UNIT 3

Business intelligence: Data Warehousing, Data Acquisition, Business Analytics & Visualization: The Nature and Sources of Data, Data Collection, **Problems and Quality, The Web/Internet and Commercial Database Services, Database Management System in Business Intelligence**, Data Warehousing, Data Marts, Business Intelligence, Online Analytical Processing, Data Mining, Data Visualization, Multidimensionality and Real Time Analytics, Business Intelligence, and the Web

Textbook 2 : Chapter 5: 5.1,5.2,5.3,5.4,5.5,5.6,5.7,5.8,5.9,5.10,5.11,5.12,5.14

All Decision support systems use data, information and/or knowledge

Data:

- Items about things, events, activities and transactions are recorded, classified and stored but **not organized to convey any specific meaning**
- Data items can be **numeric, alphanumeric, figures, sounds or images**

Information:

- Data that have been recognized in a manner that **gives them meaning for the recipient**
- They confirm **something the recipient knows** or may have "surprise" value by **revealing something not known.**
- An MSS application processes data items so that the **results are meaningful for an intended action or decision**

Knowledge:

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

- MSS data can include **documents, pictures, Maps, sound, video and animation**
- MSS data can be **stored and organized** in different eays **before and after use**
- MSS data also include **concepts, thoughts, and Opinions**
- Many MSS applications use **summary or extracted Data** that come from three primary sources:

  i) Internal
  ii) External
  iii) Personal

Internal Data:

- Internal data are **stored in one or more places**
- Internal data are **about people, products, services and processes**

  Example –
  - Data about employees , their pay
  - Data about **equipment and machinery can be stored in the maintenance** department database
  - Sales data can be stored in several places:
    Aggregate sales data in the corporate database and details at each region's data base

- An MSS can **use raw data as well as processed data**

- Internal data are available via an **organization's Intranet or other internal network**

External Data:

- There are many **sources of external data**
- They **range from commercial databases to data collected by sensors and satellites**
- Data are available **on any repositories, on the internet, as films and photographs and as music or voices**
- **Government reports and files are a major source of external data,** most of which are **available on the web** today

- External data may also be available by using **GIS, from federal census bureaus, and other demographics sources** that gather data either **directly from customers or from data suppliers**

- **Chambers of commerce, local banks, research institutions and the like, flood the environment with data and information**, resulting in information overload for the MSS user

- **Data can come from around the globe**

Personal data and knowledge:

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

5.3 DATA COLLECTION, PROBLEMS, AND QUALITY

Methods for collecting raw data:

- Raw data can be collected **manually or by instruments and sensors**
- Representative data collection methods are **time studies, surveys, Observations and soliciting information from experts**
- In addition, **sensors and scanners** are increasingly being used in data Acquisitions

- The most reliable method of data collection is from **point of purchase Inventory control**
    - When you buy something , the **register record sales** information with your personal information collected from your credit card
    - This has enabled **Wal-Mart, Sears, and other retailers** to build **complete, massive data warehouses** in which they **collect and store Business intelligence data** about their customers
    - Information is then used to identify **customer buying patterns to Manage local store inventory and identify new merchandising opportunities**
    - It also helps the **retail organization manage its suppliers**

- The need for **reliable, accurate data** for any MSS is universally accepted
    - Some methods involve dealing with data capture problem by using **Bar codes or RFID technology**
        - An RFID electronic button sends an identification signal with some data directly to a nearby receiver
    - Even **Biometric devices** are used to collect real world data.
        - Biometric systems **detect various physical and behavioral features** of individuals and access them to authenticate the identities of visitors and immigrants etc.
    - **Databases and data mining** methods are also used

Case study :

RFID TAGS HELP AUTOMATE DATA COLLECTION AND USE

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

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

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

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

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

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

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

Data Problems

- All computer-based systems **depend on data**
- The **quality and integrity of the data are critical** if the MSS is to **avoid** the **GIGO (Garbage In Garbage Out)** syndrome
- **MSS depend on data** because compiled data that make up **information and knowledge are at the heart of any decision making system**
- Data must be **available to the system or the system must include a data acquisition subsystem**
- **Data issues should be considered in the planning stage** of the system development
- If **too many problems are anticipated, the costs of solving them can be estimated**.
    - **If they are excessive, the MSS project should not be undertaken or should be put on hold until costs and problems decrease**

TABLE Data Problems

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

Data Quality

- Data Quality is an **extremely important issue** because
  - Quality **determines the usefulness of data as well as the quality of the decisions** based on them
- Data in organizational databases are frequently found to be **Inaccurate, incomplete or ambiguous**
- Data quality often generates little enthusiasm and is typically Viewed as a maintenance function
- Data quality **is a major problem in data warehouse development** and **Business intelligence/business analytics** utilization
- Data quality can **delay the implementation of a warehouse or a data mart six months or more**
- Inaccurate data stored in a data warehouse and then reported to someone will **instantly destroy a users trust in a new system**

TABLE Source of Data Quality Problems

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

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

- ***Contextual DQ:*** Relevancy, value added, timeliness, completeness, amount of data
- ***Intrinsic DQ:*** accuracy, objectivity, believability, reputation
- ***Accessibility DQ:*** accessibility, access security
- ***Representation DQ:*** interpretability, ease of understanding, concise representation, consistent representation.

- Data quality is important, especially **CRM, ERP and other enterprise** information systems
- Problem is that **data warehousing, e- business and CRM projects often expose poor quality** data because they **require companies to extract and integrate data from multiple operational systems** that are often peppered with errors, missing values and integrity problems
- Improved data quality is the **result of a business improvement process designed** to **identify and eliminate the root causes of bad data**
- Data warehouse applications **require data cleansing every time** the warehouse **is populated or updated**.

DSS IN FOCUS 5.5

A DATA QUALITY ACTION PLAN

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

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

Source: Adapted from Berg and Heagele (1997).

DSS IN FOCUS 5.6

BEST PRACTICES FOR DATA QUALITY

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

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

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

Data Integrity

- One of the **major issues of DQ is data integrity**
- **Older filing systems may lack integrity**
- **Change made in the file in on place may not be made in the file in another place or department**
  - This results in conflicting data
  - This is especially important issue in collaborative computing environment
- Gray and Watson distinguish the following 5 issues

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

Data Access and integration

- A decision maker typically needs **access to multiple sources of data** that must be integrated

- Before data warehouses, data marts and business intelligence software, **providing access to data sources was a major, laborious process**

    - With modern web based data management tools, web based data management tools, **recognizing what data to access and providing it to the decision maker is a non trivial talk**

- **Needs of Business Analytics continue to evolve,**
- In addition to **historical, cleansed, consolidated and Point in time data**, business users increasingly demand access
- To real time, **understand and/or remote data**.

- In addition, **everything has to be integrated with the contents of their existing data warehouse**

Enterprise data resources can take many different forms: Relational Database (RDB) tables, XML documents, Electronic Data Interchange (EDI) messages, COBOL records, and so on. Independent Software Vendor (ISV) applications, such as enterprise resource planning, customer relationship management software, and in-house-developed software, define their own input and output schemas.

Data integration via XML

- XML is quickly becoming the **standard language for data base integration and data transfer**
- XML is an **excellent way to exchange data among applications** and **organizations**

Issues with Integration:

- **A** critical **Issue is whether it can function well as a native data vase format in practice**
- XML is **mismatch with relational databases: it works, but us hard to maintain**
- Difficulties -> **searching large databases**

DSS IN FOCUS 5.9

WHAT TO DO AND WHAT NOT TO DO WHEN IMPLEMENTING AN ENTERPRISE-WIDE INTEGRATION PROJECT

WHAT TO DO:

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

WHAT NOT TO DO:

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

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

Data integration Software:

- Developers of document and data capture and management software are increasingly **utilizing XML** to Transport data from sources to destinations.

Example:

- Captiva Software Corp.
- RTSe USA Inc
- Kofax Image Products Inc.
- Tower Software
- RosettaNet XML solutions
- Etc.

5.4 THE WEB/INTERNET AND COMMERCIAL DATABASE SERVICES

Web/Internet:

- Many thousands of databases all over the world are accessible through the web/internet
- A decision maker **can access** the home pages of vendors, Clients, and competitors, **view** and **download** information or **conduct research**
- **Internet is the major supplier of external data for many decision situations**

Commercial data banks:

- Online (commercial) database service sells access to specialized databases.
  - Such a service can add external data to the MSS in a **timely manner and at a reasonable cost**
  - Example: GIS data must be accurate
    Regular updates are available

The collection of data from multiple external sources may be complicated. Products from leading companies, such as Oracle, IBM, and Sybase, can transfer information from external sources and put it where it is needed, when it is needed, in a usable form.

TABLE Representative Commercial Database (Data Bank) Services

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

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

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

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

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

THE WEB AND CORPORATE DATABASES AND SYSTEMS

- Developments in **document management systems (DMS)** and **content management systems (CMS)** include use of Web
- Critical issues become more critical in web based systems
- Web browsers by employees and customers to **access vital information**

- It is important to maintain accurate, up-to-date versions of the Document, data and other content, since otherwise the **value of the information will diminish**

- **Managers expect** their DMS and CMS to **produce up-to-the-minute accurate documents and information about** the **Status of the organization** as it relates to their work
  - This real time access to data introduces new complications in the design and development of data warehouses and the tools that access them

Example –
- pilotsw.com,
- blueisle.com,
- comshare.com,
- group support systems deployed via web browsers such as lotus notes/domino and groove etc,
- dbms systems that provide access through web browsers etc

Client/Server Architecture and the internet/Intranet applications:

- **Incorporate non traditional , or rich, multimedia data types**.

❑ Example:

- Oracle Developer/2000 => able to generate graphical client/server application in PL/SQL code, Oracle's implementation of SQL as well as COBOL, C++, HTML etc

- Other tools provide **web browser capabilities, multimedia authoring and content scripting, object class libraries, OLAP routines** etc

  - Microsoft's .NET strategy supports Web based BI
  - Spider Technologies (spidertech.com) => website + db integrations
  - hart.com
  - next.com
  - netobjects.com
  - oracle.com
  - onewave.com

- **Use of the web had a far-reaching impact on collaborative computing in the form of**
  - Groupware,
  - Enterprise information system – ERP/ERM, CRM, PLM, SCM
  - KMS etc

5.5 DATABASE MANAGEMENT SYSTEMS IN DECISION SUPPORT SYSTEMS/BUSINESS INTELLIGENCE

- **Complexity of corporate data bases and large scale independent MSS databases** makes computer operating systems **inadequate** for an effective and efficient interface between the user and data base
- A **DBMS** supplements standard OS by allowing greater integration of data, complex file structure, quick retrieval and changes, and better data security
- DBMS is a software program **for adding information to a Database** and **updating, deleting, manipulating, storing, and retrieving information**
- DBMS are designed to handle large amounts of information
- Data from DBMS are extracted, and put in a statistical , mathematical or financial model for further manipulation or analysis.

- Major role of DBMS is to manage data
  - ❑ By manage => **create, delete, change and display the data.**
  - ❑ DBMS **enables users to query data as well as to generate reports**

Problems with DBMS and spreadsheets are:

- Confusion about their appropriate role
- DBMS offers capabilities similar to those available in an integrated spreadsheet such as excel, and this enabled the DBMS user to **perform DSS spreadsheet work with a DBMS**
- **Many spreadsheet programs offer a rudimentary set of DBMS** capabilities
- For some applications, DBMS work with several DBs and deal with many more data than a spreadsheet can

- For DSS application**, it is often necessary to work with both data and models**
- interfaces between DBMS and excel are simple and facilitates the exchange of data between more powerful Independent programs
- Small to medium DSS can be built **either by enhanced DBMS or by integrated spreadsheets**
- Alternatively, they can be built with a DBMS program and a spreadsheet program.
- A third approach to the construction of DSS is to **use a fully integrated DSS generator.**

5.6 DATABASE ORGANIZATION AND STRUCTURES

3 conventional structures are:

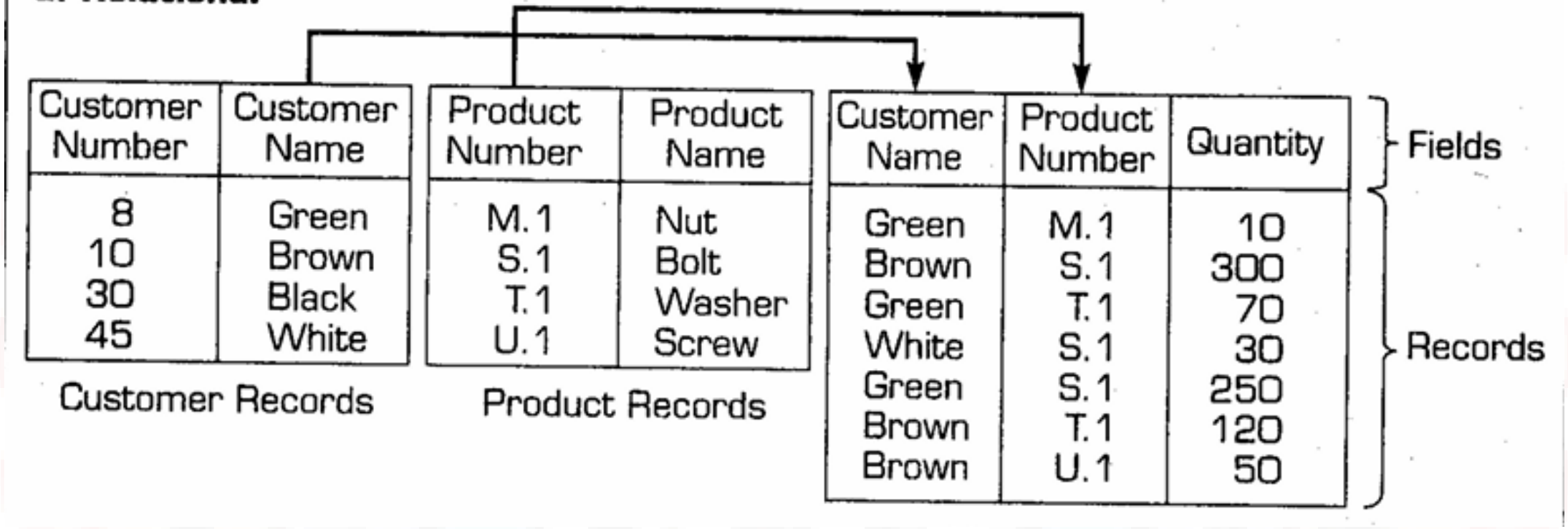
a. Relational
b. Hierarchical
c. Network

- Relational form of DSS database organization, described as flat,
- Allows the user to think in terms of two dimensional tables,
- Which is the way many people see data reports.

