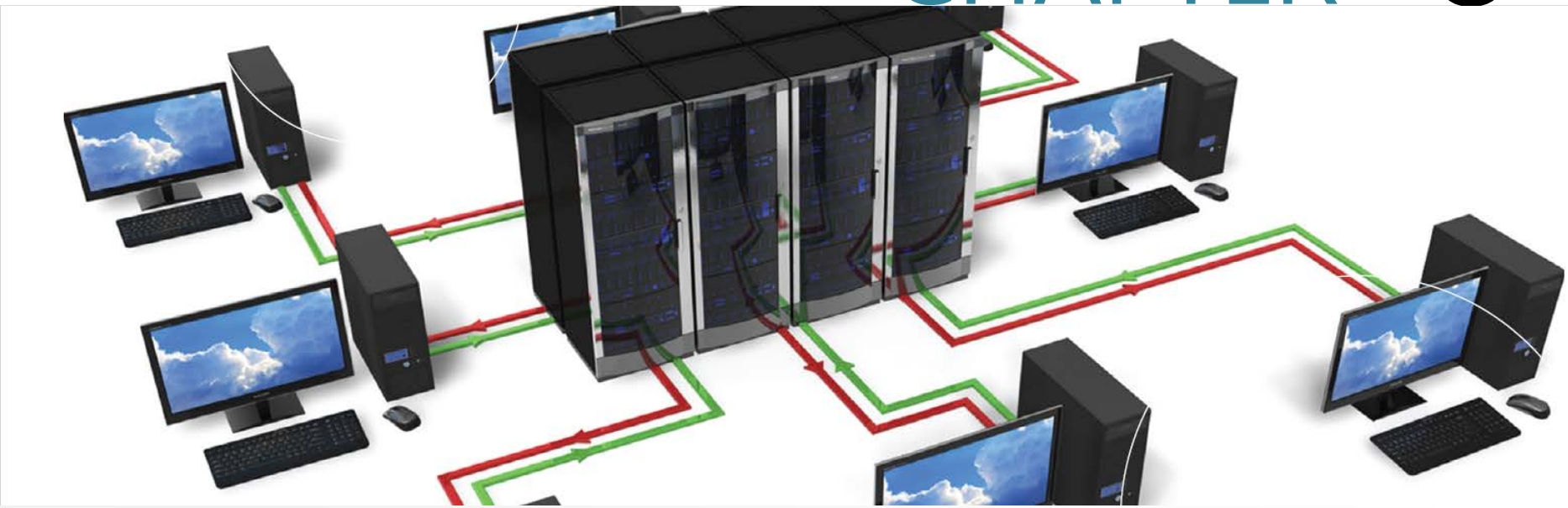
CHAPTER 3



Data and Knowledge Management

CHAPTER OUTLINE

1. Managing Data
2. The Database Approach
3. Big Data
4. Data Warehouses and Data Marts
5. Knowledge Management

LEARNING OBJECTIVES

1. Discuss ways that common challenges in managing data can be addressed using data governance.
2. Discuss the advantages and disadvantages of relational databases.
3. Define Big Data, and discuss its basic characteristics.
4. Explain the elements necessary to successfully implement and maintain data warehouses.
5. Describe the benefits and challenges of implementing knowledge management systems in organizations.

3.1 Managing Data

- Difficulties of Managing Data
- Data Governance

The Difficulties of Managing Data

- The amount of data increases exponentially over time
- Data are scattered throughout organizations
- Data are generated from multiple sources
  - internal sources (for example, corporate databases and company documents); personal sources (for example, personal thoughts, opinions, and experiences); and external sources (for example, commercial databases, government reports, and corporate Web sites)
  - **Clickstream data** are those data that visitors and customers produce when they visit a Web site and click on hyperlinks. Clickstream data provide a trail of the users' activities in the Web site, including user behavior and browsing patterns.
- New sources of data(e.g., blogs, podcasts, videocasts, and RFID tags and other wireless sensors)

The Difficulties of Managing Data (continued)

- Data Degradation
  - For example, customers move to new addresses or change their names, companies go out of business or are bought, new products are developed, employees are hired or fired, and companies expand into new countries.
- Data Rot
  - Data rot refers primarily to problems with the media on which the data are stored.
  - it is almost impossible today to find 8-track players.
- Data security, quality, and integrity are critical
- Legal requirements change frequently and differ among countries & industries
- Inconsistent data
  - organizations have developed information systems for specific business processes, such as transaction processing, supply chain management, and customer relationship management. It results in repetition and conflicts across the organization

Data Governance

- **Data Governance:**
  - an approach to managing information across an entire organization involving a formal set of unambiguous rules for creating, collecting, handling, and protecting its information.
- **Master Data Management:**
  - a process that spans all organizational business processes and applications.
  - It provides companies with the ability to store, maintain, exchange, and synchronize a consistent, accurate, and timely “single version of the truth” for the company's master data.
- **Master Data:**
  - a set of core data (e.g., customer, product, employee, vendor, geographic location, etc.) that span the enterprise information systems.
  - *Transaction data*, which are generated and captured by operational systems, describe the business's activities, or transactions.
  - In contrast, master data are applied to multiple transactions and are used to categorize, aggregate, and evaluate the transaction data.

