Name: Ahsan Bilal

What is ETL?

ETL stands for Extract, Transform, Load, and it refers to the process of collecting, cleaning, and integrating data from various sources into a single, unified database or data warehouse.
The ETL process involves the following steps:

- Extract: This involves identifying the relevant data sources, such as databases, files, APIs, or web pages, and retrieving the data from them. The extracted data is typically in its raw, unprocessed form.

- Transform: This step involves cleaning, validating, and structuring the extracted data so that it can be integrated into the target database or data warehouse. This may include removing duplicates, fixing errors, standardizing data formats, and applying business rules or data quality checks.

- Load: This step involves loading the transformed data into the target database or data warehouse, where it can be analyzed, queried, and reported on. This may involve using specialized ETL tools or writing custom scripts to load the data.

The ETL process is critical for organizations that need to combine data from multiple sources and make it available for analysis and decision-making. The process can be complex and time-consuming, but it is essential for ensuring data accuracy, consistency, and completeness.

What is ELT?

ELT stands for Extract, Load, Transform, which is a data integration process used in data warehousing and business intelligence. It is similar to the more traditional ETL (Extract, Transform, Load) process, but with a different order of execution.

In ELT, data is first extracted from the source systems, then loaded into a target database, and finally transformed into the desired format. The transformation step is usually performed within the target database using SQL or other programming languages.

The main advantage of ELT over ETL is that it allows for faster data integration because the data is loaded into the target database first, and then transformed. This eliminates the need for a separate data transformation process, which can be time-consuming and resource-intensive. Additionally, ELT allows for greater flexibility in terms of data

Date: 23/3/2023

processing and analysis because the transformation step is performed within the target database.

ELT is commonly used in cloud-based data integration solutions where the target database is hosted in the cloud, and the transformation is performed using cloud-based tools and services.

3 Tier Architecture in DE

The 3-tier architecture is a common architecture used in data engineering for designing scalable and maintainable data systems. It consists of three layers or tiers:

Presentation Tier: The presentation tier is the top layer of the 3-tier architecture and is responsible for providing a user interface to the end-users. This tier interacts with the end-users and provides them with the ability to view and interact with the data. Examples of presentation tier technologies include web applications, mobile applications, and desktop applications.

Application Tier: The application tier is the middle layer of the 3-tier architecture and is responsible for implementing business logic and processing user requests. This tier interacts with the presentation tier and the data tier. Examples of application tier technologies include web servers, application servers, and microservices.

Data Tier: The data tier is the bottom layer of the 3-tier architecture and is responsible for storing and managing data. This tier interacts with the application tier and provides access to data for the application tier to process. Examples of data tier technologies include databases, data lakes, and data warehouses.

In data engineering, the 3-tier architecture is often used for building scalable and reliable data systems. The separation of concerns between the layers allows for easier maintenance and scaling of the system. For example, changes to the presentation tier can be made without affecting the application tier or the data tier. Additionally, the data tier can be scaled separately from the application tier, allowing for more efficient use of resources.

Name: Ahsan Bilal

ETL Tools (any 3)

There are many ETL (Extract, Transform, Load) tools available in the market, and here are three popular ones:

1. **Informatica:** Informatica is a widely used ETL tool that provides a comprehensive set of features for data integration, data quality, and data governance. It supports a wide range of data sources and targets and provides a visual interface for building ETL workflows. Informatica can be used for both batch and real-time data integration.

2. **Apache NiFi:** Apache NiFi is an open-source ETL tool that provides a web-based interface for building data integration workflows. It is designed to handle large volumes of data and can be used for both batch and real-time data integration. NiFi provides a wide range of processors for data transformation and can be extended with custom processors.

3. **Talend:** Talend is an open-source ETL tool that provides a wide range of features for data integration, data quality, and data governance. It supports a wide range of data sources and targets and provides a visual interface for building ETL workflows. Talend can be used for both batch and real-time data integration and provides a wide range of connectors and components for data transformation.

Date: 23/3/2023