- Relational DBMS allow multiple access queries
  => data file consists of a number of columns preceding down a page
  => each column is considered a separate field
  => rows on a page represents individual records made up of several fields, the same design that is used by spreadsheets

- Several data files can be related by means of a common data field found in two data files
  => name of common field must be spelled exactly alike, and the fields must be the same size and type.

a. Relational



Customer Number	Customer Name
8	Green
10	Brown
30	Black
45	White

Customer Records

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

Product Records

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

Fields

Records

@table example -> data filed customer name is found in both customer and usage file and thus they are related

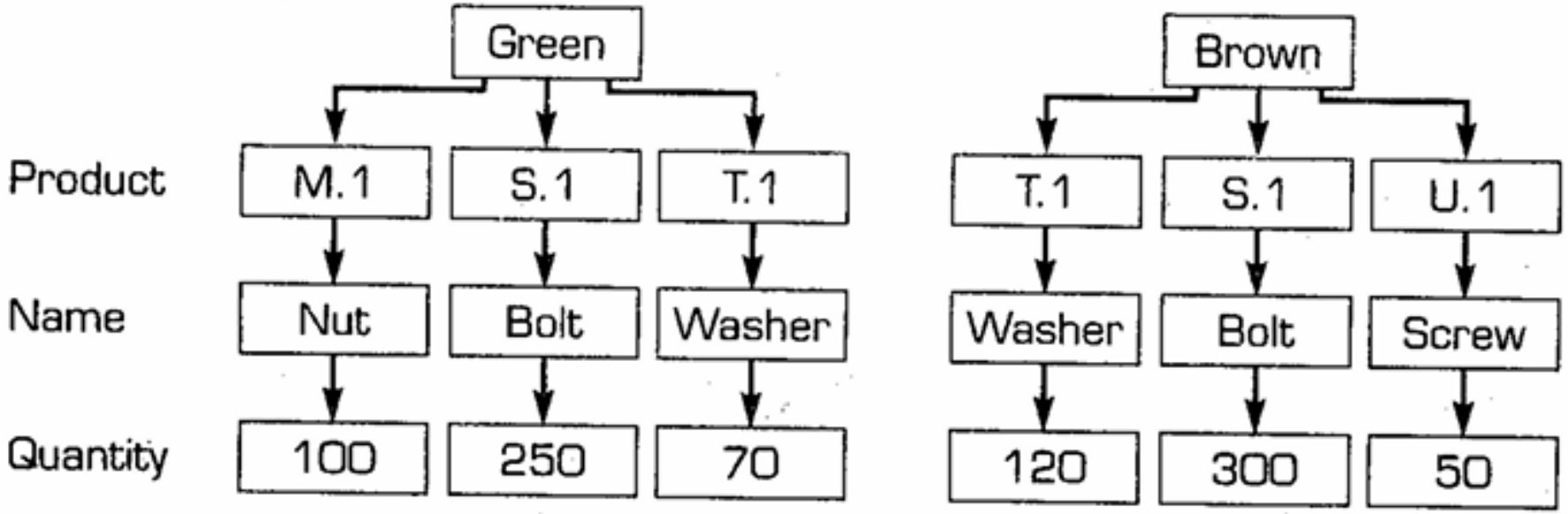
=> Product Number found in product file and usage file .

→ There is a common linkages that all three files are related and in combination form a **relational database,**

Advantage of relational data base:

- It is simple for the user to learn
- Can easily expanded or alters and cab be accessed in a number of formats
- Anticipated at the time of initial design and development of the database
- It can support large amount of data and efficient access
- Many data ware houses are organized this way

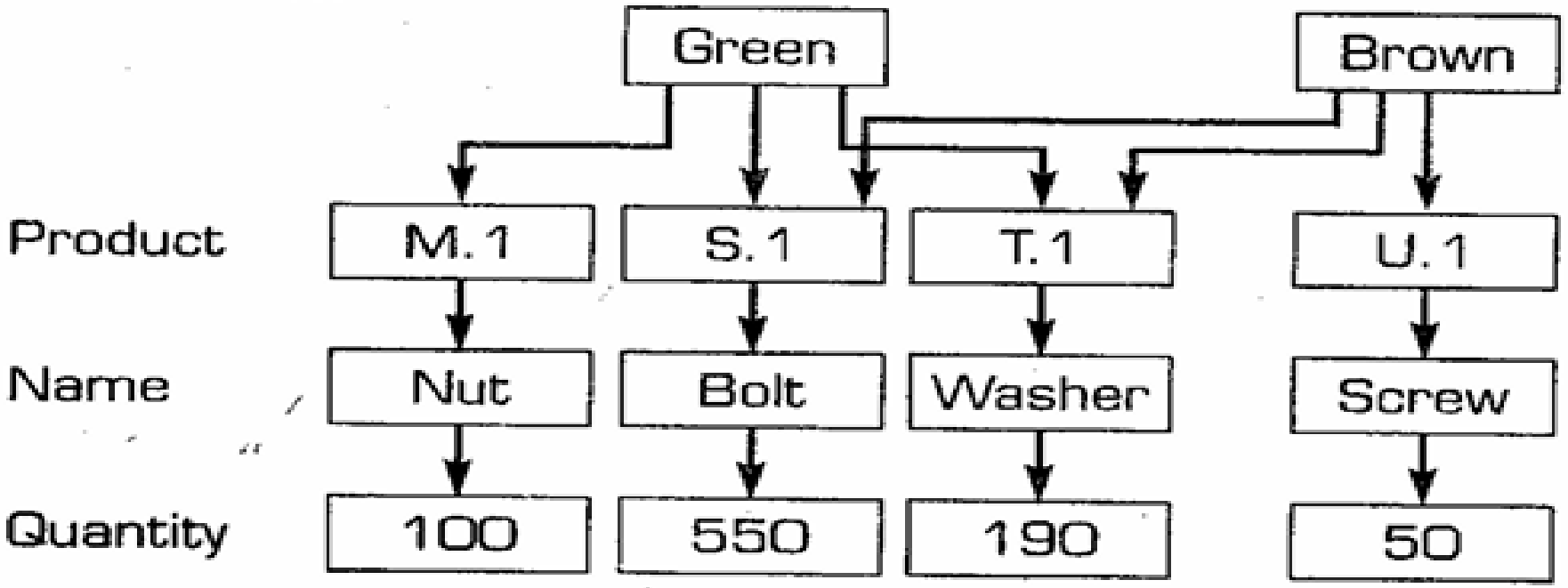
b. Hierarchical



HIERARCHICAL DATABASES

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

c. Network



NETWORK DATABASES

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

***CODASYL**, the **Conference/Committee on Data Systems Languages**,

OBJECT-ORIENTED DATABASES

- Comprehensive MSS applications, such as those involving computer integrates manufacturing (CIM), require accessibility to complex data, which may include pictures and elaborate relationships. Such situations can not be handled efficiently by hierarchical db
- Object oriented data management is based on the principle of object-oriented programming
- Object oriented data base systems combine the characteristics of OOPL, such as Veritos or UML, with a mechanism for data storage and access
- Object oriented tools focus directly on databases
- An object-oriented database management system (OODBMS) allows only one to analyze data to a conceptual level that emphasizes the natural relationships between objects.
- An object oriented data management system defines data as objects and encapsulated data along with their relevant structure and behavior. The system uses a hierarchy of classes and sub classes of objects.
- Structure int terms of relationships, and behavior, in terms methods and procedures are contained within an object

MULTIMEDIA-BASED DATABASES

- **MMDBMS manage data in a variety of format**s, in addition to the standard text or numeric field.
  - These formats include images, such as **digitized photographs, and forms of bit mapped graphics, such as maps or .PIC files, hypertext images, video clips, sound and virtual reality (multi dimensional images)**
- Cataloguing is tricky
- Accurate and known key words must ne used
- It is critical to develop effective way to manage such data for GIS and for many other applications
- Managing multimedia data continues to become more important for BI
- Binary large objects (BLOBS) – ability to bimedia datatype

DSS IN ACTION 5.13

MULTIMEDIA DATABASE MANAGEMENT SYSTEMS: A SAMPLER

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

MediaWay Inc. (mediaway.com) claims that its multimedia database management system can store, index, and retrieve multimedia data (sound, video, graphics) as easily as relational databases handle tabular data. The DBMS is aimed at companies that want to build what MediaWay calls *multimedia cataloging applications* that manage images, sound, and video across multiple back-end platforms. An advertising agency, for example, might want to use the product to build an application that accesses images of last year's advertisements stored on several servers. It is a client/server implementation. MediaWay is not the only vendor to target this niche, however. Relational database vendors, such as Oracle Corporation and Sybase Inc., have incorporated multimedia data features in their database servers. In addition, several desktop software companies promote client databases for storing scanned images. Among the industries that use this technology are health care, real estate, retailing, and insurance.

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

DOCUMENT-BASED DATABASES

- Also known as **electronic document management (EDM) systems**
- Developed to alleviate paper storage and shuffling
- They are used for **information dissemination, from storage and management, shipment tracking, expert license Processing, and workflow automation**
- Many content management systems (CMS) are based on EDM
- EDM uses both OO and MMDBMS
- Unique to EDM -> implementation and the applications.
- Web enabled document management system have become an efficient and effective delivery systems.

INTELLIGENT DATABASES

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

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

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

DSS IN FOCUS 5.15

THE BOTS OF THE FUTURE

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

One goal of software agent developers is to develop machines that perform tasks that people do not want to do. Another is to delegate to machines tasks at which they are vastly superior to humans, such as comparing the price, quality, availability, and shipping cost of items.

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

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

Source: Adapted from S. Ulfelder, "Undercover Agents," *ComputerWorld*, June 5, 2000.

5.7 DATA WAREHOUSING

Consider the scenario,
Information sharing a principal component of the national Security for homeland security

https://www.youtube.com/watch?v=hZyAY91euQg&pp=ygU4aG9tZWxhbmQgc2VjdXJpdHkgYnVzaW5lc3MgaW50ZWxsaWdlbmNlIG9wZW5pbmcgdmlnbmV0dGU%3D

For the scenario,
Information sharing a principal component of the national Security for homeland security =>

Data warehouses provide a strategic data architecture to enable decision support analysis

Data warehousing enables data mining, the ability to automatically synthesize vast amount of information in order to discover hidden truths within the data.

Data portals have emerged as next generation in Web-enabled data warehouses

One of the most significant data portals has been developed in direct response to the terrorist attacks on the US on Sep11, 2001

Because of the scenario,
Information sharing a principal component of the national Security for homeland security

5 major initiatives identified :

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

DATA WAREHOUSING CAN BE USED TO SUPPORT DECISION MAKING and TO ANALYZE LARGE AMOUNT OF DATA

https://www.youtube.com/watch?v=AHR_7jFCMeY

- **Organizations -> public + private**
  ➔ continuously collect data, information and knowledge at an increasing accelerated rate and store them in computerized systems.

- **Updating, retrieving, using and removing information becomes complicated as the data amount increases.**

- **At the same time, the number of users that interact with the information continues to increase as a result of improved reliability and availability of network access, especially including the internet**

- **Working with multiple databases is also a difficult task**

- **Data for the data warehouse are brought in from internal and external resources**

DSS IN ACTION 5.16

DATA WAREHOUSING SUPPORTS FIRST AMERICAN CORPORATION'S CORPORATE STRATEGY

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

- Identification of the top 20 percent of profitable customers
- Identification of the 40–50 percent of unprofitable customers
- Retention strategies
- Lower-cost distribution channels
- Strategies to expand customer relationships
- Redesigned information flows.

