CertyIQ

CertyIQ

Premium exam material

Get certification quickly with the CertyIQ Premium exam material.
Everything you need to prepare, learn & pass your certification exam easily. Lifetime free updates
First attempt guaranteed success.

https://www.CertyIQ.com

About CertyIQ

We here at CertyIQ eventually got enough of the industry's greedy exam paid for. Our team of IT professionals comes with years of experience in the IT industry Prior to training CertIQ we worked in test areas where we observed the horrors of the paywall exam preparation system.

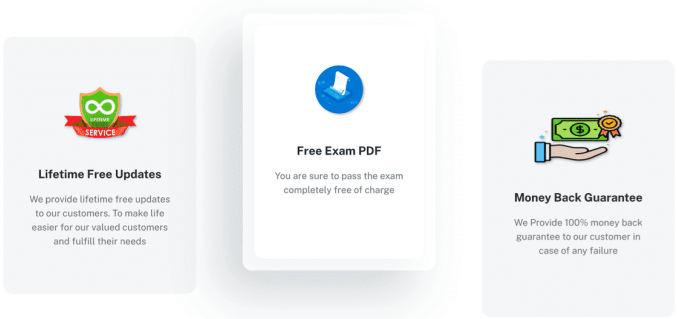
The misuse of the preparation system has left our team disillusioned. And for that reason, we decided it was time to make a difference. We had to make In this way, CertyIQ was created to provide quality materials without stealing from everyday people who are trying to make a living.

Doubt Support

We have developed a very scalable solution using which we are able to solve 400+ doubts every single day with an average rating of 4.8 out of 5.

https://www.certyiq.com

Mail us on - certyiqofficial@gmail.com



John

October 19, 2022

★★★★★

Thanks you so much for your help. I scored 972 in my exam today. More than 90% were from your PDFs!

Dana

September 04, 2022

★★★★★

Thanks a lot for this updated AZ-900 Q&A. I just passed my exam and got 974, I followed both of your Az-900 videos and the 6 PDF, the PDFs are very much valid, all answers are correct. Could you please create a similar video/PDF for DP900, your content/PDF's is really awesome. The team did a really good job. Thank You 😊.

Ahamed Shibly

2 months ago

★★★★★

Customer support is realy fast and helpful, I just finished my exam and this video along with the 6 PDF helped me pass! Definitely recommend getting the PDFs. Thank you!

October 22, 2022

★★★★★

Passed my exam today with 891 marks. Out of 52 questions, 51 were from certyiq PDFs including Contoso case study.
Thank You certyiq team!

Henry Rome

2 months ago

★★★★★

These questions are real and 100 % valid. Thank you so much for your efforts, also your 4 PDFs are awesome, I passed the DP900 exam on 1 Sept. With 968 marks. Thanks a lot, buddy!

Esmaria

2 months ago

★★★★★

Simple easy to understand explanations. To anyone out there wanting to write AZ900, I highly recommend 6 PDF's.Thank you so much, appreciate all your hard work in having such great content. Passed my exam Today - 3 September with 942 score.

Databricks

(Certified Data Engineer Associate)

Certified Data Engineer Associate

Total: **90 Questions**

Link: https://certyiq.com/papers?provider=databricks&exam=certified-data-engineer-associate

Question: 1 CertyIQ

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:

Unity Catalog in Databricks helps to eliminate Data Silos in an organization by having one single source of truth data.

Question: 2 CertyIQ

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:

Using cluster pools reduces the cluster start up time. So in this case, the reports can be refreshed quickly and not having to wait long for the cluster to start.

Question: 3 CertyIQ

Which of the following is hosted completely in the control plane of the classic Databricks architecture?

A. Worker node
B. JDBC data source
C. Databricks web application
D. Databricks Filesystem
E. Driver node

Answer: C

Explanation:

C. Data bricks web application.

In the classic Data bricks architecture, the control plane includes components like the Data bricks web application, the Data bricks REST API, and the Data bricks Workspace. These components are responsible for managing and controlling the Data bricks environment, including cluster provisioning, notebook management, access control, and job scheduling.

