Databricks

Exam Questions Databricks-Certified-Data-Engineer-Associate

Databricks Certified Data Engineer Associate Exam

NEW QUESTION 1
In which of the following scenarios should a data engineer select a Task in the Depends On field of a new Databricks Job Task?

A. When another task needs to be replaced by the new task
B. When another task needs to fail before the new task begins
C. When another task has the same dependency libraries as the new task
D. When another task needs to use as little compute resources as possible
E. When another task needs to successfully complete before the new task begins

Answer: E

NEW QUESTION 2
Which of the following commands will return the location of database customer360?

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

Answer: C

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

NEW QUESTION 3
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 4
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 5
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 6
A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.
Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?

A. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."
B. They can turn on the Auto Stop feature for the SQL endpoint.
C. They can increase the cluster size of the SQL endpoint.
D. They can turn on the Serverless feature for the SQL endpoint.
E. They can increase the maximum bound of the SQL endpoint's scaling range

Answer: C

Explanation:
https://www.databricks.com/blog/2022/03/10/top-5-databricks-performance- tips.html

NEW QUESTION 7
Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

A. None of these
B. Data lake
C. Data warehouse
D. All of these
E. Data lakehouse

Answer: E

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
Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

A. Cloud-specific integrations
B. Simplified governance
C. Ability to scale storage
D. Ability to scale workloads
E. Avoiding vendor lock-in

Answer: E

Explanation:
https://double.cloud/blog/posts/2023/01/break-free-from-vendor-lock-in-with-open-source-tech/

NEW QUESTION 10
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.tabl e.html

NEW QUESTION 10
A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.
Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

A. GRANT ALL PRIVILEGES ON TABLE sales TO team;
B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
C. GRANT SELECT ON TABLE sales TO team;
D. GRANT USAGE ON TABLE sales TO team;
E. GRANT ALL PRIVILEGES ON TABLE team TO sales;

Answer: A

NEW QUESTION 14
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 19
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 23
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 " youll 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 28
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 31
A new data engineering team team. has been assigned to an ELT project. The new data engineering team will need full privileges on the database customers to fully manage the project.
Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

A. GRANT USAGE ON DATABASE customers TO team;
B. GRANT ALL PRIVILEGES ON DATABASE team TO customers;
C. GRANT SELECT PRIVILEGES ON DATABASE customers TO teams;
D. GRANT SELECT CREATE MODIFY USAGE PRIVILEGES ON DATABASE customersTO team;
E. GRANT ALL PRIVILEGES ON DATABASE customers TO team;

Answer: E

Explanation:
To grant full privileges on the database "customers" to the new data engineering team, you can use the GRANT ALL PRIVILEGES command as shown in option E. This command provides the team with all possible privileges on the specified database, allowing them to fully manage it.

NEW QUESTION 32
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 36
A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.
Which of the following Git operations does the data engineer need to run to accomplish this task?

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

Answer: C

Explanation:
 From the docs:
In Databricks Repos, you can use Git functionality to: Clone, push to, and pull from a remote Git repository.
Create and manage branches for development work, including merging, rebasing, and resolving conflicts.
Create notebooks—including IPYNB notebooks—and edit them and other files.
Visually compare differences upon commit and resolve merge conflicts. Source: https://docs.databricks.com/en/repos/index.html

NEW QUESTION 38
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 39
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 43
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,
   A.      employees,
           FILTER (employees, i -> i.years_exp > 5) AS exp_employees
      FROM stores;

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

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

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

      SELECT
           store_id,
   E.      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 48
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 49
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 54
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 58
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 63
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 68
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 nun 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 VACUUMwith 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 69
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 74
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 79
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 83
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 84
Which of the following data workloads will utilize a Gold table as its source?

A. A job that enriches data by parsing its timestamps into a human-readable format
B. A job that aggregates uncleaned data to create standard summary statistics
C. A job that cleans data by removing malformatted records
D. A job that queries aggregated data designed to feed into a dashboard
E. A job that ingests raw data from a streaming source into the Lakehouse

Answer: D

NEW QUESTION 87
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 89
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 91
Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my_table and save the updated table?

A. SELECT * FROM my_table WHERE age > 25;
B. UPDATE my_table WHERE age > 25;
C. DELETE FROM my_table WHERE age > 25;
D. UPDATE my_table WHERE age <= 25;
E. DELETE FROM my_table WHERE age <= 25;

Answer: C

NEW QUESTION 94
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 engineeringenvironment, Delta Live Tables is specifically designed for data quality monitoring and automation within the Databricks ecosystem.

NEW QUESTION 98
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 99
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 104
A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data.
Which of the following relational objects should the data engineer create?

A. Spark SQL Table
B. View
C. Database
D. Temporary view
E. Delta Table

Answer: D

Explanation:
Temp view : session based Create temp view view_name as query All these are termed as session ended: Opening a new notebook Detaching and reattaching a cluster Installing a python package Restarting a cluster

NEW QUESTION 105
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 110

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
        employeeId STRING,
B.      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 111
......

Thank You for Trying Our Product

We offer two products:

1st - We have Practice Tests Software with Actual Exam Questions

2nd - Questons and Answers in PDF Format

Databricks-Certified-Data-Engineer-Associate Practice Exam Features:

* Databricks-Certified-Data-Engineer-Associate Questions and Answers Updated Frequently

* Databricks-Certified-Data-Engineer-Associate Practice Questions Verified by Expert Senior Certified Staff

* Databricks-Certified-Data-Engineer-Associate Most Realistic Questions that Guarantee you a Pass on Your FirstTry

* Databricks-Certified-Data-Engineer-Associate Practice Test Questions in Multiple Choice Formats and Updatesfor 1 Year

100% Actual & Verified — Instant Download, Please Click
[Order The Databricks-Certified-Data-Engineer-Associate Practice Test Here](#)