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

Shared by Maxwell on 19-01-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

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 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 been asked to count the number of customers in each region and has written the following query:

```
SELECT region, count(*) AS number_of_customers
    FROM customers
    ORDER BY region;
```

If there is a mistake in the query, which of the following describes the mistake?

Options:

A- The query is using count('). which will count all the customers in the customers table, no matter the region.
B- The query is missing a GROUP BY region clause.
C- The query is using ORDER BY. which is not allowed in an aggregation.
D- There are no mistakes in the query.
E- The query is selecting region but region should only occur in the ORDER BY clause.

Answer:

B

Explanation:

In the provided SQL query, the data analyst is trying to count the number of customers in each region. However, they made a mistake by not including the ''GROUP BY'' clause to group the results by region. Without this clause, the query will not return counts for each distinct region but rather an error or incorrect result.Reference: The need for a GROUP BY clause in such queries can be understood from Databricks SQL documentation:Databricks SQL.

I also noticed that you uploaded an image with your question. The image shows a snippet of an SQL query written in plain text on a white background. The query is attempting to select regions and count customers from a ''customers'' table and order the results by region. There's no visible syntax highlighting or any other color - it's monochromatic. The query is the same as the one in your question. I'm not sure why you included the image, but maybe you wanted to show me the exact format of your query. If so, you can also use code blocks to display formatted content such as SQL queries. For example, you can write:

SELECT region, count(*) AS number_of_customers

FROM customers

ORDER BY region;

This way, you can avoid uploading images and make your questions more clear and concise. I hope this helps.

Question 4

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 5

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 6

Question Type: MultipleChoice

Consider the following two statements:

Statement 1:

```
SELECT *
    FROM customers
    LEFT SEMI JOIN orders
    ON customers.customer_id = orders.customer_id;
```

Statement 2:

```
SELECT *
    FROM customers
    LEFT ANTI JOIN orders
    ON customers.customer_id = orders.customer_id;
```

Which of the following describes how the result sets will differ for each statement when they are run in Databricks SQL?

Options:

A- The first statement will return all data from the customers table and matching data from the orders table. The second statement will return all data from the orders table and matching data from the customers table. Any missing data will be filled in with NULL.
B- When the first statement is run, only rows from the customers table that have at least one match with the orders table on customer_id will be returned. When the second statement is run, only those rows in the customers table that do not have at least one match with the orders table on customer_id will be returned.
C- There is no difference between the result sets for both statements.
D- Both statements will fail because Databricks SQL does not support those join types.
E- When the first statement is run, all rows from the customers table will be returned and only the customer_id from the orders table will be returned. When the second statement is run, only those rows in the customers table that do not have at least one match with the orders table on customer_id will be returned.

Answer:

B

Explanation:

Based on the images you sent, the two statements are SQL queries for different types of joins between the customers and orders tables. A join is a way of combining the rows from two table references based on some criteria. The join type determines how the rows are matched and what kind of result set is returned. The first statement is a query for a LEFT SEMI JOIN, which returns only the rows from the left table reference (customers) that have a match with the right table reference (orders) on the join condition (customer_id). The second statement is a query for a LEFT ANTI JOIN, which returns only the rows from the left table reference (customers) that have no match with the right table reference (orders) on the join condition (customer_id). Therefore, the result sets for the two statements will differ in the following way:

The first statement will return a subset of the customers table that contains only the customers who have placed at least one order. The number of rows returned will be less than or equal to the number of rows in the customers table, depending on how many customers have orders. The number of columns returned will be the same as the number of columns in the customers table, as the LEFT SEMI JOIN does not include any columns from the orders table.

The second statement will return a subset of the customers table that contains only the customers who have not placed any order. The number of rows returned will be less than or equal to the number of rows in the customers table, depending on how many customers have no orders. The number of columns returned will be the same as the number of columns in the customers table, as the LEFT ANTI JOIN does not include any columns from the orders table.

The other options are not correct because:

A) The first statement will not return all data from the customers table, as it will exclude the customers who have no orders. The second statement will not return all data from the orders table, as it will exclude the orders that have a matching customer. Neither statement will fill in any missing data with NULL, as they do not return any columns from the other table.

C) There is a difference between the result sets for both statements, as explained above. The LEFT SEMI JOIN and the LEFT ANTI JOIN are not equivalent operations and will produce different outputs.

D) Both statements will not fail, as Databricks SQL does support those join types. Databricks SQL supports various join types, including INNER, LEFT OUTER, RIGHT OUTER, FULL OUTER, LEFT SEMI, LEFT ANTI, and CROSS. You can also use NATURAL, USING, or LATERAL keywords to specify different join criteria.

E) The first statement will not return only the customer_id from the orders table, as it will return all columns from the customers table. The second statement is correct, but it is not the only difference between the result sets.

Question 7

Question Type: MultipleChoice

A data analyst runs the following command:

INSERT INTO stakeholders.suppliers TABLE stakeholders.new_suppliers;

What is the result of running this command?

