DWDM

1.Differentiate OLAP/OLTP

Category	OLAP (Online Analytical Processing)	OLTP (Online Transaction Processing)
Definition	It is well-known as an online database query management system.	It is well-known as an online database modifying system.
Data source	Consists of historical data from various Databases.	Consists of only operational current data.
Method used	It makes use of a data warehouse.	It makes use of a standard database management system (DBMS).
Application	It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc.	It is application-oriented. Used for business tasks.
Normalized	In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized (3NF).
Usage of data	The data is used in planning, problem-solving, and decision-making.	The data is used to perform day-to-day fundamental operations.
Task	It provides a multi-dimensional view of different business tasks.	It reveals a snapshot of present business tasks.
Purpose	It serves the purpose to extract information for analysis and decision-making.	It serves the purpose to Insert, Update, and Delete information from the database.
Volume of data	A large amount of data is stored typically in TB, PB	The size of the data is relatively small as the historical data is archived in MB, and GB.
Queries	Relatively slow as the amount of data involved is large. Queries may take hours.	Very Fast as the queries operate on 5% of the data.
Update	The OLAP database is not often updated. As a result, data integrity is unaffected.	The data integrity constraint must be maintained in an OLTP database.
Backup and Recovery	It only needs backup from time to time as compared to OLTP.	The backup and recovery process is maintained rigorously
Processing time	The processing of complex queries can take a lengthy time.	It is comparatively fast in processing because of simple and straightforward queries.
Types of users	This data is generally managed by CEO, MD, and GM.	This data is managed by clerksForex and managers.
Operations	Only read and rarely write operations.	Both read and write operations.
Updates	With lengthy, scheduled batch operations, data is refreshed on a regular basis.	The user initiates data updates, which are brief and quick.
Nature of audience	The process is focused on the customer.	The process is focused on the market.
Database Design	Design with a focus on the subject.	Design that is focused on the application.
Productivity	Improves the efficiency of business analysts.	Enhances the user's productivity.

2.What is Data Mining ?Explain KDD process with a neat diagram.

A) Data mining refers to the process of extracting meaningful patterns, knowledge, or information from large sets of data. It involves analyzing vast amounts of data to uncover hidden patterns, correlations, and trends that might not be immediately obvious. Essentially, data mining transforms large amounts of raw data into valuable insights.

Synonyms for Data Mining include:

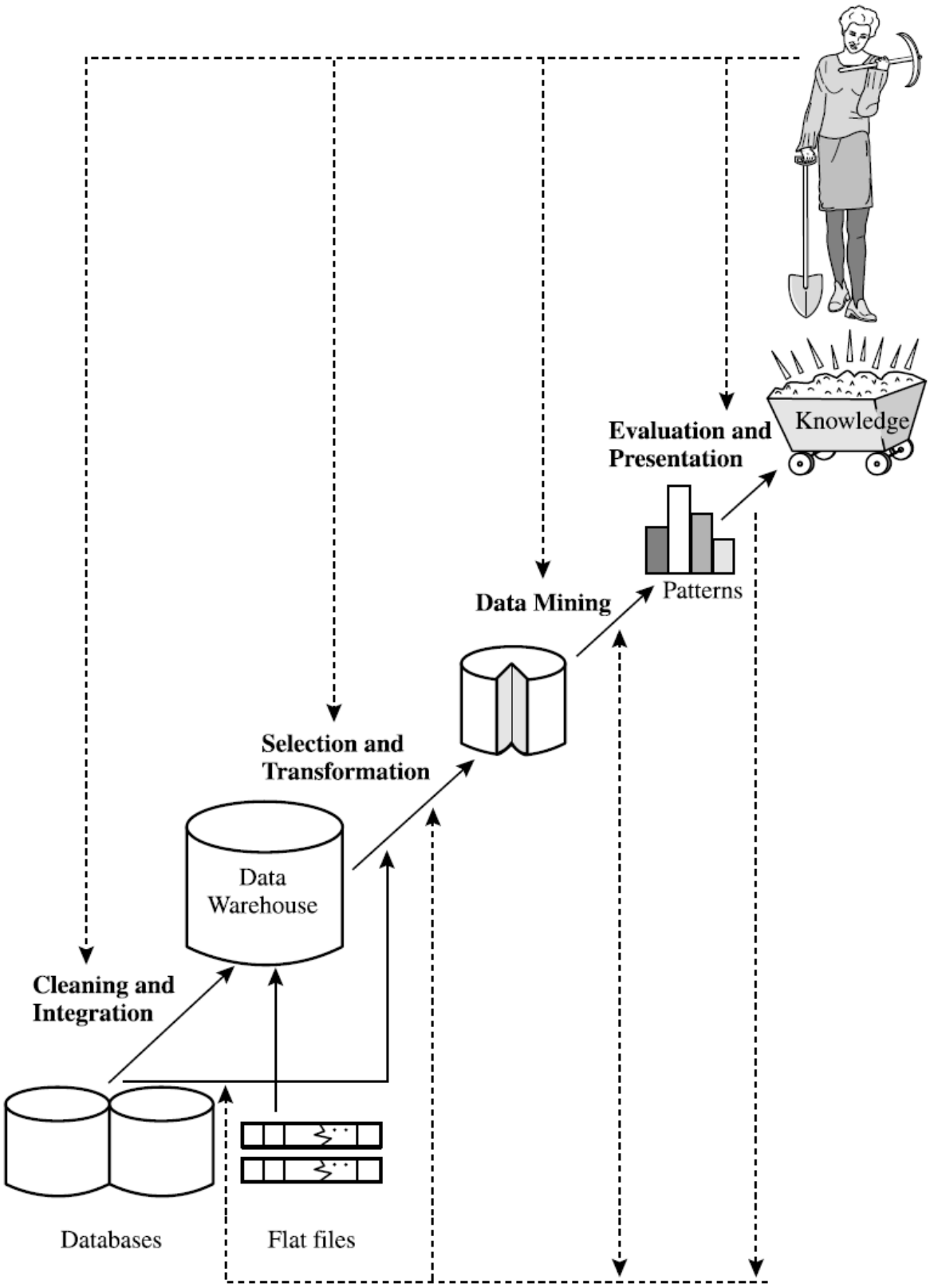
- Knowledge discovery from databases (KDD)
- Knowledge extraction
- Data pattern analysis
- Data archeology
- Data dredging

