GAQM.Databricks-Certified-Data-Engineer-Associate.v2026-11-25.q157

Exam Code:	Databricks-Certified-Data-Engineer-Associate
Exam Name:	Databricks Certified Data Engineer Associate Exam
Certification Provider:	GAQM
Free Question Number:	157
Version:	v2026-11-25
# of views:	102
# of Questions views:	1570
https://www.freecram.net/torrent/GAQM.Databricks-Certified-Data-Engineer-Associate.v2026-11-25.q157.html	

NEW QUESTION: 1
A data engineer has realized that they made a mistake when making a daily update to a table.
They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted.
Which of the following explains why the data files are no longer present?
A. The VACUUM command was run on the table
B. The TIME TRAVEL command was run on the table
C. The DELETE HISTORY command was run on the table
D. The OPTIMIZE command was nun on the table
E. The HISTORY command was run on the table
Answer: (SHOW ANSWER)
The VACUUM command in Delta Lake is used to clean up and remove unnecessary data files that are no longer needed for time travel or query purposes. When you run VACUUM with certain retention settings, it can delete older data files, which might include versions of data that are older than the specified retention period. If the data engineer is unable to restore the table to a version that is 3 days old because the data files have been deleted, it's likely because the VACUUM command was run on the table, removing the older data files as part of data cleanup.

NEW QUESTION: 2
A data engineer is joining a very large events_df (billions of rows) with a small country_lookup_df (fewer than 10,000 rows) on the country_code column. The join is slow due to data shuffling. How should the data engineer improve performance?
A. Cache both DataFrames, then run the same join
B. Persist both DataFrames, then run the same join
C. broadcast events_df when performing the join
D. broadcast country_lookup_df when performing the join
Answer: (SHOW ANSWER)
Broadcasting country_lookup_df sends the small lookup table to each executor, avoiding a large shuffle of the billions-row events_df and improving join performance.

NEW QUESTION: 3
A data engineer is setting up a new Databricks pipeline that ingests clickstream events from Kafka and daily product catalogs from cloud object storage. To ensure auditability and easy reprocessing, the engineer wants to land all source data. Later stages will handle cleaning, deduplication, and business modeling before the data is used in dashboards.
Which approach aligns to Medallion Architecture principles?
A. Land both sources in Gold with denormalized star schemas to optimize BI while retaining full source fidelity.
B. Land streaming events from Kafka in silver, and the product catalog directly in Gold to minimize layers for batch data.
C. Land both sources in the Bronze layer append-only with minimal validation, then build Silver/Gold downstream for quality and analytics.
D. Land both sources directly into the Silver layer with schema enforcement and deduplication to reduce downstream complexity.
Answer: C (LEAVE A REPLY)
The Medallion Architecture recommends landing all raw source data in the Bronze layer in an append-only, minimally transformed form to preserve source fidelity and support auditability and reprocessing, with cleaning, deduplication, and business modeling performed in downstream Silver and Gold layers.

NEW QUESTION: 4
A data engineer is configuring a batch processing job in Databricks to read large Parquet datasets from cloud storage. The engineer wants Spark to automatically distribute computation across multiple worker nodes without manually assigning tasks. Which Spark abstraction enables automatic parallel processing?
A. Driver
B. Cluster
C. DataFrame
D. Executor
Answer: (SHOW ANSWER)

NEW QUESTION: 5
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. Temporary view
B. Database
C. Delta Table
D. View
E. Spark SQL Table
Answer: (SHOW ANSWER)

NEW QUESTION: 6
A data engineer wants to inspect the schema of a Spark DataFrame during development to understand column names and data types. Which DataFrame method prints the schema in a tree format?
A. showSchema
B. getSchema
C. describeSchema

D. printSchema
Answer: (SHOW ANSWER)

NEW QUESTION: 7
A data engineer is building a Silver layer pipeline that should only execute after the Bronze layer table has been successfully updated with new records. Which trigger type in Databricks Jobs is specifically designed to orchestrate this dependency by reacting to changes in the upstream Unity Catalog table?
A. Table update trigger
B. File arrival trigger
C. Scheduled trigger
D. Continuous trigger
Answer: (SHOW ANSWER)
A Table update trigger starts the Silver job when the monitored upstream Unity Catalog table is updated, providing event-driven orchestration without polling or a fixed schedule.

NEW QUESTION: 8
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. Databricks Asset Bundles (DAB) + GitHub Integration
B. Maintain workflow_config.json and deploy it using Databricks CLI
C. Maintain workflow_config.json and deploy it using Terraform
D. Manually create and manage the workflow in UI
Answer: (SHOW ANSWER)
The best practice for deploying ETL pipelines to production is to use Databricks Asset Bundles (DAB) with GitHub integration. DAB provides a declarative YAML-based structure for defining workflows and resources, supports CI/CD, and integrates seamlessly with GitHub, enabling version-controlled, automated deployments to production.

NEW QUESTION: 9
A data engineer investigates a slow stage in the Databricks Spark UI. The engineer scrolls down to the Summary Metrics section for task durations, and they observe for that stage:
25th percentile duration: 40 seconds
Median duration: 45 seconds
75th percentile duration: 50 seconds
Max duration: 80 seconds
What is the interpretation of these metrics?
A. The stage is healthy.
B. There is a GC issue.
C. The cluster is under-provisioned.
D. The stage likely suffers from data skew.
Answer: (SHOW ANSWER)

NEW QUESTION: 10

A data engineer wants to delegate day-to-day permission management for the schema main marketing to the mkt-admins group, without making them workspace admins. They should be able to grant and revoke privileges for other users on objects within that schema.
Which approach aligns with UC's ownership and privilege model?

Freecram.net

A. Grant MANAGE permissions on the metastore to mkt-admins, which allows managing privileges for all schemas and tables globally.
B. Make mkt-admins a workspace-level admins group, then SELECT on main.marketing to allow privilege delegation.
C. Transfer ownership of the schema main.marketing to mkt-admins; owners can manage privileges on the schema and its contained objects.
D. Grant USE SCHEMA on main.marketing and MODIFY on all tables to mkt-admins, which enables managing grants within the schema.
Answer: (SHOW ANSWER)
In Unity Catalog, object owners have full control over privilege management for that object and its children. Transferring ownership of the schema to the group allows it to grant and revoke privileges within the schema without requiring workspace-level administrative rights.

NEW QUESTION: 11

A Databricks Job with four sequential tasks is executed. The job failed at Task 3. After the root cause is fixed, the data engineer needs to complete the workflow without re-running the successful Tasks 1 and 2. Which action in the Job Run details page allows the engineer to resume execution from the point of failure?
A. Repair run
B. Reset job
C. Clone task
D. Run now
Answer: (SHOW ANSWER)
Repair run reruns the failed Task 3 and any dependent downstream tasks while preserving the successful results of Tasks 1 and 2.

NEW QUESTION: 12

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

NEW QUESTION: 13

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.
```
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()
    )
```

B.
```
import dlt

dlt.table(
    name = "orders_raw",

def orders_raw():
    return (
        spark.readStream
        .format("cloud_files")
        .option("cloudFiles.format", "json")
        .option("cloudFiles.schemaLocation", "/schema/location"
        .load("kafka://broker:9092/orders_topic")
    )
```

C.
```
CREATE LIVE TABLE orders_raw
AS SELECT
    CAST(value AS STRING) AS json_data
FROM STREAM kafka.`broker:9092/orders_topic`;
```

GAQM Global Association for Quality Management
freecram.net

```
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");
```

GAQM Global Association for Quality Management
freecram.net

D.

Answer: (SHOW ANSWER)

The correct way to ingest Kafka data into a Delta Live Tables pipeline is with SQL using CREATE LIVE TABLE ... AS SELECT ... FROM STREAM kafka.... This properly connects to the Kafka broker, subscribes to the topic, and casts the Kafka message value field for downstream processing in DLT.

NEW QUESTION: 14

Which of the following must be specified when creating a new Delta Live Tables pipeline?

A. The preferred DBU/hour cost

B. A location of a target database for the written data

C. At least one notebook library to be executed

D. A key-value pair configuration

E. A path to cloud storage location for the written data

Answer: (SHOW ANSWER)

NEW QUESTION: 15

A data engineer is configuring a cluster in Databricks and wants to ensure that multiple Spark executors can run tasks concurrently across worker nodes. Which cluster component is responsible for executing tasks?

A. Executor

B. Job scheduler

C. Driver

D. Notebook

Answer: (SHOW ANSWER)

NEW QUESTION: 16

A data engineer has written a function in a Databricks Notebook to calculate the population of bacteria in a given medium.

```
def calculate_population_of_bacteria (intital_population, exponential_factor)
    return future_population = intitial_population ** expontential_factor
```

Analysts use this function in the notebook and sometimes provide input arguments of the wrong data type, which can cause errors during execution.

Which Databricks feature will help the data engineer quickly identify if an incorrect data type has been provided as input?

A. The Spark User interface has a debug tab that contains the variables that are used in this session.

B. The Databricks debugger enables breakpoints that will raise an error if the wrong data type is submitted.
C. The Databricks debugger enables the use of a variable explorer to see at a glance the value of the variables.
D. The Data Engineer should add print statements to find out what the variable is.
Answer: (SHOW ANSWER)
The Databricks debugger supports setting breakpoints that pause execution and allow inspection of variables. If the wrong data type is passed to the function, the debugger raises an error at runtime, helping the engineer quickly identify the issue.

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 17
An organization has implemented a data pipeline in Databricks and needs to ensure it can scale automatically based on varying workloads without manual cluster management. The goal is to meet the company's Service Level Agreements (SLAs), which require high availability and minimal downtime, while Databricks automatically handles resource allocation and optimization.
Which approach fulfills these requirements?
A. Deploy Job Clusters with fixed configurations, dedicated to specific tasks, without automatic scaling.
B. Use Spot Instances to allocate resources dynamically while minimizing costs, with potential interruptions.
C. Use Interactive Clusters in Databricks, adjusting cluster sizes manually based on workload demands.
D. Use Serverless compute in Databricks to automatically scale and provision resources with minimal manual intervention.
Answer: D (LEAVE A REPLY)
Serverless compute in Databricks automatically provisions and scales resources to meet workload demands, ensuring high availability and minimal downtime while reducing the need for manual cluster management, which aligns with SLA requirements.

NEW QUESTION: 18
A developer is building a data pipeline that processes records from a bronze table into a silver table. The bronze table, bronze_events, contains duplicate records due to at-least-once delivery guarantees from the upstream ingestion system. The developer writes the following PySpark code to prepare the data:

```
deduped_df = df.dropDuplicates()
deduped_df.summary("count", "mean", "stddev").show()
from pyspark.sql.functions import approx_count_distinct
deduped_df.select(approx_count_distinct("user_id")).show()
```

The developer notices that after running dropDuplicates() without arguments, some rows that differ only in the event_timestamp column remain. The developer wants to de-duplicate based only on user_id and event_type, keeping one row per unique combination of these two columns.