The other options, such as worker nodes, JDBC data sources, Data bricks Filesystem (DBFS), and driver nodes, are typically part of the data plane or the execution environment, which is separate from the control plane. Worker nodes are responsible for executing tasks and computations, JDBC data sources are used to connect to external databases, DBFS is a distributed file system for data storage, and driver nodes are responsible for coordinating the execution of Spark jobs.

Question: 4 CertyIQ

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:

D. The ability to support batch and streaming workloads.

Delta Lake is a key component of the Data bricks 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 Data bricks or the Lakehouse Platform in general, they are not specifically associated with Delta Lake.

Question: 5 CertyIQ

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:

C. Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.

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.

Question: 6 CertyIQ

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

Explanation:

DELETE FROM my_table WHERE age > 25;

Question: 7 CertyIQ

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:

A. The VACUUM command was run on the table.

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.

Question: 8 CertyIQ

Which of the following Git operations must be performed outside of Databricks Repos?

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

Answer: E

Explanation:

MERGE is the only git operation that is listed in the options that cannot be performed with Data bricks repos. CLONE is absolutely possible.

Clone can be done in Data bricks Repo. Merge not in Repos, need to be in Git.

https://learn.microsoft.com/en-us/azure/databricks/repos/

Question: 9 CertyIQ

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:

B. A data lake house supports ACID-compliant transactions.

One of the key features of a data lake house 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 lake house, by offering ACID compliance, helps maintain data quality by providing strong consistency and reliability, which is crucial for data pipelines and analytics.

Question: 10 CertyIQ

A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.
Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

A. Databricks Repos automatically saves development progress

B. Databricks Repos supports the use of multiple branches

C. Databricks Repos allows users to revert to previous versions of a notebook

D. Databricks Repos provides the ability to comment on specific changes

E. Databricks Repos is wholly housed within the Databricks Lakehouse Platform

Answer: B

Explanation:

B. Databricks Repos supports the use of multiple branches.

An advantage of using Databricks Repos over the built-in Databricks Notebooks versioning is the ability to work with multiple branches. Branching is a fundamental feature of version control systems like Git, which Databricks Repos is built upon. It allows you to create separate branches for different tasks, features, or experiments within your project. This separation helps in parallel development and experimentation without affecting the main branch or the work of other team members.

Branching provides a more organized and collaborative development environment, making it easier to merge changes and manage different development efforts. While Databricks Notebooks versioning also allows you to track versions of notebooks, it may not provide the same level of flexibility and collaboration as branching in Databricks Repos.

Question: 11 CertyIQ

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:

Correct answer is C:Workspace administrator.

Reference:

https://www.databricks.com/blog/2022/08/26/databricks-workspace-administration-best-practices-for-account-workspace-and-metastore-admins.html

Question: 12 CertyIQ

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:

Correct answer is E:spark.table("sales").

Reference:

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

Question: 13 CertyIQ

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:

C. DESCRIBE DATABASE customer360;

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.

Question: 14 CertyIQ

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:

Correct answer is D:COMMENT "Contains PII".

Question: 15 CertyIQ

Which of the following benefits is provided by the array functions from Spark SQL?

A. An ability to work with data in a variety of types at once
B. An ability to work with data within certain partitions and windows
C. An ability to work with time-related data in specified intervals
D. An ability to work with complex, nested data ingested from JSON files
E. An ability to work with an array of tables for procedural automation

Answer: D

Explanation:

D. An ability to work with complex, nested data ingested from JSON files.

Array functions in Spark SQL are primarily used for working with arrays and complex, nested data structures, such as those often encountered when ingesting JSON files. These functions allow you to manipulate and query nested arrays and structures within your data, making it easier to extract and work with specific elements or values within complex data formats.

While some of the other options (such as option A for working with different data types) are features of Spark SQL or SQL in general, array functions specifically excel at handling complex, nested data structures like those found in JSON files.

Question: 16 CertyIQ

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:

C. MERGE

The MERGE command is used to write data into a Delta table while avoiding the writing of duplicate records. It allows you to perform an "upsert" operation, which means that it will insert new records and update existing records in the Delta table based on a specified condition. This helps maintain data integrity and avoid duplicates when adding new data to the table.

