

Databricks

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam



NEW QUESTION 1

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?

- A. dbfs:/user/hive/database/customer360
- B. dbfs:/user/hive/warehouse
- C. dbfs:/user/hive/customer360
- D. More information is needed to determine the correct response

Answer: B

Explanation:

dbfs:/user/hive/warehouse - which is the default location

NEW QUESTION 2

Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?

- A. Manually programming in an alert system in each cell of the Notebook
- B. Setting up an Alert in the Job page
- C. Setting up an Alert in the Notebook
- D. There is no way to notify the Job owner in the case of Job failure
- E. MLflow Model Registry Webhooks

Answer: B

Explanation:

<https://docs.databricks.com/en/workflows/jobs/job-notifications.html>

NEW QUESTION 3

A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table. The code block used by the data engineer is below:

```
(spark.table("sales")
  .withColumn("avg_price", col("sales") / col("units"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("complete")
  .trigger(
    .table("new_sales")
  )
)
```

If the data engineer only wants the query to execute a micro-batch to process data every 5 seconds, which of the following lines of code should the data engineer use to fill in the blank?

- A. trigger("5 seconds")
- B. trigger()
- C. trigger(once="5 seconds")
- D. trigger(processingTime="5 seconds")
- E. trigger(continuous="5 seconds")

Answer: D

Explanation:

```
# ProcessingTime trigger with two-seconds micro-batch interval df.writeStream \
format("console") \ trigger(processingTime='2 seconds') \ start()
https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#triggers
```

NEW QUESTION 4

A data analyst has developed a query that runs against Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?

- A. SELECT * FROM sales
- B. spark.delta.table
- C. spark.sql
- D. There is no way to share data between PySpark and SQL.
- E. spark.table

Answer: C

Explanation:

```
from pyspark.sql import SparkSession spark = SparkSession.builder.getOrCreate()
df = spark.sql("SELECT * FROM sales") print(df.count())
```

NEW QUESTION 5

A data organization leader is upset about the data analysis team's reports being different from the data engineering team's reports. The leader believes the siloed nature of their organization's data engineering and data analysis architectures is to blame.

Which of the following describes how a data lakehouse could alleviate this issue?

- A. Both teams would autoscale their work as data size evolves
- B. Both teams would use the same source of truth for their work
- C. Both teams would reorganize to report to the same department
- D. Both teams would be able to collaborate on projects in real-time
- E. Both teams would respond more quickly to ad-hoc requests

Answer: B

Explanation:

A data lakehouse is designed to unify the data engineering and data analysis architectures by integrating features of both data lakes and data warehouses. One of the key benefits of a data lakehouse is that it provides a common, centralized data repository (the "lake") that serves as a single source of truth for data storage and analysis. This allows both data engineering and data analysis teams to work with the same consistent data sets, reducing discrepancies and ensuring that the reports generated by both teams are based on the same underlying data.

NEW QUESTION 6

Which of the following must be specified when creating a new Delta Live Tables pipeline?

- A. A key-value pair configuration
- B. The preferred DBU/hour cost
- C. A path to cloud storage location for the written data
- D. A location of a target database for the written data
- E. At least one notebook library to be executed

Answer: E

Explanation:

<https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html>

NEW QUESTION 7

A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

sales

customer_id	spend	units
a1	28.94	7
a3	874.12	23
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

The data engineer runs the following query to join these tables together:

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

Which of the following will be returned by the above query?

A.	customer_id	spend	store_id
	a1	28.94	s1
	a4	8.99	s2

B.	customer_id	spend	units	store_id
	a1	28.94	7	s1
	a4	8.99	1	s2

C.	customer_id	spend	store_id
	a1	28.94	s1
	a3	874.12	NULL
	a4	8.99	s2

D.	customer_id	spend	store_id
	a1	28.94	s1
	a2	NULL	s1
	a3	874.12	NULL
	a4	8.99	s2

E.	customer_id	spend	store_id
	a1	28.94	s1
	a2	NULL	s1
	a4	8.99	s2

- A. Option A
- B. Option B
- C. Option C
- D. Option D
- E. Option E

Answer: C

NEW QUESTION 8

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?

