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

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 2

Question Type: MultipleChoice

A data engineer is working with a nested array column products in table transactions. They want to expand the table so each unique item in products for each row has its own row where the

transaction_id column is duplicated as necessary.

They are using the following incomplete command:

```
SELECT
    transaction_id,
    _______ AS product
FROM transactions;
```

Which of the following lines of code can they use to fill in the blank in the above code block so that it successfully completes the task?

Options:

A- array distinct(produces)
B- explode(produces)
C- reduce(produces)
D- array(produces)
E- flatten(produces)

Answer:

B

Explanation:

Theexplodefunction is used to transform a DataFrame column of arrays or maps into multiple rows, duplicating the other column's values. In this context, it will be used to expand the nested array column products in the transactions table so that each unique item in products for each row has its own row and the transaction_id column is duplicated as necessary.Reference:Databricks Documentation

I also noticed that you sent me an image along with your message. The image shows a snippet of SQL code that is incomplete. It begins with ''SELECT'' indicating a query to retrieve data. ''transaction_id,'' suggests that transaction_id is one of the columns being selected. There are blanks indicated by underscores where certain parts of the SQL command should be, including what appears to be an alias for a column and part of the FROM clause. The query ends with ''FROM transactions;'' indicating data is being selected from a 'transactions' table.

If you are interested in learning more about Databricks Data Analyst Associate certification, you can check out the following resources:

Databricks Certified Data Analyst Associate: This is the official page for the certification exam, where you can find the exam guide, registration details, and preparation tips.

Data Analysis With Databricks SQL: This is a self-paced course that covers the topics and skills required for the certification exam. You can access it for free on Databricks Academy.

Tips for the Databricks Certified Data Analyst Associate Certification: This is a blog post that provides some useful advice and study tips for passing the certification exam.

Databricks Certified Data Analyst Associate Certification: This is another blog post that gives an overview of the certification exam and its benefits.

Question 3

Question Type: MultipleChoice

Which of the following is a benefit of Databricks SQL using ANSI SQL as its standard SQL dialect?

Options:

A- It has increased customization capabilities
B- It is easy to migrate existing SQL queries to Databricks SQL
C- It allows for the use of Photon's computation optimizations
D- It is more performant than other SQL dialects
E- It is more compatible with Spark's interpreters

Answer:

B

Explanation:

Databricks SQL uses ANSI SQL as its standard SQL dialect, which means it follows the SQL specifications defined by the American National Standards Institute (ANSI). This makes it easier to migrate existing SQL queries from other data warehouses or platforms that also use ANSI SQL or a similar dialect, such as PostgreSQL, Oracle, or Teradata.By using ANSI SQL, Databricks SQL avoids surprises in behavior or unfamiliar syntax that may arise from using a non-standard SQL dialect, such as Spark SQL or Hive SQL12.Moreover, Databricks SQL also adds compatibility features to support common SQL constructs that are widely used in other data warehouses, such as QUALIFY, FILTER, and user-defined functions2.Reference:ANSI compliance in Databricks Runtime,Evolution of the SQL language at Databricks: ANSI standard by default and easier migrations from data warehouses

Question 4

Question Type: MultipleChoice

A data analyst creates a Databricks SQL Query where the result set has the following schema:

region STRING

number_of_customer INT

When the analyst clicks on the "Add visualization" button on the SQL Editor page, which of the following types of visualizations will be selected by default?

Options:

A- Violin Chart
B- Line Chart
C- IBar Chart
D- Histogram
E- There is no default. The user must choose a visualization type.

Answer:

C

Explanation:

According to the Databricks SQL documentation, when a data analyst clicks on the ''Add visualization'' button on the SQL Editor page, the default visualization type isBar Chart. This is because the result set has two columns: one of type STRING and one of type INT. The Bar Chart visualization automatically assigns the STRING column to the X-axis and the INT column to the Y-axis. The Bar Chart visualization is suitable for showing the distribution of a numeric variable across different categories.Reference:Visualization in Databricks SQL,Visualization types

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

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 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

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 9

Question Type: MultipleChoice

A data team has been given a series of projects by a consultant that need to be implemented in the Databricks Lakehouse Platform.

Which of the following projects should be completed in Databricks SQL?

Options:

A- Testing the quality of data as it is imported from a source
B- Tracking usage of feature variables for machine learning projects
C- Combining two data sources into a single, comprehensive dataset
D- Segmenting customers into like groups using a clustering algorithm
E- Automating complex notebook-based workflows with multiple tasks

Answer:

C

Explanation:

Databricks SQL is a service that allows users to query data in the lakehouse using SQL and create visualizations and dashboards1.One of the common use cases for Databricks SQL is to combine data from different sources and formats into a single, comprehensive dataset that can be used for further analysis or reporting2.For example, a data analyst can use Databricks SQL to join data from a CSV file and a Parquet file, or from a Delta table and a JDBC table, and create a new table or view that contains the combined data3. This can help simplify the data management and governance, as well as improve the data quality and consistency.Reference:

Databricks SQL overview

Databricks SQL use cases

Joining data sources

Question 10

Question Type: MultipleChoice

Which of the following statements about adding visual appeal to visualizations in the Visualization Editor is incorrect?

Options:

A- Visualization scale can be changed.
B- Data Labels can be formatted.
C- Colors can be changed.
D- Borders can be added.
E- Tooltips can be formatted.

Answer:

D

Explanation:

The Visualization Editor in Databricks SQL allows users to create and customize various types of charts and visualizations from the query results. Users can change the visualization type, select the data fields, adjust the colors, format the data labels, and modify the tooltips. However, there is no option to add borders to the visualizations in the Visualization Editor.Borders are not a supported feature of the new chart visualizations in Databricks1. Therefore, the statement that borders can be added is incorrect.Reference:

New chart visualizations in Databricks | Databricks on AWS

Question 11

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 ...