Data Mining and Data Warehousing

Introduction

Course Title:Introduction to Data Mining and Data Warehousing

- **Course Code**: CoSc4152?
- Credits: 3
- Lecture Hours: 48
- **Course Objective**
  - The objective of the course is to make learner understand foundation principles and techniques of data mining and data warehousing.
  - Students will be able to select and use various data mining language and tools very useful for adding business value of an organization.
- **Course Description**
  - Introduction, Data Preprocessing- Data Integration and Transformation, Classification, Association Analysis, Cluster Analysis, Information Privacy and Data Mining, Advanced Applications, Search engines, Data Warehouses, Capacity Planning.

Course Details

- **Unit 1: Introduction**
  - 1.1. Data Mining Origin
  - 1.2. Data Mining & Data Warehousing basics
- **Unit 2: Data Preprocessing**
  - 2.1. Data Types and Attributes
  - 2.2. Data Pre-processing
  - 2.3. OLAP
  - 2.4 Characteristics of OLAP Systems
  - 2.5 Multidimensional View and Data cube
  - 2.6 Data Cube Implementation
  - 2.7 Data Cube Operations (Roll-up, Roll Down, slice and dice and pivot)
  - 2.8 Guidelines for OLAP Implementation

Contd..

- **Unit 3: Data Warehousing**
  - 3.1. Operational Data sources
  - 3.2. ETL (Extract, Transform, Load)
  - 3.3. Data Warehouse Processes, Managers and their functions
  - 3.4. Data Warehouses and Data Warehouses Design
  - 3.5. Guidelines for Data Warehouse Implementation
- **Unit 4: Association Analysis**
  - 4.1. Basics and Algorithms
  - 4.2. Frequent Item-set Pattern & Apriori Principle
  - 4.3. FP-Growth, FP-Tree
  - 4.4. Handling Categorical Attributes

Contd..

- **Unit 5: Classification**
  - 5.1. Basics and Algorithms
  - 5.2. Decision Tree Classifier
  - 5.3. Rule Based Classifier
  - 5.4. Nearest Neighbor Classifier
  - 5.5. Bayesian Classifier
  - 5.6. Artificial Neural Network Classifier
  - 5.7. Issues : Over-fitting, Validation, Model Comparison
- **Unit 6: Cluster Analysis**
  - 6.1 . Basics and Algorithms
  - 6.2 . K-means Clustering
  - 6.3 . Hierarchical Clustering
  - 6.4 . Density-based spatial clustering of applications with noise (DBSCAN) Clustering

Contd..

- **Unit 7: Information Privacy and Data Mining**
  - 7.1 Basic principles to Protect Information Privacy
  - 7.2 Uses and Misuses of Data Mining
  - 7.3 Primary Aims of data Mining
  - 7.4 Pitfalls of Data Mining
- **Unit 8: Advanced Applications**
  - 8.1. Web-mining: Web content mining, web usage mining
  - 8.2. Time-series data mining

Contd..

- **Unit 9: Search Engines**
  - 9.1 Characteristics of search engine
  - 9.2 Search Engine functionality
  - 9.3 Ranking of Web pages

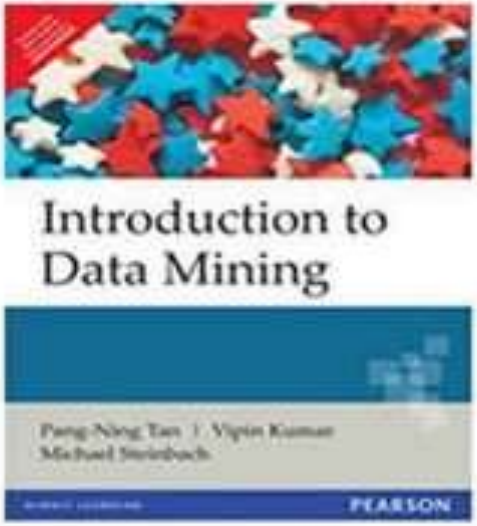
- **Unit 10 Capacity Planning**
  - 10.1 Calculating storage requirement, CPU requirements

Contd..

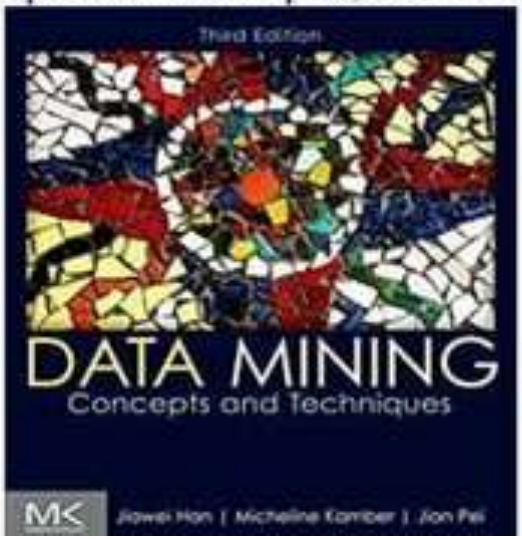
- **Practical:** Students should practice enough on real-world data intensive problems

Contd..

- **References:** Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Introductionto Data Mining, 2005, Addison- Wesley.

