Databricks

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

NEW QUESTION 1
In which of the following scenarios should a data engineer select a Task in the Depends On field of a new Databricks Job Task?

A. When another task needs to be replaced by the new task
B. When another task needs to fail before the new task begins
C. When another task has the same dependency libraries as the new task
D. When another task needs to use as little compute resources as possible
E. When another task needs to successfully complete before the new task begins

Answer: E

NEW QUESTION 2
Which of the following commands will return the location of database customer360?

A. DESCRIBE LOCATION customer360;
B. DROP DATABASE customer360;
C. DESCRIBE DATABASE customer360;
D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user'};
E. USE DATABASE customer360;

Answer: C

Explanation:
To retrieve the location of a database named "customer360" in a database management system like Hive or Databricks, you can use the DESCRIBE DATABASE command followed by the database name. This command will provide information about the database, including its location.

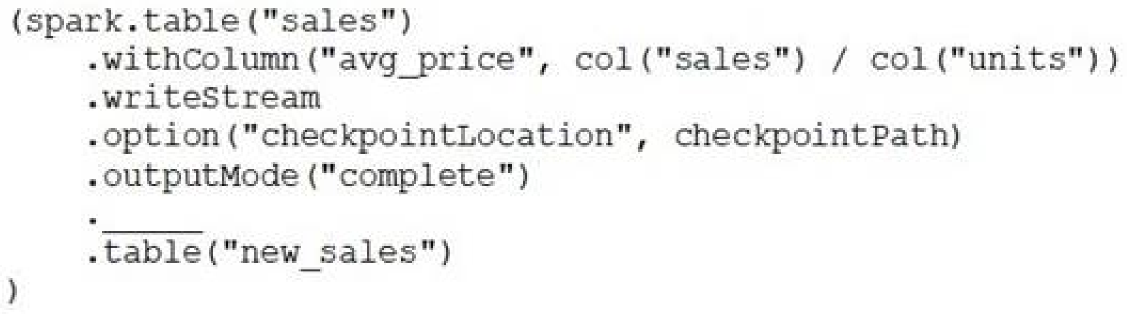
NEW QUESTION 3
A data engineer has created a new database using the following command: CREATE DATABASE IF NOT EXISTS customer360;
In which of the following locations will the customer360 database be located?

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 4
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 cade block used by the data engineer is below:



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

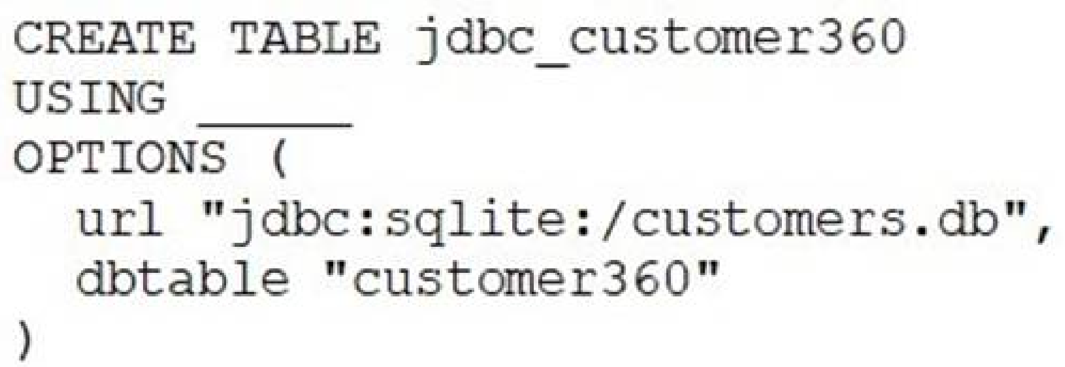
If the data engineer only wants the query to execute a micro-batch to process data every 5 seconds, which of the following lines of code should the data engineer use to fill in the blank?

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 5
A data engineer needs to create a table in Databricks using data from their organization’s existing SQLite database.
They run the following command:



```
CREATE TABLE jdbc_customer360
USING _____
OPTIONS (
   url "jdbc:sqlite:/customers.db",
   dbtable "customer360"
)
```

Which of the following lines of code fills in the above blank to successfully complete the task?

