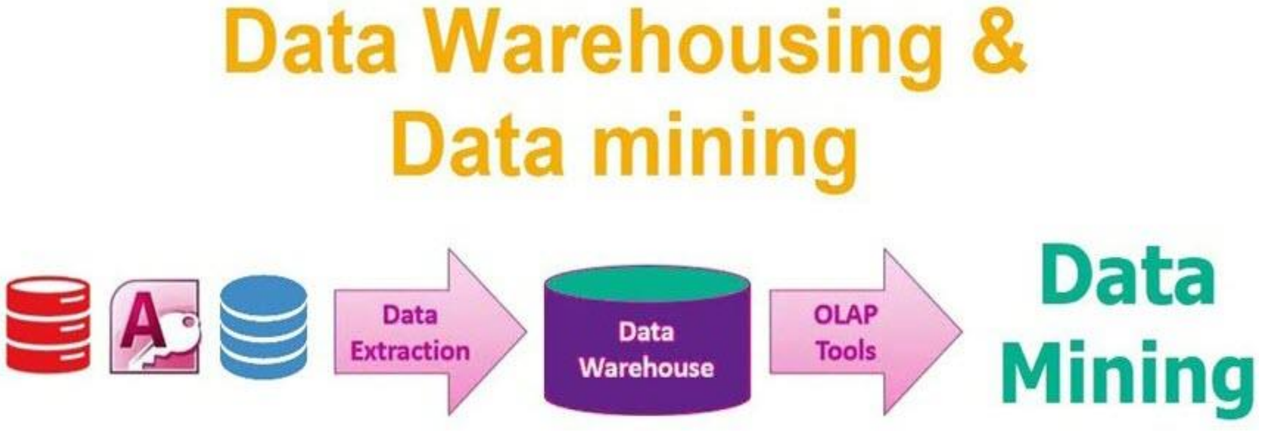
# Data Warehouse and Data Mining

# CSL608

# Unit-1

## Data Warehouse

- A data warehouse is a technique for collecting and managing data from varied sources to provide meaningful business insights.
- It is a blend of technologies and components which allows the strategic use of data.
- Data Warehouse is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing.
- It is a process of transforming data into information and making it available to users for analysis.
- A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights.
- The huge amount of data comes from multiple places such as Marketing and Finance.
- The extracted data is utilized for analytical purposes and helps in decision- making for a business organization.
- The data warehouse is designed for the analysis of data rather than transaction processing.



## Data Mining

- Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called *Knowledge Discovery in Database (KDD)*. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

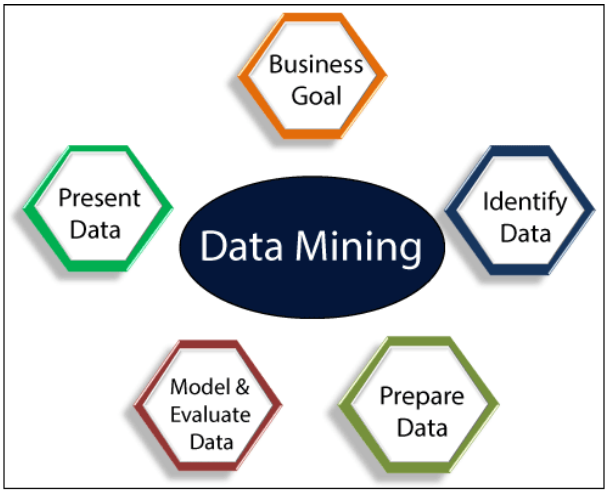
The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

In other words, we can say that Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective. This process includes various types of services such as text mining, web mining, audio and video mining, pictorial data mining, and social media mining. It is done through software that is simple or highly specific. By outsourcing data mining, all the work can be done faster with low operation costs. Specialized firms can also use new technologies to collect data that is impossible to locate manually. There are tonnes of information available on various platforms, but very little knowledge is accessible. The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.



