

***DATA ENGINEERING***

***TASK-04 & 05***

*Submitted to*

***Sir Muhammad Ahtisham***

*Submitted by*

***Aabra Malik***

***Extract, Transform, and Load (ETL) process***

A data pipeline called extract, transform, and load (ETL) is used to gather data from various sources. The data is subsequently transformed in accordance with business requirements before being loaded into a destination data repository. ETL's transformation process is carried out in a dedicated engine, and staging tables are frequently used to store data while it is being transformed and then loaded to its final destination.

The numerous procedures that are used to convert data often include filtering, sorting, aggregating, joining, cleaning, deduplicating, and validating data.

The three ETL processes are frequently carried out in simultaneously to save time. Instead of waiting for the entire extraction process to finish, for instance, a loading process could start working on the prepared data while the data is still being extracted by performing transformation on the data that has already been received.

Here's a breakdown of each step in the ETL process:

***Extract:*** The first step is to extract data from various sources, such as databases, flat files, APIs, or web services. The goal is to gather all the relevant data needed for analysis.

***Transform:*** Once the data is extracted, it needs to be transformed into a format that can be used for analysis. This includes cleaning the data, removing duplicates, filling in missing values, and converting data types. In addition, the data may need to be enriched with additional information, such as customer demographics or product categories.

***Load:*** The final step is to load the transformed data into a data warehouse or data mart. This involves creating tables and columns to store the data and ensuring that the data is loaded correctly and consistently. Once the data is loaded, it can be used for reporting, analysis, and visualization.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

***Extract, Load, and Transform (ELT) process***

ELT (Extract, Load, Transform) is a variation of the ETL (Extract, Transform, Load) process used in data integration and business intelligence systems. While both ETL and ELT involve extracting data from various sources and loading it into a central repository, they differ in the way data is transformed.

In the ELT process, the data is first extracted and loaded into a target system, such as a data warehouse or data lake, without any transformation. Once the data is loaded, it can be transformed using tools within the target system, such as SQL queries or data pipelines.

Here's a breakdown of each step in the ELT process:

***Extract:*** The first step is to extract data from various sources, such as databases, flat files, APIs, or web services. The goal is to gather all the relevant data needed for analysis.

***Load:*** Once the data is extracted, it is loaded into a target system, such as a data warehouse or data lake. This involves creating tables and columns to store the data and ensuring that the data is loaded correctly and consistently.

***Transform:*** The final step is to transform the data within the target system. This can include cleaning the data, removing duplicates, filling in missing values, and converting data types. In addition, the data can be enriched with additional information, such as customer demographics or product categories, using SQL queries or data pipelines.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

***3 Tier Architecture in DE***

In the context of data engineering, the 3-tier architecture is a common design pattern for building scalable and maintainable data processing systems.

The 3-tier architecture typically consists of the following layers:

***Presentation layer:*** This layer is responsible for presenting the data to the end-user. It can include various user interfaces such as web applications, mobile apps, or desktop applications.

***Application layer:*** This layer contains the business logic of the system. It is responsible for processing the data, applying transformations, and performing various computations. The application layer is typically implemented using programming languages such as Python, Java, or Scala.

***Data layer:*** This layer is responsible for storing and managing the data. It can include various data storage technologies such as relational databases, SQL databases, data lakes, and data warehouses.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

***ETL Tools***

Here are some popular ETL (Extract, Transform, Load) tools used in the industry:

Informatica

Hadoop

IBM InfoSphere DataStage

Azure Data Factory

Oracle Data Integrator (ODI)

Google Cloud Dataflow

Microsoft SQL Server Integration Services (SSIS)

Apache NiFi

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

***What is Historical Load***

Historical load refers to the process of loading historical data into a data warehouse or other data storage system. This data can include transactional data, log data, sensor data, or any other data that has been collected over time.

Historical load is an important step in building a data warehouse or data lake, as it allows data analysts and data scientists to access and analyze large volumes of data from a single location. This data can then be used to perform trend analysis, identify patterns, and make data-driven decisions.

The process of historical load typically involves extracting data from source systems, transforming it into a common format, and loading it into a centralized data repository.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

***What is Full Load***

Full Load refers to the process of extracting all data from a source system and loading it into a target system. This is typically done when setting up a new data warehouse or data mart, or when making major changes to an existing data architecture.

During a Full Load, all historical and current data is extracted from the source system and transformed to fit the data model of the target system. This can involve cleaning, enriching, and aggregating the data to make it more usable for analytics and reporting.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****

***What is Incremental Load***

Incremental Load refers to the process of extracting and loading only the data that has changed since the last load. This approach is often used to update a data warehouse or data mart with new or modified data, while minimizing the amount of data that needs to be processed.

During an Incremental Load, data changes are identified using timestamps, versioning, or other markers, and only the changed data is extracted and transformed. This allows for faster processing times and reduced resource usage, as compared to Full Load processes.

Incremental Load is typically used for ongoing updates to a data warehouse or data mart, such as daily, hourly, or real-time updates.

***\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\****