- A. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."
- B. They can turn on the Auto Stop feature for the SQL endpoint.
- C. They can increase the cluster size of the SQL endpoint.
- D. They can turn on the Serverless feature for the SQL endpoint.
- E. They can increase the maximum bound of the SQL endpoint's scaling range

Answer: C

Explanation:

<https://www.databricks.com/blog/2022/03/10/top-5-databricks-performance-tips.html>

NEW QUESTION 9

Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

- A. None of these
- B. Data lake
- C. Data warehouse
- D. All of these
- E. Data lakehouse

Answer: E

NEW QUESTION 10

A data engineer only wants to execute the final block of a Python program if the Python variable day_of_week is equal to 1 and the Python variable review_period is True.

Which of the following control flow statements should the data engineer use to begin this conditionally executed code block?

- A. if day_of_week = 1 and review_period:
- B. if day_of_week = 1 and review_period = "True":
- C. if day_of_week == 1 and review_period == "True":
- D. if day_of_week == 1 and review_period:
- E. if day_of_week = 1 & review_period: = "True":

Answer: D

Explanation:

This statement will check if the variable `day_of_week` is equal to 1 and if the variable `review_period` evaluates to a truthy value. The use of the double equal sign (`==`) in the comparison of `day_of_week` is important, as a single equal sign (`=`) would be used to assign a value to the variable instead of checking its value. The use of a single ampersand (`&`) instead of the keyword `and` is not valid syntax in Python. The use of quotes around `True` in options B and C will result in a string comparison, which will not evaluate to `True` even if the value of `review_period` is `True`.

NEW QUESTION 10

Which of the following describes the storage organization of a Delta table?

- A. Delta tables are stored in a single file that contains data, history, metadata, and other attributes.
- B. Delta tables store their data in a single file and all metadata in a collection of files in a separate location.
- C. Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.
- D. Delta tables are stored in a collection of files that contain only the data stored within the table.
- E. Delta tables are stored in a single file that contains only the data stored within the table.

Answer: C

Explanation:

Delta tables store data in a structured manner using Parquet files, and they also maintain metadata and transaction logs in separate directories. This organization allows for versioning, transactional capabilities, and metadata tracking in Delta Lake. Thank you for pointing out the error, and I appreciate your understanding.

NEW QUESTION 11

Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

- A. Cloud-specific integrations
- B. Simplified governance
- C. Ability to scale storage
- D. Ability to scale workloads
- E. Avoiding vendor lock-in

Answer: E

Explanation:

<https://double.cloud/blog/posts/2023/01/break-free-from-vendor-lock-in-with-open-source-tech/>

NEW QUESTION 12

Which of the following benefits is provided by the array functions from Spark SQL?

- A. An ability to work with data in a variety of types at once
- B. An ability to work with data within certain partitions and windows
- C. An ability to work with time-related data in specified intervals
- D. An ability to work with complex, nested data ingested from JSON files
- E. An ability to work with an array of tables for procedural automation

Answer: D

Explanation:

Array functions in Spark SQL are primarily used for working with arrays and complex, nested data structures, such as those often encountered when ingesting JSON files. These functions allow you to manipulate and query nested arrays and structures within your data, making it easier to extract and work with specific elements or values within complex data formats. While some of the other options (such as option A for working with different data types) are features of Spark SQL or SQL in general, array functions specifically excel at handling complex, nested data structures like those found in JSON files.

NEW QUESTION 17

A data analyst has created a Delta table `sales` that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which of the following commands could the data engineering team use to access `sales` in PySpark?

- A. `SELECT * FROM sales`
- B. There is no way to share data between PySpark and SQL.
- C. `spark.sql("sales")`
- D. `spark.delta.table("sales")`
- E. `spark.table("sales")`

Answer: E

Explanation:

<https://spark.apache.org/docs/3.2.1/api/python/reference/api/pyspark.sql.SessionCatalog.html>

NEW QUESTION 22

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?

- A. They can set up an Alert with a custom template.

- B. They can set up an Alert with a new email alert destination.
- C. They can set up an Alert with one-time notifications.
- D. They can set up an Alert with a new webhook alert destination.
- E. They can set up an Alert without notifications.

Answer: D

NEW QUESTION 24

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?

- A. They can set up separate expectations for each table when developing their DLT pipeline.
- B. They cannot determine which table is dropping the records.
- C. They can set up DLT to notify them via email when records are dropped.
- D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.
- E. They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.

