DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER$^{\text{Q\&As}}$

Databricks Certified Professional Data Engineer Exam

Pass Databricks DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

A Spark job is taking longer than expected. Using the Spark UI, a data engineer notes that the Min, Median, and Max Durations for tasks in a particular stage show the minimum and median time to complete a task as roughly the same, but the max duration for a task to be roughly 100 times as long as the minimum.

Which situation is causing increased duration of the overall job?

A. Task queueing resulting from improper thread pool assignment.

B. Spill resulting from attached volume storage being too small.

C. Network latency due to some cluster nodes being in different regions from the source data

D. Skew caused by more data being assigned to a subset of spark-partitions.

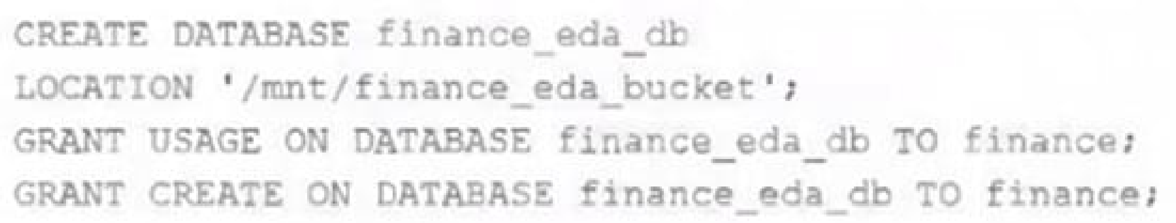
E. Credential validation errors while pulling data from an external system.

Correct Answer: D

Explanation: This is the correct answer because skew is a common situation that causes increased duration of the overall job. Skew occurs when some partitions have more data than others, resulting in uneven distribution of work among tasks and executors. Skew can be caused by various factors, such as skewed data distribution, improper partitioning strategy, or join operations with skewed keys. Skew can lead to performance issues such as long-running tasks, wasted resources, or even task failures due to memory or disk spills. Verified References: [Databricks Certified Data Engineer Professional], under "Performance Tuning" section; Databricks Documentation, under "Skew" section.

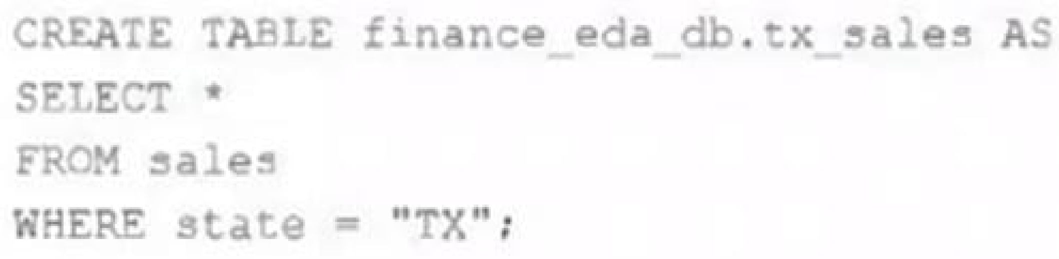
QUESTION 2

An external object storage container has been mounted to the location/mnt/finance_eda_bucket. The following logic was executed to create a database for the finance team:



```
CREATE DATABASE finance_eda_db
LOCATION '/mnt/finance_eda_bucket';
GRANT USAGE ON DATABASE finance_eda_db TO finance;
GRANT CREATE ON DATABASE finance_eda_db TO finance;
```

After the database was successfully created and permissions configured, a member of the finance team runs the following code:



```
CREATE TABLE finance_eda_db.tx_sales AS
SELECT *
FROM sales
WHERE state = "TX";
```

If all users on the finance team are members of thefinancegroup, which statement describes how thetx_salestable will be created?

A. A logical table will persist the query plan to the Hive Metastore in the Databricks control plane.

B. An external table will be created in the storage container mounted to /mnt/finance eda bucket.

C. A logical table will persist the physical plan to the Hive Metastore in the Databricks control plane.

D. An managed table will be created in the storage container mounted to /mnt/finance eda bucket.

E. A managed table will be created in the DBFS root storage container.

Correct Answer: B

Explanation: The code uses the CREATE TABLE USING DELTA command to create a Delta Lake table from an existing Parquet file stored in an external object storage container mounted to /mnt/finance_eda_bucket. The code also uses the LOCATION keyword to specify the path to the Parquet file as /mnt/finance_eda_bucket/tx_sales.parquet. By using the LOCATION keyword, the code creates an external table, which is a table that is stored outside of the default warehouse directory and whose metadata is not managed by Databricks. An external table can be created from an existing directory in a cloud storage system, such as DBFS or S3, that contains data files in a supported format, such as Parquet or CSV. Verified References: [Databricks Certified Data Engineer Professional], under "Delta Lake" section; Databricks Documentation, under "Create an external table" section.

QUESTION 3

The data engineering team is migrating an enterprise system with thousands of tables and views into the Lakehouse. They plan to implement the target architecture using a series of bronze, silver, and gold tables. Bronze tables will almost exclusively be used by production data engineering workloads, while silver tables will be used to support both data engineering and machine learning workloads. Gold tables will largely serve business intelligence and reporting purposes. While personal identifying information (PII) exists in all tiers of data, pseudonymization and anonymization rules are in place for all data at the silver and gold levels.

