Data Management Concerns

Building Blocks of Data Management:

- Data profiling – understanding the data.
- Data Quality Management – improving the quality of data.
- Data Integration – combining similar data from multiple sources.
- Data augmentation – improving the value of the data.

Data Management

IT applications cannot be done without using some kind of data
Which are at the core of management and marketing operations. However, managing data is difficult for various reasons.

- The amount of data increases exponentially with time.
- Data are scattered throughout organizations.
- Data are collected by many individuals using several methods.
- External data needs to be considered in making organizational decisions.
- Data security, quality, and integrity are critical.
- Selecting data management tools can be a major problem.

Data are an asset, when converted to information and knowledge, give the firm competitive advantages.

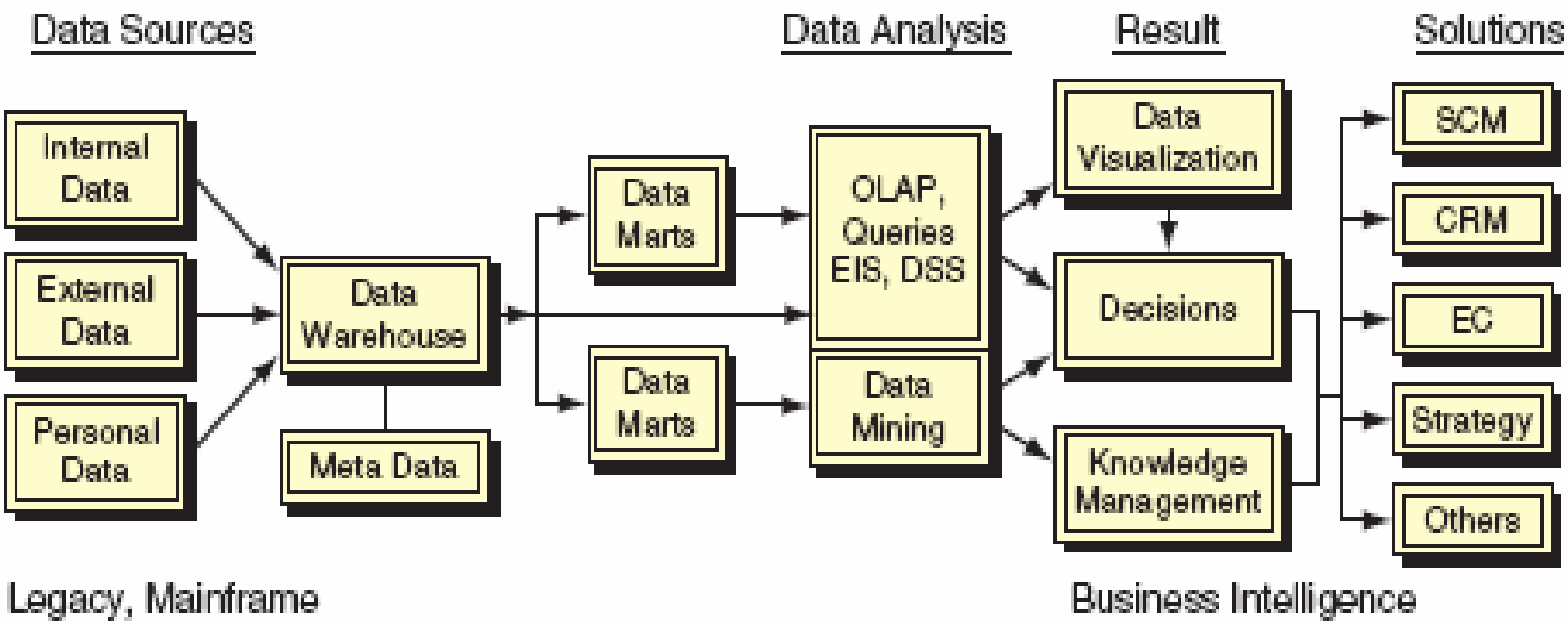
Data Life Cycle Process

Businesses run on data that have been processed to information and knowledge, which mangers apply to businesses problems and opportunities. This transformation of data into knowledge and solutions is accomplished in several ways.

1. New data collection occurs from various sources.
2. It is temporarily stored in a database then preprocessed to fit the format of the organizations data warehouse or data marts
3. Users then access the warehouse or data mart and take a copy of the needed data for analysis.
4. Analysis (looking for patterns) is done with
   - Data analysis tools
   - Data mining tools

The result of all these activities is the generating of decision support and knowledge

Data Life Cycle Process Continued



The result - generating knowledge

Data Sources

The data life cycle begins with the acquisition of data from data sources. These sources can be classified as internal, personal, and external.

- Internal Data Sources are usually stored in the corporate database and are about people, products, services, and processes.
- Personal Data is documentation on the expertise of corporate employees usually maintained by the employee. It can take the form of:
  - estimates of sales
  - opinions about competitors
  - business rules
  - Procedures
  - Etc.
- External Data Sources range from commercial databases to Government reports.
- Internet and Commercial Database Services are accessible through the Internet.

Methods for Collecting Raw Data

The task of data collection is fairly complex. Which can create data-quality problem requiring validation and cleansing of data.

- Collection can take place
  - In the field
  - From individuals
  - Via manually methods
    - Time studies
    - Surveys
    - Observations
    - Contributions from experts
  - Using instruments and sensors
  - Transaction processing systems (TPS)
  - Via electronic transfer
  - From a web site (clickstream)

Methods for managing data collection

One way to improve data collection from multiple external sources is to use a data flow manager (DFM), which takes information from external sources and puts it where it is needed, when it is needed, in a usable form.

