Task 4:

1. What is ETL? In detail.

- ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a data warehouse or other target system.
- ETL was introduced as a process for integrating and loading data for computation and analysis, eventually becoming the primary method to process data for data warehousing projects.

❖ ETL is often used by an organization to:

- Extract data from legacy systems
- Cleanse the data to improve data quality and establish consistency
- Load data into a target database

➢ **Extract**

During data extraction, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from a variety of data sources, which can be structured or unstructured. Those sources include but are not limited to:

- SQL or NoSQL servers

➢ **Transform**

In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase can involve the following tasks:

- Filtering, cleansing, de-duplicating, validating, and authenticating the data.
- Performing calculations, translations, or summarizations based on the raw data.
- Removing, encrypting, or protecting data governed by industry or governmental regulators
- Formatting the data into tables or joined tables to match the schema of the target data warehouse.

➢ **Load**

In this last step, the transformed data is moved from the staging area into a target data warehouse.

2. What is ELT? in detail.

ELT, which stands for “Extract, Load, Transform,” is another type of data integration.

- **Extract**

During data extraction, data is copied or exported from source locations to a staging area. The data set can consist of many data types and come from virtually any structured or unstructured source, including but not limited to:

- SQL or NoSQL servers

That said, it is more typically used with unstructured data.

- **Load**

In this step, the transformed data is moved from the staging area into a data storage area, such as a data warehouse or data lake.

- **Transform**

In this stage, a schema-on-write approach is employed, which applies the schema for the data using SQL, or transforms the data, prior to analysis. This stage can involve the following:

- Filtering, cleansing, de-duplicating, validating and authenticating the data.
- Performing calculations, translations, data analysis or summaries based on the raw data. Removing, encrypting, hiding, or otherwise protecting data governed by government or industry regulations.
- Formatting the data into tables or joined tables based on the schema deployed in the warehouse.

3. ETL Tools (any 3)

- Informatica PowerCenter.
- Apache Airflow.
- IBM Infosphere Datastage.
- Oracle Data Integrator.

4. 3 Tier Architecture in DE

Data tier

The data tier, sometimes called database tier, data access tier or back-end, is where the information processed by the application is stored and managed. This can be a relational database management system.

Application tier

The application tier, also known as the logic tier or middle tier, is the heart of the application. In this tier, information collected in the presentation tier is processed - sometimes against other information in the data tier - using business logic, a specific set of business rules.

Presentation tier

The presentation tier is the user interface and communication layer of the application, where the end user interacts with the application. Its main purpose is to display information to and collect information from the user.

Task 5:

1. What is Historical Load

- Historical load is the one-time initial load of data that the Source already had before the creation of the Pipeline.

- In data engineering, "Historical Load" refers to the process of copying data from a source system that covers a specific historical period, such as a previous month, year, or even several years. It is a method of data integration that allows organizations to analyze and report on past events and trends, and to perform historical analysis.

- Historical Loads are typically performed as part of a larger data integration project, such as building a data warehouse or data mart. The process involves extracting data from the source system, transforming it as necessary to meet the requirements of the target system, and then loading it into the target system.
- Historical Loads are important for organizations that need to perform trend analysis, historical reporting, or other types of retrospective analysis. They are also useful for organizations that need to migrate from one system to another, as they allow historical data to be preserved and moved to the new system.

2. What is Full Load

- During a Full Load, all the data from the source system is extracted and transformed as necessary, and then loaded into the target system. This process is typically used when creating a new data warehouse or data mart, or when updating an existing one with fresh data.

- In data engineering, "Full Load" refers to a process of copying all of the data from a source system to a target system, without any filters or conditions applied.

3. What is Incremental Load

- In data engineering, "Incremental Load" refers to the process of copying only the data that has changed since the last load from a source system to a target system. It is a method of data integration that is commonly used in ETL (Extract, Transform, Load)

processes to move smaller amounts of data between systems in a more efficient and timely manner.

- During an Incremental Load, the system identifies and extracts only the data that has been updated, inserted, or deleted since the last load, as determined by a timestamp or other marker. This data is then transformed as necessary and loaded into the target system.

- Incremental Loads are generally faster and more efficient than Full Loads because they process less data, require less storage, and consume fewer resources. However, they require more complex logic and infrastructure to manage, as they must be able to track changes to the data and identify which records have been updated since the last load.

- Incremental Loads are commonly used in data integration scenarios where data is updated frequently, such as in real-time reporting, transactional databases, or IoT (Internet of Things) systems. By only processing the changes to the data, Incremental Loads allow organizations to keep their systems up to date with minimal latency and ensure that reports and analytics are based on the latest data.