Databricks.Databricks-Certified-Data-Analyst-Associate.v2024-07-21.q15

Exam Code:	Databricks-Certified-Data-Analyst-Associate
Exam Name:	Databricks Certified Data Analyst Associate Exam
Certification Provider:	Databricks
Free Question Number:	15
Version:	v2024-07-21
# of views:	188
# of Questions views:	150
https://www.freeqas.com/qa/Databricks/Databricks-Certified-Data-Analyst-Associate/Databricks.Databricks-Certified-Data-Analyst-Associate.v2024-07-21.q15.html	

NEW QUESTION: 1
Which of the following approaches can be used to ingest data directly from cloud-based object storage?
A. It is not possible to directly ingest data from cloud-based object storage
B. Create an external table while specifying the object storage path to LOCATION
C. Create an external table while specifying the DBFS storage path to FROM
D. Create an external table while specifying the DBFS storage path to PATH
E. Create an external table while specifying the object storage path to FROM
Answer: B (LEAVE A REPLY)
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.

NEW QUESTION: 2

A data analyst runs the following command:
SELECT age, country
FROM my_table
WHERE age >= 75 AND country = 'canada';
Which of the following tables represents the output of the above command?

A.

age	country
80	canada
NULL	canada
90	*NULL*

B.

age	country
80	*NULL*
75	*NULL*
90	*NULL*

C.

id	age	country
900	80	canada
901	75	canada
902	90	canada

D.

age	country
80	canada
14	canada
90	canada

E.

age	country
80	canada
75	canada
90	canada

Answer: E (LEAVE A REPLY)
The SQL query provided is designed to filter out records from "my_table" where the age is 75 or above and the country is Canada. Since I can't view the content of the links provided directly, I need to rely on the image attached to this question for context. Based on that, Option E (the image attached) represents a table with columns "age" and "country", showing records where age is 75 or above and country is Canada. Reference: The answer can be inferred from understanding SQL queries and their outputs as per Databricks documentation: Databricks SQL

NEW QUESTION: 3

Which of the following approaches can be used to connect Databricks to Fivetran for data ingestion?

A. Use Workflows to establish a SQL warehouse (formerly known as a SQL endpoint) for Fivetran to interact with

B. Use Delta Live Tables to establish a cluster for Fivetran to interact with

C. Use Partner Connect's automated workflow to establish a cluster for Fivetran to interact with

D. Use Partner Connect's automated workflow to establish a SQL warehouse (formerly known as a SQL endpoint) for Fivetran to interact with

E. Use Workflows to establish a cluster for Fivetran to interact with

Answer: (SHOW ANSWER)

Partner Connect is a feature that allows you to easily connect your Databricks workspace to Fivetran and other ingestion partners using an automated workflow. You can select a SQL warehouse or a cluster as the destination for your data replication, and the connection details are sent to Fivetran. You can then choose from over 200 data sources that Fivetran supports and start ingesting data into Delta Lake. Reference: Connect to Fivetran using Partner Connect, Use Databricks with Fivetran