- DFM consists of
  - a decision support system
  - a central data request processor
  - a data integrity component
  - links to external data suppliers
  - the processes used by the external data suppliers.

Data Quality and Integrity

Data quality (DQ) is an extremely important issue since quality determines the data's usefulness as well as the quality of the decisions based on the data. Data integrity means that data must be accurate, accessible, and up-to-date.

- Intrinsic DQ: Accuracy, objectivity, believability, and reputation.
- Accessibility DQ: Accessibility and access security.
- Contextual DQ: Relevancy, value added, timeliness, completeness, amount of data.
- Representation DQ: Interpretability, ease of understanding, concise representation, consistent representation.

Data quality is the cornerstone of effective business intelligence.

Document Management

Document management is the automated control of electronic documents, page images, spreadsheets, word processing documents, and other complex documents through their entire life cycle within an organization, from initial creation to final archiving.

- Maintaining paper documents, requires that:
  - Everyone have the current version
  - An update schedule be determined
  - Security be provided for the document
  - The documents be distributed to the appropriate individuals in a timely manner

Transactional vs. Analytical Data Processing

Transactional processing takes place in operational systems (TPS) that provide the organization with the capability to perform business transactions and produce transaction reports. The data are organized mainly in a *hierarchical structure* and are centrally processed. This is done primarily for fast and efficient processing of routine, repetitive data.

A supplementary activity to transaction processing is called **analytical processing,** which involves the analysis of accumulated data. Analytical processing, sometimes referred to as *business intelligence*, includes data mining, decision support systems (DSS), querying, and other analysis activities. These analyses place strategic information in the hands of decision makers to enhance productivity and make better decisions, leading to greater competitive advantage.

The Data Warehouse

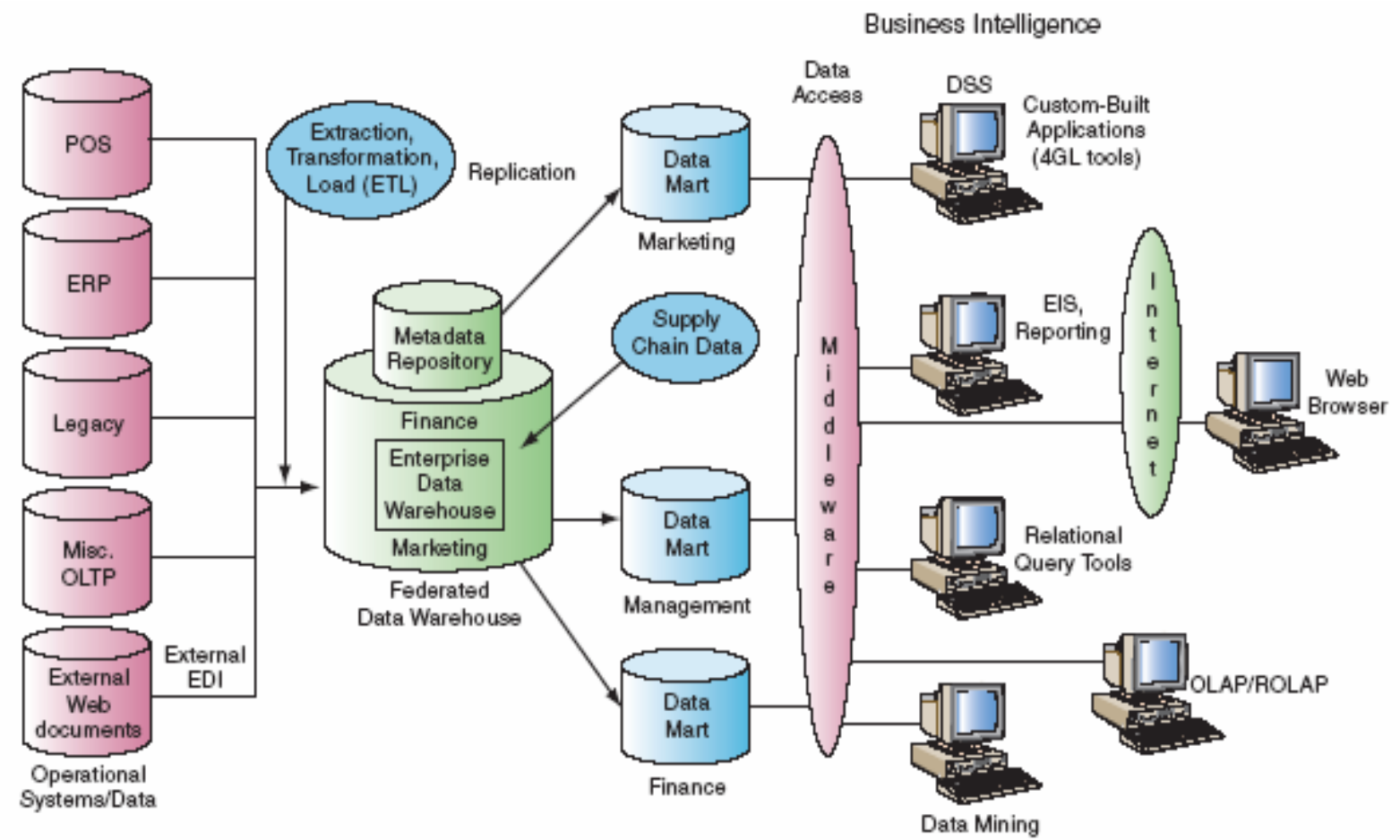
A data warehouse is a repository of subject-oriented historical data that is organized to be accessible in a form readily acceptable for analytical processing activities (*such as data mining, decision support, querying, and other applications*).

