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 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 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 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 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 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 6
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 7
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 8
A new data engineering team team. has been assigned to an ELT project. The new data engineering team will need full privileges on the database customers 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 USAGE ON DATABASE customers TO team;
B. GRANT ALL PRIVILEGES ON DATABASE team TO customers;
C. GRANT SELECT PRIVILEGES ON DATABASE customers TO teams;
D. GRANT SELECT CREATE MODIFY USAGE PRIVILEGES ON DATABASE customersTO team;
E. GRANT ALL PRIVILEGES ON DATABASE customers TO team;

Answer: E

Explanation:
To grant full privileges on the database "customers" to the new data engineering team, you can use the GRANT ALL PRIVILEGES command as shown in option E. This command provides the team with all possible privileges on the specified database, allowing them to fully manage it.

NEW QUESTION 9
A data analyst has a series of queries in a SQL program. The data analyst wants this program to run every day. They only want the final query in the program to run on Sundays. They ask for help from the data engineering team to complete this task.
Which of the following approaches could be used by the data engineering team to complete this task?

A. They could submit a feature request with Databricks to add this functionality.
B. They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.
C. They could only run the entire program on Sundays.
D. They could automatically restrict access to the source table in the final query so that it is only accessible on Sundays.
E. They could redesign the data model to separate the data used in the final query into a new table.

Answer: B

NEW QUESTION 10
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 ofversion control systems like Git, which Databricks Repos is built upon. It allows you to create separate branches for different tasks, features, or experiments within your project. This separation helps in parallel development and experimentation without affecting the main branch or the work of other team members. Branching provides a more organized and collaborative development environment, making it easier to merge changes and manage different development efforts. While Databricks Notebooks versioning also allows you to track versions of notebooks, it may not provide the same level of flexibility and collaboration as branching in Databricks Repos.

NEW QUESTION 10
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 12
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 16
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 19
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 24
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 27
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 31
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 32
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 36
.....

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