2. How would you design an ETL process to pull data from a relational database (RDS) like

MySQL or PostgreSQL? Discuss the factors you would consider to ensure efficient and

reliable data extraction.

Steps

**Extraction**

1.Discussion with BA or the Data Scientists or the Business users who needed the Data and then identify the sources of Data whether it is replicating the table in RDMS to HDFS or any other sources or Views

2.Once the Requirement is freeze then determine the logic for the data wheather any filtering is applied based on selected criteria

3.Decides how the data often the data is retrieved (hourly,daily,weekly)

**Transformation**

1.Address the inconsistency ,missing value or format change in the extracted Data

2. Restructures data to minimize redundancy and ensure data integrity in relational databases

3. Applies techniques like anonymization or pseudonymization to protect sensitive data during processing

4. Choose between updating the target data completely or updating it incrementally in response to changes in the source

**Loading**

1.Implement mechanisms to handle potential errors during the loading process due to network failure or any error by implementing the operational table to know the status of the loading

**Factors for Efficient Data Extraction**

1.Use efficient SQL queries to minimize data retrieval time from the source RDS

2.Process data in batches to optimize network usage and database operations

3.Consider parallel processing techniques if dealing with large datasets to speed up the ETL process incase of Apache spark consider the upper bound and lowerbound based on the partition to effective retrieval of Data

4.Set up alerts for critical errors or performance issues to ensure timely intervention

3. CleverTap provides event-based user data. How would you approach extracting and

transforming this event data into a usable format for further analysis? Outline the key

components of your ETL pipeline

1.Integrate the Clever Tap webhooks with Kafka for the real time integration of Data for analysis

2. Create an endpoint that listens on the specified URL you provided in the CleverTap webhook configuration

3. Use libraries or code to parse the incoming JSON data from CleverTap webhooks. The data will contain details about the user, event type, properties, and timestamps

4. Publish the transformed event data (or raw data if no transformation) to a specific Kafka topic

5. Deploy your server-side code to a reliable hosting environment that can handle inbound webhooks and Kafka communication could be cloud or containerization

5.Secure your kafka cluster with authorization mechanism like ACL or Kerberos

6.Use Stream Processing Framework Like Apache Flink or Spark to extract and transform and clean the data

1. Address inconsistencies or missing values within the event data

2. Define a schema to organize the event data consistently. This ensures easier querying and analysis

3. Calculate session duration based on event timestamps, create user segments based on event properties

7. Choose a suitable target system for storing the transformed data -Data Warehouse,Data Lake or any Analytics platform depending on the downstream use case

8. A scheduler Airflow,Azkaban to automate the ETL process at regular intervals

9. Implement mechanisms to handle potential errors during each stage of the pipeline

10. Track the ETL process execution, including successful runs, errors, and data volumes to identify and address any issues