND CPS TRACK WHERE YOU ARE AND
HELP YOU WITH WHAT YOU DO

Here are some examples of how GIS, in conjunction with GPS, helps firms and governments keep track of and improve their efforts. GIS helps companies differentiate their delivery services and meet demand for ever-shrinking delivery windows.

UltraEx, a West Coast company that specializes in same-day deliveries (of items like emergency blood supplies and computer parts), equips all of its vehicles with @Road's GPS receivers and wireless modems. In addition to giving dispatchers a big-picture view of the entire fleet, @Road helps UltraEx keep clients happy by letting *them* track the location and speed of their shipments on the Web in real-time. This Delivery 411 service, which UltraEx codeveloped with @Road, shows customers a map of the last place the satellite detected the delivery vehicle and how fast it was traveling. Dispatchers can choose the closest driver for each job, and drivers who own their vehicles are unable to falsify mileage sheets because @Road reports exact mileage for each vehicle. UltraEx spends roughly \$2 a day per vehicle to have @Road, "but if the driver can make one more pickup per day, we're way ahead," says Michael Oakes, vice president of business development at UltraEx.

Publix Direct, the online grocery service of Publix Supermarkets, uses GIS-enabled logistics software from Descartes to optimize delivery routes. When a customer places an order, the software does an on-the-fly analysis to determine the most profitable delivery windows given the customer's location, order size, other scheduled deliveries in that zone, and estimates of driving and service times based on data from Navigation Technologies. Within five to 15 seconds, the customer sees delivery-time options that would be most costeffective for Publix. Customers choose a 90-minute window, and then get a confirmation e-mail giving a 60minute estimated time of arrival on the day of the delivery. The software is so accurate that Publix Direct handles more than 7,000 orders a week-and delivers 97 percent of them on time. "The economics of delivery are a make-or-break facet of this business," says Jim Cossin, director of fulfillment operations for Publix Direct. "This allows us to balance the convenience factor with the customer, offering them as many possible'

windows as we can, while at the same time creating economically feasible routes in the background."

Location is germane to virtually every government function, and many municipalities are at the forefront of applying GIS. New York City pioneered CompStat, which uses GIS to map criminal activity and police deployment by date, time, and location. By making precinct commanders accountable for their own policing strategies, it has been a major factor in reducing the city's violent crime rate by nearly 70 percent in the past decade, says Lawrence Knafo, deputy commissioner of New York City's Department of IT and Telecom (DO ITT). In March 2003, New York expanded its use of GIS to launch a 311 (telephone) service to handle nonemergency service requests. (In most of the United States, dialing 911 on the telephone will connect you directly to the police.) Calls are entered into a CRM system that taps into GIS databases to verify callers' addresses and cross-streets before city workers are dispatched. Operators can access location-based information, such as garbage pickup times, and contact information for local elected officials. Beyond enabling efficient responses to service requests, the system allows the city to aggregate-and map, spatially and temporally-311 data across service sectors. *Geocoding* the calls makes it possible to analyze how well (or poorly) the city is providing services, helping policy-makers decide *how best to allocate scarce resources*. Knafo thinks that analysis of geocoded 311 and 911 data *could reveal previously unnoticed patterns in quality-of-life complaints that tend to precede violent crimes*. He says, "We might be able to actually stop crime before it happens."

Some police departments, neighborhood activists, and concerned citizens in other cities are utilizing GIS to fight crime. Geographical information about neighborhoods is integrated with crime reports to analyze crime patterns. By identifying trends and providing information to citizens, police are better able to set up surveillance activities, and citizens can modify behaviors, leading to lower crime in these areas.

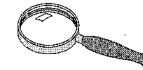
Source: Adapted partly from A. Dragoon, "Putting IT' on the Map," C/O, May 15, 2003.

branches? What if a competitor opens a branch?). Each map consolidates pages of analysis. Some pioneering banks are First Florida Banks (Tampa, Florida) and NJB Financial (Princeton, New Jersey).

For many companies, the intelligent organization of data within a GIS can provide a framework to support the process of decision-making and of designing alternative strategies. Some examples of successful GIS applications are summarized in Table 5.8. Leading companies incorporate geographical information systems into their business intelligence systems. GIS *ideally* incorporate census data (see www.census.gov) as a source of demographic data for effective decision-making (see Gimes, 2001). For many organizations, GIS and related spatial analysis are a top priority. Sears invested several million dollars in GIS technology for logistics leading to a savings of \$52 million per year (see Gonzales, 2003). The U.S. Defense Department has invested some \$21 billion in the satellite system that feeds *Geophysical positional systems (GPS)*. GPS devices detect their position on earth within a reasonable precision to couple these devices with mapping software. GPS in conjunction with GIS are making major inroads in business intelligence applications. Commercial and government uses are endless, since detection devices are relatively inexpensive. See DSS in Action 5.55 and DSS in Action 5.56 for examples of how these technologies have been used and potentially could be used.

TABLE 5.8 GIS Applications

<u>Organization</u>	<u>GIS Application</u>
Pepsi Cola Inc., Super Value, Acordia Inc.	Used in site selection for new Taco Bell and Pizza Hut restaurants; combining demographic data and traffic patterns
CIGNA (health insurance)	Uses GIS to answer such questions as, How many CIGNA-affiliated physicians are available within an 8-mile radius of a business?
Western Auto (a subsidiary of Sears)	Integrates data with GIS to create a detailed demographic profile of store's neighborhood to determine the best product mix to offer at the store
Sears, Roebuck & Co.	Uses GIS to support planning of truck routes
Health maintenance organizations	Tracks cancer rate and that of other diseases to determine expansion strategy and allocation of expensive equipment in their facilities
Wood Personnel Services (employment agencies) Wilkening & Co. (consulting services)	Maps neighborhoods where temporary workers live to locate marketing and recruiting cities
CellularOne Corporation	Designs optimal sales territories and routes for their clients, reducing travel costs by 15 percent
Sun Microsystems	Maps its entire cellular network to identify clusters of call ' disconnects and to dispatch technicians accordingly
Consolidated Rail Corporation	Manages leased property in dozens of places worldwide
Federal Emergency Management Agency	Monitors the condition of 20,000 miles of railroad track and thousands of parcels of adjoining land
Toyota (and other car manufacturers)	Assesses the damage of hurricanes, floods, and other natural disasters by relating videotapes of the damage to digitized maps of properties
	Combines GIS and GPS as a navigation tool
	Directs drivers to destinations via the best route