A. org.apache.spark.sql.jdbc
B. autoloader
C. DELTA
D. sqlite
E. org.apache.spark.sql.sqlite

Answer: A

Explanation:
CREATE TABLE new_employees_table USING JDBC
OPTIONS (
url "<jdbc_url>",
dbtable "<table_name>", user '<username>', password '<password>'
) AS
SELECT * FROM employees_table_vw https://docs.databricks.com/external-data/jdbc.html#language-sql

NEW QUESTION 6
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?

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

Answer: B

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.

NEW QUESTION 7
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 8
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 9
A data engineer is attempting to drop a Spark SQL table my_table. The data engineer wants to delete all table metadata and data.
They run the following command: DROP TABLE IF EXISTS my_table
While the object no longer appears when they run SHOW TABLES, the data files still exist.
Which of the following describes why the data files still exist and the metadata files were deleted?

A. The table’s data was larger than 10 GB
B. The table’s data was smaller than 10 GB
C. The table was external
D. The table did not have a location
E. The table was managed

Answer: C

Explanation:
The reason why the data files still exist while the metadata files were deleted is because the table was external. When a table is external in Spark SQL (or in other database systems), it means that the table metadata (such as schema information and table structure) is managed externally, and Spark SQL assumes that the data is managed and maintained outside of the system. Therefore, when you execute a DROP TABLE statement for an external table, it removes only the table metadata from the catalog, leaving the data files intact. On the other hand, for managed tables (option E), Spark SQL manages both the metadata and the data files. When you drop a managed table, it deletes both the metadata and the associated data files, resulting in a complete removal of the table.

NEW QUESTION 10
A data engineer is attempting to drop a Spark SQL table my_table and runs the following command:
DROP TABLE IF EXISTS my_table;
After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.
Which of the following describes why all of these files were deleted?

A. The table was managed
B. The table's data was smaller than 10 GB
C. The table's data was larger than 10 GB
D. The table was external
E. The table did not have a location

Answer: A

Explanation:
managed tables files and metadata are managed by metastore and will be deleted when the table is dropped . while external tables the metadata is stored in a external location. hence when a external table is dropped you clear off only the metadata and the files (data) remain.

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 15
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 18
Which of the following statements regarding the relationship between Silver tables and Bronze tables is always true?

A. Silver tables contain a less refined, less clean view of data than Bronze data.

B. Silver tables contain aggregates while Bronze data is unaggregated.
C. Silver tables contain more data than Bronze tables.
D. Silver tables contain a more refined and cleaner view of data than Bronze tables.
E. Silver tables contain less data than Bronze tables.

Answer: D

Explanation:
https://www.databricks.com/glossary/medallion-architecture

NEW QUESTION 22
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

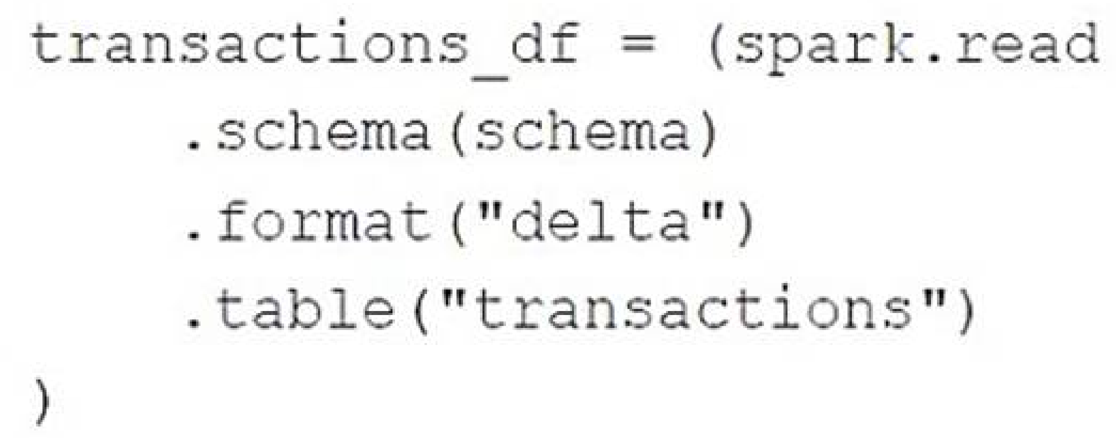
Explanation:
https://spark.apache.org/docs/3.2.1/api/python/reference/api/pyspark.sql.SparkSession.tabl e.html

NEW QUESTION 25
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?

