Databricks.Databricks-Certified-Data-Analyst-Associate.v2026-09-22.q49

Exam Code:	Databricks-Certified-Data-Analyst-Associate
Exam Name:	Databricks Certified Data Analyst Associate Exam
Certification Provider:	Databricks
Free Question Number:	49
Version:	v2026-09-22
# of views:	127
# of Questions views:	490
https://www.freepdfdumps.com/Databricks.Databricks-Certified-Data-Analyst-Associate.v2026-09-22.q49.html	

NEW QUESTION: 1
A BI analyst is building an analytical data model in Databricks using Delta Lake tables. The source system contains transactional sales data that changes frequently. The analyst chooses to apply the Data Vault 2.0 methodology to manage historical changes while ensuring scalability and auditability across multiple business domains.
Which component is used to capture the many-to-many relationship between hubs in a Data Vault v2 model?
A. Hub Table
B. Satellite Table
C. Link Table
D. Reference Table
Answer: C (LEAVE A REPLY)
Option C is correct. In Data Vault modeling, hubs represent core business entities, satellites store descriptive and historical attributes, and links represent relationships between hubs. A many-to-many relationship between business entities is therefore captured by a Link Table. The Databricks Data Analyst Associate Exam Guide includes Data Vault schemas as part of data modeling with Databricks SQL, and Databricks' Data Vault guidance explains that links represent relationships between hub entities. References: Databricks Data Analyst Associate Exam Guide and Databricks Data Vault guidance.

NEW QUESTION: 2
A distributed team of data analysts share computing resources on an interactive cluster with autoscaling configured. In order to better manage costs and query throughput, the workspace administrator is hoping to evaluate whether cluster upscaling is caused by many concurrent users or resource-intensive queries.
In which location can one review the timeline for cluster resizing events?
A. Workspace audit logs
B. Driver's log file
C. Ganglia
D. Cluster Event Log
E. Executor's log file
Answer: D (LEAVE A REPLY)

Option D is correct. Cluster resizing caused by autoscaling is recorded as a cluster activity event, so the administrator should review the Cluster Event Log. Driver and executor logs are useful for application
/runtime troubleshooting, not for the cluster resizing timeline. Ganglia shows metrics but is not the event timeline. Official Databricks documentation identifies cluster activity events as the place to retrieve cluster activity, and Databricks autoscaling documentation describes resize events such as "Cluster resize request started."

NEW QUESTION: 3
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

Answer: E (LEAVE A REPLY)
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: 4
A data analyst has developed a query that runs against a Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean. However, the data engineering team uses Python for its tests rather than SQL. Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?
A. SELECT * FROM sales
B. spark.delta.table
C. spark.sql
D. There is no way to share data between PySpark and SQL.
E. spark.table
Answer: (SHOW ANSWER)
Option C is correct because the requirement is to run the SQL query and then operate on the results in PySpark. spark.sql(...) executes SQL text and returns a Spark DataFrame that can be used in PySpark tests.
spark.table(...) can load a table directly, but it does not run an arbitrary SQL query. Option A is SQL text by itself, not a PySpark operation. Official Databricks documentation states that a SparkSession can be used to execute SQL over tables, and Databricks notebook documentation specifically references using SQL inside Python with a spark.sql command.

NEW QUESTION: 5
Which location can be used to determine the owner of a managed table?
A. Review the Owner field in the table page using Catalog Explorer
B. Review the Owner field in the database page using Data Explorer
C. Review the Owner field in the schema page using Data Explorer
D. Review the Owner field in the table page using the SQL Editor
Answer: A (LEAVE A REPLY)
In Databricks, to determine the owner of a managed table, you can utilize the Catalog Explorer feature. The steps are as follows:
* Access Catalog Explorer:
* In your Databricks workspace, click on the Catalog icon in the sidebar to open Catalog Explorer.
* Navigate to the Table:
* Within Catalog Explorer, browse through the catalog and schema to locate the specific managed table whose ownership you wish to verify.
* View Table Details:
* Click on the table name to open its details page.
* Identify the Owner:
* On the table ' s details page, review the Owner field, which displays the principal (user, service principal, or group) that owns the table.
This method provides a straightforward way to ascertain the ownership of managed tables within the Databricks environment. Understanding table ownership is essential for managing permissions and ensuring proper access control.
Reference: Manage Unity Catalog object ownership

NEW QUESTION: 6
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 (LEAVE A REPLY)
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 very different 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: 7
A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data.
Which of the following relational objects should the data engineer create?
A. Spark SQL Table
B. View
C. Database
D. Temporary view
E. Delta Table
Answer: (SHOW ANSWER)
Option D is correct. A temporary view is session-scoped and does not store physical data. It is suitable when the relational object is only needed in the current session or query context and should not be available to others in other sessions. A regular view also avoids copying physical data, but it is persistent and can be shared beyond the current session. Official Databricks extract: a view is a "virtual table that has no physical data," and temporary views are scoped to the notebook/script or query level and cannot be referenced outside that scope.