DSS IN ACTION 5.56**CIS + CPS HELPS RAILROAD COMPANIES
FIND THEIR CARS, AND MAYBE EVEN DRIVE TRAINS**

CSX Transportation Inc. has equipped 3100 locomotives with a Global Positioning System. Union Pacific Railroad has installed satellite-based monitoring devices on 100 of its freight cars (out of 155,308) to test car tracking. By combining GIS with a GPS, a freight company can identify the position of a railroad car or truck within 100 meters. Railroad companies can readily identify locomotives that have left their route, and the specific cars that have been left behind or sent with the wrong locomotive. Further benefits include the ability to prevent accidents. In the future, it may be possible to drive trains and other vehicles by using such systems in conjunction with artificial intelligence methods (as NASA uses in the Mars Rovers). For example, at The

University of Georgia's National Environmentally Sound Production Agriculture Laboratory, scientists are developing a farm tractor that is controlled by a gyroscope and GPS. Bad weather and visibility issues are not a problem in that the tractor knows where it is. Scientists have not yet developed devices that would let the system detect small obstacles. This could readily be accomplished with a robotic vision system and artificial intelligence methods to interpret what it sees.

Sources: Adapted partly from L. Rosencrance, "Railroads Hot for Satellite Monitoring," *ComputerWorld*, April, 2000; E. Minor, "Look Ma, No Hands!" *Athens Daily News* (AP), 2001, March, 2001.

CIS AND THE WEB/INTERNET/INTRANET

Most major GIS software vendors provide Web access, such as embedded browsers, or a Web/Internet/intranet server that hooks directly to their software. Thus, users can access dynamic maps and data via the Internet or a corporate intra net. GIS Web services are proliferating. These geographical systems form an information-rich global infrastructure that adds a new dimension to the GIS industry by integrating multiple and disparate application services. GIS Web services is revolutionizing how companies use and interact with geospatial information. For example, GIS can help the manager of a retail operation determine where to open a store located on a major city intersection, within a is-minute drive of a freeway exit ramp, surrounded by middle-class neighborhoods with professional families. See Gonzales (2003). Big Horn Computer Services (Buffalo, New York) uses a Web-adapted GIS to develop a custom application for a national television network that wants its affiliate stations to be able to access an intranet containing demographic information about their viewers. Using a Web browser, employees at each station can view thematically shaded maps analyzing their market.

A number of firms are deploying GIS on the Internet for internal use or for use by their customers. For example, Visa Plus, which operates a network of automated teller machines, has developed a GIS application that lets Internet users call up a map to locate any of the company's 257,000 ATM machines worldwide. As GIS Web server software is deployed by vendors, more applications will be developed. Maps, GIS data, and information about GIS are available over the Web through a number of vendors and public agencies. Related to this is the inclusion of spatial data in data warehouses, for later use with Web technology.

Some important GIS software are ArcView and ArcInfo (ESRI), AGISMap (AGIS), GeoMedia (Intergraph), and MapInfo Professional (MapInfo). ArcInfo's (www.esri.com/arcinfo/) data model provides tools to model complex spatial systems with no programming. Culpepper (2002) describes how the CommunityViz (www.communityviz.com) software integrates city-planning simulation and modeling functionality to ESRI's ArcView GIS software. The user can set up and run different

scenarios, based on user-specified variables and constraints, to determine relationships among municipal projects and social, environmental, or economic indicators. Entire sets of policies may be tested.

Current trends for GIS as a decision support/business intelligence tool involve continuing the combination or integration of GIS with other, especially Web-based, decision support/business intelligence tools, such as data warehouses, ERP, collaboration tools, and personal productivity applications. GIS data can integrate into other systems via XML through the Geography Markup Language (GML) (see Lais, 2000). One critical area that GIS have been successfully integrated into is CRM (see Dragoon, 2003a; Winslow and Lea, 2002; Sonnen, 1999; and DSS in Action 5.55). For further details on GIS, GPS, and the Web, see Dragoon (2003a), Duffy (2002), Hapgood (2001), Korte (2001), Kowal (2002), Lais (2000,2001), Leatham (2000), Price and Schweitzer (2002), and Winslow and Lea (2002).

5 .. 14 BUSINESS INTELLIGENCE AND THE WEB: WEB INTELLIGENCE/WEB ANALYTICS

BUSINESS INTELLIGENCE

Business intelligence activities—from data acquisition, through warehousing, to mining—can be performed with Web tools or are interrelated with Web technologies and electronic commerce. Specifically, business intelligence tools can be used to analyze Web site performance in real-time. Users with browsers can log onto a system, make inquiries, get reports, and so on, in a real-time setting. This is done through intranets, and for outsiders via extranets (see www.informationadvantage.com; also, for a comprehensive discussion of business intelligence on the Web, see the white paper at businessobjects.com).

A 2001 IDC survey of 500 IT managers indicated that 20 percent of organizations having 500 or more employees are linking their business intelligence activities to the Internet (see Kudyba, 2002; Dash, 2001). Users definitely want to improve the application of business intelligence on and to the Web. The number of organizations that recognize the importance of doing so is steadily growing.

Electronic commerce software vendors are providing Web tools that connect the data warehouse with the e-commerce ordering and cataloging systems. One example is Tradelink, a product of Hitachi (www.hitachi.comj). Hitachi's e-commerce tool suite combines e-commerce activities, such as catalog management, payment applications, mass customization, and order management, with data warehouses (marts) and ERP systems. Case Application 5.2 indicates how a firm provided a Web-based self-service system so that end user customers could handle their own benefits accounts.

Data warehousing and decision support vendors are integrating their products with Web technologies and e-commerce, or creating new ones for the same purpose. Examples are Comshares DecisionWeb, Brio's eWarehouse (www.brio.com), Web Intelligence from Business Objects, Cognos's DataMerchant, and Hyperion's Appsource "wired for OLAP" product, which integrates OLAP with Web tools. Pilot's Internet Publisher incorporates Internet capabilities within the Pilot Decision Support Suite. IBM's Decision Edge and MicroStrategy's DSS Web are other tools that offer OLAP capabilities on the Intranet from anywhere in the corporation, using browsers, search engines, and other Web technologies. MicroStrategy offers DSS Agent and DSS Web for help in drilling down for detailed information, providing graphical views, pushing information to users' desktops, and more. Bringing interactive querying, reporting,

and other OLAP tasks to many users (company employees and business partners) via the Web can also be facilitated by using Oracle's Financial Analyzer and Sales Analyzer, Hummingbird BIIWeb and BI/Broker, and several of the products cited above.

