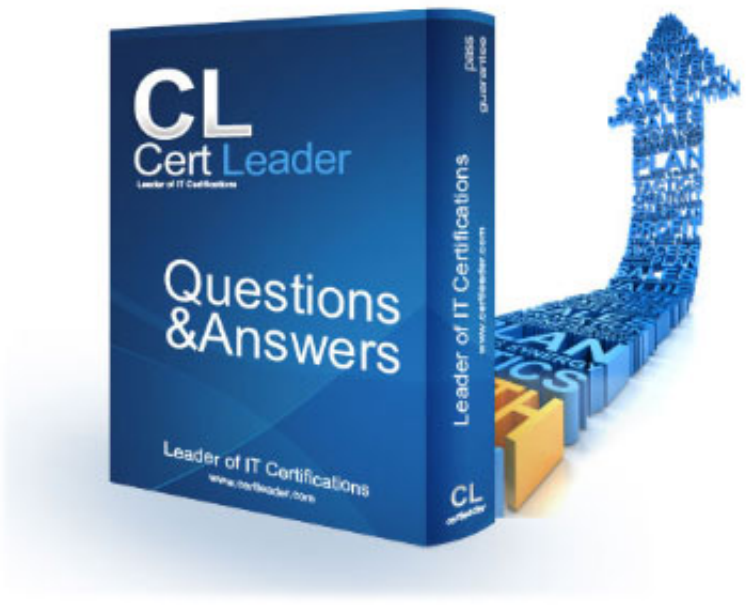
Databricks-Certified-Data-Engineer-Associate Dumps

Databricks Certified Data Engineer Associate Exam

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



NEW QUESTION 1
Which of the following commands will return the location of database customer360?

A. DESCRIBE LOCATION customer360;
B. DROP DATABASE customer360;
C. DESCRIBE DATABASE customer360;
D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user'};
E. USE DATABASE customer360;

Answer: C

Explanation:
To retrieve the location of a database named "customer360" in a database management system like Hive or Databricks, you can use the DESCRIBE DATABASE command followed by the database name. This command will provide information about the database, including its location.

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 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
Which of the following must be specified when creating a new Delta Live Tables pipeline?

A. A key-value pair configuration
B. The preferred DBU/hour cost
C. A path to cloud storage location for the written data
D. A location of a target database for the written data
E. At least one notebook library to be executed

Answer: E

Explanation:
https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html

NEW QUESTION 6
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 7
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 8
A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.
Which of the following commands can be used to grant the necessary permission on the entire database to the new team?

A. GRANT VIEW ON CATALOG customers TO team;
B. GRANT CREATE ON DATABASE customers TO team;
C. GRANT USAGE ON CATALOG team TO customers;
D. GRANT CREATE ON DATABASE team TO customers;
E. GRANT USAGE ON DATABASE customers TO team;

Answer: E

Explanation:
The GRANT statement is used to grant privileges on a database, table, or view to a user or role. The ALL PRIVILEGES option grants all possible privileges on the specified object, such as CREATE, SELECT, MODIFY, and USAGE. The syntax of the GRANT statement is:
GRANT privilege_type ON object TO user_or_role;
Therefore, to grant full permissions on the database customers to the new data engineering team, the command should be:
GRANT ALL PRIVILEGES ON DATABASE customers TO team;

NEW QUESTION 9
A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:

```
transactions_df = (spark.read
      .schema(schema)
      .format("delta")
      .table("transactions")
)
```

Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

A. Replace predict with a stream-friendly prediction function
B. Replace schema(schema) with option ("maxFilesPerTrigger", 1)
C. Replace "transactions" with the path to the location of the Delta table
D. Replace format("delta") with format("stream")
E. Replace spark.read with spark.readStream

Answer: E

Explanation:
https://docs.databricks.com/en/structured-streaming/delta-lake.html

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 10
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 13
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 14
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 17
Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?

A. When they are working interactively with a small amount of data
B. When they are running automated reports to be refreshed as quickly as possible
C. When they are working with SQL within Databricks SQL
D. When they are concerned about the ability to automatically scale with larger data
E. When they are manually running reports with a large amount of data

Answer: A

Explanation:
A Single Node cluster is a cluster consisting of an Apache Spark driver and no Spark workers. A Single Node cluster supports Spark jobs and all Spark data sources, including Delta Lake. A Standard cluster requires a minimum of one Spark worker to run Spark jobs.

NEW QUESTION 18
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 19
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 21
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 23
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 27
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
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 36
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 39
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 42
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 44
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 46
Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

A. Parquet files can be partitioned
B. CREATE TABLE AS SELECT statements cannot be used on files
C. Parquet files have a well-defined schema
D. Parquet files have the ability to be optimized
E. Parquet files will become Delta tables

Answer: C

Explanation:
https://www.databricks.com/glossary/what-is- parquet#:~:text=Columnar%20storage%20like%20Apache%20Parquet,compared%20to%20row%2Doriented%20databases. Columnar storage like Apache Parquet is designed to bring efficiency compared to row-based files like CSV. When querying, columnar storage you can skip over the non-relevant data very quickly. As a result, aggregation queries are less time-consuming compared to row-oriented databases.

NEW QUESTION 51
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
A data engineer has joined an existing project and they see the following query in the project repository:
CREATE STREAMING LIVE TABLE loyal_customers AS SELECT customer_id -
FROM STREAM(LIVE.customers) WHERE loyalty_level = 'high';
Which of the following describes why the STREAM function is included in the query?

A. The STREAM function is not needed and will cause an error.
B. The table being created is a live table.
C. The customers table is a streaming live table.
D. The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.
E. The data in the customers table has been updated since its last run.

Answer: C

Explanation:
https://docs.databricks.com/en/sql/load-data-streaming-table.html Load data into a streaming table
To create a streaming table from data in cloud object storage, paste the following into the query editor, and then click Run:
SQL
Copy to clipboardCopy
/* Load data from a volume */
CREATE OR REFRESH STREAMING TABLE <table-name> AS SELECT * FROM STREAM
read_files('/Volumes/<catalog>/<schema>/<volume>/<path>/<folder>')
/* Load data from an external location */
CREATE OR REFRESH STREAMING TABLE <table-name> AS
SELECT * FROM STREAM read_files('s3://<bucket>/<path>/<folder>')

NEW QUESTION 59
.....

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