Question: 17 CertyIQ

A data engineer needs to apply custom logic to string column city in table stores for a specific use case. In order to apply this custom logic at scale, the data engineer wants to create a SQL user-defined function (UDF).
Which of the following code blocks creates this SQL UDF?

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

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

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

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

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

Answer: A

Explanation:

The answer E is incorrect. A user defined function is never written as CREATE UDF. The correct way is CREATE FUNCTION. So that leaves us with the choices A and D. Out of that, in D, there is no such thing as RETURN CASE so the correct answer is A.

Question: 18 CertyIQ

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

Explanation:

They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.

Question: 19 CertyIQ

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:

The previous day's file has already been copied into the table.

Reference:

https://docs.databricks.com/ingestion/copy-into/tutorial-notebook.html

Question: 20 CertyIQ

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:

Correct answer is A:org.apache.spark.sql.jdbc.

Question: 21 CertyIQ

A data engineering team has two tables. The first table march_transactions is a collection of all retail transactions in the month of March. The second table april_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables.
Which of the following commands should be run to create a new table all_transactions that contains all records from march_transactions and april_transactions without duplicate records?

A. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
INNER JOIN SELECT * FROM april_transactions;

B. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
UNION SELECT * FROM april_transactions;

C. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
OUTER JOIN SELECT * FROM april_transactions;

D. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
INTERSECT SELECT * from april_transactions;

E. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
MERGE SELECT * FROM april_transactions;

Answer: B

Explanation:

CREATE TABLE all_transactions AS

SELECT * FROM march_transactions

UNION SELECT * FROM april_transactions;

Question: 22 CertyIQ

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:

D. if day_of_week == 1 and review_period.

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.

Question: 23 CertyIQ

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:

C is the correct answer. For external tables, you need to go to the specific location using DESCRIBE EXTERNAL TABLE command and delete all files.

Question: 24 CertyIQ

A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.
Which of the following data entities should the data engineer create?

A. Database
B. Function
C. View
D. Temporary view
E. Table

Answer: E

Explanation:

E. Table

To create a data entity that can be used by other data engineers in other sessions and must be saved to a physical location, you should create a table. Tables in a database are physical storage structures that hold data, and they can be accessed and shared by multiple users and sessions. By creating a table, you provide a permanent and structured storage location for the data entity that can be used across different sessions and by other users as needed.

Options like databases (A) can be used to organize tables, views (C) can provide virtual representations of data, and temporary views (D) are temporary in nature and don't save data to a physical location. Functions (B) are typically used for processing data or performing calculations, not for storing data.

Question: 25 CertyIQ

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:

D. Delta Live Tables

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.

Question: 26 CertyIQ

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. The compute resources will persist to allow for additional testing.

B. All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.

C. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped.

D. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.

E. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

Answer: C

Explanation:

All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped.

Question: 27 CertyIQ

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:

A. Checkpointing and Write-ahead Logs.

To reliably track the exact progress of processing and handle failures in Spark Structured Streaming, Spark uses both checkpointing and write-ahead logs. Checkpointing allows Spark to periodically save the state of the streaming application to a reliable distributed file system, which can be used for recovery in case of failures. Write-ahead logs are used to record the offset range of data being processed, ensuring that the system can recover and reprocess data from the last known offset in the event of a failure.

Question: 28 CertyIQ

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:

A. Gold tables are more likely to contain aggregations than Silver tables.

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.

Question: 29 CertyIQ

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:

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

In a typical data processing pipeline following a "Bronze-Silver-Gold" data lakehouse architecture, Bronze tables are the initial stage where raw data is ingested and transformed into a structured format with a schema applied. The schema provides structure and meaning to the raw data, making it more usable and accessible for downstream processing.

Therefore, Bronze tables contain the raw data but in a structured and schema-enforced format, which makes

them distinct from the unprocessed, unstructured raw data files.

Question: 30 CertyIQ

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:

B. Spark Structured Streaming.

The Auto Loader process in Data bricks 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 Data bricks 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.

Question: 31 CertyIQ

A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table.
The cade block used by the data engineer is below:

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

If the data engineer only wants the query to execute a micro-batch to process data every 5 seconds, which of the following lines of code should the data engineer use to fill in the blank?