Data marts continue to become much more popular in the Web environment. For example, Bell Canada uses its intranet extensively for fast data access to its multiple data marts (over 300 analysts; see *PCWeek*, July 28, 1997), and at Nabisco, the large food company, financial analysts track the profits and losses of 8000 products using Web browsers, saving millions of dollars (*Info World*, Sept. 28, 1998).

WEB ANALYTICS/WEB INTELLIGENCE

Web analytics and **Web intelligence** are the terms used to describe the application of business analytics/business intelligence to Web sites. The tools and methods are highly visual in nature (see Section 5.12). Schlegel (2003) describes the basics of Web analytics, and even provides a proposed clickstream analysis architecture. As clickstream operations increase, the amount of data to process will grow exponentially, and scalability issues will become critical for Web intelligence/Web analytics. See DSS in Focus 5.57 and Section 5.7. Werner and Abramson (2001) describe a method (based on sorting and aggregation) to process a billion records a day for a Web data warehouse. See also Hayes (2001) and Ruber (2003) for information about Web clickstream analysis. Langseth and Vivatrat (2002) discuss why proactive, Web-based business intelligence is a hallmark of the real-time enterprise. Sodhi and Aichlmayr (2001) discuss how and why we should embed specific analytical models in Web-based data mining tools. See DSS in Action 5.58 for information, with an example, about how Web analytics are used in practice.

Informatica Corp. has focused closely on using the Web to enable organizations to track business performance. Using the Informatica Enterprise business intelligence platform, organizations gather business performance metrics via voice systems, the Web, and wireless transmission. The Informatica Analytics Delivery Platform is an Internet-based system that provides real-time business performance results.

NetIQ Corp.'s WebTrends business intelligence tool focuses on real-time analysis of Web traffic and online transactions. WebTrends enables organizations to track con-

DSS IN FOCUS 5.57

THE CHALLENGES OF CLICKSTREAM ANALYSIS



There are many complications when dealing with Web intelligence/Web analytics. Here is a list of things to look out for when preparing to perform clickstream analysis:

- Data preparation can consume 80 percent of the project resources.
- Raw clickstream data must be obtained from multiple servers.
- Individual customer data are usually buried in a mass of other data about pages served, hosts, referring pages, and browser types.
- A single page request can generate multiple entries into server logs.
- Taking a sequence of log records and creating a session of page views involves lots of data cleansing to eliminate superfluous data.
- Identifying the sessions in the data stream is complex. It requires cookies or other session identification numbers in URLs.
- Proxy servers (where customer requests do not come from the home server) confuse the identity of a session and why it ended.

Source: Adapted from Edelstein, 2001, p. 80.

66.

DSS IN ACTION 5.58**WEB ANALYTICS IMPROVES
PERFORMANCE FOR ONLINE MERCHANTS**

Online merchants anxious to improve the return of their Web site investment must learn their visitors' actions in real-time. This goes well beyond performing clickstream analysis and collecting the transaction reports with separate tools.

Online sales grew 52 percent to \$78 billion in 2002, according to a Forrester Research report issued on January 28, 2003. E-commerce sales were fueled by growth in new product categories and retailer mastery of digital marketing, Web analytics, and multichannel marketing. Companies with traditional catalog businesses advanced online sales to reduce the burden on call center operations and to lower order-processing costs.

It is critical to understand customers' online behavior to determine how and what to market to them. Understanding and properly using the operating metrics of an e-commerce site can make or break a business. For example, special product promotions can be put online in a matter of days, versus the months required for expensive catalog revisions and nationwide mailings.

Web analytics can boost the bottom line. Yun-Hui Chong, Internet director for Newport News, says that the data her company receives about the activities of its 1.6 million monthly Web visitors enables it to look at the return on investment of all online marketing campaigns. "Based on that, we optimize our banners and the presentation of merchandise," she says. "We also use it to do clickstream analysis to understand how customers are reacting to our site." It has become particularly important to identify customers who abandon the site or just browse certain categories. The firm sends them very customized e-mail promotions about these categories. Since it began doing so, there have been significant increases in conversion and revenue per e-mail sent. Targeting browsers and abandoners via e-mail on three product categories experiencing the worst conversion rates resulted in a better than sixfold increase in revenue per e-mail sent, while the cost per order dropped some 83 percent. Web analytics clearly pays off!

Source: Adapted from Peter Ruber, "Analytics Improve Merchandising," *InternetWorld*, June 2003, pp.11-12.

DSS IN FOCUS 5.59**WEB SERVICES BUSINESS INTELLIGENCE TOOLS**

Here is a sample of business intelligence tools that support Web services, especially through XML integration:

Actuate Corp. (www.actuate.com): Actuate 6
Business Objects (www.businessobjects.com):
Business Objects, Web Intelligence,
BusinessObjects Developer Suite
ClearForest Corp. (www.clearforest.com):
Clear Research, Clear Events, ClearSight
Cognos, Inc. (www.cognos.com): Cognos Series 7,
Cognos Web Services SDK
Crystal Decisions (www.crystaldecisions.com):
Crystal Enterprise, Crystal Reports, Crystal
Analysis Professional
Dimensional Insight, Inc. (www.dimins.com):
DI-Diver, DI-ProDiver, DI-WebDiver,
DI-ReportDiver, DI-Broadcast

Hummingbird Ltd. (www.hummingbird.com):
Hummingbird BI

Information Builders Inc.
(www.informationbuilders.com): WebFocus

Insight Corp. (www.insight.com): StatServer,
Analytics Server

Microstrategy Inc. (www.microstrategy.com):
Microstrategy Web Universal, Microstrategy SDK

SQL Power Group Inc. (www.sqlpower.ca):
Power*Dashboard

Targit (www.targit.com): TargitAnalysis 2K2

Source: Adapted from Jack Vaughan, "XML Meets the Data Warehouse," *Application Development Trends*, January 2003, pp.27-30.

