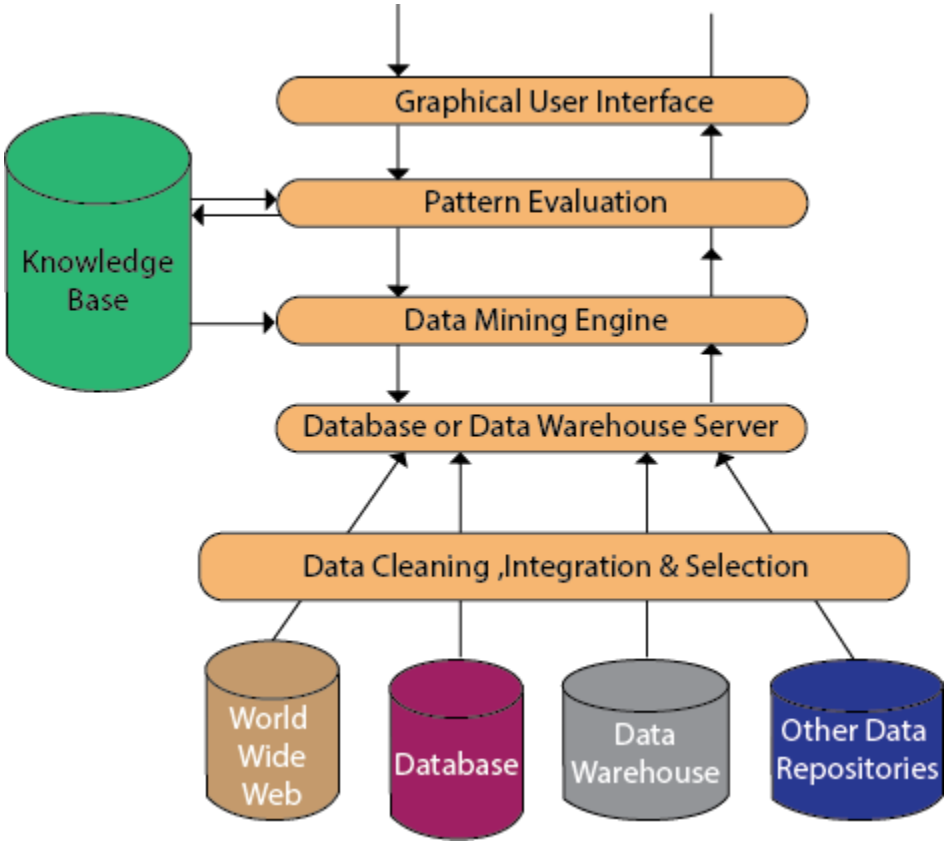
Data mining is a significant method where previously unknown and potentially useful information is extracted from the vast amount of data. The data mining process involves several components, and these components constitute a data mining system architecture.

Data Mining Architecture

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the 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. You need a huge amount of historical data 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.

Different processes:

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 can't 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. These procedures are not as easy as we think. Several methods may be performed on the data as part of selection, integration, and cleaning.

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, we can say data mining is the root of our 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

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis

- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Data Mining tools

Data Mining is the set of techniques that utilize specific algorithms, statical analysis, artificial intelligence, and database systems to analyze data from different dimensions and perspectives.

Data Mining tools have the objective of discovering patterns/trends/groupings among large sets of data and transforming data into more refined information.

We can perform various algorithms such as clustering or classification on your data set and visualize the results itself. It is a framework that provides us better insights for our data and the phenomenon that data represent. Such a framework is called a data mining tool.

These are the most popular data mining tools:

1. Orange Data Mining:

Orange is a perfect machine learning and data mining software suite. It supports the visualization and is a software-based on components written in Python computing language and developed at the bioinformatics laboratory at the faculty of computer and information science.As it is a software-based on components, the components of Orange are called "widgets." These widgets range from preprocessing and data visualization to the assessment of algorithms and predictive modeling.

Widgets deliver significant functionalities such as:

- Displaying data table and allowing to select features
- Data reading
- Training predictors and comparison of learning algorithms
- Data element visualization, etc.

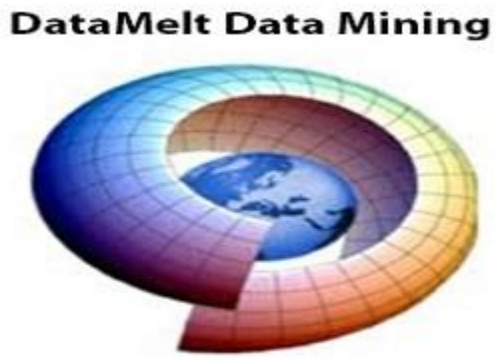
The instrument has machine learning components, add-ons for bioinformatics and text mining, and it is packed with features for data analytics. This is also used as a python library.

2. SAS Data Mining:

SAS stands for Statistical Analysis System. It is a product of the SAS Institute created for analytics and data management. SAS can mine data, change it, manage information from various sources, and analyze statistics. It offers a graphical UI for non-technical users.

SAS data miner allows users to analyze big data and provide accurate insight for timely decision-making purposes. SAS has distributed memory processing architecture that is highly scalable. It is suitable for data mining, optimization, and text mining purposes.

3. DataMelt Data Mining:



DataMelt is a computation and visualization environment which offers an interactive structure for data analysis and visualization. It is primarily designed for students, engineers, and scientists. It is also known as DMelt.

DMelt is a multi-platform utility written in JAVA. It can run on any operating system which is compatible with JVM (Java Virtual Machine). It consists of Science and mathematics libraries.

- **Scientific libraries:**
  Scientific libraries are used for drawing the 2D/3D plots.
- **Mathematical libraries:**
  Mathematical libraries are used for random number generation, algorithms, curve fitting, etc.

DMelt can be used for the analysis of the large volume of data, data mining, and statistical analysis. It is extensively used in natural sciences, financial markets, and engineering.

4. Rattle:

Ratte is a data mining tool based on GUI. It uses the R stats programming language. Rattle exposes the statical power of R by offering significant data mining features. While rattle has a comprehensive and well-developed user interface, It has an integrated log code tab that produces duplicate code for any GUI operation.

The data set produced by Rattle can be viewed and edited. Rattle gives the other facility to review the code, use it for many purposes, and extend the code without any restriction.

5. Rapid Miner:

Rapid Miner is one of the most popular predictive analysis systems created by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It offers an integrated environment for text mining, deep learning, machine learning, and predictive analysis.

The instrument can be used for a wide range of applications, including company applications, commercial applications, research, education, training, application development, machine learning.

Rapid Miner provides the server on-site as well as in public or private cloud infrastructure. It has a client/server model as its base. A rapid miner comes with template-based frameworks that enable fast delivery with few errors(which are commonly expected in the manual coding writing process)