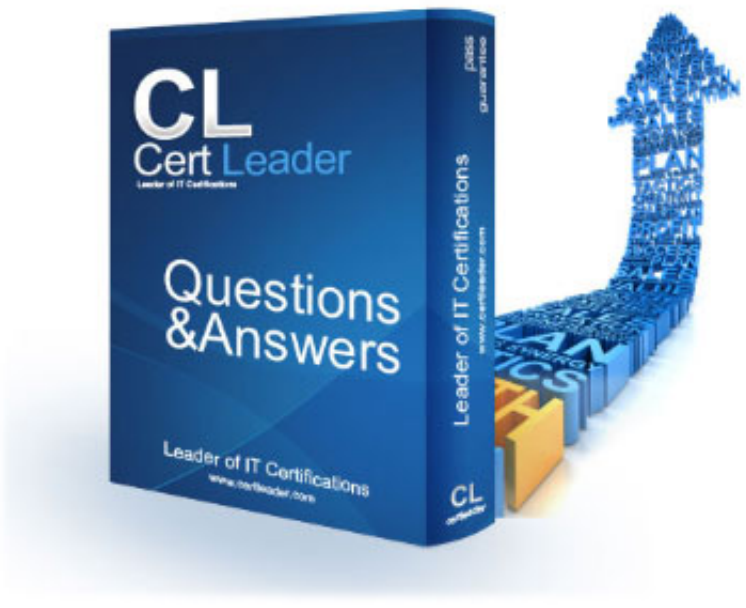
Databricks-Certified-Data-Engineer-Associate Dumps

Databricks Certified Data Engineer Associate Exam

https://www.certleader.com/Databricks-Certified-Data-Engineer-Associate-dumps.html



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
Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?

A. Manually programming in an alert system in each cell of the Notebook
B. Setting up an Alert in the Job page
C. Setting up an Alert in the Notebook
D. There is no way to notify the Job owner in the case of Job failure
E. MLflow Model Registry Webhooks

Answer: B

Explanation:
https://docs.databricks.com/en/workflows/jobs/job-notifications.html

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
Which of the following describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.
C. CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.
D. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
E. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

Answer: B

Explanation:
The CREATE STREAMING LIVE TABLE syntax is used when you want to create Delta Live Tables (DLT) tables that are designed for processing data incrementally. This is typically used when your data pipeline involves streaming or incremental data updates, and you want the table to stay up to date as new data arrives. It allows you to define tables that can handle data changes incrementally without the need for full table refreshes.

NEW QUESTION 6
A data engineer is attempting to drop a Spark SQL table my_table and runs the following command:
DROP TABLE IF EXISTS my_table;
After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.
Which of the following describes why all of these files were deleted?

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

Answer: A

Explanation:
managed tables files and metadata are managed by metastore and will be deleted when the table is dropped . while external tables the metadata is stored in a external location. hence when a external table is dropped you clear off only the metadata and the files (data) remain.

NEW QUESTION 7
A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

sales

customer_id	spend	units
a1	28.94	7
a3	874.12	23
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

The data engineer runs the following query to join these tables together:

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

Which of the following will be returned by the above query?

A.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2

B.

customer_id	spend	units	store_id
a1	28.94	7	s1
a4	8.99	1	s2

C.

customer_id	spend	store_id
a1	28.94	s1
a3	874.12	NULL
a4	8.99	s2

D.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a3	874.12	NULL
a4	8.99	s2

E.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a4	8.99	s2

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: C

NEW QUESTION 8
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 9
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 10
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 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 12
Which of the following is hosted completely in the control plane of the classic Databricks architecture?

A. Worker node
B. JDBC data source
C. Databricks web application
D. Databricks Filesystem
E. Driver node

Answer: C

Explanation:
In the classic Databricks architecture, the control plane includes components like the Databricks web application, the Databricks REST API, and the Databricks Workspace. These components are responsible for managing and controlling the Databricks environment, including cluster provisioning, notebook management, access control, and job scheduling. The other options, such as worker nodes, JDBC data sources, Databricks Filesystem (DBFS), and driver nodes, are typically part of the data plane or the execution environment, which is separate from the control plane. Worker nodes are responsible for executing tasks and computations, JDBC data sources are used to connect to external databases, DBFS is a distributed file system for data storage, and driver nodes are responsible for coordinating the execution of Spark jobs.

NEW QUESTION 14
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?