## Types of Data Mining

Data mining can be performed on the following types of data:

**Relational Database:**

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

**Data warehouses:**

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

**Data Repositories:**

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

**Object-Relational Database:**

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

**Transactional Database:**

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

## Advantages of Data Mining

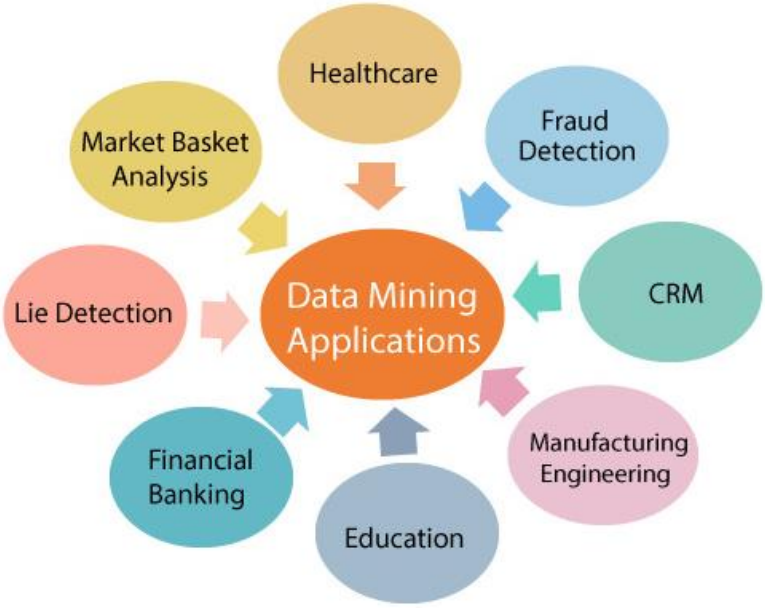
- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

## Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

# Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.



These are the following areas where data mining is widely used:

**Data Mining in Healthcare:**

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

**Data Mining in Market Basket Analysis:**

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

**Data mining in Education:**

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

### Data Mining in Manufacturing Engineering:

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

### Data Mining in CRM (Customer Relationship Management):

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

### Data Mining in Fraud detection:

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

### Data Mining in Lie Detection:

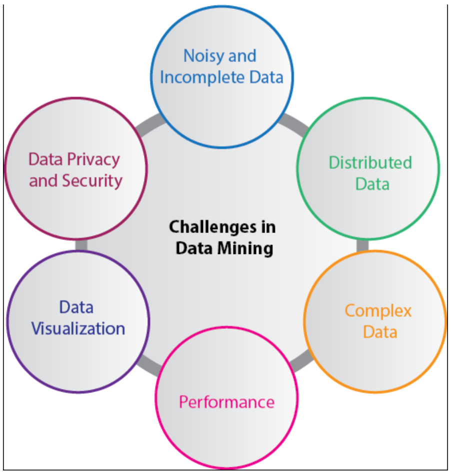
Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

### Data Mining Financial Banking:

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

## Challenges of Implementation in Data mining

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods, and techniques, etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.


Noisy and Incomplete Data
Data Privacy and Security
Distributed Data
Challenges in Data Mining
Data Visualization
Complex Data
Performance

**Incomplete and noisy data:**

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than $ 500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data. Even some customers may not be willing to disclose their phone numbers, which results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data)makes data mining challenging.

**Data Distribution:**

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, It is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

**Complex Data:**

Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

**Performance:**

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

**Data Privacy and Security:**

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

**Data Visualization:**

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

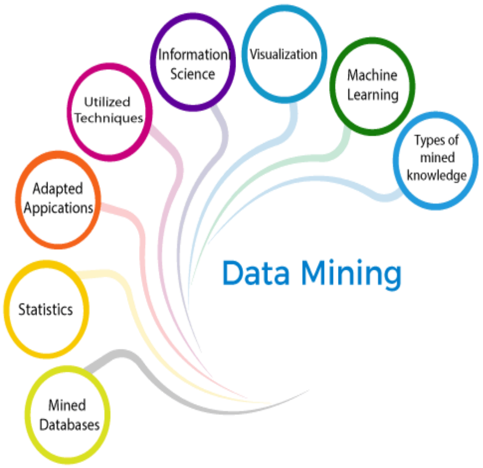
# Classification of Data Mining Systems

Data mining refers to the process of extracting important data from raw data. It analyses the data patterns in huge sets of data with the help of several software. Ever since the development of data mining, it is being incorporated by researchers in the research and development field.