The organization is interested in reducing security concerns while maximizing the ability to collaborate across diverse teams.

Which statement exemplifies best practices for implementing this system?

A. Isolating tables in separate databases based on data quality tiers allows for easy permissions management through database ACLs and allows physical separation of default storage locations for managed tables.

B. Because databases on Databricks are merely a logical construct, choices around database organization do not impact security or discoverability in the Lakehouse.

C. Storinq all production tables in a single database provides a unified view of all data assets available throughout the Lakehouse, simplifying discoverability by granting all users view privileges on this database.

D. Working in the default Databricks database provides the greatest security when working with managed tables, as these will be created in the DBFS root.

E. Because all tables must live in the same storage containers used for the database they\\'re created in, organizations should be prepared to create between dozens and thousands of databases depending on their data isolation requirements.

Correct Answer: A

Explanation: This is the correct answer because it exemplifies best practices for implementing this system. By isolating tables in separate databases based on data quality tiers, such as bronze, silver, and gold, the data engineering team can achieve several benefits. First, they can easily manage permissions for different users and groups through database ACLs, which allow granting or revoking access to databases, tables, or views. Second, they can physically separate the default storage locations for managed tables in each database, which can improve performance and reduce costs. Third, they can provide a clear and consistent naming convention for the tables in each database, which can improve discoverability and usability. Verified References: [Databricks Certified Data Engineer Professional], under "Lakehouse" section; Databricks Documentation, under "Database object privileges" section.

QUESTION 4

Which distribution does Databricks support for installing custom Python code packages?

A. sbt

B. CRAN

C. CRAM

D. nom

E. Wheels

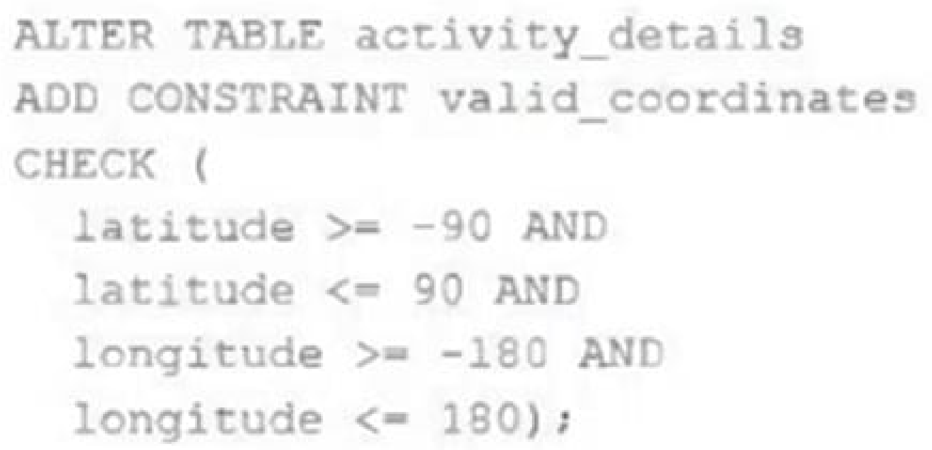
F. jars

Correct Answer: D

QUESTION 5

The downstream consumers of a Delta Lake table have been complaining about data quality issues impacting performance in their applications. Specifically, they have complained that invalidlatitudeandlongitudevalues in theactivity_detailstable have been breaking their ability to use other geolocation processes.

A junior engineer has written the following code to addCHECKconstraints to the Delta Lake table: A senior engineer has confirmed the above logic is correct and the valid ranges for latitude and longitude are provided, but the code fails when executed. Which statement explains the cause of this failure?



```
ALTER TABLE activity_details
ADD CONSTRAINT valid_coordinates
CHECK (
   latitude >= -90 AND
   latitude <= 90 AND
   longitude >= -180 AND
   longitude <= 180);
```

A. Because another team uses this table to support a frequently running application, two- phase locking is preventing the operation from committing.

B. The activity details table already exists; CHECK constraints can only be added during initial table creation.

C. The activity details table already contains records that violate the constraints; all existing data must pass CHECK constraints in order to add them to an existing table.

D. The activity details table already contains records; CHECK constraints can only be added prior to inserting values into a table.

E. The current table schema does not contain the field valid coordinates; schema evolution will need to be enabled before altering the table to add a constraint.

Correct Answer: C

Explanation: The failure is that the code to add CHECK constraints to the Delta Lake table fails when executed. The code uses ALTER TABLE ADD CONSTRAINT commands to add two CHECK constraints to a table named activity_details. The first constraint checks if the latitude value is between -90 and 90, and the second constraint checks if the longitude value is between -180 and 180. The cause of this failure is that the activity_details table already contains records that violate these constraints, meaning that they have invalid latitude or longitude values outside of these ranges. When adding CHECK constraints to an existing table, Delta Lake verifies that all existing data satisfies the constraints before adding them to the table. If any record violates the constraints, Delta Lake throws an exception and aborts the operation. Verified References: [Databricks Certified Data Engineer Professional], under "Delta Lake" section; Databricks Documentation, under "Add a CHECK constraint to an existing table" section.