Data Mining

Unit-1

What is Data Mining?

In general terms, **“Mining”** is the process of extraction of some valuable material from the earth e.g. coal mining, diamond mining etc. In the context of computer science, **“Data Mining”** refers to the extraction of useful information from a bulk of data.

Data mining is used in almost all the places where a large amount of data is stored and processed.

For example, banks typically use ‘data mining’ to find out their prospective customers who could be interested in credit cards, personal loans.

Main Purpose of Data Mining

Basically, the information gathered from Data Mining helps to predict hidden patterns, future trends and behaviours and allowing businesses to take decisions.

Data mining is the computational process of analyzing data from different perspective, dimensions, angles and categorizing/summarizing it into meaningful information.

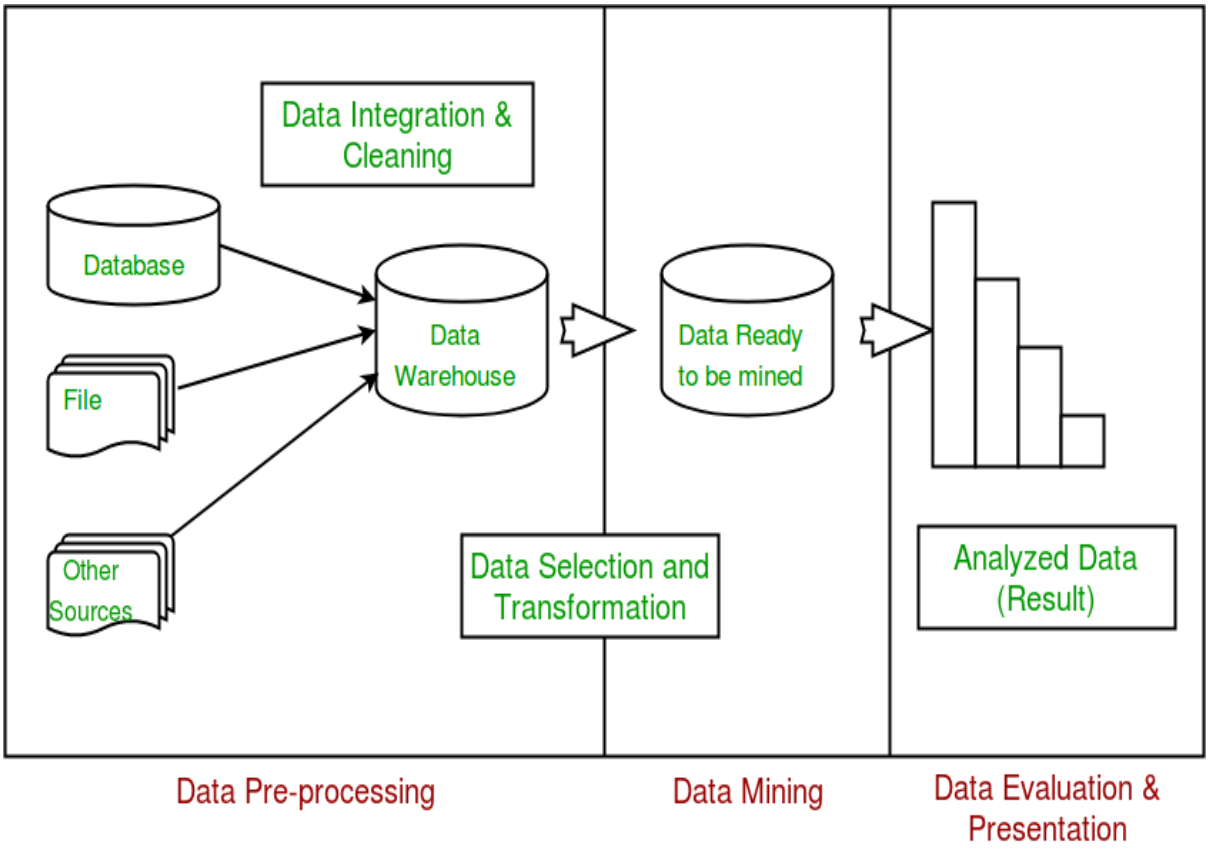
Data Mining Applications

Data mining is highly useful in the following domains −

- Market Analysis and Management
- Corporate Analysis & Risk Management
- Fraud Detection

Data Mining Process

The whole process of Data Mining comprises of three main phases:
1. Data Pre-processing – Data cleaning, integration, selection and transformation takes place
2. Data Extraction – Occurrence of exact data mining
3. Data Evaluation and Presentation – Analyzing and presenting results.