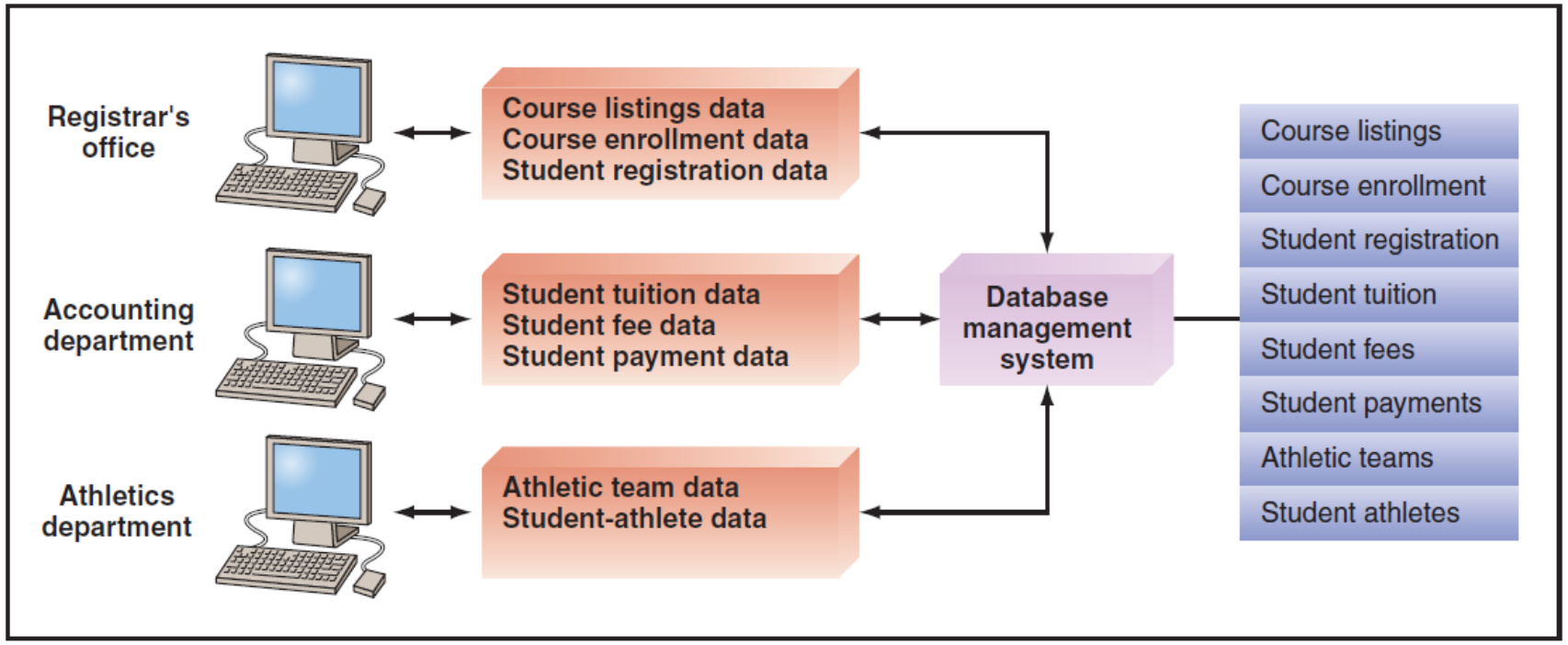
Data Governance

- **Master Data vs Transaction Data:**
  - Let's look at an example of a transaction: Mary Jones purchased one Samsung 42-inch plasma television, part number 1234, from Bill Roberts at Best Buy, for $2,000, on April 20, 2015.
  - In this example, the master data are “product sold,” “vendor,” “salesperson,” “store,” “part number,” “purchase price,” and “date.”
  - When specific values are applied to the master data, then a transaction is represented.
  - Therefore, transaction data would be, respectively, “42-inch plasma television,” “Samsung,” “Best Buy,” “Bill Roberts,” “1234,” “$2,000,” and “April 20, 2015.”

3.2 The Database Approach

- Data File
  - a collection of logically related records.
- File system vs Database system
- Database Systems Minimize & Maximize Three Things
  - *Minimizing*
  - *Data redundancy:* The same data are stored in multiple locations.
  - *Data isolation:* Applications cannot access data associated with other applications.
  - *Data inconsistency:* Various copies of the data do not agree.
  - *Maximizing*
  - *Data security:* Because data are “put in one place” in databases, there is a risk of losing a lot of data at one time. Therefore, databases must have extremely high security measures in place to minimize mistakes and deter attacks.
  - *Data integrity:* Data meet certain constraints; for example, there are no alphabetic characters in a Social Security number field.
  - *Data independence:* Applications and data are independent of one another; that is, applications and data are not linked to each other, so all applications are able to access the same data.

