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 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 3
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 4
A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.
Which of the following approaches can the data engineering team use to improve the latency of the team's queries?

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

Answer: A

Explanation:
 When many users are running small queries simultaneously on a SQL endpoint, the database can become overloaded, causing slow query execution times. By increasing the cluster size of the SQL endpoint, the database can handle more simultaneous queries, resulting in faster query execution times.

NEW QUESTION 5
Which of the following data lakehouse features results in improved data quality over a traditional data lake?

A. A data lakehouse provides storage solutions for structured and unstructured data.
B. A data lakehouse supports ACID-compliant transactions.
C. A data lakehouse allows the use of SQL queries to examine data.
D. A data lakehouse stores data in open formats.
E. A data lakehouse enables machine learning and artificial Intelligence workloads.

Answer: B

Explanation:
 One of the key features of a data lakehouse that results in improved data quality over a traditional data lake is its support for ACID (Atomicity, Consistency, Isolation, Durability) transactions. ACID transactions provide data integrity and consistency guarantees, ensuring that operations on the data are reliable and that data is not left in an inconsistent state due to failures or concurrent access. In a traditional data lake, such transactional guarantees are often lacking, making it challenging to maintain data quality,
especially in scenarios involving multiple data writes, updates, or complex transformations. A data lakehouse, by offering ACID compliance, helps maintain data quality by providing strong consistency and reliability, which is crucial for data pipelines and analytics.

NEW QUESTION 6
A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.
Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?

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

Answer: C

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

NEW QUESTION 7
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 8
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 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 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 10
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 11
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 14
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 19
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 20
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 24
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 28
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 31
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 Development 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 once and the pipeline will shut dow
B. The compute resources will be terminated.
C. All datasets will be updated at set intervals until the pipeline is shut dow
D. The compute resources will persist until the pipeline is shut down.
E. All datasets will be updated once and the pipeline will persist without any processin
F. The compute resources will persist but go unused.
G. All datasets will be updated once and the pipeline will shut dow
H. The compute resources will persist to allow for additional testing.
I. All datasets will be updated at set intervals until the pipeline is shut dow
J. The compute resources will persist to allow for additional testing.

Answer: E

Explanation:
You can optimize pipeline execution by switching between development and production modes. Use the Delta Live Tables Environment Toggle Icon buttons in the Pipelines UI to switch between these two modes. By default, pipelines run in development mode.
When you run your pipeline in development mode, the Delta Live Tables system does the following:
Reuses a cluster to avoid the overhead of restarts. By default, clusters run for two hours when development mode is enabled. You can change this with the pipelines.clusterShutdown.delay setting in the Configure your compute settings.
Disables pipeline retries so you can immediately detect and fix errors. In production mode, the Delta Live Tables system does the following:
Restarts the cluster for specific recoverable errors, including memory leaks and stale credentials.
Retries execution in the event of specific errors, for example, a failure to start a cluster. https://docs.databricks.com/en/delta-live-tables/updates.html#optimize-execution

NEW QUESTION 32
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 36
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 41
Which of the following is stored in the Databricks customer's cloud account?

A. Databricks web application
B. Cluster management metadata
C. Repos
D. Data

E. Notebooks

Answer: D

NEW QUESTION 45
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 48
In which of the following file formats is data from Delta Lake tables primarily stored?

A. Delta
B. CSV
C. Parquet
D. JSON
E. A proprietary, optimized format specific to Databricks

Answer: C

Explanation:
 https://docs.delta.io/latest/delta-faq.html

NEW QUESTION 50
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 54
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 55
......

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

100% Actual & Verified — Instant Download, Please Click
[Order The Databricks-Certified-Data-Engineer-Associate Practice Test Here](https://www.certshared.com/exam/Databricks-Certified-Data-Engineer-Associate/)