



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 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 \nformat("console") \ trigger(processingTime='2 seconds') \ start()\n<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#triggers>

NEW QUESTION 4

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 5

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 6

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 7

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 8

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.SessionTable.html>

NEW QUESTION 9

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 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 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 customers TO 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 10

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 15

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

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

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

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

SELECT
  store_id,
  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 18

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 22

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 25

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 29

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 32

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 33

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 38

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 42

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 45

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 49

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 52

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 54

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 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 61

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 62

.....

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](#)