A. GRANT ALL PRIVILEGES ON TABLE sales TO team;
B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
C. GRANT SELECT ON TABLE sales TO team;
D. GRANT USAGE ON TABLE sales TO team;
E. GRANT ALL PRIVILEGES ON TABLE team TO sales;

Answer: A

NEW QUESTION 26
A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:



```
transactions_df = (spark.read
        .schema(schema)
        .format("delta")
        .table("transactions")
)
```

Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

A. Replace predict with a stream-friendly prediction function
B. Replace schema(schema) with option ("maxFilesPerTrigger", 1)
C. Replace "transactions" with the path to the location of the Delta table
D. Replace format("delta") with format("stream")
E. Replace spark.read with spark.readStream

Answer: E

Explanation:
https://docs.databricks.com/en/structured-streaming/delta-lake.html

NEW QUESTION 27
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?

A. Checkpointing and Write-ahead Logs
B. Structured Streaming cannot record the offset range of the data being processed in each trigger.
C. Replayable Sources and Idempotent Sinks
D. Write-ahead Logs and Idempotent Sinks
E. Checkpointing and Idempotent Sinks

Answer: A

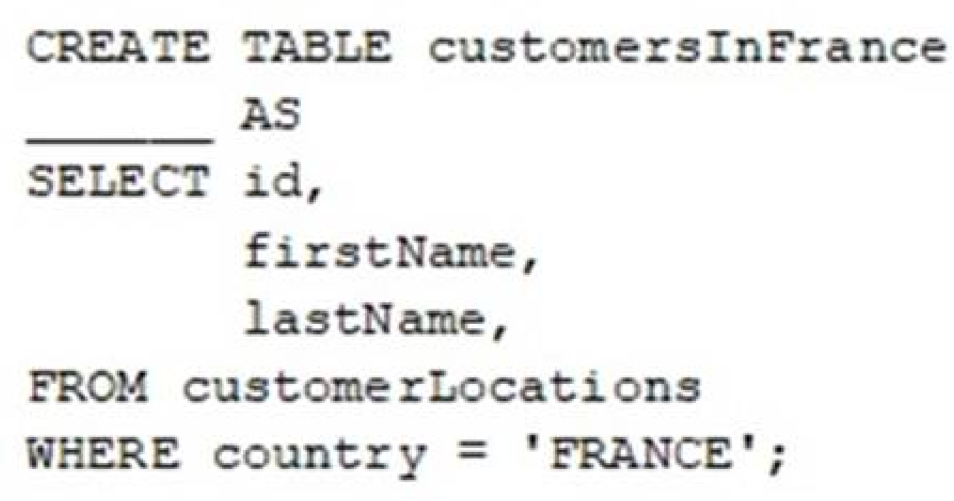
Explanation:
The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger. -- in the link search for "The engine uses " youll find the answer.https://spark.apache.org/docs/latest/structured-streaming- programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processe d%20in%20each%20trigger.

NEW QUESTION 30
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 35
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?

A. There is no way to indicate whether a table contains PII.
B. "COMMENT PII"
C. TBLPROPERTIES PII
D. COMMENT "Contains PII"
E. PII

Answer: D

Explanation:
Ref:https://www.databricks.com/discover/pages/data-quality-management CREATE TABLE my_table (id INT COMMENT 'Unique Identification Number', name STRING COMMENT 'PII', age INT COMMENT 'PII') TBLPROPERTIES
('contains_pii'=True) COMMENT 'Contains PII';

NEW QUESTION 37
Which of the following is hosted completely in the control plane of the classic Databricks architecture?

A. Worker node
B. JDBC data source
C. Databricks web application
D. Databricks Filesystem
E. Driver node

Answer: C

Explanation:
In the classic Databricks architecture, the control plane includes components like the Databricks web application, the Databricks REST API, and the Databricks Workspace. These components are responsible for managing and controlling the Databricks environment, including cluster provisioning, notebook management, access control, and job scheduling. The other options, such as worker nodes, JDBC data sources, Databricks Filesystem (DBFS), and driver nodes, are typically part of the data plane or the execution environment, which is separate from the control plane. Worker nodes are responsible for executing tasks and computations, JDBC data sources are used to connect to external databases, DBFS is a distributed file system for data storage, and driver nodes are responsible for coordinating the execution of Spark jobs.