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

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

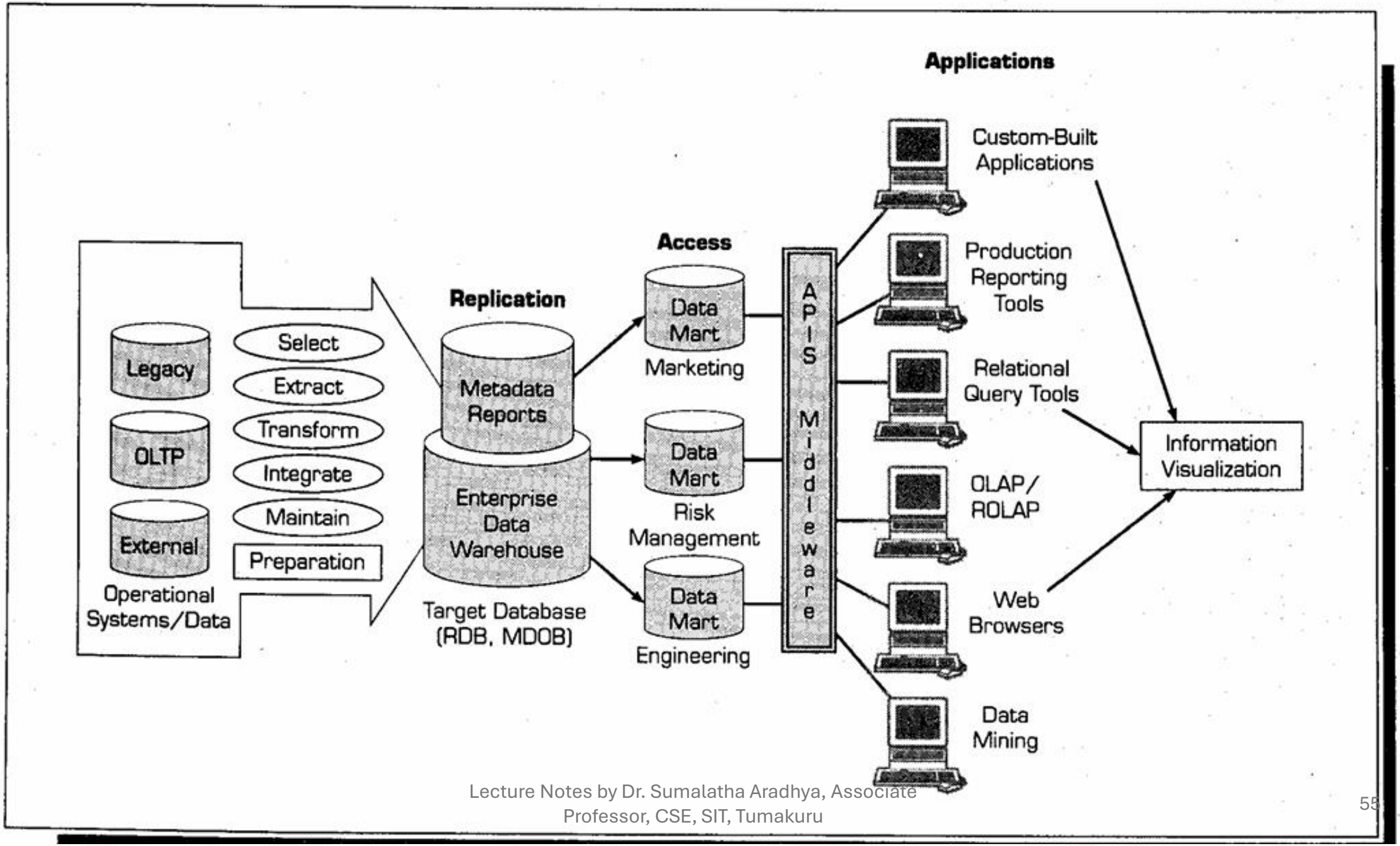
Data should be cleansed and organized in a manner consistent with the organization's needs

Once the data are populated in the data warehouse, data marts may be loaded for a specific area or department

Often, data marts are bypassed and BI tools on client PC's simply load and manipulate local data cubes

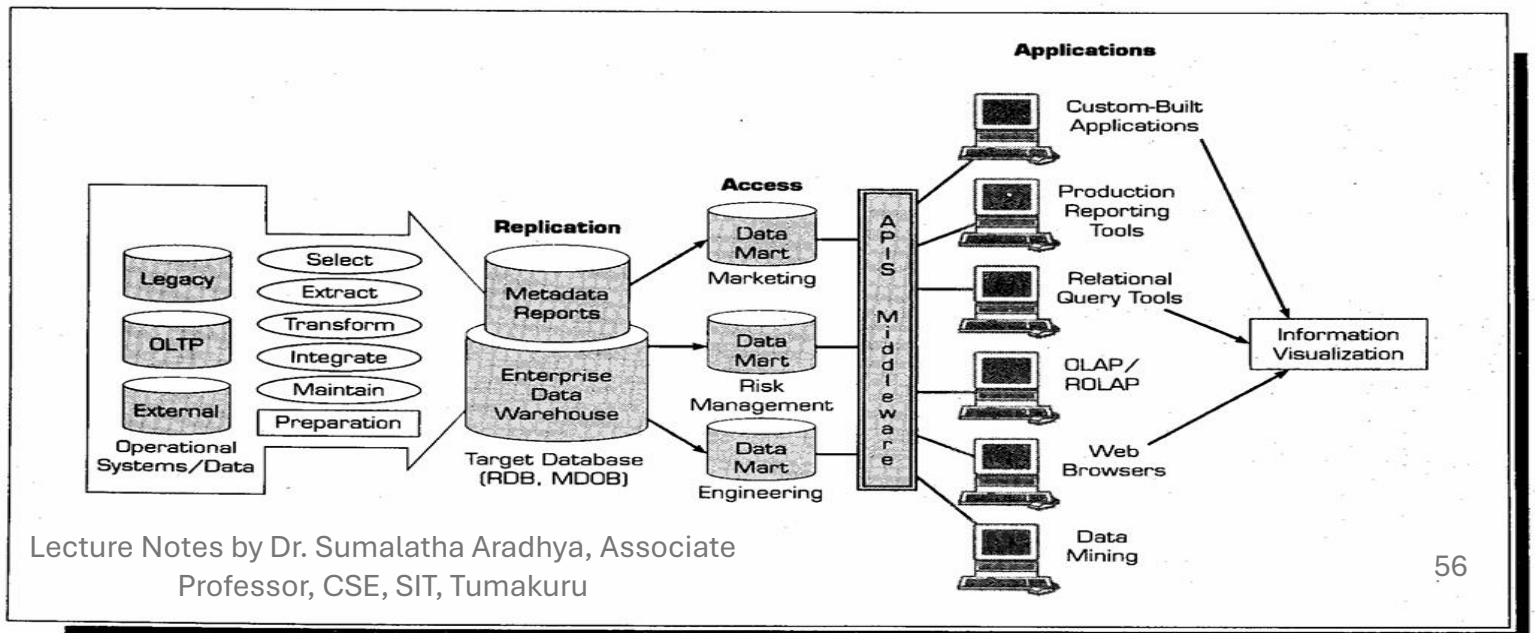
Data warehouses can be described as **subject-oriented, integrated, time-variant, non-normalized, non-volatile** collections of data that support analytical decision-making

FIGURE 5.2 DATA WAREHOUSE FRAMEWORK AND VIEWS



- Figure illustrates data warehouse framework and views
- Figure also illustrates how data work their way into the data warehouse (on the left), for further analysis by tools (to the right)
- Since enterprise information management solutions aggregate or consolidate report information and electronic documents created by any application running on any platform, and reports processed from data warehouse
- An enterprise data warehouse is a comprehensive database that supports all decision analysis required by an organization by providing summarized and detailed information
- The data warehouse has access to all information relevant to the organization, which may come from many different sources, both internal and external

FIGURE 5.2 DATA WAREHOUSE FRAMEWORK AND VIEWS



- A data warehouse begins with the physical separation of a company's operational and decision support environments
- At the heart of many companies lies a store of operational data, usually derived from **critical mainframe-based Online Transaction processing (OLTP) systems, such as order entry point of sales applications**
- Many legacy OLTP systems were implemented primarily in COBOL and still operate in a **customer information control System (CICS) environment.**
- OLTP systems for financial and inventory management and control, **also produce operational data**
- In the operational environment, ***data access, application logic tasks, and data presentation logic are tightly coupled together***, usually in non-relational databases.
- OLTP data are usually detail data that control a specific event, such as recording of a sales transaction, and are generally Not summarized
  → these relational data stores are not very conducive to data retrieval for decision support/business intelligence/Business analytic applications.
- However, decision support information must be made accessible to management
- **IT IS IMPORTANT TO PHYSICALLY SEPARATE THE DATA WAREHOUSE FROM THE OLTP SYSTEM**

CHARACTERISTICS OF DATA WAREHOUSING

1. Subject Oriented

- Data are organized by detailed subject (e.g., By customers, policy types, and claim in insurance company) , containing only INFORMATION relevant for decision support
- Subject orientation enables users to determine not only how their business is performing, but why
- A data warehouse differs from an operational database in that most operational databases have a product orientation and are tuned to handle transactions that update the database
- Subject orientation provides a more comprehensive view of the organization

2. Integrated

- Data at different source locations may be encoded differently

  e.g., gender data -> may be encoded as 0 or 1 in one place
  -> may be encoded as “m” or “f” in another

- In the warehouse, they are *scrubbed* (cleaned) into one format so that they are standardized and consistent.
- Many organizations use the same terms of data of different kinds
  e.g., netsales -> net of commission to the marketing department
  gross sales -> returns to the accounting department
- Integrated data resolve inconsistent meanings and provide uniform terminology throughout the organization
- Also data and time formats vary around the world

3. Time-Variant (time series)

- Data do not provide the current status
- They are kept for five or ten years or more and are used for trends, forecasting and comparisons
- There is a temporal quality to a data warehouse
- TIME IS THE ONE IMPORTANT DIMENSION THAT ALL DATA WAREHOUSES MUST SUPPORT
- Data for analysis from multiple sources contain multiple time points (e.g., daily, weekly, monthly views)

CHARACTERISTICS OF DATA WAREHOUSING

4. Nonvolatile

- Once entered into the warehouse, data are read-only, they cannot be changed or updated
- Obsolete data are discarded and changes are recorded as new data
- This enabled the data warehouse to be tuned almost exclusively for data access
  e.g., large amount of free space (for data growth) typically are not needed, and data base reorganizations can be scheduled in conjunction with
  the load operations of a data warehouse

5. Summarized

Operational data are aggregated when needed into summaries

6. Not normalized

Data in a data warehouse are generally not normalized and highly redundant

7. Sources

All data are present; both internal and external

8. Metadata

(Defined as data about data) are included

METADATA

- Meta data have major impacts on how data warehouses function
- Metadata refers to data about data
- Metadata describe the structure of and some meaning about the data, thereby contributing to their effective or ineffective use
- Metadata hold the key to resolving the challenge of making users comfortable with technology
- Metadata involve knowledge, and capturing and making them accessible throughout an organization have become important success factors
- With metadata and a metadata repository, organizations can dramatically improve their use of both information and application development processes
- Building a metadata repository should be mandatory for many organizations
- Business metadata benefits include the reduction of IT related problems, increased system value to the business and improved business decision making

METADATA

- Business metadata comprises information that increases our understanding of traditional data reported
- The primary purpose of metadata should be to provide context to the data; that is , enriching information leading to knowledge
- Business metadata, though difficult to provide efficiently, releases more of the potential of structured data. Context need not be the same for all users
- **In many ways, metadata assist in the conversion of data and information into knowledge**
- **Metadata form a foundation for a meta business architecture**

Semantic metadata are metadata that describe contextually relevant or domain specific information about content, in the right context. Based on industry-specific or enterprise-specific custom metadata model or ontology

→ This involves putting a level of understanding into metadata

DATA WAREHOUSING ARCHITECTURE AND PROCESS

- There are several architecture for data warehousing
- **Two-tier and three-tier architectures** are commonly used
- McFadden, Hoffer and Prescott (2003) distinguishes among 2-tier and 3-tier by dividing the data warehouse into 3 parts:

