**Task 04**

**Abdul Razaque**

**Data Engineering**

* **ETL:**

Extract, transform, and load (ETL) is the process of combining data from multiple sources into a large, central repository called a data warehouse. ETL uses a set of business rules to clean and organize raw data and prepare it for storage, data analytics, and machine learning (ML).

ETL stands for Extract, Transform, and Load, which is a process used to extract data from various sources, transform it into a format that can be easily analyzed, and load it into a target system such as a data warehouse or a database.

ETL is a process that extracts the data from different source systems, then transforms the data (like applying calculations, concatenations, etc.), and finally loads the data into the Data Warehouse system.

There are many different ETL tools and technologies available, such as Informatica, Talend, DataStage, and others, that can automate and simplify the ETL process.

ETL is a process in Data Warehousing and it stands for Extract, Transform, and Load. It is a process in which an ETL tool extracts the data from various data source systems, transforms it in the staging area, and then finally, loads it into the Data Warehouse system.

The ETL process typically involves the following steps::

* Extraction: In this step, data is extracted from various sources such as databases, spreadsheets, or flat files. The extraction process can be done using various methods such as batch processing or real-time streaming.
* Transformation: After the data has been extracted, it needs to be transformed into a format that can be easily analyzed. This step involves cleaning, sorting, filtering, and combining data from multiple sources.
* Loading: In the final step, the transformed data is loaded into a target system such as a data warehouse or a database. This step involves mapping the transformed data to the target schema and loading it into the target system.

There are many reasons for adopting ETL in the organization::

* It helps companies to analyze their business data for taking critical business decisions.
* Transactional databases cannot answer complex business questions that can be answered by ETL example.
* A Data Warehouse provides a common data repository.
* ETL provides a method of moving that from various sources into a data warehouse.
* As data sources change, the Data Warehouse will automatically update.
* Well-designed and documented ETL system is almost essential to the success of a Data Warehouse project.
* Allow verification of data transformation, aggregation, and calculation rules.
* ETL process allows sample data comparison between the source and the target system.
* ETL process can perform complex transformations and requires an extra area to store the data.
* **ETL Tools:**

There are many ETL tools are available in the market. Here, is some most prominent ones:

1. **MarkLogic:**

MarkLogic is a data warehousing solution that makes data integration easier and faster using an array of enterprise features. It can query different types of data like documents, relationships, and metadata.

2. **Oracle:**

Oracle is the industry-leading database. It offers a wide range of choice of Data Warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.

3. **Amazon RedShift**:

Amazon Redshift is a data warehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data.

Overall, the ETL process is an essential process in data warehousing that helps to ensure that the data in the data warehouse is accurate, complete, and up-to-date. However, it also comes with its own set of challenges and limitations, and organizations need to carefully consider the costs and benefits before implementing them.

* **ELT:**

ELT stands for Extract, Load, Transform, which is a data integration process similar to ETL but with a different order of steps. In ELT, data is first extracted from various sources, loaded into a target system, and then transformed into a format that can be easily analyzed.

Extract/load/transform (ELT) is the process of extracting data from one or multiple sources and loading it into a target data warehouse.

ELT is comprised of a data pipeline with three different operations being performed on data::

* Extraction: In this step, data is extracted from various sources such as databases, spreadsheets, or flat files. The extraction process can be done using various methods such as batch processing or real-time streaming.
* Loading: After the data has been extracted, it is loaded into a target system such as a data warehouse or a database. This step involves mapping the source data to the target schema and loading it into the target system.
* Transformation: In the final step, the loaded data is transformed into a format that can be easily analyzed. This step involves cleaning, sorting, filtering, and combining data from multiple sources.
* **How ELT works:**

ELT is a variation of the Extract, Transform, Load (ETL), a data integration process in which transformation takes place on an intermediate server before it is loaded into the target. In contrast, ELT allows raw data to be loaded directly into the target and transformed there.

With an ELT approach, a data extraction tool is used to obtain data from a source or sources, and the extracted data is stored in a staging area or database. Any required business rules and data integrity checks can be run on the data in the staging area before it is loaded into the data warehouse.

All data transformations occur in the data warehouse after the data is loaded.

* **Uses of ELT:**

ELT is often used in the following cases:

* When the data is structured, but the source and target database are the same type (i.e., Oracle source and target);
* When the data is unstructured and massive, such as processing and correlating data from log files and sensors'
* When the data is relatively simple, but there are large amounts of it;
* When there is a plan to use machine learning tools to process the data instead of traditional SQL queries; and schema on read.

ELT has become increasingly popular in recent years because it allows for more flexible data analysis and enables organizations to take advantage of the processing power of modern databases. With ELT, organizations can store large amounts of data in their target system and then apply complex transformations to that data in a more efficient manner.

* **3 Tier Architecture in DE:**

A three-tier architecture is a client-server architecture in which the functional process logic, data access, computer data storage and user interface are developed and maintained as independent modules on separate platforms

Three-tier architecture is a common data engineering architecture used to design, develop, and deploy large-scale data-driven applications. The architecture consists of three layers or tiers, each with a specific function in the data processing pipeline.

The three tiers are::

* **Presentation Tier**: The presentation tier, also known as the client tier, is responsible for providing a user interface for interacting with the application. This layer is typically designed for the end-users and can include web-based interfaces, mobile applications, or desktop applications.
* **Application Tier**: The application tier, also known as the middle tier, is responsible for processing and managing data. This tier contains business logic and application functionality, such as data processing, validation, and manipulation. It acts as a bridge between the presentation tier and the data tier, receiving requests from the presentation tier and processing them before returning results to the presentation tier.
* **Data Tier**: The data tier, also known as the backend tier, is responsible for storing, managing, and retrieving data. This tier includes databases, data warehouses, and data lakes. The data tier can be further divided into multiple sub-layers, such as data ingestion, storage, processing, and analysis.

Benefits include:

* **Faster development**: Because each tier can be developed simultaneously by different teams, an organization can bring the application to market faster, and programmers can use the latest and best languages and tools for each tier.
* **Improved scalability**: Any tier can be scaled independently of the others as needed.
* **Improved reliability**: An outage in one tier is less likely to impact the availability or performance of the other tiers.
* **Improved security**: Because the presentation tier and data tier can't communicate directly, a well-designed application tier can function as a sort of internal firewall, preventing SQL injections and other malicious exploits.

**Advantage**::

The advantages of using a three-tier architecture in data engineering are many. It enables the separation of concerns between different layers, allowing for independent development, testing, and deployment of each tier. It also enables scalability and performance optimizations for each tier separately, which can improve the overall performance of the system. Additionally, it allows for better security and fault tolerance by isolating the presentation layer from the data layer.

* **ETL Tools (any 3):**

There are many ETL tools are available in the market. Here, is some most prominent ones:

1. **MarkLogic:**

MarkLogic is a data warehousing solution that makes data integration easier and faster using an array of enterprise features. It can query different types of data like documents, relationships, and metadata.

2. **Oracle:**

Oracle is the industry-leading database. It offers a wide range of choice of Data Warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.

3. **Amazon RedShift**:

Amazon Redshift is a data warehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data.