***Task -4 & 5 2023/03/16***

***ETL:*** ETL stands for Extract, Transform, Load. It refers to a process of extracting data from different sources, transforming it into a structured format, and loading it into a data warehouse or another target system. The ETL process plays a critical role in data integration and allows organizations to consolidate data from various sources into a single source of truth. The three main stages of the ETL process are:

*Extraction:* In this stage, data is extracted from multiple sources such as databases, files, or web services.

*Transformation:* In this stage, the extracted data is transformed into a format that is suitable for analysis and storage. This includes data cleaning, data enrichment, and data aggregation.

*Loading:* In this stage, the transformed data is loaded into a target system such as a data warehouse or a data lake.

***ELT:*** ELT stands for Extract, Load, Transform. It is a variant of the ETL process that involves loading data into a target system before transforming it. This approach is becoming more popular with the rise of cloud computing and big data technologies, which can handle the large volumes of data involved in ELT. The main advantage of ELT is that it allows organizations to perform data transformations directly on the cloud, reducing the need for expensive data warehousing infrastructure. The three main stages of the ELT process are:

*Extraction:* In this stage, data is extracted from multiple sources such as databases, files, or web services.

*Loading:* In this stage, the extracted data is loaded into a target system such as a cloud-based data warehouse or a data lake.

*Transformation:* In this stage, the loaded data is transformed into a format that is suitable for analysis and storage. This includes data cleaning, data enrichment, and data aggregation.

***Three Tier Architecture in DE:*** Three-tier architecture in data engineering refers to the design of a software application or system that separates its functionality into three layers or tiers:

*Presentation Layer:* This layer is responsible for presenting the data to the end-user through a graphical user interface (GUI).

*Application Layer:* This layer contains the business logic and processes the data that is received from the presentation layer.

*Data Layer:* This layer is responsible for storing and retrieving data from the database.

The main advantage of the three-tier architecture is that it provides a clear separation of concerns between the different layers, making it easier to maintain and scale the application.

***ETL Tools:*** ETL tools are software applications that help organizations automate the ETL process. Here are three popular ETL tools:

*Apache NiFi*: An open-source tool that provides a web-based interface for designing and managing data flows.

*Talend:* A commercial ETL tool that provides a comprehensive set of data integration features, including data profiling, data quality, and data governance.

*Informatica:* A commercial ETL tool that provides a wide range of data integration features, including data profiling, data quality, and master data management.

***Historical Load:*** Historical load refers to the process of loading historical data into a data warehouse or another target system. This involves extracting data from various sources and loading it into a target system to provide a complete picture of historical trends and patterns. Historical data can be used for analysis, reporting, and forecasting purposes.

***Full Load:*** Full load refers to the process of loading all data from the source system into the target system. This is typically done when setting up a new data warehouse or when refreshing the entire data set. The full load process can be time-consuming and resource-intensive, but it ensures that the target system contains a complete and accurate set of data.

***Incremental Load:*** Incremental load refers to the process of loading only the data that has changed since the last time it was loaded into the target system. This is typically done on a regular basis, such as daily or weekly, to keep the target system up-to-date with the latest changes. The incremental load process involves comparing the source system data with the data that is already in the target system and only loading the new or updated records. This helps to minimize the amount of data that needs to be processed and loaded, reducing the time and resources required for the data integration process. Incremental load is a key feature of ETL/ELT tools and is essential for maintaining the accuracy and timeliness of data in a data warehouse or other target system.