NEW QUESTION: 8
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: (SHOW ANSWER)

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: 9
Which of the following data workloads will utilize a Gold table as its source?
A. A job that enriches data by parsing its timestamps into a human-readable format.
B. A job that aggregates uncleaned data to create standard summary statistics.
C. A job that cleans data by removing malformatted records.
D. A job that queries aggregated data designed to feed into a dashboard.
E. A job that ingests raw data from a streaming source into the Lakehouse.
Answer: D (LEAVE A REPLY)
Option D is correct. Gold tables are refined, business-ready datasets used for analytics, dashboards, reporting, and other downstream consumption. A workload that queries aggregated data designed to feed a dashboard is therefore using Gold-layer data. Raw streaming ingestion belongs to Bronze, cleaning malformed records belongs to Silver, and aggregating uncleaned data is not the recommended Gold pattern. Official Databricks extract: the Gold layer "consists of aggregated data tailored for analytics and reporting" and is optimized for "queries and dashboards."

NEW QUESTION: 10
In which of the following scenarios should a data engineer select a Task in the Depends On field of a new Databricks Job Task?
A. When another task needs to be replaced by the new task.
B. When another task needs to fail before the new task begins.
C. When another task has the same dependency libraries as the new task.
D. When another task needs to use as little compute resources as possible.
E. When another task needs to successfully complete before the new task begins.
Answer: E (LEAVE A REPLY)
Option E is correct. The Depends On field is used to define task dependencies in a Databricks job, so a downstream task waits for an upstream task according to the dependency rules. The standard use case is when one task must complete successfully before the next task starts. Official Databricks extract: "You can control the execution order of tasks by specifying dependencies between them," and Databricks runs upstream tasks before downstream tasks.

NEW QUESTION: 11
How can a data analyst determine if query results were pulled from the cache?
A. Go to the Query History tab and click on the text of the query. The slideout shows if the results came from the cache.
B. Go to the Alerts tab and check the Cache Status alert.
C. Go to the Queries tab and click on Cache Status. The status will be green if the results from the last run came from the cache.
D. Go to the SQL Warehouse (formerly SQL Endpoints) tab and click on Cache. The Cache file will show the contents of the cache.
E. Go to the Data tab and click Last Query. The details of the query will show if the results came from the cache.
Answer: A (LEAVE A REPLY)

Databricks SQL uses a query cache to store the results of queries that have been executed previously. This improves the performance and efficiency of repeated queries. To determine if a query result was pulled from the cache, you can go to the Query History tab in the Databricks SQL UI and click on the text of the query. A slideout will appear on the right side of the screen, showing the query details, including the cache status. If the result came from the cache, the cache status will show "Cached". If the result did not come from the cache, the cache status will show "Not cached". You can also see the cache hit ratio, which is the percentage of queries that were served from the cache. References: The answer can be verified from Databricks SQL documentation which provides information on how to use the query cache and how to check the cache status.
Reference link: Databricks SQL - Query Cache

NEW QUESTION: 12
A data analyst is troubleshooting a query in Databricks SQL that fails when processing large datasets and complex join operations. Logs indicate that the job consistently aborts due to resource constraint errors on the cluster.
Which Query Profile metric should the analyst use to identify the operator that is causing resource overuse?
A. Time spent per operator
B. Shuffle read size per operator
C. Memory peak per operator
D. Bytes spilled to disk per operator
Answer: C (LEAVE A REPLY)
The correct answer is C because the issue is a resource constraint failure, and the analyst needs to identify which operator is consuming excessive memory. In Query Profile, Memory peak shows memory usage at the operator level and helps identify the operator causing resource overuse. Time spent helps identify slow operators, shuffle read size helps analyze data movement, and bytes spilled to disk indicates spill behavior, but the most direct metric for resource overuse due to memory pressure is memory peak.
Official documentation extract used: Databricks Query Profile graph view shows metrics such as "Time spent, Memory peak, and Rows."

