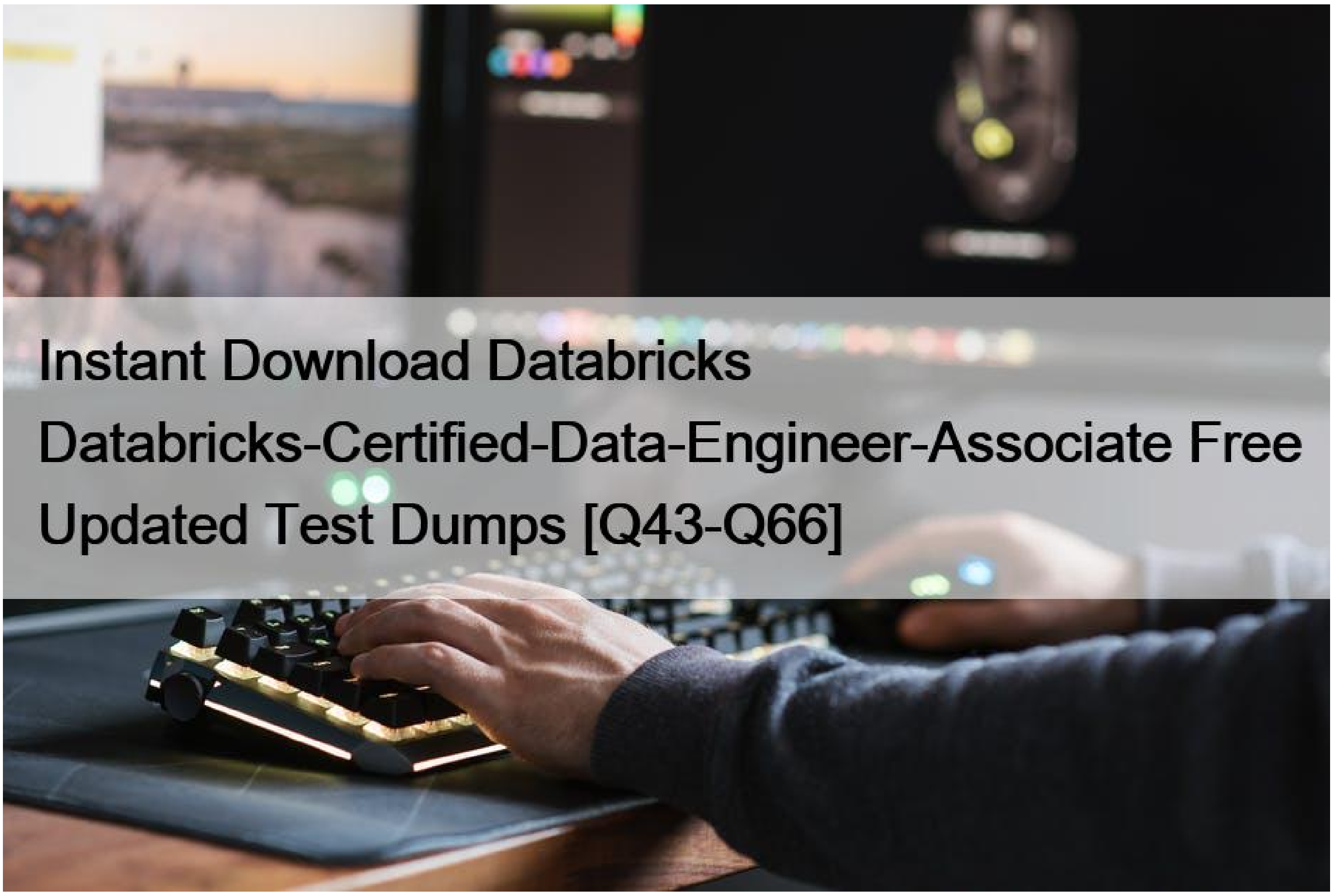
Instant Download Databricks Databricks-Certified-Data-Engineer-Associate Free Updated Test Dumps [Q43-Q66



