Unit-1

Introduction

During the phase of Web 1.0, the web applications were only read-only type. No companies were making profit out of it.

Web 2.0 (present) is Read-Write type. That is, User can interact with web applications and provide data.

Examples of User data are Login time, Personalized selections from various options, personal info (user location, age, hobbies), etc.

With this user data, companies were able to maximize their profits by providing brand/product recommendations to users based on their previous choices, etc.

Now, every national and multi-national company has tons of user data. But not all/every user data is useful for the companies.

Hence, the raw data must be filtered, so that the useless data is filtered out and useful data is fetched.

Hence, the Process of extracting the useful data from the raw collections of data is known as Data Mining.

For this extraction process, various types of techniques, statistics and machine learning algorithms are applied, and advanced scientific tools are used.

The mined data can be used to create patterns (a set of correlated data), trends (popular choices), and valuable insights.

Data mining is often used in a wide range of applications, including business intelligence, healthcare, finance, marketing, and more.

Definition:

> Data mining is the process of extracting useful data from the raw collections of data sets to discover patterns, trends, and valuable insights from it, which can be used in various applications such as business intelligence.
>
> Data mining is performed by applying various kinds of techniques such as statistical analysis, machine learning algorithms, and pattern recognition, to uncover hidden information and make insightful decisions.

Essential steps involved in the process of knowledge discovery from the raw datasets, are:

- Data Cleaning: removing noise or irrelevant data.
- Data integration: combining multiple data sources.
- Data Selection: fetching relevant data from database.
- Data transformation: transforming data into appropriate forms which can be used to perform summary and aggregation operations.
- Data mining: an essential process where intelligent methods are applied to extract data patterns and evaluate to identify the interesting patterns representing knowledge based on some interestingness measures.
- Knowledge Presentation: visualizing the interesting patterns (useful knowledge)

Types of Data (What Kinds of Data Can Be Mined?):

Data mining can be applied to any kind of data if the data is meaningful for a target application.

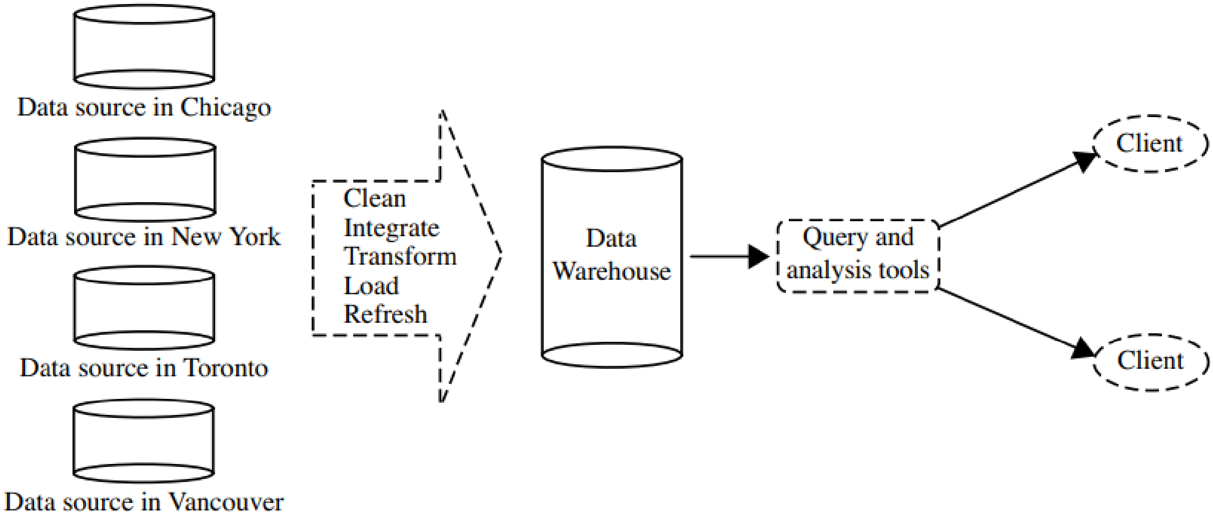
The most basic forms of data for mining applications are:

