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

Shared by Quinn on 24-05-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

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 2

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 3

Question Type: MultipleChoice

A data analyst has a managed table table_name in database database_name. They would now like to remove the table from the database and all of the data files associated with the table. The rest of the tables in the database must continue to exist.

Which of the following commands can the analyst use to complete the task without producing an error?

Options:

A- DROP DATABASE database_name;
B- DROP TABLE database_name.table_name;
C- DELETE TABLE database_name.table_name;
D- DELETE TABLE table_name FROM database_name;
E- DROP TABLE table_name FROM database_name;

Answer:

B

Explanation:

The DROP TABLE command removes a table from the metastore and deletes the associated data files. The syntax for this command is DROP TABLE [IF EXISTS] [database_name.]table_name;. The optional IF EXISTS clause prevents an error if the table does not exist. The optional database_name. prefix specifies the database where the table resides. If not specified, the current database is used. Therefore, the correct command to remove the table table_name from

the database database_name and all of the data files associated with it is DROP TABLE database_name.table_name;. The other commands are either invalid syntax or would produce undesired results.Reference:Databricks - DROP TABLE

Question 4

Question Type: MultipleChoice

A business analyst has been asked to create a data entity/object called sales_by_employee. It should always stay up-to-date when new data are added to the sales table. The new entity should have the columns sales_person, which will be the name of the employee from the employees table, and sales, which will be all sales for that particular sales person. Both the sales table and the employees table have an employee_id column that is used to identify the sales person.

Which of the following code blocks will accomplish this task?

A)

```
CREATE TEMPORARY TABLE sales_by_employee AS
    SELECT employees.employee_name sales_person,
           sales.sales
    FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id;
```

B)

```
CREATE OR REPLACE VIEW sales_by_employee USING
    SELECT employees.employee_name sales_person,
           sales.sales
    FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id;
```

C)

```
SELECT employees.employee_name sales_person,
       sales.sales
    FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id USING
    CREATE OR REPLACE VIEW sales_by_employee;
```

D)

```
CREATE OR REPLACE VIEW sales_by_employee AS
    SELECT employees.employee_name sales_person,
           sales.sales FROM sales
    JOIN employees
    ON employees.employee_id = sales.employee_id;
```

Options:

A- Option
B- Option
C- Option
D- Option

Answer:

D

Explanation:

The SQL code provided in Option D is the correct way to create a view namedsales_by_employeethat will always stay up-to-date with the sales and employees tables. The code uses theCREATE OR REPLACE VIEWstatement to define a new view that joins the sales and employees tables on the employee_id column. It selects the employee_name as sales_person and all sales for each employee, ensuring that the data entity/object is always up-to-date when new data are added to these tables.

Question 5

Question Type: MultipleChoice

How can a data analyst determine if query results were pulled from the cache?

Options:

A- Go to the Query History tab and click on the text of the query. The slideout shows if the results came from the cache.
B- Go to the Alerts tab and check the Cache Status alert.
C- Go to the Queries tab and click on Cache Status. The status will be green if the results from the last run came from the cache.

D- Go to the SQL Warehouse (formerly SQL Endpoints) tab and click on Cache. The Cache file will show the contents of the cache.
E- Go to the Data tab and click Last Query. The details of the query will show if the results came from the cache.

Answer:

A

Explanation:

Databricks SQL uses a query cache to store the results of queries that have been executed previously. This improves the performance and efficiency of repeated queries. To determine if a query result was pulled from the cache, you can go to the Query History tab in the Databricks SQL UI and click on the text of the query. A slideout will appear on the right side of the screen, showing the query details, including the cache status. If the result came from the cache, the cache status will show ''Cached''. If the result did not come from the cache, the cache status will show ''Not cached''. You can also see the cache hit ratio, which is the percentage of queries that were served from the cache.Reference: The answer can be verified from Databricks SQL documentation which provides information on how to use the query cache and how to check the cache status. Reference link: Databricks SQL - Query Cache

Question 6

Question Type: MultipleChoice

The stakeholders.customers table has 15 columns and 3,000 rows of data. The following command is run:

```
CREATE TEMP VIEW stakeholders.eur_customers AS
    SELECT * FROM stakeholders.customers
    WHERE continent = 'eur';
```

After running SELECT * FROM stakeholders.eur_customers, 15 rows are returned. After the command executes completely, the user logs out of Databricks.

