Title : Big Data Processing Challenge

Description : Design a coding challenge focused on Big Data processing using PySpark, Snowflake, and Databricks. The challenge should require candidates to analyze and manipulate large datasets efficiently, demonstrating their expertise in ETL processes and data warehousing. Candidates should also showcase their skills in using Apache Airflow for workflow automation.

Duration of the Interview : 2 hours

Subtasks:

Data Processing with PySpark

- Implement a PySpark job to process a large dataset. Use PySpark's DataFrame API to perform transformations, filtering, and aggregations on the data.
- Example: Calculate summary statistics (e.g., mean, median, standard deviation) for a dataset of customer transactions using PySpark.
- Tools: PySpark, DataFrame API

ETL with Snowflake

- Design an ETL pipeline using Snowflake for loading and transforming data. Utilize Snowflake's data loading capabilities, including COPY INTO and COPY INTO from stage, to efficiently handle data ingestion and manipulation.
- Example: Extract data from a CSV file, load it into Snowflake's staging area, and then transform and load it into a target table using SQL commands.
- Tools: Snowflake, SQL

Workflow Automation with Databricks and Airflow

- Create a workflow using Databricks notebook and Apache Airflow to orchestrate data processing tasks. Write PySpark code in Databricks notebook to perform data transformations and use Apache Airflow to schedule and monitor the workflow execution.
- Example: Implement a daily data processing workflow that extracts data from a source, performs aggregations using PySpark, and stores the results in Snowflake tables.
- Tools: Databricks notebook, Apache Airflow, PySpark