Data Mining and Data Warehousing 3 hr

Unit 1: Introduction LH 2

Unit 2: Data Preprocessing LH 6

Unit 3: Classification LH 7

Unit 4: Association Analysis LH 7

Unit 5: Cluster Analysis LH 7

Unit 6: Information Privacy and Data Mining LH 3

Unit 7: Advanced Applications LH 3

Unit 8: Search Engines LH 3

Unit 9: Data Warehousing LH7

Unit 10 Capacity Planning LH 3

Unit 1: Introduction LH 2

- 1.1. Data Mining Origin
- 1.2.Data Mining & Data Warehousing basics

1. What Is Data Mining?

- Data mining refers to extracting or mining knowledge from large amounts of data. Thus, data mining appropriately named as **knowledge mining** which emphasis on mining from large amounts of data.
- The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.
- Data mining is the process of discovering **meaningful new correlations, patterns and trends** by shifting through large amounts of data stored in repositories, **using pattern recognition technologies as well as statistical and mathematical techniques.**

Applications of Data Mining

- Market Basket Analysis, Customer Segmentation, Fraud Detection,Bio Informatics, Intrusion Detection

The key properties of data mining are

- Automatic discovery of patterns.
- Prediction of likely outcomes.
- Creation of actionable information.
- Focus on large datasets and databases.

Scope of Data Mining

- Searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining it.

- Given databases of sufficient size and quality, data mining technology can generate **new business opportunities through**
  - **Automated prediction of trends and behaviors.**
  - **Automated discovery of previously unknown patterns.**

Automated prediction of trends and behaviors. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands- on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing.

Automated discovery of previously unknown patterns. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together..

•**Process of Knowledge Discovery in Database (KDD)**

1. **Data Cleaning** - noise and inconsistent data is removed.
2. **Data Integration** - multiple data sources are combined.
3. **Data Selection** – data appropriate to the analysis task are retrieved from the database.
4. **Data Transformation** - data are
- transformed or consolidated into forms appr
mining by performing summary or
operations.
5. **Data Mining** - intelligent methods are
- applied in order to extract data patterns.
- **6.Pattern Evaluation** - data are evaluated to i
the truly interesting patterns.