With Data mining, businesses are found to gain more profit. It has not only helped in understanding customer demand but also in developing effective strategies to enforce overall business turnover. It has helped in determining business objectives for making clear decisions.

Data collection and data warehousing, and computer processing are some of the strongest pillars of data mining. Data mining utilizes the concept of mathematical algorithms to segment the data and assess the possibility of occurrence of future events.

To understand the system and meet the desired requirements, data mining can be classified into the following systems:



- Classification based on the mined Databases
- Classification based on the type of mined knowledge
- Classification based on statistics
- Classification based on Machine Learning
- Classification based on visualization
- Classification based on Information Science
- Classification based on utilized techniques
- Classification based on adapted applications

**Classification Based on the mined Databases**

A data mining system can be classified based on the types of databases that have been mined. A database system can be further segmented based on distinct principles, such as data models, types of data, etc., which further assist in classifying a data mining system.

For example, if we want to classify a database based on the data model, we need to select either relational, transactional, object-relational or data warehouse mining systems.

**Classification Based on the type of Knowledge Mined**

A data mining system categorized based on the kind of knowledge mind may have the following functionalities:

1. Characterization

2. Discrimination
3. Association and Correlation Analysis
4. Classification
5. Prediction
6. Outlier Analysis
7. Evolution Analysis

**Classification Based on the Techniques Utilized**

A data mining system can also be classified based on the type of techniques that are being incorporated. These techniques can be assessed based on the involvement of user interaction involved or the methods of analysis employed.

**Classification Based on the Applications Adapted**

Data mining systems classified based on adapted applications adapted are as follows:

1. Finance
2. Telecommunications
3. DNA
4. Stock Markets
5. E-mail

**Examples of Classification Task**

Following is some of the main examples of classification tasks:

- Classification helps in determining tumor cells as benign or malignant.
- Classification of credit card transactions as fraudulent or legitimate.
- Classification of secondary structures of protein as alpha-helix, beta-sheet, or random coil.
- Classification of news stories into distinct categories such as finance, weather, entertainment, sports, etc.

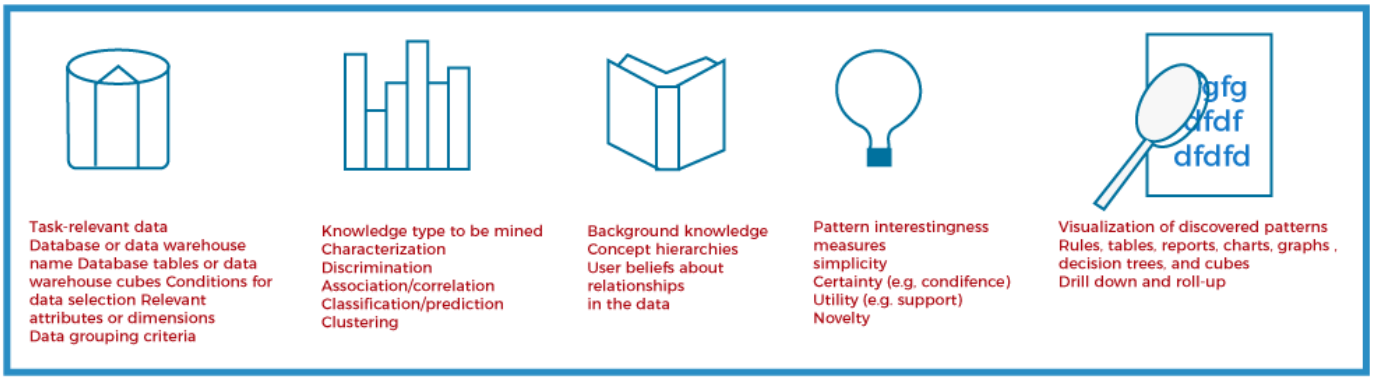
## Data Mining Task Primitives

A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process or examine the findings from different angles or depths. The data mining primitives specify the following,

1. Set of task-relevant data to be mined.
2. Kind of knowledge to be mined.
3. Background knowledge to be used in the discovery process.
4. Interestingness measures and thresholds for pattern evaluation.
5. Representation for visualizing the discovered patterns.