After logging back in two days later, what is the status of the stakeholders.eur_customers view?

Options:

A- The view remains available and SELECT * FROM stakeholders.eur_customers will execute

correctly.
B- The view has been dropped.
C- The view is not available in the metastore, but the underlying data can be accessed with SELECT * FROM delta. `stakeholders.eur_customers`.
D- The view remains available but attempting to SELECT from it results in an empty result set because data in views are automatically deleted after logging out.
E- The view has been converted into a table.

Answer:

B

Explanation:

The command you sent creates a TEMP VIEW, which is a type of view that is only visible and accessible to the session that created it. When the session ends or the user logs out, the TEMP VIEW is automatically dropped and cannot be queried anymore. Therefore, after logging back in two days later, the status of the stakeholders.eur_customers view is that it has been dropped and SELECT * FROM stakeholders.eur_customers will result in an error. The other options are not correct because:

A) The view does not remain available, as it is a TEMP VIEW that is dropped when the session ends or the user logs out.

C) The view is not available in the metastore, as it is a TEMP VIEW that is not registered in the metastore. The underlying data cannot be accessed with SELECT * FROM delta.stakeholders.eur_customers, as this is not a valid syntax for querying a Delta Lake table. The correct syntax would be SELECT * FROM delta.dbfs:/stakeholders/eur_customers, where the location path is enclosed in backticks. However, this would also result in an error, as the TEMP VIEW does not write any data to the file system and the location path does not exist.

D) The view does not remain available, as it is a TEMP VIEW that is dropped when the session ends or the user logs out. Data in views are not automatically deleted after logging out, as views do not store any data. They are only logical representations of queries on base tables or other views.

E) The view has not been converted into a table, as there is no automatic conversion between views and tables in Databricks. To create a table from a view, you need to use a CREATE TABLE AS statement or a similar command.Reference:CREATE VIEW | Databricks on AWS,Solved: How do temp views actually work? - Databricks - 20136,temp tables in Databricks - Databricks - 44012,Temporary View in Databricks - BIG DATA PROGRAMMERS,Solved: What is the difference between a Temporary View an ...

Question 7

Question Type: MultipleChoice

A data analyst is processing a complex aggregation on a table with zero null values and their query returns the following result:

group_1	group_2	sum
null	null	100
null	Y	70
null	Z	30
A	null	50
A	Y	30
A	Z	20
B	null	50
B	Y	40
B	Z	10

Which of the following queries did the analyst run to obtain the above result?

A)

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2 INCLUDING NULL;
```

B)

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2 WITH ROLLUP;
```

C)

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group 2;
```

D)

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2, (group_1, group_2);
```

E)

```
SELECT
    group_1,
    group_2,
    count(values) AS count
FROM my_table
GROUP BY group_1, group_2 WITH CUBE;
```

Options:

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

Answer:

B

Explanation:

The result set provided shows a combination of grouping by two columns (group_1 and group_2) with subtotals for each level of grouping and a grand total. This pattern is typical of a GROUP BY ... WITH ROLLUP operation in SQL, which provides subtotal rows and a grand total row in the

result set.

Considering the query options:

A) Option A: GROUP BY group_1, group_2 INCLUDING NULL - This is not a standard SQL clause and would not result in subtotals and a grand total.

B) Option B: GROUP BY group_1, group_2 WITH ROLLUP - This would create subtotals for each unique group_1, each combination of group_1 and group_2, and a grand total, which matches the result set provided.

C) Option C: GROUP BY group_1, group 2 - This is a simple GROUP BY and would not include subtotals or a grand total.

D) Option D: GROUP BY group_1, group_2, (group_1, group_2) - This syntax is not standard and would likely result in an error or be interpreted as a simple GROUP BY, not providing the subtotals and grand total.

E) Option E: GROUP BY group_1, group_2 WITH CUBE - The WITH CUBE operation produces subtotals for all combinations of the selected columns and a grand total, which is more than what is shown in the result set.

