DATA MINING

MODULE 1

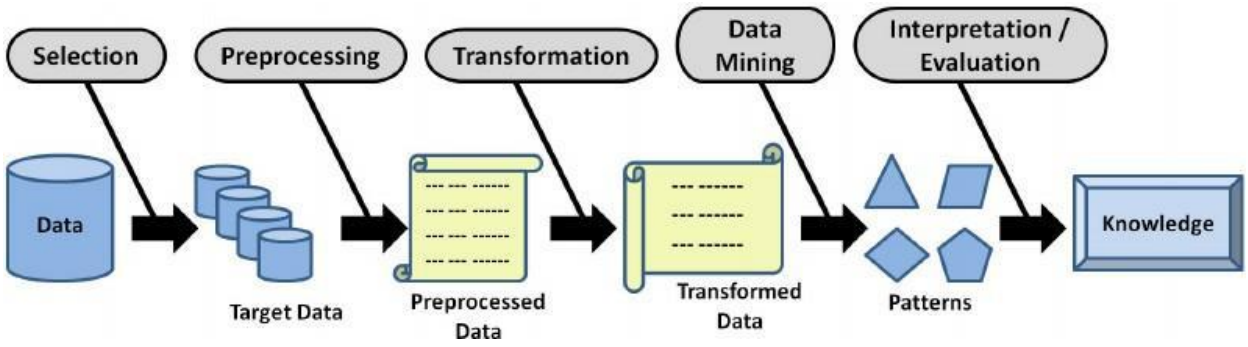
What is Data Mining?

Data mining is the process of automatically discovering useful information in large data repositories. Data mining techniques are deployed to search large databases in order to find novel and useful patterns that might otherwise remain unknown. They also provide capabilities to predict the outcome of a future observation.

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

Data Mining and Knowledge Discovery

Data mining is an integral part of K**nowledge Discovery in Databases** (KDD), which is the overall process of converting raw data into useful information. This process consists of a series of transformation steps, from data preprocessing to post processing of data mining results.



1. **Data cleaning** (to remove noise and inconsistent data)

2. **Data integration** (where multiple data sources may be combined)

3. **Data selection** (where data relevant to the analysis task are retrieved from the database)

4. **Data transformation** (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)

5. **Data mining** (an essential process where intelligent methods are applied to extract data patterns)

6. **Pattern evaluation** (to identify the truly interesting patterns representing

knowledge based on interestingness measures)

7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present mined knowledge to users)

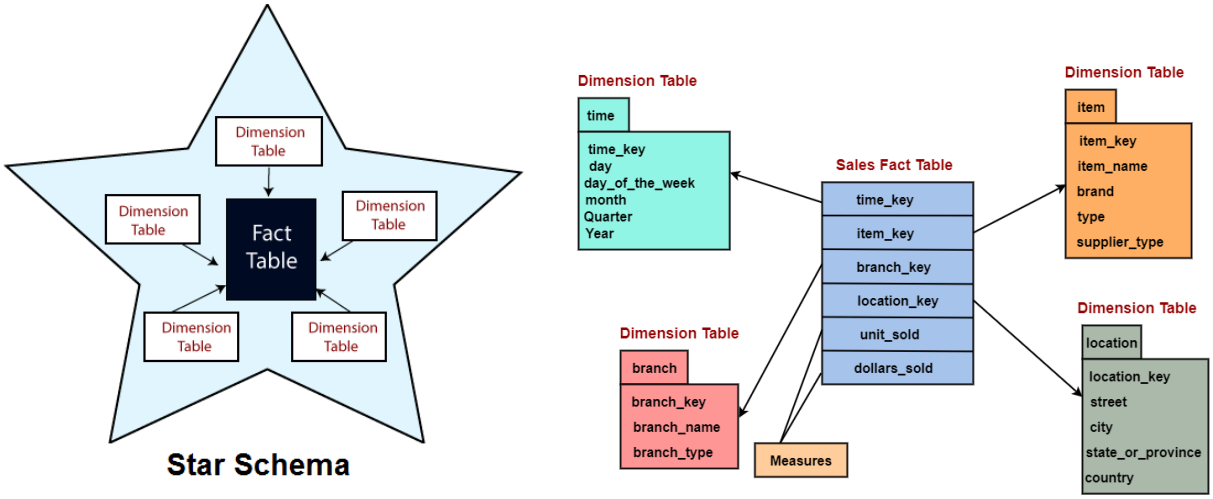
Steps1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Conceptual Modeling of Data Warehouses

Modeling data warehouses: dimensions & measures

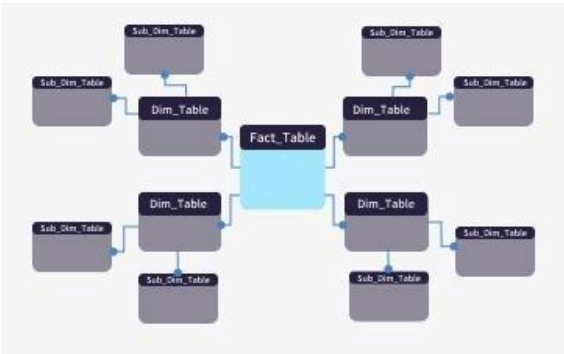
• **Star schema**: A fact table in the middle connected to a set of dimension tables.

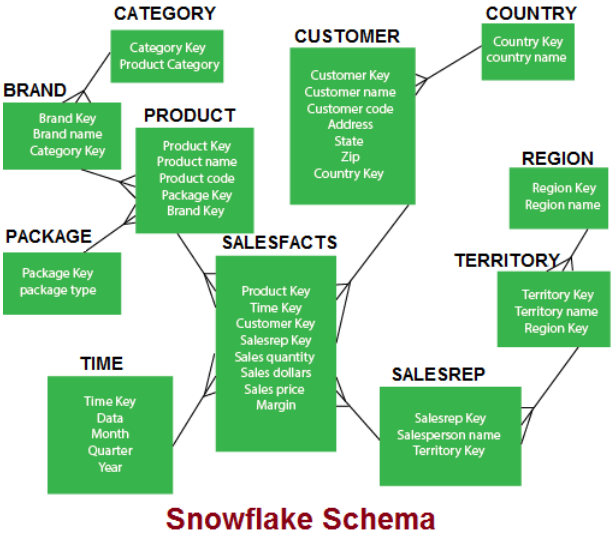


Fact Tables

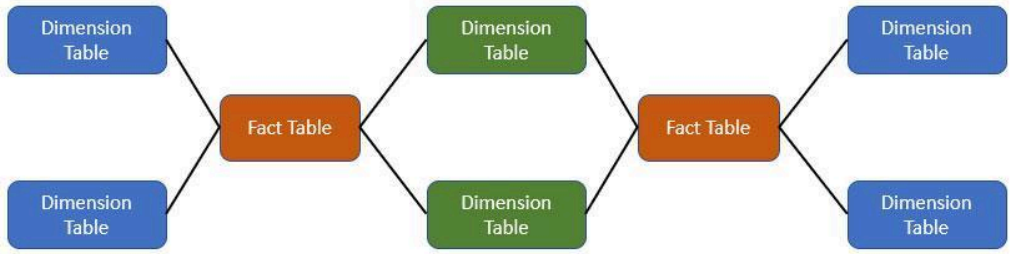
A table in a star schema which contains facts and connected to dimensions. A fact table has two types of columns: those that include fact and those that are foreign keys to the dimension table. The primary key of the fact tables is generally a composite key that is made up of all of its foreign keys.