A data mining query language can be designed to incorporate these primitives, allowing users to interact with data mining systems flexibly. Having a data mining query language provides a foundation on which user-friendly graphical interfaces can be built.

Designing a comprehensive data mining language is challenging because data mining covers a wide spectrum of tasks, from data characterization to evolution analysis. Each task has different requirements. The design of an effective data mining query language requires a deep understanding of the power, limitation, and underlying mechanisms of the various kinds of data mining tasks. This facilitates a data mining system's communication with other information systems and integrates with the overall information processing environment.


gfg
dfdf
dfdfd
Task-relevant data
Database or data warehouse name Database tables or data warehouse cubes Conditions for data selection Relevant attributes or dimensions
Data grouping criteria
Knowledge type to be mined
Characterization
Discrimination
Association/correlation
Classification/prediction
Clustering
Background knowledge
Concept hierarchies
User beliefs about relationships in the data
Pattern interestingness measures
simplicity
Certainty (e.g. condifence)
Utility (e.g. support)
Novelty
Visualization of discovered patterns
Rules, tables, reports, charts, graphs , decision trees, and cubes
Drill down and roll-up

## List of Data Mining Task Primitives

A data mining query is defined in terms of the following primitives, such as:

**1. The set of task-relevant data to be mined**

This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (the relevant attributes or dimensions).

In a relational database, the set of task-relevant data can be collected via a relational query involving operations like selection, projection, join, and aggregation.

The data collection process results in a new data relational called the ***initial data relation***. The initial data relation can be ordered or grouped according to the conditions specified in the query. This data retrieval can be thought of as a subtask of the data mining task.

This initial relation may or may not correspond to physical relation in the database. Since virtual relations are called Views in the field of databases, the set of task-relevant data for data mining is called a minable view.

**2. The kind of knowledge to be mined**

This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

**3. The background knowledge to be used in the discovery process**

This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allows data to be mined at multiple levels of abstraction.

Concept hierarchy defines a sequence of mappings from low-level concepts to higher-level, more general concepts.

- **Rolling Up - Generalization of data:** Allow to view data at more meaningful and explicit abstractions and makes it easier to understand. It compresses the data, and it would require fewer input/output operations.

- **Drilling Down - Specialization of data:** Concept values replaced by lower-level concepts. Based on different user viewpoints, there may be more than one concept hierarchy for a given attribute or dimension.

An example of a concept hierarchy for the attribute (or dimension) age is shown below. User beliefs regarding relationships in the data are another form of background knowledge.

**4. The interestingness measures and thresholds for pattern evaluation**

Different kinds of knowledge may have different interesting measures. They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. For example, interesting measures for association rules include support and confidence. Rules whose support and confidence values are below user-specified thresholds are considered uninteresting.

- **Simplicity:** A factor contributing to the interestingness of a pattern is the pattern's overall simplicity for human comprehension. For example, the more complex the structure of a rule is, the more difficult it is to interpret, and hence, the less interesting it is likely to be. Objective measures of pattern simplicity can be viewed as functions of the pattern structure, defined in terms of the pattern size in bits or the number of attributes or operators appearing in the pattern.
- **Certainty (Confidence):** Each discovered pattern should have a measure of certainty associated with it that assesses the validity or "trustworthiness" of the pattern. A certainty measure for association rules of the form "A =>B" where A and B are sets of items is confidence. Confidence is a certainty measure. Given a set of task-relevant data tuples, the confidence of "A => B" is defined as Confidence (A=>B) = # tuples containing both A and B /# tuples containing A
- **Utility (Support):** The potential usefulness of a pattern is a factor defining its interestingness. It can be estimated by a utility function, such as support. The support of an association pattern refers to the percentage of task-relevant data tuples (or transactions) for which the pattern is true. Utility (support): usefulness of a pattern Support (A=>B) = # tuples containing both A and B / total #of tuples
- **Novelty:** Novel patterns are those that contribute new information or increased performance to the given pattern set. For example -> A data exception. Another strategy for detecting novelty is to remove redundant patterns.

