GAQM.Databricks-Certified-Data-Engineer-Associate.v2026-07-28.q73

Exam Code:	Databricks-Certified-Data-Engineer-Associate
Exam Name:	Databricks Certified Data Engineer Associate Exam
Certification Provider:	GAQM
Free Question Number:	73
Version:	v2026-07-28
# of views:	104
# of Questions views:	730
https://www.freecram.net/torrent/GAQM.Databricks-Certified-Data-Engineer-Associate.v2026-07-28.q73.html	

NEW QUESTION: 1
A data engineer needs to ingest from both streaming and batch sources for a firm that relies on highly accurate data. Occasionally, some of the data picked up by the sensors that provide a streaming input are outside the expected parameters. If this occurs, the data must be dropped, but the stream should not fail.
Which feature of Delta Live Tables meets this requirement?
A. Error Handling
B. Change Data Capture
C. Monitoring
D. Expectations
Answer: (SHOW ANSWER)

NEW QUESTION: 2
Which of the following describes a scenario in which a data team will want to utilize cluster pools?
A. An automated report needs to be refreshed as quickly as possible.
B. An automated report needs to be made reproducible.
C. An automated report needs to be tested to identify errors.
D. An automated report needs to be version-controlled across multiple collaborators.
E. An automated report needs to be runnable by all stakeholders.
Answer: (SHOW ANSWER)
Databricks cluster pools are a set of idle, ready-to-use instances that can reduce cluster start and auto-scaling times. This is useful for scenarios where a data team needs to run an automated report as quickly as possible, without waiting for the cluster to launch or scale up. Cluster pools can also help save costs by reusing idle instances across different clusters and avoiding DBU charges for idle instances in the pool. References: Best practices: pools | Databricks on AWS, Best practices: pools - Azure Databricks | Microsoft Learn, Best practices: pools | Databricks on Google Cloud

NEW QUESTION: 3
A data engineer works for an organization that must meet a stringent Service Level Agreement (SLA) that demands minimal runtime errors and high availability for its data processing pipelines. The data engineer wants to avoid the operational overhead of managing and tuning clusters.
Which architectural solution will meet the requirements?

A. Deploy a dedicated, manually managed cluster optimized by in-house IT staff.
B. Utilize Databricks serverless compute that automatically optimizes resources and abstracts cluster management.
C. Use an auto-scaling cluster configured and monitored by the user.
D. Implement a hybrid approach with scheduled batch jobs on custom cloud VMs.
Answer: (SHOW ANSWER)

NEW QUESTION: 4
A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.
Which of the following code blocks successfully completes this task?

A.
```
SELECT
    store_id,
    employees,
    FILTER (employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```

B.
```
SELECT
    store_id,
    employees,
    FILTER (exp_employees, years_exp > 5) AS exp_employees
FROM stores;
```

C.
```
SELECT
    store_id,
    employees,
    FILTER (employees, years_exp > 5) AS exp_employees
FROM stores;
```

D.
```
SELECT
    store_id,
    employees,
    CASE WHEN employees.years_exp > 5 THEN employees
         ELSE NULL
    END AS exp_employees
FROM stores;
```

E.
```
SELECT
    store_id,
    employees,
    FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```


A. Option A
B. Option B
C. Option C
D. Option D
E. Option E
Answer: (SHOW ANSWER)

Option A is the correct answer because it uses the FILTER higher-order function correctly to filter out employees with more than 5 years of experience from the array column "employees". It applies a lambda function i -> i.years_exp > 5 that checks if the years of experience of each employee in the array is greater than 5. If this condition is met, the employee is included in the new array column "exp_employees".
The use of higher-order functions like FILTER can be referenced from Databricks documentation on Higher- Order Functions.

NEW QUESTION: 5
A global retail company sells products across multiple categories (e.g.. Electronics, Clothing) and regions (e.
g.. North. South, East. West). The sales team has provided the data engineer with a PySpark dataframe named sales_df as below and the team wants the data engineer to analyze the sales data to help them make strategic decisions.

sales_df

product_id	category	sales_amount	region
1	Electronics	100	North
2	Clothing	200	South
3	Electronics	150	North
4	Clothing	300	South
5	Electronics	250	East
6	Clothing	400	West

A. Category_sale: .es df -agg (sum ("sales amount") .-;r*i:rRy ("category") .alias ("total sa.en amount"))
B. Category_sales = sales_df.sum("3ales_amount"). g-1- upBy("categcryn).alias("toLal_sales_amount))
C. Category_sales = sales_df.groupBy("reqion"). agq(sum("sales_amountn).alias(ntotal_sales_amount"))
D. Category_sales = sales df.groupBy("category").agg(sum("sales amount") .alias ("total sales amount"))
Answer: (SHOW ANSWER)

NEW QUESTION: 6

Calculate the total sales amount for each region and store the results in a new dataframe called region_sales.
Given the expected result:

region	total_sales_amount
North	650
South	500
East	250
West	400

Which code will generate the expected result?

A. region sales - sales_df.agg(sum("sales_amount").groupBy("region").alias("total sales amount"))

B. region_sales = sales_df.groupBy("region").agg(sum("sales_amountM).alias("total_sales_amount"))

C. region_sales= sales_df.groupBy("category").sum(nsales_amount").alias("t_otal_sales_amounl")

D. region_sales = sales_df. sum ("salen_aiTiount") . groupBy ("region") .alias ("total_sale3_amount")

Answer: (SHOW ANSWER)

NEW QUESTION: 7

What is the maximum output supported by a job cluster to ensure a notebook does not fail?

A. 30MBS

B. 25MBS

C. 15MBS

D. 10MBS

Answer: (SHOW ANSWER)

NEW QUESTION: 8

Which of the following benefits is provided by the array functions from Spark SQL?

A. An ability to work with data in a variety of types at once

B. An ability to work with data within certain partitions and windows

C. An ability to work with time-related data in specified intervals

D. An ability to work with complex, nested data ingested from JSON files

E. An ability to work with an array of tables for procedural automation

Answer: (SHOW ANSWER)

The array functions from Spark SQL are a subset of the collection functions that operate on array columns1. They provide an ability to work with complex, nested data ingested from JSON files or other sources2. For example, the explode function can be used to transform an array column into multiple rows, one for each element in the array3. The array_contains function can be used to check if a value is present in an array column4. The array_join function can be used to concatenate all elements of an array column with a delimiter. These functions can be useful for processing JSON data that may have nested arrays or objects. References: 1: Spark SQL, Built-in Functions - Apache Spark 2: Spark SQL Array Functions Complete List - Spark By Examples 3: Spark SQL Array Functions - Syntax and Examples - DWgeek.
com 4: Spark SQL, Built-in Functions - Apache Spark : Spark SQL, Built-in Functions - Apache Spark :
[Working with Nested Data Using Higher Order Functions in SQL on Databricks - The Databricks Blog]

NEW QUESTION: 9

A data engineer wants to create a new table containing the names of customers that live in France.
They have written the following command:

```
CREATE TABLE customersInFrance
_______ AS
SELECT id,
       firstName,
       lastName,
FROM customerLocations
WHERE country = 'FRANCE';
```


A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).
Which of the following lines of code fills in the above blank to successfully complete the task?