- Benefits of a data warehouse are:
  - The ability to reach data quickly, since they are located in one place
  - The ability to reach data easily and frequently by end users with Web browsers.
- Characteristics of data warehousing are:
  - Organization. Data are organized by subject
  - Consistency. In the warehouse data will be coded in a consistent manner.

The Data Warehouse Continued

- *Characteristics of data warehousing are:*
  - Time variant. The data are kept for many years so they can be used for trends, forecasting, and comparisons over time.
  - Nonvolatile. Once entered into the warehouse, data are not updated.
  - Relational. Typically the data warehouse uses a relational structure.
  - Client/server. The data warehouse uses the client/server architecture mainly to provide the end user an easy access to its data.
  - Web-based. Data warehouses are designed to provide an efficient computing environment for Web-based applications

The Data Warehouse Continued



The Data Mart

A data mart is a small scaled-down version of a data warehouse designed for a strategic business unit (SBU) or a department. Since they contain less information than the data warehouse they provide more rapid response and are more easily navigated than enterprise-wide data warehouses.

- There are two major types of data marts:
  - Replicated (dependent) data marts are small subsets of the data warehouse. In such cases one replicates some subset of the data warehouse into smaller data marts, each of which is dedicated to a certain functional area.
  - Stand-alone data marts. A company can have one or more independent data marts without having a data warehouse. Typical data marts are for marketing, finance, and engineering applications.

The Data Cube

Multidimensional databases *(sometimes called OLAP)* are specialized data stores that organize facts by dimensions, such as geographical region, product line, salesperson, time. The data in these databases are usually preprocessed and stored in *data cubes.*

- One intersection might be the quantities of a product sold by specific retail locations during certain time periods.
- Another matrix might be Sales volume by department, by day, by month, by year for a specific region
- Cubes provide faster:
  - Queries
  - Slices and Dices of the information
  - Rollups
  - Drill Downs

Operational Data Stores

Operational data store is a database for transaction processing systems that uses data warehouse concepts to provide clean data to the TPS. It brings the concepts and benefits of a data warehouse to the operational portions of the business.

- It is typically used for short-term decisions that require time sensitive data analysis
- It logically falls between the operational data in legacy systems and the data warehouse.
- It provides detail as opposed to summary data.
- It is optimized for frequent access
- It provides faster response times.