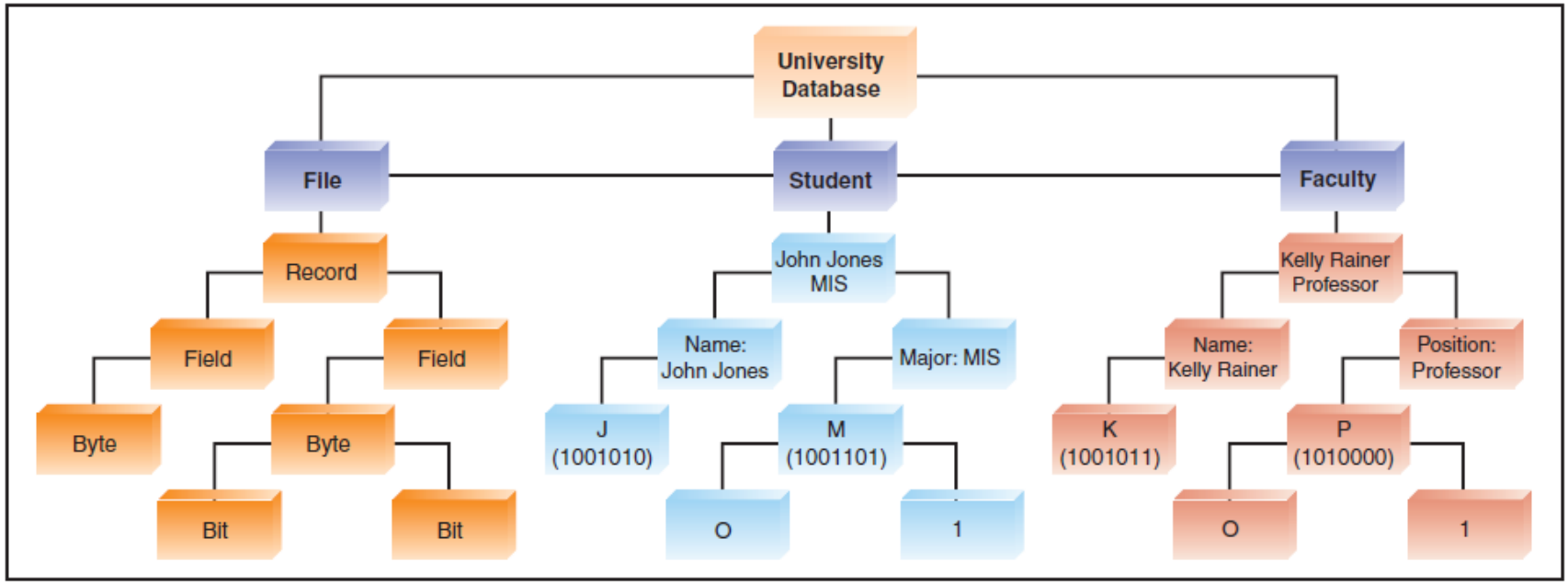
Figure 3.1: Database Management System



3.2 The Database Approach

- The Data Hierarchy
  - Data are organized in a hierarchy that begins with bits and proceeds all the way to databases.
  - A **bit** (*b*inary dig*it*) represents the smallest unit of data a computer can process. The term *binary* means that a bit can consist only of a 0 or a 1.
  - group of eight bits, called a **byte** represents a single character. A byte can be a letter, a number, or a symbol.
  - A logical grouping of characters into a word, a small group of words, or an identification number is called a **field**.
    - For example, a student's name in a university's computer files would appear in the “name” field, and her or his Social Security number would appear in the “Social Security number” field. Fields can also contain data other than text and numbers. They can contain an image, or any other type of multimedia. Examples are a motor vehicle department's licensing database that contains a driver's photograph and a field that contains a voice sample to authorize access to a secure facility.
  - A logical grouping of related fields, such as the student's name, the courses taken, the date, and the grade, comprises a **record**
    - In the Apple iTunes Store, a song is a field in a record, with other fields containing the song's title, its price, and the album on which it appears.
  - A logical grouping of related records is called a **data file** or a **table**.
  - a logical grouping of related files constitutes a **database.**

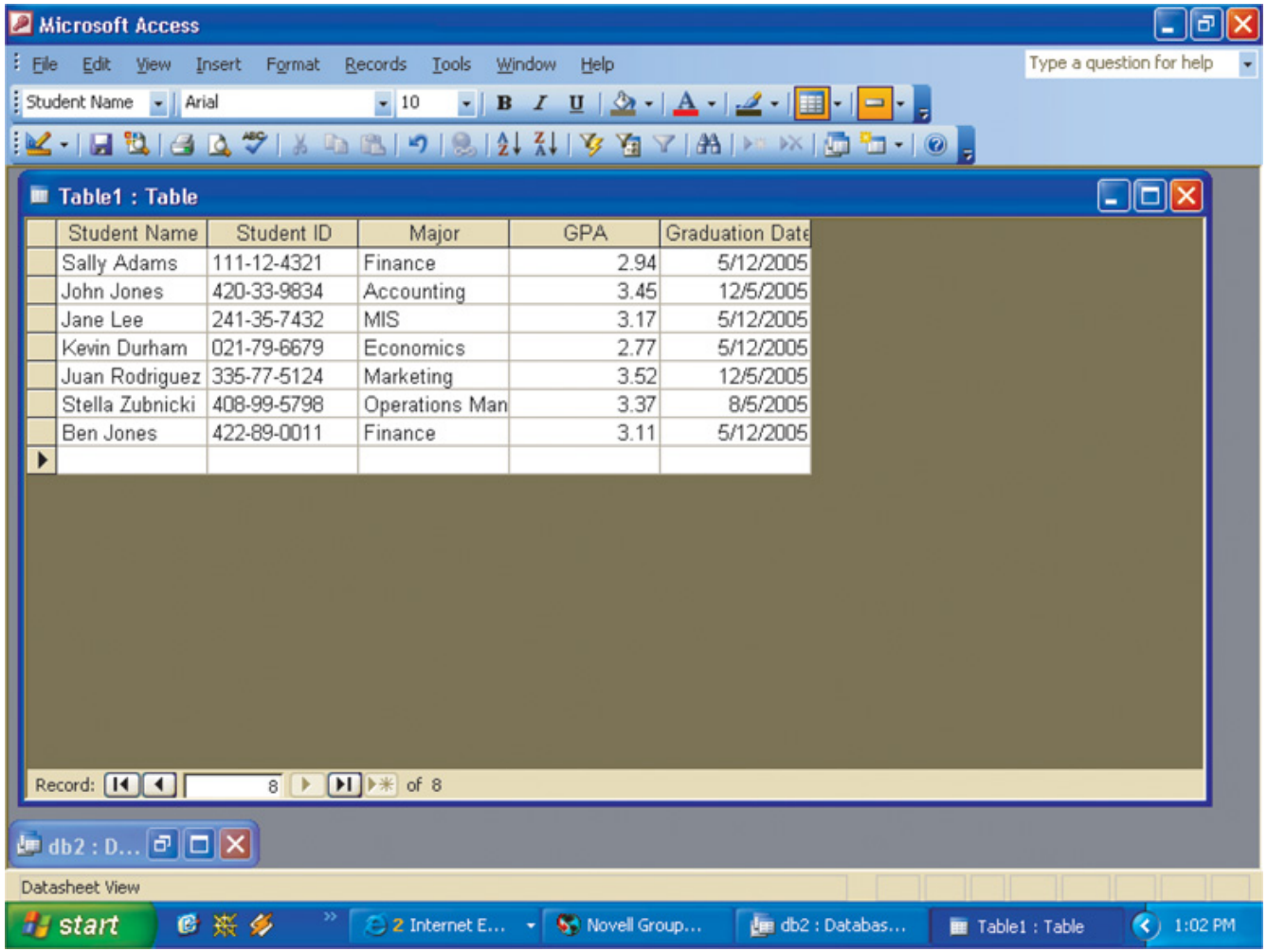
Figure 3.2: Hierarchy of Data for a Computer-based File