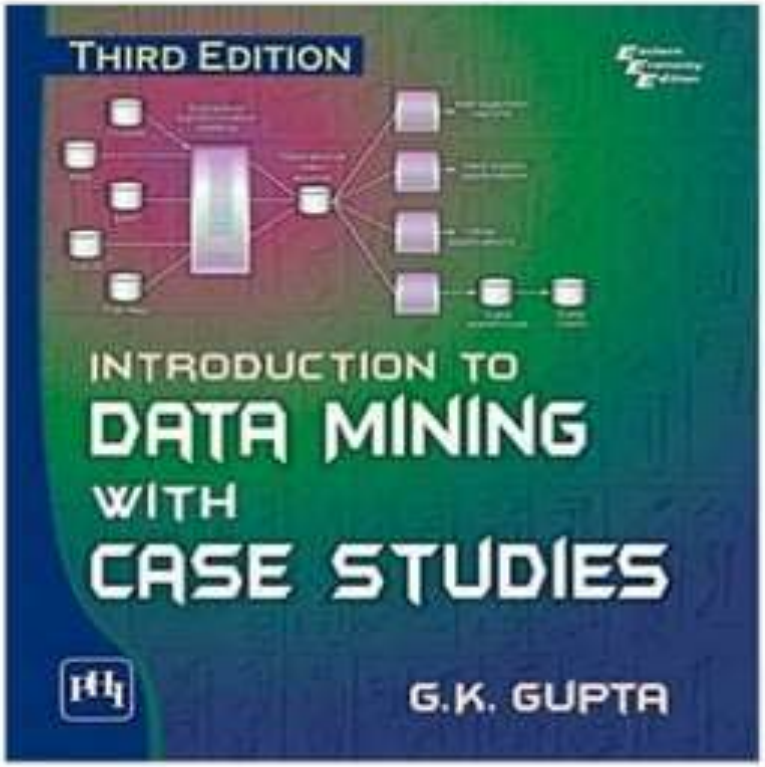


  - Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, 2nd Edition, 2006, Morgan Kaufmann.



Contd..

- G.K. Gupta, Introduction to Data Mining with Case Studies, Prentice Hall of India



- IBM, An Introduction to Building the Data Warehouse, Prentice Hall of India
- IBM, Introduction to Business Intelligence and Data Warehousing, Prentice Hall of India
- Adriaans Pieter, D. Zantige, "Data Mining", Pearson Education Asia Pub. Ltd, 2002

Unit 1 : Introduction to Data Mining and Data Warehousing

What is Data?

- A representation of facts, concepts, or instructions in a formal manner suitable for communication, interpretation, or processing by human beings or by computers.

Origin of Data mining

- The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful and affordable computers, data collection equipment, and storage media.
- This technology provides a great boost to the database and information industry, and makes a huge number of databases and information re-positories available.
- This availability of huge data repositories creates a Data explosion problem (data rich knowledge poor situation).

Contd..

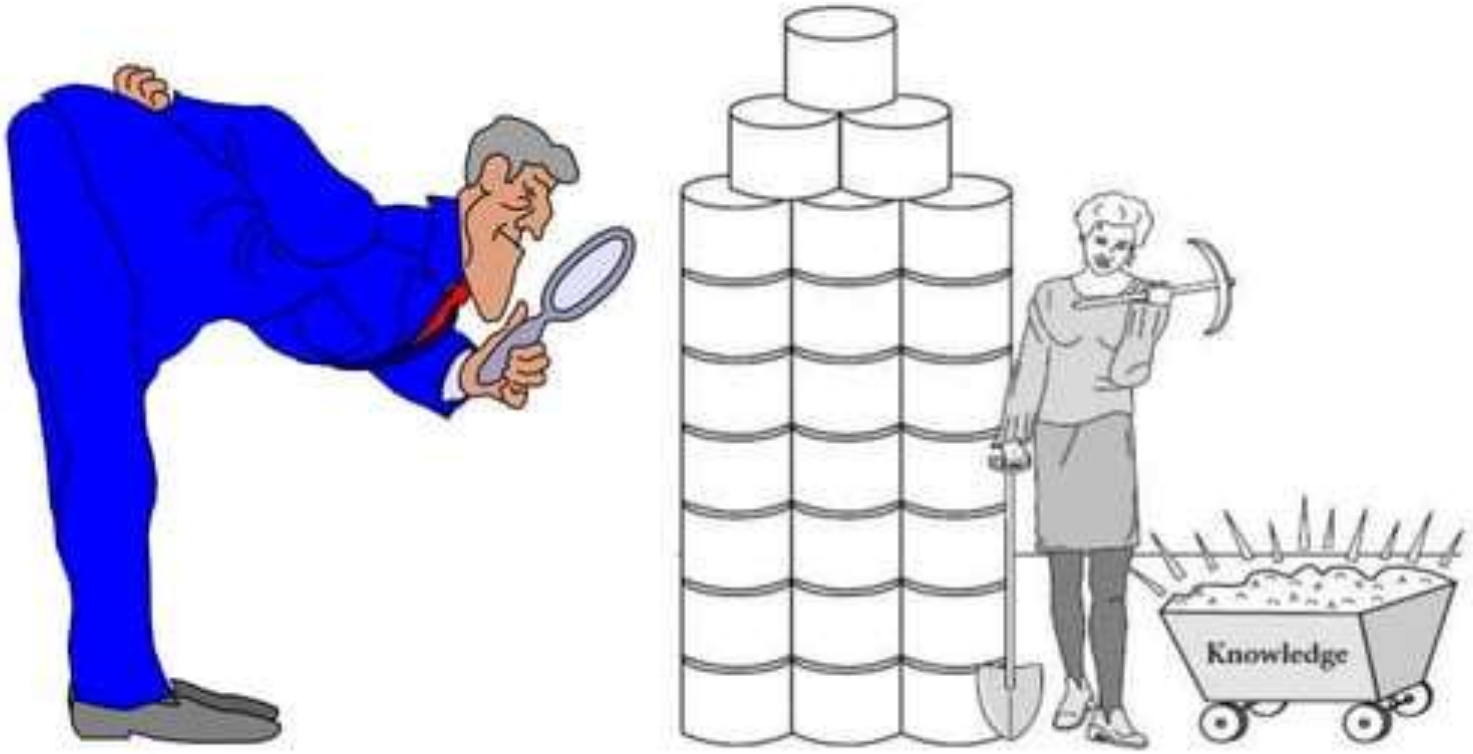
- We are drowning in data, but starving for knowledge!

