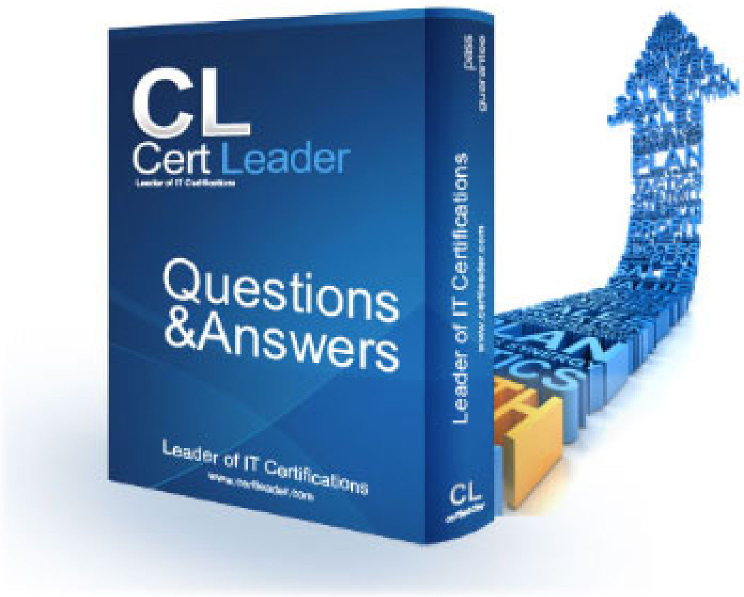
Databricks-Certified-Data-Engineer-Associate Dumps

Databricks Certified Data Engineer Associate Exam

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



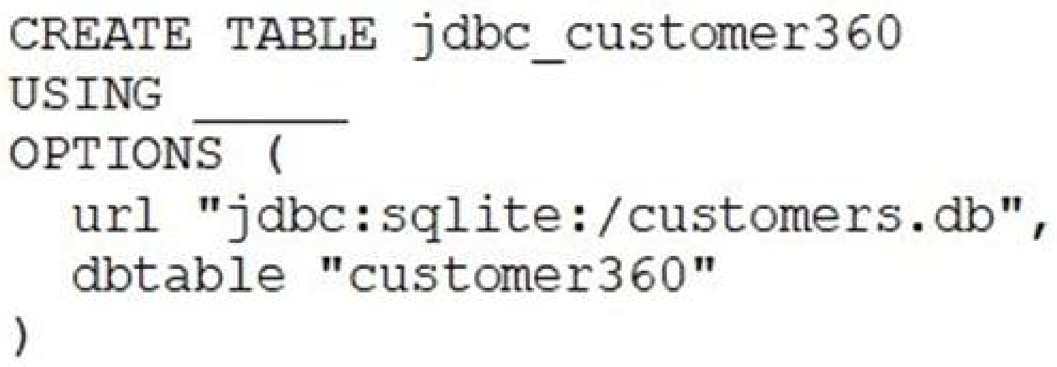
NEW QUESTION 1
A data engineer has created a new database using the following command: CREATE DATABASE IF NOT EXISTS customer360;
In which of the following locations will the customer360 database be located?

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
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

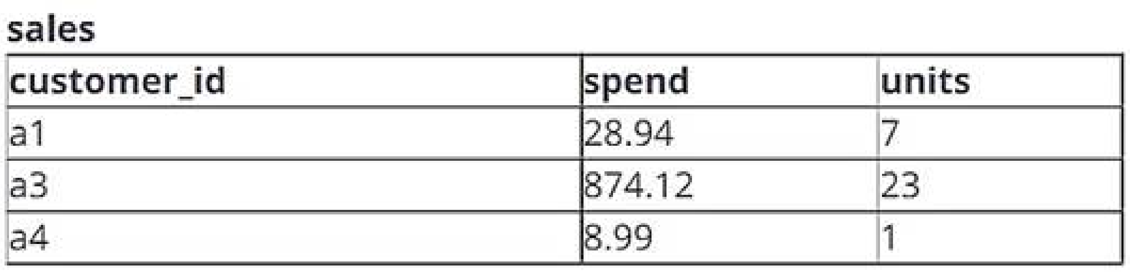
NEW QUESTION 3
Which of the following must be specified when creating a new Delta Live Tables pipeline?

