Databricks

Exam Questions Databricks-Certified-Data-Analyst-Associate

Databricks Certified Data Analyst Associate Exam

About Exambible

Your Partner of IT Exam

Found in 1998

Exambible is a company specialized on providing high quality IT exam practice study materials, especially Cisco CCNA, CCDA, CCNP, CCIE, Checkpoint CCSE, CompTIA A+, Network+ certification practice exams and so on. We guarantee that the candidates will not only pass any IT exam at the first attempt but also get profound understanding about the certificates they have got. There are so many alike companies in this industry, however, Exambible has its unique advantages that other companies could not achieve.

Our Advances

* 99.9% Uptime

All examinations will be up to date.

* 24/7 Quality Support

We will provide service round the clock.

* 100% Pass Rate

Our guarantee that you will pass the exam.

* Unique Gurantee

If you do not pass the exam at the first time, we will not only arrange FULL REFUND for you, but also provide you another exam of your claim, ABSOLUTELY FREE!

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 tabl
B. The second statement will return all data from the orders table and matching data from the customers tabl
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 returne
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 returne
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
A data engineering team has created a Structured Streaming pipeline that processes data in micro-batches and populates gold-level tables. The microbatches are triggered every minute.
A data analyst has created a dashboard based on this gold-level data. The project stakeholders want to see the results in the dashboard updated within one minute or less of new data becoming available within the gold-level tables.
Which of the following cautions should the data analyst share prior to setting up the dashboard to complete this task?

A. The required compute resources could be costly
B. The gold-level tables are not appropriately clean for business reporting
C. The streaming data is not an appropriate data source for a dashboard

D. The streaming cluster is not fault tolerant
E. The dashboard cannot be refreshed that quickly

Answer: A

Explanation:
A Structured Streaming pipeline that processes data in micro-batches and populates gold-level tables every minute requires a high level of compute resources to handle the frequent data ingestion, processing, and writing. This could result in a significant cost for the organization, especially if the data volume and velocity are large. Therefore, the data analyst should share this caution with the project stakeholders before setting up the dashboard and evaluate the trade-offs between the desired refresh rate and the available budget. The other options are not valid cautions because:
? B. The gold-level tables are assumed to be appropriately clean for business reporting, as they are the final output of the data engineering pipeline. If the data quality is notsatisfactory, the issue should be addressed at the source or silver level, not at the gold level.
? C. The streaming data is an appropriate data source for a dashboard, as it can provide near real-time insights and analytics for the business users. Structured Streaming supports various sources and sinks for streaming data, including Delta Lake, which can enable both batch and streaming queries on the same data.
? D. The streaming cluster is fault tolerant, as Structured Streaming provides end-to-end exactly-once fault-tolerance guarantees through checkpointing and write-ahead logs. If a query fails, it can be restarted from the last checkpoint and resume processing.
? E. The dashboard can be refreshed within one minute or less of new data becoming available in the gold-level tables, as Structured Streaming can trigger micro-batches as fast as possible (every few seconds) and update the results incrementally. However, this may not be necessary or optimal for the business use case, as it could cause frequent changes in the dashboard and consume more resources. References: Streaming on Databricks, Monitoring Structured Streaming queries on Databricks, A look at the new Structured Streaming UI in Apache Spark 3.0, Run your first Structured Streaming workload

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: 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 4
A data analyst has been asked to produce a visualization that shows the flow of users through a website.
Which of the following is used for visualizing this type of flow?

A. Heatmap
B. IChoropleth
C. Word Cloud
D. Pivot Table
E. Sankey

Answer: E

Explanation:
A Sankey diagram is a type of visualization that shows the flow of data between different nodes or categories. It is often used to represent the movement of users through a website, as it can show the paths they take, the sources they come from, the pages they visit, and the outcomes they achieve. A Sankey diagram consists of links and nodes, where the links represent the volume or weight of the flow, and the nodes represent the stages or steps of the flow. The width of the links is proportional to the amount of flow, and the color of the links can indicate different attributes or segments of the flow. A Sankey diagram can help identify the most common or popular user journeys, the bottlenecks or drop-offs in the flow, and the opportunities for improvement or optimization. References: The answer can be verified from Databricks documentation which provides examples and instructions on how to create Sankey diagrams using Databricks SQL Analytics and Databricks Visualizations. Reference links: Databricks SQL Analytics - Sankey Diagram, Databricks Visualizations - Sankey Diagram

NEW QUESTION 5
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 B
D. which is not allowed in an aggregation.
E. There are no mistakes in the query.
F. 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. References: 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 whyyou 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 6
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 7
A data analyst has created a user-defined function using the following line of code: CREATE FUNCTION price(spend DOUBLE, units DOUBLE)
RETURNS DOUBLE
RETURN spend / units;
Which of the following code blocks can be used to apply this function to the customer_spend and customer_units columns of the table customer_summary to create column customer_price?

A. SELECT PRICE customer_spend, customer_units AS customer_price FROM customer_summary
B. SELECT price FROM customer_summary
C. SELECT function(price(customer_spend, customer_units)) AS customer_price FROM customer_summary
D. SELECT double(price(customer_spend, customer_units)) AS customer_price FROM customer_summary
E. SELECT price(customer_spend, customer_units) AS customer_price FROM customer_summary

Answer: E

Explanation:
A user-defined function (UDF) is a function defined by a user, allowing custom logic to be reused in the user environment1. To apply a UDF to a table, the syntax is SELECT udf_name(column_name) AS alias FROM table_name2. Therefore, option E is
the correct way to use the UDF price to create a new column customer_price based on the existing columns customer_spend and customer_units from the table customer_summary. References:
? What are user-defined functions (UDFs)?
? User-defined scalar functions - SQL V

NEW QUESTION 8
Which of the following statements describes descriptive statistics?

A. A branch of statistics that uses summary statistics to quantitatively describe and summarize data.
B. A branch of statistics that uses a variety of data analysis techniques to infer properties of an underlying distribution of probability.
C. A branch of statistics that uses quantitative variables that must take on a finite or countably infinite set of values.
D. A branch of statistics that uses summary statistics to categorically describe and summarize data.
E. A branch of statistics that uses quantitative variables that must take on an uncountable set of values.

Answer: A

Explanation:
Descriptive statistics is a branch of statistics that uses summary statistics, such as mean, median, mode, standard deviation, range, frequency, or correlation, to quantitatively describe and summarize data. Descriptive statistics can help data analysts understand the main features of a data set, such as its central tendency, variability, or distribution. Descriptive statistics can also help data analysts visualize data using charts, graphs, or tables. Descriptive statistics do not make any inferences or predictions about the data, unlike inferential statistics, which use data analysis techniques to infer properties of an underlying population or probability distribution from a sample of
data. References: Databricks - Descriptive Statistics, Databricks - Data Analysis with Databricks SQL

NEW QUESTION 9
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

A. Option A
B. Option B
C. Option C
D. Option D
E. Option E

Answer: E

Explanation:
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. References: The answer can be inferred from understanding SQL queries and their outputs as per Databricks documentation: Databricks SQL

NEW QUESTION 10
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: E

Explanation:
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. References: 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 10
.....

Relate Links

100% Pass Your Databricks-Certified-Data-Analyst-Associate Exam with Exambible Prep Materials

https://www.exambible.com/Databricks-Certified-Data-Analyst-Associate-exam/

Contact us

We are proud of our high-quality customer service, which serves you around the clock 24/7.

Viste - https://www.exambible.com/