Instant Download Databricks: Databricks-Certified-Data-Engineer-Associate Free Updated Test Dumps
Valid Databricks-Certified-Data-Engineer-Associate FREE EXAM DUMPS QUESTIONS & ANSWERS

NO.43 Which of the following describes the relationship between Bronze tables and raw data?
* Bronze tables contain less data than raw data files.
* Bronze tables contain more truthful data than raw data.
* Bronze tables contain aggregates while raw data is unaggregated.
* Bronze tables contain a less refined view of data than raw data.
* Bronze tables contain raw data with a schema applied.
Bronze tables are the first layer of a medallion architecture, which is a data design pattern used to organize data in a lakehouse. Bronze tables contain raw data ingested from various sources, such as RDBMS data, JSON files, IoT data, etc. The table structures in this layer correspond to the source system table structures

&#8220;as-is&#8221;, along with any additional metadata columns that capture the load date/time, process ID, etc. The only transformation applied to the raw data in this layer is to apply a schema, which defines the column names and data types of the table. The schema can be inferred from the data source or specified explicitly. Applying a schema to the raw data enables the use of SQL and other structured query languages to access and analyze the data. Therefore, option E is the correct answer. References: What is a

Medallion Architecture?, Raw Data Ingestion into Delta Lake Bronze tables using Azure Synapse Mapping Data Flow, Apache Spark + Delta Lake concepts, Delta Lake Architecture & Azure Databricks Workspace.

NO.44 A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.

Which of the following commands can be used to grant full permissions on the database to the new data engineering team?
* GRANT ALL PRIVILEGES ON TABLE sales TO team;
* GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
* GRANT SELECT ON TABLE sales TO team;
* GRANT USAGE ON TABLE sales TO team;
* GRANT ALL PRIVILEGES ON TABLE team TO sales;

NO.45 A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.

Which of the following Git operations does the data engineer need to run to accomplish this task?
* Merge
* Push
* Pull
* Commit
* Clone

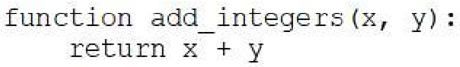
To sync a Databricks Repo with the changes from a central Git repository, the data engineer needs to run the Git pull operation. This operation fetches the latest updates from the remote repository and merges them with the local repository. The data engineer can use the Pull button in the Databricks Repos UI, or use the git pull command in a terminal session. The other options are not relevant for this task, as they either push changes to the remote repository (Push), combine two branches (Merge), save changes to the local repository (Commit), or create a new local repository from a remote one (Clone). Reference:

Run Git operations on Databricks Repos

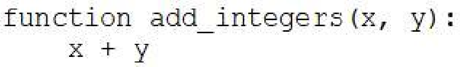
Git pull

NO.46 A data engineer that is new to using Python needs to create a Python function to add two integers together and return the sum?

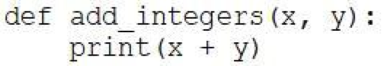
Which of the following code blocks can the data engineer use to complete this task?

*


```
function add_integers(x, y):
    return x + y
```

*


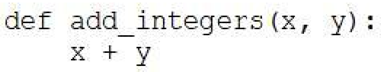
```
function add_integers(x, y):
    x + y
```

*


```
def add_integers(x, y):
    print(x + y)
```

*

c



```
def add_integers(x, y):
    x + y
```

Explanation

https://www.w3schools.com/python/python_functions.asp

NO.47 A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.