The Relational Database Model

- Database Management System (DBMS)
  - a set of programs that provide users with tools to create and manage a database.
- Relational Database Model
  - is based on the concept of two-dimensional tables and is usually designed with a number of related tables with each of these tables contains records (listed in rows) and attributes (listed in columns).
- Data Model
  - a diagram that represents entities in the database and their relationships.
- Entity
  - a person, place, thing, or event (e.g., customer, an employee, or a product).
- Instance
  - an entity refers to each row in a relational table,
- Attribute
  - each characteristic or quality of a particular entity.
  - For example, if our entities were a customer, an employee, and a product, entity attributes would include customer name, employee number, and product color.

Figure 3.3: Student Database Example



The Relational Database Model (continued)

- Primary Key
  - a field in a database that uniquely identify each record so that it can be retrieved, updated, and sorted.
  - a student record in a university would use a unique student number as its primary key.
- Secondary Key
  - a field that has some identifying information, but typically does not identify the record with complete accuracy and therefore cannot serve at the Primary Key.
  - For example, the student's major might be a secondary key if a user wanted to identify all of the students majoring in a particular field of study.
- Foreign Key
  - a field (or group of fields) in one table that uniquely identifies a row of another table. It is used to establish and enforce a link between two tables.
- Unstructured Data
  - data that does not reside in a traditional relational database. Examples of unstructured data include e-mail messages, word processing documents, videos, images, audio files, PowerPoint presentations, Facebook posts, tweets, snaps, ratings and recommendations, and Web pages.
  - Industry analysts estimate that 80 to 90 percent of the data in an organization is unstructured

3.3 Big Data

- Defining Big Data
- Characteristics of Big Data
- Issues with Big Data
- Managing Big Data
- Putting Big Data to Use
- Big Data Used in the Functional Areas of the Organization

Defining Big Data

- **Gartner (www.gartner.com)**
  - defines Big Data as diverse, high volume, high-velocity information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.
- **Big Data Institute (TBDI; www.the-bigdatainstitute.com):**
  - defines Big Data as vast datasets that:
    - Exhibit variety;
    - Include structured, unstructured, and semi-structured data;
    - Are generated at high velocity with an uncertain pattern;
    - Do not fit neatly into traditional, structured, relational databases; and
    - Can be captured, processed, transformed, and analyzed in a reasonable amount of time only by sophisticated information systems.

Big Data Generally Consist of the Following:

- Traditional Enterprise Data
  - examples are customer information from customer relationship management systems, transactional enterprise resource planning data, Web store transactions, operations data, and general ledger data.
- Machine-Generated/Sensor Data
  - examples are smart meters; manufacturing sensors; sensors integrated into smartphones, automobiles, airplane engines, and industrial machines; equipment logs; and trading systems data.
- Social Data
  - examples are customer feedback comments; microblogging sites such as Twitter; and social media sites such as Facebook, YouTube, and LinkedIn.
- Images Captured by Billions of Devices Located Throughout the World
  - from digital cameras and camera phones to medical scanners and security cameras.

Examples of Big Data

- In 2015 Google was processing more than 27 petabytes(=1,000 terabytes) of data every day.
- Facebook members upload more than 10 million new photos every hour. In addition, they click a “like” button or leave a comment nearly 3 billion times every day.
- The 800 million monthly users of Google's YouTube service upload more than an hour of video every second.
- The number of messages on Twitter is growing at 200 percent every year. By mid-2015, the volume exceeded 550 million tweets per day.

Characteristics of Big Data

- Volume
  - incredible volume of data.
- Velocity
  - The rate at which data flow into an organization is rapidly increasing and it is critical because it increases the speed of the feedback loop between a company and its customers.
- Variety
  - Big Data formats change rapidly and can include satellite imagery, broadcast audio streams, digital music files, Web page content, scans of government documents, and comments posted on social networks.

Issues with Big Data

- Untrusted data sources
- Big Data is dirty
  - Dirty data refers to inaccurate, incomplete, incorrect, duplicate, or erroneous data.
  - Examples of such problems are misspelling of words and duplicate data such as retweets or company press releases that appear numerous times in social media.
- Big Data changes, especially in data streams
  - Organizations must be aware that data quality in an analysis can change, or the data itself can change, because the conditions under which the data are captured can change.

Managing Big Data

- Big Data makes it possible to do many things that were previously impossible:
  - business trends more rapidly and accurately
  - tracking the spread of disease
  - tracking crime
  - detecting fraud

Managing Big Data (continued)

- First Step:
  - Integrate information silos into a database environment and develop data warehouses for decision making.
  - **Information Silo:** is an information system that does not communicate with other, related information systems in an organization.
- Second Step:
  - turn organization's attention to the business of information management

Managing Big Data (continued)

- Many organizations are turning to NoSQL databases to process Big Data
  - SQL: structured query language
  - RDB: Oracle, MySQL, MS Access
  - **NoSQL Database:** can manipulate structured as well as unstructured data and inconsistent or missing data providing an alternative for firms that have more and different kinds of data (Big Data) in addition to the traditional, structured data that fit neatly into the rows and columns of relational databases.
  - NoSQL DB: Hadoop, Mongo DB, Apache Cassandra, Couch DB
  - Hadoop (http://hadoop.apache.org) is not a type of database, but rather a collection of programs that allow storage, retrieval, and analysis of very large datasets using massively parallel processing. Massively parallel processing is the coordinated processing of an application by multiple processors that work on different parts of the application, with each processing using its own operating system and memory. Hadoop enables the processing of NoSQL databases,

Putting Big Data to Use

- Making Big Data Available
  - Making Big Data available for relevant stakeholders can help organizations gain value.
  - The Open Data 500 study at the GovLab at New York University found some 500 examples of U.S.-based companies whose business models depend on analyzing open government data.