- So, Powerful and versatile tools are badly needed to automatically uncover valuable information from tremendous amounts of data and to transform such data into organized knowledge.

Necessity is the Mother of invention! -plato

- This necessity has led to the birth of data mining.

What is Data Mining?



Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Contd...

- **Data mining: a misnomer?**
  - Scenario: Remember that the mining of gold from the rocks or sand is referred to as gold mining rather than rock or sand mining.
  - Thus, data mining should have been more appropriately named as "knowledge mining" which emphasis on mining knowledge from large amounts of data.
  - But, which is unfortunately somewhat long so, named "data mining"
- The overall goal of the data mining process is to extract pattern from a data set and transform it into an understandable structure for further use.

Contd..

- **Alternative names for data mining:**
  - Knowledge discovery(mining) in databases (KDD)
  - knowledge extraction
  - data/pattern analysis
  - data archeology
  - data dredging
  - information harvesting
  - business intelligence, etc.

Contd..

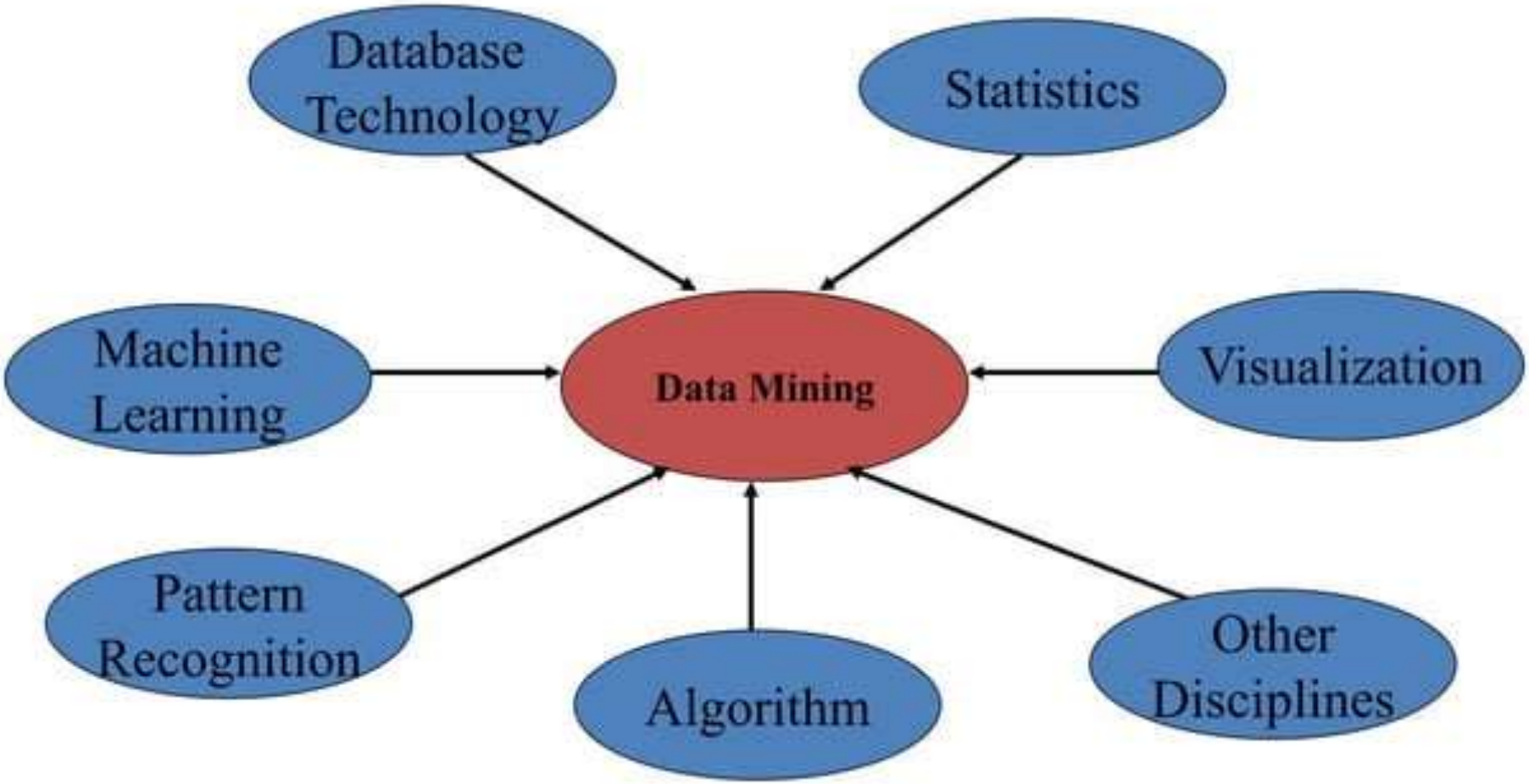
- **The key properties of data mining are:**
  - Automatic discovery of patterns
    - E.g., Market basket analysis.
  - Prediction of likely outcomes
    - E.g., weather forecasting
  - Creation of actionable information
    - E.g., Police investigation
  - Focus on large datasets and databases

Data mining is not

- Brute-force crunching of bulk data.
- "Blind" application of algorithms.
- Going to find non-existential relationships.
- Presenting data in different ways
- Queries to the database are not DM.
- A magic that will turn your data into gold.

Contd..

- **Data Mining: Confluence of Multiple Disciplines :**



Why Data Mining?—Potential Applications

- **Data analysis and decision support**
  - Market analysis and management
    - Market basket analysis, sale techniques, customer feedback on items (Opinion Mining)
  - Risk analysis and management
    - Forecasting, decision support system
  - Fraud detection and detection of unusual patterns (outliers)

Why Data Mining?—Potential Applications

- **Other Applications**
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining (mining from continuous / rapid data
    - Eg Telephone communication pattern, Web Searching, Sensor data
  - Bioinformatics and bio-data analysis

Data Mining: On What Kinds of Data?