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

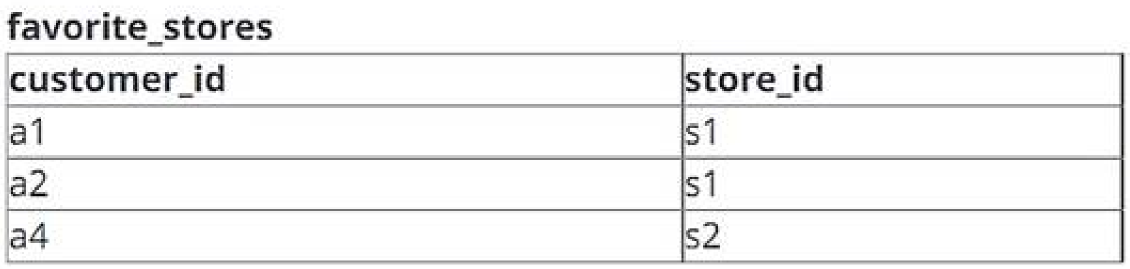
Explanation:
https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html

NEW QUESTION 4
A data engineer is working with two tables. Each of these tables is displayed below in its entirety.



sales

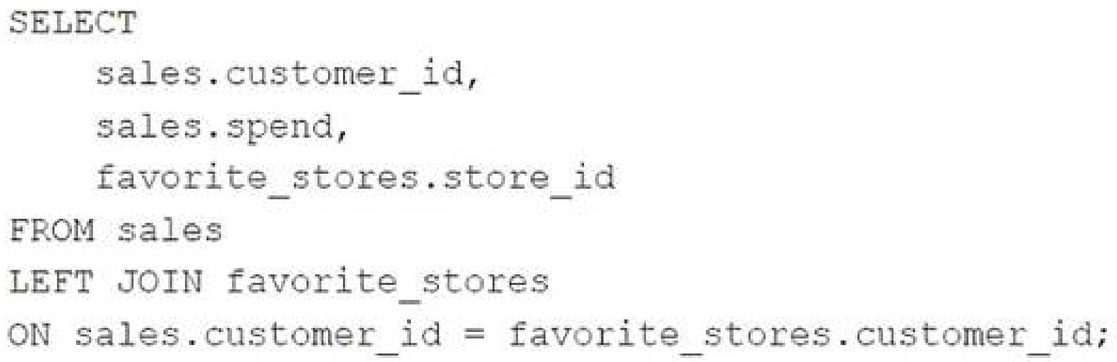
customer_id	spend	units
a1	28.94	7
a3	874.12	23
a4	8.99	1