1. Database Data
2. Data-Warehouse Data
3. Transactional Data

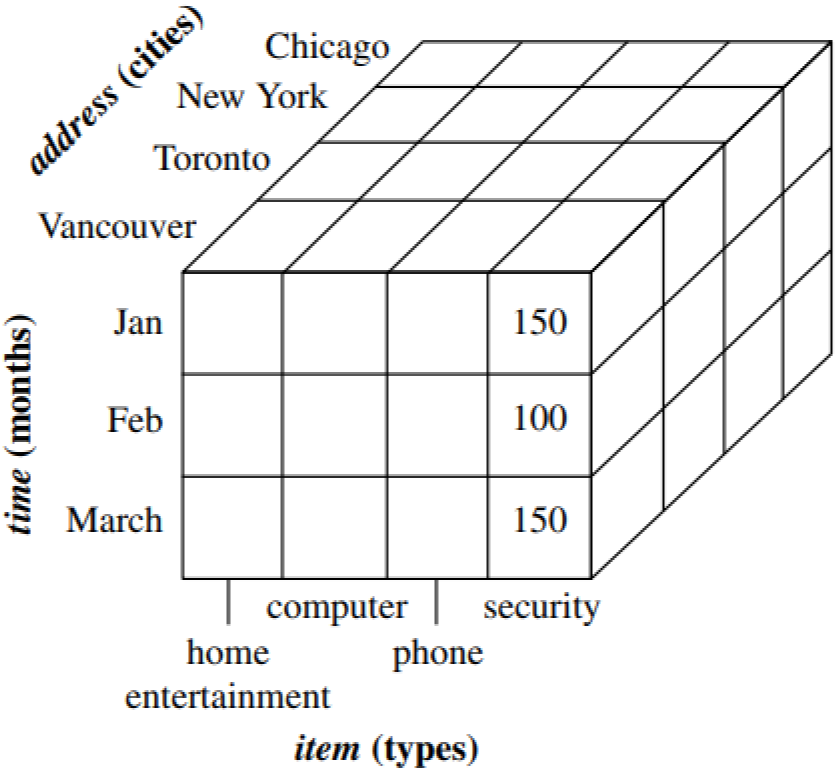
1) Database Data (Relational Database):

- Database data is structured and stored in a relational database management system (RDBMS).
- RDBMS contains tables with rows (known as tuples/records) and columns (known as attributes/fields) for structured data.
- Each table represents a collection of related data, and each row is an individual record.
- Attributes in tables describe the characteristics of entities.
- The use of semantic data models, such as the entity-relationship (ER) model, helps represent the database's entities and their relationships.
- Relational data can be queried using relational query languages like SQL and can include aggregate functions for summarizing data.
- Hence, Applying Data mining techniques on relational databases can identify trends and patterns, such as credit risk assessment, price changes, etc.

2) Data Warehouses:

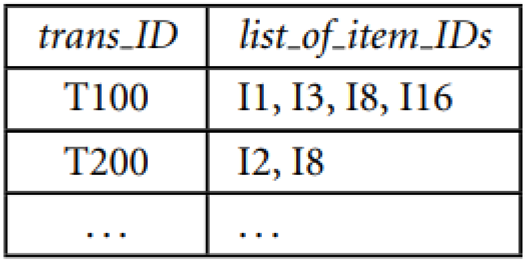


- A data warehouse is a large centralized and integrated repository of data collected from various sources within an organization.
- It is specifically designed for the purpose of enabling data analysis and reporting.
- Data warehouses store historical data and are structured in a way that makes it easy to retrieve and analyze information for business intelligence and decision-making purposes.
- Data mining tools are often used for more in-depth analysis of data warehouses.
- Data in a data warehouse is organized around major subjects and is summarized, making it easier for decision-making.
- Multidimensional data structures, such as data cubes, are used in data warehouses to provide a multidimensional view of data and allow for efficient access to summarized data.



- Hence, Data mining can be performed on data warehouses to extract insights and knowledge.

3) Transactional Data:



trans_ID	*list_of_item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
...	...

- Transactional data captures the records of individual transactions, such as customer purchases, financial transactions, online interactions, and user behavior.
- A transaction typically includes a unique transaction identity number (trans ID) and a list of items involved in the transaction.
- Each record in a transactional database represents a specific transaction, such as a purchase, flight booking, or web page interaction.
- Applying Data mining on transactional data can discover frequent item sets, helping identify items commonly purchased together (market basket analysis).
- Transactional data can be analyzed for patterns and associations, like discovering frequently co-purchased items (market basket analysis).

Functionalities of Data Mining:

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

- There are five major functionalities in data mining: characterization, discrimination, finding frequent patterns, classification, and clustering.

