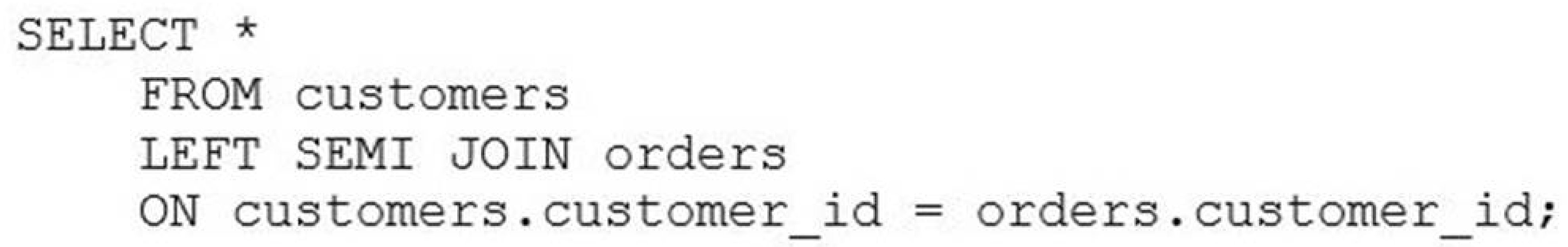
Exam Questions Databricks-Certified-Data-Analyst-Associate

Databricks Certified Data Analyst Associate Exam

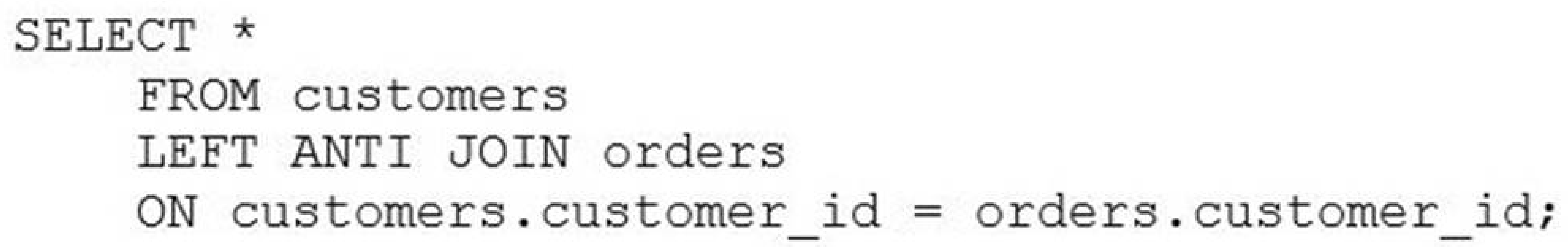
https://www.2passeasy.com/dumps/Databricks-Certified-Data-Analyst-Associate/

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 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 3
A data analyst needs to use the Databricks Lakehouse Platform to quickly create SQL queries and data visualizations. It is a requirement that the compute resources in the platform can be made serverless, and it is expected that data visualizations can be placed within a dashboard.
Which of the following Databricks Lakehouse Platform services/capabilities meets all of these requirements?

A. Delta Lake
B. Databricks Notebooks
C. Tableau
D. Databricks Machine Learning
E. Databricks SQL

Answer: E

Explanation:
Databricks SQL is a serverless data warehouse on the Lakehouse that lets you run all of your SQL and BI applications at scale with your tools of choice, all at a fraction of the cost of traditional cloud data warehouses1. Databricks SQL allows you to create SQL queries and data visualizations using the SQL Analytics UI or the Databricks
SQL CLI2. You can also place your data visualizations within a dashboard and share it with other users in your organization3. Databricks SQL is powered by Delta Lake, which provides reliability, performance, and governance for your data lake4. References:
? Databricks SQL
? Query data using SQL Analytics
? Visualizations in Databricks notebooks
? Delta Lake

