



## **Databricks**

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

Databricks Certified Data Engineer Associate Exam

## About ExamBible

### *Your Partner of IT Exam*

## Found in 1998

ExamBible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, ExamBible has its unique advantages that other companies could not achieve.

## Our Advances

### \* 99.9% Uptime

All examinations will be up to date.

### \* 24/7 Quality Support

We will provide service round the clock.

### \* 100% Pass Rate

Our guarantee that you will pass the exam.

### \* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

### 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 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 3

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 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 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 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?

	customer_id	spend	store_id
A.	a1	28.94	s1
	a4	8.99	s2

  

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

  

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

  

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

  

	customer_id	spend	store_id
E.	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

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

A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL. Which of the following commands could the data engineering team use to access sales in PySpark?

- A. SELECT \* FROM sales
- B. There is no way to share data between PySpark and SQL.
- C. spark.sql("sales")
- D. spark.delta.table("sales")
- E. spark.table("sales")

**Answer: E**

#### Explanation:

<https://spark.apache.org/docs/3.2.1/api/python/reference/api/pyspark.sql.SparkSession.table.html>

#### NEW QUESTION 10

Which of the following describes the relationship between Bronze tables and raw data?

- A. Bronze tables contain less data than raw data files.
- B. Bronze tables contain more truthful data than raw data.
- C. Bronze tables contain aggregates while raw data is unaggregated.
- D. Bronze tables contain a less refined view of data than raw data.

E. Bronze tables contain raw data with a schema applied.

**Answer:** E

**Explanation:**

The Bronze layer is where we land all the data from external source systems. The table structures in this layer correspond to the source system table structures "as-is," along with any additional metadata columns that capture the load date/time, process ID, etc. The focus in this layer is quick Change Data Capture and the ability to provide an historical archive of source (cold storage), data lineage, auditability, reprocessing if needed without rereading the data from the source system.  
<https://www.databricks.com/glossary/medallion-architecture#:~:text=Bronze%20layer%20%28raw%20data%29>

**NEW QUESTION 10**

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 15**

In order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing, which of the following two approaches is used by Spark to record the offset range of the data being processed in each trigger?

- A. Checkpointing and Write-ahead Logs
- B. Structured Streaming cannot record the offset range of the data being processed in each trigger.
- C. Replayable Sources and Idempotent Sinks
- D. Write-ahead Logs and Idempotent Sinks
- E. Checkpointing and Idempotent Sinks

**Answer:** A

**Explanation:**

The engine uses checkpointing and write-ahead logs to record the offset range of the data being processed in each trigger. -- in the link search for "The engine uses " you'll find the answer.  
<https://spark.apache.org/docs/latest/structured-streaming-programming-guide.html#:~:text=The%20engine%20uses%20checkpointing%20and,being%20processe d%20in%20each%20trigger.>

**NEW QUESTION 20**

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 22**

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 26**

Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

- A. The ability to manipulate the same data using a variety of languages
- B. The ability to collaborate in real time on a single notebook
- C. The ability to set up alerts for query failures
- D. The ability to support batch and streaming workloads
- E. The ability to distribute complex data operations

**Answer:** D

**Explanation:**

Delta Lake is a key component of the Databricks Lakehouse Platform that provides several benefits, and one of the most significant benefits is its ability to support both batch and streaming workloads seamlessly. Delta Lake allows you to process and analyze data in real-time (streaming) as well as in batch, making it a versatile choice for various data processing needs. While the other options may be benefits or capabilities of Databricks or the Lakehouse Platform in general, they are not specifically associated with Delta Lake.

**NEW QUESTION 31**

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 33**

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 34

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 35

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 37

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 39

A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.

They run the following command:

```
CREATE TABLE new_table  
  
_____  
OPTIONS (  
  header = "true",  
  delimiter = "|" )  
LOCATION "path/to/csv"
```

Which of the following lines of code fills in the above blank to successfully complete the task?

- A. None of these lines of code are needed to successfully complete the task
- B. USING CSV
- C. FROM CSV
- D. USING DELTA
- E. FROM "path/to/csv"

**Answer:** B

#### NEW QUESTION 40

A data engineer has realized that they made a mistake when making a daily update to a table. They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted.