NEW QUESTION: 13
Which open-source project helps to enable the data lakehouse by adding organization, reliability, performance, and data governance to data lake architectures?
A. MLflow
B. Apache Spark
C. Delta Lake
D. Databricks SQL
Answer: C (LEAVE A REPLY)
The correct answer is C because Delta Lake is the open-source storage layer that provides reliability and structure for data lakes. Delta Lake brings ACID transactions, scalable metadata handling, and batch
/streaming support to data lake storage, which are core capabilities behind the lakehouse architecture. MLflow is for machine learning lifecycle management, Apache Spark is a distributed processing engine, and Databricks SQL is a Databricks product for SQL analytics, not the open-source storage project described.
Official documentation extract used: Databricks states that Delta Lake is "open source software" and provides
"ACID transactions and scalable metadata handling."

NEW QUESTION: 14

Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?
A. Cloud-specific integrations
B. Simplified governance
C. Ability to scale storage
D. Ability to scale workloads
E. Avoiding vendor lock-in
Answer: E (LEAVE A REPLY)
Option E is correct. A major benefit of open source technologies and open data formats is avoiding vendor lock-in. Databricks supports open formats and interfaces so data can be used across tools and systems rather than being locked into a proprietary platform. Cloud integrations, governance, and workload scalability are Databricks benefits, but the specific benefit of embracing open source is avoiding vendor lock-in. Official Databricks extract: "Using open data formats and interfaces helps to avoid" vendor lock-in, and Databricks states that no proprietary data formats are used because Delta Lake and Iceberg are open source.

NEW QUESTION: 15
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: E (LEAVE A REPLY)
This statement is incorrect. In Databricks SQL, any user with sufficient permissions on the query or dashboard can configure a refresh schedule-workspace administrator privileges are not required.
Here is the breakdown of the correct information:
* A. True - Queries can be scheduled to refresh at intervals ranging from 1 minute to 2 weeks.
* B. True - You can configure refresh schedules in the Query Editor.
* C. False statement - A query being refreshed does use a SQL Warehouse. However, the option in question says it does not use a warehouse, which would be incorrect in a different context. Since this is a trickier one, we know that scheduled queries do require a SQL Warehouse to run.
* D. True - Refresh schedules are different from alerts; alerts are triggered based on specific conditions being met in query results.
* E. False (and thus the correct answer to this question) - You do not need to be a workspace admin to set a refresh schedule. You only need the correct permissions on the object.
Reference: Schedule a Query in Databricks SQL

NEW QUESTION: 16
Where can an admin or data owner grant database, table, and view permissions to a group?
A. Dashboard
B. Data
C. SQL Warehouses
D. Settings
Answer: B (LEAVE A REPLY)

Option B is correct. In the older Databricks SQL UI terminology used by this question, permissions for databases, tables, and views are managed from the Data area. In the current Databricks UI, this is handled through Catalog Explorer, where an admin or data owner opens the object, goes to the Permissions tab, clicks Grant, selects users or groups, and grants the required privileges. Dashboards are for reporting, SQL Warehouses provide compute, and Settings is not the object-level permission management area. Official Databricks documentation says privileges can be managed using SQL commands, CLI, Terraform, or Catalog Explorer, and the UI flow is: open the table details page, go to Permissions, click Grant, select users or groups, and assign privileges.

NEW QUESTION: 17
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?
A. Data testing
B. Ad-hoc improvements
C. Last-mile
D. Last-mile ETL
E. Data enhancement
Answer: (SHOW ANSWER)
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

NEW QUESTION: 18
Which of the following SQL keywords can be used to convert a table from a long format to a wide format?
A. TRANSFORM
B. PIVOT
C. SUM
D. CONVERT
E. WHERE
Answer: (SHOW ANSWER)
Option B is correct. PIVOT rotates unique values from rows into separate columns, which is exactly the process of converting long-format data into wide-format data. UNPIVOT does the reverse. SUM is an aggregate function, WHERE filters rows, and TRANSFORM is not the SQL keyword used for this reshaping task. Official Databricks extract: the PIVOT clause "rotat[es] unique values from a column into separate columns."

NEW QUESTION: 19
A data analyst is processing a complex aggregation on a table with zero null values and the query returns the following result:

group_1	group_2	sum
null	null	100
A	null	50
A	Y	30
A	Z	20
B	null	50
B	Y	40
B	Z	1

Which query did the analyst execute in order to get this result?
A.
```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2;
```
B.
```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2 INCLUDING NULL;
```

C.
```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2 INCLUDING NULL;
```
D.
```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2 WITH CUBE;
```
Answer: D (LEAVE A REPLY)
Option D is correct because the table has zero real null values, but the result contains null values representing subtotal and grand-total rows. That behavior is produced by WITH CUBE, which creates aggregations for combinations of grouping columns, including (group_1, group_2), (group_1), (group_2), and the grand total ().
The Databricks SQL documentation states that GROUP BY supports advanced aggregations through CUBE, and that CUBE is shorthand for grouping sets. Option A only returns detailed groups. Options B and C use invalid syntax in Databricks SQL. Reference: Databricks GROUP BY clause documentation.

