TOP 10

Intermediate Databricks and Spark Interview Questions

Abhishek Agrawal
Azure Data Engineer

Q. What are Unity Catalog in Databricks?

Ans. Unity Catalog is a centralized metadata management service in Databricks that stores and manages metadata across various components like tables, databases, views, and jobs. It provides a single point of contact for managing data access policies that apply to all workspaces.

Q. What is the difference between Delta Live table and Workflow in Databricks?

Ans. Delta Live Tables automate real-time data ingestion and processing, whereas Workflows orchestrate complex data pipelines for end-to-end processing and automation.
Using DLT, you can only trigger Notebooks, but with Workflows, you can trigger JAR, Notebook, DLT, and Python files as well.

Q. How to run a notebook from another notebook?

Ans. You can use %run or dbutils.run command.

Q. What are UDFs in Databricks?

Ans. UDFs (User-Defined Functions) allow custom logic to be reused in the user environment, reducing the human cost associated with refactoring code. They enable users to express logic in familiar languages, enhancing code readability and reusability.

Q. What is the difference between RDD and DataFrame in Spark?

RDDs (Resilient Distributed Datasets) are low-level distributed collections of objects, whereas DataFrames are higher-level distributed collections of data organized into named columns. DataFrames provide easier manipulation and optimization compared to RDDs.

Q. What are secret scopes in Databricks?

Ans. Secret Scopes in Databricks provide secure management and access control for secrets like API keys and passwords within notebooks, jobs, and clusters.

Q. How can you connect to SQL Server in Databricks?

Ans. Connect to SQL Server in Databricks using JDBC or ODBC drivers, configuring connection settings, and utilizing spark.read.jdbc() or spark.write.jdbc() methods.

Q. What are the different components of a cluster?

Ans. Azure Databricks cluster comprises Driver and Worker Nodes for computation, Apache Spark for distributed processing, Databricks Runtime for optimization, Cluster Manager for lifecycle management

Q. How to pass a parameter to a Databricks notebook from Data Factory?

Ans. You can use dbutils.widgets command to pass parameters in the notebook.

Q. Can we run multiple notebooks simultaneously in Databricks?

Ans. Yes, you can use the multiprocessing.pool library to run multiple notebooks concurrently in Databricks.

Follow for more content like this

Abhishek Agrawal

Azure Data Engineer