Unit II

DATA MINING

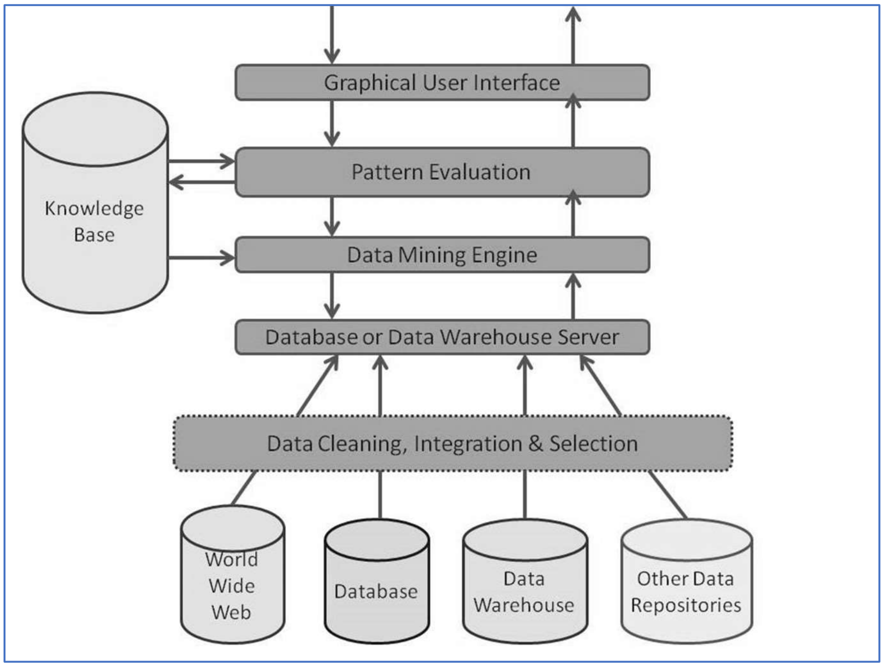
Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

Data Mining refers to the **detection and extraction** of new patterns from the already collected data. Data mining is the amalgamation of the field of statistics and computer science aiming to discover patterns in incredibly large datasets and then transforming them into a comprehensible structure for later use.

Data mining is a very important process where potentially useful and previously unknown information is extracted from large volumes of data. There are a number of components involved in the data mining process. These components constitute the architecture of a data mining system.

ARCHITECTURE OF DATA MINING

The major components of any data mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.



Data Source

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. A huge amount of historical data is needed for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases, text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

The data from multiple sources are integrated into a common source known as **Data Warehouse**.

1. ***Flat Files***
   - Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.
   - Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables.
   - Flat files are represented by data dictionary. Eg: CSV file.
   - **Application**: Used in DataWarehousing to store data, Used in carrying data to and from server, etc.
2. ***Relational Databases***
   - A Relational database is defined as the collection of data organized in tables with rows and columns.
   - Physical schema in Relational databases is a schema which defines the structure of tables.
   - Logical schema in Relational databases is a schema which defines the relationship among tables.
   - Standard API of relational database is SQL.
   - **Application**: Data Mining, ROLAP model, etc.
3. ***DataWarehouse***
   - A datawarehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.
   - There are three types of datawarehouse: **Enterprise** datawarehouse, **Data Mart** and **Virtual** Warehouse.
   - Two approaches can be used to update data in DataWarehouse: **Query-driven** Approach and **Update-driven** Approach.
   - **Application**: Business decision making, Data mining, etc.