7. Knowledge Presentation - knowledge is represe
using visualization techniques.

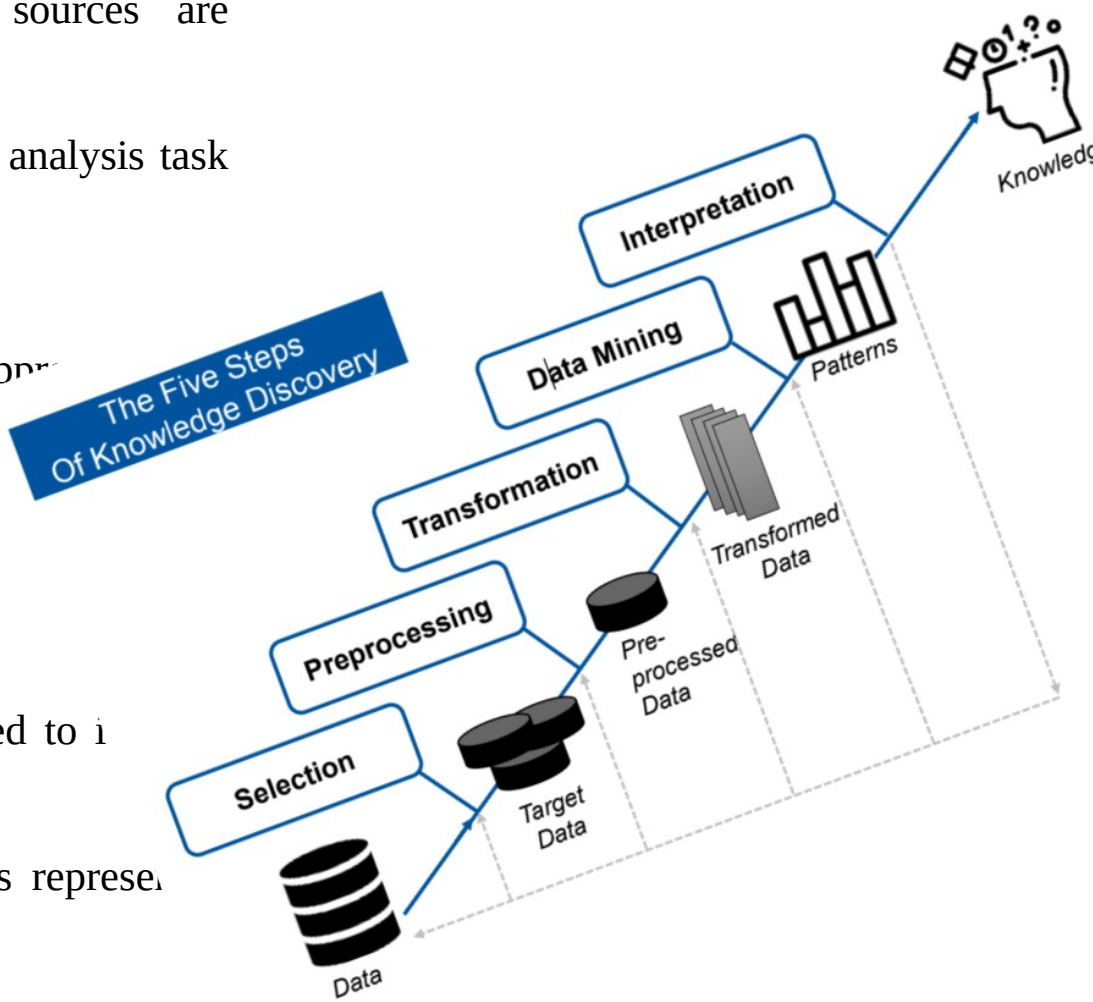
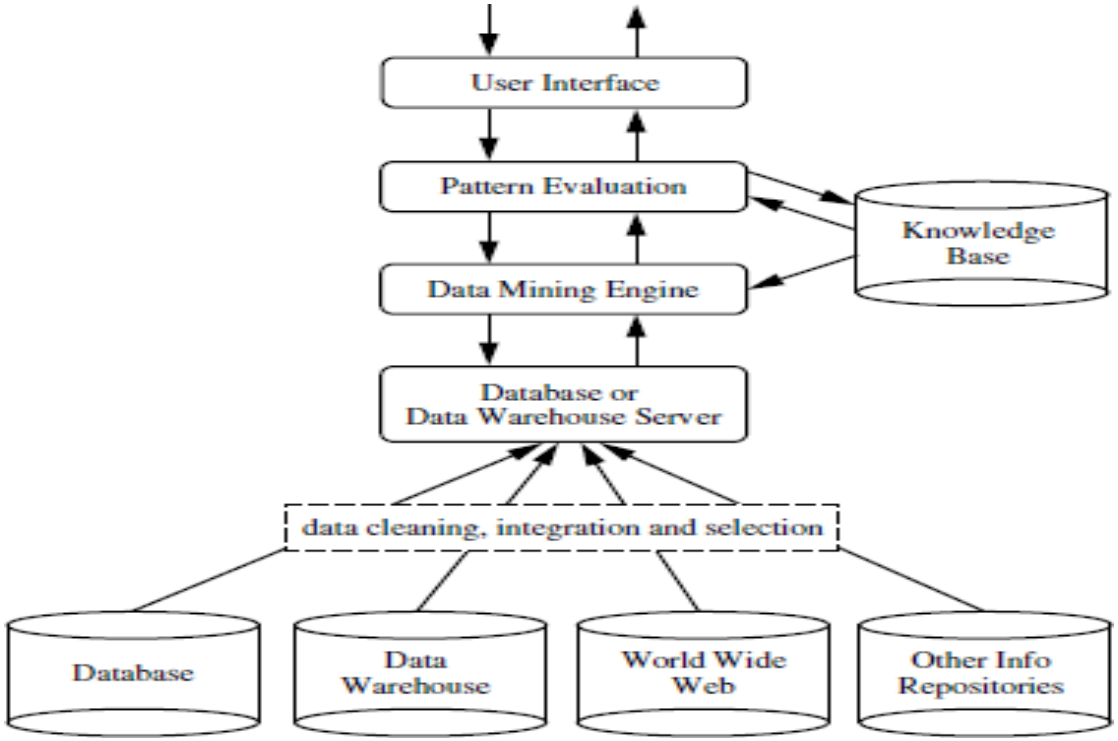