NEW QUESTION: 4

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;
```

Answer: B (LEAVE A REPLY)

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.

NEW QUESTION: 5
In which of the following situations should a data analyst use higher-order functions?
A. When custom logic needs to be applied to simple, unnested data
B. When custom logic needs to be converted to Python-native code
C. When custom logic needs to be applied at scale to array data objects
D. When built-in functions are taking too long to perform tasks
E. When built-in functions need to run through the Catalyst Optimizer
Answer: C (LEAVE A REPLY)
Higher-order functions are a simple extension to SQL to manipulate nested data such as arrays. A higher-order function takes an array, implements how the array is processed, and what the result of the computation will be. It delegates to a lambda function how to process each item in the array. This allows you to define functions that manipulate arrays in SQL, without having to unpack and repack them, use UDFs, or rely on limited built-in functions. Higher-order functions provide a performance benefit over user defined functions. Reference: Higher-order functions | Databricks on AWS, Working with Nested Data Using Higher Order Functions in SQL on Databricks | Databricks Blog, Higher-order functions - Azure Databricks | Microsoft Learn, Optimization recommendations on Databricks | Databricks on AWS

NEW QUESTION: 6
Which of the following is an advantage of using a Delta Lake-based data lakehouse over common data lake solutions?
A. ACID transactions
B. Flexible schemas
C. Data deletion
D. Scalable storage
E. Open-source formats
Answer: A (LEAVE A REPLY)
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

NEW QUESTION: 7
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?
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 (LEAVE A REPLY)
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.

NEW QUESTION: 8

A data analysis team is working with the table_bronze SQL table as a source for one of its most complex projects. A stakeholder of the project notices that some of the downstream data is duplicative. The analysis team identifies table_bronze as the source of the duplication.
Which of the following queries can be used to deduplicate the data from table_bronze and write it to a new table table_silver?
A)
CREATE TABLE table_silver AS
SELECT DISTINCT *
FROM table_bronze;
B)
CREATE TABLE table_silver AS
INSERT *
FROM table_bronze;
C)
CREATE TABLE table_silver AS
MERGE DEDUPLICATE *
FROM table_bronze;
D)
INSERT INTO TABLE table_silver
SELECT * FROM table_bronze;
E)
INSERT OVERWRITE TABLE table_silver
SELECT * FROM table_bronze;
A. Option A
B. Option B
C. Option C
D. Option D
E. Option E
Answer: (SHOW ANSWER)
Option A uses the SELECT DISTINCT statement to remove duplicate rows from the table_bronze and create a new table table_silver with the deduplicated data. This is the correct way to deduplicate data using Spark SQL12. Option B simply inserts all the rows from table_bronze into table_silver, without removing any duplicates. Option C is not a valid syntax for Spark SQL, as there is no MERGE DEDUPLICATE statement. Option D appends all the rows from table_bronze into table_silver, without removing any duplicates. Option E overwrites the existing data in table_silver with the data from table_bronze, without removing any duplicates. Reference: Delete Duplicate using SPARK SQL, Spark SQL - How to Remove Duplicate Rows

NEW QUESTION: 9
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?

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: (SHOW ANSWER)

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

NEW QUESTION: 10

A data analyst has been asked to provide a list of options on how to share a dashboard with a client. It is a security requirement that the client does not gain access to any other information, resources, or artifacts in the database.

Which of the following approaches cannot be used to share the dashboard and meet the security requirement?

A. Download the Dashboard as a PDF and share it with the client.

B. Set a refresh schedule for the dashboard and enter the client's email address in the "Subscribers" box.

C. Take a screenshot of the dashboard and share it with the client.

D. Generate a Personal Access Token that is good for 1 day and share it with the client.

E. Download a PNG file of the visualizations in the dashboard and share them with the client.

Answer: (SHOW ANSWER)

The approach that cannot be used to share the dashboard and meet the security requirement is D. Generating a Personal Access Token that is good for 1 day and sharing it with the client. This approach would give the client access to the Databricks workspace using the token owner's

identity and permissions, which could expose other information, resources, or artifacts in the database1. The other approaches can be used to share the dashboard and meet the security requirement because:
A) Downloading the Dashboard as a PDF and sharing it with the client would only provide a static snapshot of the dashboard without any interactive features or access to the underlying data2.
B) Setting a refresh schedule for the dashboard and entering the client's email address in the "Subscribers" box would send the client an email with the latest dashboard results as an attachment or a link to a secure web page3. The client would not be able to access the Databricks workspace or the dashboard itself.
C) Taking a screenshot of the dashboard and sharing it with the client would also only provide a static snapshot of the dashboard without any interactive features or access to the underlying data4.
E) Downloading a PNG file of the visualizations in the dashboard and sharing them with the client would also only provide a static snapshot of the visualizations without any interactive features or access to the underlying data5. Reference:
1: Personal access tokens
2: Download as PDF
3: Automatically refresh a dashboard
4: Take a screenshot
5: Download a PNG file

NEW QUESTION: 11
Which of the following statements about a refresh schedule is incorrect?
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 (LEAVE A REPLY)
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.

NEW QUESTION: 12
A data analyst has recently joined a new team that uses Databricks SQL, but the analyst has never used Databricks before. The analyst wants to know where in Databricks SQL they can write and execute SQL queries.

On which of the following pages can the analyst write and execute SQL queries?

A. Data page

B. Dashboards page

C. Queries page

D. Alerts page

E. SQL Editor page

Answer: E (LEAVE A REPLY)

The SQL Editor page is where the analyst can write and execute SQL queries in Databricks SQL. The SQL Editor page has a query pane where the analyst can type or paste SQL statements, and a results pane where the analyst can view the query results in a table or a chart. The analyst can also browse data objects, edit multiple queries, execute a single query or multiple queries, terminate a query, save a query, download a query result, and more from the SQL Editor page.

Reference: Create a query in SQL editor

NEW QUESTION: 13

A data analyst has been asked to use the below table sales_table to get the percentage rank of products within region by the sales:

region	product	sales
WEST	A	1880.59
EAST	A	2045.99
EAST	B	4583.23
WEST	B	3391.19

The result of the query should look like this:

region	product	sales
EAST	B	0
EAST	A	1
WEST	B	0
WEST	A	1

Which of the following queries will accomplish this task?

A)