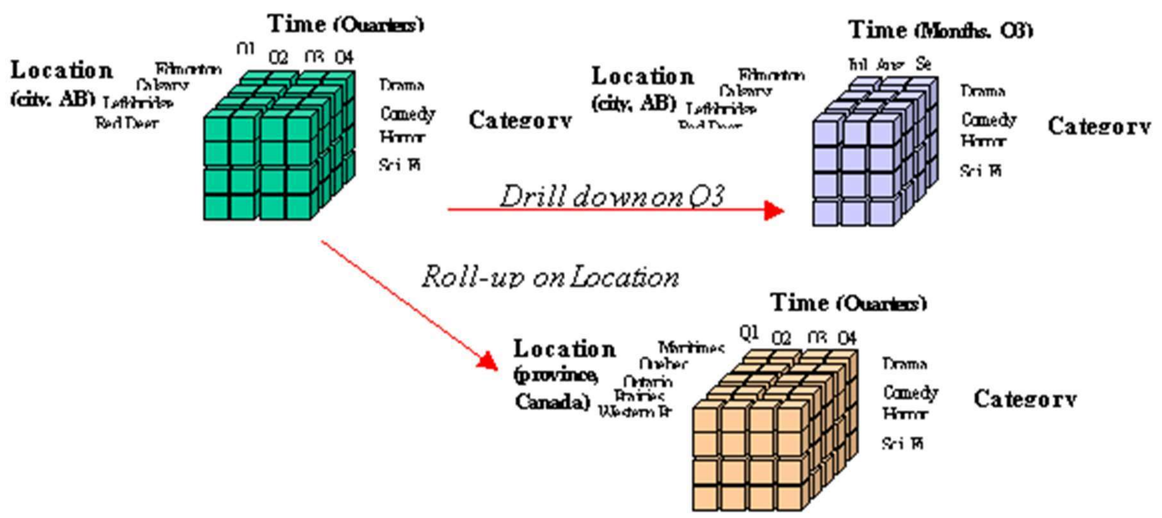


Figure 1.4: Summarized data from *Our VideoStore* before and after *drill-down* and *roll-up* operations.

4. ***Transactional Databases***
   - Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases.
   - This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
   - Highly flexible system where users can modify information without changing any sensitive information.
   - Follows ACID property (atomicity, Consistency, Isolation and Durability) of DBMS.
   - **Application**: Banking, Distributed systems, Object databases, etc.
5. ***Multimedia Databases***
   - Multimedia databases consists audio, video, images and text media.
   - They can be stored on Object-Oriented Databases.
   - They are used to store complex information in a pre-specified formats.
   - **Application**: Digital libraries, video-on demand, news-on demand, musical database, etc.
6. ***Spatial Database***
   - Store geographical information.
   - Stores data in the form of coordinates, topology, lines, polygons, etc.
   - **Application**: Maps, Global positioning, etc.

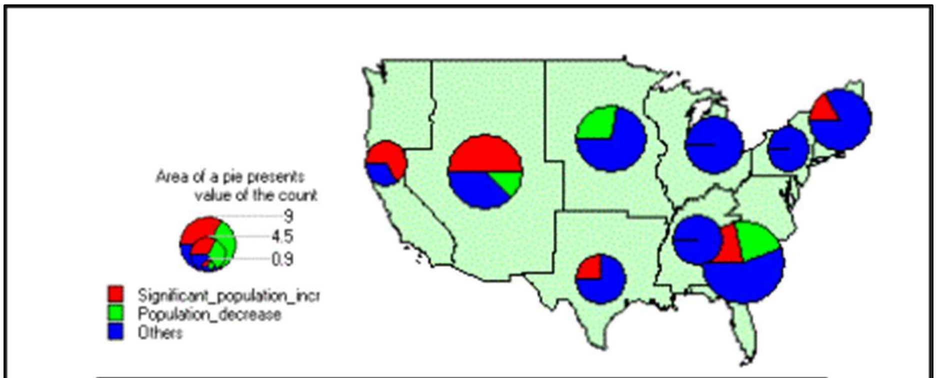


Figure 1.6: Visualization of spatial OLAP (from GeoMiner system)

7. ***Time-series Databases***

- Time series databases contains stock exchange data and user logged activities.
- Handles array of numbers indexed by time, date, etc.
- It requires real-time analysis.
- **Application**: eXtremeDB, Graphite, InfluxDB, etc.

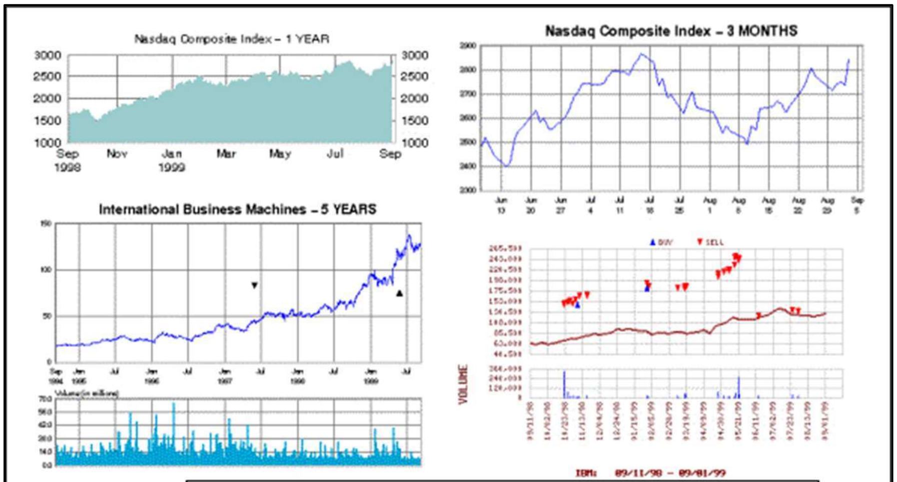


Figure 1.7: Examples of Time-Series Data (Source: Thompson Investors Group)