- As a general technology, data mining can be applied to any kind of data as long as the data are meaningful for a target application.

  - Database-oriented data sets and applications
    - Relational database, data warehouse, transactional database

Contd..

- **Advanced data sets and advanced applications**
  - Data streams and sensor data
  - Time-series data
  - graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Knowledge Discovery (KDD) Process..

- Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data stored in databases, data warehouses, or other information repositories.
- Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD.
- Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.
- Knowledge discovery consists of an iterative sequence of the following steps:

Contd..

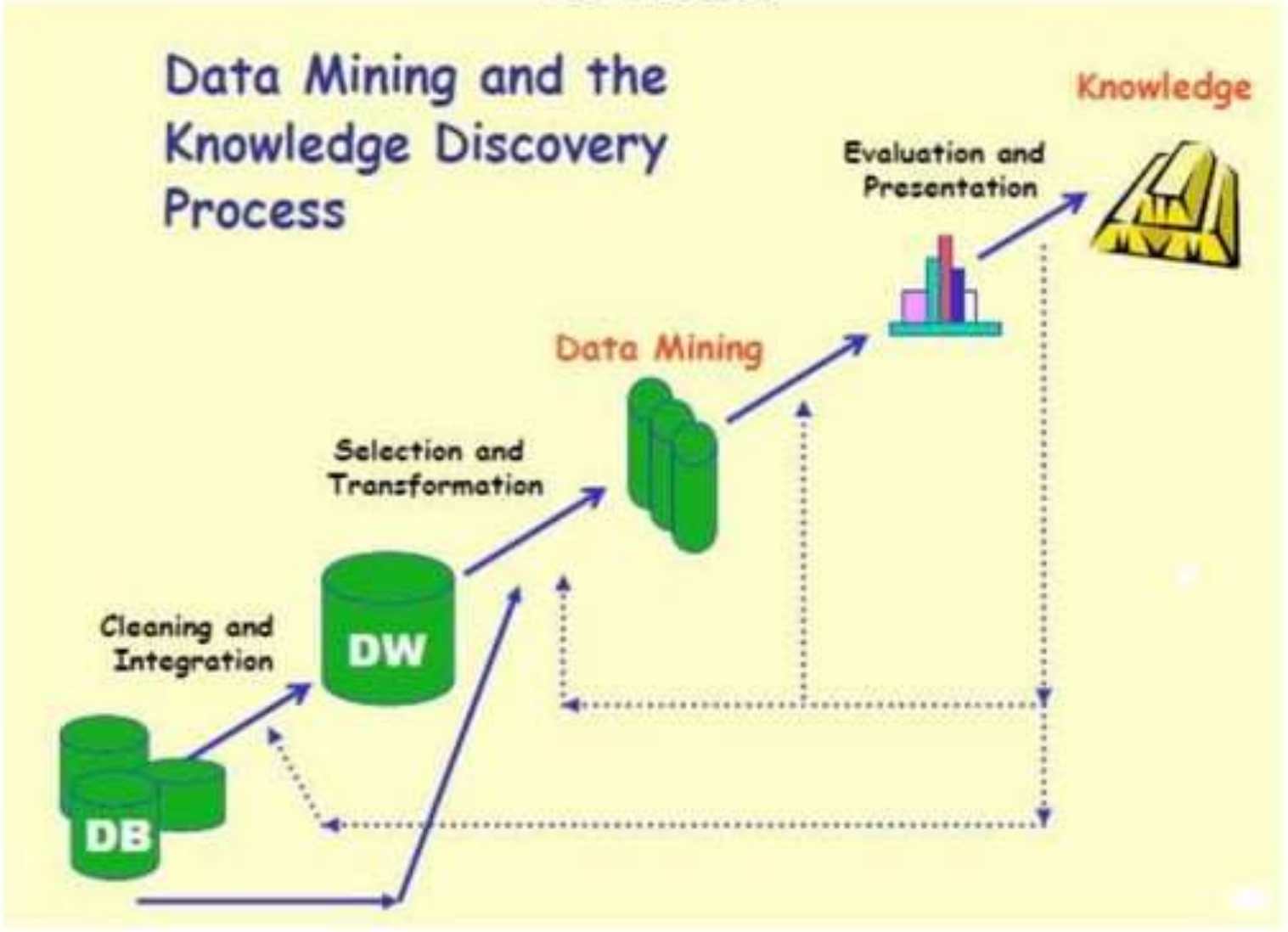


Figure: Knowledge Discovery Process (Stages of KDD)

Contd..

- **Data cleaning** :
  - It removes noise and inconsistent data
- **Data integration:**
  - This combines data from multiple data sources
- **Data selection:**
  - Data relevant to the analysis task are retrieved from the database
- **Data transformation:**
  - Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

Contd..

- **Data mining:**
  - an essential process where intelligent methods are applied in order to extract data patterns
- **Pattern evaluation:**
  - Identifies the truly interesting patterns representing knowledge based on some interestingness measures.
- **Knowledge presentation:**
  - Knowledge representation techniques are used to present the mined knowledge to the user.

Contd..

- According to this view, data mining is only one step in the knowledge discovery process.

- However, in industry, in media, and in the database research milieu, the term data mining is becoming more popular than the longer term of knowledge discovery from data.

- Therefore, we choose to use the term data mining.

- Based on this view, the architecture of a typical data mining system is described in the following slides

Architecture of Data Mining System

- A typical data mining system may have the following major components.

