By: Zia Ur Rehman

Date: 16-march-2023

**Task#4**

**What is ETL? In detail.**

ETL stands for Extract, Transform, Load, and it is a process of extracting data from various sources, transforming the data into a common format, and loading it into a target database or data warehouse.

Extract: This step involves pulling data from various sources such as databases, files, and applications. The data can be structured or unstructured, and it may be stored in different formats like CSV, XML, JSON, or database tables. In the step first data sources are identified. Then a connection is establish. Data extraction rules are defined. Once the data extraction rules are defined, the ETL tool or extraction code is used to extract the data from the sources. After the data is extracted, it is essential to verify its quality, accuracy, and completeness.

Transform: The transform part of the ETL process refers to the manipulation and processing of data from the extracted source data in a way that makes it usable for the intended destination or application. This is the stage where the extracted data is cleaned, enriched, combined, or aggregated to create a new data set that is useful for analysis, reporting, or decision-making. Transforming data involves several steps, including data cleaning, data validation, data enrichment, data aggregation, and data normalization.

Load: This step involves loading the transformed data into a target database or data warehouse. The target system could be a relational database, a NoSQL database, a data warehouse, or a cloud-based storage system. The data is loaded in batches or real-time, depending on the requirements.

**What is ELT? In detail.**

ELT stands for Extract, Load, and Transform, which is a data integration process used to move and transform data from various sources into a target data warehouse or database. ELT is a variation of the more traditional ETL (Extract, Transform, and Load) process, but the main difference is that in ELT, the transformation step is performed after the data is loaded into the target system. ELT is more flexible than ETL, as organizations can perform transformations on the data at any point in the process, depending on their needs.

Here is a breakdown of each step in the ELT process:

* Extract: The first step in ELT is to extract data from various sources, which can include databases, data lakes, APIs, or flat files. The extracted data may be structured, semi-structured, or unstructured.
* Load: After the data is extracted, it is loaded into a target system such as a data warehouse or database. The target system may be on-premises or in the cloud, depending on the organization's needs.
* Transform: Once the data is loaded into the target system, it is transformed to meet the needs of the organization. Transformation can include data cleaning, data normalization, data aggregation, and data enrichment. The transformed data can then be used for reporting, analysis, and other data-driven tasks.

**3 Tier Architecture in DE**

Three-tier architecture is a common data engineering architecture that is used to separate the components of an application into three separate layers or tiers. Each tier has a specific role in the data processing and management process, and the separation of responsibilities makes it easier to develop, test, deploy, and maintain large-scale applications. The three tiers are:

1. Presentation Tier: The presentation tier is the top-most layer of the architecture and is responsible for presenting data to users in a user-friendly format. This tier includes components like web browsers, mobile apps, and desktop applications, which allow users to interact with the data.
2. Application Tier: The application tier, also known as the logic tier or middle tier, is responsible for processing and managing the data. This tier contains the business logic, data validation, and data processing components that perform various functions on the data. It includes components like APIs, web servers, and application servers.
3. Data Tier: The data tier, also known as the back-end tier, is responsible for storing and retrieving data from various sources. This tier includes databases, data warehouses, and data lakes, which store the data in a structured or unstructured format.

**ETL Tools (any 3)**

1. Apache NiFi: A powerful and flexible ETL tool with a drag-and-drop interface that allows for the creation of data pipelines with minimal coding.
2. Microsoft SQL Server Integration Services (SSIS): A robust ETL tool for data integration and data transformation that is tightly integrated with the Microsoft SQL Server ecosystem.
3. AWS Glue: A fully managed ETL service that makes it easy to move data between data stores, clean and transform data, and load it into data warehouses.

**Task # 5:**

**- What is Historical Load?**

Historical load refers to the process of loading historical data into a data warehouse or database to create a historical record of the organization's activities. This can include data from previous months, years, or even decades, depending on the organization's requirements.

An example of historical load might involve a retail company that wants to analyze its sales data over the past five years. The company would need to load all of its sales data from the past five years into a data warehouse to perform this analysis.

**- What is Full Load?**

Full load is a term used in data engineering to describe the process of loading all the data from a source system into a target system, such as a data warehouse or a database. This is typically done when creating a new data warehouse or when updating an existing one.

An example of a full load might involve a company that wants to create a new data warehouse to store all of its customer data. The company would extract all of its customer data from various source systems, such as transactional databases, customer relationship management (CRM) systems, and marketing automation platforms.

Full load is a time-consuming process that is typically done outside of regular business hours to avoid impacting normal operations. However, it is important because it ensures that the data warehouse contains a complete and accurate copy of all the data from the source systems, which is necessary for effective data analysis and decision-making.

**- What is Incremental Load?**

Incremental load is a term used in data engineering to describe the process of loading only the data that has changed or been added since the last load, instead of loading all the data from a source system into a target system every time. This is typically done to improve the efficiency of data integration and management processes, as it reduces the amount of data that needs to be processed and loaded.

An example of incremental load might involve a company that wants to update its sales data in a data warehouse every day. Instead of loading all the sales data from the source system every day, the company would only load the new or changed sales data since the last load. This could be done by identifying the incremental changes based on timestamps, unique keys, or other criteria.

Incremental load is important because it reduces the amount of time and resources needed to load and process data, as well as the risk of data errors and inconsistencies. It is particularly useful for large data sets and frequent data updates, where full loads would be time-consuming and resource-intensive.