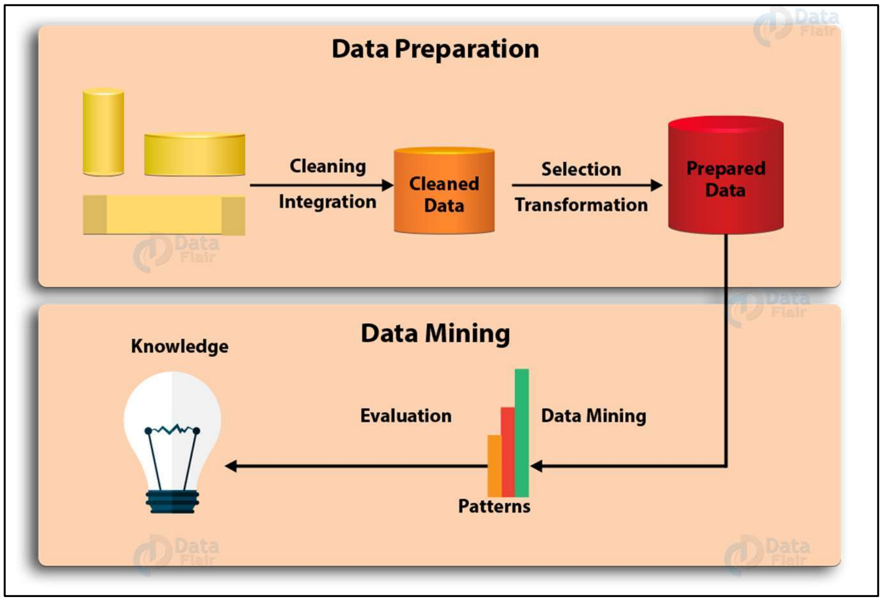
8. ***WWW***

- WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
- It is the most heterogeneous repository as it collects data from multiple resources.
- It is dynamic in nature as Volume of data is continuously increasing and changing.
- **Application**: Online shopping, Job search, Research, studying, etc.

Processes in Data Mining

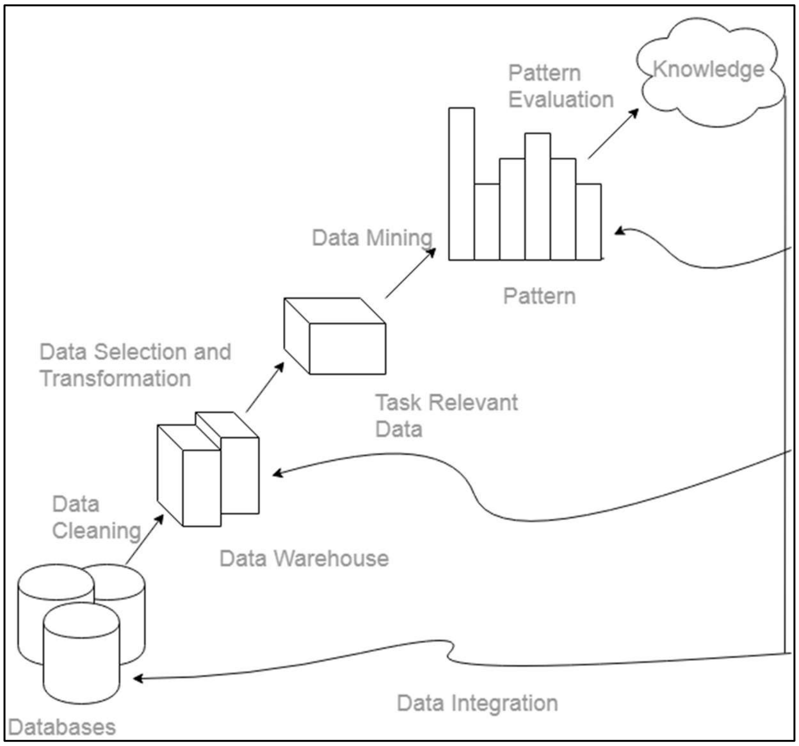
Before passing the data to the database or data warehouse server, the data must be cleaned, integrated, and selected. As the information comes from various sources and in different formats, it cannot be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources, and only the data of interest will have to be selected and passed to the server. Several methods may be performed on the data as part of selection, integration, and cleaning.

