1. **Define Clustering. Explain with example of the portioning and hierarchical methods.**

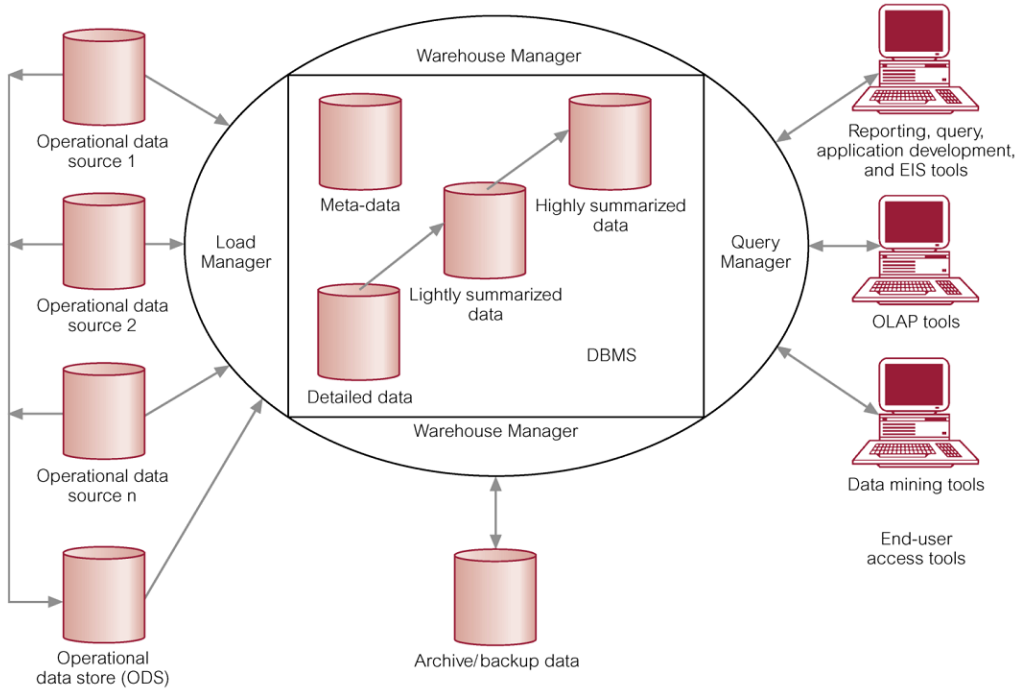
Clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis. This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

A partition clustering a simply a division of the set of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. partition clustering decomposes a data set into a set of disjoint clusters. Given a data set of N points, a partitioning method constructs K ($N \geq K$) partitions of the data, with each partition representing a cluster. That is, it classifies the data into K groups by satisfying the following requirements: (1) each group contains at least one point, and (2) each point belongs to exactly one group. Notice that for fuzzy partitioning, a point can belong to more than one group.

Many partition clustering algorithms try to minimize an objective function. For example, in K-means and K-medoids the function (also referred to as the distortion function) is
$\sum_{i=1}^{K}\sum_{j=1}^{|Ci|}Dist(xj,center(i))$

Hierarchical method(not include)

2. **Explain the architecture and implementation of data warehouse with example.**