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

DATA WAREHOUSING ARCHITECTURE AND PROCESS

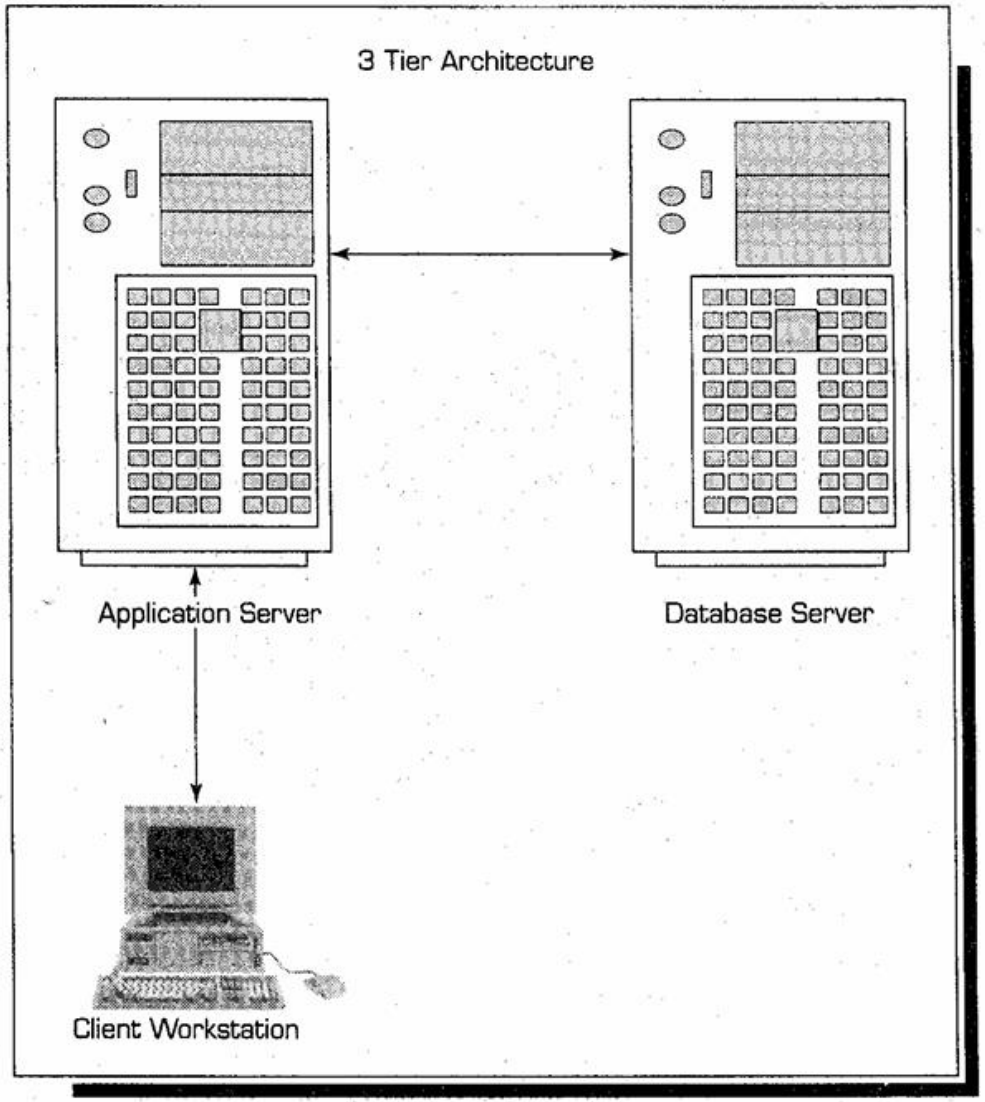


FIGURE 5.3 ARCHITECTURE OF A 3-TIER DATA WAREHOUSE

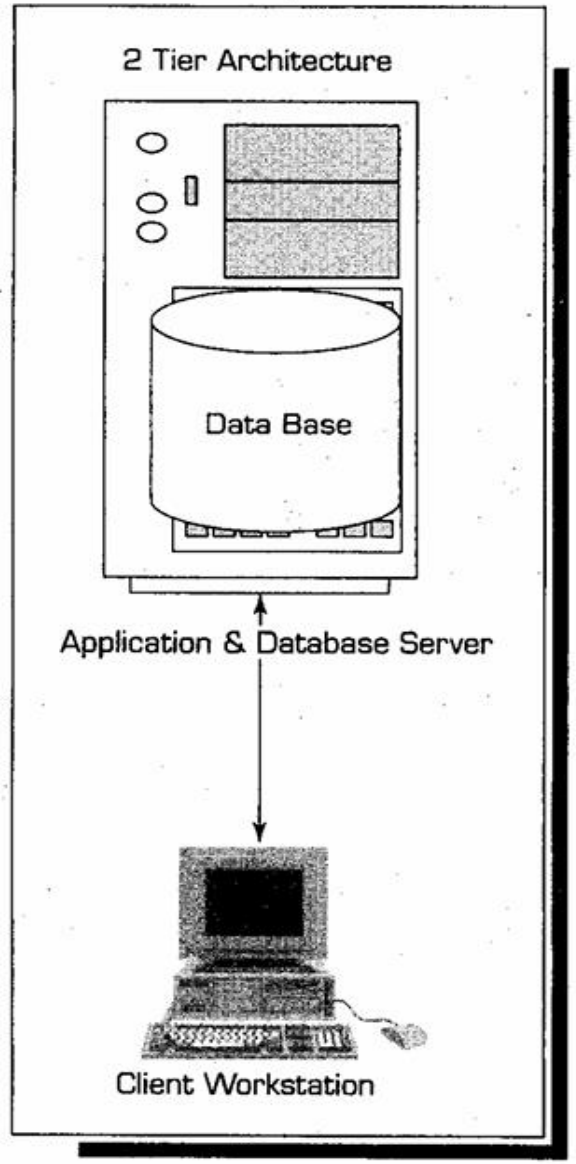


FIGURE 5.4 ARCHITECTURE OF A 2-TIER DATA WAREHOUSE

DATA WAREHOUSING ARCHITECTURE AND PROCESS

- In 3-tier architecture,
  - operational system contain the data,
  - the software for data acquisition in one tier (server),
  - data warehouse in another tier, and
  - the third tier includes the decision-support/business analytics engine (i.e., application server) and the client

Advantages of 3-tier architecture :

- Separation of the functions of the data warehouse , which eliminates resource constraints and makes it possible to easily create data marts

In 2-tier architecture,

- DSS engine is on the same platform as the warehouse

→ It is more economical than the 3-tier structure

DATA WAREHOUSING ARCHITECTURE AND PROCESS

- Web architecture are similar in structure, requiring a design choice for housing the Web data warehouse with the transaction server or as a separate server (s)

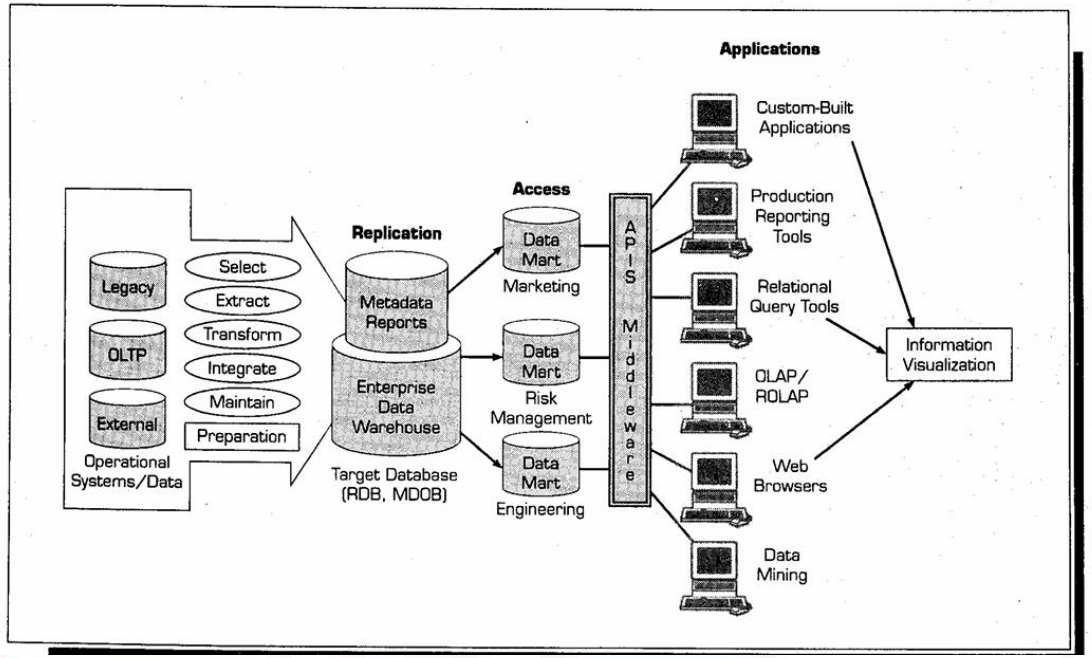
- Page loading speed is an important consideration in designing Web-based application; therefore, server capacity must be carefully planned for

Issues in selecting architecture:

- **Which data base management system to use?**
  - Most use RDBMS
  - Oracle SQL server, DB@(IBM) etc
  - Each supports both client-server architecture and web-based architecture
- **Will parallel processing and/or partitioning be utilized?**
  - Parallel processing enables multiple CPU's to process data warehouse query requests simultaneously and provides scalability
  - Data warehouse designers need to decide whether the database tables will be partitioner for access efficiency and what the criteria will be
  - This is an important consideration that is necessitated by the large amounts of data contained in a typical data warehouse.
  - Teradata has adopted this approach
- **Will data migration tools be used to load the data warehouse?**
- **What tools will be used to support data retrieval and analysis?**

DATA WAREHOUSE DEVELOPMENT

FIGURE 5.2 Data Warehouse Framework and Views



A typical data warehouse structure

- **Process of migrating data to data warehouse involves the extraction of data from all relevant sources**
- **Data sources may consist of files extracted from OLTP databases, spreadsheets, personal databases (e.g., MS Access) or external files**
- **All of the input files are written to a set of staging tables, which are designed to facilitate the load process**
- **A data warehouse contains numerous business rules that define such things as how the data will be used, summarization Rules, standardization of encoded attributes and calculation rules**
- **Any data quality issues pertaining to the source files need to be corrected before the data are loaded into data warehouse**

DATA WAREHOUSE DEVELOPMENT

Benefits of well designed data warehouse:

1. Rules can be stored in the metadata repository
2. Rules are applied to the data warehouse centrally
3. Differs from OLTP approach, which typically has data and business rules scattered throughout the system

Data loading process:

- Can be performed either through **data transformation tools** that provide GUI to aid in the development and maintenance business rule development
  or
- Through more **traditional methods by developing programs or utilities** to load the data warehouse using programming languages such as PL/SQL. C++ or .NET

Few Issues with respect to purchase tools vs build transformation process:

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

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

STAR SCHEMAS

- **Data warehouse design is based upon the concept of dimensional modeling**
- **Dimensional modeling is a retrieval based model that supports high-volume query access**
- **Star schema is the means by which dimensional modeling is implemented**
- A star schema contains a central **fact table**
- **A fact table contains the attributes needed to perform**
  - **Decision analysis, descriptive attributes used for query reporting, and foreign keys to link to dimension tables**
  - The decision analysis attributes consist of **performance measures, operational metrics, aggregated measures, and all other metrics needed to analyse the organization's performance**
- **A fact table primarily addresses what the data warehouse supports for decision analysis**
- Surrounding the central fact tables **are dimension tables**
- **Dimension tables contain attributes that describe the data contained within the fact table**
- Dimension tables address how data will be analyzed
  - Examples of dimensions that would support **a product fact table are location, time**

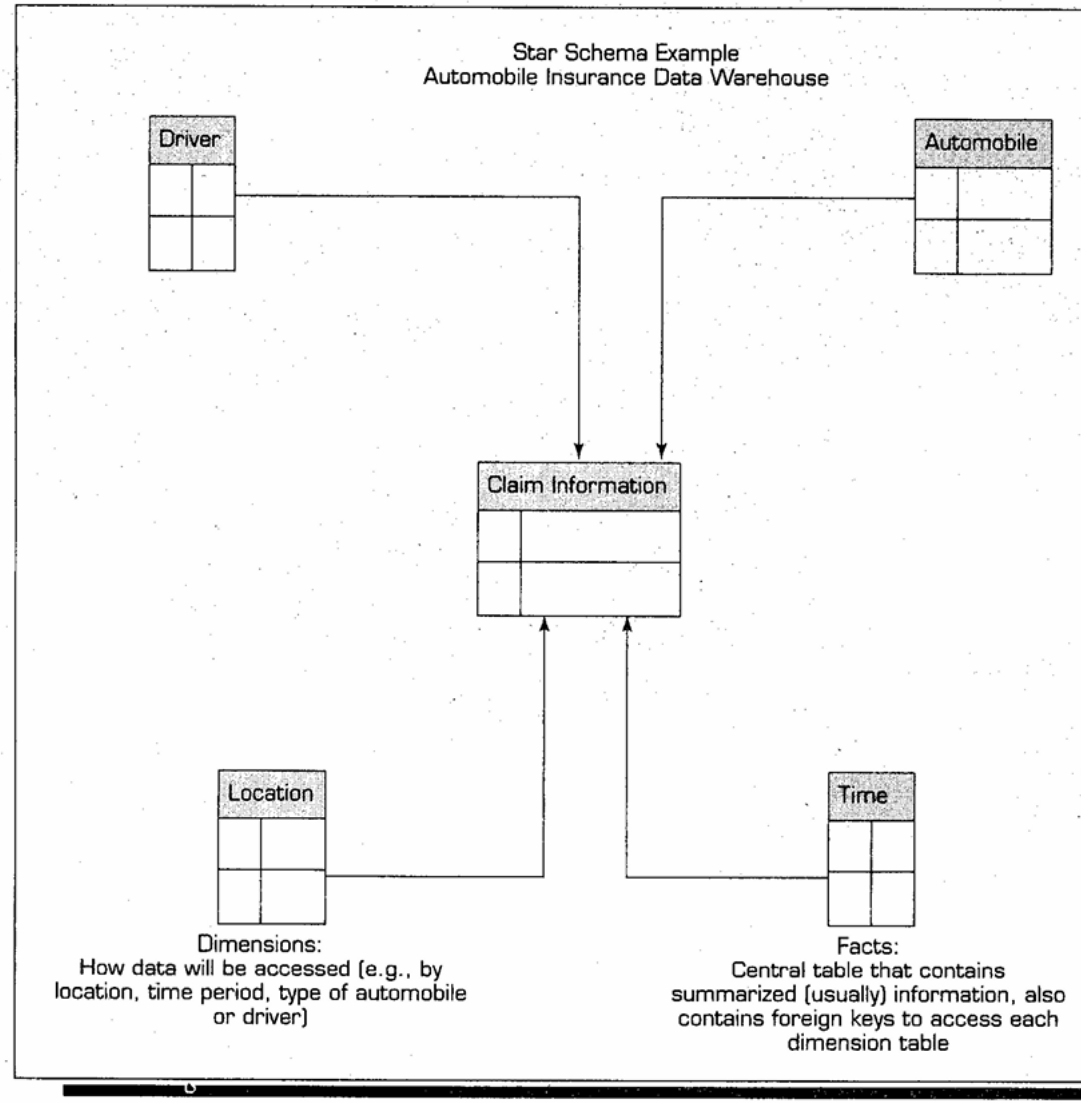


FIGURE 5.5 STAR SCHEMA

STAR SCHEMAS

- The grain of a data warehouse defines the highest level of details that is reported
- The grain will indicate whether the data warehouse is highly summarized or also includes detailed transaction data
- If the grain is defined too high, then the warehouse may not support detail requests to drill down into the data
- Drill down analysis is the process of probing beyond a summarized value to investigate each of the detail transactions that compromise the summary
- A low level of granularity will result in more data being stored in the warehouse
- Larger amounts of detail may impact the performance of queries by making the response times longer
- During the scoping of a data warehouse project, it is important to identify the right level of granularity that will be needed