The most common synonym is **"Knowledge Discovery in Databases (KDD)"**.

The KDD Process (Knowledge Discovery in Databases)

The **KDD process** involves several steps for the systematic extraction of knowledge from databases. The process is as follows:

1. **Data Cleaning**: Removing noise and inconsistent data.
2. **Data Integration**: Combining data from multiple sources.
3. **Data Selection**: Retrieving the relevant data for the analysis task.
4. **Data Transformation**: Transforming or consolidating data into appropriate forms for mining. This can involve summary or aggregation operations.
5. **Data Mining**: Applying mining methods to extract patterns from the data.
6. **Pattern Evaluation**: Identifying truly interesting patterns based on predefined measures.
7. **Knowledge Representation**: Using visualization and knowledge representation techniques to present the discovered knowledge.


Evaluation and
Presentation
Knowledge
Patterns
Data Mining
Selection and
Transformation
Data
Warehouse
Cleaning and
Integration
Databases
Flat files

This flow illustrates how raw data is processed through various stages, leading to the discovery of valuable patterns and insights.

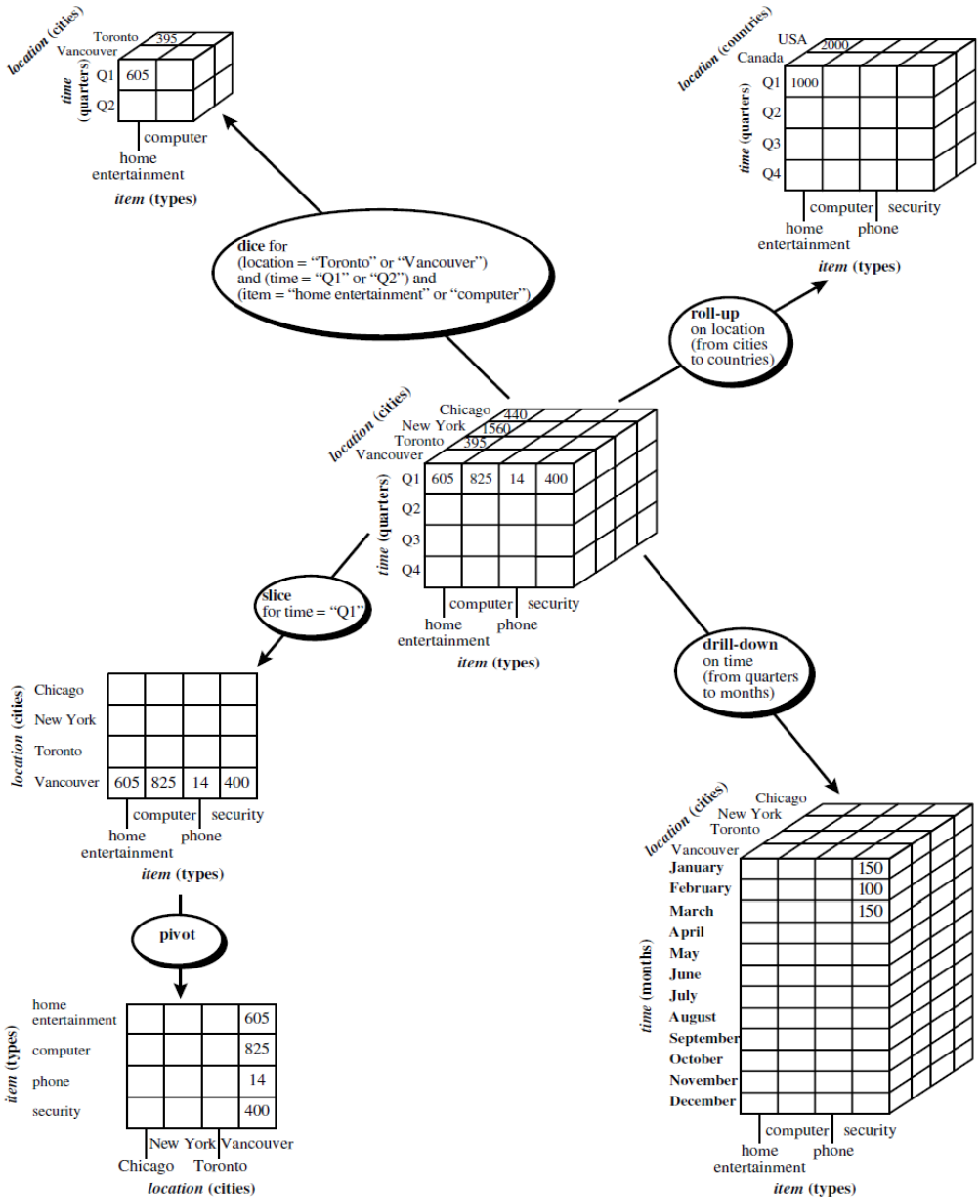
3.Explain OLAP operations with a neat sketch.