NEW QUESTION 39
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.

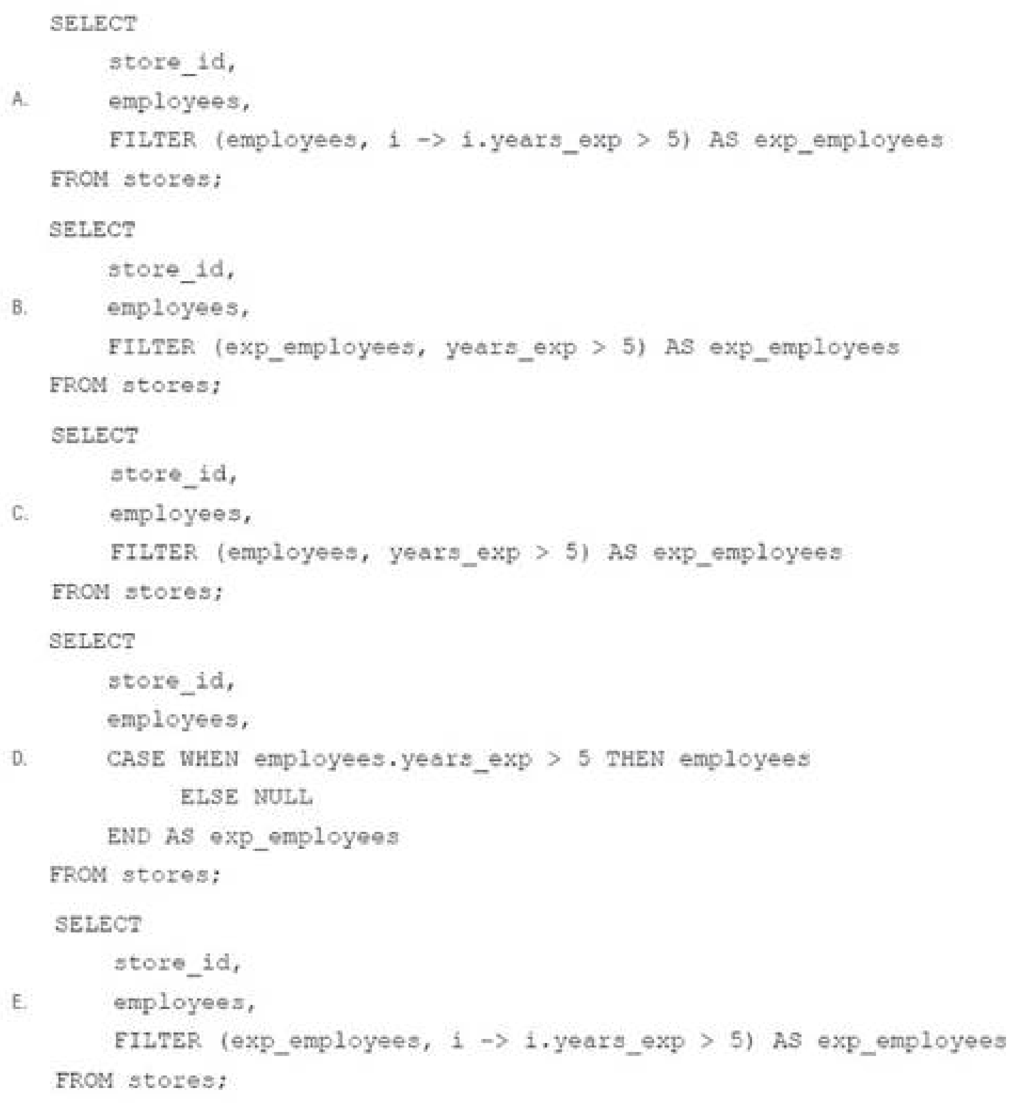
NEW QUESTION 43
A dataset has been defined using Delta Live Tables and includes an expectations clause:
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW
What is the expected behavior when a batch of data containing data that violates these constraints is processed?