A. There is no way to indicate whether a table contains PII.

B. "COMMENT PII"

C. TBLPROPERTIES PII

D. COMMENT "Contains PII"

E. PII

Answer: D (LEAVE A REPLY)

In Databricks, when creating a table, you can add a comment to columns or the entire table to provide more information about the data it contains. In this case, since it's organization policy to indicate that the new table includes personally identifiable information (PII), option D is correct. The line of code would be added after defining the table structure and before closing with a semicolon. References: Data Engineer Associate Exam Guide, CREATE TABLE USING (Databricks SQL)

NEW QUESTION: 10

Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

A. Parquet files can be partitioned

B. CREATE TABLE AS SELECT statements cannot be used on files

C. Parquet files have a well-defined schema

D. Parquet files have the ability to be optimized

E. Parquet files will become Delta tables

Answer: (SHOW ANSWER)

Option C is the correct answer because Parquet files have a well-defined schema that is embedded within the data itself. This means that the data types and column names of the Parquet files are automatically detected and preserved when creating an external table from them. This also enables the use of SQL and other structured query languages to access and analyze the data. CSV files, on the other hand, do not have a schema embedded in them, and require specifying the schema explicitly or inferring it from the data when creating an external table from them. This can lead to errors or inconsistencies in the data types and column names, and also increase the processing time and complexity.
CREATE TABLE AS SELECT, Parquet Files, CSV Files, Parquet vs. CSV

NEW QUESTION: 11

A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer 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: (SHOW ANSWER)
To sync a Databricks Repo with the changes from a central Git repository, the data engineer needs to run the Git pull operation. This operation fetches the latest updates from the remote repository and merges them with the local repository. The data engineer can use the Pull button in the Databricks Repos UI, or use the git pull command in a terminal session. The other options are not relevant for this task, as they either push changes to the remote repository (Push), combine two branches (Merge), save changes to the local repository (Commit), or create a new local repository from a remote one (Clone). References:
* Run Git operations on Databricks Repos
* Git pull

NEW QUESTION: 12
A data engineer has been given a new record of data:
id STRING = 'a1'
rank INTEGER = 6
rating FLOAT = 9.4
Which of the following SQL commands can be used to append the new record to an existing Delta table my_table?
A. INSERT INTO my_table VALUES ('a1', 6, 9.4)
B. my_table UNION VALUES ('a1', 6, 9.4)
C. INSERT VALUES ('a1' , 6, 9.4) INTO my_table
D. UPDATE my_table VALUES ('a1', 6, 9.4)
E. UPDATE VALUES ('a1', 6, 9.4) my_table
Answer: (SHOW ANSWER)
To append a new record to an existing Delta table, you can use the INSERT INTO statement with the VALUES clause. This statement will insert one or more rows into the table with the specified values.
Option A is the only code block that follows this syntax correctly. Option B is incorrect, as it uses the UNION operator, which will return a new table that is the union of two tables, not append to an existing table. Option C is incorrect, as it uses the INSERT VALUES statement, which is not a valid SQL syntax.
Option D is incorrect, as it uses the UPDATE statement, which will modify existing rows in the table, not append new rows. Option E is incorrect, as it uses the UPDATE VALUES statement, which is also not a valid SQL syntax. References: Insert data into a table using SQL | Databricks on AWS, Insert data into a table using SQL - Azure Databricks, Delta Lake Quickstart - Azure Databricks

NEW QUESTION: 13
A data engineer needs to provide access to a group named manufacturing-team. The team needs privileges to create tables in the quality schema.
Which set of SQL commands will grant a group named manufacturing-team to create tables in a schema named production with the parent catalog named manufacturing with the least privileges?
A.
```
GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE SCHEMA ON SCHEMA
manufacturing.quality TO manufacturing-team; GRANT USE CATALOG ON CATALOG manufacturing TO
manufacturing-team;
```

B.

```
GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT CREATE SCHEMA ON SCHEMA
manufacturing.quality TO manufacturing-team; GRANT CREATE CATALOG ON CATALOG manufacturing TO
manufacturing-team;
  GRANT USE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE SCHEMA ON SCHEMA
  manufacturing.quality TO manufacturing-team; GRANT USE CATALOG ON CATALOG manufacturing TO
  manufacturing-team;
```

C.

```
   GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT CREATE SCHEMA ON
  SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE CATALOG ON CATALOG manufacturing TO
  manufacturing-team;
```

D.

Answer: (SHOW ANSWER)

NEW QUESTION: 14

What is the structure of an Asset Bundle?

A. A YAML configuration file that specifies the artifacts, resources, and configurations for the project.

B. A Docker image containing runtime environments and the source code of the assets

C. A single plain text file enumerating the names of assets to be migrated to a new workspace.

D. A compressed archive (ZIP) that solely contains workspace assets without any accompanying metadata.

Answer: (SHOW ANSWER)

NEW QUESTION: 15

A data engineer is attempting to write Python and SQL in the same command cell and is running into an error The engineer thought that it was possible to use a Python variable in a select statement.

Why does the command fail?

A. Databricks supports language interoperability in the same cell but only between Scala and SQL

B. Databricks supports one language per cell.

C. Databricks supports multiple languages but only one per notebook.

D. Databricks supports language interoperability but only if a special character is used.

Answer: (SHOW ANSWER)

NEW QUESTION: 16

Identify the impact of ON VIOLATION DROP ROW and ON VIOLATION FAIL UPDATE for a constraint violation.

A data engineer has created an ETL pipeline using Delta Live table to manage their company travel reimbursement detail, they want to ensure that the if the location details has not been provided by the employee, the pipeline needs to be terminated.

How can the scenario be implemented?

A. CONSTRAINT valid_location EXPECT (location != NULL) ON VIOLATION FAIL

B. CONSTRAINT valid_location EXPECT (location != NULL) ON VIOLATION FAIL UPDATE

C. CONSTRAINT valid_location EXPECT (location = NULL)

D. CONSTRAINT valid_location EXPECT (location != NULL) ON DROP ROW

Answer: (SHOW ANSWER)

NEW QUESTION: 17
A data engineer is maintaining an ETL pipeline code with a GitHub repository linked to their Databricks account. The data engineer wants to deploy the ETL pipeline to production as a databricks workflow.
Which approach should the data engineer use?
A. Manually create and manage the workflow in UI
B. Maintain workflow_conf ig. json and deploy it using Terraform
C. Maintain workflow_config.j son and deploy it using Databricks CLI
D. Databricks Asset Bundles (DAB) + GitHub Integration
Answer: (SHOW ANSWER)

NEW QUESTION: 18
A data engineer needs to conduct Exploratory Analysis on data residing in a database that is within the company's custom-defined network in the cloud. The data engineer is using SQL for this task.
Which type of SQL Warehouse will enable the data engineer to process large numbers of queries quickly and cost-effectively?
A. Pro SQL Warehouse
B. Serverless SQL Warehouse
C. Serverless compute for notebooks
D. Classic SQL Warehouse
Answer: (SHOW ANSWER)

NEW QUESTION: 19
A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.
Which of the following describes why Auto Loader inferred all of the columns to be of the string type?
A. There was a type mismatch between the specific schema and the inferred schema
B. JSON data is a text-based format
C. Auto Loader only works with string data
D. All of the fields had at least one null value
E. Auto Loader cannot infer the schema of ingested data
Answer: (SHOW ANSWER)

JSON data is a text-based format that represents data as a collection of name-value pairs. By default, when Auto Loader infers the schema of JSON data, it treats all columns as strings. This is because JSON data can have varying data types for the same column across different files or records, and Auto Loader does not attempt to reconcile these differences. For example, a column named "age" may have integer values in some files, but string values in others. To avoid data loss or errors, Auto Loader infers the column as a string type.
However, Auto Loader also provides an option to infer more precise column types based on the sample data.
This option is called cloudFiles.inferColumnTypes and it can be set to true or false. When set to true, Auto Loader tries to infer the exact data types of the columns, such as integers, floats, booleans, or nested structures. When set to false, Auto Loader infers all columns as strings. The default value of this option is false. References: Configure schema inference and evolution in Auto Loader, Schema inference with auto loader (non-DLT and DLT), Using and Abusing Auto Loader's Inferred Schema, Explicit path to data or a defined schema required for Auto loader.

