Home Assignment: -

1.**ETL Process:**

- **Extract:** Data is collected from various sources.
- **Transform:** Data is cleaned, aggregated, and converted into a format suitable for analysis.
- **Load:** Transformed data is loaded into a target system, usually a data warehouse.

2.Data Warehouse vs. Data Lake vs. Data Mart:

- **Data Warehouse:** A structured repository of historical data designed for querying and analysis. It integrates data from various sources and is optimized for read access.
- **Data Lake:** A large, unstructured repository that stores raw data in its native format until needed. It supports various types of data, including structured, semi-structured, and unstructured data.
- **Data Mart:** A subset of a data warehouse tailored for a specific business line or team. It focuses on a specific subject area or department, providing relevant data to users.

3.**OLAP vs. OLTP:**

- **OLAP (Online Analytical Processing):** Designed for complex queries and data analysis, such as generating reports and performing data mining. It is optimized for read-heavy operations and supports multi-dimensional analysis.
- **OLTP (Online Transaction Processing):** Designed for transactional tasks and managing day-to-day operations. It is optimized for high-speed, read-write operations and supports a large number of short online transaction queries.

4.**Covariance vs. Correlation:**

- **Covariance:** Measures the degree to which two variables change together. It can be positive (variables increase together) or negative (one variable increases as the other decreases). The magnitude depends on the units of the variables.
- **Correlation:** Measures the strength and direction of a linear relationship between two variables. It is normalized (ranges from -1 to 1), making it easier to interpret regardless of the variable units.

5.**Schemas of Data Warehousing:**

- **Star Schema:** Central fact table connected to multiple dimension tables. It is simple and efficient for queries.
- **Snowflake Schema:** A normalized form of the star schema where dimension tables are split into related sub-dimensions, reducing redundancy.
- **Galaxy Schema (or Fact Constellation):** Multiple fact tables shared by multiple dimension tables, allowing complex queries and multiple data mart configurations.

6.**KDD Process (Knowledge Discovery in Databases):**

- **Selection:** Choosing relevant data from a database.
- **Preprocessing:** Cleaning and transforming data to correct inconsistencies.

- **Transformation:** Converting data into suitable formats for mining.
- **Data Mining:** Applying algorithms to extract patterns and insights from data.
- **Evaluation:** Assessing the patterns to ensure they are useful and meaningful.
- **Presentation:** Visualizing the results to make them understandable and actionable.

7.**Data Warehouse Architecture:**

- **Data Sources:** Where data originates.
- **ETL Layer:** Extracts, transforms, and loads data from sources into the data warehouse.
- **Data Warehouse:** Central repository where data is stored in a structured format.
- **OLAP Layer:** Provides tools and frameworks for data analysis and querying.
- **Data Marts:** Specialized subsets of the data warehouse for specific business needs.
- **Front-End Tools:** Visualization and reporting tools for end-users to interact with data.

8.**Five-Number Summary:**

- **Minimum:** The smallest value in the data set.
- **First Quartile (Q1):** The value below which 25% of the data falls.
- **Median (Q2):** The middle value that divides the data set into two equal halves.
- **Third Quartile (Q3):** The value below which 75% of the data falls.
- **Maximum:** The largest value in the data set.

9.**Partial Materialization Methods for Data Cube Computation:**

- **Lazy Aggregation:** Computes data cube cells on demand, only when queried.
- **Eager Aggregation:** Pre-computes and stores aggregate values for all possible cells in the cube ahead of time.

10.Simple Regression Analysis:

- **Simple Linear Regression:** Models the relationship between a dependent variable (Y) and an independent variable (X) using a linear equation.