A) **OLAP (Online Analytical Processing)** operations allow users to interactively analyze multidimensional data from multiple perspectives. OLAP systems support complex queries that help in decision-making. The primary OLAP operations include **Roll-up**, **Drill-down**, **Slice**, **Dice**, and **Pivot**. These operations enable users to navigate and manipulate the data cube.

Here are the main OLAP operations:

1. **Roll-up**: This operation performs aggregation by climbing up a concept hierarchy or by reducing the dimensions. It is a way to summarize data, showing less detailed data.
    - Example: Aggregating sales data from the city level to the country level.
2. **Drill-down**: The opposite of roll-up, drill-down navigates from less detailed data to more detailed data by either stepping down a hierarchy or adding dimensions.
    - Example: Breaking down sales data from country to city, or from year to quarter.
3. **Slice**: The slice operation selects a single dimension from the cube, creating a sub-cube. It allows a specific portion of the data to be viewed.
    - Example: Viewing sales data for a specific time period like Q1.
4. **Dice**: The dice operation creates a sub-cube by selecting two or more dimensions and applying specific selection criteria.
    - Example: Viewing sales data for specific items in certain regions during specific quarters.
5. **Pivot (Rotate)**: This operation rotates the data axes to provide a different view of the data. It helps in reorganizing the data for easier analysis.

- Example: Rotating the time and item axes in a 2D slice to view the data differently.



4.Write about basic statistical description of a data.

A) Statistical descriptions provide essential summaries of data, offering insights into central tendencies, dispersion, and data distribution. These descriptions help identify key properties of the data, and they are vital in highlighting potential outliers or noise.

Here are the key elements in basic statistical descriptions of data:

1. Measures of Central Tendency

These measures help identify the center or the typical value of a dataset.

- **Mean (Average)**: The mean is the arithmetic average of a set of values.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

WEIGHTED MEAN:

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$$

called as weighted arithmetic mean.

Use - trimmed mean because of the disadvantage of using mean is its avoid 2% of high and low sensitivity to extreme values.

For skewed (asymmetric) data, center of data is median.

(for calculating median) N values are in sorted order,

if N is even-> avg of the middle two numbers

N is odd-> center value

MEDIAN:

$$median = L_1 + \left(\frac{N/2 - (\sum freq)_l}{freq_{median}} \right) width$$

L1->lower boundary of the median interval

N-> No. of values

(Σfreq)l->sum of all the intervals that are lower than the median intervals that are lower than the median interval.

MODE:

i.e; that occurs most frequently in the set.

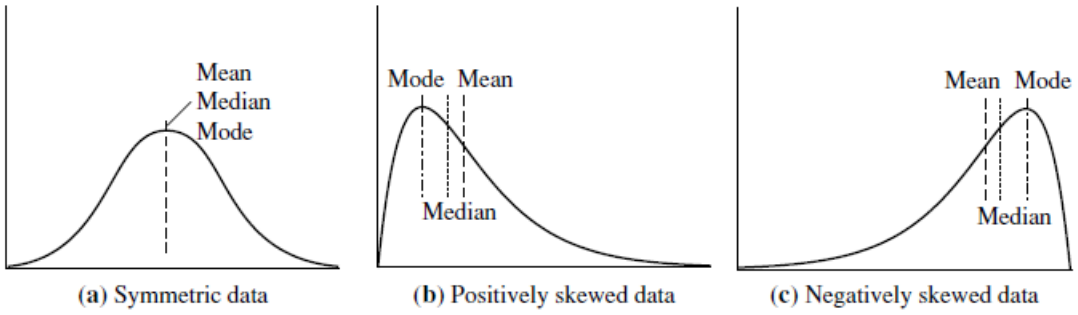
Data sets with one, two (or) three modes are called unimodal, bimodal and trimodal & multimodal

$$\text{mean-mode} = 3*(\text{mean-median})$$

Midrange:

The midrange is the average of the smallest and largest values in the data set.

$$\text{Midrange}=\text{Max Value}+\text{Min Value}/2$$



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

2. Measures of Dispersion

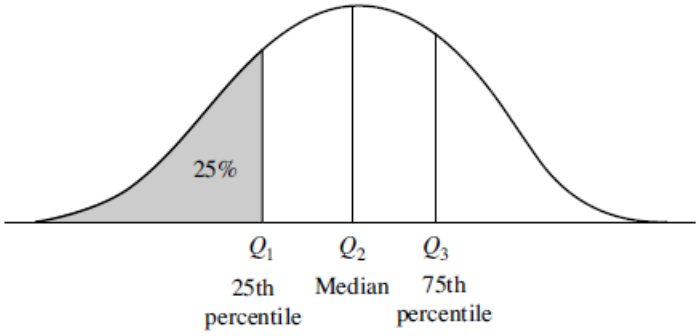
Dispersion measures describe the spread of the data points, indicating how much the data varies.

Range: The range is the difference between the largest and smallest values in the dataset.

$$\text{Range}=\text{Max Value}-\text{Min Value}$$

Quartiles: Quartiles divide the data into four equal parts:

- **Q1 (First Quartile)**: 25% of the data lies below Q1.
- **Q2 (Second Quartile/Median)**: 50% of the data lies below Q2.
- **Q3 (Third Quartile)**: 75% of the data lies below Q3.



Interquartile Range (IQR): The IQR measures the spread of the middle 50% of the data.

$$IQR=Q3-Q1$$

Variance: Variance measures how much the data points differ from the mean. It is the average of the squared differences from the mean.

$$Variance=\sum(Xi-Mean)^2/n$$

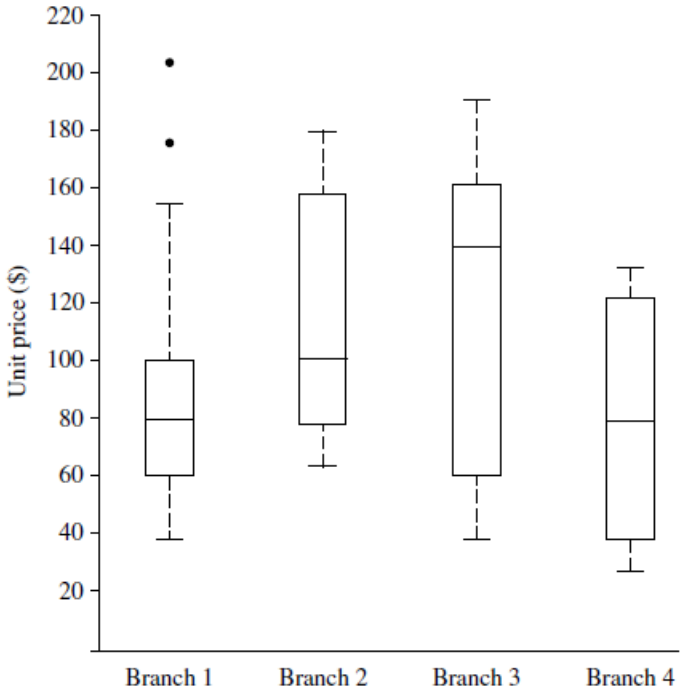
Standard Deviation: The standard deviation is the square root of the variance and is used to quantify the amount of variation or dispersion in a dataset.

$$Standard\ Deviation=\sqrt{}\ Variance$$

3. FIVE-NUMBER SUMMARY, BOXPLOTS, AND OUTLIERS

The F**IVE-NUMBER SUMMARY** of a distribution consists of the median (*Q*2), the quartiles *Q*1 and *Q*3, and the smallest and largest individual observations, written in the order of *Minimum*, *Q*1, *Median*, *Q*3, *Maximum*.