Which code change achieves this deduplication requirement?
A. Replace df.dropDuplicates() with df.groupBy("user_id", "event_type").count()
B. Replace df.dropDuplicates() with df.distinct().select("user_id", "event_type")
C. Replace df.dropDuplicates() with df.dropDuplicates(["user_id", "event_type"])
D. Replace df.dropDuplicates() with df.drop("event_timestamp").dropDuplicates()
Answer: (SHOW ANSWER)
dropDuplicates(["user_id", "event_type"]) removes rows based only on those two columns while retaining one complete row for each unique combination.

NEW QUESTION: 19
A data engineer is using Spark SQL to analyze a large dataset stored in Delta format. The engineer notices that queries filtering on partition columns run significantly faster. What is the primary reason for this performance improvement?
A. Broadcast joins
B. Partition pruning
C. Data replication
D. Data caching
Answer: B (LEAVE A REPLY)

NEW QUESTION: 20
Which of the following commands can be used to write data into a Delta table while avoiding the writing of duplicate records?
Freecram.net
A. MERGE
B. APPEND
C. DROP
D. IGNORE
E. INSERT
Answer: (SHOW ANSWER)

NEW QUESTION: 21
A data engineer is designing a medallion architecture. Raw Kafka events must be stored exactly as received for reprocessing, then cleaned and deduplicated, then aggregated for BI dashboards.
Which mapping of layers is correct?
A. Bronze = aggregated metrics, Silver = raw ingest, Gold = cleaned records
B. Bronze = raw ingest, Silver = cleaned and conformed records, Gold = business-level aggregates
C. Bronze = cleaned records, Silver = business aggregates, Gold = raw ingest
D. Bronze = external tables only, Silver = views only, Gold = temporary views only
Answer: (SHOW ANSWER)
In the medallion architecture, bronze holds raw, append-only ingested data to enable full reprocessing; silver holds validated, deduplicated, and joined data; gold holds curated, aggregated datasets optimized for analytics and reporting.

NEW QUESTION: 22

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. Change Data Capture

B. Error Handling

C. Monitoring

D. Expectations

Answer: D (LEAVE A REPLY)

Delta Live Tables Expectations allow defining data quality rules. Records that fail the expectations (such as sensor readings outside valid ranges) can be dropped without stopping the streaming pipeline, ensuring continuous and accurate processing.

NEW QUESTION: 23

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 previous step in the DLT pipeline is static.

B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.

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 the subsequent step in the DLT pipeline is static.

E. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.

Answer: (SHOW ANSWER)

NEW QUESTION: 24

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

A. 25MBs

B. 10MBs

C. 30MBs

D. 15MBs

Answer: (SHOW ANSWER)

The maximum output supported by a job cluster in Databricks is 10MB. If the output exceeds this limit, the notebook may fail.

NEW QUESTION: 25

A data engineer wants to combine two datasets that share the same schema and append rows from one dataset to another. Which Spark DataFrame operation performs this row-wise combination?

A. union

B. intersect

C. merge

D. join

Answer: (SHOW ANSWER)

NEW QUESTION: 26

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 USAGE ON CATALOG team TO customers;
B. GRANT VIEW ON CATALOG customers TO team;
C. GRANT USAGE ON DATABASE customers TO team;
D. GRANT CREATE ON DATABASE team TO customers;
E. GRANT CREATE ON DATABASE customers TO team;
Answer: C (LEAVE A REPLY)

NEW QUESTION: 27
A data engineer needs to persist a DataFrame to disk instead of memory because the dataset is too large to fit in memory. Which Spark persistence level supports disk storage?
A. MEMORY_ONLY
B. MEMORY_SERIALIZED
C. MEMORY_ONLY_SER
D. MEMORY_AND_DISK
Answer: (SHOW ANSWER)

NEW QUESTION: 28
A data engineer must deliver a trustworthy customer 360 dataset in Databricks for data scientists and BI teams. The engineer plans to join deduplicated customer records with cleaned transaction data, enforce schema and data quality checks, and create a conformed "customer_transactions"view. Later, highly aggregated, domain-specific tables (for weekly spend and executive summaries) will be produced for dashboards.
Where should the engineer build the conformed "customer_transactions" dataset, and where should the aggregated, report-ready tables reside?
A. Build both "customer_transactions" and aggregated, report-ready tables in Silver to keep the model simpler.
B. Build "customer_transactions" in Bronze and put the aggregated, report-ready tables in Silver.
C. Build "customer_transactions" in Gold and put the aggregated, report-ready tables in Silver.
D. Build "customer_transactions" in Silver and put the aggregated, report-ready tables in Gold.
Answer: (SHOW ANSWER)
The conformed customer_transactions dataset belongs in Silver because it represents cleaned, deduplicated, and validated data that is suitable for broad analytical use. Aggregated, domain- specific, and report-ready tables are best placed in Gold, as this layer is designed for highly curated datasets optimized for BI, dashboards, and executive reporting.

NEW QUESTION: 29
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 fail before the new task begins
B. When another task needs to use as little compute resources as possible
C. When another task has the same dependency libraries as the new task
D. When another task needs to be replaced by the new task

E. When another task needs to successfully complete before the new task begins
Answer: (SHOW ANSWER)

NEW QUESTION: 30
A data engineer wants to run SQL queries directly on data stored in Delta Lake using a serverless compute environment designed for analytics workloads. Which Databricks component provides this capability?
Freecram.net
A. Databricks SQL Warehouse
B. DBFS
C. MLflow
D. Spark Driver
Answer: (SHOW ANSWER)

NEW QUESTION: 31
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. Set up Azure Blob Storage permissions at the container level, allowing access to all external tables.
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. Use Unity Catalog to manage access controls and permissions for each external table individually.
Answer: (SHOW ANSWER)
The correct approach is to use Unity Catalog to manage fine-grained access controls and permissions for each external table individually. Unity Catalog centralizes governance and ensures users only have access to the specific external tables they need, instead of granting broad storage or workspace-level permissions.

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 32
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 without notifications.
B. They can set up an Alert with a new email alert destination.
C. They can set up an Alert with a custom template.
D. They can set up an Alert with one-time notifications.
E. They can set up an Alert with a new webhook alert destination.
Answer: (SHOW ANSWER)

NEW QUESTION: 33
A data engineer wants to read only a subset of columns from a Delta table to reduce memory usage and improve performance. Which Spark DataFrame operation should be used?
A. groupBy
B. select
C. orderBy
D. filter
Answer: B (LEAVE A REPLY)

NEW QUESTION: 34
A data engineer is troubleshooting two different pipeline failures:
- Pipeline A fails with a java.lang.OutOfMemoryError is immediately thrown after the command display(df.collect()) is called on a 100GB dataset.
- Pipeline B fails during a wide transformation (a join of two large tables) with an ExecutorLostFailure error message, indicating executor memory exhaustion during shuffle.
Which action should the data engineer take to fix these two issues?
A. For Pipeline A, switch to a Storage Optimized cluster; for Pipeline B, use a Compute Optimized cluster.
B. For Pipeline A, increase the spark.driver.memory; for Pipeline B, increase the spark.sql.shuffle.partitions or executor memory.
C. For Pipeline A, increase the number of executors; for Pipeline B, increase the driver memory.
D. For both pipelines, enable Adaptive Query Execution (AQE) as it automatically manages memory allocation for both the driver and executors.
Answer: (SHOW ANSWER)
collect() moves the entire 100GB dataset to the driver, causing driver memory exhaustion, so Pipeline A requires avoiding collect() or increasing driver memory. Pipeline B fails on executors during a shuffle, so increasing shuffle partitions to reduce partition size or adding executor memory addresses the issue.

NEW QUESTION: 35
A Python file is ready to go into production and the client wants to use the cheapest but most efficient type of cluster possible. The workload is quite small, only processing 10GBs of data with only simple joins and no complex aggregations or wide transformations. Which cluster meets the requirement?
A. Interactive cluster
B. Job cluster with spot instances enabled

C. Job cluster with spot instances disabled
D. Job cluster with Photon enabled
Answer: (SHOW ANSWER)
A job cluster with spot instances enabled is the cheapest and efficient option for small, production workloads. Spot instances reduce costs, and a job cluster ensures the cluster only runs during the workload execution.

NEW QUESTION: 36
A data engineer wants to optimize query performance for a Delta table storing billions of records.
Queries frequently filter on the column transaction_date. Which Delta Lake feature should be applied to physically organize the data and reduce the amount of data scanned during query execution?
A. Z-Ordering
B. VACUUM
C. Broadcast join
D. Caching
Answer: A (LEAVE A REPLY)

NEW QUESTION: 37
A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.
Which of the following approaches can the data engineering team use to improve the latency of the team's queries?
A. They can increase the cluster size of the SQL endpoint.
B. They can increase the maximum bound of the SQL endpoint's scaling range.
C. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."
D. They can turn on the Serverless feature for the SQL endpoint.
E. They can turn on the Auto Stop feature for the SQL endpoint.
Answer: (SHOW ANSWER)

NEW QUESTION: 38
A data engineer configures a Databricks Lakeflow Job for daily customer data processing:
- Entry task: Single notebook loads raw data
- Parallel tasks:
- SQL query task for cleansing
- Notebook task for feature engineering
- Pipeline task for model updates
- Exit task: Dashboard refresh (must run after all parallel tasks
complete)
- Requirement: Implement this dependency pattern using a DAG-based task graph Which task configuration ensures all parallel tasks complete before the dashboard refresh task runs?
A. Create separate Lakeflow Jobs for each parallel task and trigger them sequentially using external orchestration.
B. Add the dashboard task as dependent on all three parallel tasks using fan-in control flow.

C. Configure the parallel tasks with sequential dependencies so each task waits for the previous one to finish.
D. Set the dashboard task to depend only on the SQL query task; other tasks run independently.
Answer: (SHOW ANSWER)
A fan-in dependency makes the dashboard refresh task depend on all three parallel tasks.
Databricks runs the parallel tasks after the entry task, and the dashboard refresh starts only after every upstream dependency completes.

NEW QUESTION: 39
A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.
Which of the following data entities should the data engineer create?
A. View
B. Temporary view
C. Table
D. Function
E. Database
Answer: (SHOW ANSWER)

NEW QUESTION: 40
A data engineer maintains a Declarative Automation Bundle (formerly Databricks Asset Bundle), which deploys a dashboard.
The `databricks.yml` includes:

```
bundle:
  name: workspace_assets

variables:
  dataset_catalog:
    default: catalog_dev

targets:
  dev:
    variables:
      dataset_catalog: catalog_dev
  prod:
    variables:
      dataset_catalog: catalog_prod
```

GAQM
Global Association for Quality Management
freecram.net

Which set of actions will enable the dashboard to be deployed in prod using catalog_prod without changing the resource definition?
A. Run databricks bundle deploy --profile prod so the CLI selects targets.prod and applies dataset_catalog: catalog_prod automatically.
B. Run databricks bundle deploy --target prod so the deployment uses targets.prod and overrides dataset_catalog to catalog_prod.
C. Run databricks bundle execute --profile prod so the CLI selects targets.prod and applies dataset_catalog: catalog_prod automatically.

