CHAPTER 3

Data and Knowledge Management

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

>>>

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. Recognize the necessary environment to successfully implement and maintain data warehouses.
5. Describe the benefits and challenges of implementing knowledge management systems in organizations.

OPENING >

- **Flurry Gathers Data from Smartphone Users**

1. Do you feel that Flurry should be installed on your smartphone by various app makers without your consent? Why or why not? Support your answer.
2. What problems would Flurry encounter if someone other than the smartphone's owner uses the device? (Hint: Note how Flurry gathers data.)
3. Can Flurry survive the privacy concerns that are being raised about its business model?

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, personal, external)
- New sources of data

The Difficulties of Managing Data (continued)

- Data Degradation
- Data Rot
- Data security, quality, and integrity are critical
- Legal requirements change frequently and differ among countries & industries

’S ABOUT BUSINESS 3.1

- **New York City Opens Its Data to All**


© mura/iStockphoto

1. What are some other creative applications addressing city problems that could be developed using NYC’s open data policy?
2. List some disadvantages of providing all city data in an open, accessible format.

Data Governance

Data Governance: is an approach to managing information across an entire organization involving a formal set of unambiguous rules for creating, collecting, handling, and protecting its information.One strategy for implementing data governance is Master Data Management.

Master Data Management: a strategy for data governance involving a process that spans all organizational business processes and applications providing companies with the ability to store, maintain, exchange, and synchronize a consistent, accurate, and timely 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.

- Businesses first adopted computer applications (mid-1950s) until the early 1970s, organizations managed their data in a file management environment.

- Each application required its own data, which were organized in a data file.

- A data file is a collection of logically related records.

- In a file management environment,each application has a specific data file related to it.

- This file contains all of the data records the application requires.

Using databases eliminates many problems that arose from previous methods of storing and accessing data, such as file management systems.

3.2 The Database Approach

- Data File
- Database Systems Minimize & Maximize Three Things
- The Data Hierarchy
- The Relational Database Model

Database Management Systems (DBMS) Minimize:

- 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.

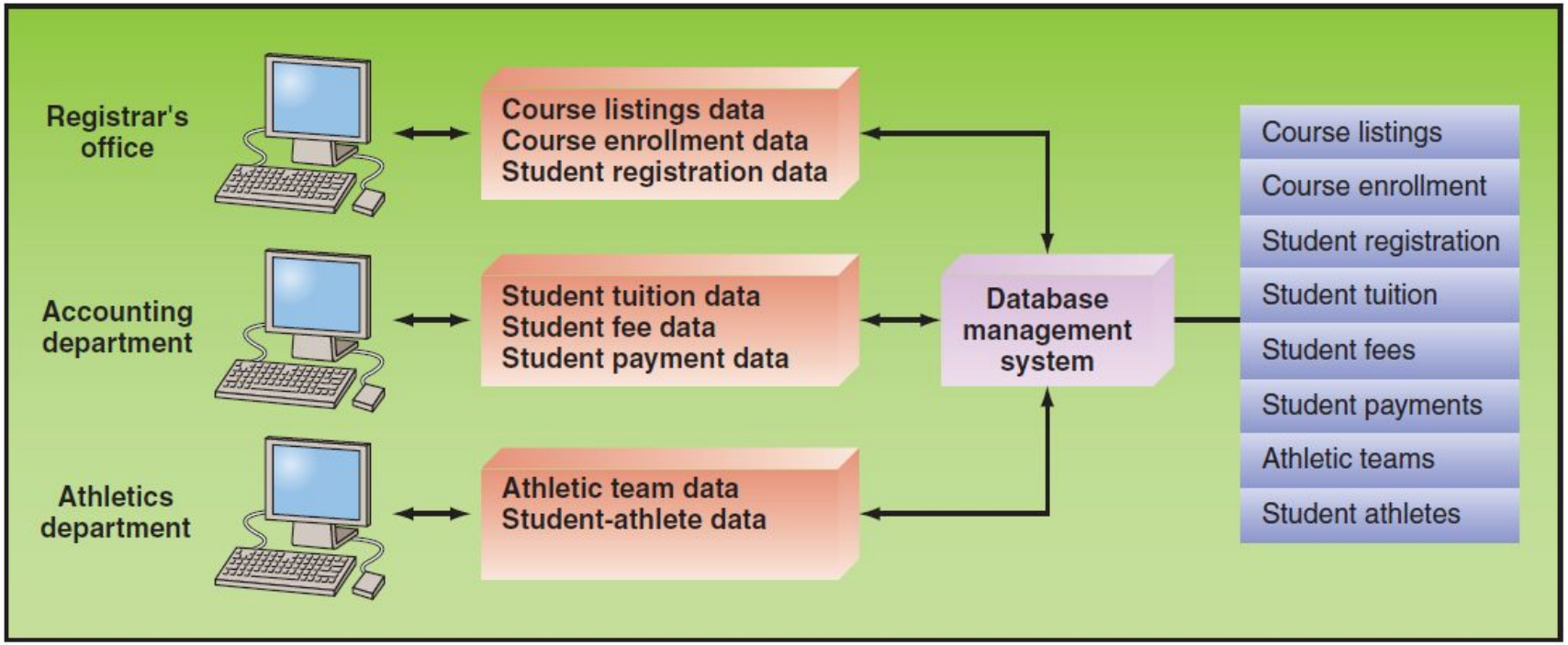
Database Management Systems (DBMS) Maximize:

- Data Security: Because data are "put in one place" in databases, there is a risk of losing a lot of data at once. Therefore, databases 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.

IT'S ABOUT BUSINESS 3.2

- **Google's Knowledge Graph**
  1. Refer to the definition of a relational database. In what way can the Knowledge Graph be considered a database? Provide specific examples to support your answer.
  2. Refer to the definition of an expert system in Plug IT In 5. Could the Knowledge Graph be considered an expert system? If so, provide a specific example to support your answer.
  3. What are the advantages of the Knowledge Graph over traditional Google searches?

Figure 3.1: Database Management System



Data Hierarchy

- Bit
- Byte
- Field
- Record
- Data File (Table)
- Database

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

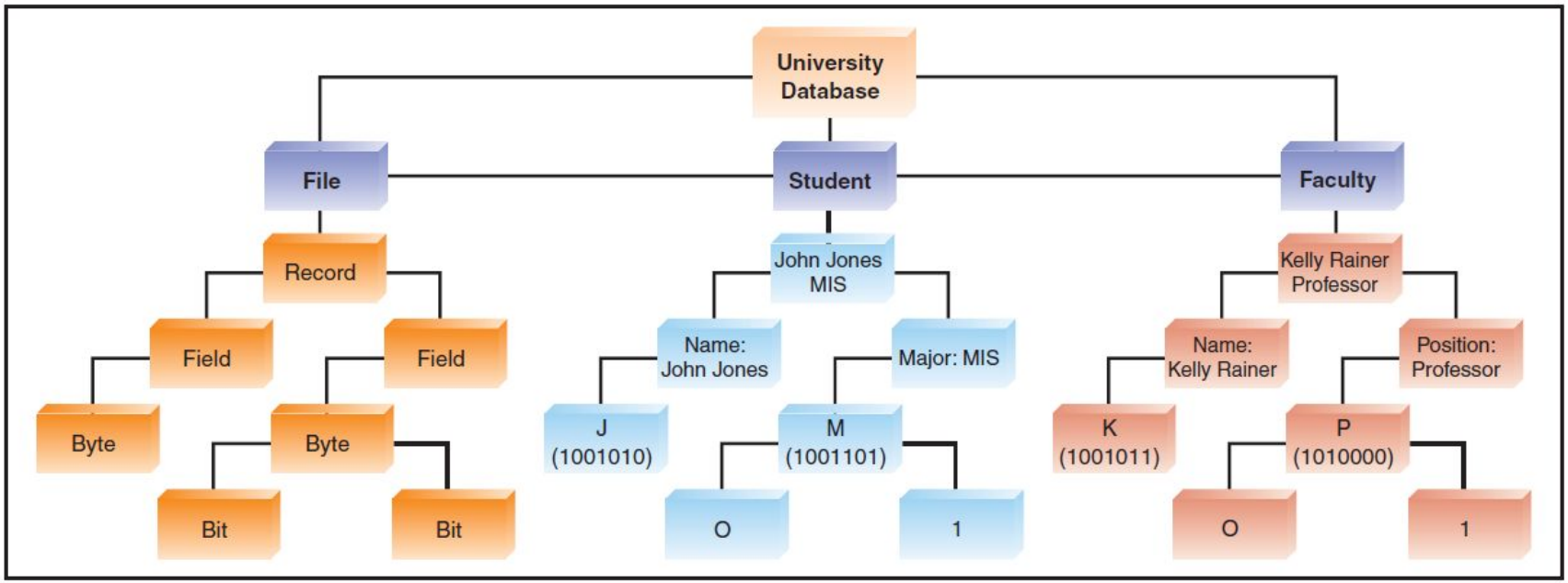


Figure 3.2 Hierarchy of data for a computer-based file.