Implementing data warehousing a massive effort that must be planned and executed according to established methods

Eckerson describes four major ways to develop a data warehouse. These include

1. top-down
2. Bottom-up
3. Hybrid
4. Federated

DSS IN FOCUS 5.17

THE FOUR MAJOR APPROACHES TO BUILDING A DATA WAREHOUSE

There are four major approaches to building a data warehousing environment: (1) top-down, (2) bottom-up, (3) hybrid, and (4) federated. Most organizations follow one or another of these approaches. In the top-down approach, the data warehouse is the center of the analytic environment. It is carefully designed and implemented. The design and implementation of all other aspects of business intelligence are based on it. This approach provides an integrated, flexible architecture to support later analytic data structures. In the bottom-up approach, the goal is to deliver business value by deploying multidimensional data marts quickly. Later these are organized into a data warehouse. The hybrid approach attempts to blend the first two approaches. The federated approach is a concession to the natural forces that undermine the best plans for developing a perfect system. It uses all possible means to integrate analytical resources to meet changing needs or business conditions. Essentially, the federated approach involves integrating disparate systems (see the Opening Vignette and DSS in Action 5.7).

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

- Federated approach is probably the least well known
- Federation is often viewed as a form of information integration
- It complements the traditional ETL and replication approaches by creating and maintaining a logical view od a single Warehouse or mart, where as the data reside in separate systems
- Semantic webs are used to wrap data into containers that reside in repositories in information space
  -> this approach may be the solution to the massive data integration problem facing the Department of Homeland security

- Data warehousing projects may be **data-centric or application- centric**
- A datacentric warehouse is based upon a data model that is **independent of any** application
- It is designed to support a **variety of user needs** and applications
- The methodological approach to designing a data-centric warehouse involves **data modeling with a group of business experts who are familiar with the different information views needed to support the business**
  -> this consists of top-down approach in producing specification of information needs so as to not leave data behind
- It is broad in scope and requires knowledge of current and anticipated data needs
- A mapping approach should be used to provide a structured approach to classification of data
- Data-centric warehouses should support flexibility because enterprise information constantly needs change based upon changes in the underlying business
- The more dynamic the business , the greater the possibility that data needs will change during the development of the data warehouse

- An **application-centric warehouse** is one initially designed to support a single initiative or small set of initiatives
- This is the preferred approach for independent data mart development
- The advantage of application-centric approach is **that it provides a more focused scope, and therefore increases the likelihood of successful data warehouse implantation**
- The disadvantage is that **critical data needs may be left out during the initial development, and therefore multiple iterations may be necessary**

DSS IN FOCUS 5.18

BEST PRACTICES FOR DATA WAREHOUSE IMPLEMENTATION

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

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

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

DSS IN FOCUS 5.19

DATA WAREHOUSE RISKS

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

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

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

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

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

The natural progression of information in a data warehouse is

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

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

Wixcom and Watson defined a research model for data warehouse success that identified following Seven implementation factors that can be categorized into 3 criteria's i.e., Organizational issues, project issues and technical issues

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

- A data warehouse will only be successful **if there is strong senior management support** for its development and a project champion

- The successful implementation of a data warehouse results in the e**stablishment of an architectural framework that may allow for decision analysis throughout an organization and in some cases also provides comprehensive supply chain management** by granting access to an organization's customers and supplier.

- The implementation of Web based data warehouses (Web housing) has **facilitated ease of access to vast amount of data**, but it is **difficult to determine the hard benefits associated with a data warehouse**.
  - **Hard benefits are defined as benefits to an organization that can be expressed in monetary terms**
- Many organizations have **limited information technology resources** and **must prioritize which projects will be worked on first**
- **Data warehouse resources can be a significant cost, in some cases requiring high-end processors and large increases in direct-access storage devices (DASD)**

- User participation in the development of data and access modeling is a critical factor in data warehouse development

- During data modeling, expertise is required to
  - determine what data are needed,
  - define business rule associated with the data, and
  - decide what aggregations and other calculations may be necessary

- Access modeling is needed
  - to determine how data are to be retrieved from a data warehouse and
  - will assist in the physical definition of the warehouse by helping to define which data require indexing
  - It may also indicate whether dependent data marts are needed to facilitate information retrieval

- The team skills needed to develop and implement a data warehouse require in-depth knowledge of the data base technology and development tools utilized

- Source systems and development technology, reference the many inputs and the process used to load and maintain a data warehouse

MASSIVE DATA WAREHOUSES AND SCALABILITY

- A data warehouse needs to support scalability
- Main issues pertaining to scalability are
  - the amount of data in the warehouse
  - how quickly the warehouse is expected to grow
  - The number of concurrent users, and the complexity of user queries
- A data warehouse must scale both horizontally and vertically
  - The warehouse will grow as a function of data growth and the need to expand the warehouse to support new business functionality
  - Data growth may be caused by the addition of current cycle data and/or historical data

MASSIVE DATA WAREHOUSES AND SCALABILITY

Issue with Scalability:

- Given that the size of data warehouses is **expanding at an exponential rate, scalability is an important issue**
- **Good scalability means that queries and other data access functions will grow (ideally) linearly with the size of the warehouses**

Issues with Creating Scalable Data warehouses:

- **Scalability is difficult in managing hundreds of terabytes or more**
- **Terabytes of data have considerable inertia, occupy a lot of physical space, and require powerful computers**
- **Some firm utilize parallel processing, others use clever indexing and search schemes to manage their data**
- **Some spread their data across different physical data stores.**
- **A data warehouses approach the petabyte size, better solutions to scalability continue to be developed**

Importance of effective indexing for data warehouses:

- Correct indexing can definitely lead to efficient searches through massive amounts of data
- As a data warehouse is designed, it is important to consider correct indexing to help solve scalability problems

USERS, CAPABILITIES, AND BENEFITS

- **Analysts, managers, executives, administrative assistants, and professionals are the major end-users of data warehouses**

- **A data warehousing solution should provide ready access to critical data, insulate operation databases from ad hoc processing that can slow TPS systems, and provide high-level summary information as well as data drill down capabilities**

- These properties can
  - **improve business knowledge,**
  - **Provide competitive advantage**
  - **enhance customer service and satisfaction**
  - **facilitate decision making**
  - **improve worker productivity**
  - **help streamline business processes**

5.8 DATA MARTS

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

The high cost of data warehouses limits their use to large companies

As an alternative, many firms use a lower-cost, scaled-down version of a data warehouse referred to as an independent data mart

An independent data mart is a small warehouse designed for a strategic business unit (SBU) or a department, but its source is not an enterprise data warehouse

Advantages of data marts include the following:

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

Several types of data marts:

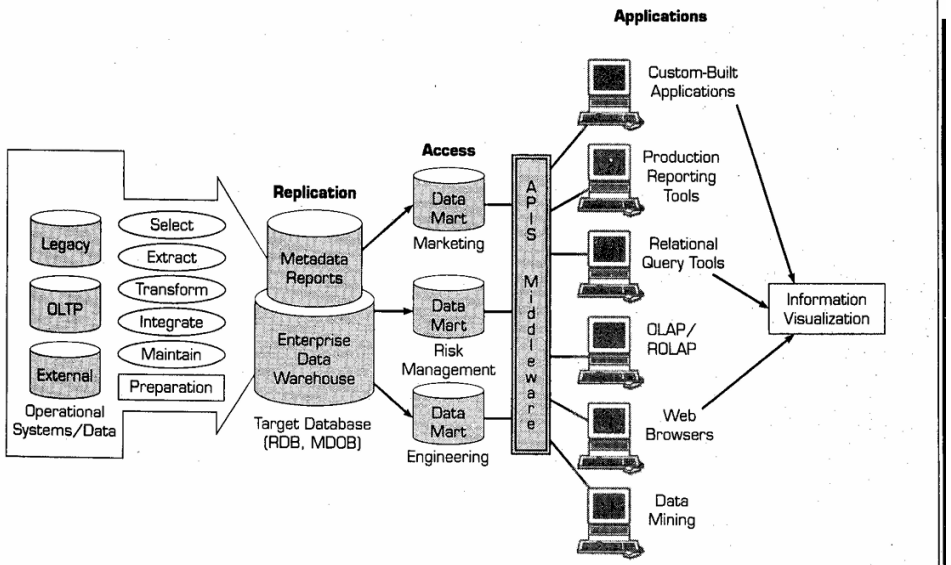
1. **Replicated (dependent) data marts**

- Sometimes it is easier to work with smaller parts of the warehouse
- In such cases one can replicate functional subsets of the data warehouse in smaller databases, each of which is dedicated to certain areas
- As shown in Figure, data mart is an addition to the data warehouse

2. **Independent data marts**

- A company can have one or more independent data marts without having a data warehouse
- In such cases there is a need to integrate the data marts
- This is possible only if each data mart is assigned a specific set of information for which it is responsible
- The IS department specifies the rules to the meta data so that the information kept by each mart is compatible with that provided by all the other marts
- When this is not done, the data marts are difficult to integrate, creating potentially fragmented problems for the organization

FIGURE 5.2 DATA WAREHOUSE FRAMEWORK AND VIEWS



5.9 BUSINESS INTELLIGENCE/BUSINESS ANALYTICS

- Business intelligence describes the basic architectural components of a business intelligence environment, ranging from traditional topics, such as business pre modeling and data modeling, to more modern topics such as business Rule system data profiling, information compliance and data quality, data Warehousing and mining

- Business intelligence involves acquiring data and information from a wide from a wide variety of sources and utilizing the decision making

- Business analytics adds an additional dimension to business intelligence i.e. Models and solution methods

Activities of Business Intelligence:

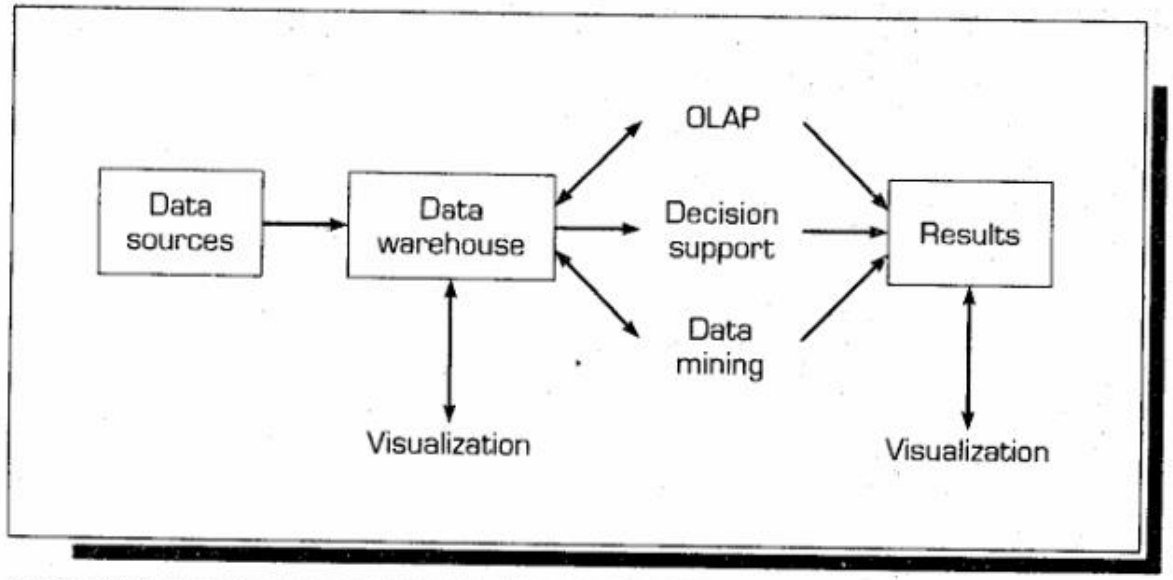


FIGURE 5.6 THE ACTIVITIES OF BUSINESS INTELLIGENCE

- BI methods and tools are highly visual in nature
  - They provide charts and graphs of multidimensional data with the click of a mouse
  - These methods generally access data from data warehouses and deposit them into a local, multidimensional database system

- Online Analytical Processing (OLAP) methods allow an analyst, or even a manager to slice and dice the data, while observing graphs and tables that reflect the dimensions being observed

- Models may be applied to the data for forecasting or to identify opportunities

- Data mining methods apply statistical and deterministic models and artificial intelligence methods to data, perhaps guided by an analyst (or manager), to identify hidden relationships or induce/discover knowledge among the various data or text elements

- Both datamining and OLAP have been used to identify e.g., white theft in organizations. They are able to identify invoices, embezzlement, customer impersonation and similar offenses.
- Patterns and anomalies become more readily identifiable
- Suspicious activities can be isolated, measured and tracked.