NEW QUESTION: 20
A Data Analyst is working on sensor_df; this DataFrame contains two columns: record_datetime timestamp and record array.
Which code fragment returns a DataFrame that splits the record column into separate columns and has one array item per row?
A. Uses withColumn, but selects sensor_id, status, and health as if they are already top-level columns.
B. Selects nested fields before correctly creating the exploded column.
C. exploded_df = sensor_df.withColumn(" record_exploded " , explode(" record "))exploded_df = exploded_df.select(" record_datetime " , " record_exploded.sensor_id " , " record_exploded.status " , " record_exploded.health ")
D. exploded_df = exploded_df.select(" record_datetime " , " record_exploded ")
Answer: (SHOW ANSWER)
Option C is correct after correcting the formatting and typing errors in the uploaded option text. The analyst needs explode(" record ") because the record column is an array, and the requirement is to return one array item per row. Then the analyst must select fields from the exploded struct using dot notation, such as record_exploded.sensor_id, record_exploded.status, and record_exploded.health. Databricks PySpark documentation states that explode "returns a new row for each element in the given array or map," and withColumn returns a new DataFrame by adding or replacing a column. The select method projects expressions or column names into the resulting DataFrame.

NEW QUESTION: 21
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 (LEAVE A REPLY)
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 not satisfactory, 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: 22
A data analyst is working in an organization that utilizes a multi-hop, medallion architecture. They have been tasked with creating a new Gold table from an existing Silver table.
Which query is performing a hop from a Silver table to a Gold table?

A.
```
CREATE TABLE cleaned_transactions AS
    SELECT
        store_id,
        customer_id,
        sales,
        units
    FROM transactions;
```

B.
```
CREATE TABLE transactions AS
    SELECT
    *
    FROM raw_transactions;
CREATE TABLE store_sales AS
```

C.

```
CREATE TABLE store_sales AS
    SELECT
        store_id,
        sum(sales) AS store_sales
    FROM cleaned_transactions
    GROUP BY store_id;
```

D.

```
CREATE TABLE cleaned_transactions AS
    SELECT
    *
    FROM transactions
    WHERE units > 0;
```

Answer: C (LEAVE A REPLY)

Option C is the correct answer because it creates a new aggregated table, store_sales, from the existing cleaned /refined table cleaned_transactions. In a medallion architecture, a table such as cleaned_transactions represents a Silver-layer table because it contains cleaned or validated transaction-level data. The query then creates store_sales by grouping by store_id and calculating SUM(sales), which is an aggregation suitable for analytics and reporting. That is exactly the Silver-to-Gold hop.

Option A creates cleaned_transactions from transactions, so it is moving toward a cleaned Silver table, not creating a Gold table. Option B creates transactions from raw_transactions, which is closer to a raw-to-Bronze or Bronze preparation step. Option D filters rows from transactions into cleaned_transactions, which is also a cleaning/validation step associated with Silver, not Gold.

Exact extract from official Databricks documentation: Databricks describes medallion architecture as improving data through "Bronze # Silver # Gold layer tables." The same official page identifies Silver as

"Data cleaning and validation" and Gold as "Dimensional modeling and aggregation." Databricks also states that the Gold layer "consists of aggregated data tailored for analytics and reporting."

NEW QUESTION: 23

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 (LEAVE A REPLY)

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, data enrichment, 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: 24
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 BY. which is not allowed in an aggregation.
D. There are no mistakes in the query.
E. The query is selecting region but region should only occur in the ORDER BY clause.
Answer: (SHOW ANSWER)
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 why you 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: 25
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 (LEAVE A REPLY)

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: 26
A data analyst has been asked to use the below table sales_table to 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)

A. Option A

B. Option B

C. Option C

D. Option D

Answer: B (LEAVE A REPLY)