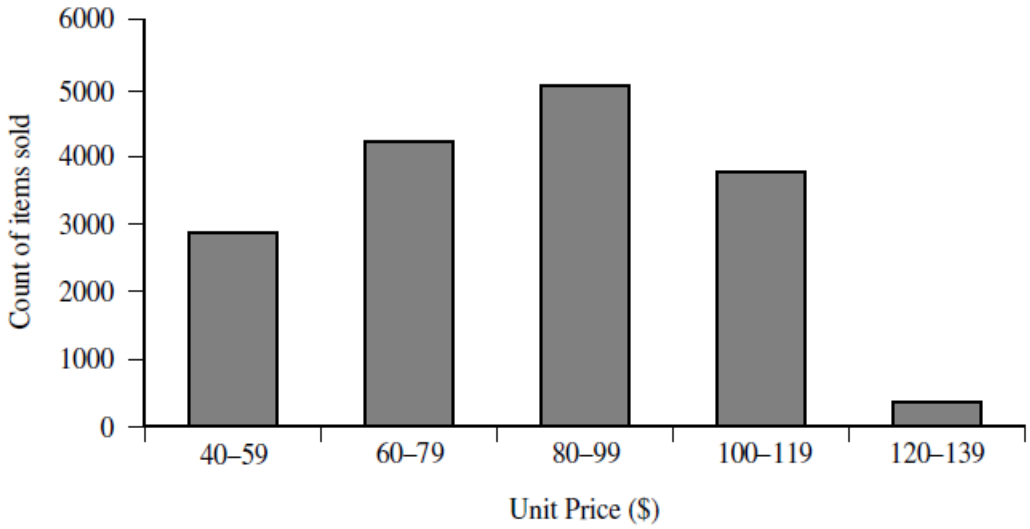
BOXPLOTS are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:



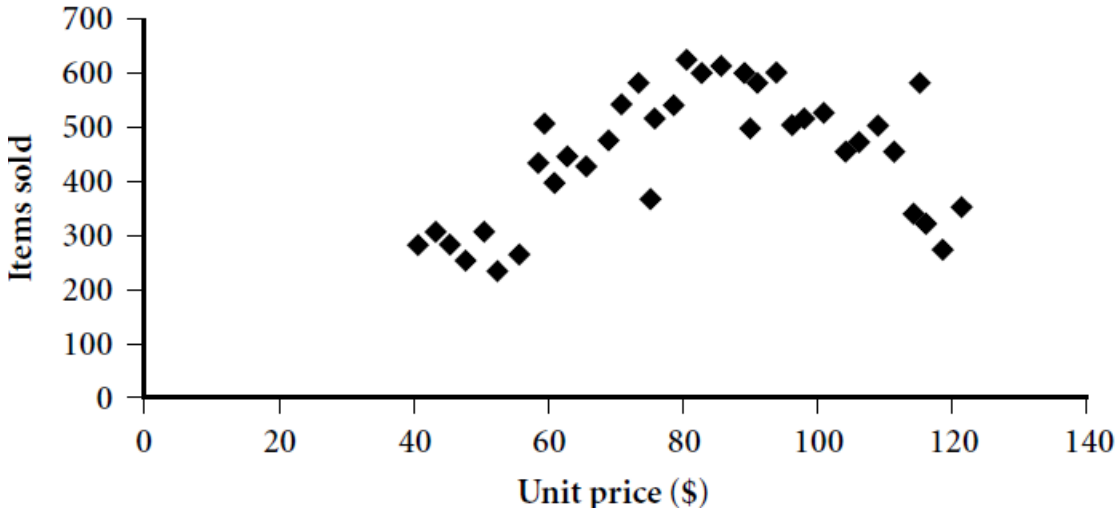
4. Graphical Representation

Basic statistical descriptions are often visualized using the following tools:

- **Boxplots**: Incorporates the five-number summary and highlights outliers.
- **Histograms**: Show the frequency distribution of data.



- **Scatter Plots**: Visualize relationships between two numerical variables



A scatter plot

- **Quantile-Quantile (Q-Q) Plots**: Compare the quantiles of two distributions.

5.what is Data warehousing ?Discuss in detail.

A) A **data warehouse** is a central repository of integrated data from multiple sources, designed to support management's decision-making processes. The data is stored in a way that allows users to analyze it from different perspectives using OLAP (Online Analytical Processing) tools. Data warehouses are constructed to allow fast, flexible, and interactive analysis of large amounts of data.

