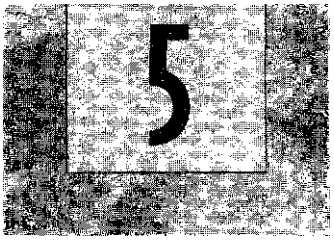
CH-APTER



M S S iNTtLLIGtHCt: 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[1]

Data warehouses 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. Then 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. For example, 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 oil 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 complying 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 RFID 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 cost-effective. 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," *InformationWeek*, 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.

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

Source: Based on Alter (1980), p. 130. Alter, S. L. (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 incorrectly 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 older 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 5.4.

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

TABLE 5.2 Source of Data Quality Problems

Source of Data Quality Problem	*Percent Response*
Data entry by employees	76
Changes to source systems	53
Data migration or conversion projects	48
Mixed expectations by users	46
External data	34
Systems errors	26
Data entry by customers !	25
Other	12

Source: Adapted from Wayne Eckerson, "Data Quality and the Bottom *Line," Application Development Trends,* May 2002, pp. 24-30.

-T*-

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 full-time positions devoted to ensuring data quality.

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.

- ***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), Dyche (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 Staekpole, "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), Qlson. (2003a, 2003b), Staekpole (2001), Stodder (2002), and Theodoratos arid Bouzeghoug (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 CIO 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 data-mart 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." *InformationWeek,* 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," *ComputerWorld,* June 16,2003, p. 7; Katherine Mclntire 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," *CRN,* No.1019, November 4,2002, pp. 3'9-40; 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)T 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 ENTERPRISE-WIDE 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.

WHAT NOT TO DO:

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

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.

CLEANING 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 supply-chain management; see Chapters 7 and 8). The Internet is a major source of data.

- The **Web/Internet.** 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-

TABLE 5.1 Ri-nriSi.nt: iUv '<>niisiui 0:il;ib;-sc (Data Hank) 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 (CMS)** 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 Bluelsle 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 Decision Web from Comshare (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 PL/SQL 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 Ofthe 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 Kroenke, 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—arc 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

Customer Number	Customer Name
B	Green
10	Brawn
30	Black
45	White

Customer Records

Product Number	Product Name
M.1	Nut
S.1	Bolt
T. 1	Washer
U.1	Screw

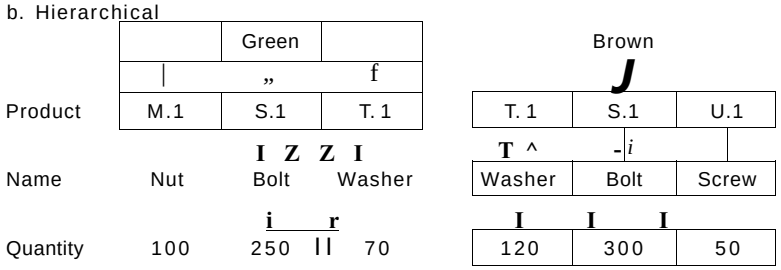
Product Records

Customer Name	Product Number	Quantity
Green	M.1	10
Brown	S.1	300
Green	11	70
White	S.1	30
Green	S.1	250
Brawn	T. 1	120
Brown	U.1	50

Usage Records

- Fields
- Records

b. Hierarchical



c. Network

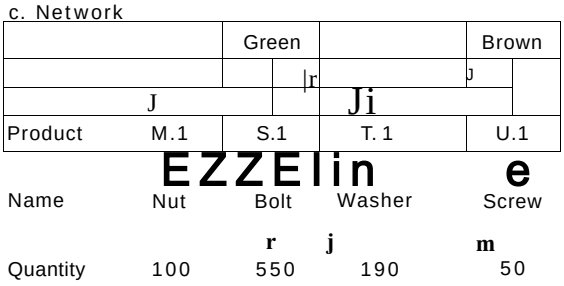


FIGURE 5.1 DATABASE STRUCTURES

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 S.l and T.l.

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 objected-oriented systems, may be useful.