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: 27
A data analyst at an e-commerce company needs to process daily sales data. The data consists of approximately 50,000 records stored in a single CSV file, totaling about 20 MB. The analyst needs to perform aggregations and generate a summary report.
Which approach could the data analyst use in this situation?
A. Deploy a real-time streaming solution using Spark Streaming to process incoming data.
B. Use a local Python script with the pandas library to read and analyze the CSV file.
C. Implement Apache Spark with a distributed cluster to process the data in parallel.
D. Set up a Hadoop ecosystem with HDFS and MapReduce for distributed processing.
Answer: B (LEAVE A REPLY)
Option B is correct. A 20 MB CSV file with about 50,000 records is small enough for a local pandas workflow. A real-time streaming solution, a distributed Spark cluster, or Hadoop MapReduce would add unnecessary complexity for a small single-file batch analysis. Databricks documentation explains that pandas is available in Databricks Runtime and can be used for data analysis, while pandas API on Spark is useful when pandas-style workloads need to scale beyond smaller datasets. Databricks also notes that pandas does not scale out to big data, which is exactly why Spark or pandas API on Spark is used for larger workloads, not for a small 20 MB file.

NEW QUESTION: 28
A stakeholder has provided a data analyst with a lookup dataset in the form of a 50-row CSV file. The data analyst needs to upload this dataset for use as a table in Databricks SQL.
Which approach should the data analyst use to quickly upload the file into a table for use in Databricks SOL?
A. Create a table by uploading the file using the Create page within Databricks SQL
B. Create a table via a connection between Databricks and the desktop facilitated by Partner Connect.
C. Create a table by uploading the file to cloud storage and then importing the data to Databricks.
D. Create a table by manually copying and pasting the data values into cloud storage and then importing the data to Databricks.
Answer: A (LEAVE A REPLY)
Databricks provides a user-friendly interface that allows data analysts to quickly upload small datasets, such as a 50-row CSV file, and create tables within Databricks SQL. The steps are as follows:
* Access the Data Upload Interface:
* In the Databricks workspace, navigate to the sidebar and click on New > Add or upload data.
* Select Create or modify a table.
* Upload the CSV File:
* Click on the browse button or drag and drop the CSV file directly onto the designated area.
* The interface supports uploading up to 10 files simultaneously, with a total size limit of 2 GB.
* Configure Table Settings:
* After uploading, a preview of the data is displayed.
* Specify the table name, select the appropriate schema, and configure any additional settings as needed.
* Create the Table:
* Once all configurations are set, click on the Create Table button to finalize the process.
This method is efficient for quickly importing small datasets without the need for additional tools or complex configurations. Options B, C, and D involve more complex or manual processes that are unnecessary for this task.
Reference: Create or modify a table using file upload

NEW QUESTION: 29
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;
A. Option A
B. Option B
C. Option C
D. Option D
E. Option E
Answer: A (LEAVE A REPLY)
Option A uses the SELECT DISTINCT statement to remove duplicate rows from the table_bronze and create a new table table_silver with the deduplicated data. This is the correct way to deduplicate data using Spark SQL12. Option B simply inserts all the rows from table_bronze into table_silver, without removing any duplicates. Option C is not a valid syntax for Spark SQL, as there is no MERGE DEDUPLICATE statement.
Option D appends all the rows from table_bronze into table_silver, without removing any duplicates. Option E overwrites the existing data in table_silver with the data from table_bronze, without removing any duplicates. References: Delete Duplicate using SPARK SQL, Spark SQL - How to Remove Duplicate Rows

NEW QUESTION: 30
A data analyst wants to generate insights from large, complex datasets. The analyst needs to quickly understand the meaning of various data columns, ask questions in natural language, and receive AI-driven recommendations for optimizing data queries and workflows. Which Databricks component is primarily responsible for enabling these capabilities?
A. Data Intelligence Engine

B. Unity Catalog
C. Genie Spaces
D. Databricks Assistant
Answer: A (LEAVE A REPLY)
Option A is correct. The Data Intelligence Engine is the platform-level intelligence layer that understands the semantics and uniqueness of an organization's data and enables AI-assisted experiences across Databricks.
Unity Catalog provides governance and metadata management, Genie Spaces provide a natural-language interface for curated business data, and Databricks Assistant is a user-facing assistant for code/query help.
However, the question asks which component is primarily responsible for enabling these capabilities across the platform; that is the Data Intelligence Engine. Databricks describes the platform as powered by a Data Intelligence Engine that understands the uniqueness and semantics of data and helps optimize performance.
References: Databricks Data Intelligence Platform documentation and Data Analyst Associate Exam Guide.