The correct answer is Option B, which uses WITH ROLLUP to generate the subtotals for each level of grouping as well as a grand total. This matches the result set where we have subtotals for each group_1, each combination of group_1 and group_2, and the grand total where both group_1 and group_2 are NULL.

Question 8

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 9

Question Type: MultipleChoice

A data analyst has created a Query in Databricks SQL, and now they want to create two data visualizations from that Query and add both of those data visualizations to the same Databricks SQL Dashboard.

Which of the following steps will they need to take when creating and adding both data visualizations to the Databricks SQL Dashboard?

Options:

A- They will need to alter the Query to return two separate sets of results.
B- They will need to add two separate visualizations to the dashboard based on the same Query.
C- They will need to create two separate dashboards.
D- They will need to decide on a single data visualization to add to the dashboard.
E- They will need to copy the Query and create one data visualization per query.

Answer:

B

Explanation:

A data analyst can create multiple visualizations from the same query in Databricks SQL by clicking the + button next to the Results tab and selecting Visualization. Each visualization can have a different type, name, and configuration. To add a visualization to a dashboard, the data analyst can click the vertical ellipsis button beneath the visualization, select + Add to Dashboard, and choose an existing or new dashboard. The data analyst can repeat this process for each visualization they want to add to the same dashboard.Reference:Visualization in Databricks SQL,Visualize queries and create a dashboard in Databricks SQL

Question 10

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 11

Question Type: MultipleChoice

A data analyst has created a user-defined function using the following line of code:

```
CREATE FUNCTION price(spend DOUBLE, units DOUBLE)

RETURNS DOUBLE

RETURN spend / units;
```

Which of the following code blocks can be used to apply this function to the customer_spend and customer_units columns of the table customer_summary to create column customer_price?

Options:

A- SELECT PRICE customer_spend, customer_units AS customer_price FROM customer_summary
B- SELECT price FROM customer_summary
C- SELECT function(price(customer_spend, customer_units)) AS customer_price FROM customer_summary
D- SELECT double(price(customer_spend, customer_units)) AS customer_price FROM customer_summary
E- SELECT price(customer_spend, customer_units) AS customer_price FROM customer_summary

Answer:

E

Explanation:

A user-defined function (UDF) is a function defined by a user, allowing custom logic to be reused in the user environment1.To apply a UDF to a table, the syntax isSELECT udf_name(column_name) AS alias FROM table_name2. Therefore, option E is the correct way to use the UDFpriceto create a new columncustomer_pricebased on the existing columnscustomer_spendandcustomer_unitsfrom the tablecustomer_summary.Reference:

What are user-defined functions (UDFs)?

User-defined scalar functions - SQL

V

Question 12

Question Type: MultipleChoice

An analyst writes a query that contains a query parameter. They then add an area chart visualization to the query. While adding the area chart visualization to a dashboard, the analyst chooses "Dashboard Parameter" for the query parameter associated with the area chart.

Which of the following statements is true?

Options:

A- The area chart will use whatever is selected in the Dashboard Parameter while all or the other visualizations will remain changed regardless of their parameter use.
B- The area chart will use whatever is selected in the Dashboard Parameter along with all of the other visualizations in the dashboard that use the same parameter.
C- The area chart will use whatever value is chosen on the dashboard at the time the area chart is added to the dashboard.
D- The area chart will use whatever value is input by the analyst when the visualization is added to the dashboard. The parameter cannot be changed by the user afterwards.
E- The area chart will convert to a Dashboard Parameter.

Answer:

B

Explanation:

A Dashboard Parameter is a parameter that is configured for one or more visualizations within a dashboard and appears at the top of the dashboard.The parameter values specified for a Dashboard Parameter apply to all visualizations reusing that particular Dashboard Parameter1. Therefore, if the analyst chooses ''Dashboard Parameter'' for the query parameter associated with the area chart, the area chart will use whatever is selected in the Dashboard Parameter along with all of the other visualizations in the dashboard that use the same parameter.This allows the user to filter the data across multiple visualizations using a single parameter widget2.Reference:Databricks SQL dashboards,Query parameters