NEW QUESTION 4
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

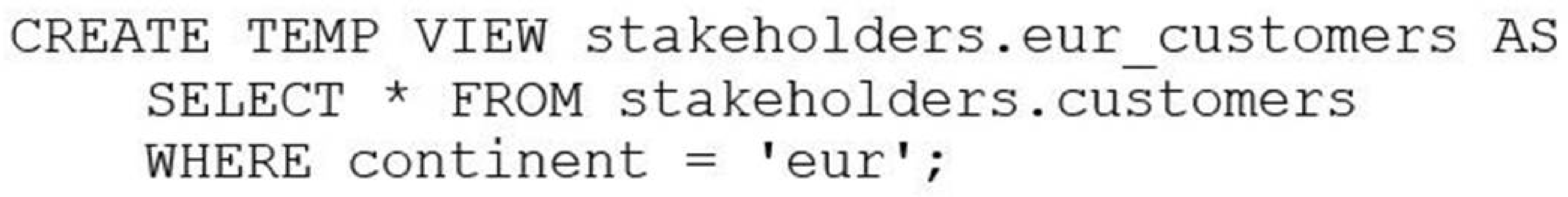
NEW QUESTION 5
A data analyst runs the following command:
INSERT INTO stakeholders.suppliers TABLE stakeholders.new_suppliers; What is the result of running this command?

A. The suppliers table now contains both the data it had before the command was run and the data from the new suppliers table, and any duplicate data is deleted.
B. The command fails because it is written incorrectly.
C. The suppliers table now contains both the data it had before the command was run and the data from the new suppliers table, includingany duplicate data.
D. The suppliers table now contains the data from the new suppliers table, and the new suppliers table now contains the data from the suppliers table.
E. The suppliers table now contains only the data from the new suppliers table.

Answer: B

Explanation:
The command INSERT INTO stakeholders.suppliers TABLE stakeholders.new_suppliers is not a valid syntax for inserting data into a table in Databricks SQL. According to the documentation12, the correct syntax for inserting data into a table is either:
? INSERT { OVERWRITE | INTO } [TABLE] table_name [PARTITION clause] [(
column_name [, ...]) | BY NAME] query
? INSERT INTO [TABLE] table_name REPLACE WHERE predicate query
The command in the question is missing the OVERWRITE or INTO keyword, and the query part that specifies the source of the data to be inserted. The TABLE keyword is optional and can be omitted. The PARTITION clause and the column list are also optional and depend on the table schema and the data source. Therefore, the command in the question will fail with a syntax error.
References:
? INSERT | Databricks on AWS
? INSERT - Azure Databricks - Databricks SQL | Microsoft Learn

NEW QUESTION 6
The stakeholders.customers table has 15 columns and 3,000 rows of data. The following command is run:



```
CREATE TEMP VIEW stakeholders.eur_customers AS
    SELECT * FROM stakeholders.customers
    WHERE continent = 'eur';
```

After runningSELECT * FROM stakeholders.eur_customers, 15 rows are returned. After the command executes completely, the user logs out of Databricks. After logging back in two days later, what is the status of thestakeholders.eur_customersview?

A. The view remains available and SELECT * FROM stakeholders.eur_customers will execute correctly.
B. The view has been dropped.
C. The view is not available in the metastore, but the underlying data can be accessed with SELECT * FROM delt
D. `stakeholders.eur_customers`.
E. The view remains available but attempting to SELECT from it results in an empty result set because data in views are automatically deleted after logging out.
F. The view has been converted into a table.

Answer: B

Explanation:
The command you sent creates a TEMP VIEW, which is a type of view that is only visible and accessible to the session that created it. When the session ends or the user logs out, the TEMP VIEW is automatically dropped and cannot be queried anymore. Therefore, after logging back in two days later, the status of the stakeholders.eur_customers view is that it has been dropped and SELECT * FROM stakeholders.eur_customers will result in an error. The other options are not correct because:
? A. The view does not remain available, as it is a TEMP VIEW that is dropped when the session ends or the user logs out.
? C. The view is not available in the metastore, as it is a TEMP VIEW that is not registered in the metastore. The underlying data cannot be accessed with SELECT * FROM delta. stakeholders.eur_customers, as this is not a valid syntax for querying a Delta Lake table. The correct syntax would be SELECT * FROM delta.dbfs:/stakeholders/eur_customers, where the location path is enclosed in backticks. However, this would also result in an error, as the TEMP VIEW does not write any data to the file system and the location path does not exist.
? D. The view does not remain available, as it is a TEMP VIEW that is dropped when the session ends or the user logs out. Data in views are not automatically deleted after logging out, as views do not store any data. They are only logical representations of queries on base tables or other views.
? E. The view has not been converted into a table, as there is no automatic conversion between views and tables in Databricks. To create a table from a view, you need to use a CREATE TABLE AS statement or a similar
command. References: CREATE VIEW | Databricks on AWS, Solved: How do temp views actually work? - Databricks - 20136, temp tables in Databricks - Databricks - 44012, Temporary View in Databricks - BIG DATA PROGRAMMERS, Solved: What is the difference between a Temporary View an ??

