Free Questions for Databricks-Certified-Data-Analyst-Associate

Shared by Maldonado on 04-10-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

Which of the following benefits of using Databricks SQL is provided by Data Explorer?

Options:

A- It can be used to run UPDATE queries to update any tables in a database.

B- It can be used to view metadata and data, as well as view/change permissions.

C- It can be used to produce dashboards that allow data exploration.

D- It can be used to make visualizations that can be shared with stakeholders.

E- It can be used to connect to third party BI cools.

Answer:

B

Explanation:

Data Explorer is a user interface that allows you to discover and manage data, schemas, tables, models, and permissions in Databricks SQL. You can use Data Explorer to view schema details, preview sample data, and see table and model details and

properties.Administrators can view and change owners, and admins and data object owners can grant and revoke permissions1.Reference:Discover and manage data using Data Explorer

Question 2

Question Type: MultipleChoice

Which of the following is an advantage of using a Delta Lake-based data lakehouse over common data lake solutions?

Options:

A- ACID transactions

B- Flexible schemas

C- Data deletion

D- Scalable storage

E- Open-source formats

Answer:

A

Explanation:

A Delta Lake-based data lakehouse is a data platform architecture that combines the scalability and flexibility of a data lake with the reliability and performance of a data warehouse. One of the key advantages of using a Delta Lake-based data lakehouse over common data lake solutions is that it supports ACID transactions, which ensure data integrity and consistency. ACID transactions enable concurrent reads and writes, schema enforcement and evolution, data versioning and rollback, and data quality checks. These features are not available in traditional data lakes, which rely on file-based storage systems that do not support transactions.Reference:

Delta Lake: Lakehouse, warehouse, advantages | Definition

Synapse -- Data Lake vs. Delta Lake vs. Data Lakehouse

Data Lake vs. Delta Lake - A Detailed Comparison

Building a Data Lakehouse with Delta Lake Architecture: A Comprehensive Guide

Question 3

Question Type: MultipleChoice

Delta Lake stores table data as a series of data files, but it also stores a lot of other information.

Which of the following is stored alongside data files when using Delta Lake?

Options:

A- None of these

B- Table metadata, data summary visualizations, and owner account information

C- Table metadata

D- Data summary visualizations

E- Owner account information

Answer:

C

Explanation:

Delta Lake stores table data as a series of data files in a specified location, but it also stores table metadata in a transaction log. The table metadata includes the schema, partitioning information, table properties, and other configuration details. The table metadata is stored alongside the data files and is updated atomically with every write operation. The table metadata can be accessed using the DESCRIBE DETAIL command or the DeltaTable class in Scala, Python, or Java. The table metadata can also be enriched with custom tags or user-defined commit messages using the TBLPROPERTIES or userMetadata options.Reference:

Enrich Delta Lake tables with custom metadata

Delta Lake Table metadata - Stack Overflow

Metadata - The Internals of Delta Lake

Question 4

Question Type: MultipleChoice

Which of the following should data analysts consider when working with personally identifiable information (PII) data?

Options:

A- Organization-specific best practices for Pll data

B- Legal requirements for the area in which the data was collected

C- None of these considerations

D- Legal requirements for the area in which the analysis is being performed

E- All of these considerations

Answer:

E

Explanation:

Data analysts should consider all of these factors when working with PII data, as they may affect the data security, privacy, compliance, and quality. PII data is any information that can be used to identify a specific individual, such as name, address, phone number, email, social security number, etc. PII data may be subject to different legal and ethical obligations depending on the context and location of the data collection and analysis. For example, some countries or regions may have stricter data protection laws than others, such as the General Data Protection Regulation (GDPR) in the European Union. Data analysts should also follow the organization-specific best practices for PII data, such as encryption, anonymization, masking, access control, auditing, etc. These best practices can help prevent data breaches, unauthorized access, misuse, or loss of PII data.Reference:

How to Use Databricks to Encrypt and Protect PII Data

Automating Sensitive Data (PII/PHI) Detection

Databricks Certified Data Analyst Associate

Question 5

Question Type: MultipleChoice

After running DESCRIBE EXTENDED accounts.customers;, the following was returned:

```
Name        accounts.customers
Location    dbfs:/stakeholders/customers
Provider    delta
Owner       root
Type        EXTERNAL
```

Now, a data analyst runs the following command:

DROP accounts.customers;

Which of the following describes the result of running this command?

