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 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 3
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 4
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?

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. References: Dashboards in notebooks, How to add text to a dashboard in Databricks

NEW QUESTION 5
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: D

Explanation:
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. References:
? New chart visualizations in Databricks | Databricks on AWS

NEW QUESTION 6

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 7
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 8
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

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.

NEW QUESTION 9
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 verydifferent 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 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 Bl 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
Which of the following layers of the medallion architecture is most commonly used by data analysts?

A. None of these layers are used by data analysts
B. Gold
C. All of these layers are used equally by data analysts
D. Silver
E. Bronze

Answer: B

Explanation:
The gold layer of the medallion architecture contains data that is highly refined and aggregated, and powers analytics, machine learning, and production applications. Data analysts typically use the gold layer to access data that has been transformed into knowledge, rather than just information. The gold layer represents the final stage of data quality and optimization in the lakehouse. References: What is the medallion lakehouse architecture?

NEW QUESTION 11
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: D

Explanation:
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. References:
? 1: Personal access tokens
? 2: Download as PDF
? 3: Automatically refresh a dashboard
? 4: Take a screenshot
? 5: Download a PNG file

NEW QUESTION 16
.....

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/