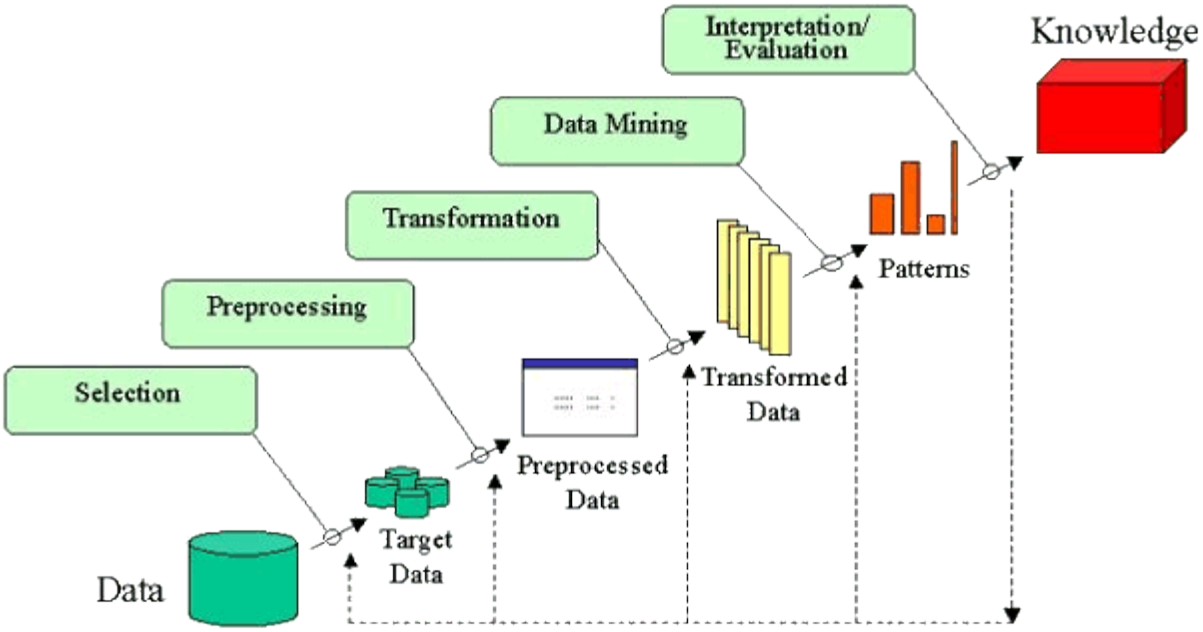


Operational Data Sources: It may include:
• Network databases.
• Departmental file systems and RDBMSs.
• Private workstations and servers.
• External systems (Internet, commercially available databases).
Operational Data Store (ODS): It is a repository of current and integrated operational data used for analysis.
• Often structured and supplied with data in same way as DW.
• May act simply as staging area for data to be moved into the warehouse.
Provides users with the ease of use of a relational database while remaining distant from decision support functions of the DW.
Warehouse Manager (Data Manager):
• Operations performed include:
– Analysis of data to ensure consistency.
– Transformation/merging of source data from temp storage into DW
– Creation of indexes.
– Backing-up and archiving data.
Query Manager (Manages User Queries):
• Operations include:
– directing queries to the appropriate tables and
– scheduling the execution of queries.
• In some cases, the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.
Meta Data: This area of the DW stores all the meta-data (data about data) definitions used by all the processes in the warehouse.
• Used for a variety of purposes:
– Extraction and loading processes
– Warehouse management process
– Query management process
• End-user access tools use meta-data to understand how to build a query.
• Most vendor tools for copy management and end-user data access use their own versions of meta-data.
Lightly and Highly Summarized Data: It stores all the pre-defined lightly and highly aggregated data generated by the warehouse manager.
• The purpose of summary info is to speed up the performance of queries.
• Removes the requirement to continually perform summary operations (such as sort or group by) in answering user queries.
Archive/Backup Data: It stores detailed and summarized data for the purposes of archiving and backup.
• May be necessary to backup online summary data if this data is kept beyond the retention period for detailed data.
• The data is transferred to storage archives such as magnetic tape or optical disk.
End-User Access Tools:
• The principal purpose of data warehousing is to provide information to business users for strategic decision-making.
• Users interact with the warehouse using end-user access tools.
• There are three main groups of access tools:

1. Data reporting, query tools
2. Online analytical processing (OLAP) tools *(Discussed later)*
3. Data mining tools *(Discussed later)*

3. What do you mean by knowledge discovery in database (KDD)?

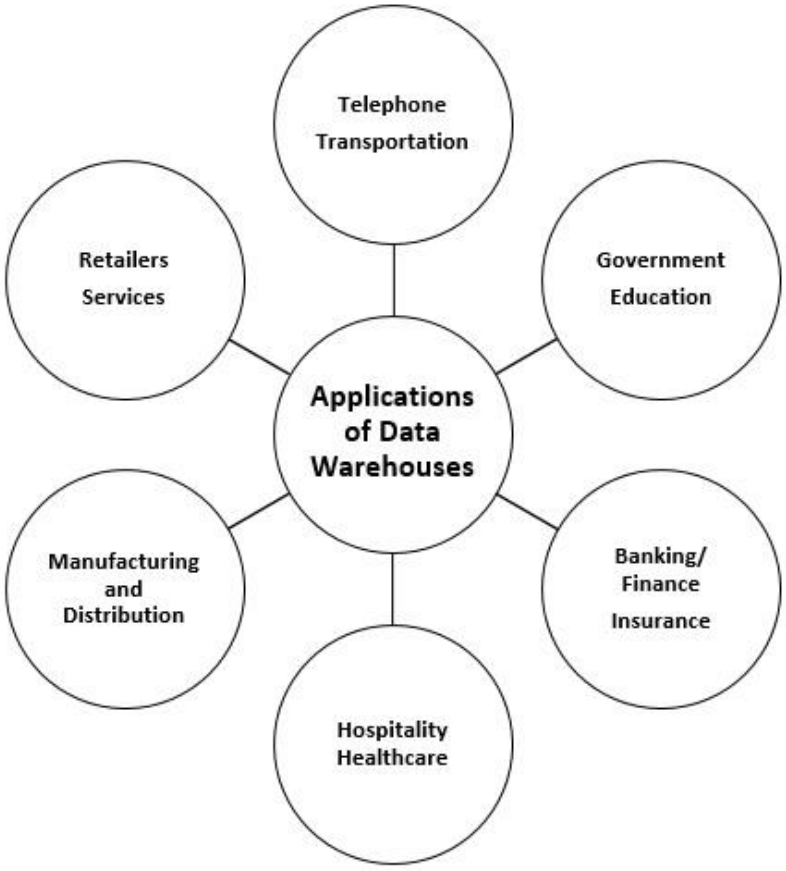
Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.