• **Snowflake schema**: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake

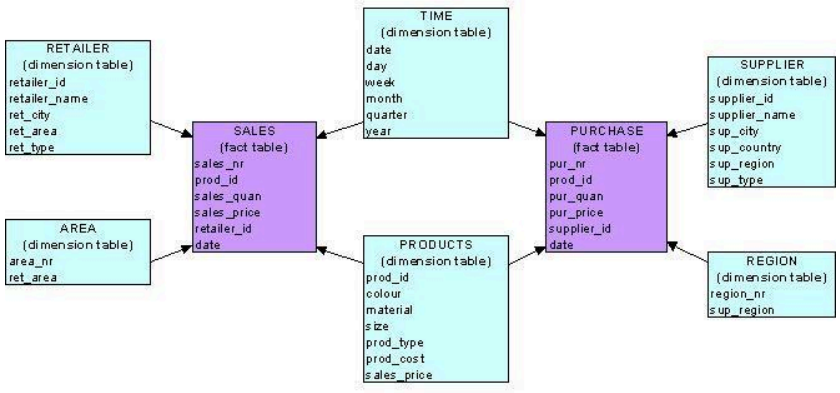




• **Fact constellations**: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation



Galaxy Model



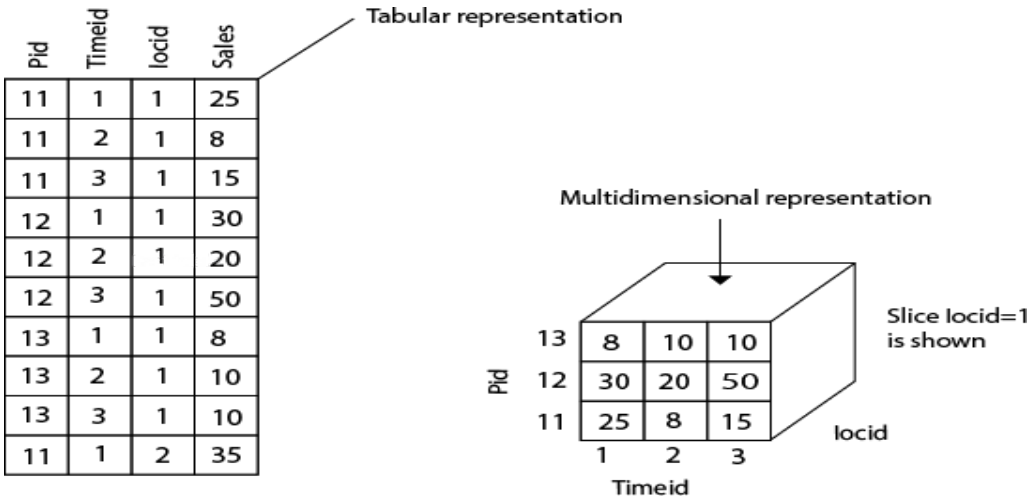
What is Multi- Dimensional Data Model?

- A multidimensional model views data in the form of a **data-cube.**
- A **data cube** enables data to be modeled and viewed in **multiple dimensions**.

- It is defined by dimensions and facts.
- The dimensions are the perspectives or entities concerning which an organization keeps records.
    - ➔ For example, a shop may create a sales data warehouse to keep records of the store's sales for the dimension, time, item, and location.
    - ➔ These dimensions allow the sale to keep track of things, for example, monthly sales of items and the locations at which the items were sold.
    - ➔ Each dimension has a table related to it, called a dimensional table, which describes the dimension further.
    - ➔ For example, a dimensional table for an item may contain the attributes item_name, brand, and type.

A multidimensional data model is organized around a central theme, for example, sales. This theme is represented by a fact table. Facts are numerical measures. The fact table contains the names of the facts or measures of the related dimensional tables.

Example:



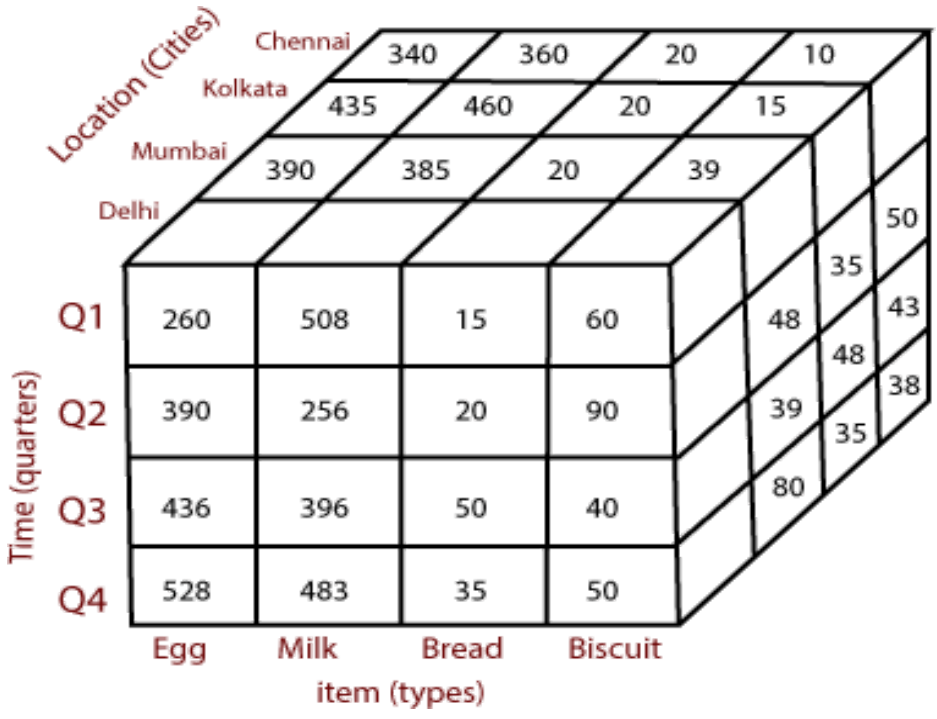
Consider the data of a shop for items sold per quarter in the city of Delhi. The data is shown in the table. In this 2D representation, the sales for Delhi are shown for the time dimension (organized in quarters) and the item dimension (classified according to the types of an item sold). The fact or measure displayed in rupee_sold (in thousands).

Location="Delhi"				
	item (type)			
Time (quarter)	Egg	Milk	Bread	Biscuit
Q1	260	508	15	60
Q2	390	256	20	90
Q3	436	396	50	40
Q4	528	483	35	50

Now, if we want to view the sales data with a third dimension, For example, suppose the data according to time and item, as well as the location is considered for the cities Chennai, Kolkata, Mumbai, and Delhi. These 3D data are shown in the table. The 3D data of the table are represented as a series of 2D tables.

	Location="Chennai"				Location="Kolkata"				Location="Mumbai"				Location="Delhi"			
	item				item				item				item			
Time	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit	Egg	Milk	Bread	Biscuit
Q1	340	360	20	10	435	460	20	15	390	385	20	39	260	508	15	60
Q2	490	490	16	50	389	385	45	35	463	366	25	48	390	256	20	90
Q3	680	583	46	43	684	490	39	48	568	594	36	39	436	396	50	40
Q4	535	694	39	38	335	365	83	35	338	484	48	80	528	483	35	50

Conceptually, it may also be represented by the same data in the form of a 3D data cube, as shown in fig:



OLAP Operations in the Multi- dimensional Data Model

OLAP stands for **On-Line Analytical Processing**. OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast,consistent, interactive access in a wide variety of **possible views of data** that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.