A. trigger("5 seconds")
B. trigger()
C. trigger(once="5 seconds")
D. trigger(processingTime="5 seconds")
E. trigger(continuous="5 seconds")

Answer: D

Explanation:

trigger(processingTime="5 seconds")

Question: 32 CertyIQ

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:

C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.

Invalid rows will be dropped as requested by the constraint and flagged as such in log files. If you need a quarantine table, you'll have to write more code.

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.

Reference:

https://docs.databricks.com/en/delta-live-tables/expectations.html

Question: 33 CertyIQ

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:

B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally. The CREATE STREAMING LIVE TABLE syntax is used to create tables that read data incrementally, while the CREATE LIVE TABLE syntax is used to create tables that read data in batch mode. Delta Live Tables support both streaming and batch modes of processing data. When the data is streamed and needs to be processed incrementally, CREATE STREAMING LIVE TABLE should be used.

Question: 34 CertyIQ

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.

Reference:

https://docs.databricks.com/en/ingestion/auto-loader/index.html

Question: 35 CertyIQ

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")
)
```

Answer: E

Explanation:

E is the right answer. The "gold layer" is used to store aggregated clean data, E is the only answer in which aggregation is performed.

Question: 36 CertyIQ

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:

D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.

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.

Question: 37 CertyIQ

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:

B. They can create a new task in the existing Job and then add it as a dependency of the original task. Adding a new task as a dependency to an existing task in the same Job allows the new task to run before the original task is executed. This ensures that the data engineer can run the new notebook prior to the original task without having to create a new Job from scratch. Cloning the existing task or creating a new Job would add unnecessary complexity to the pipeline.

Question: 38 CertyIQ

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:

The correct answer is E. They can set the query's refresh schedule to end on a certain date in the query scheduler.Databricks SQL supports a query scheduler that enables users to schedule SQL queries to run at defined intervals. By default, scheduled queries run indefinitely. However, users can configure the scheduler to stop running queries at a specific time or after a specific number of runs. In this scenario, the engineering team can set the query's refresh schedule to end on a certain date, ensuring that the query does not run beyond the first week of the project's release and potentially cost the organization more money.

Question: 39 CertyIQ

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:

A: increase the cluster size of the SQL endpoint.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.

Question: 40 CertyIQ

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

Explanation:

The data engineer can use the Auto Stop feature to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard. The Auto Stop feature allows the SQL endpoint to automatically shut down when there are no active connections, which will minimize the total running time of the SQL endpoint. By scheduling the dashboard to refresh once per day, the SQL endpoint will only be running for a short period of time each day, which will minimize the total running time and reduce costs.

Question: 41 CertyIQ

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:

The approach the data engineer can use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100 is:C. They can set up an Alert with a new webhook alert destination.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.

Question: 42 CertyIQ

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:

Correct answer is C:They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.

Reference:

https://docs.databricks.com/workflows/jobs/jobs.html

Question: 43 CertyIQ

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:

D: use clusters that are from a cluster pool.Using clusters from a cluster pool can improve the start-up time for the clusters used in the Job because the pool contains preconfigured and pre-started clusters that can be used immediately. This can save time and resources compared to starting new clusters for each task.

D. They can use clusters that are from a cluster pool. Cluster pools allow you to pre-create a pool of ready-to-use clusters that can be used for running jobs, thereby eliminating the need to start new clusters each time a job runs. This can greatly reduce the startup time for each task.

Question: 44 CertyIQ

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 customers TO team;

E. GRANT ALL PRIVILEGES ON DATABASE customers TO team;

Answer: E

Explanation:

GRANT ALL PRIVILEGES ON DATABASE customers TO team;

Question: 45 CertyIQ

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:

Correct answer is E:GRANT USAGE ON DATABASE customers TO team;

Question: 46 CertyIQ

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:

pull is required from the Databricks Repo to sync the changes b/w local and central repo.

Question: 47 CertyIQ

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:

Correct answer is E:Avoiding vendor lock-in.

Question: 48 CertyIQ

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

Explanation:

Correct answer is E:Data Explorer.

Question: 49 CertyIQ

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:

When they are working interactively with a small amount of data.

Question: 50 CertyIQ

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

Explanation:

INSERT INTO my_table VALUES ('a1', 6, 9.4).

Question: 51 CertyIQ

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.

Question: 52 CertyIQ

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:

Correct answer is C:Parquet.

Question: 53 CertyIQ

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

Explanation:

Correct answer is D:Data.

Question: 54 CertyIQ

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

Explanation:

Data Lakehouse can be used as a single source of truth for multiple specific use cases.

Question: 55 CertyIQ

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?

A.