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

Answer: C

Explanation:
With the defined constraint and expectation clause, when a batch of data is processed, any records that violate the expectation (in this case, where the timestamp is not greater than '2020-01-01') will be dropped from the target dataset. These dropped records will also be recorded as invalid in the event log, allowing for auditing and tracking of the data quality issues without causing the entire job to fail. https://docs.databricks.com/en/delta-live-tables/expectations.html

NEW QUESTION 48
A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.
Which of the following code blocks successfully completes this task?



```
    SELECT
        store_id,
A.      employees,
        FILTER (employees, i -> i.years_exp > 5) AS exp_employees
    FROM stores;

    SELECT
        store_id,
B.      employees,
        FILTER (exp_employees, years_exp > 5) AS exp_employees
    FROM stores;

    SELECT
        store_id,
C.      employees,
        FILTER (employees, years_exp > 5) AS exp_employees
    FROM stores;

    SELECT
        store_id,
        employees,
D.      CASE WHEN employees.years_exp > 5 THEN employees
             ELSE NULL
        END AS exp_employees
    FROM stores;

    SELECT
        store_id,
E.      employees,
        FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees
    FROM stores;
```

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

Answer: A

NEW QUESTION 49
Which of the following commands will return the number of null values in the member_id column?

A. SELECT count(member_id) FROM my_table;
B. SELECT count(member_id) - count_null(member_id) FROM my_table;
C. SELECT count_if(member_id IS NULL) FROM my_table;
D. SELECT null(member_id) FROM my_table;
E. SELECT count_null(member_id) FROM my_table;

Answer: C

Explanation:
https://docs.databricks.com/en/sql/language-manual/functions/count.html
Returns
A BIGINT.
If * is specified also counts row containing NULL values.
If expr are specified counts only rows for which all expr are not NULL. If DISTINCT duplicate rows are not counted.

NEW QUESTION 54
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 58
Which of the following commands can be used to write data into a Delta table while avoiding the writing of duplicate records?

A. DROP
B. IGNORE
C. MERGE
D. APPEND
E. INSERT

Answer: C

Explanation:
To write data into a Delta table while avoiding the writing of duplicate records, you can use the MERGE command. The MERGE command in Delta Lake allows you to combine the ability to insert new records and update existing records in a single atomic operation. The MERGE command compares the data being written with the existing data in the Delta table based on specified matching criteria, typically using a primary key or unique identifier. It then performs conditional actions, such as inserting new records or updating existing records, depending on the comparison results. By using the MERGE command, you can handle the prevention of duplicate records in a more controlled and efficient manner. It allows you to synchronize and reconcile data from different sources while avoiding duplication and ensuring data integrity.

NEW QUESTION 63
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

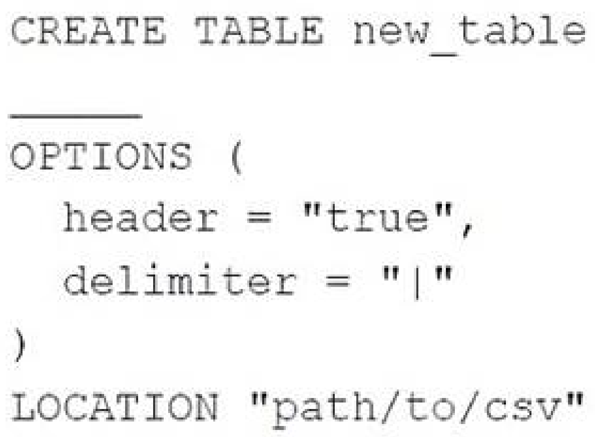
Explanation:
For following tasks, work in your Git provider:
Create a pull request. Resolve merge conflicts. Merge or delete branches. Rebase a branch.
https://docs.databricks.com/repos/index.html

NEW QUESTION 66
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 68
A data engineer needs to create a table in Databricks using data from a CSV file at location
/path/to/csv.
They run the following command:



```
CREATE TABLE new_table

______
OPTIONS (
   header = "true",
   delimiter = "|"
)
LOCATION "path/to/csv"
```

Which of the following lines of code fills in the above blank to successfully complete the task?

A. None of these lines of code are needed to successfully complete the task
B. USING CSV
C. FROM CSV
D. USING DELTA
E. FROM "path/to/csv"

Answer: B

NEW QUESTION 70
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 72
A data engineer has a Job that has a complex run schedule, and they want to transfer that schedule to other Jobs.
Rather than manually selecting each value in the scheduling form in Databricks, which of the following tools can the data engineer use to represent and submit the schedule programmatically?