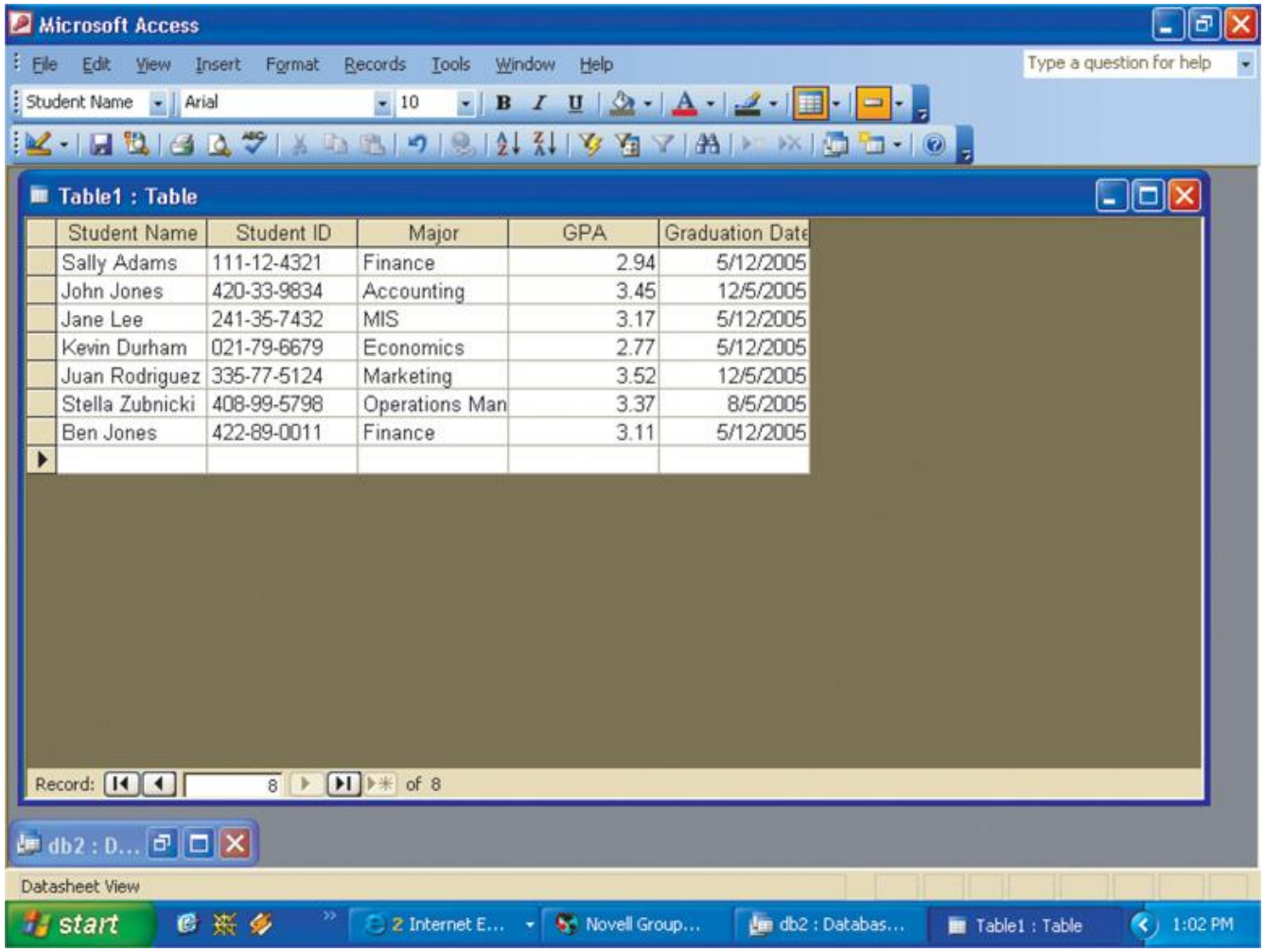
The Relational Database Model

- Database Management System (DBMS)
- Relational Database Model
- Data Model
- Entity
- Instance
- Attribute

The Relational Database Model (continued)

- Primary Key
- Secondary Key
- Foreign Key

Figure 3.3: Student Database Example



3.3 Big Data

- Defining Big Data
- Characteristics of Big Data
- Issues with Big Data
- Managing Big Data
- Putting Big Data to Use

Defining Big Data

- **Gartner (www.gartner.com)**
- **Big Data Institute**

Defining Big Data: Gartner

- Diverse, high volume, high-velocity information assets that require new forms of processing to enable enhanced decision making, insight discovery, and process optimization.

Defining Big Data: The Big Data Institute (TBDI)

- Vast Datasets that:
  - Exhibit variety
  - Include structured, unstructured, and semi-structured data
  - Generated at high velocity with an uncertain pattern
  - Do not fit neatly into traditional, structured, relational databases
  - Can be captured, processed, transformed, and analyzed in a reasonable amount of time only by sophisticated information systems.

Examples of Big Data

Big Data generally consists 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.

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 include satellite imagery, broadcast audio streams, digital music files, Web page content.

Issues with Big Data

- Big Data can come from **untrusted sources**.
- **Big Data is dirty:** Dirty data refers to inaccurate, incomplete, incorrect, duplicate, or erroneous data.
- **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 can reveal valuable patterns, trends, and information that were previously hidden:
  - 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.
- Second Step:
  - making sense of their proliferating data.

Managing Big Data (continued)

- Many organizations are turning to NoSQL databases to process Big Data

’S ABOUT BUSINESS 3.3

- **The MetLife Wall**
  1. Describe the problems that MetLife was experiencing with customer data before it implemented the MetLife Wall.
  2. Describe how these problems originated.

Leveraging Big Data:Ways to leverage big data to gain value

- Making Big Data Available
- Enabling Organizations to Conduct Experiments
- Micro-Segmentation of Customers
- Creating New Business Models
- Organizations Can Analyze Far More Data

Making Big Data Available: Making Big Data available for relevant stakeholders can help organizations gain value.

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.

Micro-Segmentation of Customers: Segmentation of a company's customers means dividing them up into groups that share one or more characteristics.

Creating New Business Models:

- Companies are able to use Big Data to create 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.

Organizations Can Analyze Far More Data: In some cases, organizations can even process all the data in a population relating to a particular phenomenon, meaning that they do not have to rely as much on sampling.