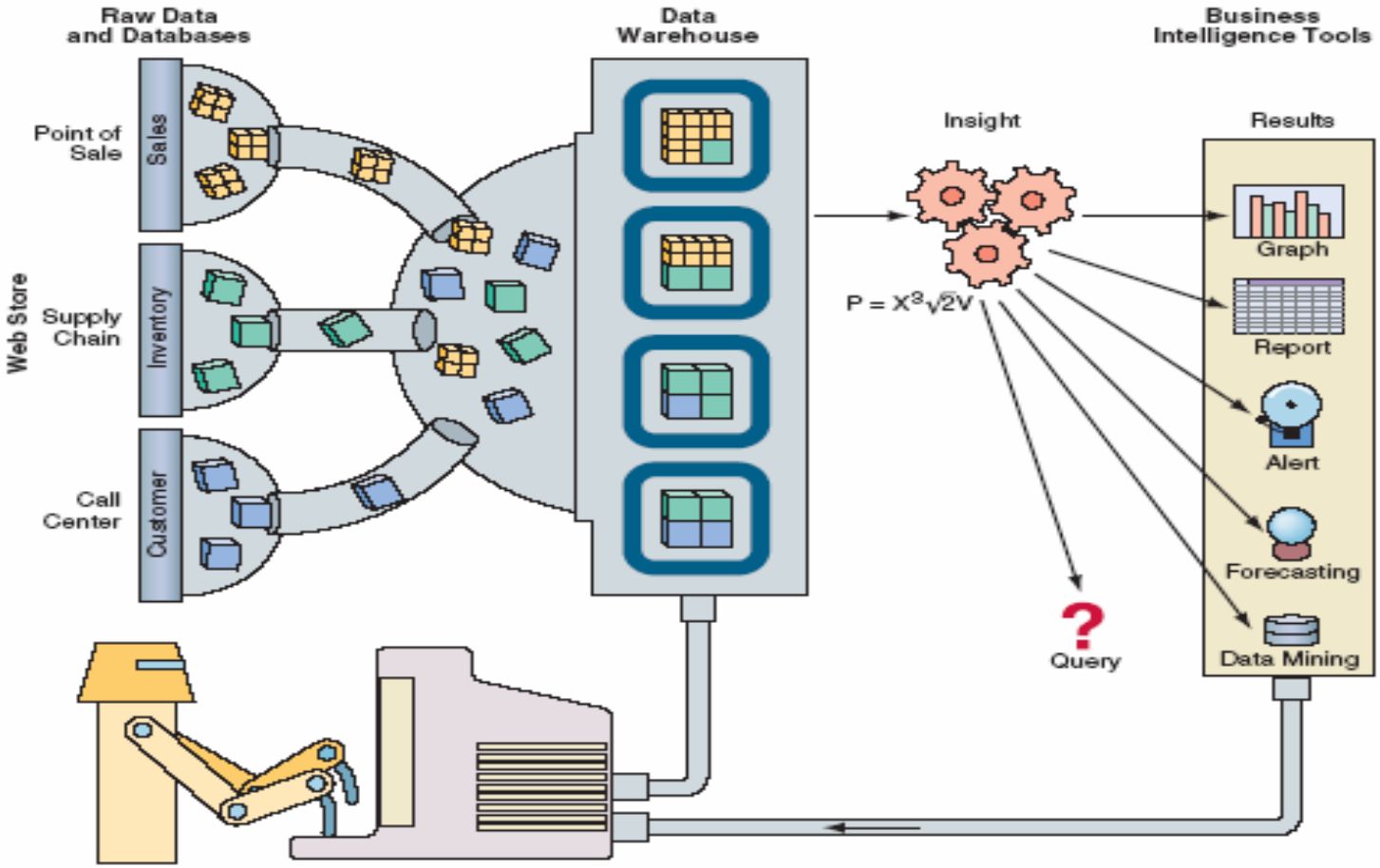
Business Intelligence

Business intelligence (BI) is a broad category of applications and techniques for gathering, storing, analyzing and providing access to data. It help's enterprise users make better business and strategic decisions. Major applications include the activities of query and reporting, online analytical processing (OLAP), DSS, data mining, forecasting and statistical analysis.

- Business intelligence includes:
  - outputs such as financial modeling and budgeting
  - resource allocation
  - coupons and sales promotions
  - Seasonality trends
  - Benchmarking (business performance)
  - competitive intelligence.

Starts with Knowledge Discovery

Business Intelligence Continued



How It Works.

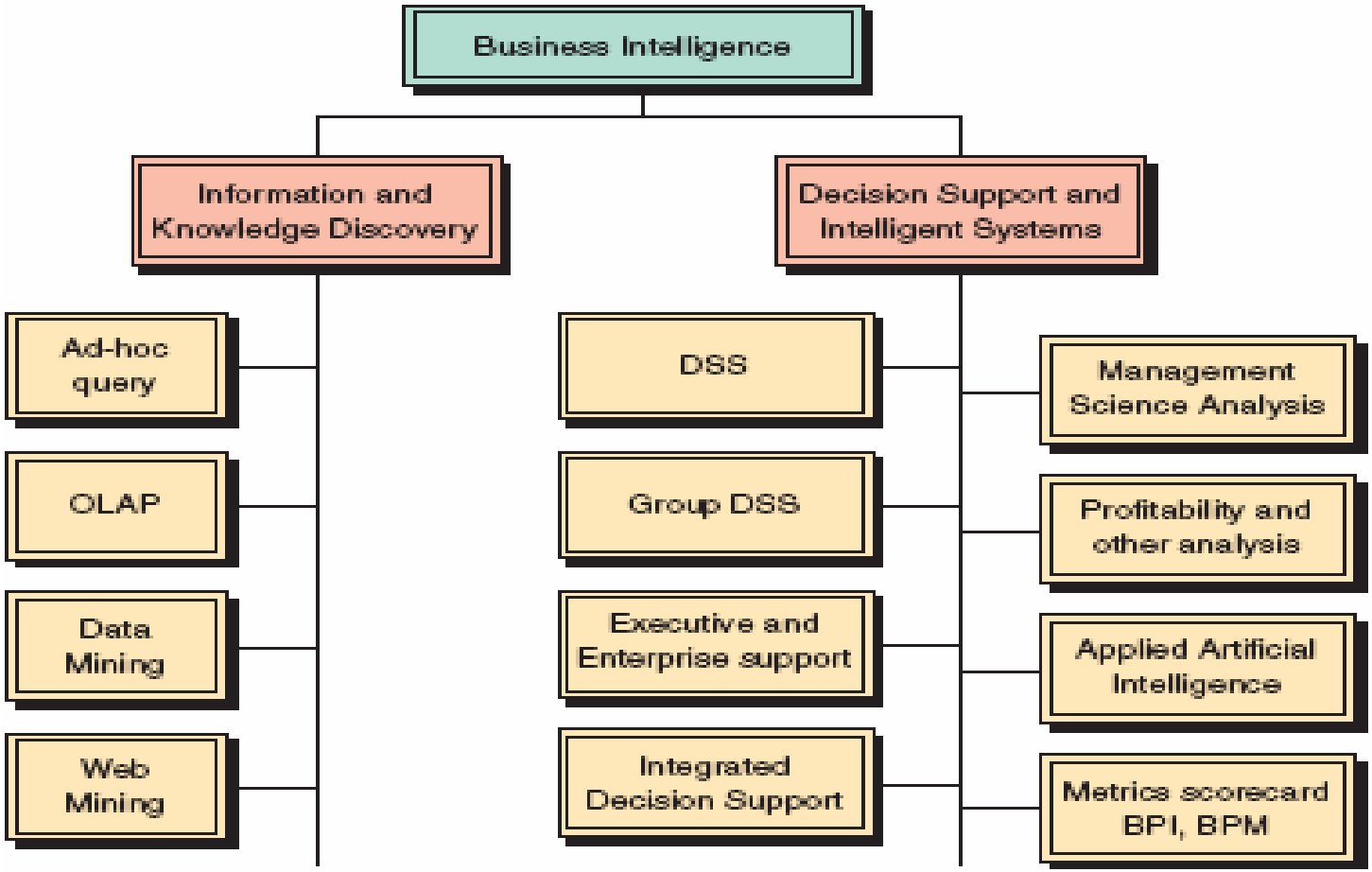
Knowledge Discovery

Before information can be processed into BI it must be discovered or extracted from the data stores. The major objective of this knowledge discovery in databases (KDD) is to identify valid, novel, potentially useful, and understandable patterns in data.

- KDD supported by three technologies:
  - massive data collection
  - powerful multiprocessor computers
  - data mining and other algorithms.
- KDD primarily employs three tools for information discovery:
  - Traditional query languages (SQL, ...)
  - OLAP
  - Data mining

Discovering useful patterns

Knowledge Discovery Continued



Discovering useful patterns

Queries

Queries allow users to request information from the computer that is not available in periodic reports. Query systems are often based on menus or if the data is stored in a database via a structured query language (SQL) or using a query-by-example (QBE) method.

- User requests are stated in a query language and the results are subsets of the relationship
  - Sales by department by customer type for specific period
  - Weather conditions for specific date
  - Sales by day of week
  - ...

Online Analytical Processing

Online analytical processing (OLAP) is a set of tools that analyze and aggregate data to reflect business needs of the company. These business structures (multidimensional views of data) allow users to quickly answer business questions. OLAP is performed on Data Warehouses and Marts.

- ROLAP (Relational OLAP) is an OLAP database implemented on top of an existing relational database. The multidimensional view is created each time for the user.
- MOLAP (Multidimensional OLAP) is a specialized multidimensional data store such as a Data Cube. The multidimensional view is physically stored in specialize data files.

Application View not a data structure or schema

Data Mining

Data mining is a tool for analyzing large amounts of data. It derives its name from the similarities between searching for valuable business information in a large database, and mining a mountain for a vein of valuable ore.

- Data mining technology can generate new business opportunities by providing:
  - Automated prediction of trends and behaviors.
  - Automated discovery of previously unknown or hidden patterns.
- Data mining tools can be combined with:
  - Spreadsheets
  - Other end-user software development tools
- Data mining creates a data cube then extracts data

Data Mining Techniques

- Case-based reasoning. uses historical cases to recognize patterns
- Neural computing is a machine learning approach which examines historical data for patterns.
- Intelligent agents retrieving information from the Internet or from intranet-based databases .
- Association analysis uses a specialized set of algorithms that sort through large data sets and express statistical rules among items.
- Decision trees
- Genetic algorithms
- Nearest-neighbor method

Data Mining Tasks

- Classification. Infers the defining characteristics of a certain group.
- Clustering. Identifies groups of items that share a particular characteristic. *Clustering differs from classification in that no predefining characteristic is given.*
- Association. Identifies relationships between events that occur at one time.
- Sequencing. Identifies relationships that exist over a period of time.
- Forecasting. Estimates future values based on patterns within large sets of data.
- Regression. Maps a data item to a prediction variable.
- Time Series analysis examines a value as it varies over time.

“Other” Mining Environments

In addition to data stored in traditional databases there are other “structures” that can be mined for patterns.

- Text Mining is the application of data mining to non-structured or less-structured text files
- Web Mining is the application of data mining techniques to data related to the World Wide Web. The data may be present in web pages or related to Web activity.
- Spatial Mining is the application of data mining techniques to data that have a location component.
- Temporal Mining is the application of data mining techniques to data that are maintained for multiple points in time.

Data Visualization

Data visualization refers to presentation of data by technologies such as digital images, geographical information systems, graphical user interfaces, multidimensional tables and graphs, virtual reality, three-dimensional presentations, videos and animation.

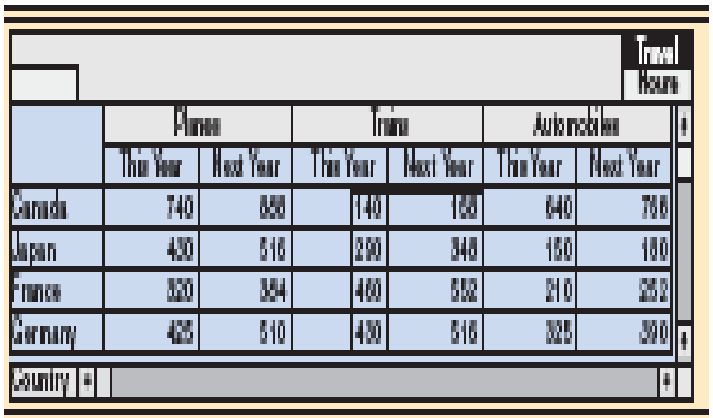
- Multidimensionality Visualization: Modern data and information may have several dimensions.
  - Dimensions:
    - Products
    - Salespeople
    - Market segments
    - Business units
    - Geographical locations
    - Distribution channels
    - Countries
    - Industries

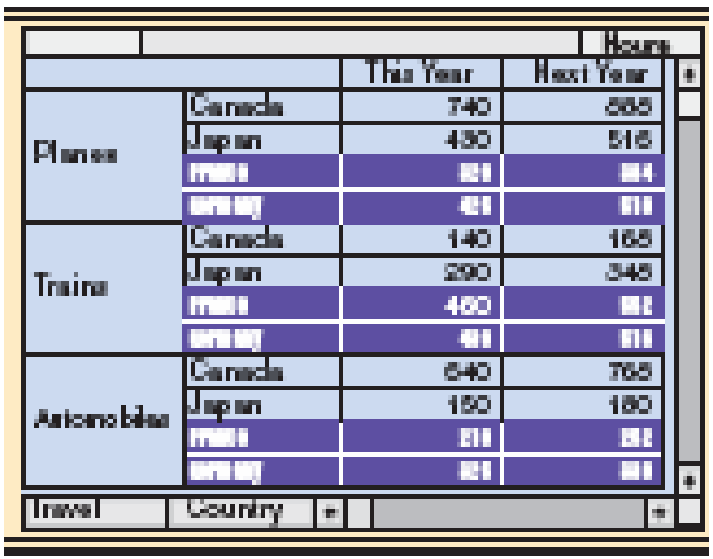
Data Visualization Continued

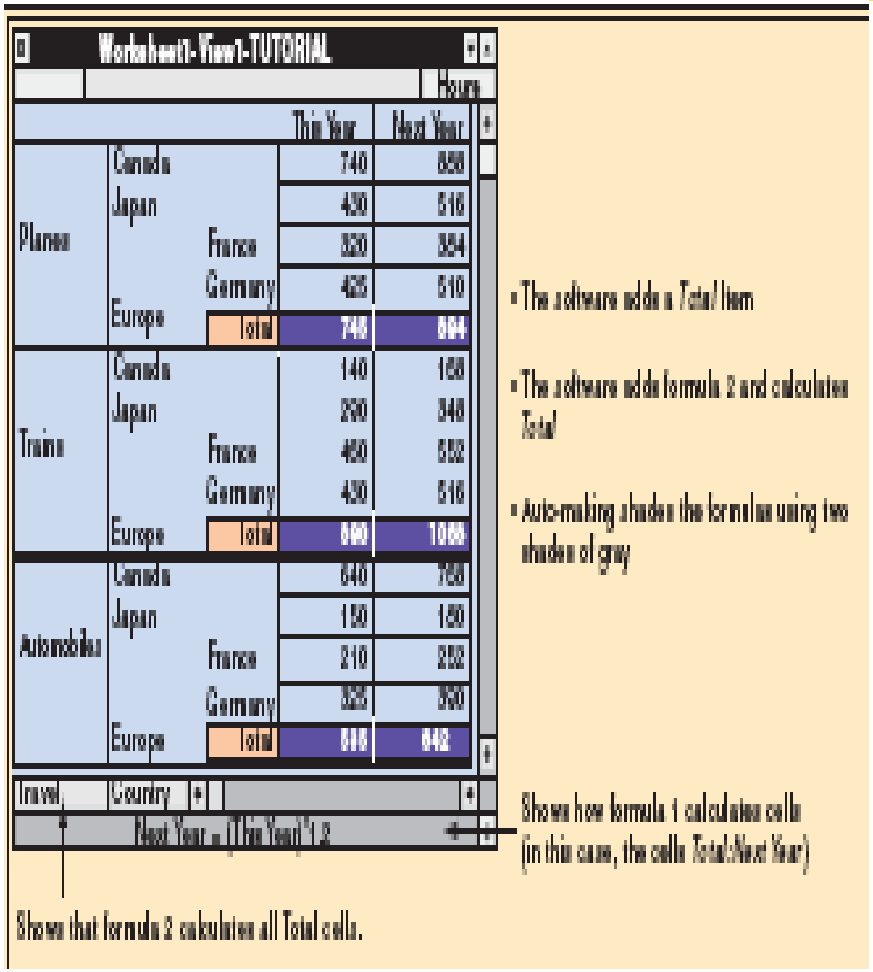
Multidimensionality Visualization:

- Measures:
  - Money
  - Sales volume
  - Head count
  - Inventory profit
  - Actual versus forecasted results.
- Time:
  - Daily
  - Weekly
  - Monthly
  - Quarterly
  - Yearly.

Data Visualization Continued







Data Visualization Continued

- A geographical information system (GIS) is a computer-based system for capturing, storing, checking, integrating, manipulating, and displaying data using digitized maps. Every record or digital object has an identified geographical location. It employs spatially oriented databases.
- Visual interactive modeling (VIM) uses computer graphic displays to represent the impact of different management or operational decisions on objectives such as profit or market share.
- Virtual reality (VR) is interactive, computer-generated, three-dimensional graphics delivered to the user. These artificial sensory cues cause the user to "believe" that what they are doing is real.

Specialized Databases

Data warehouses and data marts serve end users in all functional areas.
Most current databases are static: They simply gather and store information.
Today's business environment also requires specialized databases.

- Marketing transaction database (MTD)
  - combines many of the characteristics of the current databases and marketing data sources into a new database that allows marketers to engage in real-time personalization and target every interaction with customers
- Interactive capability
  - an interactive transaction occurs with the customer exchanging information and updating the database in real time, as opposed to the periodic (weekly, monthly, or quarterly) updates of classical warehouses and marts.