Applications of Data Mining

1. Financial Analysis
2. Biological Analysis
3. Scientific Analysis
4. Intrusion Detection
5. Fraud Detection
6. Research Analysis

Patterns used in data mining

1. Association

The items or objects in relational databases, transactional databases or any other information repositories are considered, while finding **associations or correlations.**

2. Classification

- The goal of classification is to construct a model with the help of historical data that can accurately predict the value.
- It maps the data into the predefined groups or classes and searches for the new patterns.

For example:
To predict weather on a particular day will be categorized into - sunny, rainy, or cloudy.

3. Regression

- Regression creates predictive models. Regression analysis is used to make predictions based on existing data by applying formulas.
- Regression is very useful for finding (or predicting) the information on the basis of previously known information.

4. Cluster analysis

- It is a process of portioning a set of data into a set of meaningful subclass, called as cluster.
- It is used to place the data elements into the related groups without advanced knowledge of the group definitions.

5. Forecasting

Forecasting is concerned with the discovery of knowledge or information patterns in data that can lead to reasonable predictions about the future.

Technologies used in data mining

Several techniques used in the development of data mining methods. **Some of them are mentioned below:**

1. Statistics:

- It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.
- Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

2. Machine learning

- **Arthur Samuel** defined machine learning as a field of study that gives computers the ability to learn without being programmed.
- When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.
- In machine learning, an algorithm is constructed to predict the data from the available database **(Predictive analysis).**
- It is related to computational statistics.

3. Information retrieval

Information deals with uncertain representations of the semantics of objects (text, images).
For example: Finding relevant information from a large document.

4. Database systems and data warehouse

- Databases are used for the purpose of recording the data as well as data warehousing.
- Online Transactional Processing (OLTP) uses databases for day to day transaction
  purpose.
- To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.
- **E**ntity-**R**elational modelling techniques are used for relational database management system design.

- Data warehouses are used to store historical data which helps to take strategically decision for business.
- It is used for online analytical processing (OALP), which helps to analyze the data.

5. Decision support system

- Decision support system is a category of information system. It is very useful in decision making for organizations.
- It is an interactive software based system which helps decision makers to extract useful information from the data, documents to make the decision.

Major Issue in Data Mining

Data mining systems face a lot of challenges and issues in today's world some of them are:

1 Mining methodology and user interaction issues

2 Performance issues

3 Issues relating to the diversity of database types

1 Mining methodology and user interaction issues:

Mining different kinds of knowledge in databases:
Different user - different knowledge - different way.That means different client want a different kind of information so it becomes difficult to cover vast range of data that can meet the client requirement.

Interactive mining of knowledge at multiple levels of abstraction:
Interactive mining allows users to focus the search for patterns from different angles.The data mining process should be interactive because it is difficult to know what can be discovered within a database.

Incorporation of background knowledge:
Background knowledge is used to guide discovery process and to express the discovered patterns.

Query languages and ad hoc mining:
Relational query languages (such as SQL) allow users to pose ad-hoc queries for data retrieval.The language of data mining query language should be in perfectly matched with the query language of data warehouse.

Handling noisy or incomplete data:
In a large database, many of the attribute values will be incorrect.This may be due to human error or because of any instruments fail. Data cleaning methods and data analysis methods are used to handle noise data.

2 Performance issues

Efficiency and scalability of data mining algorithms:
To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable.

Parallel, distributed, and incremental mining algorithms:
The huge size of many databases, the wide distribution of data, and complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel.

3 Issues relating to the diversity of database types:

Handling of relational and complex types of data:
There are many kinds of data stored in databases and data warehouses. It is not possible for one system to mine all these kind of data.So different data mining system should be construed for different kinds data.

Mining information from heterogeneous databases and global information systems:
Since data is fetched from different data sources on Local Area Network (LAN) and Wide Area Network (WAN).The discovery of knowledge from different sources of structured is a great challenge to data mining.