favorite_stores

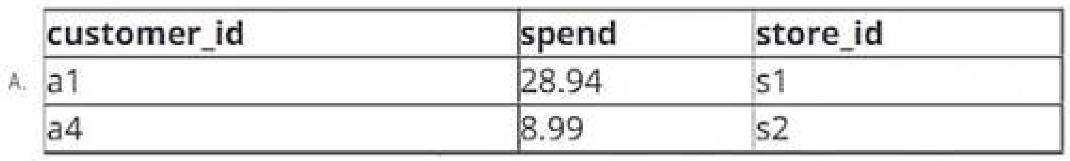
customer_id	store_id
a1	s1
a2	s1
a4	s2

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



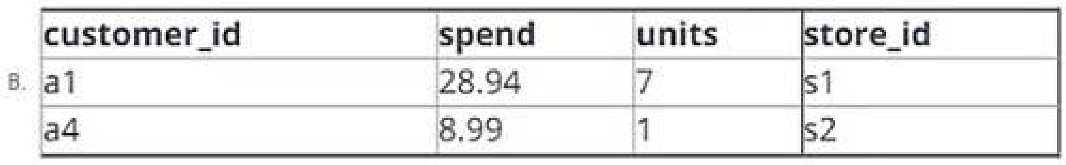
```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

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



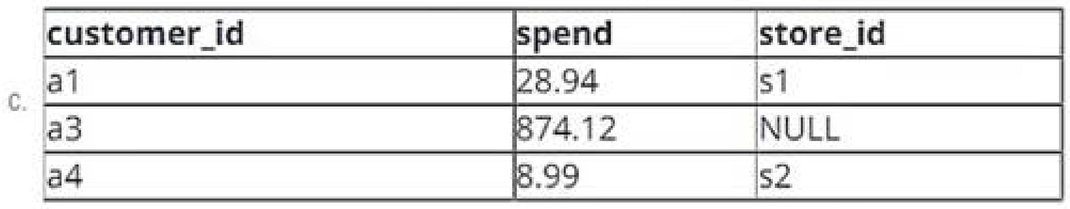
A.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2



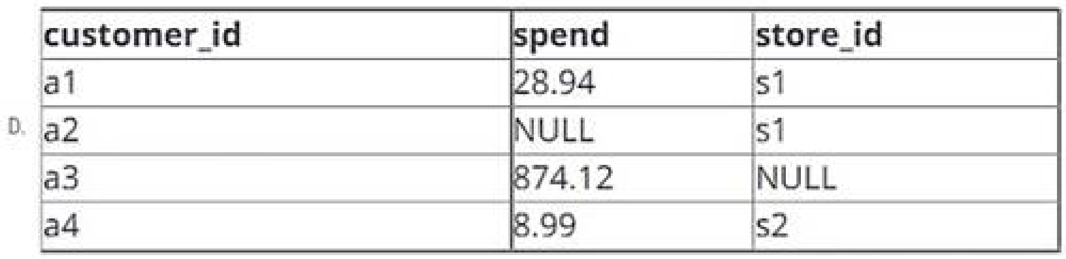
B.

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



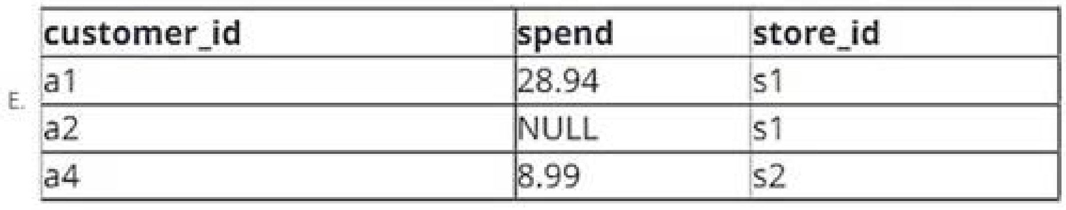
C.

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 5
Which of the following tools is used by Auto Loader process data incrementally?

A. Checkpointing
B. Spark Structured Streaming
C. Data Explorer
D. Unity Catalog
E. Databricks SQL

Answer: B

Explanation:
The Auto Loader process in Databricks is typically used in conjunction with Spark Structured Streaming to process data incrementally. Spark Structured Streaming is a real-time data processing framework that allows you to process data streams incrementally as new data arrives. The Auto Loader is a feature in Databricks that works with Structured Streaming to automatically detect and process new data files as they are added to a specified data source location. It allows for incremental data processing without the need for manual intervention.
How does Auto Loader track ingestion progress? As files are discovered, their metadata is persisted in a scalable key-value store (RocksDB) in the checkpoint location of your Auto Loader pipeline. This key-value store ensures that data is processed exactly once. In case of failures, Auto Loader can resume from where it left off by information stored in the checkpoint location and continue to provide exactly-once guarantees when writing data into Delta Lake. You don't need to maintain or manage any state yourself to achieve fault tolerance or exactly-once semantics.https://docs.databricks.com/ingestion/auto- loader/index.html

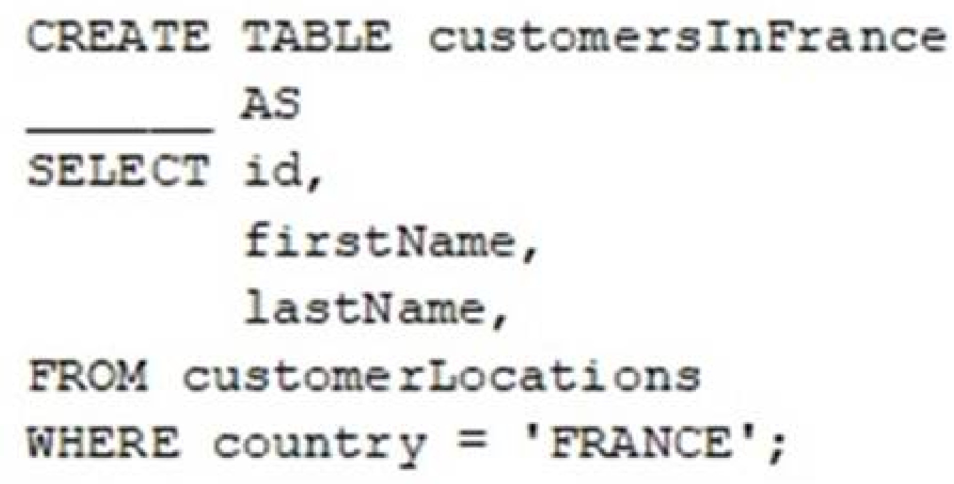
NEW QUESTION 6
Which of the following describes the relationship between Bronze tables and raw data?

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 7
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 8
A data analyst has a series of queries in a SQL program. The data analyst wants this program to run every day. They only want the final query in the program to run on Sundays. They ask for help from the data engineering team to complete this task.
Which of the following approaches could be used by the data engineering team to complete this task?

A. They could submit a feature request with Databricks to add this functionality.
B. They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.
C. They could only run the entire program on Sundays.
D. They could automatically restrict access to the source table in the final query so that it is only accessible on Sundays.
E. They could redesign the data model to separate the data used in the final query into a new table.

Answer: B

NEW QUESTION 9
A dataset has been defined using Delta Live Tables and includes an expectations clause:
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW
What is the expected behavior when a batch of data containing data that violates these constraints is processed?

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

Answer: C

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

NEW QUESTION 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 10
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 13