3.4 Data Warehouses and Data Marts

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

Describing Data Warehouses and Data Marts

- Organized by business dimension or Use online analytical processing (OLAP)
- Integrated
- Time variant
- Nonvolatile
- Multidimensional

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.

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): involves analysis of accumulated data by end users.

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).

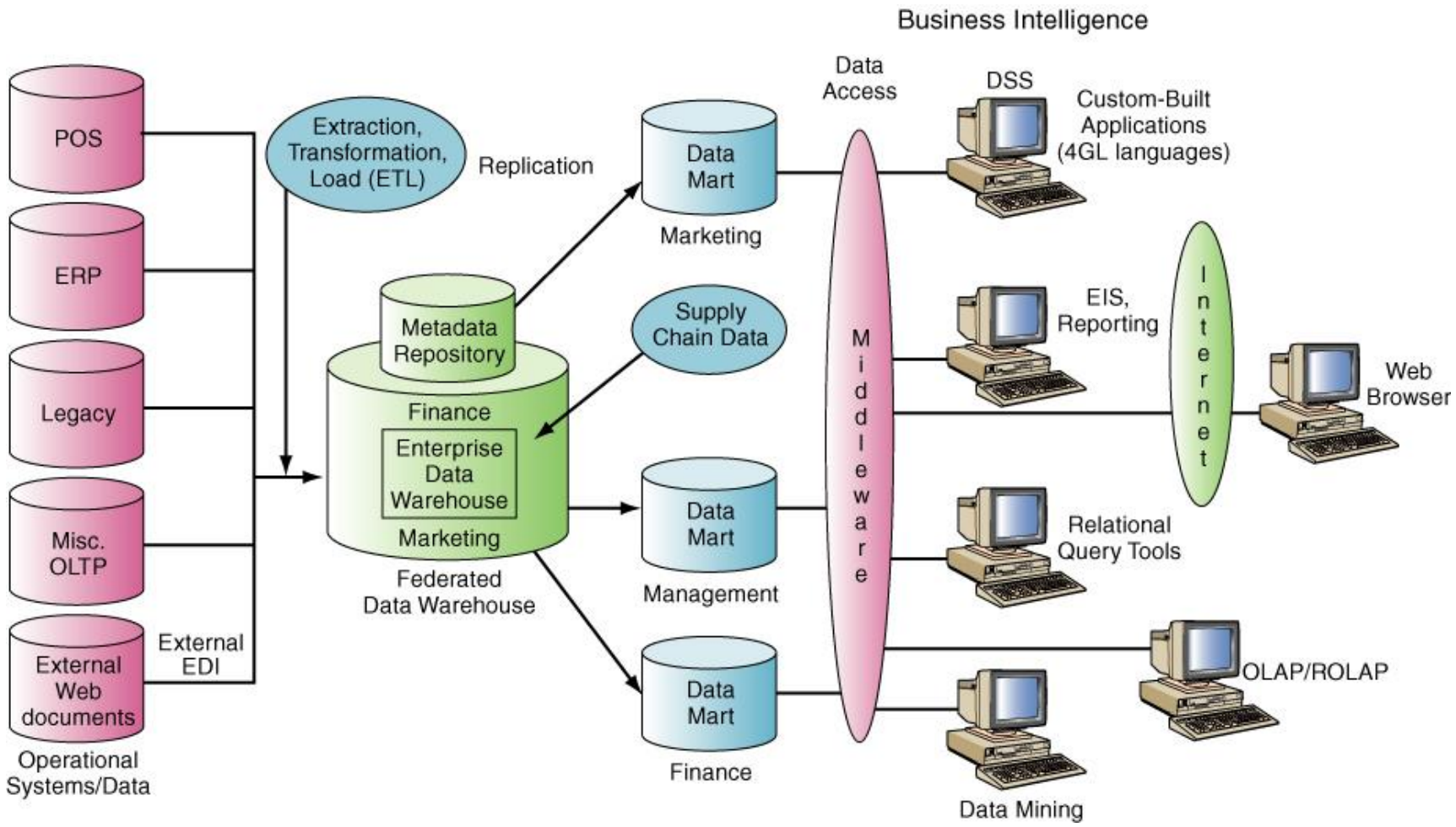
Nonvolatile - Data warehouses and data marts are nonvolatile—that is, users cannot change or update the data.