NEW QUESTION: 20

A data engineer is working on a personal laptop and needs to perform complex transformations on data stored in a Delta Lake on cloud storage. The engineer decides to use Databricks Connect to interact with Databricks clusters and work in their local IDE.
How does Databricks Connect enable the engineer to develop, test, and debug code seamlessly on their local machine while interacting with Databricks clusters?

A. By providing a local environment that mimics the Databricks runtime, enabling the engineer to develop, test, and debug code using their preferred ide
B. By allowing direct execution of Spark jobs from the local machine without needing a network connection
C. By providing a local environment that mimics the Databricks runtime, enabling the engineer to develop, test, and debug code using a specific IDE that is required by Databricks
D. By providing a local environment that mimics the Databricks runtime, enabling the engineer to develop, test, and debug code only through Databricks' own web interface
Answer: (SHOW ANSWER)

NEW QUESTION: 21
Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?
A. When they are working interactively with a small amount of data
B. When they are running automated reports to be refreshed as quickly as possible
C. When they are working with SQL within Databricks SQL
D. When they are concerned about the ability to automatically scale with larger data
E. When they are manually running reports with a large amount of data
Answer: (SHOW ANSWER)
The scenario in which a data engineer will want to use a single-node cluster is when they are working interactively with a small amount of data. A single-node cluster is a cluster consisting of an Apache Spark driver and no Spark workers1. A single-node cluster supports Spark jobs and all Spark data sources, including Delta Lake1. A single-node cluster is helpful for single-node machine learning workloads that use Spark to load and save data, and for lightweight exploratory data analysis1. A single-node cluster can run Spark locally, spawn one executor thread per logical core in the cluster, and save all log output in the driver log1. A single-node cluster can be created by selecting the Single Node button when configuring a cluster1.
The other options are not suitable for using a single-node cluster. When running automated reports to be refreshed as quickly as possible, a data engineer will want to use a multi-node cluster that can scale up and down automatically based on the workload demand2. When working with SQL within Databricks SQL, a data engineer will want to use a SQL Endpoint that can execute SQL queries on a serverless pool or an existing cluster3. When concerned about the ability to automatically scale with larger data, a data engineer will want to use a multi-node cluster that can leverage the Databricks Lakehouse Platform and the Delta Engine to handle large-scale data processing efficiently and reliably4. When manually running reports with a large amount of data, a data engineer will want to use a multi-node cluster that can distribute the computation across multiple workers and leverage the Spark UI to monitor the performance and troubleshoot the issues.

1: Single Node clusters | Databricks on AWS
2: Autoscaling | Databricks on AWS
3: SQL Endpoints | Databricks on AWS
4: Databricks Lakehouse Platform | Databricks on AWS
[Spark UI | Databricks on AWS]

NEW QUESTION: 22
A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.
Which of the following approaches can the data engineer use to set up the new task?
A. They can clone the existing task in the existing Job and update it to run the new notebook.
B. They can create a new task in the existing Job and then add it as a dependency of the original task.
C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.
D. They can create a new job from scratch and add both tasks to run concurrently.
E. They can clone the existing task to a new Job and then edit it to run the new notebook.
Answer: (SHOW ANSWER)
To set up the new task to run a new notebook prior to the original task in a single-task Job, the data engineer can use the following approach: In the existing Job, create a new task that corresponds to the new notebook that needs to be run. Set up the new task with the appropriate configuration, specifying the notebook to be executed and any necessary parameters or dependencies. Once the new task is created, designate it as a dependency of the original task in the Job configuration. This ensures that the new task is executed before the original task.

NEW QUESTION: 23
An organization plans to share a large dataset stored in a Databricks workspace on AWS with a partner organization whose Databricks workspace is hosted on Azure. The data engineer wants to minimize data transfer costs while ensuring secure and efficient data sharing.
Which strategy will reduce data egress costs associated with cross-cloud data sharing?
A. Using Delta Sharing without any additional configurations
B. Configure VPN connection between AWS and Azure for faster data sharing
C. Sharing data via pre-signed URLs without monitoring egress costs
D. Migrating the dataset to Cloudflare R2 object storage before sharing
Answer: (SHOW ANSWER)

NEW QUESTION: 24
A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.
Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?
A. Databricks Repos automatically saves development progress
B. Databricks Repos supports the use of multiple branches
C. Databricks Repos allows users to revert to previous versions of a notebook
D. Databricks Repos provides the ability to comment on specific changes
E. Databricks Repos is wholly housed within the Databricks Lakehouse Platform
Answer: (SHOW ANSWER)

Databricks Repos is a visual Git client and API in Databricks that supports common Git operations such as cloning, committing, pushing, pulling, and branch management. Databricks Notebooks versioning is a legacy feature that allows users to link notebooks to GitHub repositories and perform basic Git operations. However, Databricks Notebooks versioning does not support the use of multiple branches for development work, which is an advantage of using Databricks Repos. With Databricks Repos, users can create and manage branches for different features, experiments, or bug fixes, and merge, rebase, or resolve conflicts between them. Databricks recommends using a separate branch for each notebook and following data science and engineering code development best practices using Git for version control, collaboration, and CI/CD.
References: Git integration with Databricks Repos - Azure Databricks | Microsoft Learn, Git version control for notebooks (legacy) | Databricks on AWS, Databricks Repos Is Now Generally Available - New 'Files' Feature in
..., Databricks Repos - What it is and how we can use it | Adatis.

NEW QUESTION: 25
A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.
Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?
A. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to
"Reliability Optimized."
B. They can turn on the Auto Stop feature for the SQL endpoint.
C. They can increase the cluster size of the SQL endpoint.
D. They can turn on the Serverless feature for the SQL endpoint.
E. They can increase the maximum bound of the SQL endpoint's scaling range
Answer: (SHOW ANSWER)
Option D is the correct answer because it enables the Serverless feature for the SQL endpoint, which allows the endpoint to automatically scale up and down based on the query load. This way, the endpoint can handle more concurrent queries and reduce the time it takes to return results. The Serverless feature also reduces the cold start time of the endpoint, which is the time it takes to start the cluster when a query is submitted to a non- running endpoint. The Serverless feature is available for both AWS and Azure Databricks platforms.
Databricks SQL Serverless, Serverless SQL endpoints, New Performance Improvements in Databricks SQL

NEW QUESTION: 26
Which of the following SQL keywords can be used to convert a table from a long format to a wide format?
A. PIVOT
B. CONVERT
C. WHERE
D. TRANSFORM
E. SUM
Answer: (SHOW ANSWER)
The SQL keyword that can be used to convert a table from a long format to a wide format is PIVOT. The PIVOT clause is used to rotate the rows of a table into columns of a new table1. The PIVOT clause can aggregate the values of a column based on the distinct values of another column, and use those values as the column names of the new table1. The PIVOT clause can be useful for transforming data from a long format, where each row represents an observation with multiple attributes, to a wide format, where each row represents an observation with a single attribute and multiple values2. For example, the PIVOT clause can be used to convert a table that contains the sales of different products by different regions into a table that contains the sales of each product by each region as separate columns1.