D. Run databricks bundle deploy --var dataset_catalog=catalog_prod so the CLI selects targets.prod and applies dataset_catalog: catalog_prod automatically.
Answer: (SHOW ANSWER)
databricks bundle deploy --target prod selects the prod target, which overrides dataset_catalog with catalog_prod. A CLI profile controls authentication, not bundle target selection.

NEW QUESTION: 41
A data engineer uses the Databricks workspace UI with Unity Catalog enabled. In Data Explorer, they select catalog corp_marketing, then schema campaigns, and see table email_stats. The engineer must let the growth-analysts group read email_stats from their SQL warehouses, but not create, alter, or delete any objects in corp_marketing or campaigns. Which action sequence meets the requirement?
A. 1. On schema campaigns, open Permissions and grant USE SCHEMA to growth-analysts
2. On the table email_stats, open Permissions and grant SELECT and MODIFY to growth- analysts
B. 1. On catalog corp_marketing, open Permissions and grant USE CATALOG to growth-analysts
2. On schema campaigns, open Permissions and grant USE SCHEMA to growth-analysts
3. On the table email_stats, open Permissions and grant SELECT to growth-analysts
C. 1. On catalog corp_marketing, open Permissions and grant USE CATALOG and CREATE SCHEMA to growth-analysts
2. On schema campaigns, open Permissions and grant USE SCHEMA to growth-analysts
3. On the table email_stats, open Permissions and grant SELECT to growth-analysts
D. 1. On schema campaigns, open Permissions and grant USE SCHEMA and SELECT to growth- analysts
2. On the table email_stats, open Permissions and grant SELECT to growth-analysts
Answer: (SHOW ANSWER)
Reading a Unity Catalog table requires USE CATALOG on its catalog, USE SCHEMA on its schema, and SELECT on the table. These privileges allow read access without granting object creation or modification rights.

NEW QUESTION: 42
A data engineer notices that a Spark job repeatedly scans a large Delta table even though the dataset does not change during the session. Which technique can store the dataset in memory to speed up repeated queries?
A. MERGE
B. OPTIMIZE
C. CACHE TABLE
D. VACUUM
Answer: (SHOW ANSWER)

NEW QUESTION: 43
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

Calculate the total sales amount for each product category and store the results in a new dataframe called category_sales.
What will generate the expected result of category_sales?

category	total_sales_amount
Electronics	500
Clothing	900

A. category_sales =
sales_df.groupBy("category").agg(sum("sales_amount").alias("total_sales_amount"))
B. category_sales =
sales_df.sum("sales_amount").groupBy("category").alias("total_sales_amount"))
C. category_sales =
sales_df.agg(sum("sales_amount").groupBy("category").alias("total_sales_amount"))
D. category_sales =
sales_df.groupBy("region").agg(sum("sales_amount").alias("total_sales_amount"))
Answer: (SHOW ANSWER)
The correct approach is to group by the product category and then aggregate the sum of sales_amount, aliasing it as total_sales_amount. This produces the expected totals: 500 for Electronics and 900 for Clothing.

NEW QUESTION: 44
A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that feeds the dashboard is automatically processed using a Databricks Job.
Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?
A. They can turn on the Auto Stop feature for the SQL endpoint.
B. They can set up the dashboard's SQL endpoint to be serverless.
C. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.
D. They can reduce the cluster size of the SQL endpoint.
E. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.

Answer: (SHOW ANSWER)

NEW QUESTION: 45
A data engineer is developing a small proof of concept in a notebook. When running the entire notebook, the Cluster usage spikes. The data engineer wants to keep the development requirements and get real-time results. Which Cluster meets these requirements?
A. All Purpose Cluster with autoscaling
B. Job Cluster with Photon enabled and autoscaling
C. Job Cluster with autoscaling enabled
D. All-Purpose Cluster with a large fixed memory size
Answer: A (LEAVE A REPLY)
An All-Purpose Cluster with autoscaling is best for interactive development and proof of concept work in notebooks, since it provides real-time results and can dynamically scale resources as usage spikes.

NEW QUESTION: 46
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. Re-run the entire workflow
B. Repair run
C. Restart the cluster
D. Switch to another cluster
Answer: (SHOW ANSWER)
Repair run in Databricks allows rerunning only the failed tasks in a workflow instead of re-running the entire pipeline, minimizing downtime and reducing costs.

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 47
Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?
A. Manually programming in an alert system in each cell of the Notebook
B. There is no way to notify the Job owner in the case of Job failure
C. Setting up an Alert in the Notebook
D. Setting up an Alert in the Job page
E. MLflow Model Registry Webhooks

Answer: (SHOW ANSWER)

NEW QUESTION: 48
What is the structure of an Asset Bundle?
A. A Docker image containing runtime environments and the source code of the assets
B. A compressed archive (ZIP) that solely contains workspace assets without any accompanying metadata
C. A single plain text file enumerating the names of assets to be migrated to a new workspace
D. A YAML configuration file that specifies the artifacts, resources, and configurations for the project
Answer: (SHOW ANSWER)
An Asset Bundle in Databricks is structured as a YAML configuration file that defines the project's artifacts, resources, and configurations. This format allows reproducibility, automation, and portability of projects across workspaces, making it the standard approach for production pipelines.

NEW QUESTION: 49
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 of the following commands could the data engineering team use to access sales in PySpark?
A. There is no way to share data between PySpark and SQL.
B. spark.delta.table("sales")
C. SELECT * FROM sales
D. spark.table("sales")
E. spark.sql("sales")
Answer: D (LEAVE A REPLY)

NEW QUESTION: 50
Which two conditions are applicable for governance in Databricks Unity Catalog? (Choose two.)
A. You can have more than 1 metastore within a databricks account console but only 1 per region.
B. Both catalog and schema must have a managed location in Unity Catalog provided metastore is not associated with a location
C. You can have multiple catalogs within metastore and 1 catalog can be associated with multiple metastore
D. If catalog is not associated with location, it's mandatory to associate schema with managed locations
E. If metastore is not associated with location, it's mandatory to associate catalog with managed locations
Answer: (SHOW ANSWER)
You can have more than 1 metastore within a Databricks account console but only 1 per region:
This is correct. In Databricks Unity Catalog, you can create multiple metastores, but you can only have one metastore per region within an account.
If metastore is not associated with location, it's mandatory to associate catalog with managed locations: This is also correct. If a metastore does not have a default storage location, then the catalog must be associated with a managed location to ensure that data is stored properly within Unity Catalog.

NEW QUESTION: 51

A data engineer has been provided a PySpark DataFrame named dfwith columns product and revenue. The data engineer needs to compute complex aggregations to determine each product's total revenue, average revenue, and transaction count. Which code snippet should the data engineer use?

A.

```
from pyspark.sql import functions as F
aggregated_df = df.groupBy("product").agg( F.sum("revenue").alias("total_revenue"),
F.avg("revenue").alias("avg_revenue"), F.count("*").alias("transaction_count") )
```

B.

```
aggregated_df = df.groupby("product").agg(
    "sum(revenue)",
    "avg(revenue)",
    "count(revenue)"
)
```

C.

```
from pyspark.sql import functions as F aggregated_df = df.select("product",

"revenue").groupBy("product").agg( F.sum("revenue"), F.mean("revenue") )
```

D.

```
aggregated_df =
df.groupBy("product").agg( {"revenue": "sum", "revenue": "avg", "revenue": "count"} )
```

Answer: A (LEAVE A REPLY)

The correct code uses PySpark functions (sum, avg, count) with aliases inside .agg() after grouping by product. This produces total revenue, average revenue, and transaction count for each product.

NEW QUESTION: 52

A data engineer runs CREATE TABLE gold.revenue AS SELECT * FROM silver.orders in Unity Catalog without specifying a LOCATION.
What is the result?

A. An external table is created and files are written to the workspace root DBFS location.

B. A managed Delta table is created in the metastore/schema's default storage location.

C. A temporary view is created that is dropped at the end of the session.

D. The command fails because a LOCATION clause is mandatory in Unity Catalog.

Answer: (SHOW ANSWER)

A CTAS statement without a LOCATION clause creates a Unity Catalog managed table in Delta format, stored in the managed storage location configured for the schema, catalog, or metastore.
Unity Catalog then controls the storage lifecycle, including file cleanup on drop.

NEW QUESTION: 53

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 create a new job from scratch and add both tasks to run concurrently.

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 clone the existing task in the existing Job and update it to run the new notebook.
E. They can clone the existing task to a new Job and then edit it to run the new notebook.
Answer: (SHOW ANSWER)

NEW QUESTION: 54
A Delta Live Table pipeline includes two datasets defined using STREAMING LIVE TABLE.
Three datasets are defined against Delta Lake table sources using LIVE TABLE.
The table is configured to run in Production mode using the Continuous Pipeline Mode.
Assuming previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?
A. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.
B. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.
C. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped.
D. All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.
E. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.
Answer: (SHOW ANSWER)

NEW QUESTION: 55
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 source is not a Delta table
C. When the source table can be deleted
D. When the target table cannot contain duplicate records
E. When the target table is an external table
Answer: (SHOW ANSWER)

NEW QUESTION: 56
A data engineering team runs nightly batch ETL pipelines that transform large volumes of data stored in Delta Lake. The workloads are predictable and run on a fixed schedule. The team wants automatic scaling and cost efficiency, paying only for compute used during job execution. Which Databricks compute option is appropriate for this workload?
A. An all-purpose compute cluster configured with autoscaling.
B. A Databricks SQL warehouse configured for serverless execution.
C. A single-node compute cluster configured without autoscaling.
D. A job compute cluster triggered by a scheduled workflow.
Answer: (SHOW ANSWER)
Job compute is designed for scheduled, automated workloads. It can autoscale for the ETL workload and terminates after the job finishes, avoiding the idle cost of persistent all-purpose compute.

NEW QUESTION: 57
A data engineer is configuring Delta Sharing for a multi-team project where teams from different departments will need to access shared data. The data engineer has successfully created a Unity Catalog metastore and is now setting up the Delta Share. The goal is to ensure that internal teams can access the shared data with full permissions, while external partners can only read the data. Which action should the Data Engineer take to configure the sharing correctly?
A. Create a Delta Share, set up a secure access URL for internal teams and external partners, and distribute the URL to provide them access to the shared data.
B. Create a Delta Share, add the internal team's tables and views, and assign READ/WRITE permissions to both external partners and internal teams.
C. Assign READ permissions to external partners through the Delta Share and READ/WRITE permissions to internal teams, and ensure the correct tables and views are shared.
D. Grant READ permissions to external partners and READ/WRITE permissions to internal teams through the Delta Share.
Answer: (SHOW ANSWER)
In Delta Sharing, permissions must be managed separately for each recipient. To achieve the goal, the data engineer should create the Delta Share, add the appropriate tables/views, assign READ permissions to external partners, and assign READ/WRITE permissions to internal teams.
This ensures external users only consume data, while internal teams can both read and modify it.