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

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

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

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

E.
```
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 19
Which of the following describes the relationship between Gold tables and Silver tables?

A. Gold tables are more likely to contain aggregations than Silver tables.
B. Gold tables are more likely to contain valuable data than Silver tables.
C. Gold tables are more likely to contain a less refined view of data than Silver tables.
D. Gold tables are more likely to contain more data than Silver tables.
E. Gold tables are more likely to contain truthful data than Silver tables.

Answer: A

Explanation:
In some data processing pipelines, especially those following a typical "Bronze-Silver-Gold" data lakehouse architecture, Silver tables are often considered a more refined version of the raw or Bronze data. Silver tables may include data cleansing, schema enforcement, and some initial transformations. Gold tables, on the other hand, typically represent a stage where data is further enriched, aggregated, and processed to provide valuable insights for analytical purposes. This could indeed involve more aggregations compared to Silver tables.

NEW QUESTION 24
Which of the following commands can be used to write data into a Delta table while avoiding the writing of duplicate records?

A. DROP
B. IGNORE
C. MERGE
D. APPEND
E. INSERT

Answer: C

Explanation:
To write data into a Delta table while avoiding the writing of duplicate records, you can use the MERGE command. The MERGE command in Delta Lake allows you to combine the ability to insert new records and update existing records in a single atomic operation. The MERGE command compares the data being written with the existing data in the Delta table based on specified matching criteria, typically using a primary key or unique identifier. It then performs conditional actions, such as inserting new records or updating existing records, depending on the comparison results. By using the MERGE command, you can handle the prevention of duplicate records in a more controlled and efficient manner. It allows you to synchronize and reconcile data from different sources while avoiding duplication and ensuring data integrity.

NEW QUESTION 28
Which of the following Git operations must be performed outside of Databricks Repos?

A. Commit
B. Pull
C. Push
D. Clone
E. Merge

Answer: E

Explanation:
For following tasks, work in your Git provider:
Create a pull request. Resolve merge conflicts. Merge or delete branches. Rebase a branch.
https://docs.databricks.com/repos/index.html

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

A. They can set up an Alert with a custom template.
B. They can set up an Alert with a new email alert destination.
C. They can set up an Alert with a new webhook alert destination.
D. They can set up an Alert with one-time notifications.
E. They can set up an Alert without notifications.

Answer: C

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

NEW QUESTION 36
A 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 37
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 39
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 44
A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.
Which of the following describes why Auto Loader inferred all of the columns to be of the
string type?

A. There was a type mismatch between the specific schema and the inferred schema
B. JSON data is a text-based format
C. Auto Loader only works with string data
D. All of the fields had at least one null value
E. Auto Loader cannot infer the schema of ingested data

Answer: B

Explanation:
JSON data is a text-based format that uses strings to represent all values. When Auto Loader infers the schema of JSON data, it assumes that all values are strings. This is because Auto Loader cannot determine the type of a value based on its string representation. https://docs.databricks.com/en/ingestion/auto-loader/schema.html Forexample, the following JSON string represents a value that is logically a boolean: JSON "true" Use code with caution. Learn more However, Auto Loader would infer that the type of this value is string. This is because Auto Loader cannot determine that the value is a boolean based on its string representation. In order to get Auto Loader to infer the correct types for columns, the data engineer can provide type inference or schema hints. Type inference hints can be used to specify the types of specific columns. Schema hints can be used to provide the entire schema of the data. Therefore, the correct answer is B. JSON data is a text-based format.

NEW QUESTION 47
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 49
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 51
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 56
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 60
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 61
A data architect has determined that a table of the following format is necessary:

employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
...	...	...

Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above format regardless of whether a table already exists with this name?

A.
```
CREATE TABLE IF NOT EXISTS table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

B.
```
CREATE OR REPLACE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
USING DELTA
```

C.
```
CREATE OR REPLACE TABLE table_name WITH COLUMNS (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
) USING DELTA
```

D.
```
CREATE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
```

E.
```
CREATE OR REPLACE TABLE table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: E

NEW QUESTION 62
.....

Thank You for Trying Our Product

* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

* One year free update

You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

We currently serve more than 30,000,000 customers.

* Shop Securely

All transactions are protected by VeriSign!

100% Pass Your Databricks-Certified-Data-Engineer-Associate Exam with Our Prep Materials Via below:

https://www.certleader.com/Databricks-Certified-Data-Engineer-Associate-dumps.html