Which of the following commands can be used to grant full permissions on the database to the new data engineering team?
* GRANT ALL PRIVILEGES ON TABLE sales TO team;
* GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
* GRANT SELECT ON TABLE sales TO team;
* GRANT USAGE ON TABLE sales TO team;
* GRANT ALL PRIVILEGES ON TABLE team TO sales;

To grant full permissions on a table to a user or a group, you can use the GRANT ALL PRIVILEGES ON TABLE statement. This statement will grant all the possible privileges on the table, such as SELECT, CREATE, MODIFY, DROP, ALTER, etc. Option A is the only code block that follows this syntax correctly. Option B is incorrect, as it does not grant all the possible privileges on the table, but only a subset of them. Option C is incorrect, as it only grants the SELECT privilege on the table, which is not enough to fully manage the project. Option D is incorrect, as it grants the USAGE privilege on the table, which is not a valid privilege for tables. Option E is incorrect, as it grants all the privileges on the table team to the user or group sales, which is the opposite of what the question asks. References: Grant privileges on a table using SQL | Databricks on AWS, Grant privileges on a table using SQL &#8211; Azure Databricks, SQL Privileges &#8211; Databricks

NO.48 A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.

Which of the following approaches can the data engineer take to identify the table that is dropping the records?
* They can set up separate expectations for each table when developing their DLT pipeline.
* They cannot determine which table is dropping the records.
* They can set up DLT to notify them via email when records are dropped.
* They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.
* They can navigate to the DLT pipeline page, click on the &#8220;Error&#8221; button, and review the present errors.

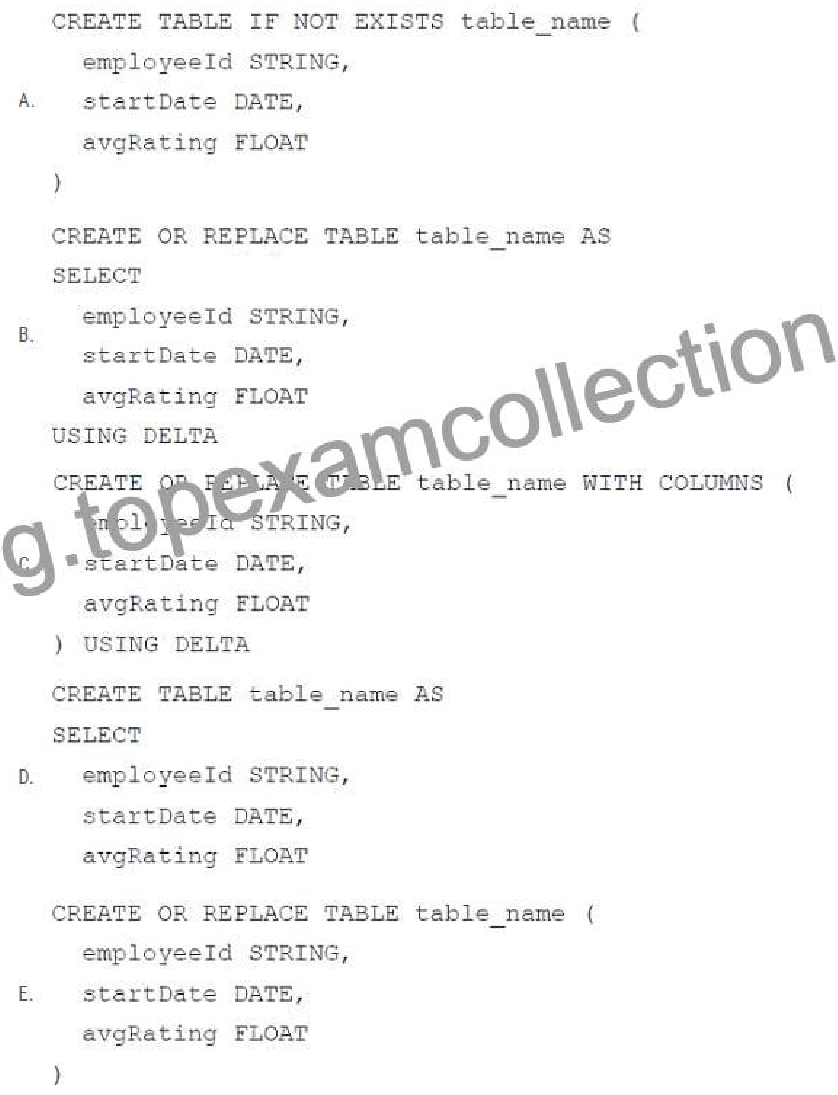
One of the features of DLT is that it provides data quality metrics for each dataset in the pipeline, such as the number of records that pass or fail expectations, the number of records that are dropped, and the number of records that are written to the target. These metrics can be accessed from the DLT pipeline page, where the data engineer can click on each table and view the data quality statistics for the latest update or any previous update. This way, they can identify which table is dropping the records and why.
References:

* Monitor Delta Live Tables pipelines

* Manage data quality with Delta Live Tables

NO.49 A data architect has determined that a table of the following format is necessary:

Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above format regardless of whether a table already exists with this name?



A.
```
CREATE TABLE IF NOT EXISTS table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

B.
```
CREATE OR REPLACE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
USING DELTA
```

C.
```
CREATE OR REPLACE TABLE table_name WITH COLUMNS (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
) USING DELTA
```

D.
```
CREATE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
```

E.
```
CREATE OR REPLACE TABLE table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

* Option A
* Option B
* Option C
* Option D
* Option E

NO.50 Which of the following describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

* CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
* CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.
* CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.
* CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
* CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

Explanation

The CREATE STREAMING LIVE TABLE syntax is used when you want to create Delta Live Tables (DLT) tables that are designed for processing data incrementally. This is typically used when your data pipeline involves streaming or incremental data updates, and you want the table to stay up to date as new data arrives.

It allows you to define tables that can handle data changes incrementally without the need for full table refreshes.

NO.51 A data engineer is maintaining a data pipeline. Upon data ingestion, the data engineer notices that the source data is starting to have a lower level of quality. The data engineer would like to automate the process of monitoring the quality level.

Which of the following tools can the data engineer use to solve this problem?
* Unity Catalog
* Data Explorer
* Delta Lake
* Delta Live Tables
* Auto Loader

Explanation

https://docs.databricks.com/delta-live-tables/expectations.html

Delta Live Tables is a tool provided by Databricks that can help data engineers automate the monitoring of data quality. It is designed for managing data pipelines, monitoring data quality, and automating workflows.

With Delta Live Tables, you can set up data quality checks and alerts to detect issues and anomalies in your data as it is ingested and processed in real-time. It provides a way to ensure that the data quality meets your desired standards and can trigger actions or notifications when issues are detected. While the other tools mentioned may have their own purposes in a data engineeringenvironment, Delta Live Tables is specifically designed for data quality monitoring and automation within the Databricks ecosystem.

NO.52 A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start.

Which of the following actions can the data engineer perform to improve the start up time for the clusters used for the Job?
* They can use endpoints available in Databricks SQL
* They can use jobs clusters instead of all-purpose clusters
* They can configure the clusters to be single-node
* They can use clusters that are from a cluster pool
* They can configure the clusters to autoscale for larger data sizes

NO.53 A data engineer has a Python notebook in Databricks, but they need to use SQL to accomplish a specific task within a cell. They still want all of the other cells to use Python without making any changes to those cells.

Which of the following describes how the data engineer can use SQL within a cell of their Python notebook?
* It is not possible to use SQL in a Python notebook
* They can attach the cell to a SQL endpoint rather than a Databricks cluster
* They can simply write SQL syntax in the cell
* They can add %sql to the first line of the cell
* They can change the default language of the notebook to SQL

NO.54 A data engineer has realized that they made a mistake when making a daily update to a table. They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted.