NEW QUESTION: 58
A data engineer is attempting to grant a user access to a view. They successfully run GRANT SELECT ON VIEW sales_catalog.market_data.summary TO analyst_group.
However, the analyst still receives an "Insufficient Permissions" error when trying to query the view.
In which order should the engineer verify the Hierarchy of Securable Objects to ensure the analyst has the necessary usage permissions?
A. CATALOG -> SCHEMA -> VIEW
B. METASTORE -> TABLE -> VIEW
C. STORAGE CREDENTIAL -> EXTERNAL LOCATION -> VIEW
D. WORKSPACE -> CLUSTER -> VIEW
Answer: (SHOW ANSWER)
To query the view, the analyst needs USE CATALOG on sales_catalog, then USE SCHEMA on market_data, and SELECT on summary. Privileges on the underlying tables are not required for querying a Unity Catalog view.

NEW QUESTION: 59
A data engineer is working with a very large dataset and wants to improve query performance by colocating related data in the same files. The engineer decides to apply Z-ordering on frequently filtered columns. Which Delta Lake command is used?
A. ANALYZE TABLE
B. VACUUM ZORDER
C. CACHE TABLE
D. OPTIMIZE ZORDER BY
Answer: (SHOW ANSWER)

NEW QUESTION: 60

A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.
Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

Freecram.net

A. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.
B. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
C. They can turn on the Auto Stop feature for the SQL endpoint.
D. They can reduce the cluster size of the SQL endpoint.
E. They can set up the dashboard's SQL endpoint to be serverless.
Answer: (SHOW ANSWER)

NEW QUESTION: 61
A data engineer is standardizing repository layouts for multiple teams adopting Databricks Asset Bundles. The engineer wants to ensure every project has a single authoritative configuration file at the repository root that defines the bundle name, targets, workspace settings, permissions, and resource mappings (for jobs and pipelines).
What strategy should the data engineer use to meet the goal?
A. Place multiple databricks.ymlfiles under each subfolder (for example, jobs/pipelines/, workspace/) and merge them at deploy time using the include mapping.
B. Place exactly one databricks.yml at the repository root; it is the main configuration file and may reference additional configuration files via the include mapping.
C. Place a databricks.ymlin a databricks/hidden folder at the repository root; only hidden locations are valid for bundle configs.
D. Place a databricks.yaml at the repository root and optional databricks.yml in subfolders; the CLI prefers .yamlover .ymlwhen both exist.
Answer: (SHOW ANSWER)
Databricks Asset Bundles require a single authoritative databricks.yml file at the repository root that defines core bundle configuration, while allowing modularization by referencing additional files through the include mapping.

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 62
A data engineer runs df.toPandas() on a wide DataFrame containing 50 million rows. The notebook cell fails with a java.lang.OutOfMemoryError on the driver. Which memory configuration is directly responsible for this failure?
A. spark.memory.fraction is too low, limiting the storage memory available for caching
B. spark.executor.memory is too low to process the DataFrame transformations
C. spark.sql.shuffle.partitions is set too high, causing excessive memory consumption during shuffles

D. spark.driver.memory is too low to hold the collected result set
Answer: D (LEAVE A REPLY)
df.toPandas() collects the entire DataFrame into the driver process. A 50-million-row wide dataset exceeds available spark.driver.memory, causing the driver-side OutOfMemoryError.

NEW QUESTION: 63
A data engineer needs to append new records to an existing Delta table while preserving existing data. The ingestion pipeline runs every hour and adds incremental data without replacing previous records. Which write mode should be used?
A. append
B. overwrite
C. errorIfExists
D. ignore
Answer: A (LEAVE A REPLY)

NEW QUESTION: 64
A data engineer wants to analyze historical versions of a Delta table to debug a data quality issue. The engineer needs to query the table as it existed two days ago. Which Delta Lake feature allows querying older snapshots of a table?
A. VACUUM
B. Z-Ordering
C. OPTIMIZE
D. Time Travel
Answer: D (LEAVE A REPLY)

NEW QUESTION: 65
A data engineer must prevent rows with a null order_id from ever being written to a Delta table, causing the write to fail if violated.
Which command enforces this?
A. ALTER TABLE orders ADD CONSTRAINT id_nn CHECK (order_id IS NOT NULL)
B. ALTER TABLE orders ALTER COLUMN order_id DROP NOT NULL
C. ALTER TABLE orders SET TBLPROPERTIES ('delta.enforceNotNull' = 'true')
D. CREATE TABLE orders (order_id INT COMMENT 'required')
Answer: A (LEAVE A REPLY)
Delta Lake supports CHECK constraints that are validated on every write; any transaction inserting a violating row fails with a constraint violation error. Existing data is validated when the constraint is added, so the command fails if current rows already violate it.

NEW QUESTION: 66
A data engineer needs Task C to run ONLY if Task A succeeds AND Task B fails. Which dependency configuration implements this conditional logic?
A. Task C depends on Task A (All done) and Task B (All done) with a condition widget.
B. Task C depends on Task A (Success) with Task B listed as a negative dependency.
C. Task C uses two dependency chains: Task A with All succeeded, Task B with at least one failed.
D. Create a Python task that checks dbutils.jobs.taskValues from Tasks A and B before running Task C.

Answer: (SHOW ANSWER)

Use separate gate tasks: one runs only when Task A succeeds, and another runs only when Task B fails. Task C then depends on both gates with All succeeded, enforcing the exact A succeeded AND B failed condition. A single Run if setting applies collectively to all dependencies and cannot express mixed per-task outcomes.

NEW QUESTION: 67

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, years_exp > 5) AS exp_employees
FROM stores;
```

B.

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

C.

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

D.

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

E.

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

Answer: (SHOW ANSWER)

NEW QUESTION: 68

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. Commit

B. Clone

C. Merge

D. Pull

E. Push

Answer: (SHOW ANSWER)

NEW QUESTION: 69

What Databricks feature can be used to check the data sources and tables used in a workspace?

A. Use the lineage feature to visualize a graph that highlights where the table is used only in reports.

B. Use the lineage feature to visualize a graph that highlights where the table is used only in notebooks.

C. Use the lineage feature to visualize a graph that shows all dependencies, including where the table is used in notebooks, other tables, and reports.

D. Do not use the lineage feature as it only tracks activity from the last 3 months and will not provide full details on dependencies.

Answer: (SHOW ANSWER)

The lineage feature in Unity Catalog provides a complete visualization of data dependencies, showing how a table is used across notebooks, reports, and other tables. This helps track data usage, dependencies, and impact analysis for governance and auditing.

NEW QUESTION: 70

A data engineer wants to delete records from a Delta table where the status column equals "inactive". The operation should modify only the affected rows without rewriting the entire dataset.

Which SQL command should be used?

A. INSERT

B. UPDATE

C. DELETE

D. MERGE

Answer: (SHOW ANSWER)

NEW QUESTION: 71
A data engineer is implementing a data pipeline using Databricks notebooks. The engineer wants to share parameters such as file paths and processing dates across multiple tasks within a job.
Which Databricks utility enables parameter passing?
A. dbutils.widgets
B. dbutils.fs
C. spark.conf
D. display
Answer: (SHOW ANSWER)

NEW QUESTION: 72
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. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.
D. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
E. There is no way to determine why a Job task is running slowly.
Answer: (SHOW ANSWER)

NEW QUESTION: 73
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 of the following locations can the data engineer review their permissions on the table?
A. Databricks Filesystem
B. Jobs
C. Data Explorer
D. Repos
E. Dashboards
Answer: C (LEAVE A REPLY)

NEW QUESTION: 74
An organization is building a data lakehouse and needs to ingest data from multiple sources into Unity Catalog-managed tables:
Source 1: Salesforce (50+ objects, frequent schema changes, OAuth
authentication required)
Source 2: On-premises SQL server database (100+ tables, CDC enabled,
requires network connectivity)
Source 3: Daily JSON files landing in Azure Data Lake Storage Gen2
The organization wants all ingested data to be governed by Unity Catalog, to minimize data engineering effort for schema changes, and to use serverless wherever possible to reduce operational overhead.

Which ingestion strategy meets these requirements?
A. Use partner connectors (Fivetran) for all three sources via Partner Connect, as they provide the most comprehensive integration with Unity Catalog and automated schema evolution.
B. Use Lakeflow Connect managed connectors for all three sources, as they provide consistent serverless execution and Unity Catalog integration across all data source types.
C. Use Lakeflow Connect managed connectors for Salesforce and SQL Server, and Auto Loader for ADLS files, all within Lakeflow Spark Declarative Pipelines.
D. Use Auto Loader for all three sources with custom Python code to handle Salesforce API calls and SQL server CDC extraction, then write to Unity Catalog tables.
Answer: (SHOW ANSWER)
Lakeflow Connect provides managed, serverless connectors for Salesforce and SQL Server CDC, including Unity Catalog governance and automated schema evolution. Auto Loader is the appropriate managed mechanism for incrementally ingesting JSON files from ADLS Gen2 and can run within Lakeflow Spark Declarative Pipelines.

NEW QUESTION: 75
A developer has configured Databricks Connect and starts running PySpark code locally. They notice that Data Frame operations are executed on the remote cluster, but actions like .show() return results directly in their local console. What does this behaviour illustrate about how Databricks Connect is used?
A. It requires manual synchronization of results from the cluster
B. It runs all Spark compute locally and only reads remote files
C. It proxies Spark commands from the local client to the remote cluster
D. It mirrors the entire cluster environment on the developer's machine
Answer: (SHOW ANSWER)
Databricks Connect forwards DataFrame transformations and actions from the local environment to the remote cluster for execution, while returning results such as .show() back to the local client, demonstrating a proxy-based interaction model.

NEW QUESTION: 76
A data provider must share a live Delta table with an external partner who uses an open-source Spark environment outside Databricks, without copying the data.
Which feature should be used?
A. Delta Sharing with a recipient and share
B. A Unity Catalog external location with a storage credential granted to the partner
C. A deep clone written to a partner-owned S3 bucket nightly
D. A materialized view exported as Parquet and delivered by SFTP
Answer: (SHOW ANSWER)
Delta Sharing is an open protocol that lets a provider share live data from Unity Catalog with recipients on any platform that supports the protocol, with no data replication. Access is governed by shares and recipients, and the recipient always reads the current version of the data.

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 77
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 with SQL within Databricks SQL
B. When they are running automated reports to be refreshed as quickly as possible
C. When they are working interactively with a small amount of data
D. When they are manually running reports with a large amount of data
E. When they are concerned about the ability to automatically scale with larger data
Answer: (SHOW ANSWER)

NEW QUESTION: 78
Which of the following tools is used by Auto Loader process data incrementally?
A. Data Explorer
B. Checkpointing
C. Spark Structured Streaming
D. Unity Catalog
E. Databricks SQL
Answer: (SHOW ANSWER)

NEW QUESTION: 79
A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table.
The code block used by the data engineer is below:

```
(spark.readStream
    .table("sales")
    .withColumn("avg_price", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .______
    .table("new_sales")
)
```

GAQM Global Association for Quality Management
freecram.net

If the data engineer only wants the query to process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

A. trigger(parallelBatch=True)

B. processingTime(1)

C. trigger(processingTime="once")

D. trigger(availableNow=True)

E. trigger(continuous="once")

Answer: (SHOW ANSWER)

NEW QUESTION: 80

Which Databricks SQL predicate correctly performs a null-safe equality comparison so that rows are matched when both sides are NULL or when both are equal non-NULL values?

Freecram.net

A. WHERE customer_id <= > :cid

B. WHERE customer_id = :cid

C. WHERE COALESCE(customer_id, '') = COALESCE(:cid, '')

D. WHERE customer_id <> :cid OR customer_id IS NULL

Answer: (SHOW ANSWER)

Databricks SQL uses the null-safe equality operator <=>. It returns TRUE when both values are NULL or when both non-NULL values are equal. The operator must be written without spaces:

customer_id <=> :cid.

NEW QUESTION: 81

In a healthcare provider organization using Delta Lake to store electronic health records (EHRs), a data analyst needs to analyze a snapshot of the patient_records table from two weeks ago before some recent data corrections were applied. What approach should the Data Engineer take to allow the analyst to query that specific prior version?

A. Truncate the table to remove all data, then reload the data from two weeks ago into the truncated table for the analyst to query.

B. Identify the version number corresponding to two weeks ago from the Delta transaction log, share that version number with the analyst to query using VERSION AS OF syntax, or export that version to a new Delta table for the analyst to query.

C. Restore the table to the version from two weeks ago using the RESTOREcommand, and have the analyst query the restored table.

D. Use the VACUUMcommand to remove all versions of the table older than two weeks, then the analyst can query the remaining version.

Answer: (SHOW ANSWER)

NEW QUESTION: 82

A data engineer observes that an ETL job in Databricks has increased its execution time from 5 minutes (baseline) to 12 minutes over the past week. Using the Lakeflow Jobs run history view, the engineer needs to identify whether the slowdown is consistent or intermittent and determine the root cause. Which action should the engineer take to analyze execution time trends against the baseline?

A. Compare cluster metrics in Lakeflow Jobs run history to spot resource spikes causing delays.

B. Archive old job runs and resets the baseline counter to ensure accurate comparison of current performance.

C. Rerun the job immediately to establish a new baseline since historical data becomes irrelevant after seven days.

D. Increase cluster size automatically whenever execution time exceeds the previous run's duration.

Answer: A (LEAVE A REPLY)

Comparing historical run durations shows whether the slowdown is recurring or sporadic, while compute metrics such as CPU, memory, and worker utilization help identify resource bottlenecks associated with slower runs.

NEW QUESTION: 83
Differentiate between all-purpose clusters and jobs clusters.
A data engineering team has created a python notebook to load data from cloud storage, this job has been tested and now needs to be scheduled in production.
Which would be the best cluster to be used in this case?
A. Serverless SQL warehouse
B. Jobs Cluster
C. Any Unity Catalog-enabled cluster
D. All purpose cluster
Answer: B (LEAVE A REPLY)

NEW QUESTION: 84
A data engineer is facing performance bottlenecks in an e-commerce transactions Delta table.
The table is a managed Unity Catalog table, and it uses partitioning and Z-ordering in its data layout scheme. The predictive optimization for Unity Catalog tables is also enabled. The table has a frequently changing query filter, and the data engineer does not observe a benefit of the Data Layout or the Predictive Optimization. How should the data engineer fix the data layout bottlenecks?
A. Re-write the Data Layout with Liquid Clustering and cluster by the Z-Ordered columns.
B. Enable Delta Caching so that query results can be read through caches.
C. Tweak the Z-Order columns and run OPTIMIZE manually.
D. Switch the Data layout from Partition+Z-Ordering to Automatic Liquid Clustering.
Answer: (SHOW ANSWER)
When queries use frequently changing filter conditions, partitioning and Z-ordering lose effectiveness because they rely on fixed columns. The correct solution is to switch to Automatic Liquid Clustering, which dynamically reorganizes data based on observed query patterns, removing the need for manual Z-ordering or static partitioning and resolving the data layout bottleneck.

NEW QUESTION: 85
A data engineer needs to access the view created by the sales team, using a shared cluster. The data engineer has been provided usage permissions on the catalog and schema. In order to access the view created by sales team. What are the minimum permissions the data engineer would require in addition?
A. Needs SELECT permission on the VIEW and the underlying TABLE.
B. Needs ALL PRIVILEGES at the SCHEMA level
C. Needs ALL PRIVILEGES on the VIEW
D. Needs SELECT permission only on the VIEW
Answer: (SHOW ANSWER)

NEW QUESTION: 86
A data engineer is writing a DataFrame to a Delta table and wants to physically divide the data into directories based on a specific column such as country. Which Spark DataFrame writer option should be used?

A. partitionBy
B. groupBy
C. distributeBy
D. orderBy
Answer: (SHOW ANSWER)

NEW QUESTION: 87
A data engineer is using the Databricks OPTIMIZE command on a Delta table. What happens when OPTIMIZE is run twice on the same table with the same data?
A. It has no effect because it is idempotent.
B. It changes the number of tuples per file significantly.
C. It further reduces file sizes by re-clustering the data.
D. It triggers a full liquid clustering process.
Answer: (SHOW ANSWER)
The OPTIMIZE command is idempotent - running it again on the same Delta table with unchanged data has no additional effect, since the files are already compacted.

NEW QUESTION: 88
Which of the following statements regarding the relationship between Silver tables and Bronze tables is always true?
A. Silver tables contain a more refined and cleaner view of data than Bronze tables.
B. Silver tables contain a less refined, less clean view of data than Bronze data.
C. Silver tables contain more data than Bronze tables.
D. Silver tables contain less data than Bronze tables.
E. Silver tables contain aggregates while Bronze data is unaggregated.
Answer: (SHOW ANSWER)

NEW QUESTION: 89
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. Unity Catalog
C. Cloud File Storage
D. Data lake
Answer: (SHOW ANSWER)
Delta Lake is the optimized storage layer that adds ACID transactions, schema enforcement, and reliability on top of cloud object storage, making it ideal for managing large-scale data workloads.

NEW QUESTION: 90
Which of the following describes the type of workloads that are always compatible with Auto Loader?
A. Serverless workloads
B. Machine learning workloads

C. Dashboard workloads
D. Batch workloads
E. Streaming workloads
Answer: (SHOW ANSWER)

NEW QUESTION: 91
A data analyst has developed a query that runs against 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.table
C. spark.delta.table
D. There is no way to share data between PySpark and SQL.
E. spark.sql
Answer: (SHOW ANSWER)

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 92
A data engineering team must version-control notebooks, review changes through pull requests, and deploy the same code to dev, staging, and production workspaces.
Which Databricks capability supports this workflow?
A. Databricks Git folders (Repos) integrated with a remote Git provider
B. Exporting notebooks as DBC archives and emailing them between workspaces
C. Storing notebook source files in DBFS /FileStore and copying them manually
D. Enabling notebook revision history and restoring snapshots as needed
Answer: (SHOW ANSWER)
Databricks Git folders provide native integration with providers such as GitHub, GitLab, Azure DevOps, and Bitbucket, supporting branching, commits, pulls, and pull-request workflows.
Combined with Databricks Asset Bundles or job Git sources, the same repository can be deployed across environments.

NEW QUESTION: 93
Which languages are supported by Serverless compute clusters? (Choose two.)

A. SQL
B. Python
C. R
D. Scala
E. Java
Answer: (SHOW ANSWER)
Serverless compute clusters in Databricks support SQL (via Serverless SQL Warehouses) and Python (for interactive and job-based execution). Other languages like R, Scala, and Java require regular clusters.

NEW QUESTION: 94
Which of the following is hosted completely in the control plane of the classic Databricks architecture?
A. Worker node
B. JDBC data source
C. Databricks web application
D. Databricks Filesystem
E. Driver node
Answer: (SHOW ANSWER)
The control plane in the classic Databricks architecture is responsible for managing the Databricks workspace, user and group management, and cluster management, among other things. The Databricks web application is a part of the control plane that enables users to interact with the workspace, create and manage clusters, and work with notebooks, jobs, and data.
Worker nodes and driver nodes are part of the data plane, which is responsible for executing data processing tasks. JDBC data sources and the Databricks Filesystem are services that are used by both the control plane and the data plane.

NEW QUESTION: 95
Which of the following benefits is provided by the array functions from Spark SQL?
A. An ability to work with complex, nested data ingested from JSON files
B. An ability to work with data in a variety of types at once
C. An ability to work with an array of tables for procedural automation
D. An ability to work with time-related data in specified intervals
E. An ability to work with data within certain partitions and windows
Answer: (SHOW ANSWER)

NEW QUESTION: 96
A data engineer needs to enforce row-level security on main.secure.events(region STRING, event_id STRING). Members of the account group all_regions need to see all rows. All other users need to see only rows where region = 'EU'. Which SQL sequence satisfies the requirement?
A.
```
CREATE FUNCTION main.secure.region_filter(region STRING)
RETURN IF(is_account_group_member('all_regions'), true, region = 'EU');
```

B.
```
ALTER TABLE main.secure.events
SET ROW FILTER main.secure.region_filter ON (region);
CREATE FUNCTION main.secure.region_filter(region STRING)
RETURN IF(is_account_group_member('all_regions'), true, region = 'EU');
```

C.
```
ALTER VIEW main.secure.events
SET ROW FILTER main.secure.region_filter ON (region);
CREATE FUNCTION main.secure.region_filter(region STRING)
RETURN IF(is_account_group_member('all_regions'), true, region = 'EU');
```

D.
```
ALTER TABLE main.secure.events
ALTER COLUMN region
SET MASK main.secure.region_filter;
CREATE FUNCTION main.secure.region_filter(region STRING)
RETURN IF(is_account_group_member('all_regions'), true, region = 'EU');