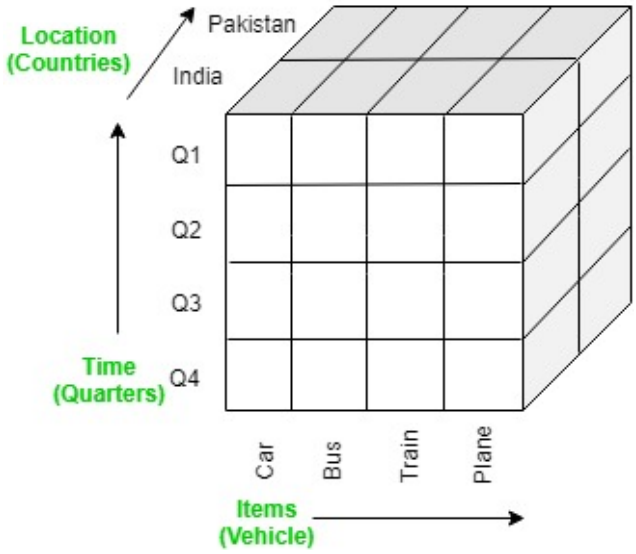
- **OLAP** implements the multidimensional analysis of business information and supports the capability for complex estimations, trend analysis, and sophisticated data modeling.
- It is rapidly enhancing the essential foundation for Intelligent Solutions containing Business Performance Management, Planning, Budgeting, Forecasting, Financial Documenting, Analysis, Simulation-Models, Knowledge Discovery, and Data Warehouses Reporting.
- OLAP enables end-clients to perform ad hoc analysis of records in multiple dimensions, providing the insight and understanding they require for better decision making.

❖ In the multidimensional model, the records are organized into various dimensions, and each dimension includes multiple levels of abstraction described by concept hierarchies. This organization supports users with the flexibility to view data from various perspectives. A number of OLAP data cube operations exist to demonstrate these different views, allowing interactive queries and search of the record at hand. Hence, OLAP supports a user-friendly environment for interactive data analysis.

❖ Consider the OLAP operations which are to be performed on multidimensional data. The figure shows data cubes for sales of a shop. The cube contains the dimensions, location, and time and item, where the **location** is aggregated with regard to city values,**time** is aggregated with respect to quarters, and an **item** is aggregated with respect to item types.

Roll-Up

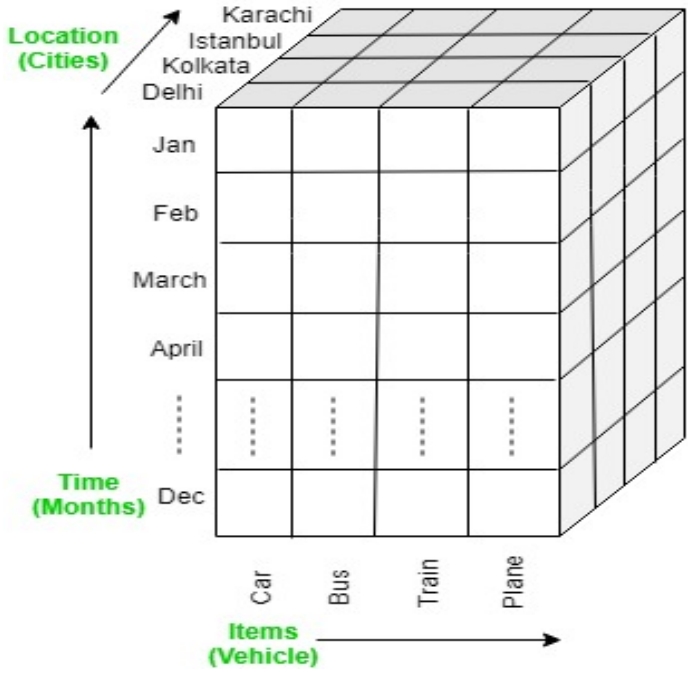
- The roll-up operation **(also known as drill-up or aggregation operation)** performs aggregation on a data cube, by climbing down concept hierarchies, i.e., dimension reduction.
- Roll-up is like **zooming-out** on the data cubes.
- In the cube given in the overview section, the **roll-up operation** is performed by climbing up in the concept hierarchy of Location dimension **(City -> Country)**



- When a roll-up is performed by **dimension reduction**, one or more **dimensions are removed** from the cube.

Drill-Down

- The drill-down operation **(also called roll-down)** is the reverse operation of **roll-up**.
- Drill-down is like **zooming-in** on the data cube.
- It navigates from less detailed records to more detailed data.
- Drill-down can be performed by either **stepping down** a concept hierarchy for a dimension or adding additional dimensions.
- In the cube given below, the drill down operation is performed by moving down in the concept hierarchy of **Time dimension (Quarter -> Month).**



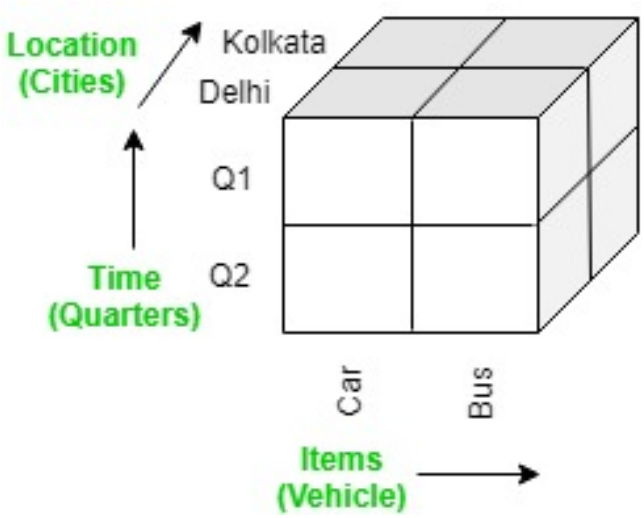
- Figure shows a drill-down operation performed on the dimension time by stepping down a concept hierarchy which is defined as day, month, quarter, and year.
- Drill- down appears by descending the time hierarchy from the level of the quarter to a more detailed level of the month.

Slice

- A **slice** is a subset of the cubes corresponding to a single value for one or more members of the dimension.
- For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site.
- So, the Slice operations perform a selection on one dimension of the given cube, thus resulting in a subcube.
- In the cube given in the overview section, Slice is performed on the dimension Time = "Q1".