Answer: D

Explanation:

To identify the table in a Delta Live Tables (DLT) pipeline where data is being dropped due to quality concerns, the data engineer can navigate to the DLT pipeline page, click on each table in the pipeline, and view the data quality statistics. These statistics often include information about records dropped, violations of expectations, and other data quality metrics. By examining the data quality statistics for each table in the pipeline, the data engineer can determine at which table the data is being dropped.

NEW QUESTION 25

Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

- A. The ability to manipulate the same data using a variety of languages
- B. The ability to collaborate in real time on a single notebook
- C. The ability to set up alerts for query failures
- D. The ability to support batch and streaming workloads
- E. The ability to distribute complex data operations

Answer: D

Explanation:

Delta Lake is a key component of the Databricks Lakehouse Platform that provides several benefits, and one of the most significant benefits is its ability to support both batch and streaming workloads seamlessly. Delta Lake allows you to process and analyze data in real-time (streaming) as well as in batch, making it a versatile choice for various data processing needs. While the other options may be benefits or capabilities of Databricks or the Lakehouse Platform in general, they are not specifically associated with Delta Lake.

NEW QUESTION 29

A data engineer runs a statement every day to copy the previous day's sales into the table transactions. Each day's sales are in their own file in the location "/transactions/raw".

Today, the data engineer runs the following command to complete this task:

```
COPY INTO transactions
FROM "/transactions/raw"
FILEFORMAT = PARQUET;
```

After running the command today, the data engineer notices that the number of records in table transactions has not changed.

Which of the following describes why the statement might not have copied any new records into the table?

- A. The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.
- B. The names of the files to be copied were not included with the FILES keyword.
- C. The previous day's file has already been copied into the table.
- D. The PARQUET file format does not support COPY INTO.
- E. The COPY INTO statement requires the table to be refreshed to view the copied rows.

Answer: C

Explanation:

<https://docs.databricks.com/en/ingestion/copy-into/index.html> The COPY

INTO SQL command lets you load data from a file location into a Delta table. This is a re- triable and idempotent operation; files in the source location that have already been loaded are skipped. if there are no new records, the only consistent choice is C no new files were loaded because already loaded files were skipped.

NEW QUESTION 33

A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.

Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- A. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- B. They can set up the dashboard's SQL endpoint to be serverless.
- C. They can turn on the Auto Stop feature for the SQL endpoint.

- D. They can reduce the cluster size of the SQL endpoint.
- E. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

Answer: C

NEW QUESTION 34

Which of the following describes the relationship between Gold tables and Silver tables?

- A. Gold tables are more likely to contain aggregations than Silver tables.
- B. Gold tables are more likely to contain valuable data than Silver tables.
- C. Gold tables are more likely to contain a less refined view of data than Silver tables.
- D. Gold tables are more likely to contain more data than Silver tables.
- E. Gold tables are more likely to contain truthful data than Silver tables.

Answer: A

Explanation:

In some data processing pipelines, especially those following a typical "Bronze-Silver-Gold" data lakehouse architecture, Silver tables are often considered a more refined version of the raw or Bronze data. Silver tables may include data cleansing, schema enforcement, and some initial transformations. Gold tables, on the other hand, typically represent a stage where data is further enriched, aggregated, and processed to provide valuable insights for analytical purposes. This could indeed involve more aggregations compared to Silver tables.

NEW QUESTION 36

Which of the following Git operations must be performed outside of Databricks Repos?

- A. Commit
- B. Pull
- C. Push
- D. Clone
- E. Merge

Answer: E

Explanation:

For following tasks, work in your Git provider:
Create a pull request. Resolve merge conflicts. Merge or delete branches. Rebase a branch.
<https://docs.databricks.com/repos/index.html>

NEW QUESTION 37

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?

- A. It is not possible to use SQL in a Python notebook
- B. They can attach the cell to a SQL endpoint rather than a Databricks cluster
- C. They can simply write SQL syntax in the cell
- D. They can add %sql to the first line of the cell
- E. They can change the default language of the notebook to SQL

Answer: D

NEW QUESTION 42

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?

- A. Databricks Repos automatically saves development progress
- B. Databricks Repos supports the use of multiple branches
- C. Databricks Repos allows users to revert to previous versions of a notebook
- D. Databricks Repos provides the ability to comment on specific changes
- E. Databricks Repos is wholly housed within the Databricks Lakehouse Platform

Answer: B

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 of version 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.

NEW QUESTION 47

Which of the following describes a scenario in which a data team will want to utilize cluster pools?