- Data mining can be used for both descriptive (past data) and predictive (future data) tasks, depending on the goal.

- Descriptive mining tasks involve analyzing the data in a database to understand its general properties and characteristics. This includes summarizing the data, identifying patterns, and gaining insights into its distribution and relationships.

- Descriptive mining aims to provide a comprehensive overview of the data without making any predictions or inferences about future outcomes.

- Predictive mining tasks, on the other hand, focus on using the current data to make predictions or inferences about future outcomes, by applying statistical and machine learning techniques

(1) Class/Concept Description: (Characterization and Discrimination)

- In data mining, you often have data associated with classes or concepts (e.g., customer classes or product categories).

- Data Characterization and Data Discrimination are two approaches used in data analysis to understand and describe a given set of data.

- Characterization involves classifying similar data (creating datasets (classes) such that the data in dataset share common properties) from mined data.

- Characterization results in concise descriptions of the class.

- Discrimination involves comparing the features of one class against other classes, aiming to highlight the differences.

- Discrimination identifies characteristics that distinguish the target class from others.

- Both characterization and discrimination use techniques like data summarization, attribute generalization, and attribute discrimination.

- The outcomes can be presented through various visual forms, including charts, tables, or rules.

(2) Mining Frequent Patterns, Associations, and Correlations:

- Frequent patterns refer to sets of items or events that frequently occur together in a dataset.

I.e., recurring patterns.

- It involves finding combinations of items that appear together more often than expected by chance.

Ex: In an e-com platform, it is found that most of the customers' combined/paired purchases are "bread and milk".

- Association mining is the process of finding associations or relationships among items or events in a dataset.

- It discovers rules that describe the dependencies between different items or events.

Ex: in a supermarket dataset, association mining can uncover rules like "If a customer buys bread and eggs, they are likely to also purchase butter."

- Association rules have support (how often the pattern occurs) and confidence (how likely the rule is true).

- Correlation mining is the process of measuring the association (statistical relationships or dependencies) between two different variables in a dataset.

- It measures how changes in one variable relate to changes in another variable.

- A correlation coefficient (r) is a statistic used for measuring the strength of a linear association between two variables.

- A Correlation Coefficient value (r) ranges from -1 (to) 1.

- (r = -1) indicates a perfect negative relationship in which high values of one variable are perfectly related to low values in the other variables, and conversely, low values of one variable are perfectly related to high values in the other variable.

- (r = 0) indicates no relationship between the two variables. That is, one cannot use the scores on one variable to tell anything about the scores on the second variable.

- (r = 1) indicates a perfect positive relationship in which high values of one variable are perfectly related to high values in the other variable, and conversely, low values of one variable are perfectly related to low values in the other variable.

Ex: in a marketing campaign dataset, correlation mining can reveal if there is a correlation between the amount spent on advertising and the increase in sales.

(3) Classification and Regression for Predictive Analysis:

- Classification is the process of predicting the class labels for the input data to categorize the data into classes.

- This can be done by building models. A classification model acts as a decision tree which separates the provided input data and gives each group of data (class) a label.

- These models are derived and can be built by providing labeled training data.

Hence, classification models are used to predict class labels for new, unlabeled data.

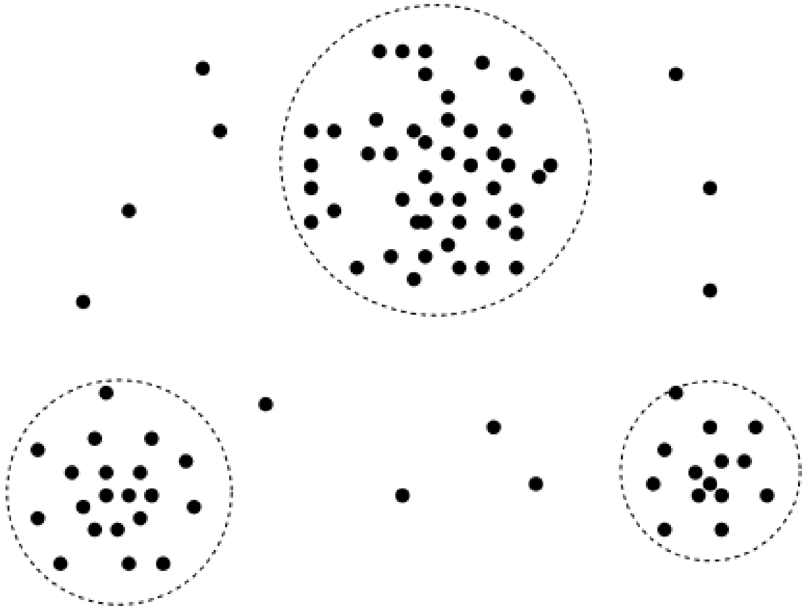
Ex: in email spam detection, a classification model can be trained on labeled emails (spam or not spam) to predict the class of incoming emails.