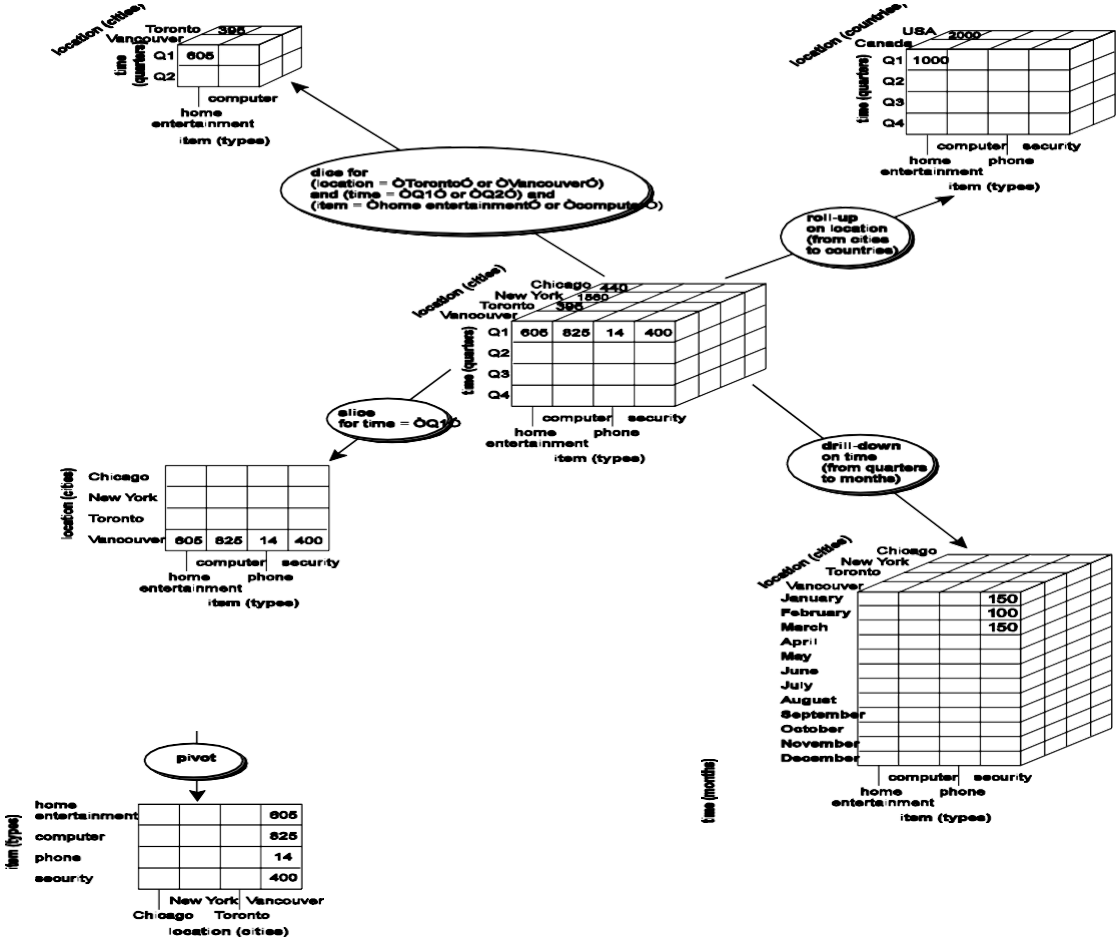
Dice

- The dice operation describes a subcube by operating a selection on two or more dimensions.
- In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:
    - ★ Location = "Delhi" or "Kolkata"
    - ★ Time = "Q1" or "Q2"
    - ★ Item = "Car" or "Bus"



Pivot

- The pivot operation is also called a rotation.
- Pivot is a visualization operation which rotates the data axes in view to provide an alternative presentation of the data.
- It may contain swapping the rows and columns or moving one of the row-dimensions into the column dimensions.



Data Mining Tasks

Data mining tasks are generally divided into two major categories:

Predictive tasks

The objective of these tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as the target or dependent variable, while the attributes used for making the prediction are known as the explanatory or independent variables.

Descriptive tasks

Here, the objective is to derive patterns (correlations, trends, clusters, trajectories, and anomalies) that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require post processing techniques to validate and explain the results.

Data Mining Functionality

- **Class/Concept Descriptions**

  Classes or definitions can be correlated with results. In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts. These class or concept definitions are referred to as class/concept descriptions.

  - **Data Characterization:**

    This refers to the summary of general characteristics or features of the class that is under the study. For example, To study the characteristics of a software product whose sales increased by 15% two years ago, anyone can collect these types of data related to such products by running SQL queries.

  - **Data Discrimination:**

    It compares common features of classes which are under study. The output of this process can be represented in many forms. Eg., barcharts, curves and pie charts.

- **Mining Frequent Patterns, Associations, and Correlations:**

  Frequent patterns are nothing but things that are found to be most common in the data.There are different kinds of frequency that can be observed in the dataset.

  - **Frequent Itemset:**

    This applies to a number of items that can see together regularly for eg: milk and sugar.
  - **Frequent Subsequence:**

    This refers to the pattern series that often occurs regularly Eg: purchasing a phone followed by a back cover.

  - **Frequent Substructure:**

    Frequent substructures are patterns that appear frequently in a dataset. These can include subgraphs, subtrees, or any other substructure depending on the context of the data. For example, in chemical informatics, molecules can be represented as graphs, and frequent substructures might refer to recurring patterns in these molecular structures.

Association Analysis:

The process involves uncovering the relationship between data and deciding the rules of the association. It is a way of discovering the relationship between various items. For example, it can be used to determine the sales of items that are frequently purchased together.

Correlation Analysis:

Correlation is a mathematical technique that shows whether and how strongly the pairs of attributes are related to each other. For Example, Highted people tend to have more weight.

Classification of Data Mining Systems

- Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science.
- Moreover, depending on the data mining approach used, techniques from

other disciplines may be applied, such as neural networks,fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high-performance computing.

- Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, Web technology, economics, business, bioinformatics, or psychology.
- Data mining systems can be categorized according to various criteria, as follows:

Classification according to the *kinds of databases* mined:

- A data mining system can be classified according to the kinds of databases mined.
- Database systems can be classified according to different criteria (such as data models, or the types of data or applications involved), each of which may require its own data mining technique.
- Data mining systems can therefore be classified accordingly.

What Kinds of Data Can Be Mined?

Data mining can be applied to any kind of data as long as the data are meaningful for a target application. The most basic forms of data for mining applications are database data, data warehouse data and transactional data. The type of data determines which tools and techniques can be used to analyze the data.

- **Database Data**
  - A Database System, also called a database management system(DBMS), consists of a collection of interrelated data, known as a

database, and a set of software programs to manage and access the data.

- The software programs provide mechanisms for defining database structures and data storage; for specifying and managing concurrent, shared, or distributed data access; and for ensuring consistency and security of the information stored despite system crashes or attempts at unauthorized access.

➔ Relational database

- It is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes(columns or fields) and usually stores a long set of **tuples** (records or rows).
- Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values.
- A semantic data model, such as an entity-relationship(ER) data model,is often constructed for relational databases. An ER data model represents the database set of entities and their relationships.

➔ Data Warehouse

- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and usually residing at a single site.Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing. To Facilitate Decision Making , the data in a data warehouse is organized around major subjects(e.g.,customer,item,supplier, and activity). The data are stored to provide information from a historical perspective, such as in the past 6 to 12 months, and are typically summarized.
- A data warehouse is usually modeled by a multidimensional data structure, called a datacube,in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum(sales amount). A data cube provides a multidimensional view of data and allows the precomputation and fast access of summarized data.

➔ Transactional Data

- In general, each record in a transactional database captures a transaction, such as a customer's purchase, a flight booking, or a user's clicks on a webpage. A transaction typically includes a unique transaction identity number (trans ID) and a list of the items making up the transaction, such as the items purchased in the transaction.
- A transactional database may have additional tables, which contain other information related to the transactions, such as item description, information

about the salesperson or the branch, and so on.

Classification according to the kinds of knowledge mined:

Data mining systems can be categorized according to the kinds of knowledge they mine,that is,based on data mining functionalities, such as characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis. A comprehensive data mining system usually provides multiple and/or integrated data mining functionalities.

- **Characterization**: Data characterization is a summarization of general features of objects in a target class, and produces what is called characteristics.The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.
- **Discrimination**:Data discrimination produces what are called *discriminant rules* and is basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class.
- **Association analysis**: Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*,identifies the frequent itemsets.Another threshold,*confidence*, which is the conditional probability that an item appears in a transaction when another item appears, is used to pinpoint association rules. Association analysis is commonly used for market basket analysis.
- **Classification**: Classification analysis is the organization of data in given classes. Also known as *supervised classification*,the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set where* all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The Model Is Used To Classify New Objects.

- **Prediction**:Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context.

- There are two major types of predictions:one can either try to predict some unavailable data values or pending trends, or predict a class label for some data.

- The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes.

Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trend in time related data. The major idea is to use a large number of past values to consider probable future values.

- **Clustering**: Similar to Classification, clustering is the organization of data in classes.However,unlike classification,and clustering,class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called *unsupervised classification*, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (*intra-class similarity*) and minimizing the similarity between objects of different classes (*inter-class similarity*).
- **Outlier analysis**: Outliers are data elements that cannot be grouped in a given class or cluster. Also Known As *exceptions or surprises*,they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains,and thus can be very significant and their analysis valuable.
- **Evolution and deviation analysis**: Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution analysis models evolutionary trends in data, which consent to characterizing, comparing, classifying or clustering of time related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

Classification according to the kinds of techniques utilized:

Data mining systems can be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved (e.g., autonomous systems,interactive exploratory systems,query-driven systems) or the methods of data analysis employed (e.g., database-oriented or data warehouse–oriented techniques,machine learning,statistics,visualization,pattern recognition, neural networks, and so on). A sophisticated data mining system will often adopt multiple data mining techniques or work out an effective, integrated technique that combines the merits of a few individual approaches.

