**Task 4**

**Question 1:**

**What is ETL? in detail**

**Answer:**

Extract, Transform, and Load, or ETL for short. In data integration and data warehousing, it is a method for extracting data from various sources, transforming it into a format that is more usable, and loading it into a target database or data warehouse.

Each step is broken down as follows:

**Extract:** Data is gathered from a variety of sources during the extract stage, including databases, web pages, and flat files. The steps in this process are finding the data sources, selecting the relevant data, and retrieving it.

**Transform:** The extracted data are cleaned, integrated, and enhanced during the transform stage to ensure that they are in the appropriate format for analysis. Filtering, sorting, aggregation, and joining are just a few examples of data transformation processes.

**Load**: The transformed data is loaded into the target database or data warehouse during the load stage. The data is typically structured, making it simple to query and analyze.

Data integration relies heavily on the ETL process because it enables businesses to combine data from multiple sources into a single system, making it easier to analyze and make informed business decisions. Due to the fact that the data is standardized and validated during the transformation process, it also ensures consistency, accuracy, and completeness of the data.

**Question 2:**

**What is ELT? in detail**

**Answer:**

Extract, Load, and Transform is short for ELT. Similar to ETL, but with a different order of operations, it is a process for data integration.

Each step is broken down as follows:

**Extract**: Data is gathered from a variety of sources during the extract stage, including databases, web pages, and flat files. The steps in this process are finding the data sources, selecting the relevant data, and retrieving it.

**Load**: The extracted data are loaded into the target database or data warehouse during the load stage. Typically, the data is loaded as-is without any processing or transformation.

**Transform:** The data are transformed into a format that can be used for analysis during the transform stage. Cleansing, integrating, and enriching the data in this stage ensures its consistency and accuracy. Filtering, sorting, aggregation, and joining are just a few examples of data transformation processes.

The order of operations is where ETL and ELT differ the most. Data is extracted from sources, converted into a format that can be used, and then loaded into the target database or data warehouse in ETL. Data is extracted from sources, loaded into the target database or data warehouse, and converted into a format that can be used in ELT.

When data volumes are too large to process in memory during the transformation stage or when the target database or data warehouse has a high processing power, ELT is frequently used. Data can be loaded into the target system and transformed as needed with ELT, which speeds up the data integration process and saves time and resources. To deal with the changes in the target system, ELT, on the other hand, necessitates a higher level of expertise and infrastructure.

**Question 3:**

**3 Tier Architecture in DE**

Answer:

In data engineering, a common design pattern called 3-Tier architecture is used to create scalable, dependable, and easy-to-maintain systems for processing large amounts of data. The system is divided into three layers in an architecture known as a "3-Tier": the application layer, the data layer, and the presentation layer.

**Layer of Presentation**: The user interface and interaction with end users are the responsibility of this layer. It includes browsers, mobile apps, and web servers, among other things. The show layer fills in as a connection point between the client and the application layer, permitting clients to get to and cooperate with the information.

**Layer of Application:** The functionality and business logic needed to process data are provided by the application layer. Data transformation and manipulation to meet specific business requirements are the responsibility of this layer. It has components that interact with the data layer, such as APIs, web services, or microservices.

**Layer of Data:** Data management and storage are the responsibility of the data layer. It incorporates parts, for example, data sets, information distribution centers, or information lakes. The application layer has access to a centralized repository for storing data that is provided by the data layer. For the application layer to be able to process and analyze data, the data layer is essential.

**Question 4:**

**ETL Tools (any 3)**

Answer:

Yes, there are three additional widely used ETL tools in data engineering:

**Spark by Apache:** Open-source Apache Spark is a distributed computing framework with built-in ETL operations support. Spark SQL, Spark Streaming, and MLlib are just a few of the many data processing APIs it provides. Flash is regularly utilized for handling enormous volumes of information progressively and group handling.

**Microsoft SSIS:** A proprietary ETL tool called Microsoft SQL Server Integration Services (SSIS) is used to integrate, transform, and load data into Microsoft SQL Server and other databases. It includes a wide range of connectors and transformations to handle complex data integration and transformation requirements, as well as a graphical user interface for designing ETL workflows.

**Pentaho:** Pentaho is an open-source ETL device utilized for incorporating and changing information from different sources. It supports a wide range of data sources and destinations and offers a drag-and-drop interface for creating ETL workflows. In addition, it has features for data quality and profiling to guarantee the consistency and accuracy of the data.