Difference between data mining and OLAP :

Data mining is highly visual in the ways results are displayed

Data Mining	OLAP
Data mining runs automatically.	OLAP is driven.
data mining looks for relationships with some direction from the analyst	Typically analysts run OLAP systems. They drive OLAP
Managers focus on visualization rather than application of appropriate and accurate analysis tools	Managers use more and more improved tools resulting in a trend to move BI from analyst to user
Managers understands BI and analytics methods easily	Managers may not fully understand BI methods
extracts knowledge from the data that was previously unknown, often used for prediction, classification, or anomaly detection.	focused on querying and analyzing multidimensional data
Data Mining is about finding patterns and predicting outcomes from data using advanced techniques	**OLAP** is more about querying and reporting from a structured database

DSS IN FOCUS 5.24

ARE BUSINESS INTELLIGENCE SYSTEMS MAKING FIRMS SMARTER?

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

- Most notably, the use of business intelligence technologies is high, and growing.
- Larger companies are somewhat more likely than smaller companies to use BI.
- In 2002, successful companies spent almost 50 percent more on BI technology than unsuccessful companies. BI seems to be necessary (but not sufficient) for success.
- The government utilizes virtually every market intelligence technology at significantly higher rates than any other sector of the economy.
- The technologies used to collect, aggregate, analyze, and report on competitive intelligence along with the percentage response in parentheses are: reporting tools (82.1), automated data/information feeds (79), intranets/portals (70.4), data warehousing (69.8), content management (63), data-visualization software (41.4), specialty search engines (41.4), work-flow software (41.4), and harvesting (e.g., intelligent agents) (38.9).
- Just 49 percent of less successful companies are happy with their competitive intelligence efforts.
- Some 88 percent of companies have confidence in the accuracy of the customer information they gather.
- Dissatisfaction with BI usually derives from difficulty in distributing the results.
- CIOs want to move firms to the real-time enterprise.

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

Computing systems are now an **indispensable infrastructure** with which we run, manage, and coordinate business operations

Decision makers throughout every enterprise need an IT architecture that serves their needs, rather than the other way around

Companies achieve success when they do the following:

- Make better decisions with greater speed and confidence
- Streamline operations
- Shorten product development cycles
- Maximize value from existing product lines and anticipate new opportunities
- Create better, more focused marketing as well as improved relationships with customers and suppliers

DSS IN FOCUS 5.25

TEN CRITICAL CHALLENGES FOR BUSINESS INTELLIGENCE SUCCESS

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

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

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

New forms of BI continue to emerge

Performance Management Systems (PMS) are one of the new forms

These are business intelligence tools that provide scorecards and other relevant information
With which decision makers can determine their level of success in reaching their goals

DSS IN ACTION 5.26

RETAIL MAKES STEADY BUSINESS INTELLIGENCE PROGRESS

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

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

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

Other retailers are looking for similar ways to obtain a competitive edge. Putting the right products in the right place at the right time at the right price (see revenue management in Chapter 4) is the goal of retailers. Doing it right determines who succeeds, and who fails.

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

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

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

DASHBOARDS

Dashboards provides managers with exactly the information they need in the correct format at the correct time

BI systems are the foundation of dashboards, which have evolved from **executive information systems into enterprise information systems that access data warehouses via OLAP systems**

Dashboards can impact on communications and company politics

Dashboards and scorecards measure and display what is important

Each individual, ideally can focus on what is important to him or her

Essentially a dashboard is a preset OLAP display

BI dashboards have spread to various non financial departments of firms, including sales and customer service

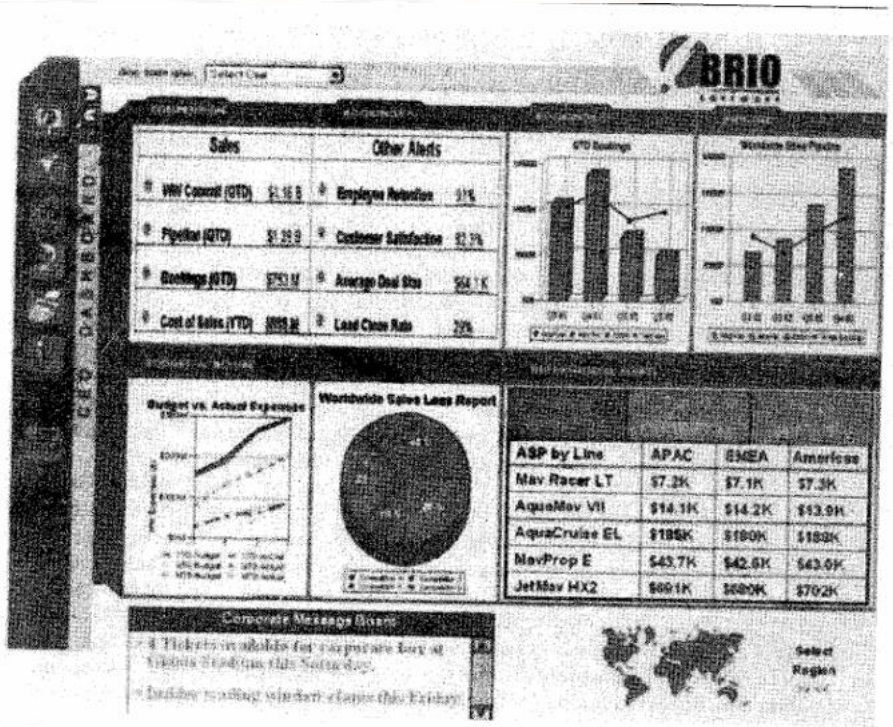


FIGURE 5.7 SCREEN SHOT FROM BIO PERFORMANCE SUITE DIGITAL DASHBOARD

BI dashboards have spread to various non financial departments of firms, including sales and customer service

Table details how dashboards have spread through the organization

TABLE 5.4 Departments with Dashboards

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

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

DSS IN FOCUS 5.27

BUSINESS INTELLIGENCE ASSESSMENT

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

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

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

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

DSS IN FOCUS 5.28

CRITICAL LESSONS IN BUSINESS INTELLIGENCE AND DATA WAREHOUSING

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

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

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

- Web has had a profound impact on how these tools function and what they are utilized for
- The visual nature of most BI tools is often based on Web browser interfaces
- As web use and e commerce increase, there is more of a

demand for gathering and analysing data from the clickstream to Identify where customers go on a web site, where they came from, where they came from, where they go afterwards, and what they buy or don't buy

TABLE 5.5 Database and Business Intelligence Technologies, and Web Impacts

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

5.10 ONLINE ANALYTICAL PROCESSING (OLAP)

- For corporate transaction processing at critical systems etc, **the systems must be virtually fault tolerant and provide rapid response**

- An effective solution was provided by **online transaction processing (OLTP)**

- OLTP centres on a **distributed relational database environment**

- The latest developments in this area are the utilization of ERP and SCM software for
  - **transaction processing tasks,**
  - **CRM applications**
  - **integration with web based technologies and intranets**

- Access to data is often needed by both OLTP and MSS application

- But, trying to use both at a time to serve request may cause problem
  - So, companies elect to separate databases into OLTP types and OLAP instead of directly utilizing pre, multidimensional data cubes

- **The database must be integrated with the centralized, cohesive and consistent control of multidimensional data across the enterprise**

- **To make database aware of the higher level objects that relate directly to OLAP and business models**.
- In effect, these objects will take the existing atomic entities and compound them to make dimensional entities such as attributes, facts, relationships, hierarchies, and dimensions.
- Once these high level objects are defined, the new information can be stored and managed as part of the catalogs.
- **In effect, managing metadata becomes part of the relational database management system in order to make it "OLAP aware".**

- The term OLAP refers to variety of activities usually performed **by end-users in online systems.**
- **Real strength of OLAP is in its analytic capabilities**

- There is no agreement on what activities are considered OLAP

- Probable Activities are:
  - **Generating queries**
  - **Requesting ad hoc reports and graphs**
  - **Conducting statistical analyses**
  - **Building DSS and multimedia applications**
  - **Executive and/or enterprise information systems and data mining**

- Essentially, **OLAP provides modeling and visualization capabilities to large data sets, either to data base management systems or more often , data warehouse systems**

- OLAP is different from data mining in that users can ask specific, open-ended questions

- Users, typically analysts run OLAP systems. They drive OLAP, where as data mining looks for relationships with some direction from the analyst

- OLAP is generally facilitated by working with data warehouse and with a set of OLAP tools.

OLAP tools can be
  - query tools,
  - spreadsheets
  - data mining tools
  - data visualization tools etc

An example for framework – **IBM's OLAP server** to analyze large amount of data to detect fraudulent claims and speed up the processing of claims
- It takes only a couple of days to analyze data that previously took several weeks . As the tools and hardware improve, claims can be analyzed instantaneously. Cost of accessing claims is greatly reduced.

SQL FOR QUERYING

- **Structured query language (SQL) is a standard data language for data access and manipulation in relational database management systems**

- It is an **English like language consisting of several layers of increasing complexity and capability**

- SQL is used for
  - **online access to databases,**
  - **DBMS operations from programs and**
  - **database administration functions.**
  - **Data access and manipulation function of some leading DBMA software products such as Oracle, IBMS;s DB2, Ingres etc**

- Since SQL is non procedural and **fairly user friendly, many end users can use it for their own queries and database operations**

- **SQL can be employed for programs written in any standard programming language thus, it facilitates software integration**

- Support of DSS/BI is accomplished in the warehouse with products from vendors such as **Brio, Business Objects, Cognos, Pilot Software, and SAS**

- SQL is a fairly simplistic OLAP tool.

OLAP TOOLS

- **Using SQL and other conventional data access and analysis tools is helpful, But not sufficient for OLAP**

- In OLAP, a special class of tools is used, known as
  - **decision support/business intelligence/Business analytic front ends,**
  - **data-access front ends, database front ends and**
  - **visual information across systems**

    - These methods go well **beyond spreadsheets in power and results**

    - These tools are intended to empower users

- **OLAP tools have characteristics that distinguish them from reporting tools designed to support traditional OLTP reporting applications**

OLAP TOOLS

The characteristics of OLAP tools were succinctly defined by E.F. Codd and Associates. They defined for types of processing that are performed by analysts within an organization:

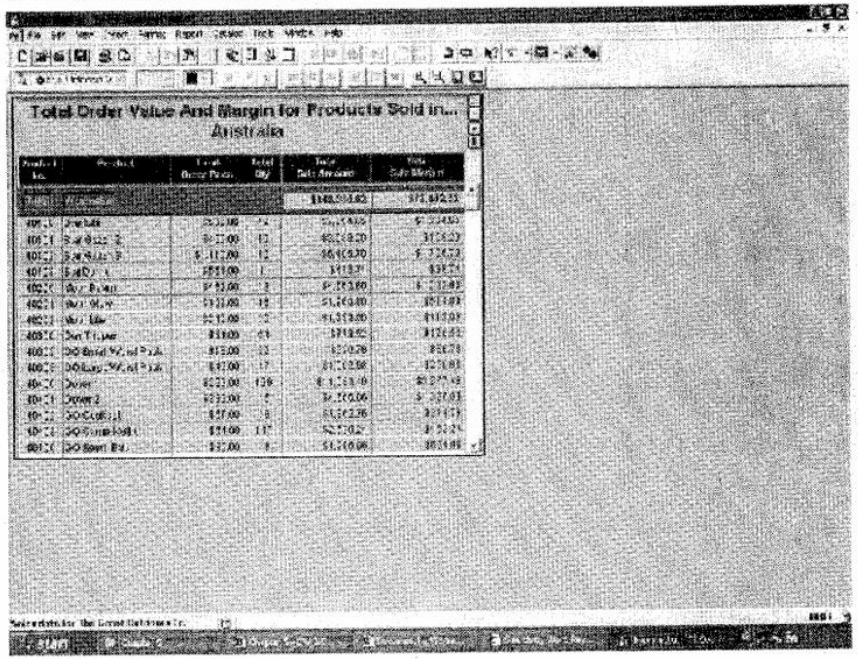
1. Categorical analysis is a static analysis based upon historical data. It is based upon the premise that past performance is an indicator of the future. This is the primary analysis supported by OLTP transaction-based databases.
2. Exegetical analysis is also based upon historical data, adding the ability to perform drill down analysis. Drill down analysis is the ability to query further into the data to determine the detail data that were used to determine a derived value.
3. Contemplative analysis allows a user to change a single value to its impact.
4. Formulaic analysis permits changes to multiple variables.