- **Statistics:**

  Statistics studies the collection, analysis, interpretation or explanation, and presentation of data. Data mining has an inherent connection with statistics. A statistical model is a set of mathematical functions that describe the behavior of the objects in a target class in terms of random variables and their associated probability distributions. Statistical models are widely used

to model data and data classes. Statistics research develops tools for prediction and forecasting using data and statistical models. Statistical methods can be used to summarize or describe a collection of data.

- **Machine Learning:**

  Machine learning investigates how computers can learn (or improve their performance)based on data. A main research area is for computer programs to automatically learn to recognize complex patterns and make intelligent decisions based on data. Machine learning is a fast-growing discipline.

Classification according to the applications adapted:

- Data mining systems can also be categorized according to the applications they adapt.
- For example, data mining systems may be tailored specifically for finance, telecommunications, DNA, stock markets, e-mail, and soon.
- Different applications often require the integration of application-specific methods. Therefore, a generic, all-purpose data mining system may not fit domain-specific mining tasks.

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

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 tasks.
- **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 the 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.

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 amounts of data in databases, data mining algorithms 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 are further processed in a parallel fashion. Then the results from the partitions are merged. The incremental algorithms update databases without mining the data again from scratch.

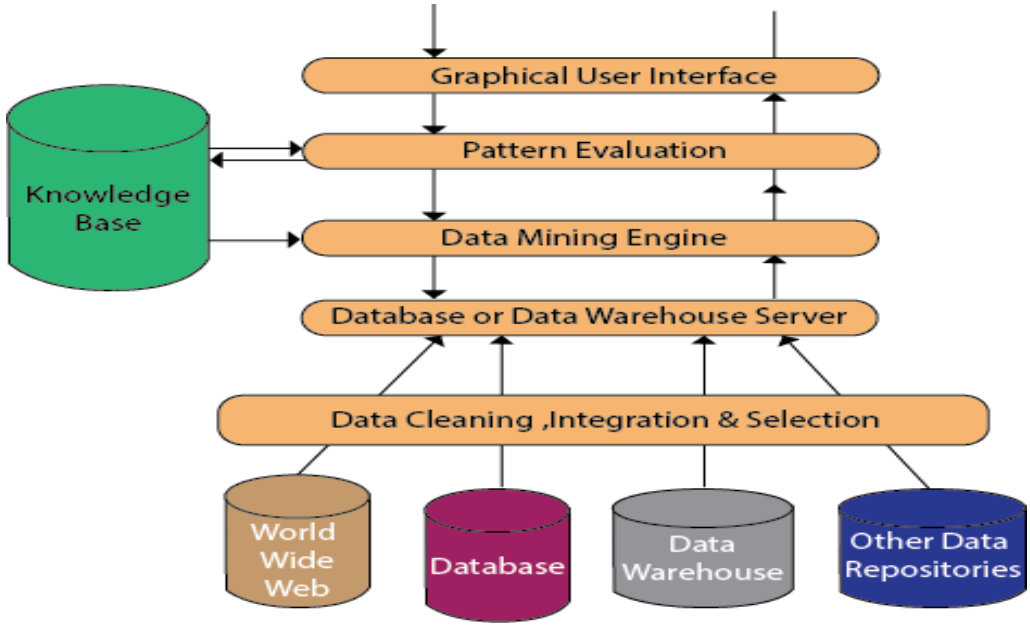
Diverse Data Types Issues

- **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 this 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 sources may be structured, semi structured or unstructured.Therefore mining the knowledge from them adds challenges to data mining.

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 responsible 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 results or 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.

Operational DBMS Vs Data Warehouses

- The Operational Database is the source of information for the data warehouse. It includes detailed information used to run the day to day operations of the business.
- The data frequently changes as updates are made and reflect the current value of the last transactions. Operational Database Management Systems also called **OLTP (Online Transactions Processing Databases),** are used to manage dynamic data in real-time.
- Data Warehouse Systems serve users or knowledge workers in the purpose of data analysis and decision-making.
- Such systems can organize and present information in specific formats to accommodate the diverse needs of various users. These systems are called **Online-Analytical Processing (OLAP) Systems.**

Operational Database Systems	Data Warehouses
Operational systems are generally designed to support high-volume transaction processing.	Data warehousing systems are generally designed to support high-volume analytical processing. (i.e. OLAP).
Operational systems focuses on Data in.	Data warehousing systems focuses on Information out.
In Operational systems data is stored with a functional or process orientation.	In Data warehousing systems data is stored with a subject orientation.
Performance is low for analysis queries.	Performance is high for analysis queries.
It is used for Online Transactional Processing (OLTP)	It is used for Online Analytical Processing (OLAP).
Operational systems represent current transactions.	Data warehousing systems reads the historical data.
Data within operational systems are generally updated regularly.	Data within a data warehouse is non-volatile, meaning when new data is added old data is not erased so rarely updates.
Complex data structures.	Multi dimensional data structures.

OLTP Vs OLAP

- **OLTP (On-Line Transaction Processing)**is featured by a large number of short on-line transactions (INSERT, UPDATE, and DELETE). The primary significance of OLTP operations is put on very rapid query processing, maintaining record integrity in multi-access environments, and effectiveness consistent by the number of transactions per second. In the OLTP database, there is an accurate and current record, and the schema used to save the transactional database is the entity model (usually 3NF).

- **OLAP (On-line Analytical Processing)** is represented by a relatively low volume of transactions. Queries are very difficult and involve aggregations. For OLAP operations, response time is an effectiveness measure. OLAP applications are generally used by Data Mining techniques. In the OLAP database there is aggregated, historical information, stored in multi-dimensional schemas(generally star schema).

OLAP	OLTP
✓ Gives a multi-dimensional view of business activities.	✓ Enables a snapshot of ongoing business processes.
✓ Helps with planning, problem solving, and decision support.	✓ Useful for controlling and running fundamental business tasks.
✓ Data source is consolidated data	✓ Data source is the operational data.
✓ Includes Periodic long-running batch jobs that refresh the data.	✓ Has short and fast inserts and updates which are initiated by end users.
✓ OLAP applications are widely used by Data Mining techniques.	✓ Large number of short on-line transactions
✓ Database design is typically de-normalized and contains fewer tables.	✓ Database design in OLTP is highly normalized.
✓ Often involves complex queries along with aggregations, which in turn compels processing speed to be dependent on the amount of data involved; batch data refreshes, etc.	✓ Involves standardized and simple queries that return relatively few records hence is faster.

DATA PREPROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

Why do we need Data Preprocessing?

A real-world data generally contains:

- **Incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data.

  Example: Customer Information with Missing Values

Original Data:
Customer 1: Name - Alice, Age - 28, Gender - Female
Customer 2: Name - Bob, Age - (missing), Gender - Male
Customer 3: Name - Claire, Age - 35, Gender - (missing)

Explanation: Incomplete data is present in the form of missing values. The age of Bob and the gender of Claire are not provided, making it challenging to perform a comprehensive analysis.

- **Noisy**: Containing errors or outliers.

  Example: Student Exam Scores with Typographical Errors

  Original Data: 90, 92, 88, 95, 105

  Noisy Data (due to a typographical error): 90, 92, 88, 95, 1050

  Explanation: The last exam score contains a typographical error, adding an extra zero. This introduces noise into the dataset, making it inaccurate and potentially misleading.

- **Inconsistent**: Containing discrepancies in codes or names

  Example: Product Prices in Different Currencies

  Original Data: $20, €15, $25, €18, ¥2000

  Inconsistent Data: $20, €15, $25, €18, ¥2000, £10

  Explanation: The inconsistency arises from mixing different currencies in the dataset. To ensure consistency, all prices should be in the same currency.

which cannot be directly used for machine learning models. Data warehouse needs consistent integration of quality data. Data preprocessing is required for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Major Tasks in Data Preprocessing

1. Data Cleaning

Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

• Missing Values

Missing data may be due to:

Equipment malfunction, inconsistent with other recorded data and thus deleted, data not entered due to misunderstanding, certain data may not be considered important at the time of entry, not register history or changes of the data.