Object-oriented data management u 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. Struct ure, 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-

DSS IN ACTION 5.12

G. 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 new systems 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.

timedia 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.[1]

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 MediaWay 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.

DSS IN ACTION 5.14

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 82-minute 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.

tems. 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.

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 (Al) 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," *ComputerWorld,* June 5,2000.

One of IBM's main initiatives in commercial Al 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 in Action 5.16). Data for the data warehouse are brought in from various external and internal

DSS IN ACTION 5.16

i s a *mm*

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.

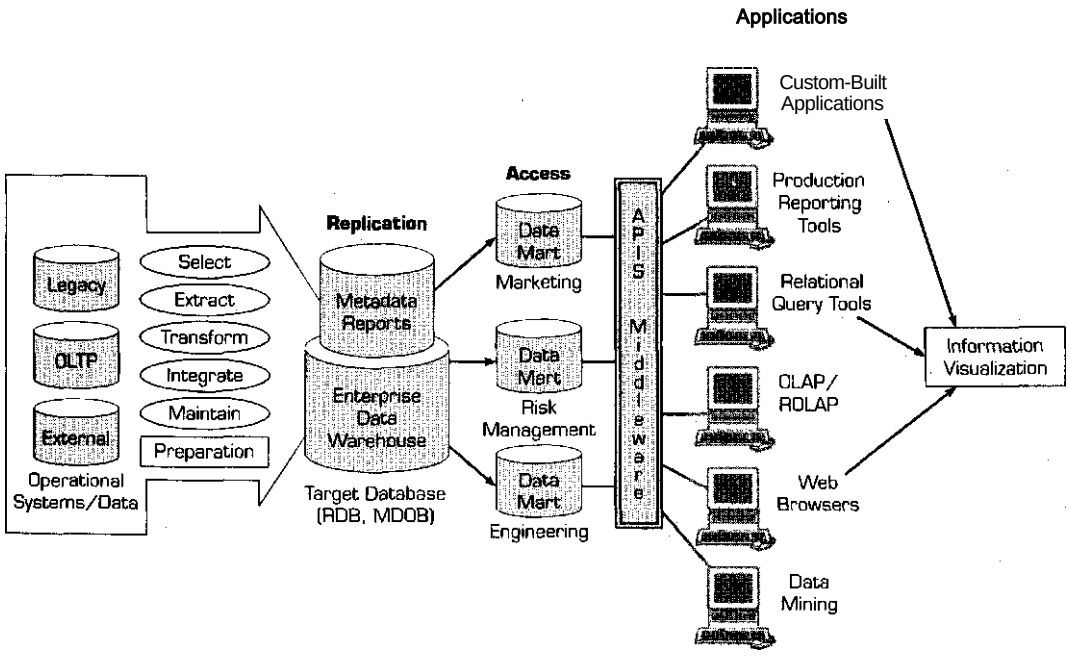
Source: Adapted from B. Cooper, H. J. Watson, B. H. Wixom, and D. Goodhue, "First American Tennessee Case Study," *MIS Quarterly,* 2004, forthcoming. Also presented as "Data Warehousing Supports Corporate Strategy at First American Corporation." SIM International's Best Paper Contest Recipients, 1999.

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 PCs 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 non-relational 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 domain-specific 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 DSS/BI/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-