Which of the following explains why the data files are no longer present?

- A. The VACUUM command was run on the table
- B. The TIME TRAVEL command was run on the table
- C. The DELETE HISTORY command was run on the table
- D. The OPTIMIZE command was run on the table
- E. The HISTORY command was run on the table

**Answer:** A

**Explanation:**

The VACUUM command in Delta Lake is used to clean up and remove unnecessary data files that are no longer needed for time travel or query purposes. When you run VACUUM with certain retention settings, it can delete older data files, which might include versions of data that are older than the specified retention period. If the data engineer is unable to restore the table to a version that is 3 days old because the data files have been deleted, it's likely because the VACUUM command was run on the table, removing the older data files as part of data cleanup.

**NEW QUESTION 45**

Which of the following describes a scenario in which a data team will want to utilize cluster pools?

- A. An automated report needs to be refreshed as quickly as possible.
- B. An automated report needs to be made reproducible.
- C. An automated report needs to be tested to identify errors.
- D. An automated report needs to be version-controlled across multiple collaborators.
- E. An automated report needs to be runnable by all stakeholders.

**Answer:** A

**Explanation:**

Cluster pools are typically used in distributed computing environments, such as cloud-based data platforms like Databricks. They allow you to pre-allocate a set of compute resources (a cluster) for specific tasks or workloads. In this case, if an automated report needs to be refreshed as quickly as possible, you can allocate a cluster pool with sufficient resources to ensure fast data processing and report generation. This helps ensure that the report is generated with minimal latency and can be delivered to stakeholders in a timely manner. Cluster pools allow you to optimize resource allocation for high-demand, time-sensitive tasks like real-time report generation.

**NEW QUESTION 46**

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 49**

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 53**

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 down
- 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 processing
- D. The compute resources will persist but go unused.
- E. All datasets will be updated at set intervals until the pipeline is shut down

- 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 down.
- H. The compute resources will be terminated.
- I. All datasets will be updated once and the pipeline will shut down.
- 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 58**

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 60**

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 63**

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 67**

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 down
- 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 processing
- F. The compute resources will persist but go unused.
- G. All datasets will be updated once and the pipeline will shut down

- 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 down.
- 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 69**

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](https://www.w3schools.com/python/python_functions.asp)

**NEW QUESTION 70**

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.>

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.

**NEW QUESTION 74**

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 engineering environment, Delta Live Tables is specifically designed for data quality monitoring and automation within the Databricks ecosystem.

**NEW QUESTION 78**

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 80**

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 84**

A data engineer needs access to a table `new_table`, but they do not have the correct permissions. They can ask the table owner for permission, but they do not know who the table owner is.

Which of the following approaches can be used to identify the owner of `new_table`?

- A. Review the Permissions tab in the table's page in Data Explorer
- B. All of these options can be used to identify the owner of the table
- C. Review the Owner field in the table's page in Data Explorer
- D. Review the Owner field in the table's page in the cloud storage solution
- E. There is no way to identify the owner of the table

**Answer: C**

**NEW QUESTION 85**

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 90**

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 92**

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?

```
CREATE TABLE IF NOT EXISTS table_name (  
    employeeId STRING,  
A.  startDate DATE,  
    avgRating FLOAT  
)  
  
CREATE OR REPLACE TABLE table_name AS  
SELECT  
B.  employeeId STRING,  
    startDate DATE,  
    avgRating FLOAT  
USING DELTA  
  
CREATE OR REPLACE TABLE table_name WITH COLUMNS (  
    employeeId STRING,  
C.  startDate DATE,  
    avgRating FLOAT  
) USING DELTA  
  
CREATE TABLE table_name AS  
SELECT  
D.  employeeId STRING,  
    startDate DATE,  
    avgRating FLOAT  
  
CREATE OR REPLACE TABLE table_name (  
    employeeId STRING,  
E.  startDate DATE,  
    avgRating FLOAT  
)
```

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

**Answer:** E

#### NEW QUESTION 93

.....

## Relate Links

**100% Pass Your Databricks-Certified-Data-Engineer-Associate Exam with Exambible Prep Materials**

<https://www.exambible.com/Databricks-Certified-Data-Engineer-Associate-exam/>

## Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - <https://www.exambible.com/>