TASK # 4 & 5

ETL (**E**xtract, **T**ransform, and **L**oad). It is a process used in data warehousing and business intelligence to move data from various sources, transform it to fit business needs, and load it into a target system.

The ETL process typically involves the following steps:

- **Extract**: Data is extracted from various sources, such as databases, flat files(simple files working as a database), or APIs.

- **Transform**: The extracted data is transformed and cleaned to fit the target system's schema and business rules. This may involve tasks such as data cleansing, data aggregation, and data enrichment.
  In general, ETL transformations can be performed using two main approaches:

  - **Push-down processing**: In push-down processing, the **transformation logic is executed directly in the source** or **target database** system, leveraging the processing power of the database server. This can result in faster and more efficient transformations, as the data does not need to be moved across the network.

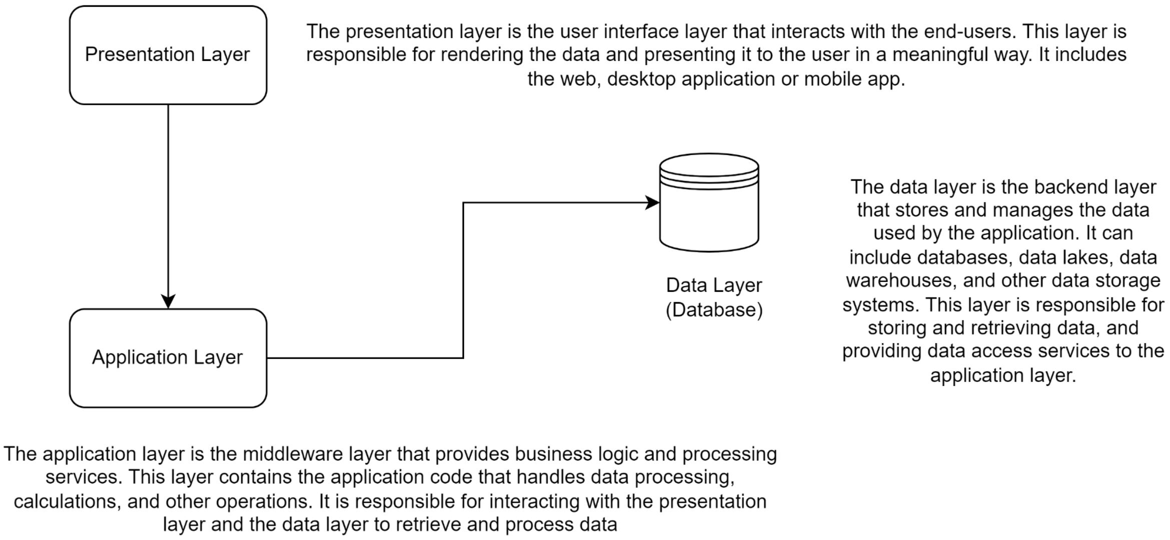
  - **Pull-up processing**: In pull-up processing, the **transformation logic is executed in the ETL tool**, after the data has been extracted from the source system and before it is loaded into the target system.

- **Load**: The transformed data is loaded into the target system, such as a data warehouse or data lake, where it can be used for reporting, analytics, and other business intelligence applications.

<u>ELT & ETL:</u>

ELT (**E**xtract, **L**oad, **T**ransform) is a **variation** of the ETL (Extract, Transform, Load) process used in data warehousing and business intelligence. The **main difference** between ETL and ELT **is the order** in which the data is transformed.

3 Tier Architecture in Data Engineering



The presentation layer is the user interface layer that interacts with the end-users. This layer is responsible for rendering the data and presenting it to the user in a meaningful way. It includes the web, desktop application or mobile app.

The data layer is the backend layer that stores and manages the data used by the application. It can include databases, data lakes, data warehouses, and other data storage systems. This layer is responsible for storing and retrieving data, and providing data access services to the application layer.

The application layer is the middleware layer that provides business logic and processing services. This layer contains the application code that handles data processing, calculations, and other operations. It is responsible for interacting with the presentation layer and the data layer to retrieve and process data

Some popular ETL tools are:

- **Informatica PowerCenter**: Informatica PowerCenter is a widely used commercial ETL tool that offers a range of features for data integration, including data profiling, data quality, and data governance.

- **Microsoft SQL Server Integration Services (SSIS)**: SSIS is a Microsoft ETL tool that comes with SQL Server. It allows you to extract data from various sources, transform it, and load it into a target system.

- **Talend Open Studio**: Talend Open Studio is a popular open-source ETL tool that provides a comprehensive set of features for data integration, including data profiling, data quality, and data governance.

Historical Load:

- A historical load, in the context of data warehousing and ETL processes, refers to the process of loading historical data into a target system. This is typically done when creating a data warehouse or data mart, where historical data is needed for analysis and reporting purposes.

- One of the challenges of performing a historical load is dealing with large volumes of data. This can require specialised tools and techniques for data extraction, transformation, and loading, such as parallel processing and data partitioning.

Full Load:

- A full load in ETL refers to the process of loading all of the data from a source system into a target system, regardless of whether the data has been previously loaded. Full loads are typically used when first setting up a data integration process or when making major changes to the source or target systems.

Incremental load:

- Incremental load in ETL is a process of extracting, transforming, and loading only the data that has changed since the last load. This approach helps to reduce the amount of data that needs to be processed and loaded, which can improve performance and reduce the risk of duplicating data.

- Instead of loading all of the data every time, the ETL process only loads the data that has been added, updated, or deleted since the last load.

This can be accomplished using a variety of techniques, such as:

- **Using timestamps or version numbers:** Many source systems include timestamps or version numbers that can be used to identify when data has changed. The ETL process can use these identifiers to extract and load only the changed data.

- **Using change data capture (CDC) technologies**: CDC technologies capture and record changes to a database in real-time, making it possible to identify and extract only the changed data.

- **Using delta tables:** Delta tables are tables that contain only the changed data since the last load. The ETL process can use these tables to extract and load only the changed data.