❖ CHAPTER HIGHLIGHTS

- Data exist in internal, external, and personal sources.
- External data are available on thousands of online Web sites, commercial databases, directories, reports, and so on.
- Data for MSS must be collected frequently in the field using one or several methods.
- MSS may have data problems, such as incorrect data, nontimely data, poorly measured and indexed data, too many data, or no data.
- Commercial online databases, such as CompuServe and Dow Jones Information Service, can be major sources of MSS/BI data.
- The Internet has become a major external data source for MSS/BI.
- Intranets are providing internal data for MSS/BI.
- Most major databases have Web links to enable direct query via Web browsers on client workstations.
- Data are usually organized in relational, hierarchical, or network architectures. For many MSSIBIIBA, the relational database type is preferable.
- Structured query language (SQL) is a standard means of access for querying relational databases.
- Multimedia databases have become increasingly more important for decision-making applications.
- Object-oriented databases are easy to use and can be accessed very quickly. They are especially useful in distributed MSS and complex DSS.
- One of the most critical objectives is to make databases intelligent so that users can find information quickly by themselves.
- A data warehouse is a specially constructed data repository where data are organized so that they can be easily accessed by end users for several applications.
- Data marts contain data on one topic (e.g., marketing). They can be a replication of a subset of data in the data warehouse. They are a less expensive solution that can be replaced by or can supplement a data warehouse.
- Data marts can be independent of or dependent on a data warehouse.
- Data multidimensionality enables people to view data quickly in different dimensions, even if the data are in different files and databases.
- Business intelligence is moving toward real-time capabilities.
- Visualization is an important business intelligence capability.
- Business intelligence methods include OLAP and data mining.
- Online analytical processing (OLAP) is a set of tools for timely data analysis. It is extremely important in MSSIBIIBA applications.
- Data mining is the discovery of knowledge in databases. It is often done on data in data warehouses.
- Data mining can be hypothesis-driven or discovery-driven.
- The Web continues to impact dramatically on how database management systems are developed and operate.

❖ KEY WORDS

- | | | |
|--------------------------------------|-------------------------------------|---|
| ◦ business analytics | • dependent data mart | • object-oriented database management system (OODBMS) |
| • business intelligence | • development technology | • online analytical processing (OLAP) |
| • client/server architecture | • discovery-driven data mining | • online (commercial) databases |
| • content-management system (CMS) | • document management systems (DMS) | • query tools |
| • data | • hypothesis-driven data mining | • relational database |
| • data mart | • independent data mart | • source systems |
| • data integrity | • information | • structured query language (SQL) |
| • data mining | • intelligent database | • user participation |
| ◦ data quality (DQ) | • Internet | • Web analytics |
| • data visualization | • knowledge | • Web intelligence |
| • data warehouse | • metadata | |
| • database management systems (DBMS) | • multidimensionality | |

❖ QUESTIONS FOR REVIEW

1. Define data, information, and knowledge. Identify two examples of each.
2. Describe the role of the Internet in MSS data management and business intelligence.

3. What is SQL? Why is it important?
4. List the major categories of data sources for an MSS/BI.
5. Why are data quality and data integrity so important?
6. Describe the benefits of commercial databases.
7. Define object-oriented database management.
8. Define document management.
9. Define a star schema.
10. What are intelligent databases, and why are they so popular?
11. How can an expert system provide a good interface to commercial databases?
12. Define data multidimensionality and a multidimensional database.
13. Describe why visualization is so important in business intelligence.
14. Define a data warehouse, and list some of its characteristics.
15. What is the difference between a database and a data warehouse?
16. Describe the role that a data warehouse can play in MSS. List its benefits.
17. Define a data mart and explain why they are important.
18. Describe OLAP.
19. Define data mining and list its major technologies.
20. What is meant by real-time business analytics?
21. Differentiate data mining, text mining, and Web mining.
22. Distinguish between KDD and data mining.
Explain how the Web is impacting business intelligence/businessanalytic methods and technologies.

• :: QUESTIONS FOR DISCUSSION

1. Relate data warehousing to OLAP and data visualization.
2. Discuss the relationship between multiple sources of data, including external data, and the data warehouse.
3. Explain the relationship between SQL and a DBMS.
4. Compare OLTP to OLAP.
5. Define and describe a commercial database (online) service. Name one or two with which you are familiar.
6. Explain the relationship between OLAP and data mining.
7. Describe multidimensionality and explain its potential benefits for MSS.
8. Identify a commercial DBMS provider. Prepare a short report that describes the services offered, the fees, and the process for obtaining the service.
9. It is said that a relational database is the best for DSS (as compared to hierarchical and network structures). Explain why.
10. Explain the issue of data quality and some of the measures one can take to improve it.
11. It is said that object-oriented DBMS are the best solutions to a complex (especially distributed) DSS. Explain.
12. What is a data warehouse, and what are its benefits? Why is Web accessibility important?
13. Describe the major dimensions of data quality.
14. Why is data quality such an important issue to an organization?
15. Discuss the benefits of DMS.
- ~6. Discuss what an organization should consider before making a decision to purchase data-mining software.
17. Distinguish data mining from other analytic tools.
18. Explain the process of text mining.
19. Describe the concepts underlying Web intelligence and Web analytics.

❖ EXERCISES

1. A university is installing a DSS for budget preparation, expense monitoring, and financial planning. There are four schools at the university and 18 departments. In addition, there are two research institutions and many administrative services. Prepare a diagram that shows how the DSS will be distributed to all users—Comment on the data and its sources for such a DSS. Suggest what decisions could be supported at each managerial level. Explain how OLAP can be utilized effectively by the university.
2. Typically, data on a university campus are stored in different physical locations for different purposes. For example, the registrar's office, the housing office, the individual departmental offices, the personnel office, the staff benefits office, and the fund-raising and development office may maintain separate unintegrated databases with student (and faculty) records.
 - a. Explain what problems can occur in obtaining data to support complex decisions.

