Task 4

Q1) What is ETL in detail?

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). You can address specific business intelligence needs through data analytics (such as predicting the outcome of business decisions, generating reports and dashboards, reducing operational inefficiency, and more).

Q2) What is ELT in detail?

ELT, which stands for "Extract, Load, Transform," is another type of data integration process, similar to its counterpart ETL, "Extract, Transform, Load". This process moves raw data from a source system to a destination resource, such as a data warehouse. While similar to ETL, ELT is a fundamentally different approach to data pre-processing which has only more recently gained adoption with the transition to cloud environments.

Q3) 3 Tier Architecture in DE?

The three-tier architecture is a common approach used in data engineering to design and organize software systems that handle data. In this architecture, the system is divided into three distinct layers, each with a specific function and set of responsibilities. The layers are as follows:

Presentation Layer:

This is the topmost layer of the architecture, which is responsible for interacting with the end-users. It includes user interfaces, such as web pages, mobile apps, or desktop applications. The primary purpose of this layer is to provide a user-friendly interface for users to interact with the system.

Application or Business Logic Layer:

This is the middle layer of the architecture and is responsible for processing and manipulating data based on the business logic or application logic of the system. It includes business rules, workflows, and data validation rules. The primary purpose of this layer is to ensure that the system functions as per the business requirements.

Data Storage Layer:

This is the bottom layer of the architecture and is responsible for storing and retrieving data from the system. It includes databases, data warehouses, data lakes, and other data storage technologies. The primary purpose of this layer is to ensure data is efficiently stored, managed, and retrieved.

Q4) ETL Tools (any 3)?

- Apache NiFi:
  Apache NiFi is a powerful open-source ETL tool that enables the automation of data flows between systems. It offers a web-based user interface for designing and managing data flows and supports various data sources and destinations. Apache NiFi is designed to handle real-time data processing and offers features such as data enrichment, transformation, and routing.
- Talend: Talend is a comprehensive ETL tool that offers a wide range of data integration and processing capabilities. It provides a drag-and-drop interface for designing data workflows, supports a variety of data sources and destinations, and offers a range of data transformation and manipulation features. Talend also supports big data technologies such as Hadoop, Spark, and NoSQL databases.
- Microsoft SQL Server Integration Services (SSIS): Microsoft SQL Server Integration Services (SSIS) is a popular ETL tool that is tightly integrated with the Microsoft SQL Server database. It offers a graphical user interface for designing data workflows and supports a range of data sources and destinations. SSIS offers a wide range of data transformation and manipulation features and supports real-time and batch processing modes. It also integrates with other Microsoft products such as Azure Data Factory and Power BI.

-