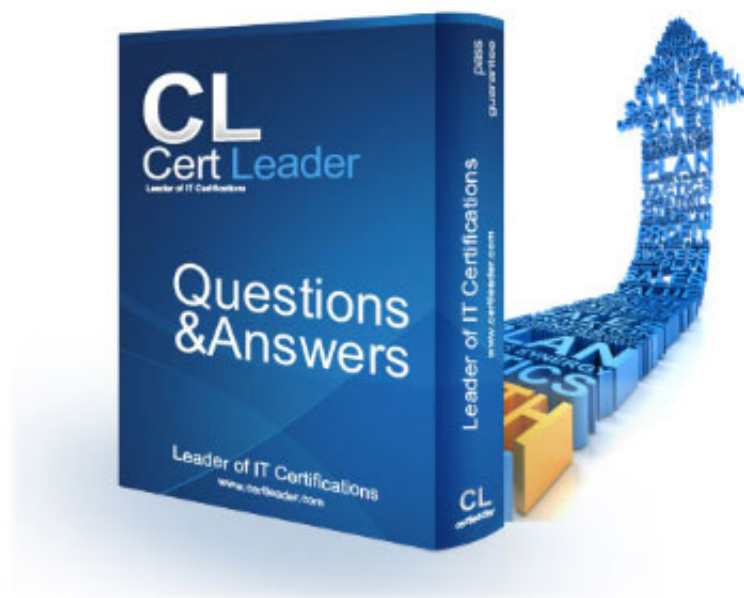


## Databricks-Certified-Data-Engineer-Associate Dumps

### Databricks Certified Data Engineer Associate Exam

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



**NEW QUESTION 1**

Which of the following 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 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 code block used by the data engineer is below:

```
(spark.table("sales")
  .withColumn("avg_price", col("sales") / col("units"))
  .writeStream
  .option("checkpointLocation", checkpointPath)
  .outputMode("complete")
  ._____
  .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**

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 6**

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 7**

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 8**

A data engineer only wants to execute the final block of a Python program if the Python variable day\_of\_week is equal to 1 and the Python variable review\_period is True.

Which of the following control flow statements should the data engineer use to begin this conditionally executed code block?

- A. if day\_of\_week = 1 and review\_period:
- B. if day\_of\_week = 1 and review\_period = "True":
- C. if day\_of\_week == 1 and review\_period == "True":
- D. if day\_of\_week == 1 and review\_period:
- E. if day\_of\_week = 1 & review\_period: = "True":

**Answer:** D

**Explanation:**

This statement will check if the variable `day_of_week` is equal to 1 and if the variable `review_period` evaluates to a truthy value. The use of the double equal sign (`==`) in the comparison of `day_of_week` is important, as a single equal sign (`=`) would be used to assign a value to the variable instead of checking its value. The use of a single ampersand (`&`) instead of the keyword `and` is not valid syntax in Python. The use of quotes around `True` in options B and C will result in a string comparison, which will not evaluate to `True` even if the value of `review_period` is `True`.

**NEW QUESTION 9**

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 10**

A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.

Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with \$0 in sales is greater than zero?

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

**Answer:** D

**NEW QUESTION 10**

A 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 12**

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 15**

A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column `employees` in table `stores`. The custom logic should create a new column `exp_employees` that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this



custom logic at scale, the data engineer wants to use the FILTER higher-order function.  
Which of the following code blocks successfully completes this task?

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

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

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

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

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

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

**Answer:** A

#### NEW QUESTION 16

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 19

A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.

Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

- A. Databricks Repos automatically saves development progress
- B. Databricks Repos supports the use of multiple branches
- C. Databricks Repos allows users to revert to previous versions of a notebook
- D. Databricks Repos provides the ability to comment on specific changes
- E. Databricks Repos is wholly housed within the Databricks Lakehouse Platform

**Answer:** B

#### Explanation:

An advantage of using Databricks Repos over the built-in Databricks Notebooks versioning is the ability to work with multiple branches. Branching is a fundamental feature of version control systems like Git, which Databricks Repos is built upon. It allows you to create separate branches for different tasks, features, or experiments within your project. This separation helps in parallel development and experimentation without affecting the main branch or the work of other team members. Branching provides a more organized and collaborative development environment, making it easier to merge changes and manage different development efforts. While Databricks Notebooks versioning also allows you to track versions of notebooks, it may not provide the same level of flexibility and collaboration as branching in Databricks Repos.

#### NEW QUESTION 23

A data engineer needs to apply custom logic to string column city in table stores for a specific use case. In order to apply this custom logic at scale, the data

engineer wants to create a SQL user-defined function (UDF).  
Which of the following code blocks creates this SQL UDF?

A.

```
CREATE FUNCTION combine_nyc(city STRING)
RETURNS STRING
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

B.

```
CREATE UDF combine_nyc(city STRING)
RETURNS STRING
CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

C.

```
CREATE UDF combine_nyc(city STRING)
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

D.

```
CREATE FUNCTION combine_nyc(city STRING)
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

E.

```
CREATE UDF combine_nyc(city STRING)
RETURNS STRING
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

A.

**Answer:** A

**Explanation:**

<https://www.databricks.com/blog/2021/10/20/introducing-sql-user-defined-functions.html>

#### NEW QUESTION 28

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 31

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 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 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 down
- 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 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 40**

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 41

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 45

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 49

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 53

A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table.

The code block used by the data engineer is below:



```
(spark.readStream
  .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 process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

- A. processingTime(1)
- B. trigger(availableNow=True)
- C. trigger(parallelBatch=True)
- D. trigger(processingTime="once")
- E. trigger(continuous="once")

**Answer:** B

**Explanation:**

<https://stackoverflow.com/questions/71061809/trigger-availablenow-for-delta-source-streaming-queries-in-pyspark-databricks>

**NEW QUESTION 58**

.....

## 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>