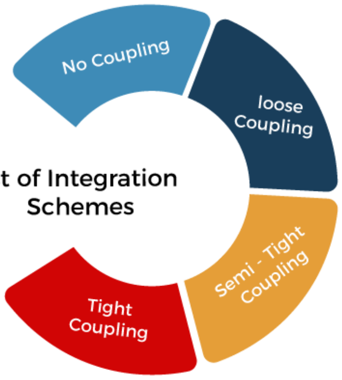
**5. The expected representation for visualizing the discovered patterns**

This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, cross tabs, charts, graphs, decision trees, cubes, or other visual representations.

Users must be able to specify the forms of presentation to be used for displaying the discovered patterns. Some representation forms may be better suited than others for particular kinds of knowledge.

For example, generalized relations and their corresponding cross tabs or pie/bar charts are good for presenting characteristic descriptions, whereas decision trees are common for classification.

# Integration schemes of Database and Data warehouse systems



### No Coupling

In no coupling schema, the data mining system does not use any database or data warehouse system functions.

### Loose Coupling

In loose coupling, data mining utilizes some of the database or data warehouse system functionalities. It mainly fetches the data from the data repository managed by these systems and then performs data mining. The results are kept either in the file or any designated place in the database or data warehouse.

### Semi-Tight Coupling

In semi-tight coupling, data mining is linked to either the DB or DW system and provides an efficient implementation of data mining primitives within the database.

### Tight Coupling

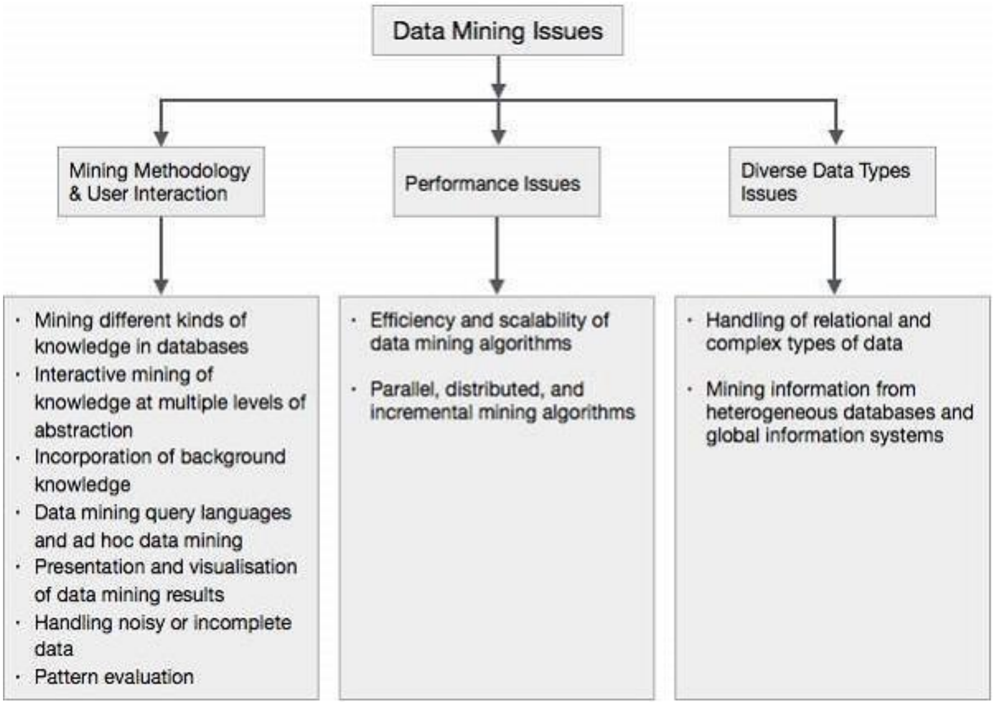
A data mining system can be effortlessly combined with a database or data warehouse system in tight coupling.

## major issues in data mining

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.


Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types Issues
· Mining different kinds of knowledge in databases
· Interactive mining of knowledge at multiple levels of abstraction
· Incorporation of background knowledge
· Data mining query languages and ad hoc data mining
· Presentation and visualisation of data mining results
· Handling noisy or incomplete data
· Pattern evaluation
· Efficiency and scalability of data mining algorithms
· Parallel, distributed, and incremental mining algorithms
· Handling of relational and complex types of data
· Mining information from heterogeneous databases and global information systems

## 1.Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

## 2. Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.

- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

## 3. Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.