Putting Big Data to Use

- Creating New Business Models
  - For example, a commercial transportation company operated a large fleet of large, long-haul trucks. The company recently placed sensors on all its trucks. These sensors wirelessly communicate large amounts of information to the company, a process called *telematics*. The sensors collect data on vehicle usage (including acceleration, braking, cornering, etc.), driver performance, and vehicle maintenance.
  - By analyzing this Big Data, the transportation company was able to improve the condition of its trucks through near-real-time analysis that proactively suggested preventive maintenance. In addition, the company was able to improve the driving skills of its operators by analyzing their driving styles.
- Organizations Can Analyze Far More Data
  - Organizations can even process all the data relating to a particular phenomenon, meaning that they do not have to rely as much on sampling.
  - For example, consider political polling using landline phones. This sample tends to exclude people who use only cell phones. This bias can seriously skew the results because cell phone users are typically younger and more liberal than people who rely primarily on landline phones.

Putting Big Data to Use

- Enabling Organizations to Conduct Experiments
  - Big Data allows organizations to improve performance by conducting controlled experiments.
  - For example, Amazon (and many other companies such as Google and LinkedIn) constantly experiments by offering slight different “looks” on its Web site. These experiments are called A/B experiments, because each experiment has only two possible outcomes.
- Micro-Segmentation of Customers
  - Segmentation of a company's customers means dividing them up into groups that share one or more characteristics.
  - For example, Paytronix Systems (www.paytronix.com) provides loyalty and rewards program software for thousands of different restaurants. Paytronix gathers restaurant guest data from a variety of sources beyond loyalty and gift programs, including social media. Paytronix analyzes this Big Data to help its restaurant clients micro-segment their guests. Restaurant managers are now able to more precisely customize their loyalty and gift programs. In doing so, the managers are noting improved performance in their restaurants in terms of profitability and customer satisfaction.

Big Data Used in the Functional Areas of the Organization

- Human Resources
  - Caesars Entertainment (www.caesars.com)
    - It analyzes health-insurance claim data for its 65,000 employees and their covered family members.
    - Managers can track thousands of variables that indicate how employees use medical services, such as the number of emergency room visits and whether employees choose a generic or brand name drug.
    - Data revealed that too many employees with medical emergencies were being treated at hospital emergency rooms rather than at less-expensive urgent-care facilities. The company launched a campaign to remind employees of the high cost of emergency room visits, and they provided a list of alternative facilities. Subsequently, 10,000 emergencies shifted to less-expensive alternatives, for a total savings of $4.5 million.
  - Catalyst IT Services (www.catalystdevworks.com)
    - a technology outsourcing company that hires teams for programming jobs.
    - It requires candidates to fill out an online assessment.
    - How the candidate responds—laboring over an answer, answering quickly and then returning later, or skipping the problem entirely—provides insight into how that candidate might deal with challenges that he or she will encounter on the job.
    - That is, someone who labors over a difficult question might be effective in an assignment that requires a methodical approach to problem solving, whereas an applicant who takes a more aggressive approach might perform better in a different job setting.
    - employee turnover at Catalyst averages about 15 percent per year, compared with more than 30 percent for its U.S. competitors

Big Data Used in the Functional Areas of the Organization

- Product Development
  - Ford Motor Company (www.ford.com)
    - It puts customer preference information to work in designing new products.
    - It was considering a “three blink” turn indicator that had been available on its European cars for years.
    - Ford decided that conducting a full-scale market research test on this blinker would be too costly and time consuming. Instead, it examined auto-enthusiast Web sites and owner forums to discover what drivers were saying about turn indicators.
    - Using text-mining algorithms, researchers culled more than 10,000 mentions and then summarized the most relevant comments.
    - Ford introduced the three-blink indicator on the new Ford Fiesta in 2010
- Operations
  - United Parcel Service (UPS)
    - It uses sensors in its delivery vehicles that can capture the truck's speed and location, the number of times it is placed in reverse, and whether the driver's seat belt is buckled.
    - These data are uploaded at the end of each day to a UPS data center, where they are analyzed overnight.
    - By combining GPS information and data from sensors installed on more than 46,000 vehicles, UPS reduced fuel consumption by 8.4 million gallons, and it cut 85 million miles off its routes.

Big Data Used in the Functional Areas of the Organization

- Marketing
  - The United Kingdom's InterContinental Hotels Group (IHG; www.ihg.com)
    - It has gathered details about the members of its Priority Club rewards program, such as income levels and whether members prefer family-style or business-traveler accommodations.
    - It then consolidated all this information with information obtained from social media into a single data warehouse.
    - Using its data warehouse and analytics software, the hotelier launched a new marketing campaign. Where previous marketing campaigns generated, on average, between 7 and 15 customized marketing messages, the new campaign generated more than 1,500.
    - One group, for instance, tends to stay on weekends, redeem reward points for gift cards, and register through IHG marketing partners. Utilizing this information, IHG sent these customers a marketing message that alerted them to local weekend events.
    - One group, for instance, tends to stay on weekends, redeem reward points for gift cards, and register through IHG marketing partners. Utilizing this information, IHG sent these customers a marketing message that alerted them to local weekend events.
- Government Operations
  - With 55 percent of the population of the Netherlands living under the threat of flooding, water management is critically important to the Dutch government.
  - The government operates a sophisticated water management system.
  - In its water management efforts, the government makes use of a vast number of sensors embedded in every physical structure used for water control. The sensors generate at least 2 petabytes of data annually.
  - The result is that Dutch authorities have reduced the costs of managing water by 15 percent.