NEW QUESTION: 31
Which of the following should data analysts consider when working with personally identifiable information (PII) data?
A. Organization-specific best practices for PII data
B. Legal requirements for the area in which the data was collected
C. None of these considerations
D. Legal requirements for the area in which the analysis is being performed
E. All of these considerations
Answer: (SHOW ANSWER)
Data analysts should consider all of these factors when working with PII data, as they may affect the data security, privacy, compliance, and quality. PII data is any information that can be used to identify a specific individual, such as name, address, phone number, email, social security number, etc. PII data may be subject to different legal and ethical obligations depending on the context and location of the data collection and analysis. For example, some countries or regions may have stricter data protection laws than others, such as the General Data Protection Regulation (GDPR) in the European Union. Data analysts should also follow the organization-specific best practices for PII data, such as encryption, anonymization, masking, access control, auditing, etc. These best practices can help prevent data breaches, unauthorized access, misuse, or loss of PII data. References:
* How to Use Databricks to Encrypt and Protect PII Data
* Automating Sensitive Data (PII/PHI) Detection
* Databricks Certified Data Analyst Associate

NEW QUESTION: 32
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 (LEAVE A REPLY)
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: 33
A data analyst is processing a complex aggregation on a table with zero null values and the query returns the following result:
Which query did the analyst execute in order to get this result?
A.
```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2 WITH ROLLUP;
```
B.
```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2 WITH CUBE;
```
C.
```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2;
```

D.

```
SELECT
    group_1,
    group_2,
    sum(values) AS sum
FROM my_table
GROUP BY group_1, group_2 INCLUDING NULL;
```


Answer: D (LEAVE A REPLY)

NEW QUESTION: 34

A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.

Which of the following Git operations does the data engineer need to run to accomplish this task?

A. Merge

B. Push

C. Pull

D. Commit

E. Clone

Answer: C (LEAVE A REPLY)

Option C is correct. To retrieve changes from a remote central Git repository into a local Databricks Git folder /Repo, the correct Git operation is Pull. Push sends local commits to the remote repository. Commit records local changes. Clone creates a new local copy. Merge combines branches but is not the operation described here. Official Databricks extract: "To pull changes from the remote Git repository, click Pull."

NEW QUESTION: 35

A data analyst wants to create a Databricks SQL dashboard with multiple data visualizations and multiple counters. What must be completed before adding the data visualizations and counters to the dashboard?

A. All data visualizations and counters must be created using Queries.

B. A SQL warehouse (formerly known as SQL endpoint) must be turned on and selected.

C. A markdown-based tile must be added to the top of the dashboard displaying the dashboard ' s name.

D. The dashboard owner must also be the owner of the queries, data visualizations, and counters.

Answer: (SHOW ANSWER)

In Databricks SQL, when creating a dashboard that includes multiple data visualizations and counters, it is imperative that each visualization and counter is based on a query. The process involves the following steps:

* Develop Queries:

* For each desired visualization or counter, write a SQL query that retrieves the necessary data.

* Create Visualizations and Counters:

* After executing each query, utilize the results to create corresponding visualizations or counters.

Databricks SQL offers a variety of visualization types to represent data effectively.

* Assemble the Dashboard:

* Add the created visualizations and counters to your dashboard, arranging them as needed to convey the desired insights.
By ensuring that all components of the dashboard are derived from queries, you maintain consistency, accuracy, and the ability to refresh data as needed. This approach also facilitates easier maintenance and updates to the dashboard elements.
Reference: Visualization in Databricks SQL

NEW QUESTION: 36
Which data lakehouse feature results in improved data quality over a traditional data lake?
A. A data lakehouse stores data in open formats.
B. A data lakehouse allows the use of SQL queries to examine data.
C. A data lakehouse provides storage solutions for structured and unstructured data.
D. A data lakehouse supports ACID-compliant transactions.
Answer: D (LEAVE A REPLY)
Option D is correct. ACID-compliant transactions improve reliability and data quality because writes are consistent and atomic rather than leaving partially written or corrupted states. Open formats and SQL access are important lakehouse capabilities, but the feature most directly tied to improved data quality over a traditional data lake is ACID transactions. Official Databricks extract: Delta Lake extends Parquet with a transaction log for "ACID transactions," and Databricks states that Delta Lake adds ACID transactions and schema evolution for reliable, high-quality data.

NEW QUESTION: 37
A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.
Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with $0 in sales is greater than zero?
A. They can set up an Alert with a custom template.
B. They can set up an Alert with a new email alert destination.
C. They can set up an Alert with one-time notifications.
D. They can set up an Alert with a new webhook alert destination.
E. They can set up an Alert without notifications.
Answer: D (LEAVE A REPLY)
Option D is correct. Databricks SQL alerts run a query on a schedule, evaluate a condition, and notify destinations when the condition is met. Since the requirement is notification through a messaging webhook, the alert should use a webhook alert destination. Official Databricks extract: alerts "run queries on a schedule and notify you when a condition that you define is met." Databricks notification documentation also states that admins can configure alternate notification destinations using webhooks.