Figure 1. Data mining as a step in Knowledge Discovery

Architecture of Data Mining

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



Based on this view, the architecture of a typical data mining system has following major components:

Data sources: Data Sources consists of one or a set of databases, data warehouses, spreadsheets, Word Wide Web or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

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.

Knowledge base (KB): a knowledge base is a centralized repository for information: a public library or a database of related information about a particular subject. A KB is not a static collection of information, but a dynamic resource that may itself have the capacity to learn, as part of an artificial intelligence (AI) expert system.

Data mining engine: The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including **association, classification, characterization, clustering, prediction, time-series analysis etc**.

Pattern evaluation module: The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the.

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, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. 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

Data Mining Techniques

Association

- Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction.

Classification

- Classification is a classic data mining technique based on machine learning. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics to classify each item in a set of data into one of a predefined set of classes or groups.

Clustering

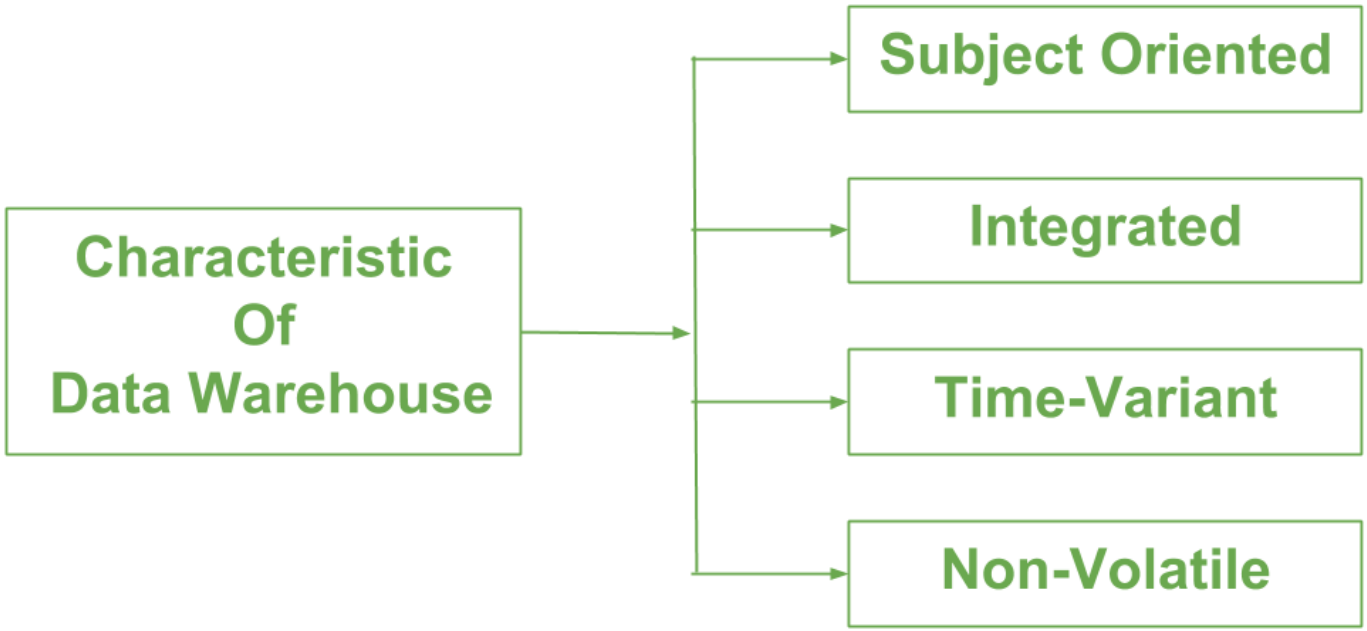
- Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes.