3.4 Data Warehouses and Data Marts

- Describing Data Warehouses and Data Marts
- A Generic Data Warehouse Environment

Describing Data Warehouses and Data Marts

- **Data Warehouse:**
  - a repository of historical data that are organized by subject to support decision makers in the organization.
- **Data Mart:**
  - a low-cost, scaled-down version of a data warehouse that is designed for the end-user needs in a strategic business unit (SBU) or an individual department.

Describing Data Warehouses and Data Marts

- Basic characteristics of data warehouses and data marts
  - Organized by business dimension or subject
    - Data are organized by subject. For example, by customer, vendor, product, price level, and region. This arrangement differs from transactional systems, where data are organized by business process, such as order entry, inventory control, and accounts receivable.
  - Use online analytical processing (OLAP)
  - Integrated
    - Data are collected from multiple systems and then integrated around subjects.
  - Time variant
    - Data warehouses and data marts maintain historical data (i.e., data that include time as a variable).
  - Nonvolatile
    - Data warehouses and data marts are nonvolatile—that is, users cannot change or update the data.
  - Multidimensional
    - Typically the data warehouse or mart uses a multidimensional data structure. Recall that relational databases store data in two-dimensional tables.
    - A common representation for this multidimensional structure is the *data cube*.

A Generic Data Warehouse Environment

- Source Systems
  - Systems that provide a source of organizational data.
  - operational/transactional systems, enterprise resource planning (ERP) systems, Web site data, third-party data (e.g., customer demographic data), operational databases
- Data Integration technology and processes
  - Reflects the growing number of ways that source system data can be handled. Typically organizations need to Extract, Transform, and Load (ETL) data from source system into a data warehouse or data mart.
- Architectures Storing the Data
  - A variety of architectures can be used to store decision-support data and the most common architecture is one central enterprise data warehouse, without data marts.
  - Another architecture is *independent data marts*. This architecture stores data for a single application or a few applications, such as marketing and finance.
  - another data warehouse architecture is the hub and spoke. This architecture contains a central data warehouse that stores the data plus multiple dependent data marts that source their data from the central repository.

A Generic Data Warehouse Environment

- Metadata
  - data maintained about the data within the data warehouse.
  - e.g., for IT personnel, database, table, and column names; refresh schedules; and data-usage measures.
  - e.g., for user, data definitions, report/query tools, report distribution information, and contact information for the help desk
- Data Quality
  - quality of the data in the warehouse must meet users' needs. If it does not, users will not trust the data and ultimately will not use it.
  - Some of the data can be improved with data-cleansing software, but the better, long-term solution is to improve the quality at the source system level.
- Governance
  - To ensure that BI is meeting their needs, organizations must implement governance to plan and control their BI activities. Governance requires that people, committees, and processes be in place.
- Users
  - There are many potential BI users, including IT developers; frontline workers; analysts; information workers; managers and executives; and suppliers, customers, and regulators.
  - Some of these users are *information producers* whose primary role is to create information for other users. Other users—including managers and executives—are *information consumers*,