Which of the following explains why the data files are no longer present?
* The VACUUM command was run on the table
* The TIME TRAVEL command was run on the table
* The DELETE HISTORY command was run on the table
* The OPTIMIZE command was nun on the table
* The HISTORY command was run on the table

The VACUUM command is used to remove files that are no longer referenced by a Delta table and are older than the retention threshold1. The default retention period is 7 days2, but it can be changed by setting the delta.logRetentionDuration and delta.deletedFileRetentionDuration configurations3. If the VACUUM command was run on the table with a retention period shorter than 3 days, then the data files that were needed to restore the table to a 3-day-old version would have been deleted. The other commands do not delete data files from the table. The TIME TRAVEL command is used to query a historical version of the table4. The DELETE HISTORY command is not a valid command in Delta Lake. The OPTIMIZE command is used to improve the performance of the table by compacting small files into larger ones5. The HISTORY command is used to retrieve information about the operations performed on the table. References: 1: VACUUM | Databricks on AWS 2: Work with Delta Lake table history | Databricks on AWS 3: [Delta Lake configuration | Databricks on AWS] 4: Work with Delta Lake table history &#8211; Azure Databricks 5: [OPTIMIZE | Databricks on AWS] : [HISTORY | Databricks on AWS]

NO.55 Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?
* The ability to manipulate the same data using a variety of languages
* The ability to collaborate in real time on a single notebook
* The ability to set up alerts for query failures
* The ability to support batch and streaming workloads
* The ability to distribute complex data operations

NO.56 A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.

Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?
* Databricks Repos automatically saves development progress
* Databricks Repos supports the use of multiple branches
* Databricks Repos allows users to revert to previous versions of a notebook
* Databricks Repos provides the ability to comment on specific changes
* Databricks Repos is wholly housed within the Databricks Lakehouse Platform

Explanation

An advantage of using Databricks Repos over the built-in Databricks Notebooks versioning is the ability to work with multiple branches. Branching is a fundamental feature ofversion control systems like Git, which Databricks Repos is built upon. It allows you to create separate branches for different tasks, features, or experiments within your project. This separation helps in parallel development and experimentation without affecting the main branch or the work of other team members. Branching provides a more organized and collaborative development environment, making it easier to merge changes and manage different development efforts. While Databricks Notebooks versioning also allows you to track versions of notebooks, it may not provide the same level of flexibility and collaboration as branching in Databricks Repos.

NO.57 In order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing, which of the following two approaches is used by Spark to record the offset range of the data being processed in each trigger?
* Checkpointing and Write-ahead Logs
* Structured Streaming cannot record the offset range of the data being processed in each trigger.
* Replayable Sources and Idempotent Sinks

* Write-ahead Logs and Idempotent Sinks
* Checkpointing and Idempotent Sinks

NO.58 A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.

Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with $0 in sales is greater than zero?
* They can set up an Alert with a custom template.
* They can set up an Alert with a new email alert destination.
* They can set up an Alert with one-time notifications.
* They can set up an Alert with a new webhook alert destination.
* They can set up an Alert without notifications.

NO.59 A data engineer and data analyst are working together on a data pipeline. The data engineer is working on the raw, bronze, and silver layers of the pipeline using Python, and the data analyst is working on the gold layer of the pipeline using SQL. The raw source of the pipeline is a streaming input. They now want to migrate their pipeline to use Delta Live Tables.

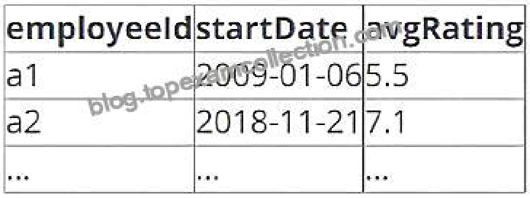
Which of the following changes will need to be made to the pipeline when migrating to Delta Live Tables?
* None of these changes will need to be made
* The pipeline will need to stop using the medallion-based multi-hop architecture
* The pipeline will need to be written entirely in SQL
* The pipeline will need to use a batch source in place of a streaming source
* The pipeline will need to be written entirely in Python
Delta Live Tables is a declarative framework for building reliable, maintainable, and testable data processing pipelines. You define the transformations to perform on your data and Delta Live Tables manages task orchestration, cluster management, monitoring, data quality, and error handling. Delta Live Tables supports both SQL and Python as the languages for defining your datasets and expectations. Delta Live Tables also supports both streaming and batch sources, and can handle both append-only and upsert data patterns. Delta Live Tables follows the medallion lakehouse architecture, which consists of three layers of data: bronze, silver, and gold. Therefore, migrating to Delta Live Tables does not require any of the changes listed in the options B, C, D, or E. The data engineer and data analyst can use the same languages, sources, and architecture as before, and simply declare their datasets and expectations using Delta Live Tables syntax. Reference:

What is Delta Live Tables?