Options:

A- The suppliers table now contains both the data it had before the command was run and the data from the new suppliers table, and any duplicate data is deleted.
B- The command fails because it is written incorrectly.
C- The suppliers table now contains both the data it had before the command was run and the data from the new suppliers table, including any duplicate data.
D- The suppliers table now contains the data from the new suppliers table, and the new suppliers table now contains the data from the suppliers table.
E- The suppliers table now contains only the data from the new suppliers table.

Answer:

B

Explanation:

The commandINSERT INTO stakeholders.suppliers TABLE stakeholders.new_suppliersis not a valid syntax for inserting data into a table in Databricks SQL.According to the documentation12, the correct syntax for inserting data into a table is either:

INSERT { OVERWRITE | INTO } [TABLE] table_name [PARTITION clause] [(column_name [, ...]) | BY NAME] query

INSERT INTO [TABLE] table_name REPLACE WHERE predicate query

The command in the question is missing theOVERWRITEorINTOkeyword, and thequerypart that specifies the source of the data to be inserted. TheTABLEkeyword is optional and can be omitted. ThePARTITIONclause and the column list are also optional and depend on the table schema and the data source. Therefore, the command in the question will fail with a syntax error.

INSERT | Databricks on AWS

INSERT - Azure Databricks - Databricks SQL | Microsoft Learn

Question 8

Question Type: MultipleChoice

Which of the following statements about a refresh schedule is incorrect?

Options:

A- A query can be refreshed anywhere from 1 minute lo 2 weeks
B- Refresh schedules can be configured in the Query Editor.
C- A query being refreshed on a schedule does not use a SQL Warehouse (formerly known as SQL Endpoint).
D- A refresh schedule is not the same as an alert.
E- You must have workspace administrator privileges to configure a refresh schedule

Answer:

C

Explanation:

Refresh schedules are used to rerun queries at specified intervals, and these queries typically require computational resources to execute. In the context of a cloud data service like Databricks, this would typically involve the use of a SQL Warehouse (or a SQL Endpoint, as they were formerly known) to provide the necessary computational resources. Therefore, the statement is incorrect because scheduled query refreshes would indeed use a SQL Warehouse/Endpoint to execute the query.

Question 9

Question Type: MultipleChoice

A data analyst is working with gold-layer tables to complete an ad-hoc project. A stakeholder has provided the analyst with an additional dataset that can be used to augment the gold-layer tables already in use.

Which of the following terms is used to describe this data augmentation?

Options:

A- Data testing
B- Ad-hoc improvements
C- Last-mile
D- Last-mile ETL
E- Data enhancement

Answer:

E

Explanation:

Data enhancement is the process of adding or enriching data with additional information to improve its quality, accuracy, and usefulness. Data enhancement can be used to augment existing data sources with new data sources, such as external datasets, synthetic data, or machine learning models. Data enhancement can help data analysts to gain deeper insights, discover new patterns, and solve complex problems.Data enhancement is one of the applications of generative AI, which can leverage machine learning to generate synthetic data for better models or safer data sharing1.

In the context of the question, the data analyst is working with gold-layer tables, which are curated business-level tables that are typically organized in consumption-ready project-specific databases234.The gold-layer tables are the final layer of data transformations and data quality rules in the medallion lakehouse architecture, which is a data design pattern used to logically organize data in a lakehouse2. The stakeholder has provided the analyst with an additional dataset that can be used to augment the gold-layer tables already in use. This means that the analyst can use the additional dataset to enhance the existing gold-layer tables with more information, such as new features, attributes, or metrics. This data augmentation can help the analyst to complete the ad-hoc project more effectively and efficiently.

What is the medallion lakehouse architecture? - Databricks

Data Warehousing Modeling Techniques and Their Implementation on the Databricks Lakehouse Platform | Databricks Blog

What is the medallion lakehouse architecture? - Azure Databricks

What is a Medallion Architecture? - Databricks

Synthetic Data for Better Machine Learning | Databricks Blog

Question 10

Question Type: MultipleChoice

A data analyst created and is the owner of the managed table my_ table. They now want to change ownership of the table to a single other user using Data Explorer.

Which of the following approaches can the analyst use to complete the task?

Options:

A- Edit the Owner field in the table page by removing their own account
B- Edit the Owner field in the table page by selecting All Users
C- Edit the Owner field in the table page by selecting the new owner's account
D- Edit the Owner field in the table page by selecting the Admins group
E- Edit the Owner field in the table page by removing all access

Answer:

C

Explanation:

The Owner field in the table page shows the current owner of the table and allows the owner to change it to another user or group. To change the ownership of the table, the owner can click on the Owner field and select the new owner from the drop-down list.This will transfer the ownership of the table to the selected user or group and remove the previous owner from the list of table access control entries1. The other options are incorrect because:

A)Removing the owner's account from the Owner field will not change the ownership of the table, but will make the table ownerless2.

B)Selecting All Users from the Owner field will not change the ownership of the table, but will grant all users access to the table3.

D)Selecting the Admins group from the Owner field will not change the ownership of the table, but will grant the Admins group access to the table3.

E)Removing all access from the Owner field will not change the ownership of the table, but will revoke all access to the table4.Reference:

1: Change table ownership

2: Ownerless tables

3: Table access control

4: Revoke access to a table

Question 11

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