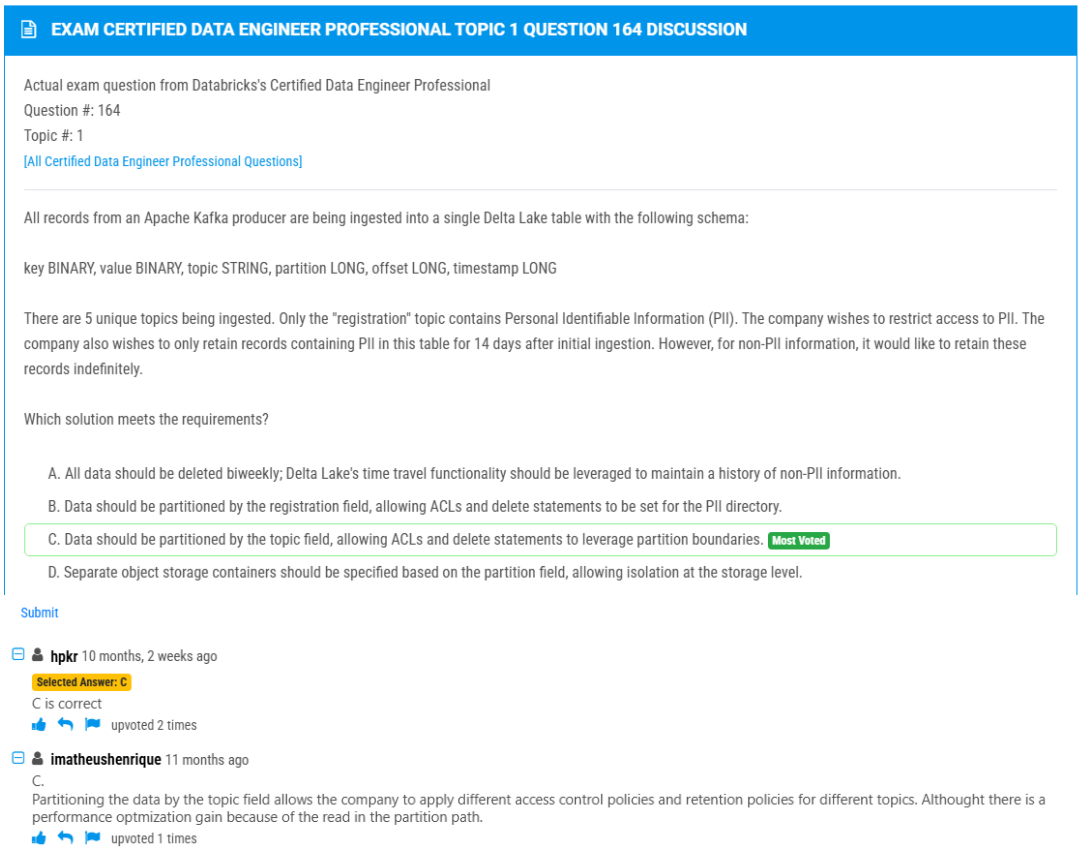
1.



EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 164 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 164
Topic #: 1
[All Certified Data Engineer Professional Questions]

All records from an Apache Kafka producer are being ingested into a single Delta Lake table with the following schema:

key BINARY, value BINARY, topic STRING, partition LONG, offset LONG, timestamp LONG

There are 5 unique topics being ingested. Only the "registration" topic contains Personal Identifiable Information (PII). The company wishes to restrict access to PII. The company also wishes to only retain records containing PII in this table for 14 days after initial ingestion. However, for non-PII information, it would like to retain these records indefinitely.

Which solution meets the requirements?