ALTER TABLE main.secure.events
SET ROW FILTER main.secure.region_filter USING COLUMNS (region);
```

GAQM Global Association for Quality Management freecram.net

Answer: (SHOW ANSWER)

The Boolean UDF must be created first and then assigned to the table using SET ROW FILTER ... ON. It returns TRUE for every row when the user belongs to all_regions; otherwise, it returns TRUE only for EU rows.

NEW QUESTION: 97

A data engineer is investigating a Lakehouse ETL job on Databricks that has suddenly become four times slower.

They need to:

- Identify the specific Spark stages that regressed.
- Drill down to per-task metrics such as shuffle read/write size,

spill, and GC time.

- Correlate these with the executor resource usage during that run.

Which Databricks-native capability is the starting point for this detailed job-level root-cause analysis?

A. Cluster Metrics (Metrics tab on the compute) to view historical Spark and hardware metrics, without using the Spark UI.

B. The Jobs run details page timeline and graph views, without opening the Spark UI, to analyze all stage and task-level Spark metrics.

C. Ganglia metrics for the cluster, which provide full visibility into Spark stages, tasks, and shuffle metrics for the job run.

D. Spark UI opened from the job run's compute, using the Jobs and Stages tabs to inspect task metrics.

Answer: (SHOW ANSWER)

The Spark UI provides detailed visibility into job execution, including stages, tasks, and granular metrics such as shuffle I/O, spill, and GC time, making it the primary tool for deep root-cause analysis of performance regressions.

NEW QUESTION: 98

A production ETL notebook runs once nightly and currently executes on a shared all-purpose cluster that stays running 24 hours a day.

Which change reduces cost while preserving the schedule?

A. Increase the all-purpose cluster's minimum worker count to speed up the run.

B. Configure the scheduled job to use a job cluster that is created at run start and terminated at run end.
C. Move the notebook to a SQL warehouse with auto-stop disabled.
D. Convert the notebook to a global temp view refreshed on demand.
Answer: B (LEAVE A REPLY)
Job clusters are provisioned when the job run starts and terminated when it completes, so compute is billed only for the run duration and at the lower Jobs Compute rate. All-purpose clusters left running continuously are billed at a higher rate even when idle.

NEW QUESTION: 99
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 AS
SELECT * FROM march_transactions
MERGE SELECT * FROM april_transactions;
B. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
INTERSECT SELECT * from april_transactions;
C. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
OUTER JOIN SELECT * FROM april_transactions;
D. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
INNER JOIN SELECT * FROM april_transactions;
E. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
UNION SELECT * FROM april_transactions;
Answer: (SHOW ANSWER)

NEW QUESTION: 100
A data engineer uploads a CSV file using the "Create or modify a table using file upload" option in Databricks. To avoid incorrect schema inference, they disable Automatically detect column types before creating a Unity Catalog-managed table. What is the outcome?
Freecram.net
A. All columns are created with the STRING data type.
B. Numeric columns are inferred as numeric types; others are STRING.
C. Previously inferred column types are preserved.
D. Table creation fails because column types are not specified.
Answer: (SHOW ANSWER)
Disabling Automatically detect column types causes Databricks to create every uploaded CSV column as STRING, avoiding automatic numeric, date, or Boolean inference.

NEW QUESTION: 101
A data engineer must support self-serve BI dashboards for hundreds of business users who run ad hoc, high-concurrency SQL queries throughout the day against governed Unity Catalog tables.
The team needs a near-instant start, autoscaling without manual tuning, and the best performance for SQL while minimizing operational overhead. Which Databricks compute should the data engineer use?
A. SQL Warehouse (Classic) with a fixed cluster size.
B. Jobs compute clusters triggered by dashboards on a schedule.
C. All-purpose classic cluster with manual autoscaling enabled.
D. SQL Warehouse (Serverless) with Photon enabled.
Answer: (SHOW ANSWER)
Serverless SQL Warehouses provide instant startup, automatic scaling, and optimized performance with Photon for high-concurrency BI workloads, while minimizing operational overhead and management effort.

NEW QUESTION: 102
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. Migrating the dataset to Cloudflare R2 object storage before sharing
B. Configure VPN connection between AWS and Azure for faster data sharing
C. Using Delta Sharing without any additional configurations
D. Sharing data via pre-signed URLs without monitoring egress costs
Answer: (SHOW ANSWER)
Delta Sharing natively supports cross-cloud data access without requiring data duplication or special networking setups. It allows secure sharing of datasets across different cloud providers while minimizing egress costs, since recipients query the data directly from its source location rather than transferring large volumes unnecessarily.

NEW QUESTION: 103
Which of the following Structured Streaming queries is performing a hop from a Silver table to a Gold table?
A.
```
(spark.readStream.load(rawSalesLocation)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```
B.
```
(spark.table("sales")
    .withColumn("avgPrice", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

```
(spark.table("sales")
    .groupBy("store")
    .agg(sum("sales"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .table("newSales")
)
```

C.

```
(spark.read.load(rawSalesLocation)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

D.

```
(spark.table("sales")
    .filter(col("units") > 0)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

E.

Answer: (SHOW ANSWER)

NEW QUESTION: 104

Which of the following describes the storage organization of a Delta table?

A. Delta tables are stored in a single file that contains only the data stored within the table.

B. Delta tables store their data in a single file and all metadata in a collection of files in a separate location.

C. Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.

D. Delta tables are stored in a collection of files that contain only the data stored within the table.

E. Delta tables are stored in a single file that contains data, history, metadata, and other attributes.

Answer: (SHOW ANSWER)

NEW QUESTION: 105

Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

A. Data lakehouse

B. Data warehouse

C. None of these

D. All of these

E. Data lake

Answer: (SHOW ANSWER)

NEW QUESTION: 106

A data engineer is reviewing the documentation on audit logs in Databricks for compliance purposes and needs to understand the format in which audit logs output events. How are events formatted in Databricks audit logs?

A. In Databricks, audit logs output events in a JSON format.

B. In Databricks, audit logs output events in a CSV format.

C. In Databricks, audit logs output events in an XML format.

D. In Databricks, audit logs output events in a plain text format.

Answer: (SHOW ANSWER)

Databricks audit logs output events in JSON format, which provides structured, machine-readable data suitable for compliance monitoring and integration with security tools.

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 107

A data engineer needs to find all customers who have never placed an order. The customers table contains customer_id, customer_name, and other customer attributes. The orders table contains order_id, customer_id, and other order attributes. Both tables share customer_id as the join key. The result must include only customer columns with no order-related columns in the output. Which expression exactly achieves this requirement?

A. Inner join, then filter for NULL order_id

B. Left anti-join from customers to orders

C. Left outer join from orders to customers

D. Full outer join, then filter where order_id IS NULL

Answer: (SHOW ANSWER)

A left anti-join returns only rows from customers that have no matching customer_id in orders, and it outputs only columns from the left-side customers table.

NEW QUESTION: 108

A data engineer ingests semi-structured JSON logs into a Delta table using Auto Loader with schema evolution enabled. A new string field userAgent appears in the JSON. What happens to the new userAgent field?

A. All existing data is backfilled with a default value for userAgent and made non-nullable.

B. The new userAgent field is added as a new nullable column in the Delta table, populated for new data.

C. Ingestion fails with a schema-mismatch error until the table is altered.

D. The new userAgent field is silently ignored, and its data is dropped for all new files.

Answer: (SHOW ANSWER)

With schema evolution enabled, Auto Loader adds userAgent as a nullable column. Existing rows have NULL, while new records populate the field when present.

NEW QUESTION: 109
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. Repair the task
B. Rerun the pipeline
C. Restart the cluster
D. Switch the cluster
Answer: (SHOW ANSWER)
For a single-task workflow, the correct action is to repair the task, which reruns only the failed task after fixing the error, avoiding unnecessary re-execution of the whole workflow.

NEW QUESTION: 110
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. FROM CSV
B. None of these lines of code are needed to successfully complete the task
C. FROM "path/to/csv"
D. USING DELTA
E. USING CSV
Answer: (SHOW ANSWER)

NEW QUESTION: 111
Which of the following describes the relationship between Bronze tables and raw data?
A. Bronze tables contain a less refined view of data than raw data.
B. Bronze tables contain raw data with a schema applied.
C. Bronze tables contain less data than raw data files.

D. Bronze tables contain more truthful data than raw data.
E. Bronze tables contain aggregates while raw data is unaggregated.
Answer: (SHOW ANSWER)

NEW QUESTION: 112
A data engineer has a Delta table containing old data files that are no longer referenced after several update and delete operations. To remove these unused files and reclaim storage space, which Delta Lake command should be executed?
A. CACHE TABLE
B. OPTIMIZE
C. VACUUM
D. DESCRIBE HISTORY
Answer: (SHOW ANSWER)

NEW QUESTION: 113
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. Dashboards
C. Catalog Explorer
D. Repos
Answer: (SHOW ANSWER)
https://docs.databricks.com/en/delta-live-tables/tutorial-pipelines.html

NEW QUESTION: 114
A data engineer investigates a slow stage in the Databricks Spark UI. The engineer scrolls down to the Summary Metrics section for task durations, and they observe for that stage:
25th percentile duration: 40 seconds
Median duration: 45 seconds
75th percentile duration: 50 seconds
Max duration: 80 seconds
What is the interpretation of these metrics?
A. There is a GC issue.
B. The stage likely suffers from data skew.
C. The stage is healthy.
D. The cluster is under-provisioned.
Answer: (SHOW ANSWER)
The maximum task duration is 60% higher than the 75th percentile (80s versus 50s). Databricks identifies this pattern as a likely indication that a small number of tasks process disproportionately large partitions-data skew.

NEW QUESTION: 115

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. Automatic data quality checks, built-in support for schema evolution, and declarative pipeline development

B. Manual schema enforcement, high operational overhead, and limited scalability

C. Requires custom code for data quality checks, no support for streaming data, and complex pipeline maintenance

D. Supports only batch processing, no data versioning, and high infrastructure costs

Answer: (SHOW ANSWER)

Delta Live Tables (DLT) is appropriate because it provides automatic data quality checks through expectations, built-in support for schema evolution, and a declarative approach to pipeline development, making it easier to build and maintain both streaming and batch ETL pipelines.

NEW QUESTION: 116

A new data engineering team team. has been assigned to an ELT project. The new data engineering team will need full privileges on the database customers 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 DATABASE customers TO team;

B. GRANT USAGE ON DATABASE customers TO team;

C. GRANT SELECT PRIVILEGES ON DATABASE customers TO teams;

D. GRANT ALL PRIVILEGES ON DATABASE team TO customers;

E. GRANT SELECT CREATE MODIFY USAGE PRIVILEGES ON DATABASE customers TO team;

Answer: (SHOW ANSWER)

NEW QUESTION: 117

A company uses Delta Sharing to collaborate with partners across different cloud providers and geographic regions. What will result in additional costs due to cross-region or egress fees?

A. Sharing data within the same cloud provider and region

B. Transferring data via Delta Sharing across clouds and across different geographic regions

C. Accessing Delta Sharing data using a VPN within the same data center

D. Utilizing Delta Sharing for internal data analytics within a single cloud environment

Answer: (SHOW ANSWER)

Transferring data via Delta Sharing across clouds and across different geographic regions incurs additional costs because cloud providers charge cross-region and egress fees when data leaves a region or moves between providers.

NEW QUESTION: 118

A data engineer has finished developing a new Declarative Automation Bundle (formerly Databricks Asset Bundle) for a streaming pipeline. They need to ensure the configuration is syntactically correct, push the code to the workspace, and finally trigger the pipeline to verify it runs as expected. Which sequence of Databricks CLI commands should the engineer use to perform these steps in the correct order?

A. databricks bundle init -> databricks bundle deploy -> databricks bundle run

B. databricks bundle validate -> databricks bundle deploy -> databricks bundle run

C. databricks bundle validate -> databricks bundle sync -> databricks bundle execute

D. databricks bundle check -> databricks bundle upload -> databricks bundle start
Answer: (SHOW ANSWER)
databricks bundle validate checks the bundle configuration, databricks bundle deploy uploads and deploys its resources to the workspace, and databricks bundle run triggers the deployed pipeline.

NEW QUESTION: 119
A data engineer needs to ingest JSON change data from Salesforce into Unity Catalog-governed Delta tables with a low-code, fully managed experience. Which Databricks capability should the data engineer use?
A. A Lakeflow Connect managed Salesforce connector that writes into Unity Catalog Delta tables
B. Use the Salesforce REST API in a PySpark notebook to query changed records and write them to Delta tables
C. Auto Loader reading JSON files from an S3 bucket exported by Salesforce
D. Auto Loader with a file arrival trigger on DBFS
Answer: (SHOW ANSWER)
Lakeflow Connect's managed Salesforce connector provides a low-code, fully managed ingestion pipeline that incrementally loads Salesforce data into Unity Catalog-governed Delta tables.

NEW QUESTION: 120
A data engineer is attempting to drop a Spark SQL table my_table and runs the following command:
DROP TABLE IF EXISTS my_table;
After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.
Freecram.net
Which of the following describes why all of these files were deleted?
A. The table did not have a location
B. The table was managed
C. The table was external
D. The table's data was larger than 10 GB
E. The table's data was smaller than 10 GB
Answer: (SHOW ANSWER)

NEW QUESTION: 121
A data engineer needs to ingest streaming data from Kafka into a Delta Lake table using Databricks. The pipeline must guarantee exactly-once processing semantics and maintain checkpoint information to recover from failures. Which Spark Structured Streaming feature ensures reliable processing and fault tolerance?
A. Checkpointing
B. Partition pruning
C. Watermarking
D. Caching
Answer: (SHOW ANSWER)

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 122
In order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing, which of the following two approaches is used by Spark to record the offset range of the data being processed in each trigger?
A. Checkpointing and Write-ahead Logs
B. Replayable Sources and Idempotent Sinks
C. Structured Streaming cannot record the offset range of the data being processed in each trigger.
D. Write-ahead Logs and Idempotent Sinks
E. Checkpointing and Idempotent Sinks
Answer: (SHOW ANSWER)

NEW QUESTION: 123
An analyst reports that a dashboard showed incorrect numbers before a data correction was applied at 09:00 today. The engineer needs to query the table exactly as it existed at version 118.
Which statement returns that data?
A. SELECT * FROM sales VERSION AS OF 118
B. DESCRIBE HISTORY sales LIMIT 118
C. RESTORE TABLE sales TO VERSION AS OF 118
D. SELECT * FROM sales WHERE _commit_version = 118
Answer: (SHOW ANSWER)
Delta time travel supports VERSION AS OF <n> (or TIMESTAMP AS OF) to read a prior snapshot without modifying the table. RESTORE would actually roll the table back, which changes the current state rather than simply inspecting history.

NEW QUESTION: 124
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;
```