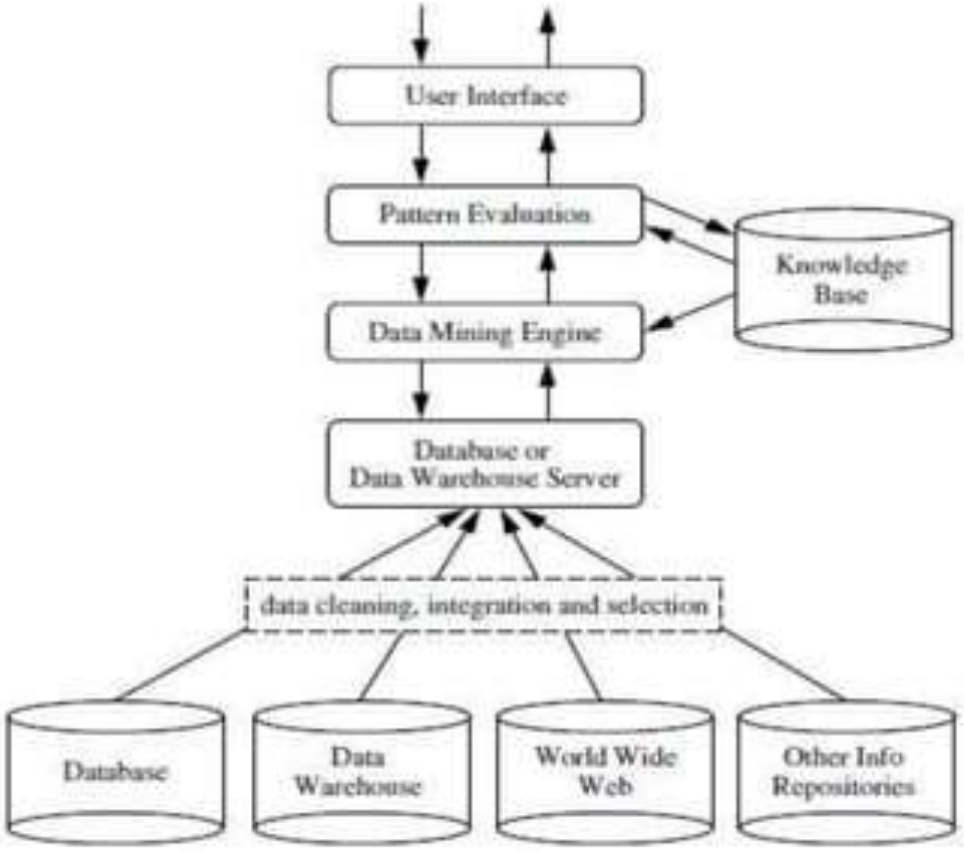


Figure: Architecture of Data Mining System

Contd..

- **Database, Data Warehouse, World Wide Web, or Other Information Repository**:
  - This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories.
  - Data cleaning and data integration techniques may be performed on the data.

Contd..

- **Database or Data Warehouse Server:**
  - The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.

Contd..

- **Knowledge Base:**
  - This is the domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. It is simply stored in the form of set of rules.
  - Such knowledge can include concept hierarchies, used to organize attributes or attribute values into different levels of abstraction.

Contd..

- **Data Mining Engine:**
  - This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as association and correlation analysis, classification, prediction, cluster analysis, outlier analysis and etc.

Contd..

- **Pattern Evaluation Module**:
  - This component typically employs interestingness measures and interacts with the data mining modules so as to *focus* the search toward interesting patterns.
  - It may use interestingness thresholds to filter out discovered patterns.

Contd..

- **User interface:**
  - This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task.
  - In addition, this component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

What is a Data Warehouse?

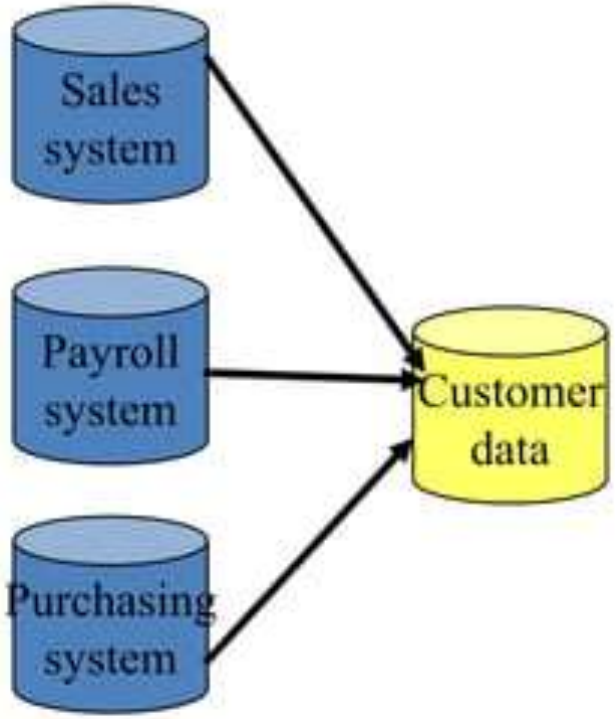
- A warehouse in general terms is a historic repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Data Warehouses are constructed via a process of
  - DATA CLEANING,
  - DATA INTEGRATION,
  - DATA TRANSFORMATION,
  - DATA LOADING, and
  - PERIODIC DATA REFRESHING.
- A data warehouse stores historical data of an organization so that they can analyze their performance over the past time (days, weeks, months or years) and plan for the future.