Data Mining Process is classified into two stages: Data preparation or data preprocessing and data mining



Data preparation process includes data cleaning, data integration, data selection and data transformation. Whereas the second phase includes data mining, pattern evaluation, and knowledge representation.

1. ***Data Cleaning***: Data cleaning is defined as removal of noisy and irrelevant data from collection.
   - Cleaning in case of ***Missing values***.
   - Cleaning ***noisy*** data, where noise is a random or variance error.
   - Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.



2. ***Data Integration***: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
   - Data integration using ***Data Migration tools***.
   - Data integration using ***Data Synchronization tools***.
   - Data integration using ***ETL***(Extract-Load-Transformation) process.
3. ***Data Selection***: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   - Data selection using ***Neural network***.
   - Data selection using ***Decision Trees***.
   - Data selection using ***Naive bayes***.
   - Data selection using ***Clustering***, ***Regression***, etc.
4. ***Data Transformation***: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.

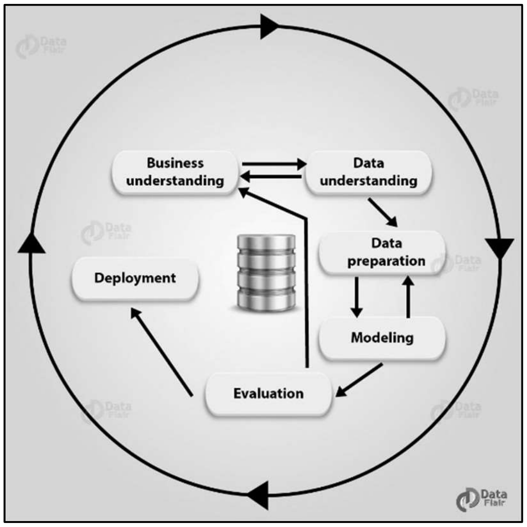
   Data Transformation is a two step process:
   - ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
   - ***Code generation***: Creation of the actual transformation program.

5. ***Data Mining***: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
    - Transforms task relevant data into ***patterns***.
    - Decides purpose of model using ***classification*** or ***characterization***.
6. ***Pattern Evaluation***: Pattern Evaluation is defined as as identifying strictly increasing patterns representing knowledge based on given measures.
    - Find ***interestingness score*** of each pattern.
    - Uses ***summarization*** and ***Visualization*** to make data understandable by user.
7. ***Knowledge representation***: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
    - Generate ***reports***.
    - Generate ***tables***.
    - Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.

Note:

- KDD is an ***iterative* process** where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.
- ***Preprocessing* of databases** consists of **Data cleaning** and **Data Integration**.

Cross-Industry Standard Process For Data Mining (CRISP-DM) - Cross-Industry Standard Process consists of six phases. Also, it’s a cyclical process.



DATABASE OR DATA WAREHOUSE SERVER

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

DATA MINING ENGINE

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, data mining is the root of data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

PATTERN EVALUATION MODULE

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns. This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

GRAPHICAL USER INTERFACE

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

KNOWLEDGE BASE

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.

TYPES OF DATA MINING ARCHITECTURE

1. **No Coupling**

   The no coupling data mining architecture retrieves data from particular data sources. It does not use the database for retrieving the data which is otherwise quite an efficient and accurate way to do the same. The no coupling architecture for data mining is poor and only used for performing very simple data mining processes.

2. **Loose Coupling**

   In loose coupling architecture data mining system retrieves data from the database and stores the data in those systems. This mining is for memory-based data mining architecture.

3. **Semi Tight Coupling**

   It tends to use various advantageous features of the data warehouse systems. It includes sorting, indexing, aggregation. In this architecture, an intermediate result can be stored in the database for better performance.

4. **Tight coupling**

   In this architecture, a data warehouse is considered as one of it's most important components whose features are employed for performing data mining tasks. This architecture provides scalability, performance, and integrated information

ADVANTAGES OF DATA MINING

- Assists in preventing future adversaries by accurately predicting future trends.
- Contributes to the making of important decisions.
- Compresses data into valuable information.
- Provides new trends and unexpected patterns.
- Helps to analyze huge data sets.
- Aids companies to find, attract and retain customers.
- Helps the company to improve its relationship with the customers.
- Assists Companies to optimize their production according to the likability of a certain product thus saving cost to the company.

DISADVANTAGES OF DATA MINING

- Excessive work intensity requires high-performance teams and staff training.
- The requirement of large investments can also be considered as a problem as sometimes data collection consumes many resources that suppose a high cost.
- Lack of security could also put the data at huge risk, as the data may contain private customer details.
- Inaccurate data may lead to the wrong output.
- Huge databases are quite difficult to manage.