freecram.net GAQM Global Association for Quality Management

B.
```
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;
```

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

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

Answer: (SHOW ANSWER)

The correct approach is to use a DLT expectation with constraint CONSTRAINT valid_date EXPECT (order_datetime IS NOT NULL) ON VIOLATION DROP ROW This ensures rows with NULL order_datetime are filtered out automatically during streaming ingestion while keeping the pipeline valid.

NEW QUESTION: 125

A data engineer converts an external table to a managed table. A structured streaming job that reads from this table continues running during the conversion. After the conversion completes, the streaming job stops processing new records. How should the data engineer resolve this?

A. Grant the streaming job additional permissions on the new managed storage location

B. Restart the streaming job, so it picks up the new managed table location

C. Run REFRESH TABLE on the converted table to update the streaming checkpoint

D. Delete the streaming checkpoint directory and let the job reprocess from the beginning

Answer: (SHOW ANSWER)

Converting the table changes its underlying storage location. Databricks requires affected streaming jobs to be restarted; the stream then resumes from its existing checkpoint and reads from the new managed location. The checkpoint should not be deleted.

NEW QUESTION: 126

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 schema(schema) with option ("maxFilesPerTrigger", 1)
B. Replace predict with a stream-friendly prediction function
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)

NEW QUESTION: 127
A data engineer needs to optimize the data layout and query performance for an e-commerce transactions Delta table. The table is partitioned by "purchase_date" a date column which helps with time-based queries but does not optimize searches on user statistics "customer_id", a high- cardinality column.
The table is usually queried with filters on "customer_id" within specific date ranges, but since this data is spread across multiple files in each partition, it results in full partition scans and increased runtime and costs.
How should the data engineer optimize the Data Layout for efficient reads?
A. Alter table implementing liquid clustering on "customer_id" while keeping the existing partitioning.
B. Alter the table implementing liquid clustering by "customer_id" and "purchase_date".
C. Alter the table to partition by "customer_id".
D. Enable delta caching on the cluster so that frequent reads are cached for performance.
Answer: (SHOW ANSWER)
Since queries filter by customer_id within purchase_date ranges, partitioning only by purchase_date causes inefficient scans across partitions. The correct optimization is to use Liquid Clustering on both customer_id and purchase_date, which dynamically reorganizes data to optimize for both dimensions and avoids the drawbacks of high-cardinality static partitioning.

NEW QUESTION: 128
A data engineer wants to update specific rows in a Delta table based on matching keys from another dataset. The operation must support both insert and update logic in a single statement.
Which Delta Lake command supports this functionality?
A. UPDATE
B. INSERT INTO
C. MERGE INTO
D. DELETE
Answer: (SHOW ANSWER)

NEW QUESTION: 129
A data engineering team uses Declarative Automation Bundles (formerly Databricks Asset Bundles) to deploy the same codebase across dev, test, and prod environments. The team wants environment-specific behavior to be applied only through bundle configuration, without modifying notebooks, job definitions, or deployment logic during promotion. Which approach applies environment-specific configuration while meeting this requirement?
A. Use separate Git branches per environment so each branch can contain environment-specific configuration values.
B. Define bundle variables and reference them from bundle resources, then override variable values per environment using bundle targets.

C. Parameterize environment-specific values inside notebooks and set those parameters when each job run is triggered.
D. Deploy the same bundle once and edit job settings in the workspace UI for each environment after deployment.
Answer: (SHOW ANSWER)
Bundle variables hold reusable configuration values, and bundle targets override those values for dev, test, and prod at deployment time. This preserves one codebase and avoids modifying notebooks or workspace resources during promotion.

NEW QUESTION: 130
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?
Freecram.net
A. failOnNewColumns
B. none
C. rescue
D. addNewColumns
Answer: (SHOW ANSWER)
The failOnNewColumns option enforces strict schema control during ingestion. If new columns appear in the source, the ingestion job fails immediately, ensuring downstream changes are reviewed and validated before processing continues.

NEW QUESTION: 131
A data engineer is designing a Gold layer table for a BI team that needs to query complex aggregations. The requirements state that the data should be precomputed to ensure low latency for dashboard users, but the underlying data changes only once every few hours. Which object type should be implemented in Unity Catalog to meet these requirements?
A. Streaming table
B. External table
C. Materialized view
D. View
Answer: (SHOW ANSWER)
A materialized view precomputes and stores the aggregation results, providing low-latency BI queries. It can be refreshed on a schedule to reflect source changes occurring every few hours.

NEW QUESTION: 132
A data engineer is decommissioning a sandbox schema in Unity Catalog. Some tables are ephemeral staging outputs that can be safety removed entirely, but a few tables point at shared cloud storage used by downstream jobs outside Databricks. The engineer must avoid deleting any shared files when cleaning up catalog objects.
How does Unity Catalog behave when dropping Managed vs External tables?
A. DROP all tables; Databricks will only remove metadata for both managed and external tables.
B. DROP managed tables that are ephemeral and DROP external tables; files for both remain for 7 days.
C. DROP managed staging tables to remove data and metadata, and DROP external tables to remove only metadata
D. DROP external tables first to delete their files, then DROP managed tables to keep their files for recovery.
Answer: (SHOW ANSWER)

In Unity Catalog, dropping a managed table deletes both the table metadata and the underlying data files, while dropping an external table removes only the catalog metadata and leaves the referenced cloud storage data intact.

NEW QUESTION: 133
A data engineer is working on a Databricks project that utilizes cloud storage. The data engineer wants to load several JSON files from containers on a storage account as soon as the file arrives within the storage account. Which syntax should the data engineer follow to first load the files into a dataframe and check that it is working as expected using Python?
A. df = spark.read.json("input/path")
B. df = spark.readStream.format("cloud").option("json").load("/input/path")
C. df = spark.readStream.format("json".load("input/path")
D. df = spark.readStream.format("cloudFiles").option("cloudFiles.format", "json").load("/input/path")
Answer: (SHOW ANSWER)
To automatically ingest JSON files as soon as they arrive in cloud storage, the correct syntax is to use Auto Loader with spark.readStream.format("cloudFiles").option("cloudFiles.format", "json").load("/input/path"). This enables continuous ingestion of new files into a streaming dataframe.

NEW QUESTION: 134
A data engineer is building a streaming pipeline that reads JSON files arriving in a cloud storage location. The pipeline must automatically process new files as they arrive without manual intervention. Which Databricks feature is designed for this use case?
A. Broadcast Join
B. Delta Time Travel
C. Structured Streaming Auto Loader
D. Spark Cache
Answer: (SHOW ANSWER)

NEW QUESTION: 135
A data engineer needs to parse only png files in a directory that contains files with different suffixes. Which code should the data engineer use to achieve this task?
A. df = spark.readStream.format("cloudFiles") \ .option("cloudFiles.format", "binaryFile") \
.append("/*.png")
B. df = spark.readstream. format("cloudFiles") \
.option("cloudFiles.format", "binaryFile") \
.option("pathGlobfilter", "*.png") \
.load(<base-path>)
C. df = spark.readStream.format("cloudFiles") \ .option("cloudFiles.format", "binaryFile") \
.option("pathGlobfilter", "*.png") \ .append()
D. df = spark.readstream.format("cloudFiles") \ .option("cloudFiles.format", "binaryFile") \
.load("/*.png")
Answer: (SHOW ANSWER)
The correct way to filter files by suffix in Auto Loader is to use the pathGlobFilter option. For parsing only .png files, the code should be:
df = spark.readStream.format("cloudFiles") \

.option("cloudFiles.format", "binaryFile") \
.option("pathGlobfilter", "*.png") \
.load(<base-path>)
This ensures only .png files are ingested from the directory.

NEW QUESTION: 136
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 configure the clusters to autoscale for larger data sizes
B. They can use endpoints available in Databricks SQL
C. They can use jobs clusters instead of all-purpose clusters
D. They can use clusters that are from a cluster pool
E. They can configure the clusters to be single-node
Answer: (SHOW ANSWER)

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 137
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. Serverless compute for notebooks
B. Pro SQL Warehouse
C. Classic SQL Warehouse
D. Serverless SQL Warehouse
Answer: B (LEAVE A REPLY)
A Pro SQL Warehouse is designed for high-performance, cost-effective query execution at scale.
It is optimized for running large volumes of queries quickly, making it ideal for exploratory analysis on enterprise datasets.