1. **Ignore the tuple:**

This is usually done when the class label is missing (assuming the mining task involves classification).This method is not very effective,unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably. By ignoring the tuple, we do not make use of the remaining attributes' values in the tuple. Such data could have been useful to the task at hand.

2. **Fill in the missing value manually:**

In general, this approach is time consuming and may not be feasible given a large data set with many missing values.

3. **Use a global constant to fill in the missing value:**

Replace all missing attribute values by the same constant such as a label like "Unknown" or $-\infty$. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown". Hence, although this method is simple, it is not foolproof.

4. **Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value:**

Central tendency, which indicates the "middle" value of a data distribution. For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median.

5. **Use the attribute mean or median for all samples belonging to the same class as the given tuple:**

For example, if classifying customers according to credit risk, we may replace the missing value with the mean income value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.

6. **Use the most probable value to fill in the missing value:**

This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

Noisy Data

Noise is a random error or variance in a measured variable.

Incorrect attribute values may due to:

Faulty data collection instruments, data entry problems, data transmission problems, technology limitation, inconsistency in naming conventions

Binning: Binning, also known as discretization, is a data preprocessing technique used in statistics and machine learning to transform continuous numerical variables into discrete categories or bins. This process involves grouping a set of continuous or numerical data points into a smaller number of intervals or bins. Binning is often used to simplify the data, reduce noise, and make it more suitable for analysis or modeling.

- In smoothing by **bin means**, each value in a bin is replaced by the mean value of the bin.
- Similarly, smoothing by **bin median can** be employed, in which each bin value is replaced by the bin median.
- In smoothing by **bin boundaries**, the minimum and maximum values in a given bin are identified as the bin boundaries.Each bin value is then replaced by the closest boundary value.
- In general, the larger the width, the greater the effect of the smoothing. Alternatively, bins may be **equal width**, where the interval range of values in each bin is constant.
- Binning is also used as a discretization technique.
- ***There are several reasons why binning data can be useful -***
- ***Simplification of data*** *- Binning reduces the complexity of data by grouping values into a smaller number of categories or intervals, which makes it easier to understand, summarize and visualize.*
- ***Reduction of noise*** *- In some cases, binning can help reduce noise in the data by smoothing out variations in individual data points and highlighting larger trends or patterns.*
- ***Facilitation of data analysis*** *- Binning can make it easier to perform statistical analysis and create visualizations, such as histograms, by reducing the number of unique values in the data.*
- ***Improvement of model performance*** *- Binning can also be used to create new features or input variables for predictive models. By grouping similar values, binning can strengthen the relationship between attributes and improve the performance of machine learning models.*

Simple discretization methods: Binning

Equal-width(distance) Binning: It divides the range into N intervals of equal size, uniform grid. If A and B are the lowest and the highest values of the attribute, the width of intervals will be:W =(B-A)/N. The most straightforward but outliers may dominate presentation. Skewed data is not handled well.

Example : {55,60,65,70,75,80,85,90,95,100}

Decide on the number of bins you want. For this example, let's use 3 bins.

Bin Width= (Max-Min)/No. of Bins

= (100-55)/3=15

Start with the minimum value and create intervals by adding the bin width successively.

The Bin 1 = 55 - 69

Bin 2 =70 - 84

Bin 3= 85 - 100

Equal-Depth/Frequency Binning:

Divides the data into bins with approximately the same number of data points in each bin.

Example : {55,60,65,70,75,80,85,90,95,100}

Decide on the number of bins you want. For this example, let's use 3 bins.

Bin size = Total no of data points/No. of bins

= 10/3 = 3.33 : Take it as 4

Now, assign each data point to a bin. In this case, each bin should contain approximately 4 data points.

Bin 1: 55,60,65,70

Bin 2: 75,80,85,90

Bin 3: 95,100

Custom Binning:

Involves manually defining bin boundaries based on domain knowledge or business rules.

Example: Custom bins might be: [Low: 0-30K], [Medium: 31-50K], [High: 51-80K], based on what's considered low, medium, or high income for a specific analysis.

Quantile Binning (Frequency Binning):

Divides the data into bins based on specific quantiles (percentiles) of the data distribution.

Example: If we want three quantiles, quantile bins might be: [Q1: 18-35K], [Q2: 36-50K], [Q3: 51-68K].

Decision-Tree-Based Binning:

Uses decision tree algorithms to determine optimal split points for binning.

Example: A decision tree might decide to split income into bins like [Low: 0-30K], [Medium: 31-50K], [High: 51-80K], based on its evaluation of income's impact on some outcome.

Clustering-Based Binning:

Applies clustering algorithms to group similar values together.

Example: Clustering might result in bins like [Cluster 1: 18-40K], [Cluster 2: 41-55K], based on income patterns.

Regression

- Regression refers to a predictive modeling technique that is used to establish the relationship between a dependent variable and one or more independent variables.
- The goal of regression analysis is to understand how the independent variables affect the dependent variable and to make predictions based on this understanding. It's a form of supervised learning, where the algorithm is trained on a labeled dataset containing input-output pairs.
- Data smoothing can also be done by regression, a technique that conforms data values to a function.
- **Linear regression** involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.
- **Multiple linear** regression is an extension of linear regression, where more

than two attributes are involved and the data are fit to a multidimensional surface.

Outlier analysis: Outliers may be detected by clustering, for example,where similar values reorganized into groups, or "clusters. "Intuitively, values that fall outside of the set of clusters may be considered outliers.

2. Data Integration:

Data mining often requires data integration—the merging of data from multiple data sources into a coherent datasource, as in data warehousing. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting dataset. This can help improve the accuracy and speed of the subsequent data mining process.

Schema integration: Integrate metadata from different sources.

Entity identification problem: Identify real world entities from multiple data sources.

Detecting and resolving data value conflicts: for the same real world entity, attribute values from different sources are different. Possible reasons: different representations, different scales.

Redundancy and correlation analysis: Redundancy is another important issue in data integration. An attribute may be redundant if it can be "derived" from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set. Some redundancies can be detected by **correlation analysis**. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data. For nominal data, we use the $\chi 2$(chi-square) test. For numeric attributes, we can use the correlation coefficient and covariance, both of which assess how one attribute's values vary from those of another.

3. Data Transformation

In this preprocessing step, the data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

Strategies for data transformation includes the following:

1. **Smoothing**: which works to remove noise from the data. Techniques include binning, regression, and clustering.

2. **Attribute construction** (or feature construction): where new attributes are constructed and added from the given set of attributes to help the mining process.

3. **Aggregation**: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for data analysis at multiple abstraction levels.

4. **Normalization:** where the attribute data are scaled so as to fall within a smaller range, such as −1.0 to 1.0, or 0.0 to 1.0. Normalizing the data attempts to give all attributes an equal weight. Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest- neighbor classification and clustering. If using the neural network backpropagation algorithm for classification mining, normalizing the input values for each attribute measured in the training tuples will help speed up the learning phase. For distance- based methods, normalization helps prevent attributes with initially large ranges (e.g.,income) from out-weighing attributes with initially smaller ranges(e.g.,binary attributes). It is also useful when given no prior knowledge of the data

Other methods:

- **Min-Max**: This technique scales the values of a feature to a range between 0 and 1. This is done by subtracting the minimum value of the feature from each value, and then dividing by the range of the feature.

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)}(new\_max(A) - new\_min(A)) + new\_min(A)$$

Min(A) - It is the minimum absolute value A.
Max(A) - It is the maximum absolute value of A.
v' - It is the new value of each attribute data.
v - It is the old value of each attribute data.
new_max(A), new_min(A) - It is the max and min value within the range (i.e boundary value of range) required respectively.