```
CREATE TABLE IF NOT EXISTS table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

B.

```
CREATE OR REPLACE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
USING DELTA
```

C.

```
CREATE OR REPLACE TABLE table_name WITH COLUMNS (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
) USING DELTA
```

D.

```
CREATE TABLE table_name AS
SELECT
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
```

E.

```
CREATE OR REPLACE TABLE table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

Answer: E

Explanation:

```
CREATE OR REPLACE TABLE table_name (
  employeeId STRING,
  startDate DATE,
  avgRating FLOAT
)
```

Question: 56 CertyIQ

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

Explanation:

They can add %sql to the first line of the cell.

Question: 57 CertyIQ

Which of the following SQL keywords can be used to convert a table from a long format to a wide format?

A.TRANSFORM
B.PIVOT
C.SUM

D.CONVERT

E.WHERE

Answer: B

Explanation:

Correct answer is B:PIVOT.

Question: 58 CertyIQ

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:

CTAS - CTAS automatically infer schema information from queryresults and do not support manual schema declaration.This meansthat CTAS statements are useful for external data ingestion fromsources with well-defined schema, such as Parquet files andtables.CTAS statements also do not support specifying additionalfile options.

Question: 59 CertyIQ

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:

D is correct.

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.

Question: 60 CertyIQ

A data analyst has developed a query that runs against Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?

A.SELECT * FROM sales

B.spark.delta.table

C.spark.sql

D.There is no way to share data between PySpark and SQL.

E.spark.table

Answer: C

Explanation:

spark.sql() should be used to execute a SQL query with Pysparkspark.table() can only be used to load a table and not run a query.

Question: 61 CertyIQ

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:

Correct answer is C:SELECT count_if(member_id IS NULL) FROM my_table;

Question: 62 CertyIQ

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?

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

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

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

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

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

Answer: A

Explanation:

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

Question: 63 CertyIQ

A data engineer has a Python variable table_name that they would like to use in a SQL query. They want to construct a Python code block that will run the query using table_name.

They have the following incomplete code block:

____(f"SELECT customer_id, spend FROM table_name ")

Which of the following can be used to fill in the blank to successfully complete the task?

A.spark.delta.sql
B.spark.delta.table
C.spark.table
D.dbutils.sql
E.spark.sql

Answer: E

Explanation:

E is correct you use spark.sql to execute python comamands.

Question: 64 CertyIQ

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
E.dbfs:/user/hive/database

Answer: B

Explanation:

dbfs:/user/hive/warehouse.

Question: 65 CertyIQ

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:

Correct answer is A:The table was managed.

Question: 66 CertyIQ

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
```

Answer: D

Explanation:

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

Question: 67 CertyIQ

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:

When the target table cannot contain duplicate records.

Question: 68 CertyIQ

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

Answer: C

Explanation:

A typical LEFT JOIN scenario.

The LEFT JOIN keyword returns all records from the left table, even if there are no matches in the right table.

Question: 69 CertyIQ

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

Explanation:

Correct answer is B:USING CSV.

Question: 70 CertyIQ

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

The code block used by the data engineer is below:

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

If the data engineer only wants the query to process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

A.processingTime(1)

B.trigger(availableNow=True)

C.trigger(parallelBatch=True)

D.trigger(processingTime="once")

E.trigger(continuous="once")

Answer: B

Explanation:

Correct answer is B:trigger(availableNow=True).

Question: 71 CertyIQ

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:

The correct answer is: B. JSON data is a text-based format 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.htmlFor example, 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.

Question: 72 CertyIQ

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. The compute resources will be terminated.

B.All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist until the pipeline is shut down.

C.All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.

D.All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

E.All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

Answer: E

Explanation:

E. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

Question: 73 CertyIQ

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

Explanation:

A job that queries aggregated data designed to feed into a dashboard.