```
SELECT
    region,
    product,
    RANK() OVER (
        PARTITION BY region
        ORDER BY sales DESC
    ) AS rank
FROM sales_table;
GROUP BY region, product;
```

B)

```
SELECT
    region,
    product,
    PERCENT_RANK () OVER (
        PARTITION BY region
        ORDER BY sales DESC
    ) AS rank
FROM sales_table;
GROUP BY region, product;
```

C)

```
SELECT
    region,
    product,
    PERCENT_RANK () OVER (
        ORDER BY sales DESC
    ) AS rank
FROM sales_table;
```

```
SELECT
    region,
    product,
    PERCENT_RANK () OVER (
        PARTITION BY product
        ORDER BY sales DESC
    ) AS rank
FROM sales_table;
GROUP BY region, product;
```

A. Option A

B. Option B

C. Option C

D. Option D

Answer: B (LEAVE A REPLY)

The correct query to get the percentage rank of products within region by the sales is option B. This query uses the PERCENT_RANK() window function to calculate the relative rank of each product within each region based on the sales amount. The window function is partitioned by region and ordered by sales in descending order. The result is aliased as rank and displayed along with the region and product columns. The other options are incorrect because:

A) Option A uses the RANK() window function instead of the PERCENT_RANK() function. The RANK() function returns the rank of each row within the partition, but not the percentage rank. Also, the query does not have a GROUP BY clause, which is required for aggregate functions like SUM().

C) Option C uses the DENSE_RANK() window function instead of the PERCENT_RANK() function. The DENSE_RANK() function returns the rank of each row within the partition, but not the percentage rank. Also, the query does not have a GROUP BY clause, which is required for aggregate functions like SUM().

D) Option D uses the ROW_NUMBER() window function instead of the PERCENT_RANK() function. The ROW_NUMBER() function returns the sequential number of each row within the partition, but not the percentage rank. Also, the query does not have a GROUP BY clause, which is required for aggregate functions like SUM(). Reference:

1: PERCENT_RANK (Transact-SQL)

2: Window functions in Databricks SQL

3: Databricks Certified Data Analyst Associate Exam Guide

NEW QUESTION: 14

Which of the following describes how Databricks SQL should be used in relation to other business intelligence (BI) tools like Tableau, Power BI, and looker?

A. As an exact substitute with the same level of functionality

B. As a substitute with less functionality

C. As a complete replacement with additional functionality

D. As a complementary tool for professional-grade presentations

E. As a complementary tool for quick in-platform BI work

Answer: (SHOW ANSWER)

Databricks SQL is not meant to replace or substitute other BI tools, but rather to complement them by providing a fast and easy way to query, explore, and visualize data on the lakehouse using the built-in SQL editor, visualizations, and dashboards. Databricks SQL also integrates seamlessly with popular BI tools like Tableau, Power BI, and Looker, allowing analysts to use their preferred tools to access data through Databricks clusters and SQL warehouses. Databricks SQL offers low-code and no-code experiences, as well as optimized connectors and serverless compute, to enhance the productivity and performance of BI workloads on the lakehouse. Reference: Databricks SQL, Connecting Applications and BI Tools to Databricks SQL, Databricks

integrations overview, Databricks SQL: Delivering a Production SQL Development Experience on the Lakehouse

NEW QUESTION: 15
Which of the following statements about adding visual appeal to visualizations in the Visualization Editor is incorrect?
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: (SHOW ANSWER)
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