

Databricks-Certified-Data-Analyst-Associate Dumps

Databricks Certified Data Analyst Associate Exam

<https://www.certleader.com/Databricks-Certified-Data-Analyst-Associate-dumps.html>



NEW QUESTION 1

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?

- A. The first statement will return all data from the customers table and matching data from the orders table.
- B. The second statement will return all data from the orders table and matching data from the customers table.
- C. Any missing data will be filled in with NULL.
- D. 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.
- E. 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.
- F. There is no difference between the result sets for both statements.
- G. Both statements will fail because Databricks SQL does not support those join types.
- H. 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.
- I. 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.

References: JOIN | Databricks on AWS, JOIN - Azure Databricks - Databricks SQL | Microsoft Learn, array_join function | Databricks on AWS, Hints | Databricks on AWS

NEW QUESTION 2

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?

- A. The area chart will use whatever is selected in the Dashboard Parameter while all of the other visualizations will remain unchanged 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 dashboar
E. The parameter cannot be changed by the user afterwards.
F. 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. References: Databricks SQL dashboards, Query parameters

NEW QUESTION 3

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?

- 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. References: Databricks - DROP TABLE

NEW QUESTION 4

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: C

Explanation:

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. References: Connect to Fivetran using Partner Connect, Use Databricks with Fivetran

NEW QUESTION 5

After runningDESCRIBE 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?

- A. Running SELECT * FROM delt
B. `dbfs:/stakeholders/customers` results in an error.
C. Running SELECT * FROM accounts.customers will return all rows in the table.
D. All files with the .customers extension are deleted.
E. The accounts.customers table is removed from the metastore, and the underlying data files are deleted.
F. 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. References: DROP TABLE | Databricks on AWS, Best practices for dropping a managed Delta Lake table - Databricks

NEW QUESTION 6

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

- 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 tools.

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. References: Discover and manage data using Data Explorer

NEW QUESTION 7

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?

- A. Violin Chart
- B. Line Chart
- C. IBar Chart
- D. Histogram
- E. There is no default
- F. 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 is Bar 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. References: Visualization in Databricks SQL, Visualization types

NEW QUESTION 8

A data analyst has been asked to use the below 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;
```

D)

```
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

Explanation:

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(). References:

? 1: PERCENT_RANK (Transact-SQL)

? 2: Window functions in Databricks SQL

? 3: Databricks Certified Data Analyst Associate Exam Guide

NEW QUESTION 9

Data professionals with varying titles use the Databricks SQL service as the primary touchpoint with the Databricks Lakehouse Platform. However, some users will use other services like Databricks Machine Learning or Databricks Data Science and Engineering.

Which of the following roles uses Databricks SQL as a secondary service while primarily using one of the other services?

- A. Business analyst
- B. SQL analyst
- C. Data engineer
- D. Business intelligence analyst
- E. Data analyst

Answer: C

Explanation:

Data engineers are primarily responsible for building, managing, and optimizing data pipelines and architectures. They use Databricks Data Science and Engineering service to perform tasks such as data ingestion, transformation, quality, and governance. Data engineers may use Databricks SQL as a secondary service to query, analyze, and visualize data from the lakehouse, but this is not their main

focus. References: Databricks SQL overview, Databricks Data Science and Engineering overview, Data engineering with Databricks

NEW QUESTION 10

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?

- 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. References: Databricks SQL endpoints, Serverless SQL endpoints, SQL endpoint clusters

NEW QUESTION 10

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?

- 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. References:

? Databricks SQL overview

? Databricks SQL use cases

? Joining data sources

NEW QUESTION 13

In which of the following situations will the mean value and median value of variable be meaningfully different?

- A. When the variable contains no outliers
- B. When the variable contains no missing values
- C. When the variable is of the boolean type
- D. When the variable is of the categorical type
- E. When the variable contains a lot of extreme outliers

Answer: E

Explanation:

The mean value of a variable is the average of all the values in a data set, calculated by dividing the sum of the values by the number of values. The median value of a variable is the middle value of the ordered data set, or the average of the middle two values if the data set has an even number of values. The mean value is sensitive to outliers, which are values that are very different from the rest of the data. Outliers can skew the mean value and make it less representative of the central tendency of the data. The median value is more robust to outliers, as it only depends on the middle values of the data. Therefore, when the variable contains a lot of extreme outliers, the mean value and the median value will be meaningfully different, as the mean value will be pulled towards the outliers, while the median value will remain close to the majority of the data¹. References: Difference Between Mean and Median in Statistics (With Example) - BYJU??S

NEW QUESTION 14

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?

- 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. References: DROP TABLE, Drop Delta table features, Best practices for dropping a managed Delta Lake table

NEW QUESTION 17

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

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. References:

? 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 20

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?

- 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.

References:

? 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

NEW QUESTION 21

.....

Thank You for Trying Our Product

* 100% Pass or Money Back

All our products come with a 90-day Money Back Guarantee.

* One year free update

You can enjoy free update one year. 24x7 online support.

* Trusted by Millions

We currently serve more than 30,000,000 customers.

* Shop Securely

All transactions are protected by VeriSign!

100% Pass Your Databricks-Certified-Data-Analyst-Associate Exam with Our Prep Materials Via below:

<https://www.certleader.com/Databricks-Certified-Data-Analyst-Associate-dumps.html>