A. pyspark.sql.types.DateType
B. datetime
C. pyspark.sql.types.TimestampType
D. Cron syntax
E. There is no way to represent and submit this information programmatically

Answer: D

NEW QUESTION 74
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 dow
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 processin
D. The compute resources will persist but go unused.
E. All datasets will be updated at set intervals until the pipeline is shut dow
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 dow
H. The compute resources will be terminated.
I. All datasets will be updated once and the pipeline will shut dow
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 79
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 80
A data engineer has left the organization. The data team needs to transfer ownership of the data engineer's Delta tables to a new data engineer. The new data engineer is the lead engineer on the data team.
Assuming the original data engineer no longer has access, which of the following individuals must be the one to transfer ownership of the Delta tables in Data Explorer?

A. Databricks account representative
B. This transfer is not possible
C. Workspace administrator
D. New lead data engineer
E. Original data engineer

Answer: C

Explanation:
https://docs.databricks.com/sql/admin/transfer-ownership.html

NEW QUESTION 81
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 84
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 88

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 Forexample, 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 93

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 98

A data engineer has joined an existing project and they see the following query in the project repository:
CREATE STREAMING LIVE TABLE loyal_customers AS SELECT customer_id -
FROM STREAM(LIVE.customers) WHERE loyalty_level = 'high';
Which of the following describes why the STREAM function is included in the query?

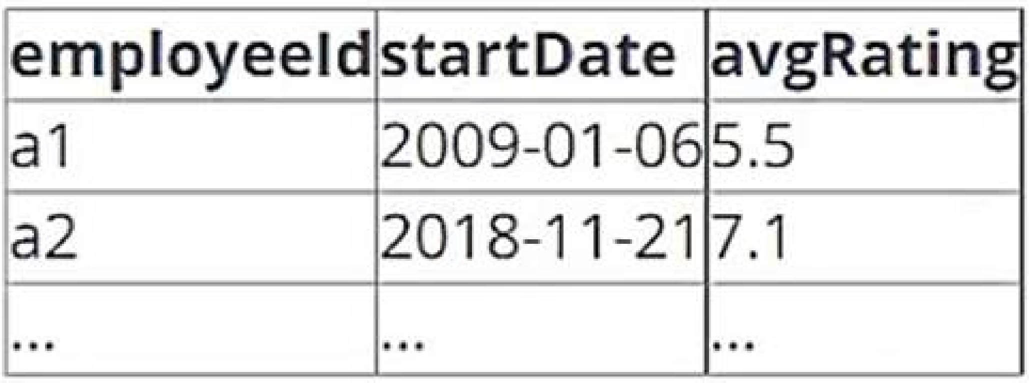
A. The STREAM function is not needed and will cause an error.
B. The table being created is a live table.
C. The customers table is a streaming live table.
D. The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.
E. The data in the customers table has been updated since its last run.

Answer: C

Explanation:
https://docs.databricks.com/en/sql/load-data-streaming-table.html Load data into a streaming table
To create a streaming table from data in cloud object storage, paste the following into the query editor, and then click Run:
SQL
Copy to clipboardCopy
/* Load data from a volume */
CREATE OR REFRESH STREAMING TABLE <table-name> AS SELECT * FROM STREAM
read_files('/Volumes/<catalog>/<schema>/<volume>/<path>/<folder>')
/* Load data from an external location */