Transform data with Delta Live Tables

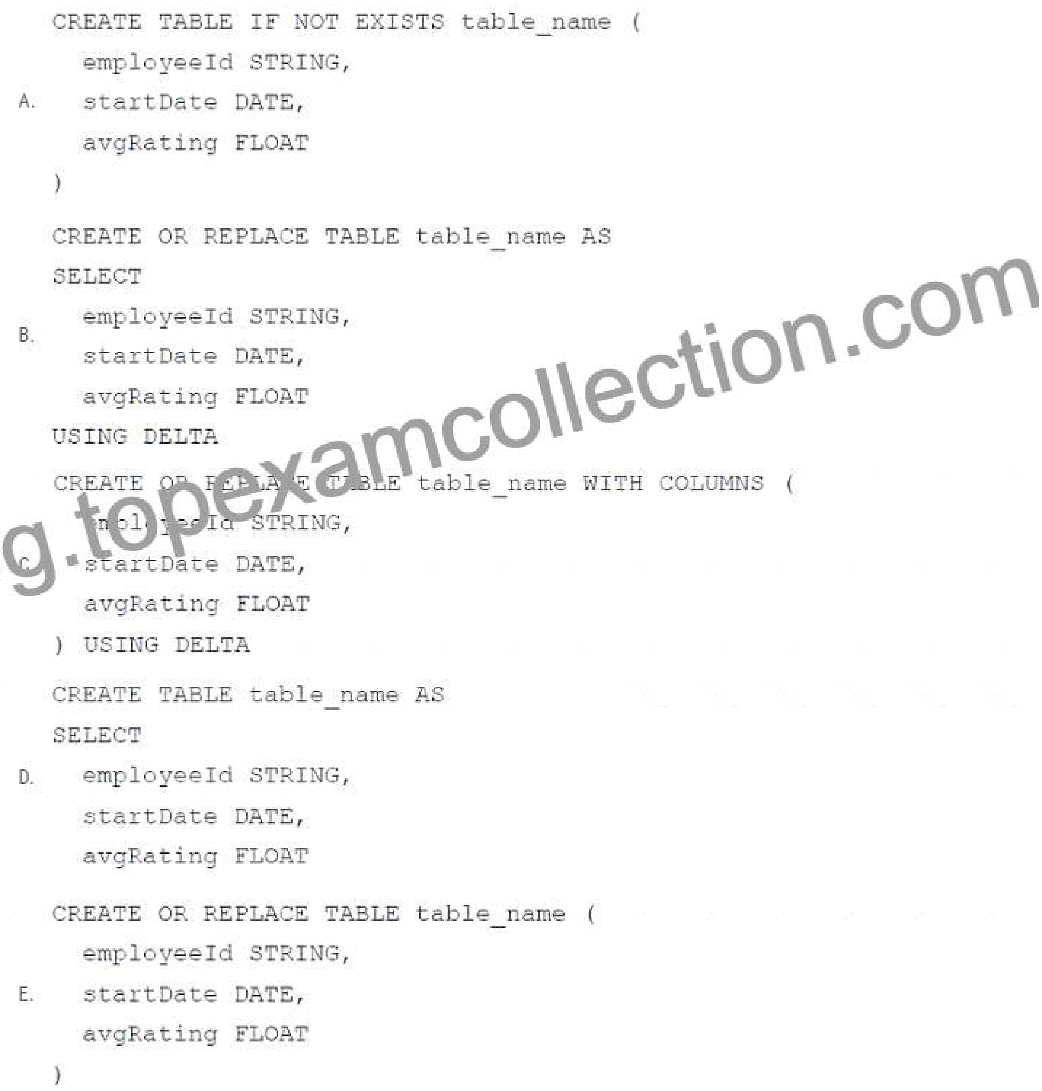
What is the medallion lakehouse architecture?