Figure 3.4: Data Warehouse Framework

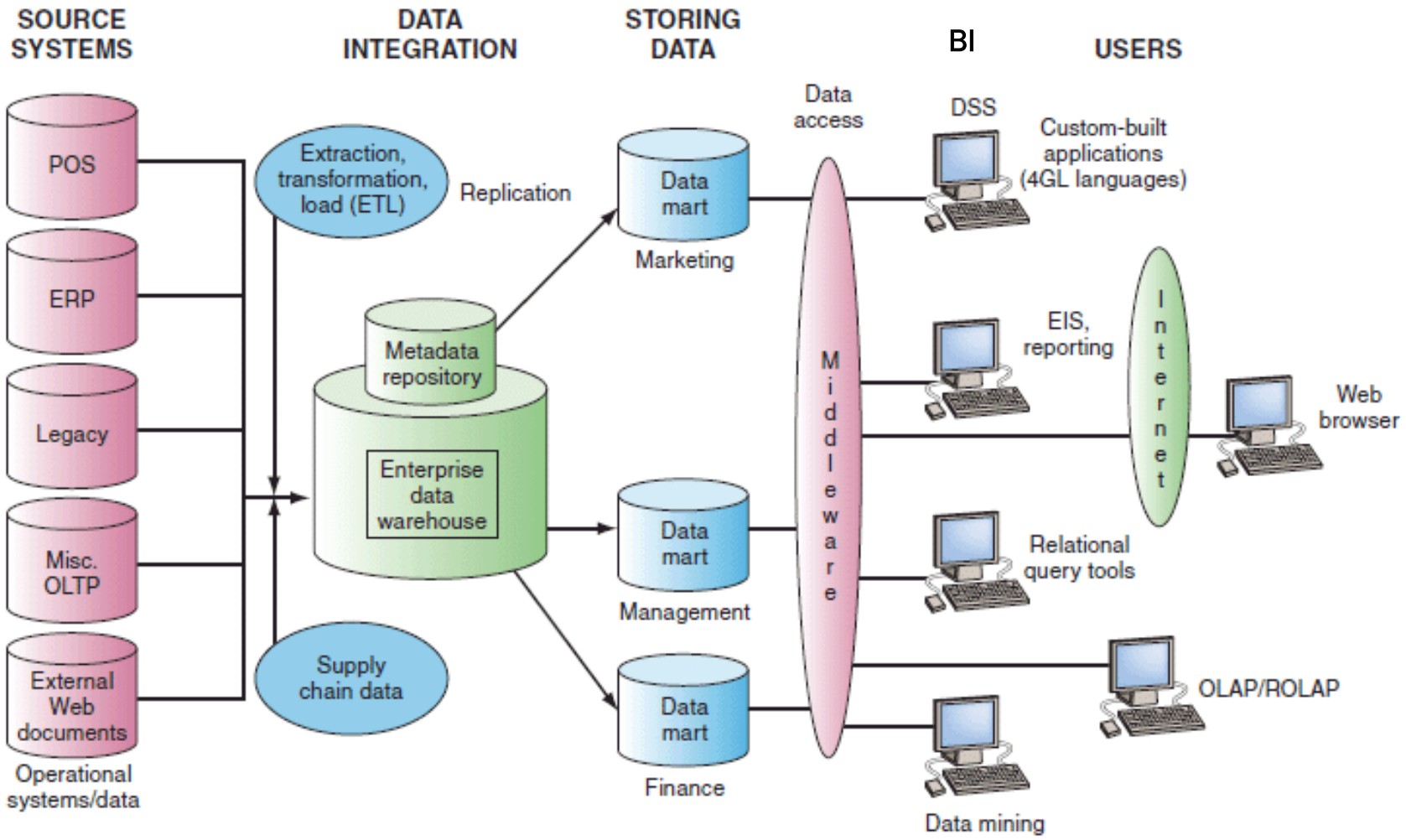


Figure 3.5: Relational Databases

To differentiate between relational databases and multidimensional data warehouses and marts, imagine your company manufactures four products—nuts, screws, bolts, and washers—and has sold them in three territories—East, West, and Central—for the previous three years—2013, 2014, and 2015.

(a) 2012

Product	Region	Sales
Nuts	East	50
Nuts	West	60
Nuts	Central	100
Screws	East	40
Screws	West	70
Screws	Central	80
Bolts	East	90
Bolts	West	120
Bolts	Central	140
Washers	East	20
Washers	West	10
Washers	Central	30

(b) 2013

Product	Region	Sales
Nuts	East	60
Nuts	West	70
Nuts	Central	110
Screws	East	50
Screws	West	80
Screws	Central	90
Bolts	East	100
Bolts	West	130
Bolts	Central	150
Washers	East	30
Washers	West	20
Washers	Central	40

(c) 2014

Product	Region	Sales
Nuts	East	70
Nuts	West	80
Nuts	Central	120
Screws	East	60
Screws	West	90
Screws	Central	100
Bolts	East	110
Bolts	West	140
Bolts	Central	160
Washers	East	40
Washers	West	30
Washers	Central	50

Figure 3.6: Data Cube

In a multidimensional database, in contrast, these data would be represented by a three-dimensional matrix (or data cube)

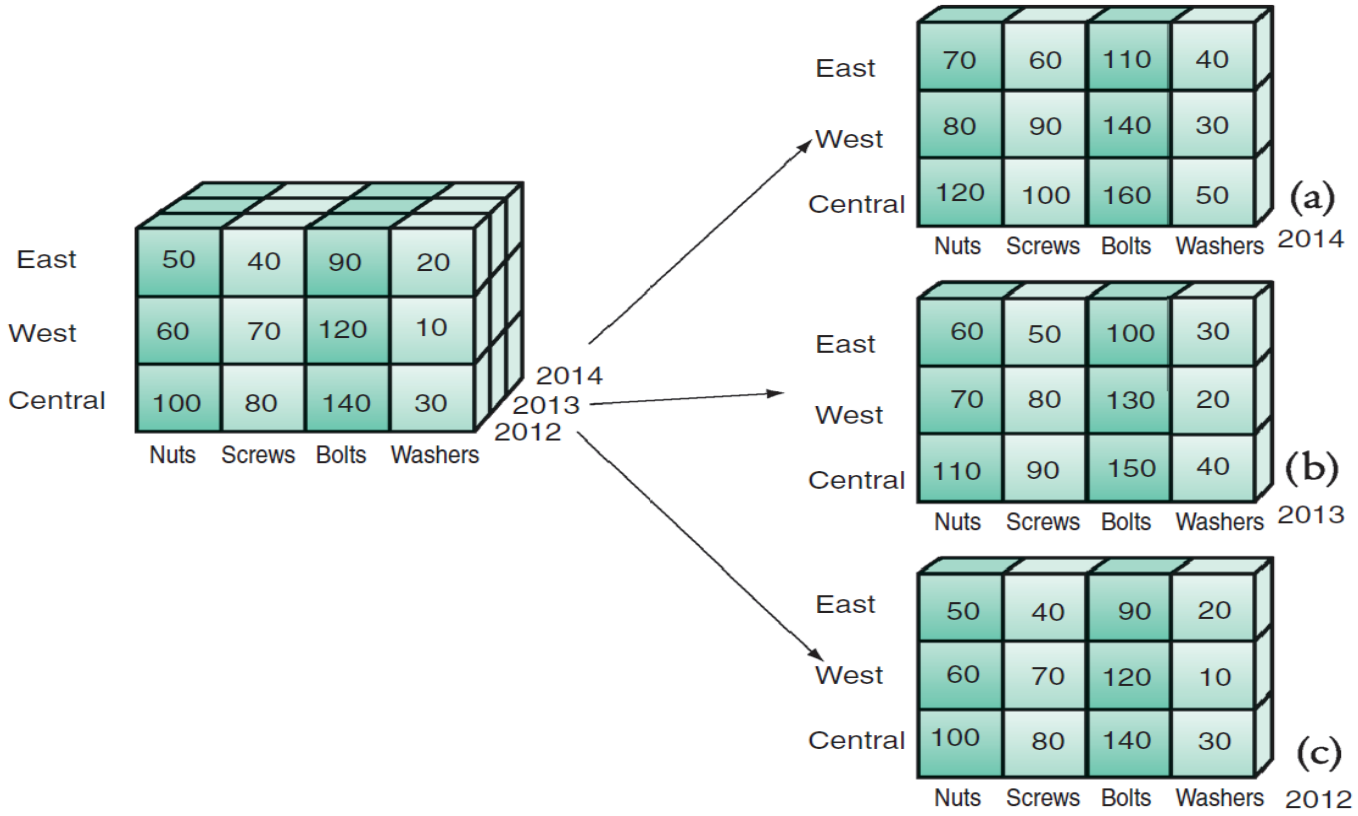
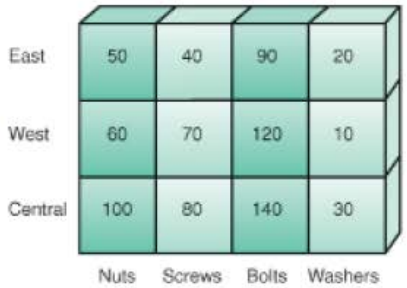