NEW QUESTION: 38
An analyst writes a query that contains a query parameter. They then add an area chart visualization to the query. While adding the area chart visualization to a dashboard, the analyst chooses " Dashboard Parameter " for the query parameter associated with the area chart.
Which of the following statements is true?
A. The area chart will use whatever is selected in the Dashboard Parameter while all or the other visualizations will remain changed regardless of their parameter use.

B. The area chart will use whatever is selected in the Dashboard Parameter along with all of the other visualizations in the dashboard that use the same parameter.
C. The area chart will use whatever value is chosen on the dashboard at the time the area chart is added to the dashboard.
D. The area chart will use whatever value is input by the analyst when the visualization is added to the dashboard. The parameter cannot be changed by the user afterwards.
E. The area chart will convert to a Dashboard Parameter.
Answer: B (LEAVE A REPLY)
A Dashboard Parameter is a parameter that is configured for one or more visualizations within a dashboard and appears at the top of the dashboard. The parameter values specified for a Dashboard Parameter apply to all visualizations reusing that particular Dashboard Parameter1. Therefore, if the analyst chooses "Dashboard Parameter" for the query parameter associated with the area chart, the area chart will use whatever is selected in the Dashboard Parameter along with all of the other visualizations in the dashboard that use the same parameter. This allows the user to filter the data across multiple visualizations using a single parameter widget2. References: Databricks SQL dashboards, Query parameters

NEW QUESTION: 39
A data analyst has produced a visualization. A stakeholder has viewed the visualization and is complaining that the visualization is difficult to interpret. After looking at the visualization, the analyst determines that the scale of the y-axis must be changed.
Where are the controls for changing the scale of the y-axis in Databricks SQL?
A. Query Editor # Y Axis
B. Dashboard Editor # Axes # Y Axis
C. Visualization Editor # Y Axis
D. Settings # User Settings # Scaling
Answer: C (LEAVE A REPLY)
The correct answer is C because axis settings are configured in the visualization editor, specifically under the Y-axis settings. The Query Editor is for writing SQL, not configuring chart axes. User Settings do not control individual chart scales. Dashboard editing may contain visualizations, but the specific control is in the visualization editor's Y Axis configuration.
Official documentation extract used: Databricks chart options state that to configure formatting for the X axis or Y axis, click the corresponding axis tab, and the Y-axis supports scale and range controls.

NEW QUESTION: 40
A data analyst has two data sources that are providing similar but complementary information. The analyst wants to combine these sources of data into a single, comprehensive dataset for ongoing use for their team in a variety of different projects.
Which term is used to describe this type of work?
A. Last-mile ETL
B. Ad-hoc improvements
C. Data testing
D. Data blending
Answer: D (LEAVE A REPLY)
Option D is correct. The scenario describes combining multiple complementary data sources into one comprehensive dataset. That is data blending. Last-mile ETL is usually project-specific final transformation near the end of an analytics workflow, while this question emphasizes combining two source datasets for broader ongoing team use. The current official Databricks exam guide describes the same kind of

capability as creating unified datasets by joining data from multiple sources. That aligns with the concept of data blending. Reference: Databricks Certified Data Analyst Associate Exam Guide.

NEW QUESTION: 41
A data team has been given a series of projects by a consultant that need to be implemented in the Databricks Lakehouse Platform.
Which of the following projects should be completed in Databricks SQL?
A. Testing the quality of data as it is imported from a source
B. Tracking usage of feature variables for machine learning projects
C. Combining two data sources into a single, comprehensive dataset
D. Segmenting customers into like groups using a clustering algorithm
E. Automating complex notebook-based workflows with multiple tasks
Answer: C (LEAVE A REPLY)
Databricks SQL is a service that allows users to query data in the lakehouse using SQL and create visualizations and dashboards1. One of the common use cases for Databricks SQL is to combine data from different sources and formats into a single, comprehensive dataset that can be used for further analysis or reporting2. For example, a data analyst can use Databricks SQL to join data from a CSV file and a Parquet file, or from a Delta table and a JDBC table, and create a new table or view that contains the combined data3.
This can help simplify the data management and governance, as well as improve the data quality and consistency. References:
* Databricks SQL overview
* Databricks SQL use cases
* Joining data sources