Example:

Normalize the following group of data –

1000, 2000, 3000, 9000
using min-max normalization by setting min:0 and max:1

Solution:

max(A)=9000, as the maximum data among 1000,2000,3000,9000 is 9000
min(A)=1000, as the minimum data among 1000,2000,3000,9000 is 1000

new_max(A)=1, as given in question- max=1
new_min(A)=0, as given in question- min=0

Case-1: normalizing 1000 –

```
v  =  1000 , putting all values in the formula,we get

v' =  (1000-1000) X (1-0)
      ----------------    + 0  =0
         9000-1000
```

Case-2: normalizing 2000 –

```
v   = 2000, putting all values in the formula,we get

v'  = (2000-1000) X (1-0)
        ---------------- + 0 =0 .125
          9000-1000
```

Case-3: normalizing 3000 –

```
v=3000, putting all values in the formula,we get

v'=(3000-1000) X (1-0)
     ---------------- + 0 =0 .25
       9000-1000
```

Case-4: normalizing 9000 –

```
v=9000, putting all values in the formula, we get

v'=(9000-1000) X (1-0)
  ----------------         + 0 =1
  9000-1000
```

Outcome :

Hence, the normalized values of 1000,2000,3000,9000 are 0, 0.125, .25, 1.

- **Z-Score**: In z-score normalization(or zero-mean normalization),the values for an attribute, A, are normalized based on the mean(i.e.,average) and standard deviation of A. This method of normalization is useful when the

actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

$$Z = \frac{X - \mu}{\sigma}$$

Z	→	Standard (Normal) or Z score
X	→	member element of group
μ	→	mean of expectation
σ	→	standard deviation

Example:

Normalize the following group of data –

200, 300, 400, 600, 1000

Solution:

Mean, $\mu = (200+ 300+ 400+ 600+ 1000)/5=500$

Standard Deviation, $\sigma = \sqrt{(xi- \mu)^2/m}$

$= \sqrt{(200-500)^2+ (300-500)^2 + (400-500)^2 +(600-500)^2+ (1000-500)^2 /5}$

$= 282.8$

200:

z = (200 - 500)/ 282.8= -1.06

300:

z = (300 -500)/ 282.8=-0.707

400: -0.354

600: 0.354

1000: 1.77

- **Decimal Scaling**: It normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A.

$$v' = v / 10^{\wedge}d$$

- v' is the new value after decimal scaling is applied.
- The attribute's value is represented by V.

- d is the smallest integer such that Max($vi/10^d$)<=1

Example:

X=[150,250,450,900]

Find the Smallest Integer (d):

$$d = \lceil \log_{10}(\max(|X|)) \rceil$$

- $d = \lceil \log_{10}(900) \rceil = 3$

Normalize each value using the formula:

$$X' = \frac{X}{10^d}$$

- For 150: $\frac{150}{10^3} = 0.15$
- For 250: $\frac{250}{10^3} = 0.25$
- For 450: $\frac{450}{10^3} = 0.45$
- For 900: $\frac{900}{10^3} = 0.90$

Normalized Data : X=[0.15,0.25,0.45,0.90]

5. **Discretization:**where the raw values of a numeric attribute(e.g.,age)are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior). The labels, in turn, can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numeric attribute.

6. **Concept hierarchy generation for nominal data**:where attributes such as street can be generalized to higher-level concepts, like city or country. Many hierarchies for nominal attributes are implicit within the database schema and can be automatically defined at the schema definition level.

4. Data Reduction:

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the

original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results.

Data reduction strategies include dimensionality reduction, numerosity reduction, and data compression.

- **Dimensionality reduction** is the process of reducing the number of random variables or attributes under consideration. Dimensionality Reduction Methods include **Wavelet Preprocessing transforms** and **Principal components analysis**, which transform or project the original data onto a smaller space. **Attribute subset selection** is a method of dimensionality reduction in which irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.
- **Numerosity reduction** techniques replace the original data volume by alternative, smaller forms of data representation. These techniques may be parametric or non- parametric. For parametric methods, a model is used to estimate the data, so that typically only the data parameters need to be stored, instead of the actual data. **Regression and log-linear models** are examples. Nonparametric methods for storing reduced representations of the data include **histograms, clustering, sampling, and data cube aggregation.**

In **data compression**, transformations are applied so as to obtain a reduced or "compressed" representation of the original data. If the original data can be reconstructed from the compressed data without any information loss, the **data reduction is called lossless**. If, instead, we can reconstruct only an approximation of the original data, then the **data reduction is called lossy.** There are several lossless algorithms for string compression; however, they typically allow only limited data manipulation. Dimensionality reduction and numerosity reduction techniques can also be considered forms of data compression.

Wavelet Transforms

- The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector X, transforms it to a numerically different vector, X0, of wavelet coefficients.
- The two vectors are of the same length. When applying this technique to data reduction, we consider each tuple as an n-dimensional data vector, that is, X = (x1, x2, ..., xn), depicting n measurements made on the tuple from n database attributes.
- The general procedure for applying a discrete wavelet transform uses a hierarchical pyramid algorithm that halves the data at each iteration, resulting in fast computational speed.

Wavelet transforms can be applied to multi dimensional data such as a data cube. This is done by first applying the transform to the first dimension, then to the second, and so on. The computational complexity involved is linear with respect to the number of cells in the cube. Wavelet transforms give good results on sparse or skewed data and on data with ordered attributes. Lossy compression by wavelets is reportedly better than JPEG compression,the current commercial standard. Wavelet transforms have many real world applications, including the compression of finger print images, computer vision, analysis of time-series data, and data cleaning.

Principal Components Analysis

- Suppose that the data to be reduced consist of tuples or data vectors described by n attributes or dimensions.
- Principal components analysis (PCA) searches for k n- dimensional orthogonal vectors that can best be used to represent the data, where $k \leq n$.
- The original data are thus projected onto a much smaller space, resulting in dimensionality reduction.
- Unlike attribute subset selection, which reduces the attribute set size by retaining a subset of the initial set of attributes, PCA"combines" the essence of attributes by creating an alternative, smaller set of variables.
- The initial data can then be projected onto this smaller set. PCA often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result.

PCA can be applied to ordered and unordered attributes,and can handle sparse data and skewed data. Multidimensional data of more than two dimensions can be handled by reducing the problem to two dimensions. Principal components may be used as inputs to multiple regression and cluster analysis.

STEPS FOR PRINCIPAL COMPONENT ANALYSIS

- Standardize the range of continuous initial variables
- Compute the covariance matrix to identify correlations
- Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
- Create a feature vector to decide which principal components to keep
- Recast the data along the principal components axes

In comparison with wavelet transforms, PCA tends to be better at handling sparse data,whereas wavelet transforms are more suitable for data of high dimensionality.

By finding the eigenvalues and eigenvectors of the covariance matrix, we

find that the eigenvectors with the largest eigenvalues correspond to the dimensions that have the strongest correlation in the dataset. This is PCA. It is a useful statistical technique that has found application in:

-fields such as face recognition and image compression.

-finding patterns in data of high dimension.

Attribute Subset Selection

- Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes(or dimensions).
- The goal of attribute subset selection is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes.
- Mining on a reduced set of attributes has an additional benefit: It reduces the number of attributes disappearing in the discovered patterns, helping to make the patterns easier to understand.
- The heuristic methods that explore a reduced search space are commonly used for attribute subset selection.
- Basic heuristic methods of attribute subset selection include the techniques that follow:

1. **Stepwise forward selection**: The procedure starts with an empty set of attributes as the reduced set. The best of the original attributes is determined and added to the reduced set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.

2. **Stepwise backward elimination**: The procedure starts with the full set of attributes. At each step, it removes the worst attribute remaining in the set.

3. **Combination of forward selection and backward elimination:** The stepwise forward selection and backward elimination methods can be combined so that, at each step, the procedure selects the best attribute and removes the worst from among the remaining attributes.

4. **Decision tree induction**: Decision tree algorithms were originally intended for classification. Decision tree induction constructs a flowchart like structure where each internal (non leaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction. At each node, the algorithm chooses the “best” attribute to partition the data into individual classes. When decision tree induction is used for attribute subset selection, a tree is constructed from the given data. All attributes that do not appear in the tree are assumed to be irrelevant. The set of attributes appearing in the tree form the reduced subset of attributes.