NEW QUESTION 7
Which of the following is a benefit of Databricks SQL using ANSI SQL as its standard SQL dialect?

A. It has increased customization capabilities
B. It is easy to migrate existingSQL queries to Databricks SQL
C. It allows for the use of Photon's computation optimizations
D. It is more performant than other SQL dialects
E. It is more compatible with Spark's interpreters

Answer: B

Explanation:
Databricks SQL uses ANSI SQL as its standard SQL dialect, which means it follows the SQL specifications defined by the American National Standards Institute (ANSI). This makes it easier to migrate existing SQL queries from other data warehouses or platforms that also use ANSI SQL or a similar dialect, such as PostgreSQL, Oracle, or Teradata. By using ANSI SQL, Databricks SQL avoids surprises in behavior or unfamiliar syntax that may arise from using anon-standard SQL dialect, such as Spark SQL or Hive SQL12. Moreover, Databricks SQL also adds compatibility features to support common SQL constructs that are widely used in other data warehouses, such as QUALIFY, FILTER, and user-defined functions2. References: ANSI compliance in Databricks
Runtime, Evolution of the SQL language at Databricks: ANSI standard by default and easier migrations from data warehouses

NEW QUESTION 8
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 9
A data analyst has been asked to configure an alert for a query that returns the income in the accounts_receivable table for a date range. The date range is configurable using a Date query parameter.
The Alert does not work.
Which of the following describes why the Alert does not work?

A. Alerts don't work with queries that access tables.
B. Queries that return results based on dates cannot be used with Alerts.
C. The wrong query parameter is being use
D. Alerts only work with Date and Time query parameters.
E. Queries that use query parameters cannot be used with Alerts.
F. The wrong query parameter is being use
G. Alerts only work with drogdown list query parameters, not dates.

Answer: D

Explanation:
According to the Databricks documentation1, queries that use query parameters cannot be used with Alerts. This is because Alerts do not support user input or dynamic values. Alerts leverage queries with parameters using the default value specified in the SQL editor for each parameter. Therefore, if the query uses a Date query parameter, the alert will always use the same date range as the default value, regardless of the actual date. This may cause the alert to not work as expected, or to not trigger at all. References:
? Databricks SQL alerts: This is the official documentation for Databricks SQL alerts,
where you can find information about how to create, configure, and monitor alerts, as well as the limitations and best practices for using alerts.

NEW QUESTION 10
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 10
Which of the following approaches can be used to ingest data directly from cloud-based object storage?

A. Createan external table while specifying the DBFS storage path to FROM
B. Create anexternal table while specifying the DBFS storage path to PATH
C. It is not possible to directly ingest data from cloud-based object storage
D. Create an external table while specifying the object storage path to FROM
E. Create an external table while specifying the object storage path to LOCATION

Answer: E

Explanation:
External tables are tables that are defined in the Databricks metastore using the information stored in a cloud object storage location. External tables do not manage the data, but provide a schema and a table name to query the data. To create an external table, you can use the CREATE EXTERNAL TABLE statement and specify the object storage path to the LOCATION clause. For example, to create an external table named ext_table on a Parquet file stored in S3, you can use the following statement:
SQL

```
CREATE EXTERNAL TABLE ext_table ( col1 INT,
col2 STRING
)
STORED AS PARQUET
LOCATION 's3://bucket/path/file.parquet'
```

AI-generated code. Review and use carefully. More info on FAQ.
References: External tables

NEW QUESTION 12
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 17
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 18
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 19
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 24

Which of the following is an advantage of using a Delta Lake-based data lakehouse over common data lake solutions?

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. References:
? 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

NEW QUESTION 27

......