Multidimensional - Typically the data warehouse or mart uses a multidimensional data structure. Recall that relational databases store data in two-dimensional tables.

A Generic Data Warehouse Environment

- Source Systems
- Data Integration
- Storing the Data
- Metadata
- Data Quality
- Governance
- Users

Figure 3.4: Data Warehouse Framework



Source Systems: Systems that provide a source of organizational data. Common Examples of Source Systems Include:

- operational/transactional systems
- enterprise resource planning (ERP) systems
- Web site data
- third-party data (e.g., customer demographic data)
- operational databases

Data Integration: 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.

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.

Metadata: data maintained about the data within the data warehouse. (e.g., database, table, and column names; refresh schedules; and data-usage measures.

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.

Example: To demonstrate difference between Relational database and Multidimensional data warehouses and data marts

Figure 3.5: Relational Databases

(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: Multidimensional database as Data Cube

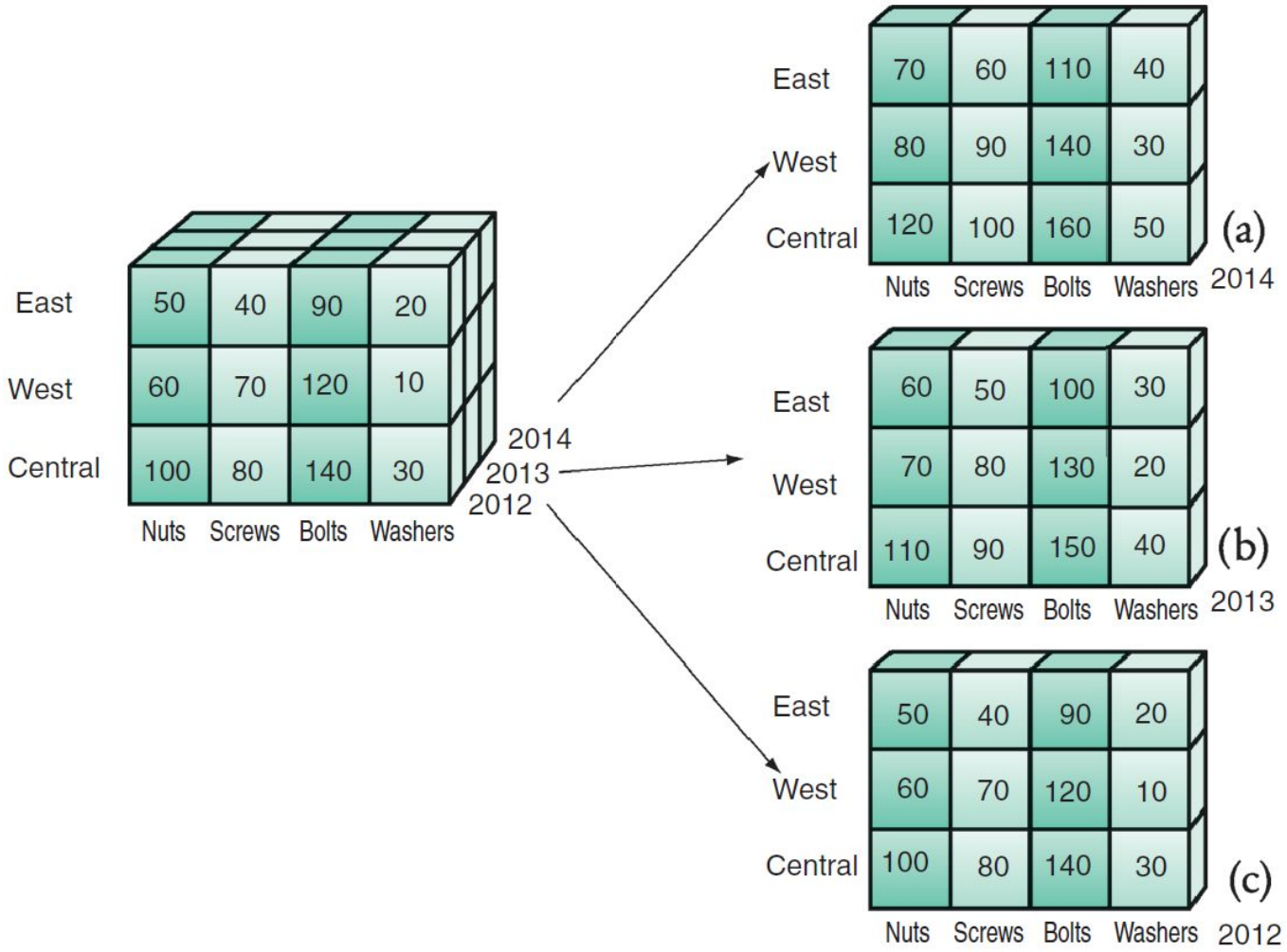
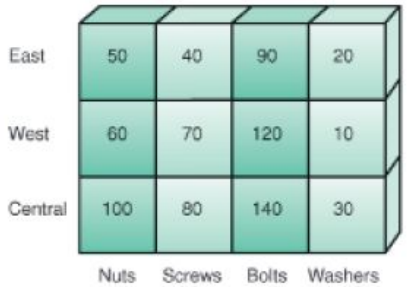


