



Download Free Databricks-Certified-Data-Analyst-Associate Exam PDF | PrepBolt

Don't miss out! Download the latest free Databricks Certified Data Analyst Associate Exam PDF questions. Access real Databricks-Certified-Data-Analyst-Associate dumps with verified answers and boost your chances to pass your certification on the first try with [PrepBolt](#) Databricks-Certified-Data-Analyst-Associate exam pdf questions and answers.

Thank you for Downloading Databricks-Certified-Data-Analyst-Associate exam PDF Demo

<https://prepbolt.com/Databricks-Certified-Data-Analyst-Associate.html>

QUESTIONS & ANSWERS
DEMO VERSION
(LIMITED CONTENT)

Question 1

Question Type: MultipleChoice

A stakeholder has provided a data analyst with a lookup dataset in the form of a 50-row CSV file. The data analyst needs to upload this dataset for use as a table in Databricks SQL.

Which approach should the data analyst use to quickly upload the file into a table for use in Databricks SQL?

Options:

- A- Create a table by uploading the file using the Create page within Databricks SQL
- B- Create a table via a connection between Databricks and the desktop facilitated by Partner Connect.
- C- Create a table by uploading the file to cloud storage and then importing the data to Databricks.
- D- Create a table by manually copying and pasting the data values into cloud storage and then importing the data to Databricks.

Databricks provides a user-friendly interface that allows data analysts to quickly upload small datasets, such as a 50-row CSV file, and create tables within Databricks SQL. The steps are as follows:

Access the Data Upload Interface:

In the Databricks workspace, navigate to the sidebar and click on New > Add or upload data.

Select Create or modify a table.

Upload the CSV File:

Click on the browse button or drag and drop the CSV file directly onto the designated area.

The interface supports uploading up to 10 files simultaneously, with a total size limit of 2 GB.

Configure Table Settings:

After uploading, a preview of the data is displayed.

Specify the table name, select the appropriate schema, and configure any additional settings as needed.

Create the Table:

Once all configurations are set, click on the Create Table button to finalize the process.

This method is efficient for quickly importing small datasets without the need for additional tools or complex configurations. Options B, C, and D involve more complex or manual processes that are unnecessary for this task.

Answer:

A

Question 2

Question Type: MultipleChoice

Query History provides Databricks SQL users with a lot of benefits. A data analyst has been asked to share all of these benefits with their team as part of a training exercise. One of the benefit statements the analyst provided to their team is incorrect.

Which statement about Query History is incorrect?

Options:

- A- It can be used to view the query plan of queries that have run.
- B- It can be used to debug queries.
- C- It can be used to automate query execution on multiple warehouses (formerly endpoints).
- D- It can be used to troubleshoot slow running queries.

Answer:

C

Explanation:

Query History in Databricks SQL is intended for reviewing executed queries, understanding their execution plans, and identifying performance issues or errors for debugging purposes. It allows users to analyze query duration, resources used, and potential bottlenecks. However, Query History does not provide any capability to automate the execution of queries across multiple warehouses; automation must be handled through jobs or external orchestration tools, not through the Query History feature itself.

Question 3

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.

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.

Answer:

B

Question 4

Question Type: MultipleChoice

Which statement about subqueries is correct?

Options:

A- Subqueries are not available in Databricks SQL

B- Subqueries can be used like other user-defined functions to transform data into different data types.

C- Subqueries can retrieve data without requiring the creation of a table or view.

D- Subqueries can be used like other built-in functions to transform data into different data types.

In Databricks SQL, a subquery is a nested query within a larger SQL query that allows for the retrieval of data without the necessity of creating a table or view. This is particularly useful for simplifying complex queries by breaking them down into more manageable parts. Subqueries can be employed in various clauses such as SELECT, FROM, and WHERE to perform operations like filtering, transforming, and aggregating data on-the-fly. This flexibility enhances query efficiency and readability without the overhead of persisting intermediate results as separate tables or views.

Answer:

C

Question 5

Question Type: MultipleChoice

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

Options:

- 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

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 data. Reference: Difference Between Mean and Median in Statistics (With Example) - BYJU'S

Answer:

E

Question 6

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

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

Answer:

A

Question 7

Question Type: MultipleChoice

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

Which step will the data analyst need to take when creating and adding both data visualizations to the Databricks SQL Dashboard?

Options:

- A- Copy the Query and create one data visualization per query.
- B- Add two separate visualizations to the dashboard based on the same Query.
- C- Decide on a single data visualization to add to the dashboard.
- D- Alter the Query to return two separate sets of results.

Answer:

B

Question 8

Question Type: MultipleChoice

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?

Options:

- 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

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. Reference: Databricks SQL endpoints, Serverless SQL endpoints, SQL endpoint clusters

Answer:

E

Question 9

Question Type: MultipleChoice

What does Partner Connect do when connecting Power BI and Tableau?

Options:

- A- Creates a Personal Access Token, downloads and installs an ODBC driver, and downloads a configuration file for connection by Power BI or Tableau to a SQL Warehouse (formerly known as a SQL Endpoint).
- B- Creates a Personal Access Token for authentication into Databricks SQL and emails it to you.
- C- Downloads a configuration file for connection by Power BI or Tableau to a SQL Warehouse (formerly known as a SQL Endpoint).
- D- Downloads and installs an ODBC driver.

When connecting Power BI and Tableau through Databricks Partner Connect, the system automates several steps to streamline the integration process:

Personal Access Token Creation: Partner Connect generates a Databricks personal access token, which is essential for authenticating and establishing a secure connection between Databricks and the BI tools.

ODBC Driver Installation: The appropriate ODBC driver is downloaded and installed. This driver facilitates communication between the BI tools and Databricks, ensuring compatibility and optimal performance.

Configuration File Download: A configuration file tailored for the selected BI tool (Power BI or Tableau) is provided. This file contains the necessary connection details, simplifying the setup process within the BI tool.

By automating these steps, Partner Connect ensures a seamless and efficient integration, reducing manual configuration efforts and potential errors.

Answer:

A

Thank You for trying Databricks-Certified-Data-Analyst-Associate PDF Demo

To try our Databricks-Certified-Data-Analyst-Associate practice exam software visit link below

<https://prepbolt.com/Databricks-Certified-Data-Analyst-Associate.html>

Start Your Databricks-Certified-Data-Analyst-Associate Preparation

Use Coupon “**SAVE50**” for extra 50% discount on the purchase of Practice Test Software. Test your Databricks-Certified-Data-Analyst-Associate preparation with actual exam questions.