- b. Explain how a data warehouse might help solve these problems.
 - c. Discuss some of the behavioral (political) and technical problems that can occur in developing and implementing a data warehouse in such an environment.
 - d. Visit or call several offices and departments at your university (or at your place of business) and determine how basic data on students and faculty (or customers and employees) are stored, maintained, and manipulated. Find out whether they have multiple databases and what chronic problems they encounter.
3. Review the list of data problems in Table 5.1. Provide additional suggestions for each category.
 4. The U.S. government spends millions of dollars gathering data on its population every 10 years (plus some mid-decade adjustments). Census data are critical in determining the representation of each state in the House of Representatives and the number of Electoral College votes to which each state is entitled in presidential elections. More important, census data provide information about U.S. markets. The demographics indicate family and gender makeup, income, educational level, and other information for states, metropolitan statistical areas (MSAs), and counties. Such data are available from various sources, including books, disks, CD-ROMs, and the World Wide Web (see Internet Exercise 6). In this exercise, we take a real-world view of external but readily available data.
 - a. Find an electronic source of standard census data files for states and MSAs.
 - b. Access the data and examine the file structures. Do the contents and organization of each make sense? Why or why not? If not, suggest improvements.
 - c. Load the state P1 data population table into a spreadsheet file (Excel if possible) and into a database file (Access if possible). How difficult was this? How could it have been made easier? Don't forget to delete the comments and U.S. totals (if present) at the top, for later use. Note that Washington, D.C., is listed as well. Print the table.
 - d. Using the state P1 population data, sort the data based on population size. What are the five most populous states and the five least populous states? Which five states have the greatest and least population densities? Which state has the most males, and which state has the most females? Which three states have the most people living on farms, and which state has the fewest lonely people? Which file type (spreadsheet or database) did you use, and why? What features made it easy to do these analyses?
 - e. Load the state basic Table P6 (household income) into a spreadsheet or database file. Which five states have the most people earning \$100,000 or more per year? Which five states have the highest percentages of people earning \$100,000 or more per year? Combine these data with data from Table P1 to determine which five states have the most people per square mile earning \$100,000 or more per year? Which file type (spreadsheet or database) did you use, and why? What features made it easy to do these analyses?
 - f. Data warehousing and data mining are used to combine data and identify patterns. Use data (load and save them in spreadsheet or database files) from the following files: P1: Population; P3: Persons by Age; P4: Households by Size; P6: Household Income; P8: Other Income Measures; and P9: Level of Education. Synthesize these tables into a usable set and determine whether there are any relationships at the state level between
 - i. **Population per square mile and education n.**
 - ii. **Income and age**
 - iii. Household size and education. Can you think of any other relationships to explore? If you can, do so. What made this task difficult or easy? Explain.
 - g. Examine the MSA data tables and see whether any of the relationships found for the state data above hold.
 - h. How does the profile of your MSA (or the one closest to where you live) compare with your state's census profile and with that of the entire United States? How did you determine this?
5. Given the following list of employees in a manufacturing company, use DBMS software or a spreadsheet to
 - a. Sort the employees by department
 - b. Sort the employees by salary in ascending order
 - c. Sort the employees by department and sort the employees of each department by age in ascending order
 - d. Calculate the average salary
 - e. Calculate the average salary of female employees
 - f. Calculate the average/age in Department A
 - g. List the names of females who were hired after December 31, 1995
 - h. Show the age distribution graphically (use a 5-year grouping) as a pie chart
 - i. Compute the linear regression relationship of salary vs. age for all employees
 - j. Compute the relationship for females and males independently. Is there a significant difference?

<i>Name</i>	<i>Gender</i>	<i>Age</i>	<i>Date Hired</i>	<i>Dept.</i>	<i>Salary</i>
Martin Dean	M	28	06 Jan. 88	A	\$22,000
Jane Hanson	F	35	15 Mar. 96	D	\$33,200
Daniel Smith	M	19	06 Dec. 90	C	\$18,500
Emily Brosmer	F	26	10 Jan. 88	B	\$27,000
Jessica Stone	F	45	26 May 83	A	\$38,900
Tom Obudzinski	M	38	01 Dec. 98	B	\$29,800
Kathleen Braun	F	32	18 Apr. 92	B	\$35,600
Lisa Gregory	F	48	03 Sept. 91	C	\$32,400
Timothy Parker	M	29	03 Aug. 93	A	\$21,200
Jessica Hibscher	F	53	30 July 94	D	\$38,900
Adam Handel	M	62	29 Nov. 97	A	\$40,250
Melissa Black	F	42	01 Dec. 89	B	\$26,400
Ray Ernster	M	29	02 July 87	C	\$23,200
Daniel Bairn	M	38	26 Feb. 88	C	\$31,000
Amy Melnikov	F	45	30 Apr. 86	A	\$36,400
Adrienne Cammizzo	F	30	15 June 86	A	\$25,400
Steven Knowless	M	48	22 Oct. 85	D	\$33,200
Patricia Salisbury	F	56	26 Feb. 84	B	\$42,600
Matthew Broekhuizen	M	44	01 Jan. 88	C	\$45,400
Sarah Parent	F	64	01 Jan. 99	A	\$38,200

6. Investigate the integration of data warehouse and GIS. Start with www.mapinfo.com.
7. Take a test drive of demos of Decision Web (Comshare) and of business intelligence from sterling.com, Temtec, Brio, and Cognos. Do not miss Sybase's free interactive CD on business intelligence (hosted by soccer star Alexi Lalas). Prepare a report.
8. Examine how new data capture devices such as RFID tags (see DSS in Action 5.3) help organizations to accurately identify and segment their customers for activities such as targeted marketing. Scan the literature and the Web, and develop five potential new applications (not in this text) of RFID technology. What issues could arise if a country's laws required such devices to be embedded in everyone's body for a national identification system?
9. Consider the problem facing the city of London (U.K.). Since February 17, 2003, the city has instituted

.. a fee for automobiles and trucks in the central city district. There are 816 cameras digitally photographing the license plate of every vehicle passing by. Computers read the plate numbers and match them against records in a database of cars for which the fee has been paid for that day. If a match is not found (and the system was initially only 90 percent accurate), the car owner receives a citation by mail. The citations range from about \$128 to \$192 depending upon when they are paid. Examine the issues pertaining to how this is done, the mistakes that are made, and the size of the databases involved, including that of the images from the license plates. Also examine how well the system is working by investigating press reports. (This exercise was inspired by Ray Hutton, "London on \$8 a Day!" *Car and Driver*, August 2003, pp. 130-131.)