NEW QUESTION: 42
A Data Analyst is working on employees_df and needs to add a new column where a 10% tax is calculated on the salary. Additionally, the DataFrame contains the column age, which is not needed.
Which code fragment adds the tax column and removes the age column?
A. employees_df = employees_df.withColumn(" tax " , employees_df.salary * 10).dropField(" age ")
B. employees_df = employees_df.withColumn(" tax " , employees_df.salary * 0.1).drop(" age ")
C. employees_df = employees_df.withColumn(" tax " , lit(0.1) * col(" salary ")).dropField(" age ")
D. employees_df = employees_df.withColumn(" tax " , employees_df.salary * 10).drop(" age ")
Answer: B (LEAVE A REPLY)
Option B is correct after correcting the uploaded typo withcolumn to withColumn. A 10% tax is calculated by multiplying salary by 0.1, not by 10. The age column should be removed using .drop(" age "). Options A and D multiply salary by 10, which calculates 1000%, not 10%. Options A and C also use dropField, which is not the correct DataFrame method for removing a top-level column. Official Databricks PySpark documentation states that withColumn adds or replaces a column and that drop(*cols) returns a new DataFrame without the specified columns.

NEW QUESTION: 43
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 (LEAVE A REPLY)
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: 44
A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.
Which command could the data engineering team use to access sales in PySpark?
A. SELECT * FROM sales
B. spark.table(" sales ")
C. spark.sql(" sales ")
D. spark.delta.table(" sales ")
Answer: B (LEAVE A REPLY)
Option B is correct because spark.table(" sales ") returns the named table as a Spark DataFrame, which the data engineering team can then test using PySpark. SELECT * FROM sales is SQL text, not a PySpark command by itself. spark.sql(" sales ") is invalid because spark.sql expects a SQL statement, not just a table name. spark.delta.table is not the standard PySpark API for loading a table. Official Databricks extract:
DataFrameReader.table "returns the specified table as a DataFrame," and Databricks also lists spark.table as a Spark operation that returns a DataFrame.

NEW QUESTION: 45
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 (LEAVE A REPLY)
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: 46
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 (LEAVE A REPLY)
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: 47
What describes the variance of a set of values?
A. Variance is a measure of how far a single observed value is from a set ot va IN
B. Variance is a measure of how far an observed value is from the variable ' s maximum or minimum value.
C. Variance is a measure of central tendency of a set of values.
D. Variance is a measure of how far a set of values is spread out from the sets central value.
Answer: D (LEAVE A REPLY)
Variance is a statistical measure that quantifies the dispersion or spread of a set of values around their mean (central value). It is calculated by taking the average of the squared differences between each value and the mean of the dataset. A higher variance indicates that the data points are more spread out from the mean, while a lower variance suggests that they are closer to the mean. This measure is fundamental in statistics to understand the degree of variability within a dataset.WikipediaWikipedia+1Investopedia+1 Reference: Variance - Wikipedia

NEW QUESTION: 48
A table named user_ltv is being used to create a view that will be used by data analysts on various teams.
Users in the workspace are configured into groups, which are used for setting up data access using ACLs.
The user_ltv table has the following schema:
email STRING, age INT, ltv INT
The following view definition is executed:
CREATE VIEW user_ltv_no_minors AS
SELECT email, age, ltv
FROM user_ltv
WHERE
CASE
WHEN is_member(" auditing ") THEN TRUE
ELSE age > = 18
END;
An analyst who is not a member of the auditing group executes the following query:
SELECT * FROM user_ltv_no_minors;
Which statement describes the results returned by this query?
A. All columns will be displayed normally for those records that have an age greater than 17; records not meeting this condition will be omitted.
B. All age values less than 18 will be returned as null values, all other columns will be returned with the values in user_ltv.
C. All values for the age column will be returned as null values, all other columns will be returned with the values in user_ltv.

D. All records from all columns will be displayed with the values in user_ltv.
E. All columns will be displayed normally for those records that have an age greater than 18; records not meeting this condition will be omitted.
Answer: (SHOW ANSWER)
Option A is correct. The user is not a member of the auditing group, so is_member(" auditing ") evaluates to false and the ELSE age > = 18 branch controls the filter. Rows with age > = 18 are returned; rows under 18 are omitted. "Age greater than 17" is equivalent to age > = 18 for integer ages. Official Databricks extract:
is_member() "returns TRUE if the current user is a member" of the specified group, and Databricks describes dynamic views as views that can filter rows based on group membership.

NEW QUESTION: 49
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 (LEAVE A REPLY)
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