- Regression, on the other hand, is the process of predicting continuous numerical values based on the provided input features.

- It aims to find a mathematical relationship between the input variables and the output variable.

Ex: in house price prediction, regression can be used to build a model that predicts the price of a house based on factors like its area, number of rooms, location, etc.

(4) Cluster Analysis:



- Cluster analysis is a data mining technique used to group similar data points together based on their characteristics or attributes.

- It helps discover natural groupings or clusters in the data.

- The goal is to maximize similarity within clusters and minimize similarity between different clusters.

- Cluster analysis is used to create class labels for data classification, when they are not available.

- Clusters can be visualized to understand data patterns and relationships better.

Ex: Imagine you have a dataset of customer data, including information such as age, income, and purchasing behavior. Cluster analysis can be used to group customers with similar characteristics into clusters, such as creating a cluster for young, high-income customers who frequently make online purchases

(5) Outlier Analysis:

- Outlier analysis focuses on detecting and identifying data points that are significantly different from the majority of the data (outliers).

- In some cases, they can be more interesting than the regular data points.

- Outlier analysis is essential for detecting rare and unusual events, like fraud or anomalies.

- Outliers can be identified using statistical tests, distance measures, or density-based methods.

- These methods help find data points that stand out or are different from the norm.

Interestingness Patterns:

A data mining system has the potential to generate large no.of patterns. But we only need the application specific patterns, which are known as interesting patterns.

Therefore, A Pattern is said to be interesting <=> the pattern is:

- Human understandable.
- Valid on new or test data.
- Application specific (Potentially useful)

Other factors that are considered for identifying the interesting patterns, are:

-> Support (a measure of interestingness): The pattern should occur frequently enough in the dataset to be considered interesting. High support indicates that the pattern is not an isolated occurrence.

-> Confidence (a measure of interestingness): The pattern should have a high level of confidence or accuracy, indicating that the observed relationships are reliable and not due to chance.

-> Novelty: The pattern should reveal new or previously unknown information, providing insights that were not evident before.

-> Actionability: The pattern should be actionable, meaning it can be utilized to make informed decisions or drive actions that lead to desirable outcomes.

-> Interpretability: The pattern should be easily interpretable and understandable by domain experts or end-users.

Generation of only interesting patterns can make this process more efficient, but it is more challenging than finding the interesting patterns.

Data Mining Systems and Generating Interesting Patterns:

● Data mining systems can generate patterns based on predefined measures of interestingness, such as support and confidence thresholds.

● However, it is not feasible for data mining systems to generate all possible patterns from a large dataset due to computational constraints and the potential overwhelming number of patterns.

● Instead, data mining systems focus on generating patterns that surpass a certain threshold of interestingness, as defined by the user or domain expert.

Ex: if confidence < 50%, pattern is interesting.

If confidence >= 50%, patter is uninteresting.

Classification of Data Mining Systems:

Data mining systems can be classified based on various criteria. Here are some common classification approaches:

Classification according to the type of data source mined:

This classification categorizes data mining systems based on the type of data they handle, such as spatial data, multimedia data, time-series data, text data, or data from the World Wide Web.

Ex: A data mining system designed specifically for analyzing social media data would fall into the category of text data mining systems.

Classification according to the data model used:

This classification categorizes data mining systems based on the data model they employ, such as relational databases, object-oriented databases, data warehouses, or transactional data.

Ex: A data mining system that focuses on analyzing data stored in a data warehouse would belong to the data warehouse-oriented classification.

Classification according to the type of knowledge discovered:

This classification categorizes data mining systems based on the kind of knowledge they uncover or the data mining functionalities they provide, such as characterization, discrimination, association, classification, clustering, etc.

Ex: A data mining system that specializes in discovering associations between items in a transactional database would fall into the association mining category.

Classification according to the mining techniques used:

Data mining systems employ various techniques for analyzing data. This classification categorizes them based on the data analysis approaches they utilize, such as machine learning, neural networks, genetic algorithms, statistics, visualization, or database oriented and data warehouse-oriented methods.

