# Databricks

## Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

**NEW QUESTION 1**
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 2**
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 3**
A data engineer is attempting to drop a Spark SQL table my_table. The data engineer wants to delete all table metadata and data.
They run the following command: DROP TABLE IF EXISTS my_table
While the object no longer appears when they run SHOW TABLES, the data files still exist.
Which of the following describes why the data files still exist and the metadata files were deleted?

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

**Answer:** C

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

**NEW QUESTION 4**
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 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 can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

A. None of these
B. Data lake
C. Data warehouse
D. All of these
E. Data lakehouse

**Answer:** E

**NEW QUESTION 8**
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 9**
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 10**
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 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 15**
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 17**
A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.
In which of the following locations can the data engineer review their permissions on the table?

A. Databricks Filesystem
B. Jobs
C. Dashboards
D. Repos
E. Data Explorer

**Answer:** E

**NEW QUESTION 19**
A dataset has been defined using Delta Live Tables and includes an expectations clause:
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW
What is the expected behavior when a batch of data containing data that violates these constraints is processed?

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

**Answer:** C

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

**NEW QUESTION 22**
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 27**
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 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 30**
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 31**
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 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 36**
A Delta Live Table pipeline includes two datasets defined using STREAMING LIVE TABLE. Three datasets are defined against Delta Lake table sources using LIVE TABLE.
The table is configured to run in Production mode using the Continuous Pipeline Mode. Assuming previously unprocessed data exists and all definitions are valid, what is the
expected outcome after clicking Start to update the pipeline?

A. All datasets will be updated at set intervals until the pipeline is shut dow
B. The compute resources will persist to allow for additional testing.
C. All datasets will be updated once and the pipeline will persist without any processin
D. The compute resources will persist but go unused.
E. All datasets will be updated at set intervals until the pipeline is shut dow
F. The compute resources will be deployed for the update and terminated when the pipeline is stopped.
G. All datasets will be updated once and the pipeline will shut dow
H. The compute resources will be terminated.
I. All datasets will be updated once and the pipeline will shut dow
J. The compute resources will persist to allow for additional testing.

**Answer:** C

**Explanation:**
In a Delta Live Table pipeline running in Continuous Pipeline Mode, when you click Start to update the pipeline, the following outcome is expected: All datasets defined using STREAMING LIVE TABLE and LIVE TABLE against Delta Lake table sources will be updated at set intervals. The compute resources will be deployed for the update process and will be active during the execution of the pipeline. The compute resources will be terminated when the pipeline is stopped or shut down. This mode allows for continuous and periodic updates to the datasets as new data arrives or changes in the
underlying Delta Lake tables occur. The compute resources are provisioned and utilized during the update intervals to process the data and perform the necessary operations.

**NEW QUESTION 40**
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 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 47**
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 52**
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 53**
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%20dat

aset%20containing%20the%20new,new%20dataset%2C%20it%20is%20inserted.

**NEW QUESTION 55**
A data engineer is maintaining a data pipeline. Upon data ingestion, the data engineer notices that the source data is starting to have a lower level of quality. The data engineer would like to automate the process of monitoring the quality level.
Which of the following tools can the data engineer use to solve this problem?

A. Unity Catalog
B. Data Explorer
C. Delta Lake
D. Delta Live Tables
E. Auto Loader

**Answer:** D

**Explanation:**
https://docs.databricks.com/delta-live-tables/expectations.html
Delta Live Tables is a tool provided by Databricks that can help data engineers automate the monitoring of data quality. It is designed for managing data pipelines, monitoring data quality, and automating workflows. With Delta Live Tables, you can set up data quality checks and alerts to detect issues and anomalies in your data as it is ingested and processed in real-time. It provides a way to ensure that the data quality meets your desired standards and can trigger actions or notifications when issues are detected. While the other tools mentioned may have their own purposes in a data engineeringenvironment, Delta Live Tables is specifically designed for data quality monitoring and automation within the Databricks ecosystem.

**NEW QUESTION 60**
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 61**
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 63**
.....

# 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