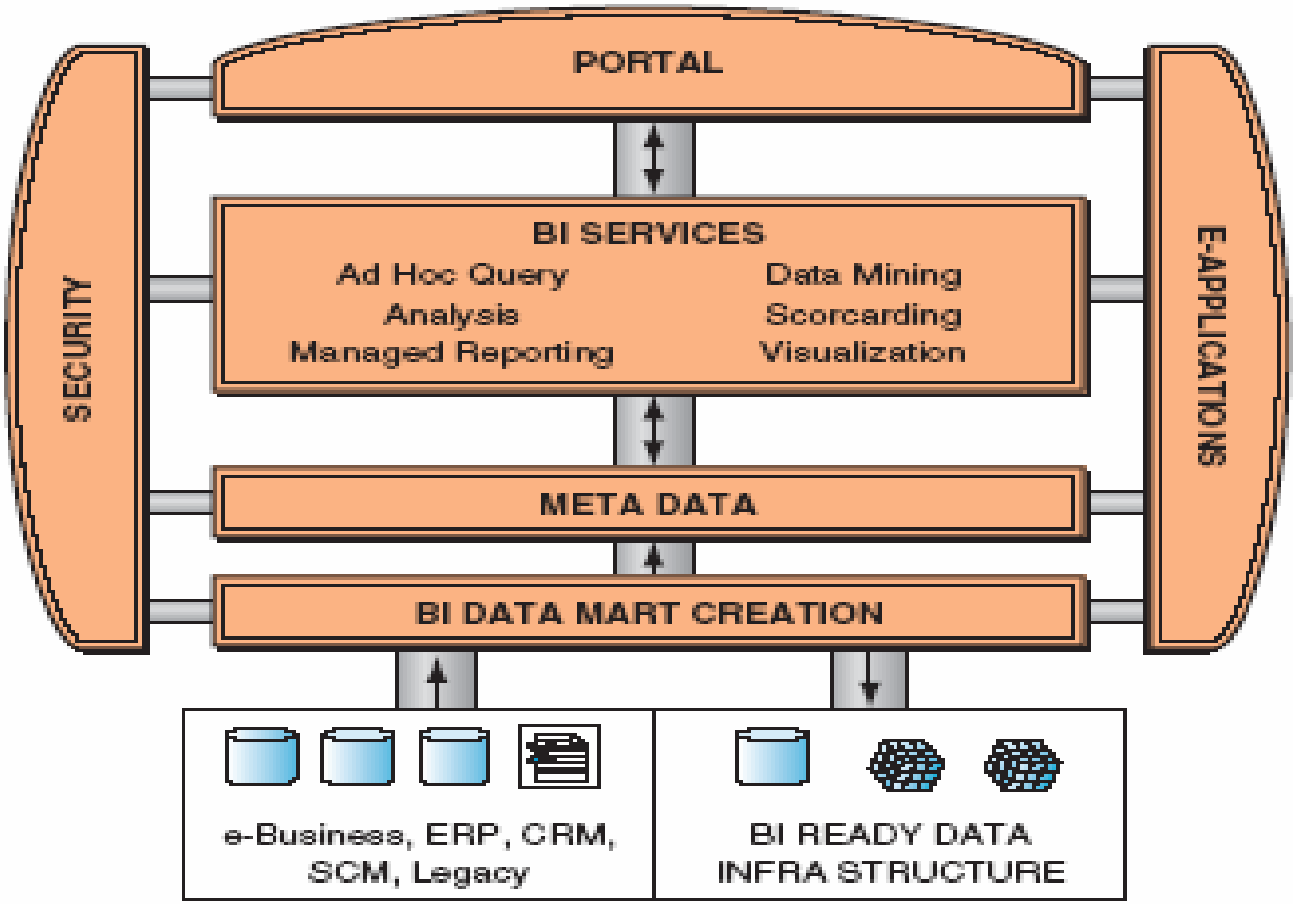
Web-based Data Management Systems

Data management and business intelligence activities—from data acquisition to mining—are often performed with Web tools, or are interrelated with Web technologies and e-business. This is done through intranets, and for outsiders via extranets.

- Enterprise BI suites and Corporate Portals integrate query, reporting, OLAP, and other tools
- Intelligent Data Warehouse Web-based Systems employ a search engine for specific applications which can improve the operation of a data warehouse
- Clickstream Data Warehouse occur inside the Web environment, when customers visit a Web site.

