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
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 3
A data engineering team has two tables. The first table march_transactions is a collection of all retail transactions in the month of March. The second table april_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables.
Which of the following commands should be run to create a new table all_transactions that contains all records from march_transactions and april_transactions without duplicate records?

A. CREATE TABLE all_transactions AS SELECT * FROM march_transactionsINNER JOIN SELECT * FROM april_transactions;
B. CREATE TABLE all_transactions AS SELECT * FROM march_transactions UNION SELECT * FROM april_transactions;
C. CREATE TABLE all_transactions AS SELECT * FROM march_transactionsOUTER JOIN SELECT * FROM april_transactions;
D. CREATE TABLE all_transactions AS SELECT * FROM march_transactionsINTERSECT SELECT * from april_transactions;
E. CREATE TABLE all_transactions AS SELECT * FROM march_transactions MERGE SELECT * FROM april_transactions;

Answer: B

Explanation:
To create a new table all_transactions that contains all records from march_transactions and april_transactions without duplicate records, you should use the UNION operator, as shown in option B. This operator combines the result sets of the two tables while automatically removing duplicate records.

NEW QUESTION 4
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 5
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 6
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 7
A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that feeds the dashboard is automatically processed using a Databricks Job.
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 turn on the Auto Stop feature for the SQL endpoint.
B. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.
C. They can reduce the cluster size of the SQL endpoint.
D. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
E. They can set up the dashboard's SQL endpoint to be serverless.

Answer: A

NEW QUESTION 8
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 9
Which of the following tools is used by Auto Loader process data incrementally?

A. Checkpointing
B. Spark Structured Streaming
C. Data Explorer
D. Unity Catalog
E. Databricks SQL

Answer: B

Explanation:
The Auto Loader process in Databricks is typically used in conjunction with Spark Structured Streaming to process data incrementally. Spark Structured Streaming is a real-time data processing framework that allows you to process data streams incrementally as new data arrives. The Auto Loader is a feature in Databricks that works with Structured Streaming to automatically detect and process new data files as they are added to a specified data source location. It allows for incremental data processing without the need for manual intervention.
How does Auto Loader track ingestion progress? As files are discovered, their metadata is persisted in a scalable key-value store (RocksDB) in the checkpoint

location of your Auto Loader pipeline. This key-value store ensures that data is processed exactly once. In case of failures, Auto Loader can resume from where it left off by information stored in the checkpoint location and continue to provide exactly-once guarantees when writing data into Delta Lake. You don't need to maintain or manage any state yourself to achieve fault tolerance or exactly-once semantics.https://docs.databricks.com/ingestion/auto- loader/index.html

NEW QUESTION 10
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 10
Which of the following describes the relationship between Bronze tables and raw data?

A. Bronze tables contain less data than raw data files.
B. Bronze tables contain more truthful data than raw data.
C. Bronze tables contain aggregates while raw data is unaggregated.
D. Bronze tables contain a less refined view of data than raw data.
E. Bronze tables contain raw data with a schema applied.

Answer: E

Explanation:
The Bronze layer is where we land all the data from external source systems. The table structures in this layer correspond to the source system table structures "as-is," along with any additional metadata columns that capture the load date/time, process ID, etc. The focus in this layer is quick Change Data Capture and the ability to provide an historical archive of source (cold storage), data lineage, auditability, reprocessing if needed without rereading the data from the source system.https://www.databricks.com/glossary/medallion- architecture#:~:text=Bronze%20layer%20%28raw%20data%29

NEW QUESTION 15
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 19
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 21
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 23
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 24
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 29
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 32
A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.
Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?

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

Answer: C

Explanation:
To achieve this, the data engineer can set up an Alert in the Databricks workspace that triggers when the query results exceed the threshold of 100 NULL values. They can create a new webhook alert destination in the Alert's configuration settings and provide the necessary messaging webhook URL to receive notifications. When the Alert is triggered, it will send a message to the configured webhook URL, which will then notify the entire team of the issue.

NEW QUESTION 36
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 40
An engineering manager wants to monitor the performance of a recent project using a Databricks SQL query. For the first week following the project's release, the manager wants the query results to be updated every minute. However, the manager is concerned that the compute resources used for the query will be left running and cost the organization a lot of money beyond the first week of the project's release.
Which of the following approaches can the engineering team use to ensure the query does not cost the organization any money beyond the first week of the project's release?