1. Identify the goal of the KDD process from the customer's perspective.
2. Understand application domains involved and the knowledge that's required
3. Select a target data set or subset of data samples on which discovery is be performed.
4. Cleanse and preprocess data by deciding strategies to handle missing fields and alter the data as per the requirements.
5. Simplify the data sets by removing unwanted variables. Then, analyze useful features that can be used to represent the data, depending on the goal or task.
6. Match KDD goals with data mining methods to suggest hidden patterns.
7. Choose data mining algorithms to discover hidden patterns. This process includes deciding which models and parameters might be appropriate for the overall KDD process.
8. Search for patterns of interest in a particular representational form, which include classification rules or trees, regression and clustering.
9. Interpret essential knowledge from the mined patterns.
10. Use the knowledge and incorporate it into another system for further action.
11. Document it and make reports for interested parties.

4. Explain the application of data warehouse and data mining.

Data Warehouses owing to their potential have deep-rooted applications in every industry which use historical data for prediction, statistical analysis, and decision making.



Banking Industry

In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.

Finance Industry

Similar to the applications seen in banking, mainly revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

Consumer Goods Industry

They are used for prediction of consumer trends, inventory management, market and advertising research. In-depth analysis of sales and production is also carried out. Apart from these, information is exchanged business partners and clientele.

Government and Education

The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.

Healthcare

One of the most important sectors which utilize data warehouses is the Healthcare sector. All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track

and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.

Data mining application:

Future Healthcare

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics.

Market Basket Analysis

Market basket analysis is a modeling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behavior of a buyer.

Education

There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behavior, studying the effects of educational support, and advancing scientific knowledge about learning.

Manufacturing Engineering

Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process.

5. Explain the data mining technique.

There are several major ***data mining*** ***techniques*** have been developing and using in data mining projects recently including *association*, *classification*, *clustering*, *prediction*, *sequential patterns* and *decision tree*. We will briefly examine those data mining techniques in the following sections.

Association

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. That's is the reason why association technique is also known as *relation technique*. The association technique is used in *market basket analysis* to identify a set of products that customers frequently purchase together.

Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that "given all records of employees who left the company, predict who will probably leave the company in a future period." In this case, we divide the records of employees into two groups that named "leave" and "stay". And then we can ask our data mining software to classify the employees into separate groups.

Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.

Prediction

The prediction, as its name implied, is one of a data mining techniques that discovers the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

Decision trees

The A decision tree is one of the most commonly used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers

The A decision tree is one of the most commonly used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers.

6. Explain the K-method algorithm.

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result.

Algorithmic steps for k-means clustering

Let $X = \{x_1,x_2,x_3,……..,x_n\}$ be the set of data points and $V = \{v_1,v_2,…….,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

Where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

7. Explain the data mining language.

The Data Mining Query Language is actually based on the Structured Query Language (SQL). Data Mining Query Languages can be designed to support ad hoc and interactive data mining. This DMQL provides commands for specifying primitives. The DMQL can work with databases and data warehouses as well. DMQL can be used to define data mining tasks. Particularly we examine how to define data warehouses and data marts in DMQL.

Here is the syntax of DMQL for specifying task-relevant data −

```
use database database_name

or

use data warehouse data_warehouse_name
in relevance to att_or_dim_list
from relation(s)/cube(s) [where condition]
order by order_list
group by grouping_list
```

Characterization

The syntax for characterization is −

```
mine characteristics [as pattern_name]
```

```
analyze  {measure(s) }
```

The analyze clause, specifies aggregate measures, such as count, sum, or count%.

Discrimination

The syntax for Discrimination is −

```
mine comparison [as {pattern_name]}
For {target_class } where  {t arget_condition }
{versus  {contrast_class_i }
where {contrast_condition_i}}
analyze  {measure(s) }
```

Association

The syntax for Association is−

```
mine associations [ as {pattern_name} ]
{matching {metapattern} }
```

Prediction

The syntax for prediction is −

```
mine prediction [as pattern_name]
analyze prediction_attribute_or_dimension
{set {attribute_or_dimension_i= value_i}}
```