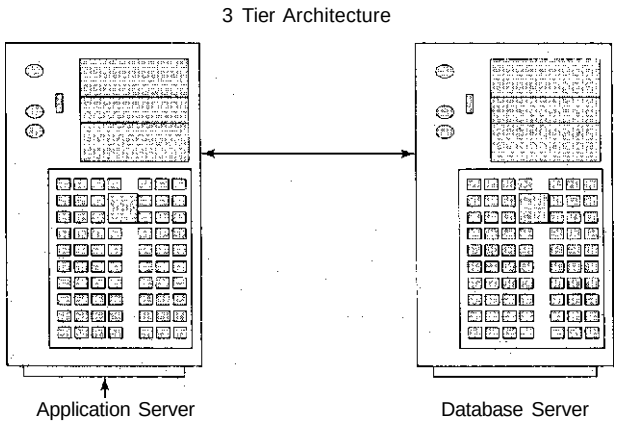


FIGURE 5.3 ARCHITECTURE or **A 3-IIIR DATA WAREHOUSE**

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),

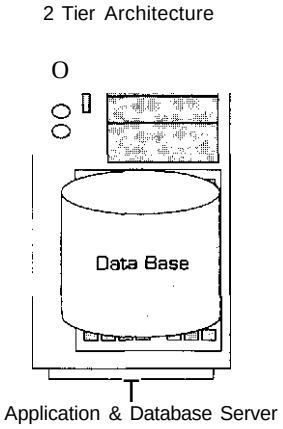


FIGURE 5.4 ARCHITECTURE OF A 2TIEK DATA WAREHOUSE

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 OLTP 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 ngeded 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.

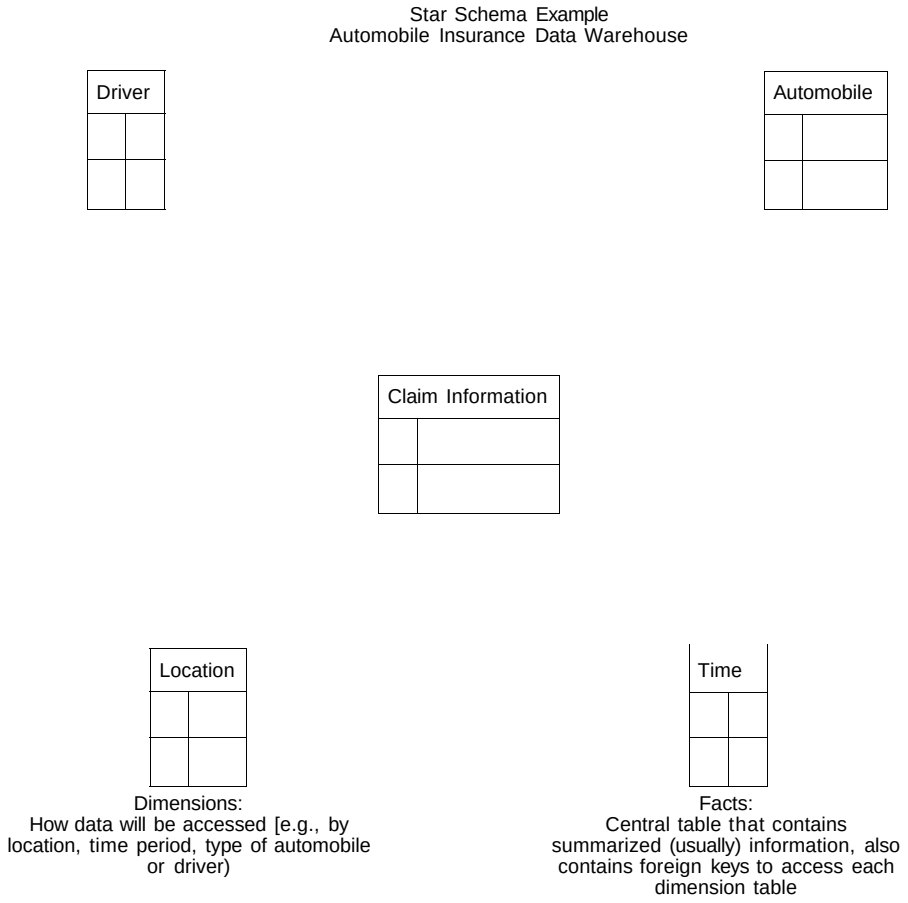


FIGURE 5.5 STAR SC HEMA

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 ETL 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.

DSS IN FOCUS 5.20

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 organization-wide. 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 supply-chain 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 information-technology 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 employees 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," *InformationWeek,* 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), Allasi (2001), Barquin and Edelstein (1997a, 1997b), Barquin, Paller, and Edelstein (1997), Deng (2003), Eckerson (2002b, 2003), Edelstein (4997), 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