Task # 4:

What is ETL? in detail.

ETL is the process of transforming data from an unstructured format to a structured one. ETL stands for Extract, Transform, and Load. We extract data from different sources which are in the unstructured format then we transform that data into their respective format so that when we apply any query it will give the desired output and in the end, we load that data onto some cloud platform or some other data platforms. ETL is the essential step without performing it we'll face a lot of anomalies. It is also an important step in Business Intelligence because they need cooked data that do not contain any errors. Data Engineer develops the whole ecosystem to gain access to the information. The main task is to obtain data, then decide how it should look in the staging area, and then make it consumable by storing it somewhere else.

What is ELT? in detail.

ELT is the modern approach to handling big data problems. ELT stands for Extract, Load, and Transform. We extract the data, load it, and then transform the data simultaneously with that we do not have to make a copy instead we transform it in the target destination. With the use of the ELT approach, we can have large amounts of data in store and that provides the business intelligence team can re-query data to develop new transformations using comprehensive data stored. ELT can be beneficial because of its ability to handle large data and we do not require extra storage to hold the data first instead we dump data in a bunch and then a transformation occurs. ELT approach is widespread in businesses/organizations.

3 Tier Architecture in DE

Three-Tier architecture in DE includes the presentation, application, and data tier. The presentation tier is the communication layer where the user interacts with the application. The main goal is to collect information and display information to the user. The application tier is the logic or middle tier where the information which is collected in the presentation tier is processed. Their tier can add delete and modify data in the data tier. The data tier is the database tier where the information is stored and managed that is processed by the application tier. This database can be Relational e.g., MySQL, PostgreSQL, etc. or it will be NoSQL-based Databases such as Mongo dB, or Cassandra.

ETL Tools (any 3)

- Talend
- AWS Glue
- SSIS
- Apache Airflow

Task # 5:

What is Historical Load

Historical load is the one-time initial load of data before the creation of the pipeline. Historical data is loaded the first time we run the pipeline. If the historical load option is not selected then the event older than the pipeline creation data are not loaded. Historical data comes with three methods for ingesting data i.e., recent data first, historical load parallelization, and earliest data first.

What is Full Load

The full load will load the entire dataset and then later be completely replaced with the new or updated dataset. When the new records were generated, it will dump that data alongside the previous record. In the full load, we don’t need to care about the time to carry out a Full Data Load because we do not associate any primary column so that when new records came they will be replaced with the entire dataset i.e., Historic and Fresh. Key benefits of Full load include low maintenance and Simple design.

What is Incremental Load

Incremental load is a method of moving data from one system to another and it will compare the incoming data from the source system to the existing data. A column with unique values for every record is used for comparing the existing with the new dataset. The column we used is the primary key because when the data is to be migrated it is selected based on the time it is not easy to identify new or modified data hence data must be compared with data already in the destination.