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

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

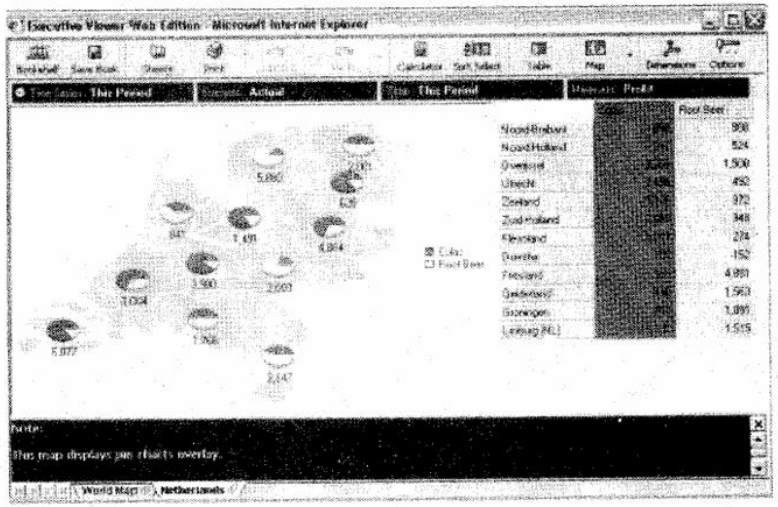
Source: Adapted from "Providing OLAP to User-Analysts: An IT Mandate," Codd & Associates, White Paper, hyperion.com. Also see Radin (1997).

FIGURE 5.8 COGNOS IMPROMPTU SAMPLE OUTPUT



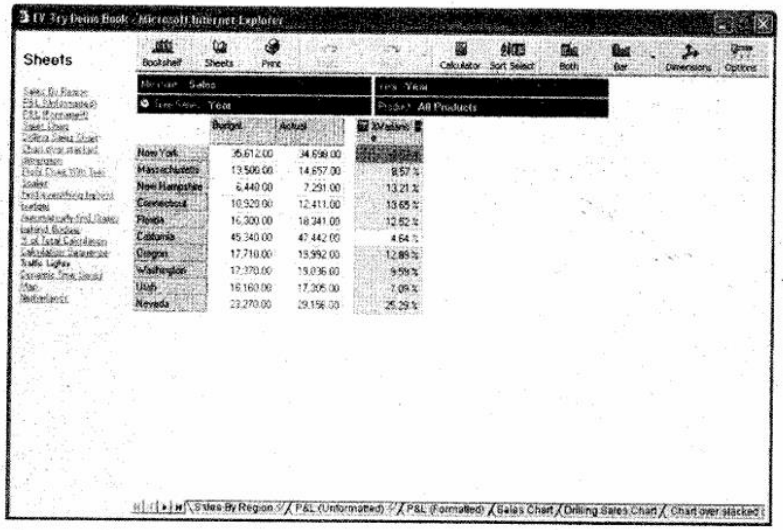
Courtesy of Cognos, Inc.

FIGURE 5.9 TEMTEC EXECUTIVE VIEWER ADVANCED MAP DATA DISPLAY



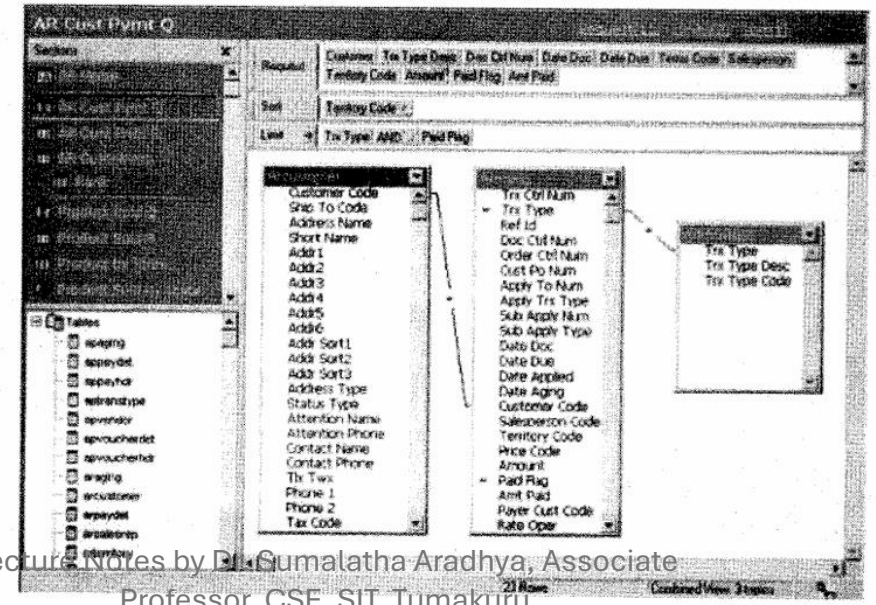
Courtesy of Temtec Inc.

FIGURE 5.10 TEMTEC EXECUTIVE VIEWER TRAFFIC LIGHT DISPLAY



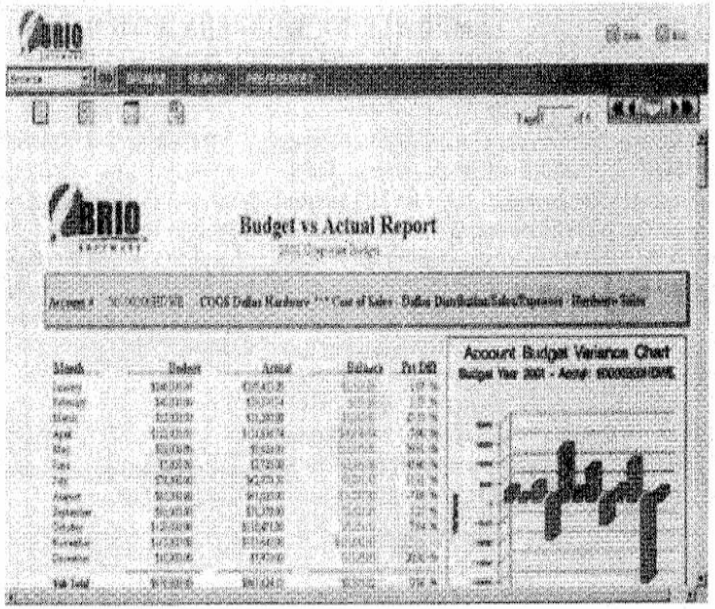
Courtesy of Temtec Inc.

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



Courtesy of Brio Software

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



Courtesy of Brio Software Inc.

5.11 DATA MINING

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

- **Datamining is a process that uses Statistical, Mathematical, Artificial intelligence and Machine learning techniques to extract and Identify useful information and subsequent knowledge from large Databases**

- Formerly, the term was used to describe the process through which undiscovered patterns in data were identified
- Later, many types of data analysis was included.

As per Gartner group,

- **Data mining is the process of engineering mathematical patterns from usually large set of data**.
- **These patterns can be rules, affinities, correlations, trends or prediction models**
- **Datamining is on the interface of computer science and statistics, utilizing advances in both disciplines to make progress in extracting information from large databases**

Data mining includes tasks known as **knowledge extraction, data archaeology, Data exploration, data pattern processing, data dredging, and information harvesting**

All these activities are conducted automatically and allow quick discovery even by non programmers

Following are the major characteristics and objectives of data mining:

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

HOW DATA MINING WORKS

- **Intelligent data mining, discovers information within data warehouses that queries and reports cannot effectively reveal**

- **Datamining tools find patterns in data and may even infer rules from them**

- **3 types of methods are used to identify pattern in data:**
  - Simple models (SQL based query, OLAP, human judgement)
  - Intermediate models (regression, decision trees, clustering)
  - Complex models (neural networks, other rule induction)

  These patterns and rules can be used to guide decision-making and forecast the effect of decisions

- **Data mining can speed analysis of focusing attention on the most important variables**

- Each data mining application class is supported by a set of algorithmic approaches to extract the relevant relationships in the data. These approaches differ in the classes of problems they are able to solve.

- The classes are:
  - Classification
  - Clustering
  - Association
  - Sequencing
  - Regression
  - Forecasting

Classes explained:

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

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

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

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

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

Data mining can be either **hypothesis driven or discovery driven**

Hypothesis driven:

Begins with proposition by the user, who then seeks to validate the truthfulness of the proposition

e.g.,

marketing manager may begin with the proposition, “Are DVD players sales related to dales of television sets?”

Discovery driven:

Finds patterns, associations, and relationships among the data

It can uncover facts that were previously unknown or not even contemplated by an organization

As per Buck, taxonomy of the classes of datamining tools and techniques as they relate to information and BI can be

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

- In data mining, the scalability of the methods and of the data warehouse are critical issues
- This is so because of the amount of data and searching required
- Seven steps are necessary for successful data mining

DSS IN FOCUS 5.34

THE SEVEN STEPS OF DATA MINING

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

1. Define the business problem.
2. Build (find or acquire) the data-mining database.
3. Explore the data.
4. Prepare the data for modeling.
5. Build (or find) the models.
6. Evaluate the models.
7. Act on the results.

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

- Data mining is iterative because data miners make mistakes because they often do not understand the process but do understand the expected results
- Actually, It is the process of discovery that is iterative
- Data mining is an experimental process that requires sound experimental design

DSS IN FOCUS 5.35

DATA MINING MYTHS

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

- ***Data mining provides instant, crystal-ball predictions.*** Data mining is a multi-step process that requires deliberate, proactive design and use.
- ***Data mining is not yet viable for business applications.*** The current state-of-the-art is ready to go for almost any business.
- ***Data mining requires a separate, dedicated database.*** Because of advances in database technology, a dedicated database is not required, even though it may be desirable.
- ***Only Ph.D.s can do data mining.*** Newer Web-based tools make data mining by managers possible.
- ***Data mining is only for large firms with lots of customer data.*** If the data accurately reflect the business or its customers, a company can utilize data mining.

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

DSS IN FOCUS 5.36

DATA MINING BLUNDERS

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

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

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

DATA MINING TOOLS AND TECHNIQUES

There are many methods for performing data mining

Data mining tools and techniques can be classified based upon the structure of the data and the algorithms uses. The main ones are:

- Statistical Methods
- Decision Tress
- Case based Reasoning
- Neural computing
- Intelligent agents
- Genetic algorithms
- Other tools

DATA MINING TOOLS AND TECHNIQUES

- ***Statistical methods.*** These include linear and nonlinear regression, point estimation, Bayes's theorem (probability distribution), correlations, and cluster analysis.
- ***Decision trees.*** Decision trees are used in classification and clustering methods. Decision trees break problems down into increasingly discrete subsets, by working from generalizations to increasingly more specific information. A decision tree can be defined as a root followed by internal nodes. Each node (including the root) is labeled with a question. The arcs associated with each node cover all possible responses. Each response represents a probable outcome
- ***Case-based reasoning.*** Using historical cases, the case-based reasoning approach can be used to recognize patterns. For example, customers of Cognitive Systems Inc. use such an approach for help desk applications. One customer has a 50,000-query case library. New cases can be matched quickly against the 50,000 samples in the library, providing automatic answers to queries with more than 90 percent accuracy.
- ***Neural computing.*** Neural networks utilize many connected nodes (which operate in a manner similar to how the neurons of the human brain function). This approach examines a massive amount of historical data for patterns. Thus, one can go through large databases and, for example, identify potential customers for a new product or companies whose profiles suggest that they are heading for bankruptcy. Many applications are in financial services and in manufacturing.
- ***Intelligent agents.*** One of the most promising approaches to retrieving information from databases, especially external ones, is the use of intelligent agents. With the availability of a vast and growing amount of information through the Internet, finding the right information is becoming more difficult. Web-based data mining applications are typically enabled by intelligent software agents.
- ***Genetic algorithms.*** Genetic algorithms work on the principle of expansion of possible outcomes. Given a fixed number of possible outcomes, genetic algorithms seek to define new and better solutions. Genetic algorithms are used for clustering and association rules.
- ***Other tools.*** Several other tools can be used. These include rule induction and data visualization. The best source of new tool development is vendor Web sites.

- Data mining algorithms are important
- When dealing with customers behavioral data, which can encompass a hundred dimensions or more, algorithms should be c**apable of dealing effectively with high-dimensional data**

- These algorithms must also be able to work with business constraints and rules

- Simple statistics do not work

- Knowledge of the business constraints, of the relations between products, and of the various behavioural segments of customers is a must

TEXT MINING

- Text mining is the application of data mining to non structured or less structured text files

- Data mining takes advantage of the infrastructure of stored data to extract additional useful information

  e.g., by data mining a customer database, an analyst might discover that everyone who buys product A also buys products B and C, but after 6 months.