Contd……

- The popular definition of the data warehouse is given by WH Inmon:
  - "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."
- Data warehousing:
  - The process of constructing and using data warehouses.
  - Is the process of extracting & transferring operational data into informational data & loading it into a central data store (warehouse)

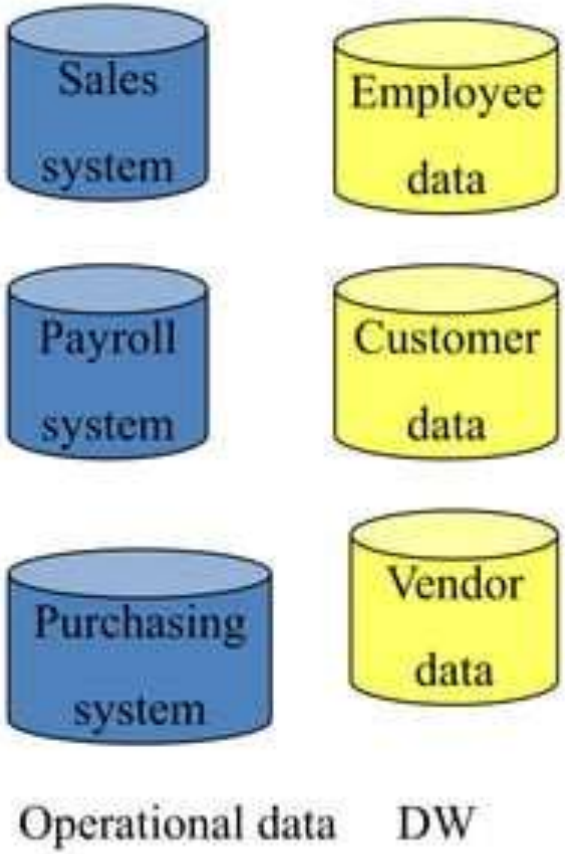
Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, etc.



Data Warehouse—Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

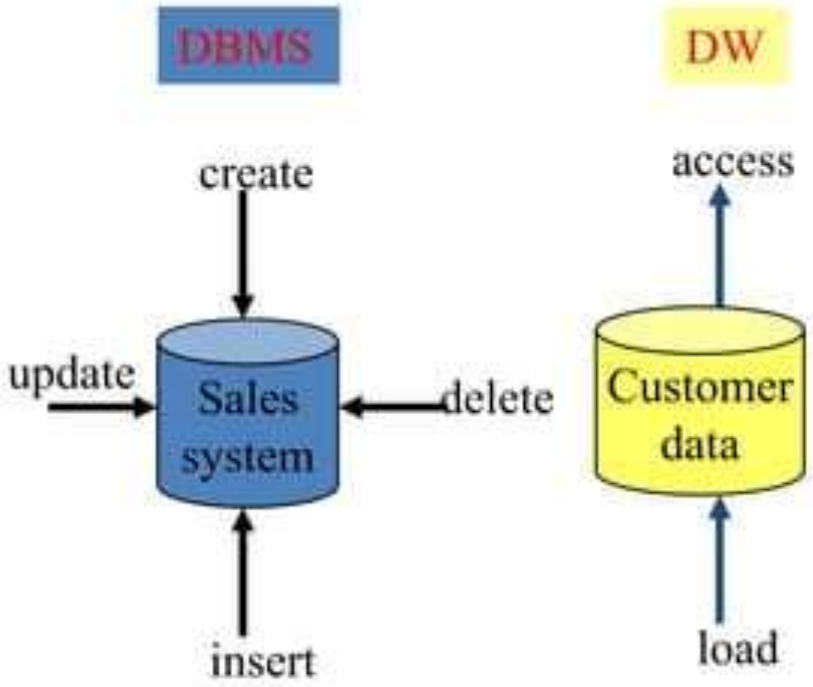


Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems.
  - Operational database: current value data.
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)

Data Warehouse—Non-Volatile

- A physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment.
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data.*



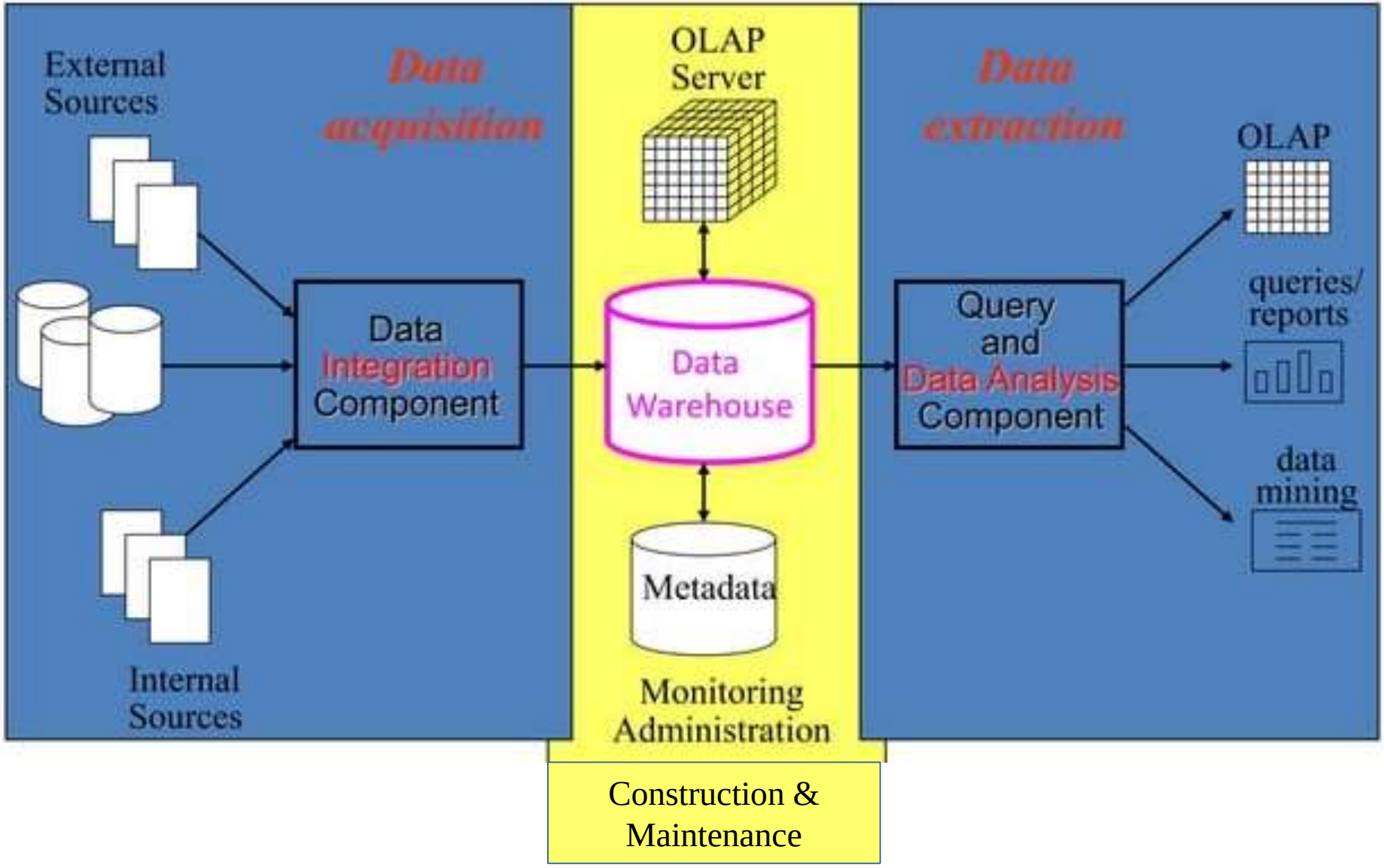
Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
    - supports querying, basic statistical analysis, and reporting using tables, charts and graphs.