Figure 3.7: Equivalence Between Relational and Multidimensional Databases

Product	Region	Sales
Nuts	East	50
Nuts	West	60
Nuts	Central	100
Screws	East	40
Screws	West	70
Screws	Central	80
Bolts	East	90
Bolts	West	120
Bolts	Central	140
Washers	East	20
Washers	West	10
Washers	Central	30

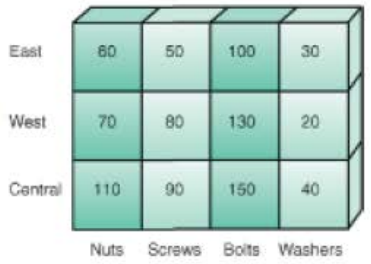
2012



(a)
2012

Product	Region	Sales
Nuts	East	60
Nuts	West	70
Nuts	Central	110
Screws	East	50
Screws	West	80
Screws	Central	90
Bolts	East	100
Bolts	West	130
Bolts	Central	150
Washers	East	30
Washers	West	20
Washers	Central	40

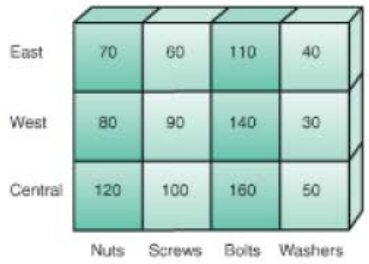
2013



(b)
2013

Product	Region	Sales
Nuts	East	70
Nuts	West	80
Nuts	Central	120
Screws	East	60
Screws	West	90
Screws	Central	100
Bolts	East	110
Bolts	West	140
Bolts	Central	160
Washers	East	40
Washers	West	30
Washers	Central	50

2014



(c)
2014

The benefits and limitations of data warehousing

- Benefits
  - End users can access needed data quickly and easily via Web browsers because these data are located in one place.
  - End users can conduct extensive analysis with data in ways that were not previously possible.
  - End users can obtain a consolidated view of organizational data.
  - These benefits can improve business knowledge, provide competitive advantage, enhance customer service and satisfaction, facilitate decision making, and streamline business processes.
- Limitations
  - they can be very expensive to build and to maintain.
  - incorporating data from obsolete mainframe systems can be difficult and expensive.
  - people in one department might be reluctant to share data with other departments.

3.5 Knowledge Management

- Concepts and Definitions
- Knowledge Management Systems
- The KMS Cycle

Concepts and Definitions

- Knowledge Management
  - a process that helps organizations manipulate important knowledge that comprises part of the organization's memory, usually in an unstructured format.
- Knowledge
  - data are a collection of facts, measurements, and statistics; information is organized or processed data that are timely and accurate.
  - Knowledge is information that is contextual, relevant, and useful. It is information in action. Intellectual capital (or intellectual assets) is another term for knowledge.
- Explicit Knowledge
  - more objective, rational, and technical knowledge.
  - In an organization, explicit knowledge consists of the policies, procedural guides, reports, products, strategies, goals, core competencies, and IT infrastructure of the enterprise.
  - explicit knowledge is the knowledge that has been codified (documented) in a form that can be distributed to others or transformed into a process or a strategy
- Tacit Knowledge
  - the cumulative store of subjective or experiential learning.
  - In an organization, tacit knowledge consists of an organization's experiences, insights, expertise, know-how, trade secrets, skill sets, understanding, and learning.
  - It also includes the organizational culture, which reflects the past and present experiences of the organization's people and processes, as well as the organization's prevailing values.
  - It is generally imprecise, costly to transfer, and highly personal.
  - because it is unstructured, it is difficult to formalize or codify

Concepts and Definitions

- Knowledge Management Systems
  - refer to the use of modern information technologies—the Internet, intranets, extranets, databases—to systematize, enhance, and expedite intra-firm and inter-firm knowledge management.
  - KMSs are intended to help an organization cope with turnover, rapid change, and downsizing by making the expertise of the organization's human capital widely accessible.

Concepts and Definitions

- The KMS Cycle
- **Create knowledge:** Knowledge is created as people determine new ways of doing things or develop know-how. Sometimes external knowledge is brought in.
- **Capture knowledge:** New knowledge must be identified as valuable and be represented in a reasonable way.
- **Refine knowledge:** New knowledge must be placed in context so that it is actionable. This is where tacit qualities (human insights) must be captured along with explicit facts.
- **Store knowledge:** Useful knowledge must then be stored in a reasonable format in a knowledge repository so that other people in the organization can access it.
- **Manage knowledge:** Like a library, the knowledge must be kept current. It must be reviewed regularly to verify that it is relevant and accurate.
- **Disseminate knowledge:** Knowledge must be made available in a useful format to anyone in the organization who needs it, anywhere and anytime.

Figure 3.8: The Knowledge Management System Cycle