Ex: A data mining system that employs machine learning algorithms to predict Customer churn in a telecommunications dataset would belong to the machine learning based classification.

Classification according to the degree of user interaction:

This classification considers the level of user involvement in the data mining process, such as query-driven systems where users actively specify mining tasks, interactive exploratory systems where users interactively explore the data and analyze patterns, or autonomous systems that automate the mining process without much user intervention.

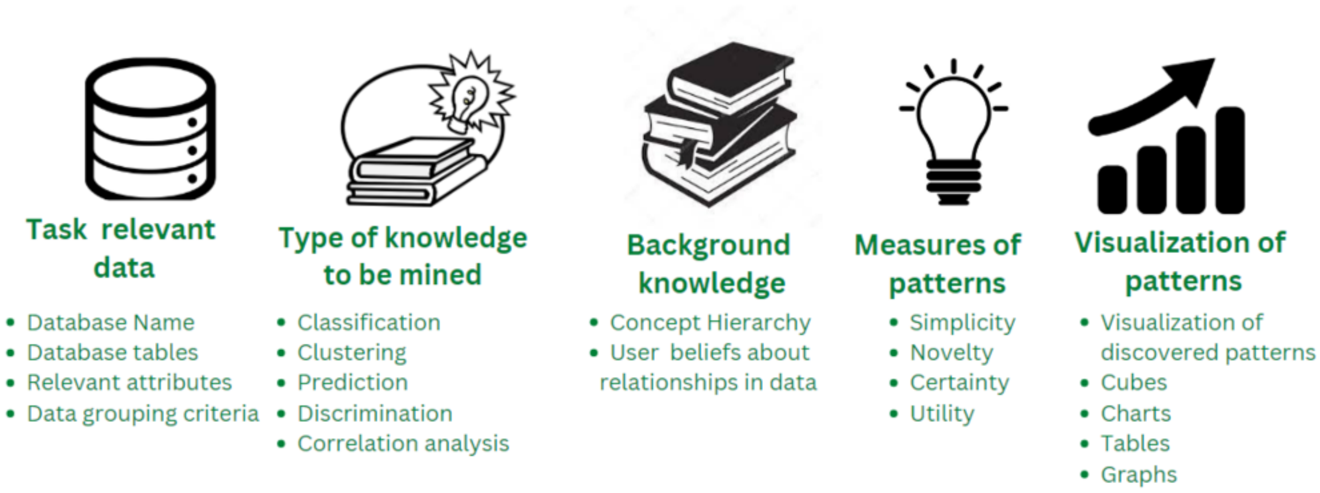
Ex: A data mining system that allows users to iteratively explore and visualize patterns in a dataset would be classified as an interactive exploratory system.

Data Mining Task Primitives:

Data mining primitives refer to the fundamental components/elements involved in the data mining process.

These primitives provide the necessary instructions and specifications to perform effective data mining tasks.

The primitives are:



-> <u>Task-relevant data</u>:

● This primitive focuses on identifying the specific data that will be used for mining.

● It involves choosing the relevant database or data warehouse, specifying conditions to choose the appropriate data, determining the relevant attributes or dimensions for exploration, and providing instructions for data ordering or grouping.

-> <u>Type of knowledge to be mined</u>:

● This primitive defines the specific type of knowledge or patterns that are to be discovered through data mining.

● It includes specifying the data mining function, such as characterization, discrimination, association, classification, clustering, or evolution analysis.

● Users can also provide pattern templates or meta patterns to guide the mining process and shape the expected patterns.

-> <u>Background Knowledge</u>:

● This primitive allows users to incorporate their domain knowledge or existing knowledge about the data being mined.

● Users can provide additional information, rules, or constraints to guide the knowledge discovery process and evaluate the patterns that are discovered.

- For example, incorporating concept hierarchies or domain-specific rules can assist in pattern interpretation.

-> Measure of interestingness of patterns:

- This primitive involves defining measures or criteria to determine the interestingness of discovered patterns.
- Users specify functions that help differentiate between uninteresting and valuable patterns.
- Interestingness measures can be based on factors such as simplicity, certainty, utility, novelty, or domain-specific requirements.

-> Visualization of discovered patterns:

- This primitive focuses on how the discovered patterns should be presented or visualized to users.
- It involves choosing appropriate visual representations, such as rules, tables, cross tabs, charts, decision trees, or other visual formats.
- Effective visualization helps users understand and interpret the discovered patterns more easily.