NEW QUESTION: 138
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 provides the ability to comment on specific changes
C. Databricks Repos supports the use of multiple branches
D. Databricks Repos is wholly housed within the Databricks Data Intelligence Platform
E. Databricks Repos allows users to revert to previous versions of a notebook
Answer: (SHOW ANSWER)

NEW QUESTION: 139
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 allows the use of SQL queries to examine data.
C. A data lakehouse enables machine learning and artificial Intelligence workloads.
D. A data lakehouse supports ACID-compliant transactions.
E. A data lakehouse stores data in open formats.
Answer: (SHOW ANSWER)

NEW QUESTION: 140
Which method should a Data Engineer apply to ensure Workflows are being triggered on schedule?
A. Scheduled Workflows require an always-running cluster, which is more expensive but reduces processing latency.
B. Scheduled Workflows process data as it arrives at configured sources.
C. Scheduled Workflows can reduce resource consumption and expense since the cluster runs only long enough to execute the pipeline.
D. Scheduled Workflows run continuously until manually stopped.
Answer: (SHOW ANSWER)
Scheduled Workflows in Databricks are designed to trigger jobs at specified intervals. These workflows can be configured to start a cluster when the job is scheduled, and the cluster will only run for the duration of the task. Once the workflow completes, the cluster can be terminated, which helps to reduce resource consumption and costs since you are only paying for the cluster while the workflow is running.

NEW QUESTION: 141
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 USAGE ON TABLE sales TO team;
B. GRANT ALL PRIVILEGES ON TABLE team TO sales;
C. GRANT ALL PRIVILEGES ON TABLE sales TO team;
D. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
E. GRANT SELECT ON TABLE sales TO team;
Answer: (SHOW ANSWER)

NEW QUESTION: 142
A data engineering project involves processing large batches of data on a daily schedule using ETL. The jobs are resource-intensive and vary in size, requiring a scalable, cost-efficient compute solution that can automatically scale based on the workload. Which compute approach will satisfy the needs described?

A. Job Cluster
B. Dedicated Cluster
C. All-Purpose Cluster
D. Databricks SQL Serverless
Answer: (SHOW ANSWER)
Job Clusters are ephemeral clusters created for scheduled or automated workloads. They are cost-efficient, scale automatically to match workload demands, and terminate after job completion, making them the ideal choice for daily large-batch ETL processing.

NEW QUESTION: 143
What are the transformations typically included in building the Bronze layer?
A. Include columns Load date/time, process ID
B. Business rules and transformations
C. Perform extensive data cleansing
D. Aggregate data from multiple sources
Answer: A (LEAVE A REPLY)
The Bronze layer stores raw ingested data with minimal transformations, typically adding metadata columns such as load date/time and process ID for traceability and auditing.

NEW QUESTION: 144
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. COMMENT "Contains PII"
B. TBLPROPERTIES PII
C. PII
D. "COMMENT PII"
E. There is no way to indicate whether a table contains PII.
Answer: (SHOW ANSWER)

NEW QUESTION: 145
A data engineer is getting a partner organization up to speed with Databricks account. Both the 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. Share required datasets and notebooks via Delta Sharing. Manage permissions via Unity Catalog.
C. Data and Notebooks can be shared simply using Unity Catalog.
D. Share access to codebase via Github and allow them to ingest datasets from your Datalake.
Answer: (SHOW ANSWER)
The correct approach is to use Delta Sharing to share Unity Catalog-managed Delta tables and notebooks securely with the partner organization. Permissions are managed through Unity Catalog, ensuring governance, access control, and seamless collaboration without manual file transfers or duplications.

NEW QUESTION: 146
A data engineer needs to support a Unity Catalog Delta table catalog.schema.sales that is currently registered as an external table. New requirements state:
- Automatic file cleanup after a drop is handled by the platform.
- Minimal downtime during migration while keeping the same table name,
permissions, and history.
- Enforced name-based access through Unity Catalog (no path-based
access).
Which action should the data engineer take to meet the requirements?
A. Run ALTER TABLE catalog.schema.sales SET MANAGED to convert and meet the goals
B. Run DROP TABLE catalog.schema.sales PURGE to delete files and meet goals
C. Run ALTER TABLE catalog.schema.sales SET LOCATION to switch external paths now
D. Create a new managed table with CTAS and migrate workloads to that table today
Answer: (SHOW ANSWER)
ALTER TABLE catalog.schema.sales SET MANAGED converts the external Delta table in place while retaining its name, permissions, configuration, views, and history, with minimal downtime.
As a managed table, Unity Catalog controls its storage lifecycle, including cleanup after deletion, and workloads should access it by its three-part table name.

NEW QUESTION: 147
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. Use Lakehouse Federation to query both data sources directly
C. Manually synchronize data from both sources into a single database
D. Develop custom ETL pipelines to ingest data into Databricks
Answer: (SHOW ANSWER)
Lakehouse Federation allows Databricks to directly query external data sources like PostgreSQL and Azure Synapse without duplicating data. This ensures the report always uses up-to-date information while avoiding the overhead and cost of data movement.

NEW QUESTION: 148

A data engineer must create a copy of a production Delta table for a one-time schema experiment. The copy must be fully independent so that later VACUUM operations on the source cannot invalidate it.
Which command should be used?
A. CREATE TABLE test_orders SHALLOW CLONE prod_orders
B. CREATE TABLE test_orders DEEP CLONE prod_orders
C. CREATE VIEW test_orders AS SELECT * FROM prod_orders
D. CREATE TABLE test_orders LIKE prod_orders
Answer: (SHOW ANSWER)
A deep clone copies both the metadata and the data files, producing a fully independent table. A shallow clone copies only metadata and references the source table's files, so vacuuming or overwriting the source can break the clone.

NEW QUESTION: 149
In a Lakeflow Declarative Pipeline (Delta Live Tables), a data engineer must ensure that records with a null order_id are discarded from the silver table while the pipeline continues to run.
Which expectation clause should be used?
A. CONSTRAINT valid_id EXPECT (order_id IS NOT NULL) ON VIOLATION DROP ROW
B. CONSTRAINT valid_id EXPECT (order_id IS NOT NULL) ON VIOLATION FAIL UPDATE
C. CONSTRAINT valid_id EXPECT (order_id IS NOT NULL)
D. ALTER TABLE silver_orders ADD CONSTRAINT valid_id CHECK (order_id IS NOT NULL)
Answer: A (LEAVE A REPLY)
ON VIOLATION DROP ROW removes failing records from the target dataset and records the violation in the event log while the pipeline continues. An expectation with no ON VIOLATION clause retains bad records and only reports metrics, and FAIL UPDATE stops the pipeline.

NEW QUESTION: 150
A Databricks single-task workflow fails 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?
Freecram.net
A. Repair the task
B. Repair the run
C. Restart the Cluster
D. Swap the cluster
Answer: (SHOW ANSWER)
Rerunning a failed workflow after fixing the notebook requires repairing the run so the task is re- executed with the updated code, without recreating the entire workflow or changing cluster configuration.

NEW QUESTION: 151
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 use a batch source in place of a streaming source.
D. The pipeline will need to be written entirely in Python.
Answer: (SHOW ANSWER)
When migrating to Delta Live Tables (DLT), you can use both SQL and Python in the same pipeline. DLT supports multi-language notebooks, allowing data engineers to continue using Python for certain transformations and data analysts to use SQL for their specific requirements.

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)

NEW QUESTION: 152
Which of the following is stored in the Databricks customer's cloud account?
A. Data
B. Cluster management metadata
C. Databricks web application
D. Repos
E. Notebooks
Answer: (SHOW ANSWER)

NEW QUESTION: 153
A data analyst has a series of queries in a SQL program. The data analyst wants this program to run every day. They only want the final query in the program to run on Sundays. They ask for help from the data engineering team to complete this task.
Which of the following approaches could be used by the data engineering team to complete this task?
A. They could redesign the data model to separate the data used in the final query into a new table.
B. They could submit a feature request with Databricks to add this functionality.
C. They could only run the entire program on Sundays.
D. They could automatically restrict access to the source table in the final query so that it is only accessible on Sundays.
E. They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.
Answer: (SHOW ANSWER)

NEW QUESTION: 154
A data engineer is designing a bronze-to-silver pipeline on the Databricks Data Intelligence Platform. The source system sends daily CSV files, and over time, new optional columns are added.
The data engineer wants a storage format and table feature that together ensure the following:
- The table prevents writes that do not conform to the defined schema.

- The table can evolve its schema to include new optional columns
without having to manually recreate the table.
- Previous versions of the table can be queried later for debugging and auditing.
Which solution fulfills the requirements?
A. Use a Parquet table with Spark's default schema inference and rely on re-running the job when the schema changes.
B. Use an external table with Auto Loader schema inference for CSV files.
C. Use a Delta table with its native capabilities.
D. Use a Delta table with schema enforcement and recreate the table whenever new columns are added.
Answer: (SHOW ANSWER)
Delta Lake provides schema enforcement, supports schema evolution for newly added columns, and retains table history for time travel and auditing.

NEW QUESTION: 155
Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?
A. The ability to set up alerts for query failures
B. The ability to support batch and streaming workloads
C. The ability to distribute complex data operations
D. The ability to collaborate in real time on a single notebook
E. The ability to manipulate the same data using a variety of languages
Answer: (SHOW ANSWER)

NEW QUESTION: 156
A data engineer has realized that the data files associated with a Delta table are incredibly small.
They want to compact the small files to form larger files to improve performance.
Which of the following keywords can be used to compact the small files?
A. REPARTITION
B. COMPACTION
C. REDUCE
D. OPTIMIZE
E. VACUUM
Answer: (SHOW ANSWER)

NEW QUESTION: 157
A data engineer needs to read files from cloud object storage into a Spark DataFrame in Databricks. The files are stored in CSV format with headers and comma delimiters. Which Spark DataFrame reader option ensures that column names are correctly inferred from the first row?
A. inferSchema
B. delimiter
C. mode
D. header
Answer: (SHOW ANSWER)

Valid Databricks-Certified-Data-Engineer-Associate Dumps shared by ExamDiscuss.com for Helping Passing Databricks-Certified-Data-Engineer-Associate Exam! ExamDiscuss.com now offer the **newest Databricks-Certified-Data-Engineer-Associate exam dumps**, the ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate exam **questions have been updated** and **answers have been corrected** get the **newest** ExamDiscuss.com Databricks-Certified-Data-Engineer-Associate dumps with Test Engine here: https://www.examdiscuss.com/GAQM/exam/Databricks-Certified-Data-Engineer-Associate/premium/ (**322** Q&As Dumps, **35%OFF** **Special Discount Code:** **freecram**)