.. INTERNET EXERCISES

1. Surf the Internet to find information about data warehousing. Identify some newsgroups that have an interest in this concept. Explore ABI/Inform in your library, e-library, and Yahoo for recent articles on the topic, including the areas of data mining, multidimensionality, and OLAP. Begin with www.dw-institute.org and the major vendors: sas.com, oracle.com, and ncr.com. Also check cio.com, dmreview.com, dssresources.com, and pwp.starnetic.com.
2. Survey some data-mining tools and vendors. Start with fairisaac.com, and www.egain.com. Also consult dmreview.com.
3. Contact some DBMS vendors and obtain information about their products. Special attention should be given to vendors that provide tools for multiple purposes, such as Cognos, Software A&G, SAS Institute, and Oracle. Free demos are available from some of these vendors over the Web (e.g., brio.com).

Download a demo or two and try them. Write a report describing your experience.

4. America Online, stock brokerages, and portals provide a free service that shows the status of investors' stock market portfolios, including profits (losses) and prices (with a 15-minute delay or even in real-time). How is such individualized information retrieved so quickly? Why must such data be updated so quickly?
5. What economic data are available from government agencies? Who provides what types of data? How easy are these data to access, download, and use? (Provide 10 examples.)
6. Search the Internet to identify sources of U.S. government census data files. Download and examine some of the files. Are they flat ASCII text data files, spreadsheet files, or database files (and what are their formats)? Were they compressed or archived? If so,

how easy was it to extract the data? Which tables (files) would be useful, and what kinds of analyses could be performed with such data (e.g., for a consumer product marketing firm, a financial services firm, an insurance company, and a real estate developer)?

7. Find recent cases of successful business intelligence applications. Try business intelligence vendors and look for cases or success stories.
8. Visit teradatauniversitynetwork.com. Identify cases on business intelligence, data warehousing, and data mining. Describe recent developments in the field.
9. Go to Web sites (especially, SAS; SPSS, Cognos, TemTec, Business Objects) and look at success stories for business intelligence (OLAP and data mining) tools. What do you find in common among the various success stories? How do they differ?

.. GROUP EXERCISES

1. Each group member will check a major DBMS vendor (Oracle, Sybase, Informix, and so on). Examine their major Web-related products. Explain the connection of the databases to data mining and to electronic commerce.
2. Data visualization is offered by all major data warehouse vendors, as well as by other companies, such as www.ilog.com. Students are assigned one to each vendor to find the products and their capabilities. (For a list of vendors, see www.dw-institute.org.) Each group summarizes the products and their capabilities.

3. Interview administrators in your college or executives in your organization to determine how data warehousing, data mining, OLAP, and visualization business intelligence/DSS tools could assist them in their work. Write up a proposal describing your findings. Include cost estimates and benefits in your report.
4. Go through the list of data warehousing risks in DSS in Action 5.19 and find three examples of each in practice.

DATA WAREHOUSING AND OLAP AT (ABELA'S

Cabela's, "the world's foremost outfitter," (Sidney, Nebraska) is the world's largest mail order distributor of products for outdoor enthusiasts. Cabela's has 6,000 employees. Every year Cabela's mails more than 60 million catalogs in 60 editions to customers across the entire United States and in 135 other countries. Cabela's also owns eight stores, an e-commerce Web site, and four telemarketing centers in the United States

In the mid-1990s, executives needed to develop a greater understanding of their customers' behaviors, individual tastes, and purchase preferences. They needed to characterize the different segments of the company's customer base. Essentially they needed a way to *cluster* or group their customers to understand them, and to target market specific products to the members of each cluster (segment)

At the time, Cabela's relied on outsourced business intelligence and in-house packaged solutions to build separate mailing lists for every single catalog and promotion. This process was costly and slow. In addition, the data's integrity was called into question.

Cabela's adopted IBM's DB2 Universal Data Enterprise Edition and IBM DB2 Warehouse Manager as its platform. Query response times are now 80 percent

faster than before; maintenance time and costs have been cut in half. The knowledge gleaned from the data warehouse has enabled the firm's marketing team to improve catalog hit rates and has led to far-reaching improvements to printed catalogs and the e-commerce Web site, enhancing the customer experience and boosting customer loyalty.

About 30 users (including four full-time statisticians and their staff, and senior managers) access the data warehouse with BRIO Explorer as the front-end querying and reporting tool, and SAS as the statistical analysis tool (both are OLAP tools). The warehouse contains 11 years of information stored in about 700 gigabytes.

Within a few months of deployment, sales in most market segments rose significantly. Since it has succeeded, improvements are already underway on the OLAP side to help better understand the crucial relationships among customers, markets, products, prices, and geography—the key factors that drive the business.

By leveraging data assets with additional data management and business intelligence technologies, Cabela's will achieve deeper and more powerful insights that will bring added value to its customers; and to the bottom line.

CASE QUESTIONS

1. Describe how Cabela's ran its marketing process before the system was developed.
2. Why is it important for a firm like Cabela's to segment its customers? What benefits can the firm obtain? Are there any disadvantages? Explain.
3. Why was it important for Cabela's to have kept 11 years of sales data on hand? Could the firm have used more?
4. How have OLAP tools helped Cabela's improve the performance of the business?
5. Describe how OLAP tools could help Cabela's do even better than it is doing with the system described.
6. Describe how data mining tools could help Cabela's do even better than it is with the system described.
7. Go to the Web sites of the vendors mentioned in the case and examine the current OLAP, data mining, and data warehousing features and capabilities of each. Describe in detail how Cabela's could use each one.
8. Describe the competitive nature of the system.

Adapted from Jerry Rivkind, "Cabela's Data Warehouse Satisfies Hunt for Business Insight." *What Works. Best Practices in Business Intelligence and Data Warehousing*, Vol. 15, The Data Warehousing Institute, Chatsworth, CA, June, 2003, pp. 8-9. '

CASE APPLICATION 5.2

BLUE CROSS AND BLUE SHIELD OF MINNESOTA'S PAIN-FREE CRM SAVES THE DAY THROUGH DATA INTEGRATION AND PLANNING

By taking the time to integrate data for its online CRM system, a regional health plan has succeeded where other large insurers have not. In 2001, John Ounjian, senior vice president and CIO of Blue Cross and Blue Shield (BCBS) of Minnesota, convinced General Mills, an \$8 billion consumer goods firm, to join his regional health plan on the promise that he would install a Web-based customer service system so that subscribers could manage their health benefits online. Subscribers could select health plans tailored to their individual needs and budgets, calculate their own contributions to their coverage, research information on prescription drugs and other treatments, locate participating physicians, and check the status of claims.