The other options are not suitable for converting a table from a long format to a wide format. CONVERT is a function that can be used to change the data type of an expression3. WHERE is a clause that can be used to filter the rows of a table based on a condition4. TRANSFORM is a keyword that can be used to apply a user- defined function to a group of rows in a table5. SUM is a function that can be used to calculate the total of a numeric column.
1: PIVOT | Databricks on AWS
2: Reshaping Data - Long vs Wide Format | Databricks on AWS
3: CONVERT | Databricks on AWS
4: WHERE | Databricks on AWS
5: TRANSFORM | Databricks on AWS
[SUM | Databricks on AWS]

NEW QUESTION: 27
A data engineer manages multiple external tables linked to various data sources. The data engineer wants to manage these external tables efficiently and ensure that only the necessary permissions are granted to users for accessing specific external tables.
How should the data engineer manage access to these external tables?
A. Use Unity Catalog to manage access controls and permissions for each external table individually.
B. Create a single user role with full access to all external tables and assign it to all users.
C. Grant permissions on the Databricks workspace level, which will automatically apply to all external tables.
D. Set up Azure Blob Storage permissions at the container level, allowing access to all external tables.
Answer: (SHOW ANSWER)

NEW QUESTION: 28
A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:

```
transactions_df = (spark.read
    .schema(schema)
    .format("delta")
    .table("transactions")
)
```


Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?
A. Replace predict with a stream-friendly prediction function
B. Replace schema(schema) with option ("maxFilesPerTrigger", 1)
C. Replace "transactions" with the path to the location of the Delta table
D. Replace format("delta") with format("stream")
E. Replace spark.read with spark.readStream
Answer: (SHOW ANSWER)
To read from a stream source, the data engineer needs to use the spark.readStream method instead of the spark.

read method. The spark.readStream method returns a DataStreamReader object that can be used to specify the details of the input source, such as the format, the schema, the path, and the options. The spark.read method is only suitable for batch processing, not streaming processing. The other changes are not necessary or correct for reading from a stream source. References: Structured Streaming Programming Guide, Read a stream, Databricks Data Sources

NEW QUESTION: 29
A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.
Which command can be used to grant full permissions on the database to the new data engineering team?
A. grant all privileges on table sales TO team;
B. GRANT SELECT ON TABLE sales TO team;
C. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
D. GRANT ALL PRIVILEGES ON TABLE team TO sales;
Answer: (SHOW ANSWER)
To grant full privileges on a table such as 'sales' to a group like 'team', the correct SQL command in Databricks is:
GRANT ALL PRIVILEGES ON TABLE sales TO team;
This command assigns all available privileges, including SELECT, INSERT, UPDATE, DELETE, and any other data manipulation or definition actions, to the specified team.
This is typically necessary when a team needs full control over a table to manage and manipulate it as part of a project or ongoing maintenance.
References:Databricks documentation on SQL permissions: SQL Permissions in Databricks

NEW QUESTION: 30
Which type of workloads are compatible with Auto Loader?
A. Streaming workloads
B. Machine learning workloads
C. Batch workloads
D. Serverless workloads
Answer: (SHOW ANSWER)

NEW QUESTION: 31
Which of the following Git operations must be performed outside of Databricks Repos?
A. Commit
B. Pull
C. Push
D. Clone
E. Merge
Answer: (SHOW ANSWER)
Databricks Repos is a visual Git client and API in Databricks that supports common Git operations such as commit, pull, push, branch management, and visual comparison of diffs when committing1. However, merge is not supported in the Git dialog2. You need to use the Repos UI or your Git provider to merge branches3. Merge is a way to combine the commit history from one branch into another branch1. During a merge, a merge conflict is encountered when Git cannot automatically combine code from one branch into another. Merge conflicts require manual resolution before a merge can be completed1. References: 4: Run Git operations on Databricks Repos4, 1: CI/CD techniques with Git and Databricks Repos1, 3: Collaborate in Repos3, 2: Databricks Repos - What it is and how we can use it2.

Databricks Repos is a visual Git client and API in Databricks that supports common Git operations such as commit, pull, push, merge, and branch management. However, to clone a remote Git repository to a Databricks repo, you must use the Databricks UI or API. You cannot clone a Git repo using the CLI through a cluster's web terminal, as the files won't display in the Databricks UI1. References: 1: Run Git operations on Databricks Repos | Databricks on AWS2

NEW QUESTION: 32
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)
A webhook alert destination is a notification destination that allows Databricks to send HTTP POST requests to a third-party endpoint when an alert is triggered. This enables the data engineer to integrate Databricks alerts with their preferred messaging or collaboration platform, such as Slack, Microsoft Teams, or PagerDuty. To set up a webhook alert destination, the data engineer needs to create and configure a webhook connector in their messaging platform, and then add the webhook URL to the Databricks notification destination. After that, the data engineer can create an alert for their Databricks SQL query, and select the webhook alert destination as the notification destination. The alert can be configured with a custom condition, such as when the number of stores with $0 in sales is greater than zero, and a custom message template, such as "Alert: {number_of_stores} stores have $0 in sales". The alert can also be configured with a recurrence interval, such as every hour, to check the query result periodically. When the alert condition is met, the data engineer and their team will receive a notification via the messaging webhook, with the custom message and a link to the Databricks SQL query. The other options are either not suitable for sending notifications via a messaging webhook (A, B, E), or not suitable for sending recurring notifications ©. References: Databricks Documentation - Manage notification destinations, Databricks Documentation - Create alerts for Databricks SQL queries, Databricks Documentation - Configure alert conditions and messages.

NEW QUESTION: 33
Which of the following data lakehouse features results in improved data quality over a traditional data lake?
A. A data lakehouse provides storage solutions for structured and unstructured data.
B. A data lakehouse supports ACID-compliant transactions.
C. A data lakehouse allows the use of SQL queries to examine data.
D. A data lakehouse stores data in open formats.

E. A data lakehouse enables machine learning and artificial Intelligence workloads.
Answer: (SHOW ANSWER)
A data lakehouse is a data management architecture that combines the flexibility, cost-efficiency, and scale of data lakes with the data management and ACID transactions of data warehouses, enabling business intelligence (BI) and machine learning (ML) on all data12. One of the key features of a data lakehouse is that it supports ACID-compliant transactions, which means that it ensures data integrity, consistency, and isolation across concurrent read and write operations3. This feature results in improved data quality over a traditional data lake, which does not support transactions and may suffer from data corruption, duplication, or inconsistency due to concurrent or streaming data ingestion and processing . References: 1: What is a Data Lakehouse? - Databricks 2: What is a Data Lakehouse? Definition, features & benefits. - Qlik 3: ACID Transactions - Databricks : [Data Lake vs Data Warehouse: Key Differences] : [Data Lakehouse: The Future of Data Engineering]

NEW QUESTION: 34
A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.
They run the following command:

```
CREATE TABLE new_table
_____
OPTIONS (
  header = "true",
  delimiter = "|"
)
LOCATION "path/to/csv"
```

Which of the following lines of code fills in the above blank to successfully complete the task?
A. USING DELTA
B. FROM CSV
C. None of these lines of code are needed to successfully complete the task
D. FROM "path/to/csv"
E. USING CSV
Answer: (SHOW ANSWER)

NEW QUESTION: 35
Which tool is used by Auto Loader to process data incrementally?
A. Spark Structured Streaming
B. Unity Catalog
C. Checkpointing
D. Databricks SQL
Answer: (SHOW ANSWER)
Auto Loader in Databricks utilizes Spark Structured Streaming for processing data incrementally. This allows Auto Loader to efficiently ingest streaming or batch data at scale and to recognize new data as it arrives in cloud storage. Spark Structured Streaming provides the underlying engine that supports various incremental data loading capabilities like schema inference and file notification mode, which are crucial for the dynamic nature of data lakes.