Question: 74 CertyIQ

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: C

Explanation:

A path to a cloud storage location for the written data - considering this option is talking about the source data being stored in cloud storage and being ingested to DLT using an autoloader.

Question: 75 CertyIQ

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:

The customers table is a streaming live table.

Reference:

https://docs.databricks.com/en/sql/load-data-streaming-table.html

Question: 76 CertyIQ

Which of the following describes the type of workloads that are always compatible with Auto Loader?

A.Streaming workloads
B.Machine learning workloads
C.Serverless workloads
D.Batch workloads
E.Dashboard workloads

Answer: A

Explanation:

Correct answer is A:Streaming workloads.

Question: 77 CertyIQ

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

Explanation:

None of these changes will need to be made.

Question: 78 CertyIQ

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:

Replace spark.read with spark.readStream.

Question: 79 CertyIQ

Which of the following queries is performing a streaming hop from raw data to a Bronze table?

A.

```
(spark.table("sales")
      .groupBy("store")
      .agg(sum("sales"))
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("complete")
      .table("newSales")
)
```

B.

```
(spark.table("sales")
      .filter(col("units") > 0)
      .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.read.load(rawSalesLocation)
      .write
      .mode("append")
      .table("newSales")
)
```

E.

```
(spark.readStream.load(rawSalesLocation)
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("append")
      .table("newSales")
)
```

Answer: E

Explanation:

```
(spark.readStream.load(rawSalesLocation)
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("append")
      .table("newSales")
)
```

Question: 80 CertyIQ

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:

Records that violate the expectation cause the job to fail.

Reference:

https://docs.databricks.com/en/delta-live-tables/expectations.html

Question: 81 CertyIQ

Which of the following statements regarding the relationship between Silver tables and Bronze tables is always true?

A.Silver tables contain a less refined, less clean view of data than Bronze data.

B.Silver tables contain aggregates while Bronze data is unaggregated.

C.Silver tables contain more data than Bronze tables.

D.Silver tables contain a more refined and cleaner view of data than Bronze tables.

E.Silver tables contain less data than Bronze tables.

Answer: D

Explanation:

Silver tables contain a more refined and cleaner view of data than Bronze tables.

Question: 82 CertyIQ

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: E

Explanation:

They can increase the maximum bound of the SQL endpoint's scaling range.

Question: 83 CertyIQ

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

Explanation:

Ans D : Cron Syntax with that you can easily copy all the syntax.

Question: 84 CertyIQ

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:

Setting up an Alert in the Job page.

Reference:

https://docs.databricks.com/en/workflows/jobs/job-notifications.html

Question: 85 CertyIQ

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

Explanation:

They can schedule the query to refresh every 1 day from the query's page in Databricks SQL.

Question: 86 CertyIQ

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

Explanation:

When another task needs to successfully complete before the new task begins.

Question: 87 CertyIQ

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

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

A.They can set up an Alert with a custom template.

B.They can set up an Alert with a new email alert destination.

C.They can set up an Alert with one-time notifications.

D.They can set up an Alert with a new webhook alert destination.

E.They can set up an Alert without notifications.

Answer: D

Explanation:

They can set up an Alert with a new web hook alert destination.

Question: 88 CertyIQ

A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that feeds the dashboard is automatically processed using a Databricks Job.

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 turn on the Auto Stop feature for the SQL endpoint.

B.They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

C.They can reduce the cluster size of the SQL endpoint.

D.They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.

E.They can set up the dashboard's SQL endpoint to be serverless.

Answer: A

Explanation:

They can turn on the Auto Stop feature for the SQL endpoint.

Question: 89 CertyIQ

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

Explanation:

Review the Owner field in the table's page in Data Explorer.

Question: 90 CertyIQ

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

Explanation:

GRANT ALL PRIVILEGES ON TABLE sales TO team;.

Thank you

Thank you for being so interested in the premium exam material.
I'm glad to hear that you found it informative and helpful.

If you have any feedback or thoughts on the bumps, I would love to hear them.
Your insights can help me improve our writing and better understand our readers.

Best of Luck

You have worked hard to get to this point, and you are well-prepared for the exam
Keep your head up, stay positive, and go show that exam what you're made of!

Feedback

More Papers