To implement this system, he needed to install a brand-new infrastructure to integrate his Web and call center operation and provide timely, accurate information to customers. He also had to migrate megabytes of data stored in back-end, legacy databases to the Web front end—massaging and reformatting the data so consumers could understand them.

The online customer self-service system (deployed in January 2002) not only met the specifications of General Mills, but also managed to beat national providers Aetna, Cigna, and Humana out of several very large accounts, such as 3M, Northwest Airlines and Target.

Ounjian says that his company's membership grew by 10 percent, or 200,000 new members, in 2002, largely because of its online customer self-service system, when several national insurers lost millions of members. In addition to offering a viable solution to customer problems, Web self-service also provides the necessary foundation for delivering health plans tailored to the needs of individual consumers—a direction in which the industry is moving to decrease managed care costs.

Ounjian realized that the plan needed to lay down a whole new infrastructure, or *chassis*. It would also need a sound data management strategy to overcome problems when it tried to move raw data from back-end systems to the Web front end. Ounjian uses Oracle databases on the front end to reassemble and synchronize back-end data so that consumers find consistent, timely information regardless of whether they use the Web channel or the call center channel. Ounjian indicates that he uses different vendors for the several components of the Web self-service architecture because there was no single vendor that could supply everything for an integrated CRM system. He wanted the flexibility to layer different applications and functionality from different vendors on top of his infrastructure.

Ounjian's biggest problem was to devise a tactical strategy for moving data and transactions from the front end to the back end, and vice versa. Making megabytes of back-end data available and understandable to users on the front end is one of the biggest challenges for any successful CRM project, regardless of industry. If you cannot get accurate information to customers in a format they understand, they will not use the system. The millions of records that had to be migrated made the task even more daunting.

Ounjian believes the reason why so many CRM projects—not just in health care but across industries—run into problems or fail altogether is because they aren't grounded by an underlying plan for transferring data that originates in one system and in one form to another system in a different form.

In early 2003, the system was being used by 61 employers with 450,000 individual employees. BCBS plans to increase that number throughout 2004.

CASE QUESTIONS

1. Describe how the BCBS system works.
2. Why was it so important to develop the infrastructure in advance of deploying the systems?
3. How did the BCBS system integrate various data sources?
4. How can users utilize the system as a decision-making/business intelligence system?
5. Describe how managers at BCBS could use the system in an OLAP and in a data mining framework.
6. Describe the competitive nature of the system.

Adapted from Meredith Levinson, "Pain-free CRM," *CIO*, May 15, 2003, pp. 79-80.

CASE APPLICATION 5.3

CLUSTER ANALYSIS FOR DATA MINING

INTRODUCTION

Cluster analysis is a very important set of methods for classifying items into common groupings called *clusters*. The methods are common in biology, medicine, genetics, the social sciences, anthropology, archaeology, astronomy, character recognition, and even in MIS development. As data mining has increased in popularity, the methods have been applied to business, especially to marketing. Cluster analysis has been used extensively for fraud detection, both credit card and e-commerce fraud, and market segmentation of customers in CRM systems. More applications in business continue to be developed as the strength of cluster analysis is understood and utilized.

CLUSTER ANALYSIS FOR DATA ANALYSIS

Cluster analysis is an exploratory data analysis tool for solving classification problems. The object is to sort cases (people, things, events, etc.) into groups, or clusters, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. Each cluster describes the class to which its members belong. An obvious one-dimensional example of cluster analysis is to establish score ranges into which to assign class grades for a college class. This is similar to the cluster analysis problem that the U.S. Treasury Department faced when establishing new tax brackets in the 1980s. A fictional example of clustering occurs in J. K. Rowling's *Harry Potter* books. The *Sorting Hat* determines which House (e.g., dormitory) to which to assign first-year students at the Hogwarts School. Another example involves how to seat guests at a wedding. As far as data mining goes, the importance of cluster analysis is that it may reveal associations and structures in data that were not previously apparent but are sensible and useful once found.

Cluster analysis results may be used to

- Help identify a classification scheme (e.g., types of customers)
- Suggest statistical models to describe populations
- Indicate rules for assigning new cases to classes for identification, targeting, and diagnostic purposes
- Provide measures of definition, size, and change in what were previously broad concepts
- Find typical cases to represent classes

CLUSTER ANALYSIS METHODS

Cluster analysis may be based on one or more of the following, general methods:

- Statistical (including both hierarchical and nonhierarchical)
- Optimal
- Neural networks
- Fuzzy logic
- Genetic algorithms

Each of these methods generally work with one of the following two general method classes:

- *Divisive*: all items start in one cluster and are broken apart.
- *Agglomerative*: all items start in individual clusters, and the clusters are joined together.

Most cluster analysis methods involve the use of a *distance* between pairs of items. That is, there is a measure of similarity between every pair of items to be clustered. Often they are based on true distances that are measured, but this need not be so, as is typically the case in IS development. Weighted averages may be used to establish these distances. For example, in an IS development project, individual modules of the system may be related by the similarity between their inputs, outputs, processes, and the specific data used. These factors are then aggregated, pairwise by item, into a single distance measure.

CLUSTERING EXAMPLE

Consider the similarity (distance) matrix that represents the similarities among eight items shown in Table 5.9. Items 4 and 5 have a lot in common, as do items 1 and 3, and 3 and 10; though 1 and 10 are moderately related, and 1 and 5 have little in common. To evaluate a solution, we add the pairwise values of all the items in each cluster. If we want three balanced clusters (between 2 and 3 items per cluster), the solution of clusters {1, 3, 6}, {2, 8}, and {4, 5, 7} have a value of $(9 + 6 + 10) + 8 + (10 + 8 + 9) = 60$. Can we do better? Try it!

Now that we have a data set, some critical issues to address are

TABLE 5.9 Similarity (Distance) Matrix

The values below the diagonal equal the values above; that is, the distance from 1 to 2 is the same as from 2 to 1. Diagonal values do not exist.