References:Databricks documentation on Auto Loader: Auto Loader Overview

NEW QUESTION: 36
A Databricks single-task workflow fails at the last task due to an error in a notebook. The data engineer fixes the mistake in the notebook. What should the data engineer do to rerun the workflow?
A. Restart the Cluster
B. Switch the cluster
C. Repair the task
D. Rerun the pipeline
Answer: (SHOW ANSWER)

NEW QUESTION: 37
Which two components function in the DB platform architecture's control plane? (Choose two.)
A. Compute Orchestration
B. Unity Catalog
C. Virtual Machines
D. Compute
E. Serverless Compute
Answer: (SHOW ANSWER)

NEW QUESTION: 38
A data engineer is designing an ETL pipeline to process both streaming and batch data from multiple sources The pipeline must ensure data quality, handle schema evolution, and provide easy maintenance. The team is considering using Delta Live Tables (DLT) in Databricks to achieve these goals. They want to understand the key features and benefits of DLT that make it suitable for this use case.
Why is Delta Live Tables (DLT) an appropriate choice?
A. Requires custom code for data quality checks, no support for streaming data, and complex pipeline maintenance
B. Automatic data quality checks, built-in support for schema evolution, and declarative pipeline development
C. Supports only batch processing, no data versioning, and high infrastructure costs
D. Manual schema enforcement, high operational overhead, and limited scalability
Answer: B (LEAVE A REPLY)

NEW QUESTION: 39
Which of the following must be specified when creating a new Delta Live Tables pipeline?
A. A key-value pair configuration
B. The preferred DBU/hour cost
C. A path to cloud storage location for the written data
D. A location of a target database for the written data
E. At least one notebook library to be executed
Answer: (SHOW ANSWER)
Option E is the correct answer because it is the only mandatory requirement when creating a new Delta Live Tables pipeline. A pipeline is a data processing workflow that contains materialized views and streaming tables declared in Python or SQL source files. Delta Live Tables infers the dependencies between these tables and ensures

updates occur in the correct order. To create a pipeline, you need to specify at least one notebook library to be executed, which contains the Delta Live Tables syntax. You can also specify multiple libraries of different languages within your pipeline. The other options are optional or not applicable for creating a pipeline. Option A is not required, but you can optionally provide a key-value pair configuration to customize the pipeline settings, such as the storage location, the target schema, the notifications, and the pipeline mode.
Option B is not applicable, as the DBU/hour cost is determined by the cluster configuration, not the pipeline creation. Option C is not required, but you can optionally specify a storage location for the output data from the pipeline. If you leave it empty, the system uses a default location. Option D is not required, but you can optionally specify a location of a target database for the written data, either in the Hive metastore or the Unity Catalog.
Tutorial: Run your first Delta Live Tables pipeline, What is Delta Live Tables?, Create a pipeline, Pipeline configuration.

NEW QUESTION: 40

A data engineer needs to combine sales data from an on-premises PostgreSQL database with customer data in Azure Synapse for a comprehensive report. The goal is to avoid data duplication and ensure up-to-date information How should the data engineer achieve this using Databricks?

A. Export data from both sources to CSV files and upload them to Databricks
B. Develop custom ETL pipelines to ingest data into Databricks
C. Manually synchronize data from both sources into a single database
D. Use Lakehouse Federation to query both data sources directly
Answer: (SHOW ANSWER)

NEW QUESTION: 41

A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.
Which of the following commands can be used to grant the necessary permission on the entire database to the new team?
A. GRANT VIEW ON CATALOG customers TO team;
B. GRANT CREATE ON DATABASE customers TO team;
C. GRANT USAGE ON CATALOG team TO customers;
D. GRANT CREATE ON DATABASE team TO customers;
E. GRANT USAGE ON DATABASE customers TO team;
Answer: (SHOW ANSWER)
The correct command to grant the necessary permission on the entire database to the new team is to use the GRANT USAGE command. The GRANT USAGE command grants the principal the ability to access the securable object, such as a database, schema, or table. In this case, the securable object is the database customers, and the principal is the group team. By granting usage on the database, the team will be able to see what tables already exist in the database. Option E is the only option that uses the correct syntax and the correct privilege type for this scenario. Option A uses the wrong privilege type (VIEW) and the wrong securable object (CATALOG). Option B uses the wrong privilege type (CREATE), which would allow the team to create new tables in the database, but not necessarily see the existing ones. Option C uses the wrong securable object (CATALOG) and the wrong principal (customers). Option D uses the wrong securable object (team) and the wrong principal (customers). References: GRANT, Privilege types, Securable objects, Principals

NEW QUESTION: 42

A departing platform owner currently holds ownership of multiple catalogs and controls storage credentials and external locations. The data engineer wants to ensure continuity: transfer catalog ownership to the platform team group, delegate ongoing privilege management, and retain the ability to receive and share data via Delta Sharing.
Which role must be in place to perform these actions across the metastore?

A. Account Admin, because account admins can only create metastores but cannot change ownership of catalogs.
B. Workspace Admin, because workspace admins can transfer ownership of any Unity Catalog object.
C. Metastore Admin, because metastore admins can transfer ownership and manage privileges across all metastore objects, including shares and recipients.
D. Catalog Owner, because catalog owners can transfer any object in any catalog in the metastore.
Answer: (SHOW ANSWER)
In Unity Catalog, the Metastore Admin role has the highest level of authority within a metastore and is responsible for governing all data assets registered in that metastore. Metastore admins can transfer ownership of catalogs, manage privileges across all securable objects (catalogs, schemas, tables, views), and administer storage credentials, external locations, and Delta Sharing objects such as shares and recipients.
This makes the role essential when a platform owner departs and continuity must be maintained at an organizational level. Workspace admins (option B) manage workspace-level settings but do not automatically have authority over Unity Catalog ownership unless explicitly granted. Catalog owners (option D) are limited to a single catalog and cannot perform actions across the entire metastore. Account admins (option A) operate at the account level and are primarily responsible for creating metastores and assigning admins, not for day-to-day governance within an existing metastore. Databricks documentation clearly defines the Metastore Admin as the role required to centrally manage ownership, privileges, and data sharing across the metastore in Databricks.

NEW QUESTION: 43
In which of the following scenarios should a data engineer use the MERGE INTO command instead of the INSERT INTO command?
A. When the location of the data needs to be changed
B. When the target table is an external table
C. When the source table can be deleted
D. When the target table cannot contain duplicate records
E. When the source is not a Delta table
Answer: (SHOW ANSWER)
The MERGE INTO command is used to perform upserts, which are a combination of insertions and updates, based on a source table into a target Delta table1. The MERGE INTO command can handle scenarios where the target table cannot contain duplicate records, such as when there is a primary key or a unique constraint on the target table. The MERGE INTO command can match the source and target rows based on a merge condition and perform different actions depending on whether the rows are matched or not. For example, the MERGE INTO command can update the existing target rows with the new source values, insert the new source rows that do not exist in the target table, or delete the target rows that do not exist in the source table1.
The INSERT INTO command is used to append new rows to an existing table or create a new table from a query result2. The INSERT INTO command does not perform any updates or deletions on the existing target table rows. The INSERT INTO command can handle scenarios where the location of the data needs to be changed, such as when the data needs to be moved from one table to another, or when the data needs to be partitioned by a certain column2. The INSERT INTO command can also handle scenarios where the target table is an external table, such as when the data is stored in an external storage system like Amazon S3 or Azure Blob Storage3. The INSERT INTO command can also handle scenarios where the source table can be deleted, such as when the source table is a temporary table or a view4. The INSERT INTO command can also handle scenarios where the source is not a Delta table, such as when the source is a Parquet, CSV, JSON, or Avro file5.
1: MERGE INTO | Databricks on AWS
2: [INSERT INTO | Databricks on AWS]
3: [External tables | Databricks on AWS]
4: [Temporary views | Databricks on AWS]
5: [Data sources | Databricks on AWS]

NEW QUESTION: 44

A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.
Which of the following approaches can the data engineer take to identify the table that is dropping the records?
A. They can set up separate expectations for each table when developing their DLT pipeline.
B. They cannot determine which table is dropping the records.
C. They can set up DLT to notify them via email when records are dropped.
D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.
E. They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.
Answer: (SHOW ANSWER)
One of the features of DLT is that it provides data quality metrics for each dataset in the pipeline, such as the number of records that pass or fail expectations, the number of records that are dropped, and the number of records that are written to the target. These metrics can be accessed from the DLT pipeline page, where the data engineer can click on each table and view the data quality statistics for the latest update or any previous update. This way, they can identify which table is dropping the records and why.
References:
* Monitor Delta Live Tables pipelines
* Manage data quality with Delta Live Tables