Figure 3.7: Equivalence Between Relational and Multidimensional Databases

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

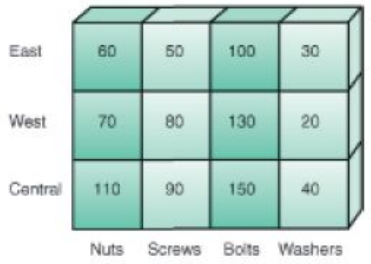
2012



(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

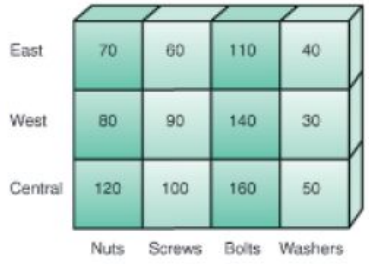
2013



(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

IT'S ABOUT BUSINESS 3.4

Data Warehouse Gives Nordea Bank a Single Version of the Truth


© halbergman/iStockphoto

1. What are other advantages (not mentioned in the case) that Nordea Bank might realize from its data warehouse?
2. What recommendations would you give to Nordea Bank about incorporating Big Data into their bank's data management? Provide specific examples of what types of Big Data you think Nordea should consider.

MANAGING DATA RESOURCES

Need:ESTABLISHING AN INFORMATION POLICY

- Every business, large and small, needs an information policy.
- Firm's data are an important resource
- 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.

INFORMATION POLICY

- 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.
- 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

- It 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.
  Data governance used to describe many of these activities.

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.
- 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.
- Database must be properly designed and enterprise-wide data standards established, so that the 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.

3.5 Knowledge Management

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

Concepts and Definitions

- Knowledge Management
- Knowledge
- Explicit and Tacit Knowledge
- Knowledge Management Systems
- The KMS Cycle

Knowledge management (KM): a process that helps organizations manipulate important knowledge that comprises part of the organization's **Knowledge:** 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.
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 is generally imprecise and costly to transfer.
Knowledge management systems (KMSs): 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.

The KMS Cycle Consists of Six Steps:

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 Cycel