Item	1	2	3	4	5	6	7	8
1		3	9	2	1	6	4	5
2			4	5	6	2	3	8
3				5	7	10	4	2
4					10	2	8	1
5						4	9	3
6							3	3
7								5
8								

- How many clusters are to be found (when do we stop)?
- Should all the clusters have an approximately equal number of items?
- How do we handle dimensional scaling when different measures are used in establishing the distance value?
- Can distance really be measured?

HIERARCHICAL CLUSTERING METHOD AND EXAMPLE

We start with a set of items, each within its own cluster. We determine the maximum number of clusters we want to have. The basic method is to

1. Decide which data to record from your items (measures of similarity).
2. Calculate the distance between all initial clusters. Store the results in a distance matrix.
3. Search through the distance matrix and find the two most similar clusters.
4. Fuse these two clusters together to produce a cluster that now has at least two items.
5. Calculate the distances between this new cluster and all the other clusters (some of which may contain one item).
6. Repeat steps 3 to 5 until you have reached the pre-specified maximum number of clusters.

Note that some methods go all the way to a single cluster of all items. To identify the solution you want, identify where you have obtained the desired number of clusters, and stop.

Applying the hierarchical method to our matrix above with a goal of three balanced clusters, the initial solution is

11,2,3,4,5,6,7,8} at a value of 134. Though this is an excellent value, we want three clusters.

We first combine items 4 and 5 (value = 10) to get

14,5} 11,2,3,6,7, 8} at $10 + 74 = 94$, which is pretty good.

We next combine 3 and 6 (value = 10) to get

13,6} 14,5} 11,2,7, 8} at $10 + 10 + 28 = 48$.

Next, we combine 7 into the cluster with 4 and 5 to get

13,6} 14,5, 7} 11,2, 8} at = 53.

We stop because we have three clusters as balanced as can be with these data (two groups of three items, one group of two items).

The above example may actually be formulated as an optimization problem and solved with an efficient algorithm (see Aronson and Klein, 1991).

CLUSTER ANALYSIS SOFTWARE

Aside from data mining methods in which cluster analysis methods are embedded, there are several specialized packages for cluster analysis. These include:

- ClustanGraphics (Clustan)
- DecisionWORKS Suite (Advanced Software Applications)
- SPSS (SPSS)
- PolyAnalyst Cluster Engine (Megaputer)
- Sokal code (see Hand, 1981)

There are also many free codes available from academic sites. Do a Web search to find them.

EFECTIVE CLUSTER ANALYSIS/ DATA MINING APPLICATIONS

Goulet and Wishart (1996) provide an excellent example of how a bank was able to classify its customers to dramatically improve their financial services. The Co-operative Desjardin's Movement is the largest banking institution in Quebec (Canada). When this analysis was done, there were 1,329 branches and 4.2 million members. The organization had combined assets in excess of (Canada) \$80 billion. It was in the process of reducing teller services, increasing ATM use and other IT methods, and reducing staff. The organization, in addition to banking, offered products and services that included life and property insurance, and several others. Since each branch is independent, the Confederation needed to market its products to both its branches and its members. At the start of the study, the bank executives realized that they needed a typology of its members not only to retain customer loyalty, but also to capture more market share by identifying profitable services to satisfy members' needs and improve market penetration.

The bank performed a *cluster analysis* of a sample of 16,000 members. By doing so, it identified 16 variables that reflected the characteristics of financial transaction patterns. Thirty *member types* were identified. Next, all 4.2 million members were classified with best fits of the 16 measures, which were used to place them into *one or more* of the 30 member type clusters.

Now financial managers and analysts can identify members whose financial transactions fall into one or more of the 30 clusters. Given a particular member's cluster, the profitability of each transaction cluster and individual customer accounts can be measured. Each branch manager can view his or her customers as investments in a portfolio, and particular market segments for the branch's products and services are readily identified.

The results are impressive. The bank can identify members with large transaction volumes in one account by matching them to their other loan or insurance accounts. The managers can then suggest more economical consolidations of members' investments and loans, thus leading to a higher level of customer satisfaction. Additionally, managers can suggest better diversification of members' investments. But the most impressive results

are in the bank's marketing efforts. The bank can focus on products and services that have the best financial performance and target them to appropriate customers. This reduces mailing and other contact costs. Response rates have been improved by targeting product promotions achieving better branding and customer retention. In fact, more profitable customers are retained at lower costs.

FURTHER READING

For more details on cluster analysis, algorithms, and software, see Aldenderfer and Blashfield (1984), Aronson and Iyer (2001), Goulet and Wishart (1996), Hand (1981), Klein and Aronson (1991), Romesburg (1984), and Zupan (1982). Also, since Web sites change almost daily, we recommend that you perform a Web search on *cluster*, *cluster analysis*, and *cluster methods*. There are excellent academic sites, many of which include free computer codes.

A BETTER SOLUTION TO THE EXAMPLE

Though not balanced, the cluster solution 11,3, 4}, 13, 5, 6}, and 15} has a value *ofn*, better than the solutions described earlier.

Cluster Analysis References

- Aldenderfer, M.S., and R.K. Blashfield. (1984). *Cluster Analysis*. Thousand Oaks, CA: Sage.
- Aronson, IE., and L.S. Iyer. (2001). "Cluster Analysis." In *Encyclopedia of Operations Research & Management Science*, Norwell, MA: Kluwer.
- Goulet, M., and D. Wishart. (1996, June). "Classifying a Bank's Customers to Improve Their Financial Services." In *Proceedings of the Conference of the Classification Society of North America (CSNA)*, Amherst, MA.
- Hand, D. (1981). *Discrimination and Classification*. New York: Wiley.
- Klein, G., and J.E. Aronson. (1991). "Optimal Clustering: A Model and Method." *Naval Research Logistics*, Vol. 38, No. 1.
- Romesburg, H. (1984). *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning.
- Zupan, I (1982). *Clustering of Large Data Sets*. New York: Research Studies Press.

CASE QUESTIONS

1. Explain why cluster analysis is important for data mining.
2. Identify the different methods of performing cluster analysis. Study the literature and Web sites to determine the kinds of problems to which each method can appropriately be applied.
3. Explain how cluster analysis works.
4. Search the Web for free and cheap cluster analysis software. Download one and describe how it works and the types of problems it solves.
5. Describe how the Co-operative Desjardin's Movement performed its cluster analysis.
6. Describe the benefits obtained from cluster analysis by the Co-operative Desjardin's Movement.