Regression

- Regression uses **existing values to forecast what other values will** be. A regression task begins with a data set in which the target values are known. For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time. The data might track age, height, weight, developmental milestones, family history, and so on. Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.

Data Warehouse

- A data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data to support management's decision-making, according to William H. Inmon.



Subject-oriented:

Focus is on Subject Areas rather than Applications

Organized around major subjects, such as customer, product, sales.

Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated:

- When data resides in many separate applications in the operational environment, the encoding of data is often inconsistent. For instance, in one application, gender might be coded as “m” and “f” in another by 0 and 1. When data are moved from the operational environment into the data warehouse, they assume a consistent coding convention e.g. gender data is transformed to “m” and “f”.

- Integration tasks handles naming conventions as well as physical attributes of data.

Time-variant:

The time horizon for the data warehouse is significantly longer than that of operational systems.

  - Operational database: current value data. (60 to 90 days)

- The data warehouse contains a place for storing data that are five to 10 years old, or older, to be used for comparisons, trends, and forecasting. These data are not updated.

Non-volatile:

- Data is not updated or changed in any way once they enter the data warehouse, but are only loaded and accessed.

Data Warehouse Features

- It is separate from the Operational Database.
- Integrates data from heterogeneous systems.
- Stores HUGE amount of data, more historical than current data.
- Does not require data to be highly accurate.
- Queries are generally complex.
- The goal is to execute statistical queries and provide results that can influence decision- making in favor of the Enterprise.
- These systems are thus called Online Analytical Processing Systems (OLAP).

Data Warehouse Design Process:

- A data warehouse can be built using a *top-down approach*, a *bottom-up approach*, or a *combination of both*.

- The top-down approach starts with the overall design and planning. It is useful in cases where the technology is mature and well known, and where the business problems that must be solved are clear and well understood. **The top-down approach goes from the general to the specific**

- The bottom-up approach starts with experiments and prototypes. This is useful in the early stage of business modeling and technology development. It allows an organization to move forward at considerably less expense and to evaluate the benefits of the technology before making significant commitments. **Bottom-up approach begins at the specific and moves to the general**.

- In the combined approach, an organization can exploit the planned and strategic nature of the top-down approach while retaining the rapid implementation and opportunistic application of the bottom-up approach.

Architecture of a Data Warehouse System

A typical data warehouse system has three main phases:

- **Data acquisition**
  - Relevant data collection
  - Recovering: transformation into the data warehouse model from existing models
  - Loading: cleaning and loading in the DW
- **Storage:**
  - **Metadata, Data Mart**
- **Data extraction**
  - Tool examples: Query report, SQL, multidimensional analysis (OLAP tools), data m