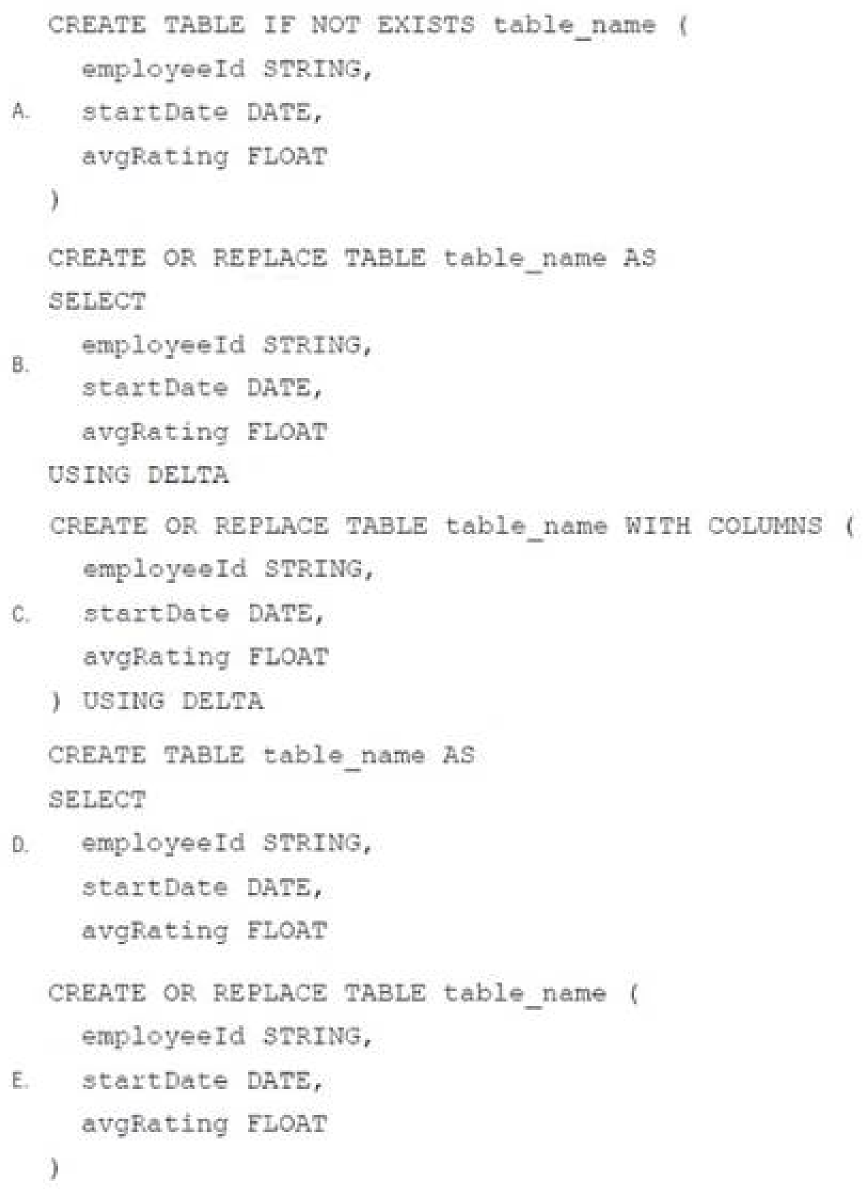
CREATE OR REFRESH STREAMING TABLE <table-name> AS
SELECT * FROM STREAM read_files('s3://<bucket>/<path>/<folder>')

NEW QUESTION 99
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?

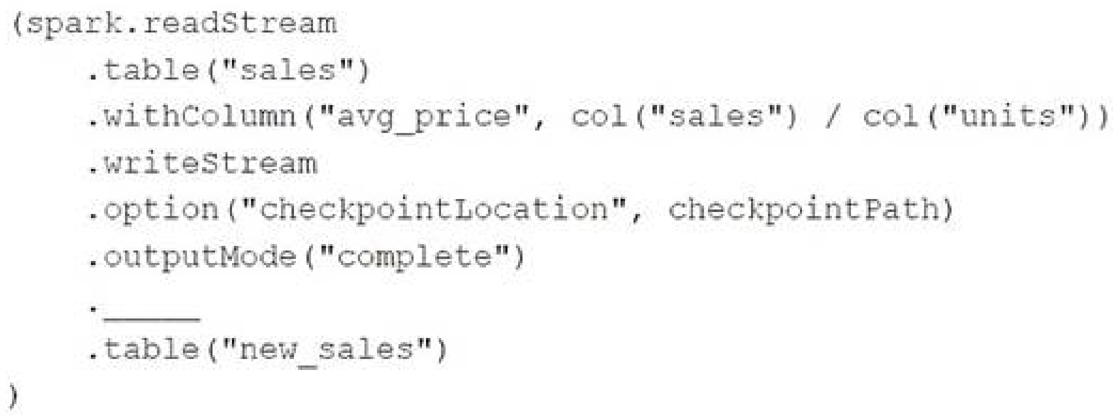


```
A.
CREATE TABLE IF NOT EXISTS table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)

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
)
```

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

Answer: E

NEW QUESTION 103
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 107
......