Regression and Log-Linear Models:

For parametric methods, data is represented using some model. The model is used to estimate the data, so that only parameters of data are required to be stored, instead of actual data. **Regression** and **Log-Linear methods** are used for creating such models.

Regression analysis

- It is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to access the strength t=of the relationship between variables and for modeling the future relationship between them.
- In (simple) **linear regression**, the data are modeled to fit a straight line,$Y=\alpha+\beta X$.
- Two parameters , α and β specify the line and are to be estimated by using the data at hand.
- **Multiple Linear Regression** is an extension of (simple) linear regression, which allows a response variable, y, to be modeled as a linear function of two or more predictor variables, Y = b0 + b1 X1 + b2 X2.

Log-linear models

- A log-linear model is a type of statistical model that is used to analyze relationships between categorical variables. The aim of these models is to provide a statistical framework for exploring and modeling relationships in categorical data.
- This model approximates discrete multidimensional probability distributions. Log-linear models can be used to estimate the probability of each point in a multidimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations.
- This allows a higher-dimensional data space to be constructed from lower-dimensional spaces. Log-linear models are therefore also useful for dimensionality reduction and data smoothing. The multi-way table of joint probabilities is approximated by a product of lower-order tables.

In Non- Parametric, these methods are used for storing reduced representations of the data including **histograms, clustering, sampling and data cube aggregation.**

Histograms

Histograms use binning to approximate data distributions and are a popular form of data reduction. Histograms were introduced. A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, referred to as buckets or bins. If each bucket represents only a single attribute–value/frequency pair, the buckets are called singleton buckets. Often, buckets instead represent continuous ranges for the given attribute.

- **Equal-width**:In an equal-width histogram, the width of each bucket range is uniform.
- **Equal-frequency (or equal-depth):** In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant.

Clustering

- Clustering techniques consider data tuples as objects.
- They partition the objects into groups, or clusters, so that objects within a cluster are "similar" to one another and "dissimilar" to objects in other clusters.
- In data reduction, the cluster representations of the data are used to replace the actual data. The effectiveness of this technique depends on the data's nature. It is much more effective for data that can be organized into distinct clusters than for smeared data.

Sampling

- Sampling can be used as a data reduction technique because it allows a large dataset to be represented by a much smaller random data sample(or subset).
- An advantage of sampling for data reduction is that the cost of obtaining a sample is proportional to the size of the sample, s, as opposed to N, the data set size.
- Hence, sampling complexity is potentially sublinear to the sizeof the data. When applied to data reduction, sampling is most commonly used to estimate the answer to an aggregate query.
- Simple random sampling (SRS) is a method of selection of a sample comprising of n a number of sampling units out of the population having
- N number of sampling units such that every sampling unit has an equal chance of being chosen.
- The samples can be drawn in two possible ways.

  The sampling units are chosen without replacement because the units, once chosen, are not placed back in the population.

  The sampling units are chosen with replacement because the selected units

are placed back in the population.

- **SRSWOR(Simple Random Sampling Without Replacement)**: probability of drawing any tuple is 1/N, that is, all samples are equally likely to be sampled.
- **SRSWR(Simple Random Sampling With Replacement)**:each tuple is drawn, it is recorded and is placed back in so that it may be drawn again.
- **Cluster sample**: if all the tuples are grouped into mutually disjoint clusters; reduced data representation can be obtained resulting in a cluster sample of the tuples.
- **Stratified sample:** If D is divided into mutually disjoint parts called strata; SRS at each strata.

Data Cube Aggregation

- Data cubes store multidimensional aggregated information.
- **Concept hierarchies** may exist for each attribute, allowing the analysis of data at multiple abstraction levels.
- Data cubes provide fast access to precomputed, summarized data, thereby benefiting online analytical processing as well as data mining.
- The cube created at the lowest abstraction level is referred to as the **base cuboid**.
- The base cuboid should correspond to an individual entity of interest such as sales or customer. In other words, the lowest level should be usable, or useful for the analysis.
- A cube at the highest level of abstraction is the **apex cuboid**.
- Data cubes created for varying levels of abstraction are often referred to as cuboids**,** so that a datacube may instead refer to a lattice of cuboids. Each higher abstraction level further reduces the resulting data size. When replying to data mining requests, the smallest available cuboid relevant to the given task should be used.

5. Data Discretization and Concept Hierarchy Generation

- Data discretization transforms numeric data by mapping values to interval or concept labels. Such methods can be used to automatically generate concept hierarchies for the data, which allows for mining at multiple levels of granularity. Discretization techniques include binning, histogram analysis, cluster analysis, decision tree analysis, and correlation analysis.
- For nominal data, concept hierarchies may be generated based on schema definitions as well as the number of distinct values per attribute. Concept hierarchies reduce the data by collecting and replacing low level concepts by higher level concepts.
- Discretization divides the range of a continuous attribute into intervals. Some classification algorithms only accept categorical attributes. It reduces data size by discretization and prepares for further analysis.

Discretization and concept hierarchy generation for numeric data

- **Binning**: Binning is a top-down splitting technique based on a specified number of bins. attribute values can be discretized by applying equal-width or equal-frequency binning, and then replacing each bin value by the bin mean or median, as in smoothing by bin means or smoothing by bin medians, respectively. These techniques can be applied recursively to the resulting partitions to generate concept hierarchies. Binning does not use class information and is therefore an unsupervised discretization technique. It is sensitive to the user-specified number of bins, as well as the presence of outliers.
- **Histogram analysis**: Like binning, histogram analysis is an unsupervised discretization technique because it does not use class information. A histogram partitions the values of an attribute, A, into disjoint ranges called buckets or bins. The histogram analysis algorithm can be applied recursively to each partition in order to automatically generate a multilevel concept hierarchy, with the procedure terminating once a prespecified number of concept levels has been reached.
- **Clustering analysis**: Cluster analysis is a popular data discretization method. A clustering algorithm can be applied to discretize a numeric attribute,A,by partitioning the values of A into clusters or groups. Clustering takes the distribution of A into consideration, as well as the closeness of data points, and therefore is able to produce high-quality discretization results.

Concept Hierarchy Generation for Nominal Data

Specification of a partial ordering of attributes explicitly at the schema level by users or experts : Concept hierarchies for nominal attributes or dimensions typically involve a group of attributes. A user or expert can easily define a concept hierarchy by specifying a partial or total ordering of the attributes at the schema level.

Specification of a portion of a hierarchy by explicit data grouping: This is essentially the manual definition of a portion of a concept hierarchy. In a large database, it is unrealistic to define an entire concept hierarchy by explicit value enumeration. On the contrary, we can easily specify explicit groupings for a small portion of intermediate-level data.

Specification of a set of attributes, but not of their partial ordering: A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state theirpartialordering.Thesystemcanthentrytoautomaticallygeneratetheattribute ordering so as to construct a meaningful concept hierarchy.

Consider the observation that since higher-level concepts generally cover several subordinate lower-level concepts, an attribute defining a high concept level (e.g., country) will usually contain a smaller number of distinct values than an attribute defining a lower concept level (e.g., street). Based on this observation, a concept hierarchycanbeautomaticallygeneratedbasedonthenumberofdistinctvaluesper attribute in the given attribute set. The attribute with the most distinct values is placed at the lowest hierarchy level. The lower the number of distinct values an attribute has, the higher it is in the generated concept hierarchy. This heuristic rule workswellinmanycases.Somelocal-levelswappingoradjustmentsmaybeapplied by users or experts, when necessary, after examination of the generated hierarchy.

Specification of only a partial set of attributes: Sometimes a user can be careless when defining a hierarchy, or have only a vague idea about what should be included in in a hierarchy. Consequently, the user may have included only a small subset of the relevant attributes in the hierarchy specification.