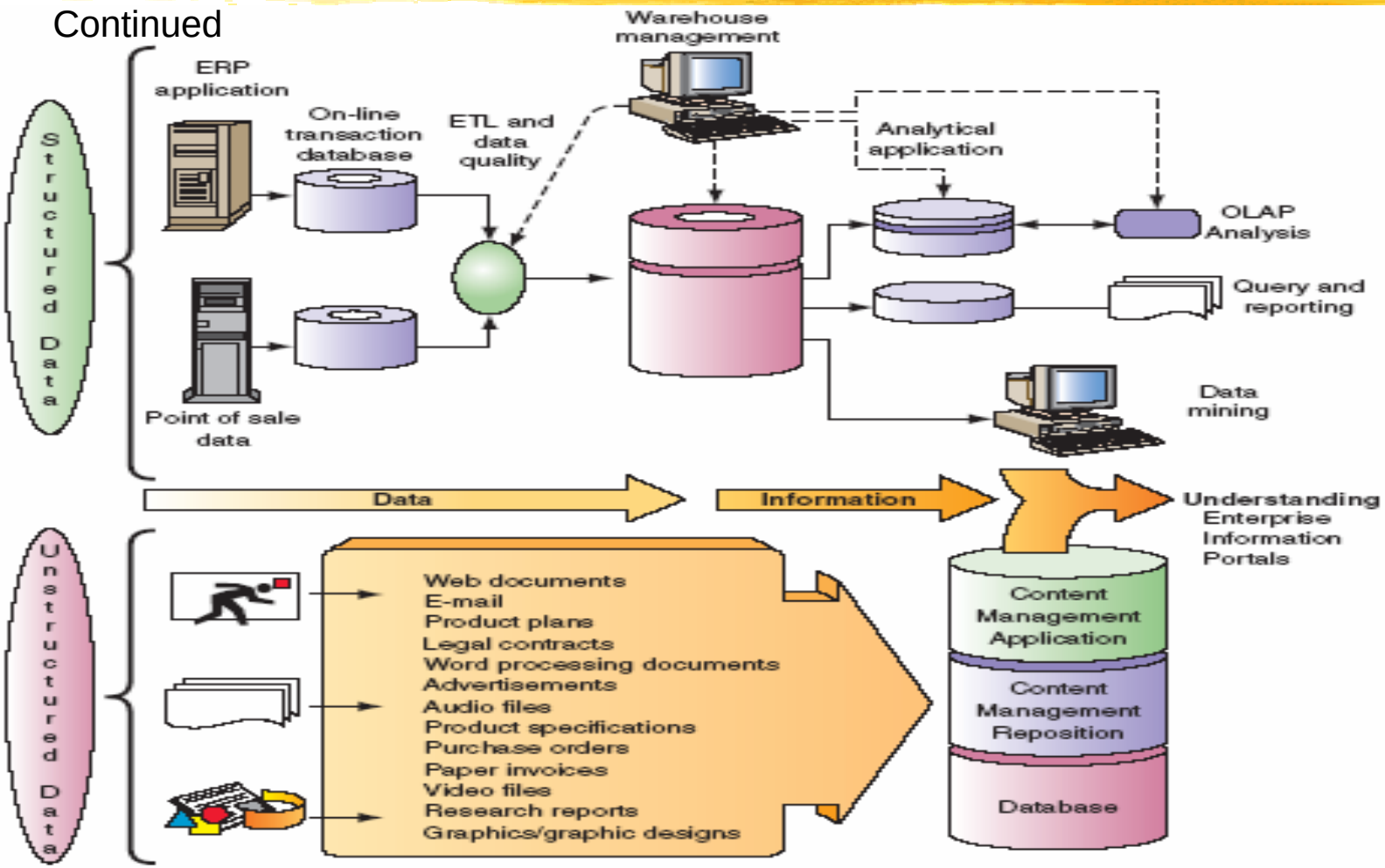
Web-based Data Management Systems

Continued



Web-based Data Management Systems

Continued



MANAGERIAL ISSUES

- **Cost-benefit issues and justification.** Some of the data management solutions discussed are very expensive and justifiable only in large corporations. Smaller organizations can make the solutions cost effective if they leverage existing databases rather than create new ones. A careful cost-benefit analysis must be undertaken before any commitment to the new technologies is made.
- **Where to store data physically.** Should data be distributed close to their users? This could potentially speed up data entry and updating, but adds replication and security risks. Or should data be centralized for easier control, security, and disaster recovery? This has communications and single point of failure risks.
- **Legal issues.** Data mining may suggest that a company send catalogs or promotions to only one age group or one gender. *A man sued Victoria's Secret Corp. because his female neighbor received a mail order catalog with deeply discounted items and he received only the regular catalog (the discount was actually given for volume purchasing).* Settling discrimination charges can be very expensive.
- **Internal or external?** Should a firm invest in internally collecting, storing, maintaining, and purging its own databases of information? Or should it subscribe to external databases, where providers are responsible for all data management and data access?

MANAGERIAL ISSUES Continued

- **Disaster recovery.** Can an organization's business processes, which have become dependent on databases, recover and sustain operations after a natural or other type of information system disaster? How can a data warehouse be protected? At what cost?
- **Data security and ethics.** Are the company's competitive data safe from external snooping or sabotage? Are confidential data, such as personnel details, safe from improper or illegal access and alteration? Who owns such personal data?
- **Ethics:** Paying for use of data. Compilers of public-domain information, such as Lexis-Nexis, face a problem of people lifting large sections of their work without first paying royalties. The Collection of Information Antipiracy Act (Bill HR 2652 in the U.S. Congress) will provide greater protection from online piracy. This, and other intellectual property issues, are being debated in Congress and adjudicated in the courts.
- **Privacy.** Collecting data in a warehouse and conducting data mining may result in the invasion of individual privacy. What will companies do to protect individuals? What can individuals do to protect their privacy?

MANAGERIAL ISSUES Continued

- **Legacy data.** One very real issue, often known as the legacy data acquisition problem, is what to do with the mass of information already stored in a variety of systems and formats,. Data in older, perhaps obsolete, databases still need to be available to newer database management systems. Many of the legacy application programs used to access the older data simply cannot be converted into new computing environments without considerable expense. Basically, there are three approaches to solving this problem. One is to create a database front end that can act as a translator from the old system to the new. The second is to cause applications to be integrated with the new system, so that data can be seamlessly accessed in the original format. The third is to cause the data to migrate into the new system by reformatting it.
- **Data delivery.** Moving data efficiently around an enterprise is often a major problem. The inability to communicate effectively and efficiently among different groups, in different geographical locations is a serious roadblock to implementing distributed applications properly, especially given the many remote sites and mobility of today's workers.