**HASNAIN AJMAL**

**Data Engineering**

**Week 1**

**Day 03: 15/02/23**

**TASK 04**

**ETL:**

ETL stands for Extract, Transform, Load. It is a process of extracting data from multiple sources, transforming it into a consistent format, and then loading it into a target database or data warehouse. The ETL process is used to integrate data from various sources such as databases, flat files, or web services, and prepare it for analysis.

The ETL process has three main stages:

* **Extraction**: In this stage, data is extracted from multiple sources such as databases, flat files, or web services. The data can be extracted in various formats like CSV, XML, JSON, or any other format supported by the source systems.
* **Transformation**: In this stage, the extracted data is transformed into a consistent format that can be loaded into the target system. This involves cleaning, filtering, aggregating, and formatting the data. Transformation can also involve enriching the data by combining it with other data sources or applying business rules.
* **Loading**: In this stage, the transformed data is loaded into the target system, which can be a data warehouse, a database, or any other data storage system.

**ELT:**

ELT stands for Extract, Load, Transform. It is a process of extracting data from various sources and loading it into a target system, where it is transformed for analysis. The ELT process differs from ETL in that the transformation of data occurs within the target system, rather than in a separate transformation layer.

The ELT process has three main stages:

* **Extraction**: In this stage, data is extracted from multiple sources such as databases, flat files, or web services. The data can be extracted in various formats like CSV, XML, JSON, or any other format supported by the source systems.
* **Loading**: In this stage, the extracted data is loaded into the target system, which can be a data warehouse, a database, or any other data storage system.
* **Transformation**: In this stage, the loaded data is transformed within the target system. This involves cleaning, filtering, aggregating, and formatting the data. Transformation can also involve enriching the data by combining it with other data sources or applying business rules.

**3 Tier Architecture in DE**

The 3-tier architecture is a data engineering architecture that separates the data processing into three layers:

* **Presentation Layer**: This layer is responsible for presenting the data to end-users through various interfaces such as dashboards, reports, or web applications. The presentation layer provides the user interface for interacting with the data.
* **Application Layer**: This layer is responsible for processing the data and performing business logic. It includes various applications that process and manipulate the data such as ETL tools, data modeling tools, and data mining tools.
* **Data Layer**: This layer is responsible for storing and managing the data. It includes various data storage systems such as databases, data warehouses, and data lakes.

**ETL Tools:**

There are many ETL tools available in the market, some of which are:

* **Informatica PowerCenter**: Informatica PowerCenter is a widely used ETL tool for integrating data from various sources. It provides a graphical user interface for designing ETL workflows and supports various data sources such as databases, flat files, and web services.
* **Microsoft SQL Server Integration Services (SSIS)**: SSIS is an ETL tool provided by Microsoft for integrating data from various sources. It provides a visual interface for designing ETL workflows and supports various data sources such as databases, flat files, and web services.
* **Talend:** Talend is an open-source ETL tool for integrating data from various sources. It provides a visual interface for designing ETL workflows and supports various data sources such as databases, flat files, and web services.

**TASK 05**

**Historical Load:**

Historical Load refers to the process of loading all the historical data from the source system into the target system. The purpose of historical load is to create a baseline of the data that can be used for analysis and reporting. The historical load is usually performed once, before the incremental or delta loads begin. The historical load can be a time-consuming process, as it involves loading a large amount of data.

**Full Load:**

Full Load refers to the process of loading all the data from the source system into the target system. The purpose of a full load is to ensure that all the data is available in the target system. A full load is usually performed once, at the beginning of the ETL process, to create a baseline of the data. The full load can be a time-consuming process, as it involves loading a large amount of data.

**Incremental Load:**

Incremental Load refers to the process of loading only the changed or new data from the source system into the target system. The purpose of an incremental load is to reduce the processing time and resources required for loading the data. Incremental load is usually performed after the full load or the historical load, and it involves only loading the new or changed data since the last load.

There are two types of Incremental Load:

* **Append Only**: In this type of incremental load, new data is added to the existing data in the target system. No changes are made to the existing data.
* **Upsert**: In this type of incremental load, new data is added to the existing data, and any changes to the existing data are updated. This type of load is useful when there are updates to the existing data in the source system, and those changes need to be reflected in the target system.

In conclusion, Historical Load, Full Load, and Incremental Load are the three essential processes in ETL. Historical Load is the process of loading all the historical data from the source system into the target system. Full Load is the process of loading all the data from the source system into the target system. Incremental Load is the process of loading only the new or changed data since the last load. Each of these processes has its use cases, depending on the requirements of the ETL process.