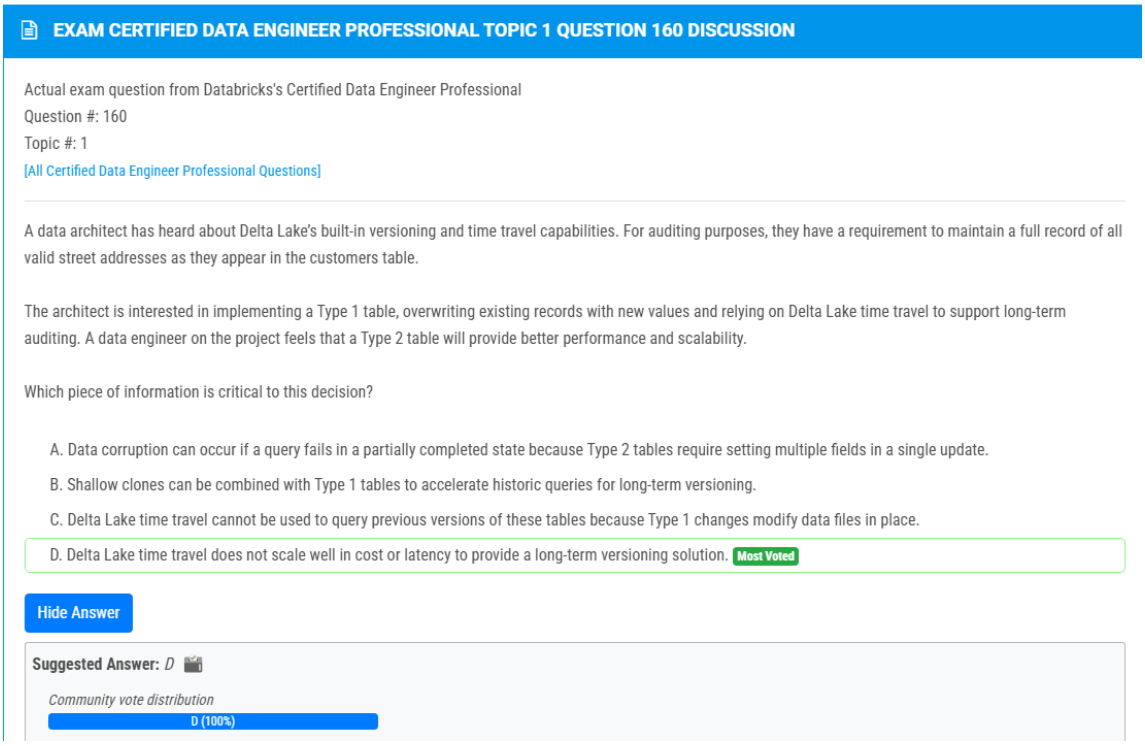
A. All data should be deleted biweekly; Delta Lake's time travel functionality should be leveraged to maintain a history of non-PII information.
B. Data should be partitioned by the registration field, allowing ACLs and delete statements to be set for the PII directory.
C. Data should be partitioned by the topic field, allowing ACLs and delete statements to leverage partition boundaries. Most Voted
D. Separate object storage containers should be specified based on the partition field, allowing isolation at the storage level.

Submit

hpkr 10 months, 2 weeks ago
Selected Answer: C
C is correct
upvoted 2 times

imatheushenrique 11 months ago
C.
Partitioning the data by the topic field allows the company to apply different access control policies and retention policies for different topics. Althought there is a performance optmization gain because of the read in the partition path.
upvoted 1 times

2.



EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 160 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 160
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data architect has heard about Delta Lake's built-in versioning and time travel capabilities. For auditing purposes, they have a requirement to maintain a full record of all valid street addresses as they appear in the customers table.

The architect is interested in implementing a Type 1 table, overwriting existing records with new values and relying on Delta Lake time travel to support long-term auditing. A data engineer on the project feels that a Type 2 table will provide better performance and scalability.

Which piece of information is critical to this decision?

A. Data corruption can occur if a query fails in a partially completed state because Type 2 tables require setting multiple fields in a single update.
B. Shallow clones can be combined with Type 1 tables to accelerate historic queries for long-term versioning.
C. Delta Lake time travel cannot be used to query previous versions of these tables because Type 1 changes modify data files in place.
D. Delta Lake time travel does not scale well in cost or latency to provide a long-term versioning solution. Most Voted

Hide Answer

Suggested Answer: D

Community vote distribution
D (100%)

3.

Question #: 152
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team maintains the following code:

```
accountDF = spark.table("accounts")
orderDF = spark.table("orders")
itemDF = spark.table("items")

orderWithItemDF = (orderDF.join(
    itemDF,
    orderDF.itemID == itemDF.itemID)
  .select(
    orderDF.accountID,
    orderDF.itemID,

    itemDF.itemName))

finalDF = (accountDF.join(
    orderWithItemDF,
    accountDF.accountID == orderWithItemDF.accountID)
  .select(
    orderWithItemDF["*"],

    accountDF.city))

(finalDF.write
  .mode("overwrite")
  .table("enriched_itemized_orders_by_account"))
```

Assuming that this code produces logically correct results and the data in the source tables has been de-duplicated and validated, which statement describes what will occur when this code is executed?

