Title : Building a Scalable Big Data Solution on AWS Cloud

Description : In this coding challenge, you will simulate building a scalable big data solution on AWS Cloud using PySpark, Databricks, AWS Glue, AWS Lambda, and other relevant tools. The primary focus will be on designing, implementing, and optimizing data pipelines and analytics solutions.

Duration of the Interview : 2 hours

Subtasks:

Data Synchronization with AWS Datasync

- Design and implement a data synchronization solution using AWS Datasync.
- Instructions: Create a data synchronization process to transfer data between on-premises storage and AWS S3 using AWS Datasync.
- Features: Set up scheduling, monitoring, and logging for the data transfer process. Ensure data integrity and security during the transfer.
- Examples with related information: Sync CSV files from an on-premises server to an S3 bucket. Transform the data using PySpark.
- Tools: AWS Datasync, AWS S3, PySpark.

Optimizing S3 Data Storage

- Optimize S3 data storage and access patterns for efficiency.
- Instructions: Review the current S3 data storage structure and access patterns. Identify areas for optimization and implement improvements to enhance performance.
- Features: Implement proper bucket organization, lifecycle policies, and access control mechanisms. Utilize partitioning and indexing for faster data retrieval.
- Examples with related information: Partition data based on date for time-based queries. Use encryption and compression for data security and storage efficiency.
- Tools: AWS S3 management console, AWS CLI, SQL queries.

PySpark Data Pipelines Optimization

- Develop and optimize PySpark data pipelines for analytics.
- Instructions: Design and build PySpark data pipelines to process and analyze large datasets efficiently. Incorporate transformations, aggregations, and data cleansing steps.
- Features: Parallelize data processing tasks, handle schema evolution, and optimize data shuffling for performance.
- Examples with related information: Extract data from Delta tables, apply machine learning algorithms for predictive analytics, and store results in a data lake.
- Tools: PySpark, Delta tables, Databricks.