NO.60 A data architect has determined that a table of the following format is necessary:



employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
...	...	...

Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above format regardless of whether a table already exists with this name?



```
   CREATE TABLE IF NOT EXISTS table_name (
     employeeId STRING,
A.   startDate DATE,
     avgRating FLOAT
   )

   CREATE OR REPLACE TABLE table_name AS
   SELECT
     employeeId STRING,
B.   startDate DATE,
     avgRating FLOAT
   USING DELTA

   CREATE OR REPLACE TABLE table_name WITH COLUMNS (
     employeeId STRING,
C.   startDate DATE,
     avgRating FLOAT
   ) USING DELTA

   CREATE TABLE table_name AS
   SELECT
D.   employeeId STRING,
     startDate DATE,
     avgRating FLOAT

   CREATE OR REPLACE TABLE table_name (
     employeeId STRING,
E.   startDate DATE,
     avgRating FLOAT
   )
```

* Option A
* Option B
* Option C
* Option D
* Option E
References: Create a table using SQL | Databricks on AWS, Create a table using SQL &#8211; Azure Databricks, Delta Lake Quickstart &#8211; Azure Databricks

NO.61 A data engineer has a Python variable table_name that they would like to use in a SQL query. They want to construct a Python code block that will run the query using table_name.

They have the following incomplete code block:

____(f&#8221;SELECT customer_id, spend FROM {table_name}&#8221;)

Which of the following can be used to fill in the blank to successfully complete the task?
* spark.delta.sql
* spark.delta.table
* spark.table
* dbutils.sql
* spark.sql

NO.62 A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.

Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?
* They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to

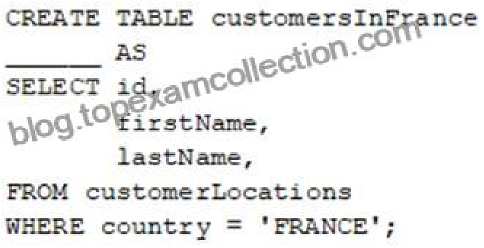
&#8220;Reliability Optimized.&#8221;
* They can turn on the Auto Stop feature for the SQL endpoint.
* They can increase the cluster size of the SQL endpoint.
* They can turn on the Serverless feature for the SQL endpoint.
* They can increase the maximum bound of the SQL endpoint&#8217;s scaling range
Explanation

Databricks SQL endpoints can run in two modes: Serverless and Dedicated. Serverless mode allows you to run queries without managing clusters, while Dedicated mode allows you to run queries on a specific cluster.

Serverless mode is faster and more cost-effective for ad-hoc queries, especially when the SQL endpoint is not running. Dedicated mode is more suitable for predictable and consistent performance, especially for long-running queries. By turning on the Serverless feature for the SQL endpoint, the data engineering team can reduce the time it takes to start the SQL endpoint and return results. The other options are not relevant or effective for this scenario. References: Databricks SQL endpoints, New Performance Improvements in Databricks SQL, Slowness when fetching results in Databricks SQL

NO.63 A data engineer wants to create a new table containing the names of customers that live in France.

They have written the following command:



```
CREATE TABLE customersInFrance
______ AS
SELECT id,
       firstName,
       lastName,
FROM customerLocations
WHERE country = 'FRANCE';
```

A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).

Which of the following lines of code fills in the above blank to successfully complete the task?
* There is no way to indicate whether a table contains PII.

