Free Questions for Databricks-Certified-Data-Analyst-Associate

Shared by Burks on 22-07-2024

For More Free Questions and Preparation Resources

Check the Links on Last Page

Question 1

Question Type: MultipleChoice

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?

Options:

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.

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

Question 2

Question Type: MultipleChoice

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?

Options:

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 isSELECT udf_name(column_name) AS alias FROM table_name2. Therefore, option E is the correct way to use the UDFpriceto create a new columncustomer_pricebased on the existing

columnscustomer_spendandcustomer_unitsfrom the tablecustomer_summary.Reference:

What are user-defined functions (UDFs)?

User-defined scalar functions - SQL

V

Question 3

Question Type: MultipleChoice

A data analysis team is working with the table_bronze SQL table as a source for one of its most complex projects. A stakeholder of the project notices that some of the downstream data is duplicative. The analysis team identifies table_bronze as the source of the duplication.

Which of the following queries can be used to deduplicate the data from table_bronze and write it to a new table table_silver?

A)

CREATE TABLE table_silver AS

SELECT DISTINCT *

FROM table_bronze;

B)

CREATE TABLE table_silver AS

INSERT *

FROM table_bronze;

C)

CREATE TABLE table_silver AS

MERGE DEDUPLICATE *

FROM table_bronze;

D)

INSERT INTO TABLE table_silver

SELECT * FROM table_bronze;

E)

INSERT OVERWRITE TABLE table_silver

SELECT * FROM table_bronze;

Options:

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

Answer:

A

Explanation:

Option A uses theSELECT DISTINCTstatement to remove duplicate rows from thetable_bronzeand create a new tabletable_silverwith the deduplicated data.This is the correct way to deduplicate data using Spark SQL12. Option B simply inserts all the rows fromtable_bronzeintotable_silver, without removing any duplicates. Option C is not a valid syntax for Spark SQL, as there is noMERGE DEDUPLICATEstatement. Option D appends all the rows fromtable_bronzeintotable_silver, without removing any duplicates. Option E overwrites the existing data intable_silverwith the data fromtable_bronze, without removing any duplicates.Reference:Delete Duplicate using SPARK SQL,Spark SQL - How to Remove Duplicate Rows

Question 4

Question Type: MultipleChoice

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?

Options:

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

Question 5

Question Type: MultipleChoice

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?

Options:

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.Reference:

1: Change table ownership

2: Ownerless tables

3: Table access control

4: Revoke access to a table

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

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.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