- A. An automated report needs to be refreshed as quickly as possible.
- B. An automated report needs to be made reproducible.
- C. An automated report needs to be tested to identify errors.
- D. An automated report needs to be version-controlled across multiple collaborators.

E. An automated report needs to be runnable by all stakeholders.

Answer: A

Explanation:

Cluster pools are typically used in distributed computing environments, such as cloud-based data platforms like Databricks. They allow you to pre-allocate a set of compute resources (a cluster) for specific tasks or workloads. In this case, if an automated report needs to be refreshed as quickly as possible, you can allocate a cluster pool with sufficient resources to ensure fast data processing and report generation. This helps ensure that the report is generated with minimal latency and can be delivered to stakeholders in a timely manner. Cluster pools allow you to optimize resource allocation for high-demand, time-sensitive tasks like real-time report generation.

NEW QUESTION 49

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"SELECT customer_id, spend FROM {table_name}")
```

Which of the following can be used to fill in the blank to successfully complete the task?

- A. `spark.delta.sql`
- B. `spark.delta.table`
- C. `spark.table`
- D. `dbutils.sql`
- E. `spark.sql`

Answer: E

NEW QUESTION 51

A Delta Live Table pipeline includes two datasets defined using STREAMING LIVE TABLE. Three datasets are defined against Delta Lake table sources using LIVE TABLE.

The table is configured to run in Production mode using the Continuous Pipeline Mode. Assuming previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?

- A. All datasets will be updated at set intervals until the pipeline is shut down
- B. The compute resources will persist to allow for additional testing.
- C. All datasets will be updated once and the pipeline will persist without any processing
- D. The compute resources will persist but go unused.
- E. All datasets will be updated at set intervals until the pipeline is shut down
- F. The compute resources will be deployed for the update and terminated when the pipeline is stopped.
- G. All datasets will be updated once and the pipeline will shut down
- H. The compute resources will be terminated.
- I. All datasets will be updated once and the pipeline will shut down
- J. The compute resources will persist to allow for additional testing.

Answer: C

Explanation:

In a Delta Live Table pipeline running in Continuous Pipeline Mode, when you click Start to update the pipeline, the following outcome is expected: All datasets defined using STREAMING LIVE TABLE and LIVE TABLE against Delta Lake table sources will be updated at set intervals. The compute resources will be deployed for the update process and will be active during the execution of the pipeline. The compute resources will be terminated when the pipeline is stopped or shut down. This mode allows for continuous and periodic updates to the datasets as new data arrives or changes in the underlying Delta Lake tables occur. The compute resources are provisioned and utilized during the update intervals to process the data and perform the necessary operations.

NEW QUESTION 53

A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.

Which of the following approaches can the tech lead use to identify why the notebook is running slowly as part of the Job?

- A. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.
- B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
- C. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.
- D. There is no way to determine why a Job task is running slowly.
- E. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

Answer: C

Explanation:

The job run details page contains job output and links to logs, including information about the success or failure of each task in the job run. You can access job run details from the Runs tab for the job. To view job run details from the Runs tab, click the link for the run in the Start time column in the runs list view. To return to the Runs tab for the job, click the Job ID value.

If the job contains multiple tasks, click a task to view task run details, including: the cluster that ran the task

the Spark UI for the task logs for the task

metrics for the task

<https://docs.databricks.com/en/workflows/jobs/monitor-job-runs.html#job-run-details>

NEW QUESTION 58

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?

- A. None of these changes will need to be made
- B. The pipeline will need to stop using the medallion-based multi-hop architecture
- C. The pipeline will need to be written entirely in SQL
- D. The pipeline will need to use a batch source in place of a streaming source
- E. The pipeline will need to be written entirely in Python

Answer: A

NEW QUESTION 62

Which of the following data workloads will utilize a Gold table as its source?

- A. A job that enriches data by parsing its timestamps into a human-readable format
- B. A job that aggregates uncleaned data to create standard summary statistics
- C. A job that cleans data by removing malformed records
- D. A job that queries aggregated data designed to feed into a dashboard
- E. A job that ingests raw data from a streaming source into the Lakehouse

Answer: D

NEW QUESTION 66

A dataset has been defined using Delta Live Tables and includes an expectations clause:

CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION FAIL UPDATE

What is the expected behavior when a batch of data containing data that violates these constraints is processed?

- A. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- B. Records that violate the expectation cause the job to fail.
- C. Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table.
- D. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
- E. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.

Answer: B