- Text mining operates with less structured information
- Documents rarely have a strong internal infrastructure and when they do, it is frequently focuses on document format rather than document content

Text mining helps organizations to

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

Text mining helps to predict expected claims and understand why outcomes deviate from the predictions

Text mining is used to extract entities and objects for frequency analysis, identify files with certain attributes for further statistical analysis, and create entirely new data captures for predictive modeling

Some popular text mining tools and vendors are:

- SAS Text Miner (www.sas.com)
- IBM Intelligent Miner for Text (www.ibm.com)
- SPSSLexiquest (www.spss.com)
- Insightful Miner for Text (www.insightful.com)
- Megaputer Intelligence TextAnalyst (www.megaputer.com)

DSS IN FOCUS 5.39

HOW TO MINE TEXT

Term extraction is the most basic form of text mining. Like all text mining techniques, it maps information from unstructured data into a structured format. The simplest data structure in text mining is the feature vector, or weighted list of words. The most important words in a text are listed, along with a measure of their relative importance. Text reduces to a list of terms and weights. The entire semantics of the text may not be present, but the key concepts are identified. To do this, text mining performs the following:

1. Eliminate commonly used words (the, and, other).
2. Replace words with their stems or roots (e.g., eliminate plurals, and various conjugations and declensions). Thus the terms "phoned," "phoning," and "phones" are mapped to "phone."
3. Calculate the weights of the remaining terms. The most common method is to calculate the frequency with which the word appears. There are two common measures: the term frequency, or *tf factor*, measures the actual number of times a word appears in a document, while the inverse document frequency, or *idf factor*, indicates the number of times the word appears in all documents in a set. The reasoning is that a large *tf factor* increases the weight, while a large *idf factor* decreases it, because terms that occur frequently in all documents would be common words to the industry and not be considered important.

For example, consider the first paragraph of this DSS in Focus box up to the colon. There were some 20 terms with 28 occurrences once we factored out common words. Here is a list of terms that appeared more than once, along with their relative frequencies (tf factors) out of a total of 28:

Term	*Term Factor*
data	.0714
structure	.0714
term	.0714
text	.0714
text mining	.1429
weight	.0714

When you consider all the important words in the paragraph, they comprise one-half of its total importance and could be used to identify its semantics. Clearly the paragraph is about text mining (weight = 0.1429) and involves text and data with structure and weight.

Source: Adapted partly from Martin Ellingsworth and Dan Sullivan, "Text Mining Improves Business Intelligence and Predictive Modeling in Insurance," *DM Review*, Vol. 13, No. 7, July 2003, pp. 42–44.

DSS IN ACTION 5.40

TEXT MINING

Text mining is a very effective approach to automatically performing analysis on standard and Web documents. For example, an international pharmaceutical firm used text mining to evaluate 500 text-based responses from patients participating in a clinical study of a new allergy medication. Text mining software detected a cluster of 50 patients who used specific words that described negative side effects. Further examination indicated that these patients all received a high dosage of the drug, and that women older than 40 were especially sensitive to the high dosage. Consequently, dosage levels are adjusted, and warnings to women over 40 are included with the medicine.

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

DSS IN ACTION 5.41

DATA MINING AT PFIZER

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

The Pfizer system searches the database of documents and extracts a set of documents characterized by simple search criteria based on a combination of keywords. Next, the set of documents is further segmented into topics. Topics are characterized by lists of keywords extracted from the free-format text contained in the documents. The scientists choose topics of interest by examining keyword lists. Pfizer has realized several benefits. First, the company has discovered that text mining is not only a technology for the categorization of information. The results of text mining also permit the building of new applications for further navigation of data and decision support. These new applications can take a prototype to complete development much faster than ever before. It is now possible to rapidly assemble powerful, easy-to-use analytical applications that address the full gamut of requirements.

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

SAMPLER OF DATA MINING APPLICATIONS

Data mining can be very helpful as shown by the following representative examples:

- ***Marketing:*** predicting which customers will respond to Internet banners or buy a particular product; segmenting customer demographics.
- ***Banking:*** forecasting levels of bad loans and fraudulent credit card usage, credit card spending by new customers, and which kinds of customers will best respond to new loan offers or other products and services.
- ***Retailing and sales:*** predicting sales and determining correct inventory levels and distribution schedules among outlets.
- ***Manufacturing and production:*** predicting when to expect machinery failures, finding key factors that control the optimization of manufacturing capacity.
- ***Brokerage and securities trading:*** predicting when bond prices will change, forecasting the range of stock fluctuation for particular issues and the overall market; determining when to trade stocks.
- ***Insurance:*** forecasting claim amounts and medical coverage costs, classifying the most important elements that affect medical coverage, predicting which customers will buy new policies with special features.
- ***Computer hardware and software:*** predicting disk drive failure, forecasting how long it will take to create new chips, predicting potential security violations.
- ***Government and defense:*** forecasting the cost of moving military equipment, testing strategies for military engagements, predicting resource consumption.
- ***Airlines:*** capturing data not only on where customers are flying but also the ultimate destination of passengers who change carriers in mid-flight. With this information airlines can identify popular locations they are not currently serving so as to add routes and capture lost business.
- ***Health care:*** correlating demographics of patients with critical illnesses; using data mining, doctors can develop better insights on symptoms and how to provide proper treatments.
- ***Broadcasting:*** predicting what programs are best shown during prime time and how to maximize returns by inserting advertisements.
- ***Police:*** tracking crime patterns, locations, criminal behavior, and attributes to help solve criminal cases

KDD AND DATA MINING

- Data mining and knowledge discovery in databases (KDD) are frequently used as synonyms
- KDD is defined as a process of using data mining methods to find useful information and patterns in the data, Where as data mining is the use of algorithms to identify patterns in data derived by the KDD process
- KDD is a comprehensive process that encompasses data mining
- The input to the KDD process consists of organizational data
- The enterprise data warehouse enables KDD to be implemented efficiently because it provides a single source for data to be mined

Dunham summarizes the KDD process as consisting of following steps:

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

DSS IN FOCUS 5.44

DATA MINING TO IDENTIFY CUSTOMER BEHAVIOR

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

- Demographic data
- Transaction data
- Online interaction data

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

Retail uses of data mining evolve as:

Step 1: *Web analytics.* Gather Web site statistics that track customers' online behavior: hits, pages, sales volume, etc. This helps adjust a Web site to meet customer needs.

Step 2: *Customer analytics.* These add depth to understanding customer interactions. Firms gather data from multiple sources, including Web site interactions, transaction data from offline purchases, and demographic data. This is critical in CRM and revenue management in that a better understanding allows an organization to cluster customers into groupings.

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

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

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

INTELLIGENT DATA MINING AND TEXT MINING

- New intelligent data and text mining methods, based on AI methods like ANN and Intelligent agents, continue to be developed and applied in practice
- These methods often prove to be very effective on specific kinds of problems and sets of data and text
- Many are applied to identifying information and knowledge on Web pages scattered around the world
- Data mining tools can effectively detect patterns in data e.g., in financial transactions, e-commerce for fraud detection
- A team of Norwegian biologists developed intelligent methods to search and mine the web for genetic studies that contain information relevant to their endeavors
- Intelligent agents can be used for intelligent mining e,g., scientific applications or smooth running of businesses etc

DSS IN FOCUS 5.45

NEW INTELLIGENT METHODS TO MINE DATA

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

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

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

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

What's the Answer? (Verity Inc.). The smart software puts human learning (rules) into the search software, enabling it to learn through logistic-regression classification. For example, instead of responding with a list of Web sites, a search engine could simply scan through several of them and answer the question that is posed (e.g., "What is the population of the world?").

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

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

DATA MINING SOFTWARE

Data mining software features more complicated algorithms for Neural networking, clustering, segmentation, and classifications that are generally more sophisticated

Many software vendors provide powerful data mining tools

Angoos Knowledge Engineering
Cognos
Cytel Statistical software
Data mind corporation
IBM
PolyAnalyst
DB2

5.12 DATA VISUALIZATION, MULTIDIMENSIONALITY, AND REAL-TIME ANALYTICS

DATA VISUALIZATION

- Data visualization refers to technologies that support visualization and sometimes interpretation of data and information at several points along with data processing chain

- It includes digital images, geographic information systems, graphical user interfaces, multidimensions, tables and graphs, virtual reality, Three dimensional presentations and animation

- Visual tools can help identify relationships directly
- The ability to identify important trends in corporate and market data provides enormous advantages

- More accurate predictive models provide significant business advantages in applications that drive content, transactions, or processes

- Confident action based on superior methods of visual data analysis, helps companies improve income and avoid costly mistakes

- Data visualization enables OLAP and data mining, especially utilizing web based tools
- Rather than having to wait for reports or compare sterile columns of numbers, a manager can utilize a browser interface in real time to look at vital organizational performance data

- By using visual analysis technologies, managers, engineers, and other professional have spotted problems that for years went undetected by standard analysis methods

- Visual technologies can be integrated to create different information presentations, especially with VR methods
- Data visualization enables problem solving methods in addition to providing graphic features to OLAP and data Mining tools
- Data visualization is easier to implement when the necessary data are in data warehouse, or better yet in a Multidimensional server

MULTIDIMENSIONALITY

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

MULTIDIMENSIONAL PRESENTATION

- Summary data can be organized in different ways for analysis and presentation
- An efficient way to do this is called multidimensionality
- The major advantage of multidimensionality is that data can be organized the way managers rather than system analysis like to see them.
- Different presentations of the same data can be arranged easily and quickly
- Underlying every OLAP system is a conceptual data model often referred to as the multidimensional data model or multidimensional modeling
- This technique helps conceptualize business models as a set of measures described by ordinary facets of business.
- The method is particularly useful for sifting, summarizing, and arranging data to facilitate analysis
- In contrast to the techniques for designing online transaction processing systems, which rely on entities, relationships, functional decomposition, and state transition analysis, MDM utilizes the constructs of facts, dimensions, hierarchies and sparsity

Three factors are considered in multidimensionality:

1. Dimensions
2. Measures
3. Time

Some examples:

Dimensions: products, salesperson, market segments, business units, geographic locations, distribution channels, countries, industries

Measures: money, sales volume, head count, inventory profit, actual vs forecasted

Time: daily, weekly, monthly, quarterly, yearly

Manager may want to know the sales of a product in a certain geographic area, by a specific salesperson, during a specified month, or in terms of units.

One can find the answers - >
If the data are organized in multidimensional databases?
or
if the query or related software products are designated for multidimensionaly ?

Multidimensionality has limitations:

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

REAL-TIME ANALYTICS

- A recent research study shows that humans will record more information in the next 3 years than since the dawn of civilization
- We need specialized methods to store our information in many formats, and to quickly retrieve and exploit it
- Business users increasingly demand access to real time, unstructured, or remote data, integrated with the contents of their data warehouse
- Data warehousing and BI tools traditionally focus on assisting managers in making strategic and tactical decisions

- When a process that require instantaneous updates are necessary for answering analytical questions, a real time response is necessary
- Query, OLAP, and data mining response times must be close to zero

- Real time data warehouses are updated on a regular basis, not just weekly or monthly

- With business analytic application, one can instantaneously identify customer buying patterns based on store displays, and recommend immediate changes to placement or the display itself

- Other applications include call-centre support, fraud detection, revenue management, transportation and many financial transactions

- An important issue in real time computing is that not all should be updated continuously. This may cause problems when reports are generated in real-time, because one person's results may not match another person's

- Real time requirements change the way we view the design of databases, data warehouses, OLAP and data mining tools, since they are literally updated concurrently while queries are active. On the other hand, the substantial business value in doing so has been demonstrated, so it is crucial that organizations adopt these methods in their business processes

- Examples of web-based, real time BI software include:
  - BusinessObjects WebIntelligence
  - Cognos Supply chain analytics and BI series
  - DatMirror Livebusiness
  - IBM DB2 etc..

5.14 BUSINESS INTELLIGENCE AND THE WEB: WEB INTELLIGENCE/WEB ANALYTICS

- BI activities
  1. acquisition
  2. warehousing
  3. mining

- These can be performed using Web tools or interrelated web technologies and electronic commerce

- BI tools can be used to analyze Web site performance in real time.

- Electronic commerce software vendors are providing web tools that connect the data warehouse with the e-commerce ordering and cataloguing systems

- Data warehousing and decision support vendors are integrating their product with web technologies and e commerce, or creating new ones for the same purpose

- Data marts continue to become much more popular in the web environment

- Exercise: List trending BI tools, Data marts

WEB ANALYTICS/WEB INTELLIGENCE

- Web analytics and web intelligence are the terms used to describe the application of business analytics/business intelligence to web sites
- The tools and methods are highly visual in nature
- Exercise:
  Identify and list web analytics and web intelligence tools and methods

THE CHALLENGES OF CLICKSTREAM ANALYSIS

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

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

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

DSS IN FOCUS 5.59

WEB SERVICES BUSINESS INTELLIGENCE TOOLS

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

Actuate Corp. (www.actuate.com): Actuate 6

Business Objects (www.businessobjects.com): Business Objects, WebIntelligence, BusinessObjects Developer Suite

ClearForest Corp. (www.clearforest.com): ClearResearch, ClearEvents, ClearSight

Cognos, Inc. (www.cognos.com): Cognos Series 7, Cognos Web Services SDK

Crystal Decisions (www.crystaldecisions.com): Crystal Enterprise, Crystal Reports, Crystal Analysis Professional

Dimensional Insight, Inc. (www.dimins.com): DI-Diver, DI-ProDiver, DI-WebDiver, DI-ReportDiver, DI-Broadcast

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

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

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

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

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

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

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