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.

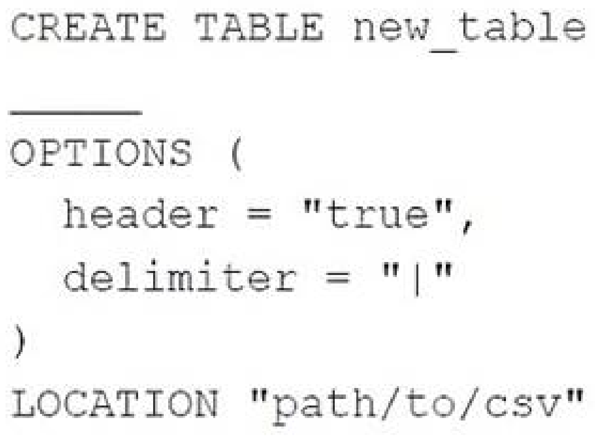
NEW QUESTION 17
Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?

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 22
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 27
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 30
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?

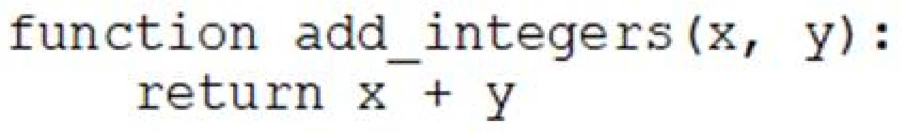
A. All datasets will be updated once and the pipeline will shut dow
B. The compute resources will be terminated.
C. All datasets will be updated at set intervals until the pipeline is shut dow
D. The compute resources will persist until the pipeline is shut down.
E. All datasets will be updated once 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.

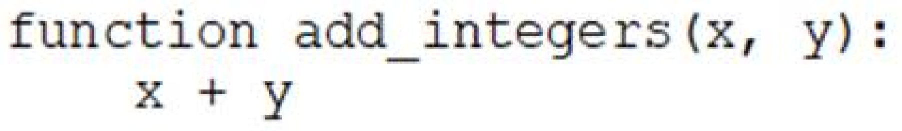
Disables pipeline retries so you can immediately detect and fix errors. In production mode, the Delta Live Tables system does the following:
Restarts the cluster for specific recoverable errors, including memory leaks and stale credentials.
Retries execution in the event of specific errors, for example, a failure to start a cluster. https://docs.databricks.com/en/delta-live-tables/updates.html#optimize-execution

NEW QUESTION 34
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)



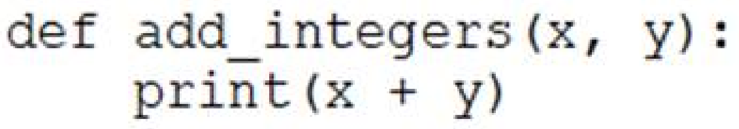
```
function add_integers(x, y):
    return x + y
```

B)



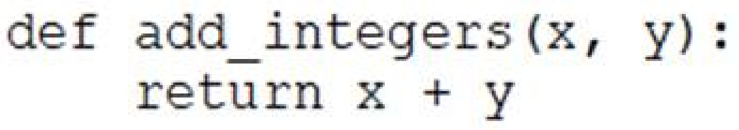
```
function add_integers(x, y):
    x + y
```

C)



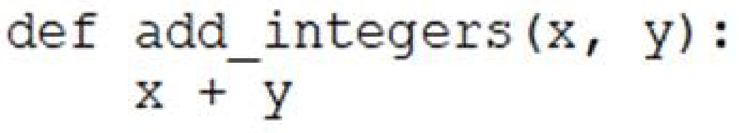
```
def add_integers(x, y):
    print(x + y)
```

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 39
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 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 46
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 48
A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start.
Which of the following actions can the data engineer perform to improve the start up time for the clusters used for the Job?

A. They can use endpoints available in Databricks SQL
B. They can use jobs clusters instead of all-purpose clusters
C. They can configure the clusters to be single-node
D. They can use clusters that are from a cluster pool
E. They can configure the clusters to autoscale for larger data sizes

Answer: D

Explanation:
Cluster pools are a way to pre-provision clusters that are ready to use. This can reduce the start up time for clusters, as they do not have to be created from scratch. All-purpose clusters are not pre-provisioned, so they will take longer to start up. Jobs clusters are a type of cluster pool, but they are not the best option for this use case. Jobs clusters are designed for long-running jobs, and they can be more expensive than other types of cluster pools. Single-node clusters are the smallest type of cluster, and they will start up the fastest. However, they may not be powerful enough to run the Job's tasks. Autoscaling clusters can scale up or down based on demand. This can help to improve the start up time for clusters, as they will only be created when they are needed. However, autoscaling clusters can also be more expensive than other types of cluster pool https://docs.databricks.com/en/clusters/pool-best-practices.html

NEW QUESTION 53
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%2
0row%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 57
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 61
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 65
......