NEW QUESTION: 45
A data engineer wants to create a new table containing the names of customers who live in France.
They have written the following command:
CREATE TABLE customersInFrance
_____ AS
SELECT id,
firstName,
lastName
FROM customerLocations
WHERE country = 'FRANCE';
A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).
Which line of code fills in the above blank to successfully complete the task?
A. COMMENT "Contains PIT
B. 511
C. "COMMENT PII"
D. TBLPROPERTIES PII
Answer: D (LEAVE A REPLY)
To include a property indicating that a table contains personally identifiable information (PII), the TBLPROPERTIES keyword is used in SQL to add metadata to a table. The correct syntax to define a table property for PII is as follows:
CREATE TABLE customersInFrance
USING DELTA
TBLPROPERTIES ('PII' = 'true')
AS
SELECT id,
firstName,

lastName
FROM customerLocations
WHERE country = 'FRANCE';
The TBLPROPERTIES ('PII' = 'true') line correctly sets a table property that tags the table as containing personally identifiable information. This is in accordance with organizational policies for handling sensitive information.
References:Databricks documentation on Delta Lake: Delta Lake on Databricks

NEW QUESTION: 46
A team creates YAML manifests that declare jobs, resources, and dependencies, then deploys them to Databricks using the Databricks CLI. The deployment succeeds.
Which feature are they using?
A. Databricks Asset Bundles
B. GitHub
C. Terraform
D. DataOps
Answer: (SHOW ANSWER)
Databricks Asset Bundles are a Databricks CLI feature that lets teams define Databricks projects "as code" using YAML configuration files. In a bundle, metadata is expressed in YAML (for example, a required databricks.yml at the root), and the configuration can declare resources (such as Databricks Jobs/workflows and other supported assets), artifacts, and deployment targets. After defining these YAML manifests, teams use the Databricks CLI to validate, deploy, and run the bundle into a Databricks workspace, which matches the scenario described (YAML + CLI deployment). This approach standardizes how workflows and related resources are packaged and promoted across environments (dev/test/prod) while keeping definitions version-controlled and reproducible. GitHub is only a hosting/version-control platform (not the YAML deployment feature itself), Terraform is a separate IaC tool, and "DataOps" is a methodology rather than a specific Databricks feature. (Databricks Documentation)

NEW QUESTION: 47
A data engineer has joined an existing project and they see the following query in the project repository:
CREATE STREAMING LIVE TABLE loyal_customers AS
SELECT customer_id -
FROM STREAM(LIVE.customers)
WHERE loyalty_level = 'high';
Which of the following describes why the STREAM function is included in the query?
A. The STREAM function is not needed and will cause an error.
B. The table being created is a live table.
C. The customers table is a streaming live table.
D. The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.

E. The data in the customers table has been updated since its last run.
Answer: (SHOW ANSWER)
The STREAM function is used to process data from a streaming live table or view, which is a table or view that contains data that has been added only since the last pipeline update. Streaming live tables and views are stateful, meaning that they retain the state of the previous pipeline run and only process new data based on the current query. This is useful for incremental processing of streaming or batch data sources. The customers table in the query is a streaming live table, which means that it contains the latest data from the source. The STREAM function enables the query to read the data from the customers table incrementally and create another streaming live table named loyal_customers, which contains the customer IDs of the customers with high loyalty level. References: Difference between LIVE TABLE and STREAMING LIVE TABLE, CREATE STREAMING TABLE, Load data using streaming tables in Databricks SQL.

NEW QUESTION: 48
Identify how the count_if function and the count where x is null can be used Consider a table random_values with below data.
What would be the output of below query?
select count_if(col > 1) as count_a. count(*) as count_b.count(col1) as count_c from random_values col1
0
1
2
NULL -
2
3
A. 4 6 5
B. 3 6 6
C. 4 6 6
D. 3 6 5
Answer: (SHOW ANSWER)

NEW QUESTION: 49
A data engineer is processing ingested streaming tables and needs to filter out NULL values in the order_datetime column from the raw streaming table orders_raw and store the results in a new table orders_valid using DLT.
Which code snippet should the data engineer use?
A.

```
CREATE OR REFRESH STREAMING TABLE orders_valid(
  CONSTRAINT valid_date
  EXPECT (order_datetime IS NOT NULL)
  ON VIOLATION DROP ROW
)
AS SELECT * FROM STREAM orders_raw;
```

B.
```
CREATE OR REFRESH STREAMING TABLE orders_valid
AS
SELECT *
FROM STREAM(orders_raw)
WHERE order_datetime IS NOT NULL;
```

C.
```
CREATE OR REFRESH STREAMING TABLE orders_valid
(
  CONSTRAINT valid_date EXPECT (order_datetime IS NOT NULL)
  ON VIOLATION DROP ROW
)
AS SELECT * FROM orders_raw;
```

D.
```
CREATE OR REPLACE STREAMING TABLE orders_valid
(
  FILTER (order_datetime IS NOT NULL)
)
AS SELECT * FROM STREAM(orders_raw);
```

Answer: (SHOW ANSWER)

NEW QUESTION: 50

A data engineer at a company that uses Databricks with Unity Catalog needs to share a collection of tables with an external partner who also uses a Databricks workspace enabled for Unity Catalog. The data engineer decides to use Delta Sharing to accomplish this.
What is the first piece of information the data engineer should request from the external partner to set up Delta Sharing?

A. The IP address of their Databricks workspace
B. The sharing identifier of their Unity Catalog metastore
C. The name of their Databricks cluster
D. Their Databricks account password
Answer: (SHOW ANSWER)

NEW QUESTION: 51

A data engineering team has two tables. The first table march_transactions is a collection of all retail transactions in the month of March. The second table april_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables.
Which of the following commands should be run to create a new table all_transactions that contains all records from march_transactions and april_transactions without duplicate records?

A. CREATE TABLE all_transactions ASSELECT * FROM march_transactionsINNER JOIN SELECT * FROM april_transactions;
B. CREATE TABLE all_transactions ASSELECT * FROM march_transactionsUNION SELECT * FROM april_transactions;

C. CREATE TABLE all_transactions ASSELECT * FROM march_transactionsOUTER JOIN SELECT * FROM april_transactions;
D. CREATE TABLE all_transactions ASSELECT * FROM march_transactionsINTERSECT SELECT * from april_transactions;
E. CREATE TABLE all_transactions ASSELECT * FROM march_transactionsMERGE SELECT * FROM april_transactions;
Answer: B (LEAVE A REPLY)
The correct command to create a new table that contains all records from two tables without duplicate records is to use the UNION operator. The UNION operator combines the results of two queries and removes any duplicate rows. The INNER JOIN, OUTER JOIN, and MERGE operators do not remove duplicate rows, and the INTERSECT operator only returns the rows that are common to both tables. Therefore, option B is the only correct answer. References: Databricks SQL Reference - UNION, Databricks SQL Reference - JOIN, Databricks SQL Reference - MERGE, [Databricks SQL Reference - INTERSECT]

NEW QUESTION: 52
A data engineer needs access to a table new_uable, but they do not have the correct permissions. They can ask the table owner for permission, but they do not know who the table owner is.
Which approach can be used to identify the owner of new_table?
A. There is no way to identify the owner of the table
B. Review the Owner field in the table ' s page in the cloud storage solution
C. Review the Permissions tab in the table ' s page in Data Explorer
D. Review the Owner field in the table's page in Data Explorer
Answer: (SHOW ANSWER)
To find the owner of a table in Databricks, one can utilize the Data Explorer feature. The Data Explorer provides detailed information about various data objects, including tables. By navigating to the specific table ' s page in Data Explorer, a data engineer can review the Owner field, which identifies the individual or role that owns the table. This information is crucial for obtaining the necessary permissions or for any administrative actions related to the table.
References:Databricks documentation on Data Explorer: Using Data Explorer in Databricks

NEW QUESTION: 53
An organization is looking for an optimized storage layer that supports ACID transactions and schema enforcement. Which technology should the organization use?
A. Delta Lake
B. Data lake
C. Unity Catalog
D. Cloud File Storage
Answer: (SHOW ANSWER)

NEW QUESTION: 54
A data engineer streams customer orders into a Kafka topic (orders_topic) and is currently writing the ingestion script of a DLT pipeline. The data engineer needs to ingest the data from Kafka brokers to DLT using Databricks What is the correct code for ingesting the data?