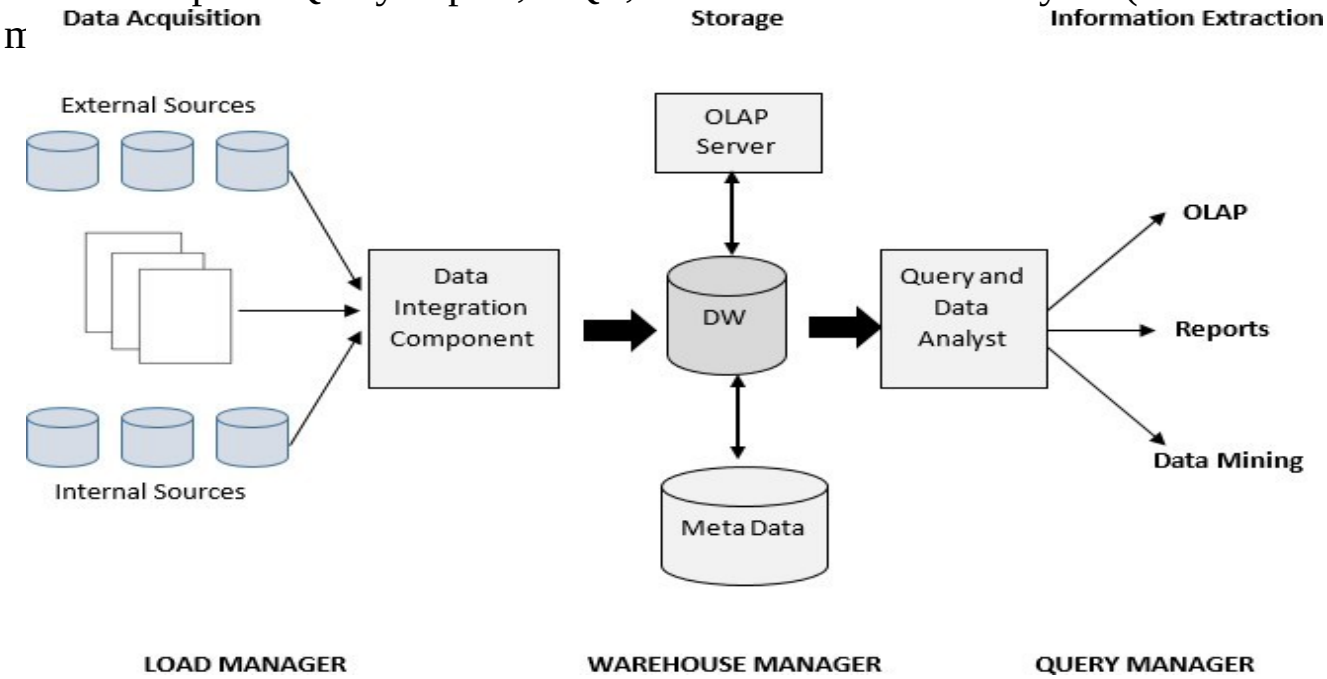


Figure 3. Data Warehouse Architecture

These three tasks are performed by following personnel:

Load Manager:

- The system components that perform all the operations necessary to support the extract and load process. It fast loads the extracted data into a temporary data store and performs simple transformations into a structure similar to the one in the data warehouse. Also called ETL (Extract Transform and Load).

Warehouse Manager (Data Manager):

- It is the system component that performs analysis of data to ensure consistency. The data from various sources and temporary storage are merged into data warehouse by the warehouse manager. The job of backing-up and archiving data as well as creation of index is performed by this manager.

Query Manager:

- Performs all the operations necessary to support the query management process by directing queries to the appropriate tables. In some cases it also stores query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate

Origin of Data Mining

- The term "Data mining" was introduced in the 1990s, but data mining is the evolution of a field with a long history. The traditional data analysis techniques have encountered following practical difficulties in meeting the challenges that motivated the development of data mining:
- **Scalability:** because of advances in data generation and collection, data sets with sizes of gigabytes, terabytes or even petabytes are becoming common. If the data mining algorithms are to handle these massive data sets, then they must be scalable. For example scalability can be improved by using sampling or developing parallel and distributed algorithms.
- **High dimensionality**: it is now common to encounter data sets with hundreds of attributes instead of the handful common a few decades ago. Data sets with temporal or spatial components often have high dimensionality. The traditional data analysis techniques that were developed for low- dimensional data often do not work well for such high dimensional data.
- **Heterogeneous and complex data:** as the role of data mining in business, science, medicine and other field has grown, so has the need for technique that can handle heterogeneous attributes. The traditional data analysis techniques only deals with data sets containing attributes of the same type, either continuous or categorical.

Data ownership and distribution: sometimes the data needed for an analysis is not stored in one location or owned by one organization. Instead, the data is geographically distributed among resources belonging to multiple entities. This requires the development of distributed data mining techniques.

Non-traditional analysis: the traditional statistical approach is based on a hypothesize-and-test paradigm. In this approach, a hypothesis is purposed, an experiment is designed to gather data and then the data is analyzed with respect to hypothesis. Unfortunately, this process is extremely labor intensive. Current data mining techniques require this task to be done in huge scale. So, the data mining techniques have been motivated by the desire to automate the process of hypothesis generation and evaluation.

So, to meet all those challenges, researchers from different disciplined began to focus on developing more efficient and scalable tools that could handle diverse types of data. This work, which culminated in the field of data mining, built upon the methodology and algorithms that researchers had previously used.