



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

A data organization has a team of engineers developing data pipelines following the medallion architecture using Delta Live Tables. While the data analysis team working on a project is using gold-layer tables from these pipelines, they need to perform some additional processing of these tables prior to performing their analysis.

Which of the following terms is used to describe this type of work?

- A. Data blending
- B. Last-mile
- C. Data testing
- D. Last-mile ETL
- E. Data enhancement

Answer: D

Explanation:

Last-mile ETL is the term used to describe the additional processing of data that is done by data analysts or data scientists after the data has been ingested, transformed, and stored in the lakehouse by data engineers. Last-mile ETL typically involves tasks such as data cleansing, dataenrichment, data aggregation, data filtering, or data sampling that are specific to the analysis or machine learning use case. Last-mile ETL can be done using Databricks SQL, Databricks notebooks, or Databricks Machine Learning. References: Databricks - Last-mile ETL, Databricks - Data Analysis with Databricks SQL

NEW QUESTION 2

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

Which of the following steps will they need to take when creating and adding both data visualizations to the Databricks SQL Dashboard?

- A. They will need to alter the Query to return two separate sets of results.
- B. They will need to add two separate visualizations to the dashboard based on the same Query.
- C. They will need to create two separate dashboards.
- D. They will need to decide on a single data visualization to add to the dashboard.
- E. They will need to copy the Query and create one data visualization per query.

Answer: B

Explanation:

A data analyst can create multiple visualizations from the same query in Databricks SQL by clicking the + button next to the Results tab and selecting Visualization. Each visualization can have a different type, name, and configuration. To add a visualization to a dashboard, the data analyst can click the vertical ellipsis button beneath the visualization, select + Add to Dashboard, and choose an existing or new dashboard. The data analyst can repeat this process for each visualization they want to add to the same dashboard. References: Visualization in Databricks SQL, Visualize queries and create a dashboard in Databricks SQL

NEW QUESTION 3

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

Explanation:

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. References: Create a query in SQL editor

NEW QUESTION 4

After running `DESCRIBE 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 5

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 6

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 7

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: C

Explanation:

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. References:
- ? 1: Change table ownership
 - ? 2: Ownerless tables
 - ? 3: Table access control
 - ? 4: Revoke access to a table

NEW QUESTION 8

A data analyst has been asked to use the below tablesales_tableto 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

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 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;
```

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

Answer: B**Explanation:**

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 10

Which of the following statements about a refresh schedule is incorrect?

- A. A query can be refreshed anywhere from 1 minute to 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 13

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 15

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 19

.....

Relate Links

100% Pass Your Databricks-Certified-Data-Analyst-Associate Exam with Examible 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/>