A. A batch job will update the enriched_itemized_orders_by_account table, replacing only those rows that have different values than the current version of the table, using accountID as the primary key.


```
    accountDF.accountID == orderWithItemDF.accountID)
  .select(
    orderWithItemDF["*"],

    accountDF.city))

(finalDF.write
  .mode("overwrite")
  .table("enriched_itemized_orders_by_account"))
```

Assuming that this code produces logically correct results and the data in the source tables has been de-duplicated and validated, which statement describes what will occur when this code is executed?

A. A batch job will update the enriched_itemized_orders_by_account table, replacing only those rows that have different values than the current version of the table, using accountID as the primary key.

B. The enriched_itemized_orders_by_account table will be overwritten using the current valid version of data in each of the three tables referenced in the join logic.

Most Voted

C. No computation will occur until enriched_itemized_orders_by_account is queried; upon query materialization, results will be calculated using the current valid version of data in each of the three tables referenced in the join logic.

D. An incremental job will detect if new rows have been written to any of the source tables; if new rows are detected, all results will be recalculated and used to overwrite the enriched_itemized_orders_by_account table.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

by hpkr at *June 12, 2024, 6:15 p.m.*

4.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 102 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 102
Topic #: 1
[All Certified Data Engineer Professional Questions]

What is the first line of a Databricks Python notebook when viewed in a text editor?

A. %python
B. // Databricks notebook source
C. # Databricks notebook source **Most Voted**
D. -- Databricks notebook source
E. # MAGIC %python

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

by 60ties at *Nov. 15, 2023, 5:19 p.m.*

Comments

Chosen Answer: ☐A ☐B ☐C ☐D ☐E

This is a voting comment (?). It is better to Upvote an existing comment if you don't have anything to add.

5.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 168 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 168
Topic #: 1
[All Certified Data Engineer Professional Questions]

What is the retention of job run history?

A. It is retained until you export or delete job run logs
B. It is retained for 30 days, during which time you can deliver job run logs to DBFS or S3
C. It is retained for 60 days, during which you can export notebook run results to HTML **Most Voted**
D. It is retained for 60 days, after which logs are archived

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

by imatheushenrique at *June 1, 2024, 3:33 a.m.*

6.

Actual exam question from Databricks's Certified Data Engineer Professional

Question #: 163

Topic #: 1

[All Certified Data Engineer Professional Questions]

A table named user_ltv is being used to create a view that will be used by data analysts on various teams. Users in the workspace are configured into groups, which are used for setting up data access using ACLs.

The user_ltv table has the following schema:

email STRING, age INT, ltv INT

The following view definition is executed:

```
CREATE VIEW user_ltv_no_minors AS
SELECT email, age, ltv
FROM user_ltv
WHERE
  CASE
    WHEN is_member("auditing") THEN TRUE
    ELSE age >= 18
  END
```

An analyst who is not a member of the auditing group executes the following query:

SELECT * FROM user_ltv_no_minors

Which statement describes the results returned by this query?

```
    CASE
      WHEN is_member("auditing") THEN TRUE
      ELSE age >= 18
    END
```

An analyst who is not a member of the auditing group executes the following query:

SELECT * FROM user_ltv_no_minors

Which statement describes the results returned by this query?

A. All columns will be displayed normally for those records that have an age greater than 17; records not meeting this condition will be omitted. **Most Voted**

B. All age values less than 18 will be returned as null values, all other columns will be returned with the values in user_ltv.

C. All values for the age column will be returned as null values, all other columns will be returned with the values in user_ltv.

D. All columns will be displayed normally for those records that have an age greater than 18; records not meeting this condition will be omitted.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

by MDWPartners at *May 29, 2024, 7:57 p.m.*

BrianNguyen95 10 months, 3 weeks ago

Selected Answer: A

Greater than 17

upvoted 2 times

Freyr 11 months ago

Selected Answer: A

Correct Answer: A
(>17) qual to (>=18). So, all records above 17 years will get in result and other records will be omitted.

upvoted 1 times

imatheushenrique 11 months ago

A. All columns will be displayed normally for those records that have an age greater than 17; records not meeting this condition will be omitted.

Because the condition of age>=18 only is respected in option A.

upvoted 1 times

MDWPartners 11 months ago

Selected Answer: A

Nope, A greater than 18 is 19. D is incorrect.

upvoted 1 times

7.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 139 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 139
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Structured Streaming job deployed to production has been resulting in higher than expected cloud storage costs. At present, during normal execution, each microbatch of data is processed in less than 3s; at least 12 times per minute, a microbatch is processed that contains 0 records. The streaming write was configured using the default trigger settings. The production job is currently scheduled alongside many other Databricks jobs in a workspace with instance pools provisioned to reduce start-up time for jobs with batch execution.