Integration of Data mining system with a Data warehouse:

A system of software suites that applies various data analysis and modeling techniques to discover patterns, relationships, and insights within large databases is known as Data Mining System.

A data warehouse/database is a centralized repository that collects, integrates and stores large volumes of data from various sources within an organization.

Hence, data warehouse acts as a database for data mining system which initially provides data to work on, for mining and is also used to store the mined patterns at the end of mining process.

Additionally, data warehouse also provides some functions/operations that can be used as primitives (making data ready to be mined) for data mining process, which could make data mining system efficient.

Hence, data warehouse is a prerequisite of data mining system to perform data mining process.

Based on the range/extent to which the data warehouse is used by data mining system, the integrations are classified into (4) types:

(1) No Coupling
(2) Loose Coupling
(3) Semi-tight Coupling
(4) Tight Coupling

(1) No Coupling:

No Data warehouse/database is used to integrate with the DMS.

The DMS obtains the initial data set from the flat files.

Flat files are simple data files in text/binary format with a structure known by the data mining algorithm. The data in these files can be transactions data, time-series data, scientific measurements data, etc.

The DMS applies DM algorithms on files which are designed specifically to give a structure to data stored on these files.

This architecture of DMS is inefficient due to lack of primitive functions and implementation of additional algorithms for structural fetching and storing the data in flat files.

(2) Loose Coupling:

A data warehouse is used to integrate with the DMS but without any primitive functions.

This architecture is more efficient than No Coupling architecture, because no additional algorithms are used to fetch and store the data in the warehouse.

This architecture is not optimized enough and may face performance issues due to the absence of primitive functions.

(3) Semi-tight Coupling:

A data warehouse with primitive functions is used to integrate with DMS.

Primitive functions pre-computational operations for data mining process, which include aggregation, sorting, etc.

Frequently used pre-computed results can be stored in the warehouse to improve the performance of the DMS.

(4) Tight Coupling:

The data warehouse is fully integrated into the DMS.

Additional Functionalities offered by this architecture along with the primitive functions is query processing which results in more efficiency and high performance of DMS.

=> Difference b/w Data Warehousing and Data Mining:

S. No.	Basis of Comparison	Data Warehousing	Data Mining
1.	**Definition**	A data warehouse is a database system that is designed for analytical analysis instead of transactional work.	Data mining is the process of analyzing data patterns.
2.	**Process**	Data is stored periodically.	Data is analyzed regularly.
3.	**Purpose**	Data warehousing is the process of extracting and storing data to	Data mining is the use of pattern recognition logic to identify patterns.

		allow easier reporting.	
4.	**Managing Authorities**	Data warehousing is solely carried out by engineers.	Data mining is carried out by business users with the help of engineers.
5.	**Data Handling**	Data warehousing is the process of pooling all relevant data together.	Data mining is considered as a process of extracting data from large data sets.
6.	**Functionality**	Subject-oriented, integrated, time-varying and non-volatile constitute data warehouses.	AI, statistics, databases, and machine learning systems are all used in data mining technologies.
7.	**Task**	Data warehousing is the process of extracting and storing data in order to make reporting more efficient.	Pattern recognition logic is used in data mining to find patterns.

8.	**Uses**	It extracts data and stores it in an orderly format, making reporting easier and faster.	This procedure employs pattern recognition tools to aid in the identification of access patterns.
9.	**Examples**	When a data warehouse is connected with operational business systems like CRM (Customer Relationship Management) systems, it adds value.	Data mining aids in the creation of suggestive patterns of key parameters. Customer purchasing behavior, items, and sales are examples. As a result, businesses will be able to make the required adjustments to their operations and production.

Major Issues in Data Mining:

Data Pre-Processing:

The process of transforming raw data into a format that is suitable for mining and analysis is known as data pre-processing.

Pre-processing involves various tasks, such as Data Cleaning, Data Integration, Data Transformation, Data Reduction and Data Discretization.

The Major Reasons for performing data preprocessing are:

Data Quality Improvement:

Handling Incomplete Data:

Dealing with Noisy Data:

Resolving Inconsistent Data:

Tasks involved in Data Pre-Processing are:

(1) Data Cleaning:

(2) Data Integration:

(3) Data Transformation:

(4) Data Reduction:

(5) Data Discretization: