**Foundations of Business Intelligence: Databases and**

**Information Management**

1. ORGANIZING DATA IN A TRADITIONAL FILE

ENVIRONMENT

An effective information system provides users with accurate, timely, and relevant information. Accurate information is free of errors. Information is timely when it is available to decision makers when it is needed. Information is relevant when it is useful and appropriate for the types of work and decisions that require it. You might be surprised to learn that many businesses don’t have timely, accurate, or relevant information because the data in their information systems have been poorly organized and maintained. That’s why data management is so essential. To understand the problem, let’s look at how information systems arrange data in computer files and traditional methods of file management.

FILE ORGANIZATION TERMS AND CONCEPTS

A computer system organizes data in a hierarchy that starts with bits and bytes and progresses to fields, records, files, and databases (see Figure 6-1). A bit represents the smallest unit of data a computer can handle. A group of bits, called a byte, represents a single character, which can be a letter. A computer system organizes data in a hierarchy that starts with the bit, which represents either a0 or a1. Bits can be grouped to form a byte to represent one character, number, or symbol. Bytes can be grouped to form a field, and related fields can be grouped to form a record. Related records can be collected to form a file, and related files can be organized into a database.

PROBLEMS WITH THE TRADITIONAL FILE ENVIRONMENT

In most organizations, systems tended to grow independently without a company-wide plan. Accounting, finance, manufacturing, human resources, and sales and marketing all developed their own systems and data files. Each application, of course, required its own files and its own computer program to operate. For example, the human resources functional area might

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

Data Redundancy and Inconsistency

Data redundancy is the presence of duplicate data in multiple data files so that the same data are stored in more than place or location. Data redundancy occurs when different groups in an organization independently collect the same piece of data and store it independently of each other. Data redundancy wastes storage resources and also leads to data inconsistency,

where the same attribute may have different values.

2 THE DATABASE APPROACH TO DATA

MANAGEMENT

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

separate files for personnel, payroll, and benefits, the corporation could create a single common human resources database.

DATABASE MANAGEMENT SYSTEMS

A database management system (DBMS) is software that permits an organization to centralize data, manage them efficiently, and provide access to the stored data by application programs. The DBMS acts as an interface between application programs and the physical data files. When the application program calls for a data item, such as gross pay, the DBMS finds this item in the database and presents it to the application program. Using traditional data files, the programmer would have to specify the size and format of each data element used in the program and then tell the computer where they were located. The DBMS relieves the programmer or end user from the task of understanding where and how the data are actually stored by separating the logical and physical views of the data. The logical view presents data as they would be perceived by end users or business specialists, whereas the physical view shows how data are actually organized and structured on physical storage media. The database management software makes the physical database available for different logical views required by users. For example, for the human resources database illustrated in Figure 6-3, a benefits specialist might require a view consisting of the employee’s name, social security number, and health insurance coverage. A payroll department member might need data such as the employee’s name, social security number, gross pay, and net pay. The data for

CAPABILITIES OF DATABASE MANAGEMENT SYSTEMS

A DBMS includes capabilities and tools for organizing, managing, and accessing the data in the database. The most important are its data definition language, data dictionary, and data manipulation language. DBMS have a data definition capability to specify the structure of the content of the database. It would be used to create database tables and to define the characteristics of the fields in each table. This information about the database would be documented in a data dictionary. A data dictionary is an automated or manual file that stores definitions of data elements and their characteristics.

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

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

3. USING DATABASES TO IMPROVE BUSINESS

PERFORMANCE AND DECISION MAKING

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

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

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

DATA WAREHOUSES

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

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

What Is a Data Warehouse?

A data warehouse is a database that stores current and historical data of potential interest to decision makers throughout the company. The data originate in many core operational transaction systems, such as systems for sales, customer accounts, and manufacturing, and may include data from Web site transactions. The data warehouse consolidates and standardizes information from different operational databases so that the information can be used across the enterprise for management analysis and decision making.

TOOLS FOR BUSINESS INTELLIGENCE: MULTIDIMENSIONAL DATA ANALYSIS AND DATA

MINING

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

Principal tools for business intelligence include software for database querying and reporting, tools for multidimensional data analysis (online analytical processing), and tools for data mining.

Online Analytical Processing (OLAP)

Suppose your company sells four different products—nuts, bolts, washers, and screws—in the East, West, and Central regions. If you wanted to ask a fairly straightforward question, such as how many washers were sold during the past quarter, you could easily find the answer by querying your sales database.

But what if you wanted to know how many washers sold in each of your sales regions and compare actual results with projected sales?

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

INTERACTIVE SESSION: TECHNOLOGY

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

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

For example, in 2007, JetBlue experience unprecedented levels of customer discontent in the

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

Fortunately, the company had recently contracted with Attensity, a leading vendor of text analytics software, and was able to use the software to analyze all of the e-mail it had received within two days.

According to JetBlue research analyst Bryan Jeppsen, Attensity Analyze for Voice of the Customer (VoC) enabled JetBlue to rapidly extract customer sentiments, preferences, and requests it couldn’t find any other way. This tool uses a proprietary technol- ogy to automatically identify facts, opinions, requests, trends, and trouble spots from the unstruc-

tured text of survey responses, service notes, e-mail messages, Web forums, blog entries, news articles, and other customer communications. The technology is able to accurately and automatically identify the many different “voices” customers use to express their feedback (such as a negative voice, positive voice, or conditional voice), which helps organizations pinpoint key events and relationships, such as intent to buy, intent to leave, or customer “wish”events. It can reveal specific product and service issues, reactions to marketing and public relations efforts, and even buying signals.

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

Hotel chains like Gaylord Hotels and Choice Hotels are using text mining software to glean insights from thousands of customer satisfaction surveys provided by their guests. Gaylord Hotels is using Clarabridge’s text analytics solution delivered via the Internet as a hosted software service to gather and analyze customer feedback from surveys, e-mail, chat messaging, staffed call centers, and online forums associated with guests’ and meeting planners’ experiences at the company’s convention resorts. The Clarabridge software sorts through the hotel chain’s customer surveys and gathers positive and negative comments, organizing them into a variety of categories to reveal less obvious insights.

For example, guests complained about many things more frequently than noisy rooms, but complaints of noisy rooms were most frequently correlated with surveys indicating an unwillingness to return to the hotel for another stay.

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

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

4. MANAGING DATA RESOURCES

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

ESTABLISHING AN INFORMATION POLICY

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

An information policy specifies the organization’s rules for sharing, disseminating, acquiring, standardizing, classifying, and inventorying information. Information policy lays out specific procedures and accountabilities, identifying which users and organizational units can share information, where information can be distributed, and who is responsible for updating and

maintaining the information. For example, a typical information policy would specify that only selected members of the payroll and human resources department would have the right to change and view sensitive employee data, such as an employee’s salary or social security number, and that these departments are responsible for making sure that such employee data are accurate.

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

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

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

database administration.

ENSURING DATA QUALITY

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

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

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

Gartner Inc. reported that more than 25 percent of the critical data in large Fortune 1000 companies’ databases is inaccurate or incomplete, including bad product codes and product descriptions, faulty inventory descriptions, erroneous financial data, incorrect supplier information, and incorrect employee data. (Gartner, 2007).

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

If a database is properly designed and enterprise-wide data standards established, duplicate or inconsistent data elements should be minimal. Most data quality problems, however, such as misspelled names, transposed numbers, or incorrect or missing codes, stem from errors during data input.

The incidence of such errors is rising as companies move their businesses to the Web and allow customers and suppliers to enter data into their Web sites that directly update internal systems.

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

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

Data quality problems are not just business problems. They also pose serious problems for individuals, affecting their financial condition and even their jobs.

The Interactive Session on Organizations describes some of these impacts, as it details the data quality problems found in the companies that collect and report consumer credit data. As you read this case, look for the management, organization, and technology factors behind this problem, and whether existing solutions are adequate.

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

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

promotion decisions.

U.S. credit bureaus collect personal information and financial data from a variety of sources,

including creditors, lenders, utilities, debt collection agencies, and the courts. These data are

aggregated and stored in massive databases maintained by the credit bureaus. The credit

bureaus then sell this information to other companies to use for credit assessment.

The credit bureaus claim they know which credit cards are in each consumer’s wallet, how much is due on the mortgage, and whether the electric bill is paid on time. But if the wrong

information gets into their systems, whether through identity theft or errors transmitted by

creditors, watch out! Untangling the mess can be almost impossible.

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

CREDIT BUREAU ERRORS—BIG PEOPLE PROBLEMS

The sheer volume of information being transmitted from creditors to credit bureaus increases the likelihood of mistakes. Experian, for example, updates 30 million credit reports each day and roughly 2 billion credit reports each month. It matches the identifying personal information in a credit application or credit account with the identifying personal information

in a consumer credit file. Identifying personal information includes items such as name (first

name, last name and middle initial), full current address and ZIP code, full previous address and ZIP code, and social security number. The new credit information goes into the consumer credit file that it best matches.

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

Perhaps the consumer didn’t write clearly on the account application. Name variations on different credit accounts can also result in less than-perfect matches. Take the name Edward Jeffrey Johnson. One account may say Edward Johnson. Another may say Ed Johnson. Another

might say Edward J. Johnson. Suppose the last two digits of Edward’s social security number get transposed—more chance for mismatches.

If the name or social security number on another person’s account partially matches the data in your file, the computer might attach that person’s data to your record. Your record might

likewise be corrupted if workers in companies supplying tax and bankruptcy data from court and government records accidentally transpose a digit or misread a document.

The credit bureaus claim it is impossible for them to monitor the accuracy of the 3.5 billion pieces of credit account information they receive each month. They must continually contend with bogus claims from consumers who falsify lender to include the background banks need to understand a complaint.

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

5. HANDS-ON MIS PROJECTS

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

Management Decision Problems

1. Emerson Process Management, a global supplier of measurement, analytical, and monitoring instruments and services based in Austin, Texas, had a new data warehouse designed for analyzing customer activity to improve service and marketing that was full of inaccurate and redundant data. The data in the warehouse came from numerous transaction processing systems in Europe, Asia, and other locations around the world. The team that designed the warehouse had assumed that sales groups in all these areas would enter customer names and addresses the same way, regardless of their location. In fact, cultural differences combined with complications from absorbing companies that Emerson had acquired led to multiple ways of entering quotes, billing, shipping, and other data. Assess the potential business impact of these data quality problems. What decisions have to be made and steps taken to reach a solution?

2. Your industrial supply company wants to create a data warehouse where management can obtain a single corporate-wide view of critical sales information to identify best-selling products in specific geographic areas, key customers, and sales trends. Your sales and product information are stored in several different systems: a divisional sales system running on a Unix server and a corporate sales system running on an IBM mainframe. You would like to create a single standard format that consolidates these data from both systems. The following format has been proposed.

• What business problems are created by not having these data in a single standard format?

• How easy would it be to create a database with a single standard format that could store the data from both systems? Identify the problems that would have to be addressed.

• Should the problems be solved by database specialists or general business managers? Explain.

• Who should have the authority to finalize a single company-wide format for this information in the data warehouse?

Achieving Operational Excellence: Building a Relational Database for Inventory Management

Software skills: Database design, querying, and reporting

Business skills: Inventory management

Businesses today depend on databases to provide reliable information about items in inventory, items that need restocking, and inventory costs. In this exercise, you’ll use database software to design a database for managing inventory for a small business.

Sylvester’s Bike Shop, located in San Francisco, California, sells road, mountain, hybrid, leisure, and children’s bicycles. Currently, Sylvester’s purchases bikes from three suppliers but plans to add new suppliers in the near future. This rapidly growing business needs a database system to manage this information.

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

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

• Prepare a report that identifies the five most expensive bicycles. The report should list the bicycles in descending order from most expensive to least expensive, the quantity on hand for each, and the markup percentage for each.

• Prepare a report that lists each supplier, its products, the quantities on hand, and associated reorder levels. The report should be sorted alphabetically by supplier. Within each supplier category, the products should be sorted alphabetically.

• Prepare a report listing only the bicycles that are low in stock and need to be reordered. The report should provide supplier information for the items identified.

• Write a brief description of how the database could be enhanced to further improve management of the business. What tables or fields should be added?

What additional reports would be useful?

Improving Decision Making: Searching Online

Databases for Overseas Business Resources

Software skills: Online databases

Business skills: Researching services for overseas operations

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

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

You will soon be traveling to the area to make plans to actually set up an office, and you want to meet with organizations that can help you with your operation. You will need to engage people or organizations that offer many services necessary for you to open your office, including lawyers, accountants, import-export experts, telecommunications equipment and support, and even trainers who can help you to prepare your future employees to work for you.

Start by searching for U.S. Department of Commerce advice on doing business in Australia. Then try the following online databases to locate companies that you would like to meet with during your coming trip: Australian Business Register (abr.business.gov.au/), Australia Trade Now (australiatradenow.com/), and the Nationwide Business Directory of Australia (www.nationwide.com.au).

If necessary, you could also try search engines such as Yahoo and Google. Then perform the following activities:

• List the companies you would contact to interview on your trip to determine whether they can help you with these and any other functions you think vital to establishing your office.

• Rate the databases you used for accuracy of name, completeness, ease of use, and general helpfulness.

• What does this exercise tell you about the design of databases?

LEARNING TRACK MODULES

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

1. Database Design, Normalization, and Entity-Relationship Diagramming

2. Introduction to SQL

3. Hierarchical and Network Data Models

Review Summary

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

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

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

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

The relational database is the primary method for organizing and maintaining data today in

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

3. What are some important database design principles?

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

4. What are the principal tools and technologies for accessing information from databases to

improve business performance and decision making?

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

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

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

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

CASE STUDY

In the aftermath of the 9-11 attacks, the FBI’s Terrorist Screening Center, or TSC, was established to consolidate information about suspected terrorists from multiple government

agencies into a single list to enhance inter-agency communication. A database of suspected terrorists known as the terrorist watch list was created.

Multiple U.S. government agencies had been maintaining separate lists and these agencies lacked a consistent process to share relevant information.

Records in the TSC database contain sensitive but unclassified information on terrorist identities, such as name and date of birth, that can be shared with other screening agencies. Classified information about the people in the watch list is maintained in other law enforcement and intelligence agency databases. Records for the watchlist database are pro-

vided by two sources: The National Counterterrorism Center (NCTC) managed by the Office of the Director of National Intelligence provides identifying information on individuals with ties to international terrorism. The FBI provides identifying information on individuals with ties to purely domestic terrorism.

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

from the database The TSC watch list database is updated daily with new nominations, modifications to existing records, and deletions. Since its creation, the list has ballooned to 400,000 people, recorded as 1.1 million names and aliases, and is continuing to grow at a rate

of 200,000 records each year. Information on the list is distributed to a wide range of government agency systems for use in efforts to deter or detect the movements of known or suspected terrorists.

Recipient agencies include the FBI, CIA, National Security Agency (NSA), Transportation Security Administration (TSA), Department of Homeland Security, State Department, Customs and Border Protection, Secret Service, U.S. Marshals Service, and the White House. Airlines use data supplied by the TSA system in their NoFly and Selectee lists for prescreening passengers, while the U.S. Customs and Border Protection system uses the watchlist data to

help screen travelers entering the United States. The State Department system screens applicants for visas to enter the United States and U.S. residents applying for passports, while state and local law enforcement agencies use the FBI system to help with arrests, detentions, and other criminal justice activities. Each of these agencies receives the subset of data in the

watch list that pertains to its specific mission.

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

The process of consolidating information from disparate agencies has been a slow and painstaking one, requiring the integration of at least 12 different databases. Two years after the process of integration took place, 10 of the 12 databases had been processed. The remaining two databases (the U.S. Immigration and Customs Enforcement’s Automatic Biometric Identification System and the FBI’s Integrated Automated Fingerprint Identification

System) are both fingerprint databases. There is still more work to be done to optimize the list’s usefulness.

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

Information about the process for inclusion on the list must necessarily be carefully protected if the list is to be effective against terrorists. The specific criteria for inclusion are not public knowledge. We do know, however, that government agencies popu-

late their watch lists by performing wide sweeps of information gathered on travelers, using many misspellings and alternate variations of the names of suspected terrorists. This often leads to the inclusion of people who do not belong on watch lists, known as “false positives.” It also results in some people being listed multiple times under different spellings of

their names.

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

or similar to a different suspected terrorist.

These incidents call attention to the quality and accuracy of the data in the TSC consolidated

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

weaknesses.

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

The number of requests for removal from the watch list continues to mount, with over 24,000

requests recorded (about 2,000 each month) and only 54 percent of them resolved. The average time to process a request in 2008 was 40 days, which was not (and still is not) fast enough to keep pace with the number of requests for removal coming in. As a result, law-abiding travelers that inexplicably find themselves on the watch list are left with no easy way to remove themselves from it.In February 2007, the Department of Homeland Security instituted its Traveler Redress Inquiry Program (TRIP) to help people that have been erroneously added to terrorist watch lists remove themselves and avoid extra screening and questioning. John Anderson’s mother claimed that despite her best efforts, she was unable to remove her son from the watch lists. Senator Kennedy reportedly was only able to remove himself from the list by personally bringing up the matter to Tom Ridge, then the Director of the Department of Homeland Security.

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

the likelihood of false matches.

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

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

A report released in May 2009 by Department of Justice Inspector General Glenn A. Fine found that the FBI had incorrectly kept nearly 24,000 people on its own watch list that supplies data to the terrorist watch list on the basis of outdated or irrelevant information. Examining nearly 69,000 referrals to the FBI list, the report found that 35 percent of those people remained on the list despite inadequate justification.

Even more worrisome, the list did not contain the names of people who should have been listed because of their terrorist ties.

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

result is unclear.).

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

no-fly list that had added Shahzad’s name.

Sources: Scott Shane, “Lapses Allowed Suspect to Board Plane,” The New York Times, May 4, 2010; Mike McIntire, “Ensnared by Error on Growing U.S. Watch List,” The New York Times, April 6, 2010; Eric Lipton, Eric Schmitt, and Mark Mazzetti, “Review of Jet Bomb Plot Shows More Missed Clues,” The New York Times, January 18, 2010; Lizette Alvarez, “Meet Mikey, 8: U.S. Has Him on Watch List,” The New York Times, January 14, 2010; Eric Lichtblau, “Justice Dept. Finds Flaws in F.B.I. Terror List,” The New York Times, May 7, 2009; Bob Egelko, “Watch-list Name Confusion Causes Hardship,” San Francisco Chronicle, March 20, 2008;

“Reports Cite Lack of Uniform Policy for Terrorist Watch List,” The Washington Post, March 18, 2008; Siobhan Gorman, “NSA’s Domestic Spying Grows as Agency Sweeps Up Data,” The Wall Street Journal, March 10, 2008; Ellen Nakashima, and Scott McCartney, “When Your Name is Mud at the Airport,” The Wall Street Journal, January 29, 2008.