Options:

A- Running SELECT * FROM delta. `dbfs:/stakeholders/customers` results in an error.

B- Running SELECT * FROM accounts.customers will return all rows in the table.

C- All files with the .customers extension are deleted.

D- The accounts.customers table is removed from the metastore, and the underlying data files are deleted.

E- The accounts.customers table is removed from the metastore, but the underlying data files are untouched.

Answer:

E

Explanation:

the accounts.customers table is an EXTERNAL table, which means that it is stored outside the default warehouse directory and is not managed by Databricks. Therefore, when you run the DROP command on this table, it only removes the metadata information from the metastore, but does not delete the actual data files from the file system. This means that you can still access the data using the location path (dbfs:/stakeholders/customers) or create another table pointing to the same location. However, if you try to query the table using its name (accounts.customers), you will get an error because the table no longer exists in the metastore.Reference:DROP TABLE | Databricks on AWS,Best practices for dropping a managed Delta Lake table - Databricks

Question 6

Question Type: MultipleChoice

A data analyst is attempting to drop a table my_table. The analyst 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?

Options:

A- The table's data was larger than 10 GB

B- The table did not have a location

C- The table was external

D- The table's data was smaller than 10 GB

E- The table was managed

Answer:

C

Explanation:

An external table is a table that is defined in the metastore, but its data is stored outside of the Databricks environment, such as in S3, ADLS, or GCS. When an external table is dropped, only the metadata is deleted from the metastore, but the data files are not affected. This is different from a managed table, which is a table whose data is stored in the Databricks environment, and whose data files are deleted when the table is dropped. To delete the data files of an external table, the analyst needs to specify the PURGE option in the DROP TABLE command, or manually delete the files from the storage system.Reference:DROP TABLE,Drop Delta table features,Best practices for dropping a managed Delta Lake table

Question 7

Question Type: MultipleChoice

A data analyst needs to use the Databricks Lakehouse Platform to quickly create SQL queries and data visualizations. It is a requirement that the compute resources in the platform can be made serverless, and it is expected that data visualizations can be placed within a dashboard.

Which of the following Databricks Lakehouse Platform services/capabilities meets all of these requirements?

Options:

A- Delta Lake

B- Databricks Notebooks

C- Tableau

D- Databricks Machine Learning

E- Databricks SQL

Answer:

E

Explanation:

Databricks SQL is a serverless data warehouse on the Lakehouse that lets you run all of your SQL and BI applications at scale with your tools of choice, all at a fraction of the cost of traditional cloud data warehouses1.Databricks SQL allows you to create SQL queries and data visualizations using the SQL Analytics UI or the Databricks SQL CLI2.You can also place your data visualizations within a dashboard and share it with other users in your organization3.Databricks SQL is powered by Delta Lake, which provides reliability, performance, and governance for your data lake4.Reference:

Databricks SQL

Query data using SQL Analytics

Visualizations in Databricks notebooks

Delta Lake

Question 8

Question Type: MultipleChoice

A data analyst wants to create a dashboard with three main sections: Development, Testing, and Production. They want all three sections on the same dashboard, but they want to clearly designate the sections using text on the dashboard.

Which of the following tools can the data analyst use to designate the Development, Testing, and Production sections using text?

Options:

A- Separate endpoints for each section

B- Separate queries for each section

C- Markdown-based text boxes

D- Direct text written into the dashboard in editing mode

E- Separate color palettes for each section

Answer:

C

Explanation:

Markdown-based text boxes are useful as labels on a dashboard. They allow the data analyst to add text to a dashboard using the %md magic command in a notebook cell and then select the dashboard icon in the cell actions menu. The text can be formatted using markdown syntax and can include headings, lists, links, images, and more. The text boxes can be resized and moved around on the

dashboard using the float layout option.Reference:Dashboards in notebooks,How to add text to a dashboard in Databricks

Question 9

Question Type: MultipleChoice

Which of the following approaches can be used to ingest data directly from cloud-based object storage?