Explanation:

<https://docs.databricks.com/en/delta-live-tables/expectations.html> Action

Result

warn (default)

Invalid records are written to the target; failure is reported as a metric for the dataset. drop

Invalid records are dropped before data is written to the target; failure is reported as a metrics for the dataset.

fail

Invalid records prevent the update from succeeding. Manual intervention is required before re-processing.

NEW QUESTION 71

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?

A)

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

B)

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

C)

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

D)

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

E)

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

- A. Option A
- B. Option B
- C. Option C
- D. Option D
- E. Option E

Answer: D

Explanation:

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

NEW QUESTION 76

An engineering manager uses a Databricks SQL query to monitor ingestion latency for each data source. The manager checks the results of the query every day, but they are manually rerunning the query each day and waiting for the results.

Which of the following approaches can the manager use to ensure the results of the query are updated each day?

- A. They can schedule the query to refresh every 1 day from the SQL endpoint's page in Databricks SQL.
- B. They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL.
- C. They can schedule the query to refresh every 1 day from the query's page in Databricks SQL.
- D. They can schedule the query to run every 1 day from the Jobs UI.
- E. They can schedule the query to run every 12 hours from the Jobs UI.

Answer: C

NEW QUESTION 80

In which of the following scenarios should a data engineer use the MERGE INTO command instead of the INSERT INTO command?

- A. When the location of the data needs to be changed
- B. When the target table is an external table
- C. When the source table can be deleted
- D. When the target table cannot contain duplicate records
- E. When the source is not a Delta table

Answer: D

Explanation:

With merge , you can avoid inserting the duplicate records. The dataset containing the new logs needs to be deduplicated within itself. By the SQL semantics of merge, it matches and deduplicates the new data with the existing data in the table, but if there is duplicate data within the new dataset, it is inserted.
<https://docs.databricks.com/en/delta/merge.html#:~:text=With%20merge%20%2C%20you%20can%20avoid%20inserting%20the%20duplicate%20records.&text=The%20dataset%20containing%20the%20new,new%20dataset%2C%20it%20is%20inserted.>

NEW QUESTION 81

Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my_table and save the updated table?

- A. SELECT * FROM my_table WHERE age > 25;
- B. UPDATE my_table WHERE age > 25;
- C. DELETE FROM my_table WHERE age > 25;
- D. UPDATE my_table WHERE age <= 25;
- E. DELETE FROM my_table WHERE age <= 25;

Answer: C

NEW QUESTION 83

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?

- A. Unity Catalog
- B. Data Explorer
- C. Delta Lake
- D. Delta Live Tables
- E. Auto Loader

Answer: D

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 engineering environment, Delta Live Tables is specifically designed for data quality monitoring and automation within the Databricks ecosystem.

NEW QUESTION 87

A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.

Which of the following data entities should the data engineer create?

- A. Database
- B. Function
- C. View
- D. Temporary view

E. Table

Answer: E

Explanation:

In the context described, creating a "Table" is the most suitable choice. Tables in SQL are data entities that exist independently of any session and are saved in a physical location. They can be accessed and manipulated by other data engineers in different sessions, which aligns with the requirements stated. A "Database" is a collection of tables, views, and other database objects. A "Function" is a stored procedure that performs an operation. A "View" is a virtual table based on the result-set of an SQL statement, but it is not stored physically. A "Temporary view" is a feature that allows you to store the result of a query as a view that disappears once your session with the database is closed.

NEW QUESTION 92

A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.

Which of the following describes why Auto Loader inferred all of the columns to be of the string type?

- A. There was a type mismatch between the specific schema and the inferred schema
- B. JSON data is a text-based format
- C. Auto Loader only works with string data
- D. All of the fields had at least one null value
- E. Auto Loader cannot infer the schema of ingested data

Answer: B

Explanation:

JSON data is a text-based format that uses strings to represent all values. When Auto Loader infers the schema of JSON data, it assumes that all values are strings. This is because Auto Loader cannot determine the type of a value based on its string representation. <https://docs.databricks.com/en/ingestion/auto-loader/schema.html> Foreexample, the following JSON string represents a value that is logically a boolean: JSON "true" Use code with caution. Learn more However, Auto Loader would infer that the type of this value is string. This is because Auto Loader cannot determine that the value is a boolean based on its string representation. In order to get Auto Loader to infer the correct types for columns, the data engineer can provide type inference or schema hints. Type inference hints can be used to specify the types of specific columns. Schema hints can be used to provide the entire schema of the data. Therefore, the correct answer is B. JSON data is a text-based format.

NEW QUESTION 97

Which of the following Structured Streaming queries is performing a hop from a Silver table to a Gold table?

A.