A.
```
CREATE STREAMING LIVE TABLE orders_raw
AS SELECT
    value:order_id AS order_id,
    value:customer_id AS customer_id,
    value:amount AS amount,
    value:order_status AS order_status,
    value:order_timestamp AS order_timestamp
FROM cloud_files("kafka://broker:9092/orders_topic", "json");
```
B.
```
CREATE LIVE TABLE orders_raw
AS SELECT
    CAST(value AS STRING) AS json_data
FROM STREAM kafka.`broker:9092/orders_topic`;
```
C.
```
import dlt

@dlt.table(
    name = "orders_raw",
)
def orders_raw():
    return (
        spark.readStream
            .format("kafka")
            .option("kafka.bootstrap.servers", "broker:9092")
            .option("subscribe", "orders_topic")
            .option("startingOffsets", "earliest")
            .load()
    )
```
D.
```
CREATE LIVE TABLE orders_raw
AS SELECT
    CAST(value AS STRING) AS json_data
FROM STREAM kafka.`broker:9092/orders_topic`;
```

Answer: (SHOW ANSWER)

NEW QUESTION: 55

Which SQL keyword can be used to convert a table from a long format to a wide format?

A. TRANSFORM

B. CONVERT

C. SUM

D. PIVOT

Answer: D (LEAVE A REPLY)

NEW QUESTION: 56
A Data Engineer is building a simple data pipeline using Delta Live Tables (DLT) in Databricksto ingest customer data. The raw customer data is stored in a cloud storage location in JSON format. The task is to create a DLT pipeline that reads the rawJSON data and writes it into a Delta table for further processing.
Which code snippet will correctly ingest the raw JSON data and create a Delta table using DLT?

A.
```
import dlt

@dlt.table
def raw_customers():
    return spark.read.json("s3://my-bucket/raw-customers/")
```

B.
```
import dlt

@dlt.table
def raw_customers():
    return spark.read.format("csv").load("s3://my-bucket/raw-customers/")
```

C.
```
import dlt

@dlt.table
def raw_customers():
    return spark.read.format("parquet").load("s3://my-bucket/raw-customers/")
```

D.
```
import dlt

@dlt.view
def raw_customers():
    return spark.format.json("s3://my-bucket/raw-customers/")
```

Answer: A (LEAVE A REPLY)

NEW QUESTION: 57
A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start.
Which of the following actions can the data engineer perform to improve the start up time for the clusters used for the Job?
A. They can use endpoints available in Databricks SQL
B. They can use jobs clusters instead of all-purpose clusters
C. They can configure the clusters to be single-node
D. They can use clusters that are from a cluster pool
E. They can configure the clusters to autoscale for larger data sizes

Answer: (SHOW ANSWER)
The best action that the data engineer can perform to improve the start up time for the clusters used for the Job is to use clusters that are from a cluster pool. A cluster pool is a set of idle clusters that can be used by jobs or interactive sessions. By using a cluster pool, the data engineer can avoid the cluster creation time and reduce the latency of the tasks. Cluster pools also offer cost savings and resource efficiency, as they can be shared by multiple users and jobs.
Option A is not relevant, as endpoints available in Databricks SQL are used for creating and managing SQL analytics workloads, not for improving cluster start up time.
Option B is not correct, as jobs clusters and all-purpose clusters have similar start up times. Jobs clusters are clusters that are dedicated to run a single job and are terminated when the job is completed. All-purpose clusters are clusters that can be used for multiple purposes, such as interactive sessions, notebooks, or multiple jobs. Both types of clusters can benefit from using a cluster pool.
Option C is not advisable, as configuring the clusters to be single-node will reduce the parallelism and performance of the tasks. Single-node clusters are clusters that have only one worker node and are typically used for testing or development purposes. They are not suitable for running production jobs that require high scalability and fault tolerance.
Option E is not helpful, as configuring the clusters to autoscale for larger data sizes will not affect the start up time of the clusters. Autoscaling is a feature that allows clusters to dynamically adjust the number of worker nodes based on the workload. It can help optimize the resource utilization and cost efficiency of the clusters, but it does not speed up the cluster creation process.
Cluster Pools
Jobs
Clusters
[Databricks Data Engineer Professional Exam Guide]

NEW QUESTION: 58
A data engineer is writing a script that is meant to ingest new data from cloud storage. In the event of the Schema change, the ingestion should fail. It should fail until the changes downstream source can be found and verified as intended changes.
Which command will meet the requirements?
A. addNewColumns
B. none
C. failOnNewColumns
D. rescue
Answer: (SHOW ANSWER)

NEW QUESTION: 59
Which of the following describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?
A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.
C. CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.
D. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
E. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.
Answer: (SHOW ANSWER)
A streaming live table or view processes data that has been added only since the last pipeline update.
Streaming tables and views are stateful; if the defining query changes, new data will be processed based on the new query and existing data is not recomputed. This is useful when data needs to be processed incrementally, such as when ingesting streaming data sources or performing incremental loads from batch data sources. A live

table or view, on the other hand, may be entirely computed when possible to optimize computation resources and time. This is suitable when data needs to be processed in full, such as when performing complex transformations or aggregations that require scanning all the data. References: Difference between LIVE TABLE and STREAMING LIVE TABLE, CREATE STREAMING TABLE, Load data using streaming tables in Databricks SQL.

NEW QUESTION: 60

A data engineer has a Python variable table_name that they would like to use in a SQL query. They want to construct a Python code block that will run the query using table_name.
They have the following incomplete code block:
____(f"SELECT customer_id, spend FROM {table_name}")
Which of the following can be used to fill in the blank to successfully complete the task?

A. spark.delta.sql
B. spark.delta.table
C. spark.table
D. dbutils.sql
E. spark.sql
Answer: (SHOW ANSWER)
The spark.sql method can be used to execute SQL queries programmatically and return the result as a DataFrame. The spark.sql method accepts a string argument that contains a valid SQL statement. The data engineer can use a formatted string literal (f-string) to insert the Python variable table_name into the SQL query. The other methods are either invalid or not suitable for running SQL queries. References: Running SQL Queries Programmatically, Formatted string literals, spark.sql

NEW QUESTION: 61
A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.
Which of the following commands can be used to grant full permissions on the database to the new data engineering team?
A. GRANT ALL PRIVILEGES ON TABLE sales TO team;
B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
C. GRANT SELECT ON TABLE sales TO team;
D. GRANT USAGE ON TABLE sales TO team;
E. GRANT ALL PRIVILEGES ON TABLE team TO sales;
Answer: A (LEAVE A REPLY)
To grant full permissions on a table to a user or a group, you can use the GRANT ALL PRIVILEGES ON TABLE statement. This statement will grant all the possible privileges on the table, such as SELECT, CREATE, MODIFY, DROP, ALTER, etc. Option A is the only code block that follows this syntax correctly. Option B is incorrect, as it does not grant all the possible privileges on the table, but only a subset of them. Option C is incorrect, as it only grants the SELECT privilege on the table, which is not enough to fully manage the project. Option D is incorrect, as it grants the USAGE privilege on the table, which is not a valid privilege for tables. Option E is incorrect, as it grants all the privileges on the table team to the user or group sales, which is the opposite of what the question asks. References: Grant privileges on a table using SQL | Databricks on AWS, Grant privileges on a table using SQL - Azure Databricks, SQL Privileges - Databricks