* &#8220;COMMENT PII&#8221;
* TBLPROPERTIES PII
* COMMENT &#8220;Contains PII&#8221;
* PII

Explanation

Ref:https://www.databricks.com/discover/pages/data-quality-management

CREATE TABLE my_table (id INT COMMENT &#8216;Unique Identification Number&#8217;, name STRING COMMENT &#8216;PII&#8217;, age INT COMMENT &#8216;PII&#8217;) TBLPROPERTIES (&#8216;contains_pii&#8217;=True) COMMENT &#8216;Contains PII&#8217;;

NO.64 A data engineer has created a new database using the following command:

CREATE DATABASE IF NOT EXISTS customer360;

In which of the following locations will the customer360 database be located?

* dbfs:/user/hive/database/customer360
* dbfs:/user/hive/warehouse
* dbfs:/user/hive/customer360
* More information is needed to determine the correct response

dbfs:/user/hive/warehouse Thereby showing &#8220;dbfs:/user/hive/warehouse/customer360.db The location of the customer360 database depends on the value of the spark.sql.warehouse.dir configuration property, which specifies the default location for managed databases and tables. If the property is not set, the default value is dbfs:/user/hive/warehouse. Therefore, the customer360 database will be located in dbfs:/user/hive/warehouse/customer360.db. However, if the property is set to a different value, such as dbfs:/user/hive/database, then the customer360 database will be located in dbfs:/user/hive/database/customer360.db. Thus, more information is needed to determine the correct response.

Option A is not correct, as dbfs:/user/hive/database/customer360 is not the default location for managed databases and tables, unless the spark.sql.warehouse.dir property is explicitly set to dbfs:/user/hive/database.

Option B is not correct, as dbfs:/user/hive/warehouse is the default location for the root directory of managed databases and tables, not for a specific database. The database name should be appended with .db to the directory path, such as dbfs:/user/hive/warehouse/customer360.db.

Option C is not correct, as dbfs:/user/hive/customer360 is not a valid location for a managed database, as it does not follow the directory structure specified by the spark.sql.warehouse.dir property.

References:

* Databases and Tables

* [Databricks Data Engineer Professional Exam Guide]

NO.65 Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

* Cloud-specific integrations
* Simplified governance
* Ability to scale storage
* Ability to scale workloads
* Avoiding vendor lock-in

One of the benefits of the Databricks Lakehouse Platform embracing open source technologies is that it avoids vendor lock-in. This means that customers can use the same open source tools and frameworks across different cloud providers, and migrate their data and workloads without being tied to a specific vendor. The Databricks Lakehouse Platform is built on open source projects such as Apache Spark, Delta Lake, MLflow, and Redash, which are widely used and trusted by millions of developers. By supporting these open source technologies, the Databricks Lakehouse Platform enables customers to leverage the innovation and community of the open source ecosystem, and avoid the risk of being locked into proprietary or closed solutions. The other options are either not related to open source technologies (A, B, C, D), or not benefits of the Databricks Lakehouse Platform (A, B). References: Databricks Documentation &#8211; Built on open source, Databricks Documentation &#8211; What is the Lakehouse Platform?, Databricks Blog &#8211; Introducing the Databricks Lakehouse Platform.

NO.66 Which of the following tools is used by Auto Loader process data incrementally?
* Checkpointing
* Spark Structured Streaming
* Data Explorer
* Unity Catalog
* Databricks SQL

Auto Loader provides a Structured Streaming source called cloudFiles that can process new data files as they arrive in cloud storage without any additional setup. Auto Loader uses a scalable key-value store to track ingestion progress and ensure exactly-once semantics. Auto Loader can ingest various file formats and load them into Delta Lake tables. Auto Loader is recommended for incremental data ingestion with Delta Live Tables, which extends the functionality of Structured Streaming and allows you to write declarative Python or SQL code to deploy a production-quality data pipeline. Reference: What is Auto Loader?, What is Auto Loader? | Databricks on AWS, Solved: How does Auto Loader ingest data? &#8211; Databricks &#8211; 5629

Free Databricks-Certified-Data-Engineer-Associate Exam Braindumps Databricks Pratice Exam:
https://www.topexamcollection.com/Databricks-Certified-Data-Engineer-Associate-vce-collection.html]