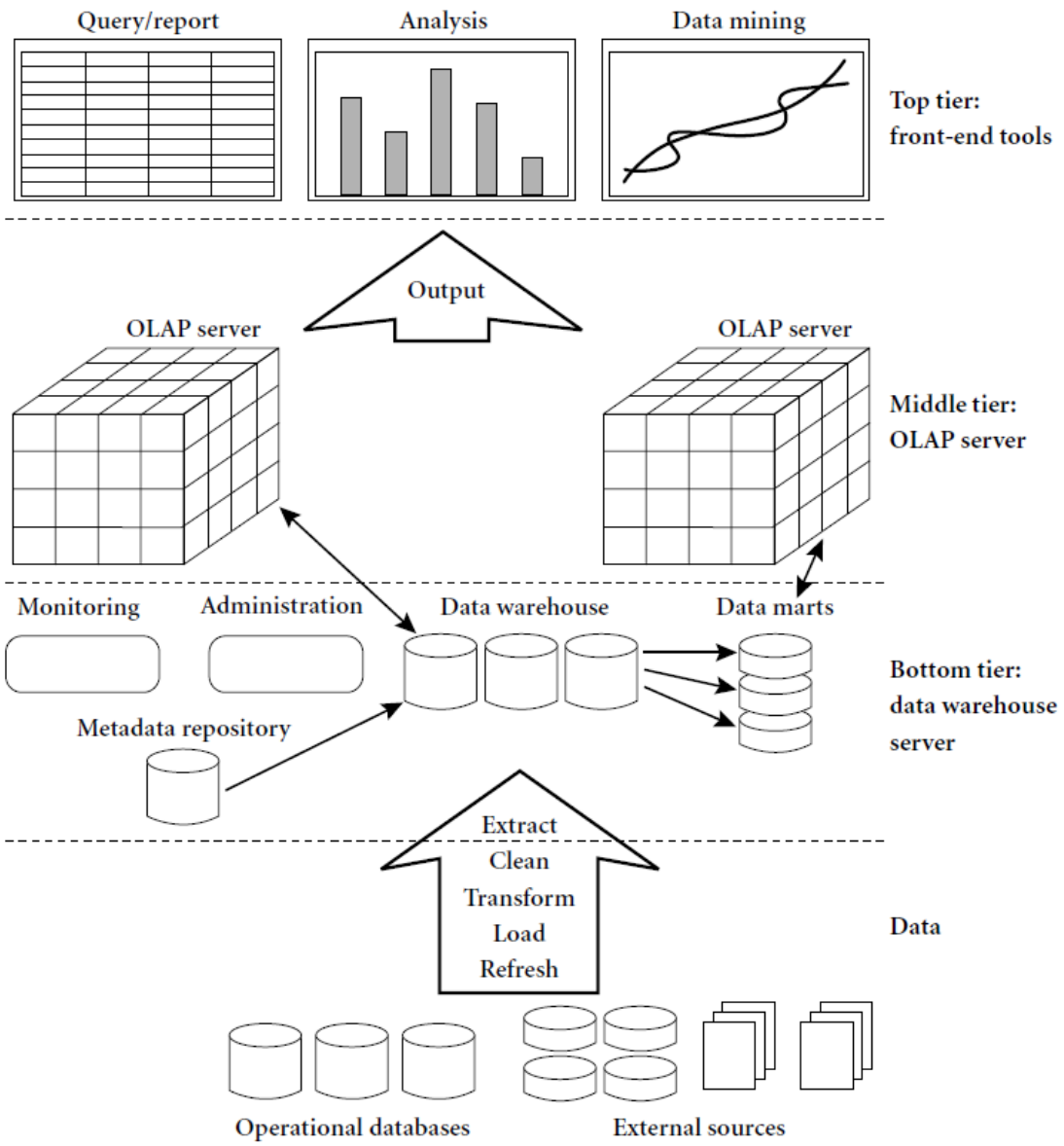
Characteristics of Data Warehousing

1. **Subject-Oriented**: A data warehouse is organized around major subjects or areas of interest for a business, such as customers, sales, or inventory. This makes it easier to analyze data related to a particular topic.
2. **Integrated**: Data in the warehouse is collected from various sources, cleaned, and integrated to maintain consistency across all formats. This involves processes like data cleaning, data integration, and data transformation.
3. **Time-Variant**: Data is stored with a historical perspective, allowing for time-based analysis (e.g., sales trends over time). This enables users to track changes and analyze historical data for better decision-making.

4. **Non-Volatile**: Once data is entered into the data warehouse, it does not change. It is read-only and does not require frequent updates like transactional databases. The data warehouse supports two main operations: data loading and data access.

Data Warehouse Architecture

The typical architecture of a data warehouse is divided into three tiers:



What is Data Warehousing?

A **data warehouse** is a central repository of integrated data from multiple sources, designed to support management's decision-making processes. The data is stored in a way that allows users to analyze it from different perspectives using OLAP (Online Analytical Processing) tools. Data warehouses are constructed to allow fast, flexible, and interactive analysis of large amounts of data.

Characteristics of Data Warehousing

1. **Subject-Oriented**: A data warehouse is organized around major subjects or areas of interest for a business, such as customers, sales, or inventory. This makes it easier to analyze data related to a particular topic.
2. **Integrated**: Data in the warehouse is collected from various sources, cleaned, and integrated to maintain consistency across all formats. This involves processes like data cleaning, data integration, and data transformation.
3. **Time-Variant**: Data is stored with a historical perspective, allowing for time-based analysis (e.g., sales trends over time). This enables users to track changes and analyze historical data for better decision-making.
4. **Non-Volatile**: Once data is entered into the data warehouse, it does not change. It is read-only and does not require frequent updates like transactional databases. The data warehouse supports two main operations: data loading and data access.

Data Warehouse Architecture

The typical architecture of a data warehouse is divided into three tiers:

1. **Bottom Tier (Warehouse Database Server)**: This is the **storage layer**, which is almost always a relational database. It holds the data extracted from operational databases and other external sources. The data is cleaned, integrated, and transformed before loading into the warehouse. Tools such as **ETL (Extract, Transform, Load)** are used for this purpose.