Contd..

- Analytical processing
  - multidimensional analysis of data warehouse data
  - supports basic OLAP operations(drill-down, roll-up, slice-dice, drilling, pivoting, which allows the user to view the data at differing degree of summarization.
- Data mining
  - knowledge discovery from hidden patterns
  - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools.

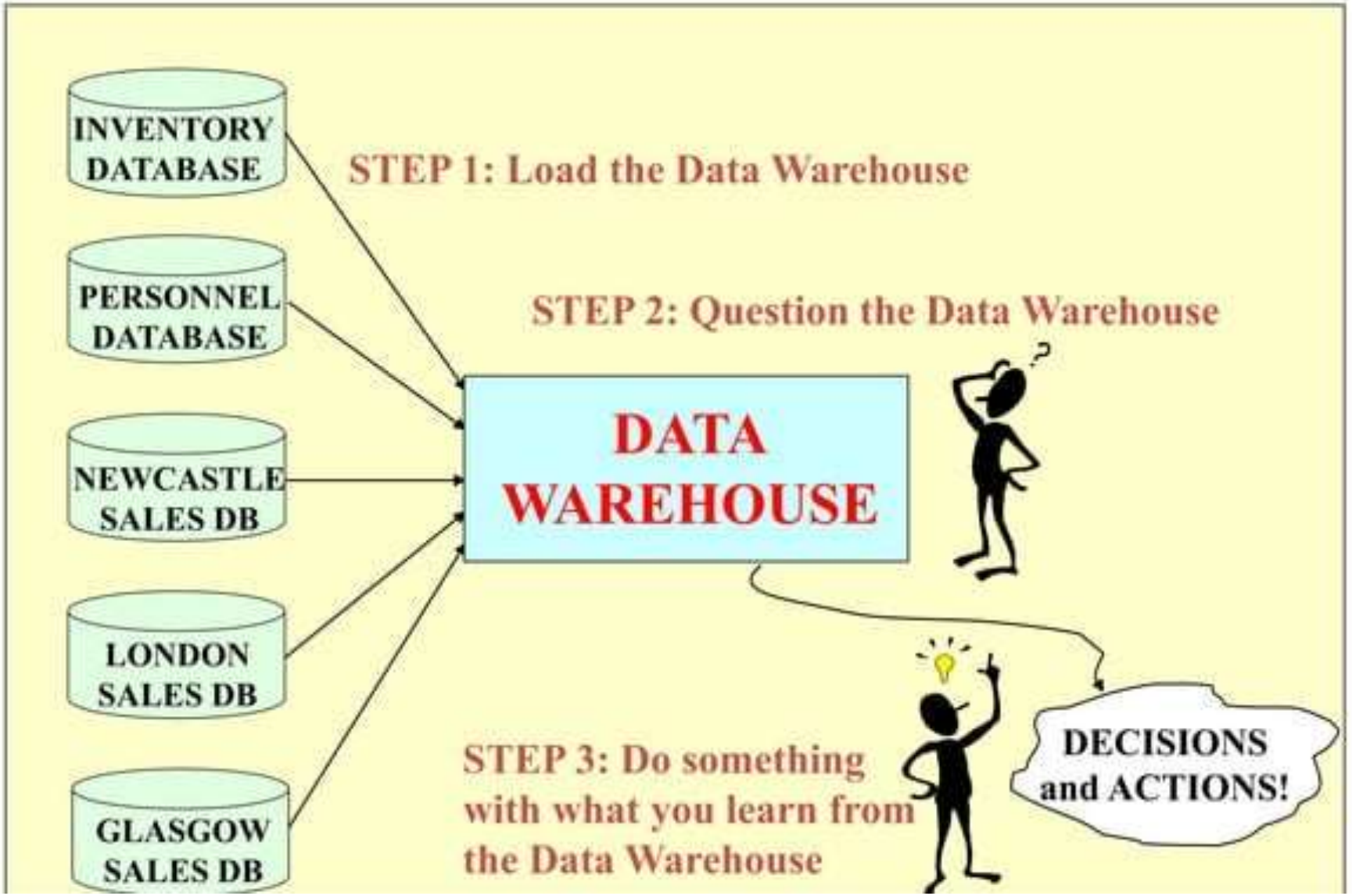
General Architecture



3 main phases

- Data acquisition:
  - relevant data collection
  - Recovering: transformation into the data warehouse model from existing models
  - Loading: cleaning and loading in the Data Warehouse.
- Storage
- Data extraction
  - Tool examples: Query/report, SQL, multidimensional analysis (OLAP tools), data mining
- Maintenance(Optional)