Options:

A- Create an external table while specifying the DBFS storage path to FROM

B- Create an external table while specifying the DBFS storage path to PATH

C- It is not possible to directly ingest data from cloud-based object storage

D- Create an external table while specifying the object storage path to FROM

E- Create an external table while specifying the object storage path to LOCATION

Answer:

E

Explanation:

External tables are tables that are defined in the Databricks metastore using the information stored in a cloud object storage location. External tables do not manage the data, but provide a schema and a table name to query the data. To create an external table, you can use the CREATE EXTERNAL TABLE statement and specify the object storage path to the LOCATION clause. For example, to create an external table named ext_table on a Parquet file stored in S3, you can use the following statement:

SQL

```
CREATE EXTERNAL TABLE ext_table (

col1 INT,

col2 STRING

)

STORED AS PARQUET

LOCATION 's3://bucket/path/file.parquet'
```

AI-generated code. Review and use carefully.More info on FAQ.

Question 10

Question Type: MultipleChoice

A data engineering team has created a Structured Streaming pipeline that processes data in micro-batches and populates gold-level tables. The microbatches are triggered every minute.

A data analyst has created a dashboard based on this gold-level data. The project stakeholders want to see the results in the dashboard updated within one minute or less of new data becoming available within the gold-level tables.

Which of the following cautions should the data analyst share prior to setting up the dashboard to complete this task?

Options:

A- The required compute resources could be costly

B- The gold-level tables are not appropriately clean for business reporting

C- The streaming data is not an appropriate data source for a dashboard

D- The streaming cluster is not fault tolerant

E- The dashboard cannot be refreshed that quickly

Answer:

A

Explanation:

A Structured Streaming pipeline that processes data in micro-batches and populates gold-level tables every minute requires a high level of compute resources to handle the frequent data ingestion, processing, and writing. This could result in a significant cost for the organization, especially if the data volume and velocity are large. Therefore, the data analyst should share this caution with the project stakeholders before setting up the dashboard and evaluate the trade-offs between the desired refresh rate and the available budget. The other options are not valid cautions because:

B) The gold-level tables are assumed to be appropriately clean for business reporting, as they are the final output of the data engineering pipeline. If the data quality is not satisfactory, the issue should be addressed at the source or silver level, not at the gold level.

C) The streaming data is an appropriate data source for a dashboard, as it can provide near real-time insights and analytics for the business users. Structured Streaming supports various sources and sinks for streaming data, including Delta Lake, which can enable both batch and streaming queries on the same data.

D) The streaming cluster is fault tolerant, as Structured Streaming provides end-to-end exactly-once fault-tolerance guarantees through checkpointing and write-ahead logs. If a query fails, it can be restarted from the last checkpoint and resume processing.

E) The dashboard can be refreshed within one minute or less of new data becoming available in the gold-level tables, as Structured Streaming can trigger micro-batches as fast as possible (every few seconds) and update the results incrementally. However, this may not be necessary or optimal for the business use case, as it could cause frequent changes in the dashboard and consume more resources.Reference:Streaming on Databricks,Monitoring Structured Streaming queries on Databricks,A look at the new Structured Streaming UI in Apache Spark 3.0,Run your first Structured Streaming workload

Question 11

Question Type: MultipleChoice

A data analyst has set up a SQL query to run every four hours on a SQL endpoint, but the SQL endpoint is taking too long to start up with each run.

Which of the following changes can the data analyst make to reduce the start-up time for the endpoint while managing costs?

Options:

A- Reduce the SQL endpoint cluster size

B- Increase the SQL endpoint cluster size

C- Turn off the Auto stop feature

D- Increase the minimum scaling value

E- Use a Serverless SQL endpoint

Answer:

E

Explanation:

A Serverless SQL endpoint is a type of SQL endpoint that does not require a dedicated cluster to run queries. Instead, it uses a shared pool of resources that can scale up and down automatically based on the demand. This means that a Serverless SQL endpoint can start

up much faster than a SQL endpoint that uses a cluster, and it can also save costs by only paying for the resources that are used. A Serverless SQL endpoint is suitable for ad-hoc queries and exploratory analysis, but it may not offer the same level of performance and isolation as a SQL endpoint that uses a cluster. Therefore, a data analyst should consider the trade-offs between speed, cost, and quality when choosing between a Serverless SQL endpoint and a SQL endpoint that uses a cluster.Reference:Databricks SQL endpoints,Serverless SQL endpoints,SQL endpoint clusters

To Get Premium Files for Databricks-Certified-Data-Analyst-Associate Visit

https://www.p2pexams.com/products/databricks-certified-data-analyst-associate

For More Free Questions Visit

https://www.p2pexams.com/databricks/pdf/databricks-certified-data-analyst-associate