```
(spark.readStream.load(rawSalesLocation)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```

B.

```
(spark.read.load(rawSalesLocation)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```

C.

```
(spark.table("sales")
  .withColumn("avgPrice", col("sales") / col("units"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```

D.

```
(spark.table("sales")
  .filter(col("units") > 0)
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("append")
  .table("newSales")
)
```

E.

```
(spark.table("sales")
  .groupBy("store")
  .agg(sum("sales"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("complete")
  .table("newSales")
)
```

A.

Answer: E

NEW QUESTION 102

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?

- A. They can use endpoints available in Databricks SQL
- B. They can use jobs clusters instead of all-purpose clusters
- C. They can configure the clusters to be single-node
- D. They can use clusters that are from a cluster pool
- E. They can configure the clusters to autoscale for larger data sizes

Answer: D

Explanation:

Cluster pools are a way to pre-provision clusters that are ready to use. This can reduce the start up time for clusters, as they do not have to be created from scratch. All-purpose clusters are not pre-provisioned, so they will take longer to start up. Jobs clusters are a type of cluster pool, but they are not the best option for this use case. Jobs clusters are designed for long-running jobs, and they can be more expensive than other types of cluster pools. Single-node clusters are the smallest type of cluster, and they will start up the fastest. However, they may not be powerful enough to run the Job's tasks. Autoscaling clusters can scale up or down based on demand. This can help to improve the start up time for clusters, as they will only be created when they are needed. However, autoscaling clusters can also be more expensive than other types of cluster pool <https://docs.databricks.com/en/clusters/pool-best-practices.html>

NEW QUESTION 103

A data engineer needs access to a table new_table, but they do not have the correct permissions. They can ask the table owner for permission, but they do not know who the table owner is. Which of the following approaches can be used to identify the owner of new_table?

- A. Review the Permissions tab in the table's page in Data Explorer
- B. All of these options can be used to identify the owner of the table
- C. Review the Owner field in the table's page in Data Explorer
- D. Review the Owner field in the table's page in the cloud storage solution
- E. There is no way to identify the owner of the table

Answer: C

NEW QUESTION 107

Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

- A. Parquet files can be partitioned
- B. CREATE TABLE AS SELECT statements cannot be used on files
- C. Parquet files have a well-defined schema
- D. Parquet files have the ability to be optimized
- E. Parquet files will become Delta tables

Answer: C

Explanation:

<https://www.databricks.com/glossary/what-is-parquet#:~:text=Columnar%20storage%20like%20Apache%20Parquet,compared%20to%20row%2Doriented%20databases.> Columnar storage like Apache Parquet is designed to bring efficiency compared to row-based files like CSV. When querying, columnar storage you can skip over the non-relevant data very quickly. As a result, aggregation queries are less time-consuming compared to row-oriented databases.

NEW QUESTION 111

A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table. The code block used by the data engineer is below:

```
(spark.readStream
  .table("sales")
  .withColumn("avg_price", col("sales") / col("units"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("complete")
  * _____
  .table("new_sales")
)
```

If the data engineer only wants the query to process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

- A. processingTime(1)
- B. trigger(availableNow=True)
- C. trigger(parallelBatch=True)
- D. trigger(processingTime="once")
- E. trigger(continuous="once")

Answer: B

Explanation:

<https://stackoverflow.com/questions/71061809/trigger-availablenow-for-delta-source-streaming-queries-in-pyspark-databricks>

NEW QUESTION 113

.....

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questions and Answers in PDF Format

Databricks-Certified-Data-Engineer-Associate Practice Exam Features:

- * Databricks-Certified-Data-Engineer-Associate Questions and Answers Updated Frequently
- * Databricks-Certified-Data-Engineer-Associate Practice Questions Verified by Expert Senior Certified Staff
- * Databricks-Certified-Data-Engineer-Associate Most Realistic Questions that Guarantee you a Pass on Your FirstTry
- * Databricks-Certified-Data-Engineer-Associate Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The Databricks-Certified-Data-Engineer-Associate Practice Test Here](#)