A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
B. They can set the query's refresh schedule to end after a certain number of refreshes.
C. They cannot ensure the query does not cost the organization money beyond the first week of the project's release.
D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.
E. They can set the query's refresh schedule to end on a certain date in the query scheduler.

Answer: E

Explanation:
If a dashboard is configured for automatic updates, it has a Scheduled button at the top, rather than a Schedule button. To stop automatically updating the dashboard and remove its subscriptions:
Click Scheduled.
In the Refresh every drop-down, select Never.
Click Save. The Scheduled button label changes to Schedule. Source:https://learn.microsoft.com/en-us/azure/databricks/sql/user/dashboards/

NEW QUESTION 43
A data engineer needs to apply custom logic to string column city in table stores for a specific use case. In order to apply this custom logic at scale, the data engineer wants to create a SQL user-defined function (UDF).
Which of the following code blocks creates this SQL UDF?
A.
```
CREATE FUNCTION combine_nyc(city STRING)
RETURNS STRING
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```
B.
```
CREATE UDF combine_nyc(city STRING)
RETURNS STRING
CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```
C.
```
CREATE UDF combine_nyc(city STRING)
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```
D.

```
CREATE FUNCTION combine_nyc(city STRING)
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

E.

```
CREATE UDF combine_nyc(city STRING)
RETURNS STRING
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

A.

Answer: A

Explanation:
https://www.databricks.com/blog/2021/10/20/introducing-sql-user-defined- functions.html

NEW QUESTION 46
A data engineer has been given a new record of data:
id STRING = 'a1'
rank INTEGER = 6 rating FLOAT = 9.4
Which of the following SQL commands can be used to append the new record to an existing Delta table my_table?

A. INSERT INTO my_table VALUES ('a1', 6, 9.4)
B. my_table UNION VALUES ('a1', 6, 9.4)
C. INSERT VALUES ('a1' , 6, 9.4) INTO my_table
D. UPDATE my_table VALUES ('a1', 6, 9.4)
E. UPDATE VALUES ('a1', 6, 9.4) my_table

Answer: A

NEW QUESTION 48
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 53
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 malformatted 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 55
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 58
A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.
Which of the following approaches can the data engineer use to set up the new task?

A. They can clone the existing task in the existing Job and update it to run the new notebook.
B. They can create a new task in the existing Job and then add it as a dependency of the original task.
C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.
D. They can create a new job from scratch and add both tasks to run concurrently.
E. They can clone the existing task to a new Job and then edit it to run the new notebook.

Answer: B

Explanation:
To set up the new task to run a new notebook prior to the original task in a single-task Job, the data engineer can use the following approach: In the existing Job, create a new task that corresponds to the new notebook that needs to be run. Set up the new task with the appropriate configuration, specifying the notebook to be executed and any necessary parameters or dependencies. Once the new task is created, designate it as a dependency of the original task in the Job configuration. This ensures that the new task is executed before the original task.

NEW QUESTION 62
A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data.
Which of the following relational objects should the data engineer create?

A. Spark SQL Table
B. View
C. Database
D. Temporary view
E. Delta Table

Answer: D

Explanation:
Temp view : session based Create temp view view_name as query All these are termed as session ended: Opening a new notebook Detaching and reattaching a cluster Installing a python package Restarting a cluster

NEW QUESTION 67
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 72
A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.
Which of the following keywords can be used to compact the small files?

A. REDUCE
B. OPTIMIZE
C. COMPACTION
D. REPARTITION
E. VACUUM

Answer: B

Explanation:
OPTIMIZE can be used to club small files into 1 and improve performance.

NEW QUESTION 73
A data engineer is designing a data pipeline. The source system generates files in a shared directory that is also used by other processes. As a result, the files should be kept as is and will accumulate in the directory. The data engineer needs to identify which files are new since the previous run in the pipeline, and set up the pipeline to only ingest those new files with each run.
Which of the following tools can the data engineer use to solve this problem?

A. Unity Catalog
B. Delta Lake
C. Databricks SQL
D. Data Explorer
E. Auto Loader

Answer: E

Explanation:
Auto Loader incrementally and efficiently processes new data files as they arrive in cloud storage without any additional setup.https://docs.databricks.com/en/ingestion/auto-loader/index.html

NEW QUESTION 78
.....

Thank You for Trying Our Product

We offer two products:

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

2nd - Questons 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