2. **Middle Tier (OLAP Server)**: The middle tier consists of an OLAP server that provides a multidimensional view of data. The OLAP server can be either **ROLAP (Relational OLAP)** or **MOLAP (Multidimensional OLAP)**. ROLAP is an extended relational database system that handles multidimensional data, while MOLAP directly operates on multidimensional data storage.
3. **Top Tier (Client Layer)**: The top layer is the **front-end** client interface, which consists of query, reporting, and analysis tools. Users interact with the data warehouse through dashboards, reports, or data mining tools.

Data Warehouse Models

1. **Enterprise Warehouse**: This is a comprehensive data warehouse that covers all data related to the entire organization. It integrates data from multiple operational systems and external sources, providing a unified, corporate-wide view.
2. **Data Mart**: A data mart is a subset of the data warehouse designed for a specific group of users. Data marts are smaller, subject-specific, and cater to departments like marketing or finance.
    - **Dependent Data Mart**: Sourced from an enterprise data warehouse.
    - **Independent Data Mart**: Built from operational data sources directly.
3. **Virtual Warehouse**: A virtual warehouse consists of views created over operational databases. It provides a logical representation of data rather than physically storing it, which reduces storage costs but may impact query performance.

Multidimensional Data Model

The **multidimensional data model** is the core of data warehousing and OLAP systems. This model organizes data into **facts** and **dimensions**:

- **Fact Table**: Contains the quantitative data (e.g., sales revenue, units sold) that users want to analyze.
- **Dimension Tables**: Contain descriptive attributes related to the facts (e.g., time, location, product).

A **data cube** is a common representation of multidimensional data, where each dimension represents a different attribute, and facts are aggregated at different levels of granularity. For example, sales data can be analyzed by time (daily, monthly), location (city, country), and product.

6.Discuss about Data mining Functionalities.

A) Data Mining Functionalities

Data mining functionalities describe the various types of patterns that can be discovered through the data mining process. These functionalities provide a broad framework to extract useful and actionable insights from vast datasets. The main categories of data mining functionalities are:

1. Concept/Class Description

Concept or class description involves the summarization and characterization of the data associated with a class or concept. There are two main functionalities within this category:

- **Data Characterization**: This is a summarization of the general characteristics of a target class. It provides a concise description of the data, which can be presented in the form of graphs, tables, charts, or multidimensional data cubes.
    - **Example**: Summarizing the characteristics of high-income customers (e.g., their age distribution, spending habits).
- **Data Discrimination**: This compares the target class data with another class or set of classes, providing a contrast between the two.
    - **Example**: Comparing the purchasing behavior of high-income customers with that of middle-income customers.

2. Association Analysis

Association analysis uncovers patterns that reveal the relationship between different variables in the dataset. It discovers association rules that show attribute-value conditions that frequently occur together.

- **Association Rules**: These are rules of the form **X ⇒ Y**, meaning that whenever **X** occurs, **Y** is likely to occur. It is often used in market basket analysis.
    - **Example**: If a customer buys bread, they are likely to buy butter (bread ⇒ butter).
- **Support**: This measures how frequently the rule occurs in the dataset.
    - **Example**: "Bread ⇒ Butter" might have 1% support if it occurs in 1% of the transactions.
- **Confidence**: This indicates how often the rule is true.
    - **Example**: The rule "Bread ⇒ Butter" might have 50% confidence if, in 50% of the cases where bread is bought, butter is also bought.

3. Classification and Prediction

Classification and prediction involve analyzing a dataset to build a model that can classify future data points or predict missing or future values.

- **Classification**: It is the process of finding a model that describes and distinguishes data classes or concepts. The model is used to predict the categorical class labels of new data points.
    - **Example**: Classifying email messages as spam or non-spam.
- **Prediction**: It involves predicting a numerical value based on the relationships identified in the data.
    - **Example**: Predicting house prices based on features like location, size, and number of bedrooms.
- **Decision Trees** and **Neural Networks** are common techniques used for classification. In a decision tree, each node represents a test on an attribute, and branches represent the outcome of the test.

4. Cluster Analysis

Clustering involves grouping a set of data objects so that objects in the same group (called a cluster) are more similar to each other than to objects in other groups.

- **Clustering**: Unlike classification, clustering is an unsupervised learning process where the class labels are not known in advance. It finds natural groupings within the data.
    - **Example**: Grouping customers based on purchasing behavior without prior knowledge of their categories.
- **Maximizing Intra-Class Similarity and Minimizing Inter-Class Similarity**: The goal of clustering is to maximize the similarity of objects within the same cluster while minimizing similarity between different clusters.

5. Outlier Analysis

Outliers are data objects that do not fit into the general pattern of the dataset. These deviations from the norm are often seen as anomalies or exceptions.