NEW QUESTION: 62
Which SQL code snippet will correctly demonstrate a Data Definition Language (DDL) operation used to create a table?
A. ALTFR TABIF employees add column salary DECTMA(10,2);
B. INSERT INTO employees (id, name) VALUES (1, 'Alice');
C. DROP TABLE employees;
D. CRFATF tabif employees (id INT, name suing
Answer: (SHOW ANSWER)

NEW QUESTION: 63
A data engineer is inspecting an ETL pipeline based on a Pyspark job that consistently encounters performance bottlenecks. Based on developer feedback, the data engineer assumes the job is low on compute resources. To pinpoint the issue, the data engineer observes the Spark UI and finds out the job has a high CPU time vs Task time.
Which course of action should the data engineer take?
A. High CPU time vs Task time means an under-utilized cluster. The data engineer may need to repartition data to spread the jobs more evenly throughout the cluster.
B. High CPU time vs Task time means over-utilized memory and the need to increase parallelism
C. High CPU time vs Task time means a CPU over-utilized job. The data engineer may need to consider executor and core tuning or resizing the cluster
D. High CPU time vs Task time means efficient use of cluster and no change needed
Answer: C (LEAVE A REPLY)

NEW QUESTION: 64
A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.
Which of the following approaches can the tech lead use to identify why the notebook is running slowly as part of the Job?
A. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.
B. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.
C. There is no way to determine why a Job task is running slowly.
D. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
E. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.
Answer: (SHOW ANSWER)

NEW QUESTION: 65
Which method should a Data Engineer apply to ensure Workflows are being triggered on schedule?
A. Scheduled Workflows can reduce resource consumption and expense since the cluster runs only long enough to execute the pipeline.
B. Scheduled Workflows run continuously until manually stopped.
C. Scheduled Workflows require an always-running cluster, which is more expensive but reduces processing latency.
D. Scheduled Workflows process data as it arrives at configured sources.
Answer: (SHOW ANSWER)

NEW QUESTION: 66
A data engineer is getting a partner organization up to speed with Databricks account. Both teams share some business use cases. The data engineer has to share some of your Unity-Catalog managed delta tables and the notebook jobs creating those tables with the partner organization.
How can the data engineer seamlessly share the required information?
A. Zip all the code and share via email and allow data ingestion from your data lake
B. Data and Notebooks can be shared simply using Unity Catalog.
C. Share required datasets and notebooks via Delta Sharing. Manage permissions via Unity Catalog.
D. Share access to codebase via Github and allow them to ingest datasets from your Datalake.
Answer: (SHOW ANSWER)

NEW QUESTION: 67
A data engineer and data analyst are working together on a data pipeline. The data engineer is working on the raw, bronze, and silver layers of the pipeline using Python, and the data analyst is working on the gold layer of the pipeline using SQL The raw source of the pipeline is a streaming input. They now want to migrate their pipeline to use Delta Live Tables.
Which change will need to be made to the pipeline when migrating to Delta Live Tables?
A. The pipeline can have different notebook sources in SQL & Python.
B. The pipeline will need to be written entirely in SQL.
C. The pipeline will need to be written entirely in Python.
D. The pipeline will need to use a batch source in place of a streaming source.
Answer: (SHOW ANSWER)
When migrating to Delta Live Tables (DLT) with a data pipeline that involves different programming languages across various data layers, the migration does not require unifying the pipeline into a single language. Delta Live Tables support multi-language pipelines, allowing data engineers and data analysts to work in their preferred languages, such as Python for data engineering tasks (raw, bronze, and silver layers) and SQL for data analytics tasks (gold layer). This capability is particularly beneficial in collaborative settings and leverages the strengths of each language for different stages of data processing.
References:Databricks documentation on Delta Live Tables: Delta Live Tables Guide

NEW QUESTION: 68
A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.
Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?
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 a new webhook alert destination.
D. They can set up an Alert with one-time notifications.
E. They can set up an Alert without notifications.
Answer: (SHOW ANSWER)
A webhook alert destination is a way to send notifications to external applications or services via HTTP requests. A data engineer can use a webhook alert destination to notify their entire team via a messaging webhook, such as Slack or Microsoft Teams, whenever the number of NULL values in the input data reaches
100. To set up a webhook alert destination, the data engineer needs to do the following steps:
* In the Databricks SQL workspace, navigate to the Settings gear icon and select SQL Admin Console.

* Click Alert Destinations and click Add New Alert Destination.
* Select Webhook and enter the webhook URL and the optional custom template for the notification message.
* Click Create to save the webhook alert destination.
* In the Databricks SQL editor, create or open the query that returns the number of input records containing unexpected NULL values.
* Click the Create Alert icon above the editor window and configure the alert criteria, such as the value column, the condition, and the threshold.
* In the Notification section, select the webhook alert destination that was created earlier and click Create Alert. References: What are Databricks SQL alerts?, Monitor alerts, Monitoring Your Business with Alerts, Using Automation Runbook Webhooks To Alert on Databricks Status Updates.

NEW QUESTION: 69
A data engineer needs to process SQL queries on a large dataset with fluctuating workloads. The workload requires automatic scaling based on the volume of queries, without the need to manage or provision infrastructure. The solution should be cost-efficient and charge only for the compute resources used during query execution.
Which compute option should the data engineer use?
A. Databricks Jobs
B. Serverless SQL Warehouse
C. Databricks Runtime for ML
D. Databricks SQL Analytics
Answer: (SHOW ANSWER)

NEW QUESTION: 70
A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.
In which location can the data engineer review their permissions on the table?

A. Jobs
B. Catalog Explorer
C. Dashboards
D. Repos
Answer: (SHOW ANSWER)

NEW QUESTION: 71
A Databricks workflow fails at the last stage due to an error in a notebook. This workflow runs daily. The data engineer fixes the mistake and wants to rerun the pipeline. This workflow is very costly and time- intensive to run.
Which action should the data engineer do in order to minimise downtime and cost?
A. Restart the cluster
B. Repair run
C. Re-run the entire workflow
D. Switch to another cluster
Answer: (SHOW ANSWER)

NEW QUESTION: 72
What is stored in a Databricks customer's cloud account?
A. Data

B. Cluster management metadata
C. Databricks web application
D. Notebooks
Answer: (SHOW ANSWER)
In a Databricks customer's cloud account, the primary elements stored include:
* Data: This is the central type of content stored in the customer's cloud account. Data might include various datasets, tables, and files that are used and managed through Databricks platforms.
* Notebooks: These are also stored within a customer's cloud account. Notebooks include scripts, notes, and other information necessary for data analysis and processing tasks.
Cluster management metadata is indeed managed through the cloud, but it's primarily handled by Databricks rather than stored directly in the customer's account. The Databricks web application itself is not stored within the customer's cloud account; rather, it's a service provided by Databricks.
References:Databricks documentation: Data in Databricks

NEW QUESTION: 73
A data engineer wants to reduce costs and optimize cloud spending. The data engineer has decided to use Databricks Serverless for lowering cloud costs while maintaining existing SLAs.
What is the first step in migrating to Databricks Serverless?
A. Legacy Ingestion pipelines that include ingestion from sources API's, files, JDBC/ODBC connections
B. A frequently running and efficient Scala-based data transformation pipeline compatible with the latest Databricks runtime and Unity Catalog
C. A frequently running and efficient Python-based data transformation pipeline compatible with the latest Databricks runtime and Unity Catalog
D. Low frequency BI Dashboarding and Adhoc SQL Analytics
Answer: (SHOW ANSWER)