THE USE OF A DATA WAREHOUSE



Benefits of Data Warehousing

- Queries do not impact Operational systems
- Provides quick response to queries for reporting
- Enables Subject Area Orientation
- Integrates data from multiple, diverse sources
- Enables multiple interpretations of same data by different users or groups
- Provides thorough analysis of data over a period of time
- Accuracy of Operational systems can be checked
- Provides analysis capabilities to decision makers

- Increase customer profitability
- Cost effective decision making
- Manage customer and business partner relationships
- Manage risk, assets and liabilities
- Integrate inventory, operations and manufacturing
- Reduction in time to locate, access, and analyze information (Link multiple locations and geographies)
- Identify developing trends and reduce time to market
- Strategic advantage over competitors

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers
- Provide reliable, High performance access
- Consistent view of Data: Same query, same data. All users should be warned if data load has not come in.
- Quality of data is a driver for business re-engineering.

Applications of Data Mining

- Data mining is an interdisciplinary field with wide and diverse applications
  - There exist nontrivial gaps between data mining principles and domain-specific applications
- Some application domains
  - Financial data analysis
  - Retail industry
  - Telecommunication industry
  - Biological data analysis

Data Mining for Financial Data Analysis

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of data warehouses for multidimensional data analysis and data mining
  - View the debt and revenue changes by month, by region, by sector, and by other factors
  - Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
  - feature selection and attribute relevance ranking
  - Loan payment performance
  - Consumer credit rating

- Classification and clustering of customers for targeted marketing
  - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
  - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
  - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail Industry

- Retail industry: huge amounts of data on sales, customer shopping history, etc.
- Applications of retail data mining
  - Identify customer buying behaviors
  - Discover customer shopping patterns and trends
  - Improve the quality of customer service
  - Achieve better customer retention and satisfaction
  - Enhance goods consumption ratios
  - Design more effective goods transportation and distribution policies

- Example 1. Design and construction of data warehouses based on the benefits of data mining
  - Multidimensional analysis of sales, customers, products, time, and region
- Example 2. Analysis of the effectiveness of sales campaigns
- Example 3. Customer retention: Analysis of customer loyalty
  - Use customer loyalty card information to register sequences of purchases of particular customers
  - Use sequential pattern mining to investigate changes in customer consumption or loyalty
  - Suggest adjustments on the pricing and variety of goods
- Example 4. Purchase recommendation and cross-reference of items

Data Mining for Telecommunication Industry

- A rapidly expanding and highly competitive industry and a great demand for data mining
  - Understand the business involved
  - Identify telecommunication patterns
  - Catch fraudulent activities
  - Make better use of resources
  - Improve the quality of service
- Multidimensional analysis of telecommunication data
  - Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.

- Fraudulent pattern analysis and the identification of unusual patterns
  - Identify potentially fraudulent users and their typical usage patterns
  - Detect attempts to gain fraudulent entry to customer accounts
  - Discover unusual patterns which may need special attention
- Multidimensional association and sequential pattern analysis
  - Find usage patterns for a set of communication services by customer group, by month, etc.
  - Promote the sales of specific services
  - Improve the availability of particular services in a region
- Use of visualization tools in telecommunication data analysis

Biomedical Data Analysis

- DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T).
- Gene: a sequence of hundreds of individual nucleotides arranged in a particular order
- Humans have around 30,000 genes
- Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
- Semantic integration of heterogeneous, distributed genome databases
  - Current: highly distributed, uncontrolled generation and use of a wide variety of DNA data
  - Data cleaning and data integration methods developed in data mining will help

- Similarity search and comparison among DNA sequences
  - Compare the frequently occurring patterns of each class (e.g., diseased and healthy)
  - Identify gene sequence patterns that play roles in various diseases
- Association analysis: identification of co-occurring gene sequences
  - Most diseases are not triggered by a single gene but by a combination of genes acting together
  - Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples
- Path analysis: linking genes to different disease development stages
  - Different genes may become active at different stages of the disease
  - Develop pharmaceutical interventions that target the different stages separately
- Visualization tools and genetic data analysis

Problems in Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long duration projects
- Complexity of integration

Major Challenges in Data Warehousing

- Data mining requires single, separate, clean, integrated, and self-consistent source of data.
  - A DW is well equipped for providing data for mining.
- Data quality and consistency is essential to ensure the accuracy of the predictive models.
  - DWs are populated with clean, consistent data
- Advantageous to mine data from multiple sources to discover as many interrelationships as possible.
  - DWs contain data from a number of sources.
- Selecting relevant subsets of records and fields for data mining
  - requires query capabilities of the DW.
- Results of a data mining study are useful if can further investigate the uncovered patterns.
  - DWs provide capability to go back to the data source.

- The largest challenge a data miner may face is the sheer volume of data in the data warehouse.
- It is quite important, then, that summary data also be available to get the analysis started.
- A major problem is that this sheer volume may mask the important relationships the data miner is interested in.
- The ability to overcome the volume and be able to interpret the data is quite important.

Major Challenges in Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality
- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration
- Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
- Application-oriented and domain-specific data mining
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

Homework

- Briefly explain data mining and define it. Why data mining being used more widely now?
- State and explain the major applications of data mining.
- Explain briefly some limitations of data mining.
- What is the future of data mining?
- Can data mining in some areas assist in identifying corruption? Select one area and study the possibilities.
- How is a data warehouse different from a database? How are they similar.
- Explain what data warehousing and OLAP aim to achieve that can not be achieved by OLTP systems.