- **Outlier Detection**: Identifies data points that significantly differ from the rest of the data.
    - **Example**: In fraud detection, outlier analysis can be used to detect unusual transactions.
- **Interesting Outliers**: In many cases, outliers are discarded as noise, but in some applications like fraud detection, outliers can represent very important and interesting data points.

UNIT-2

1.why Data preprocessing ? Discuss in detail.

A) Data preprocessing is essential because real-world data is often **incomplete, noisy, and inconsistent**. It is important to improve the quality of the data before applying data mining techniques. If the data is not properly cleaned and transformed, the results of any data mining process might be inaccurate or misleading.

The **quality of data** directly impacts the success of data mining. The goal of data preprocessing is to ensure that the data satisfies certain criteria such as **accuracy, completeness, consistency, timeliness, believability, and interpretability**.

Major Tasks in Data Preprocessing

The four main tasks in data preprocessing are:

1. **Data Cleaning**
2. **Data Integration**
3. **Data Reduction**
4. **Data Transformation**

1. Data Cleaning

Data cleaning addresses problems like **missing values, noise, and inconsistencies** in the data. Since real-world data is often incomplete or contains errors, data cleaning techniques are essential for improving data quality. Key processes involved in data cleaning include:

- **Handling Missing Values**:
    - **Ignore the tuple**: Discarding records with missing values (ineffective if many records are incomplete).
    - **Fill in missing values manually**: Time-consuming and impractical for large datasets.
    - **Use a global constant**: Filling with a default value (e.g., "unknown" or 0).
    - **Use a measure of central tendency**: Filling missing values with the mean, median, or mode.
    - **Use the most probable value**: Predict the missing value using statistical models.
- **Handling Noisy Data**: Noise refers to random errors or variability in a dataset. Techniques for smoothing noisy data include:
    - **Binning**: Data values are sorted and then divided into bins. Smoothing is done using bin means, medians, or boundaries.
    - **Regression**: Fits data to a function (e.g., linear regression) to reduce noise.

    - **Clustering**: Groups similar data points together, and values that fall outside the clusters can be considered outliers.

2. Data Integration

Data integration involves merging data from multiple sources to create a unified dataset. It helps to reduce redundancies and inconsistencies, leading to more accurate and comprehensive data for analysis.

Challenges in data integration include:

- **Entity Identification Problem**: Matching equivalent real-world entities from different sources.
    - Example: Different databases may use different names for the same customer (e.g., “Customer_ID” in one database and “Client_ID” in another).
- **Redundancy and Correlation Analysis**: Detecting and handling redundancies between attributes, ensuring that unnecessary or highly correlated attributes are removed.
    - For nominal data, the **Chi-Square Test** can be used to detect redundancies.
    - For numeric data, the **correlation coefficient** measures how strongly one attribute implies another.
- **Tuple Duplication**: Detecting and removing duplicate records that occur in different datasets or databases.
- **Data Value Conflict Detection and Resolution**: Resolving differences in data values across sources, such as differences in scaling or encoding.

3. Data Reduction

Data reduction reduces the data volume while maintaining the integrity and quality of the original dataset. This is important for improving the efficiency of data mining algorithms by reducing the computational cost.

Data reduction strategies include:

- **Dimensionality Reduction**: Reducing the number of attributes in the dataset.
    - **Wavelet Transforms**: A mathematical function used to break down data into different frequency components, making it easier to manage.
    - **Principal Component Analysis (PCA)**: Transforms data into a set of linearly uncorrelated components, reducing the dataset's dimensionality while preserving as much variance as possible.
    - **Attribute Subset Selection**: Methods like **stepwise forward selection**, **stepwise backward elimination**, and **decision tree induction** are used to remove irrelevant or redundant attributes.
- **Numerosity Reduction**: Replacing the original data with a smaller representation, such as:
    - **Parametric Methods**: Using models like **regression** and **log-linear models** to approximate the data.
    - **Non-Parametric Methods**: Techniques such as **histograms**, **clustering**, and **sampling** are used for data reduction.

4. Data Transformation

Data transformation converts the data into a format suitable for data mining. It includes processes like **smoothing, aggregation, normalization, discretization**, and **concept hierarchy generation**.

- **Smoothing**: Removing noise from data (e.g., using binning or regression).
- **Aggregation**: Summarizing or combining data (e.g., daily sales data can be aggregated into monthly or yearly totals).
- **Normalization**: Scaling data into a smaller range, such as 0.0 to 1.0 or -1.0 to 1.0. This is often done using methods like:
    - **Min-Max Normalization**
    - **Z-Score Normalization**
    - **Decimal Scaling**

- **Discretization**: Converting continuous attributes into categorical ones by grouping values into intervals (e.g., converting age into ranges like 0–10, 11–20).
    - **Binning**, **histogram analysis**, and **clustering** can be used for discretization.
- **Concept Hierarchy Generation**: Organizing data into a hierarchy by grouping values into higher-level concepts. For instance, a geographic hierarchy can move from **street → city → state → country**.