Holding all other variables constant and assuming records need to be processed in less than 10 minutes, which adjustment will meet the requirement?

A. Set the trigger interval to 3 seconds; the default trigger interval is consuming too many records per batch, resulting in spill to disk that can increase volume costs.

B. Use the trigger once option and configure a Databricks job to execute the query every 10 minutes; this approach minimizes costs for both compute and storage.

C. Set the trigger interval to 10 minutes; each batch calls APIs in the source storage account, so decreasing trigger frequency to maximum allowable threshold should minimize this cost. Most Voted

D. Set the trigger interval to 500 milliseconds; setting a small but non-zero trigger interval ensures that the source is not queried too frequently.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

Submit

lsio05 10 months, 2 weeks ago

Selected Answer: C

C,
A - incorrect explanation
B - trigger once is not correct option here
D - 500 miliseconds is already used, it's default trigger interval

upvoted 4 times

hpkr 10 months, 2 weeks ago

Selected Answer: C

Option C

upvoted 1 times

8.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 151 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 151
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior data engineer is working to implement logic for a Lakehouse table named silver_device_recordings. The source data contains 100 unique fields in a highly nested JSON structure.

The silver_device_recordings table will be used downstream to power several production monitoring dashboards and a production model. At present, 45 of the 100 fields are being used in at least one of these applications.

The data engineer is trying to determine the best approach for dealing with schema declaration given the highly-nested structure of the data and the numerous fields.

Which of the following accurately presents information about Delta Lake and Databricks that may impact their decision-making process?

A. The Tungsten encoding used by Databricks is optimized for storing string data; newly-added native support for querying JSON strings means that string types are always most efficient.

B. Because Delta Lake uses Parquet for data storage, data types can be easily evolved by just modifying file footer information in place.

C. Schema inference and evolution on Databricks ensure that inferred types will always accurately match the data types used by downstream systems.

D. Because Databricks will infer schema using types that allow all observed data to be processed, setting types manually provides greater assurance of data quality enforcement. Most Voted

Hide Answer

9.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 167 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 167
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team has been tasked with configuring connections to an external database that does not have a supported native connector with Databricks. The external database already has data security configured by group membership. These groups map directly to user groups already created in Databricks that represent various teams within the company.

A new login credential has been created for each group in the external database. The Databricks Utilities Secrets module will be used to make these credentials available to Databricks users.

Assuming that all the credentials are configured correctly on the external database and group membership is properly configured on Databricks, which statement describes how teams can be granted the minimum necessary access to using these credentials?

A. No additional configuration is necessary as long as all users are configured as administrators in the workspace where secrets have been added.

B. "Read" permissions should be set on a secret key mapped to those credentials that will be used by a given team.

C. "Read" permissions should be set on a secret scope containing only those credentials that will be used by a given team. Most Voted

D. "Manage" permissions should be set on a secret scope containing only those credentials that will be used by a given team.

Hide Answer

Suggested Answer: *C*

Community vote distribution

hpkr 10 months, 2 weeks ago

Selected Answer: C

C is correct. Read permission on secret scope should work here.

upvoted 3 times

Freyr 11 months ago

Selected Answer: C

Correct Answer: C
This option is the best practice for managing access to sensitive data. By creating a secret scope dedicated to each team and setting "Read" permissions on the scope, you ensure that only the intended team members can access their respective credentials. This method aligns with security best practices by tightly controlling access based on group membership and reducing the risk of unauthorized access.

upvoted 3 times

MDWPartners 11 months ago

Selected Answer: C

Seems C

upvoted 2 times

10.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 24 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 24
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which configuration parameter directly affects the size of a spark-partition upon ingestion of data into Spark?

A. spark.sql.files.maxPartitionBytes Most Voted
B. spark.sql.autoBroadcastJoinThreshold
C. spark.sql.files.openCostInBytes
D. spark.sql.adaptive.coalescePartitions.minPartitionNum
E. spark.sql.adaptive.advisoryPartitionSizeInBytes

Hide Answer

Suggested Answer: *A*

Community vote distribution
A (100%)

8605246 Highly Voted 1 year, 2 months ago

correct; The maximum number of bytes to pack into a single partition when reading files. This configuration is effective only when using file-based sources such as Parquet, JSON and ORC.
https://spark.apache.org/docs/latest/sql-performance-tuning.html

upvoted 5 times

Jay_98_11 Most Recent 9 months, 2 weeks ago

Selected Answer: A

correct

upvoted 3 times

sturcu 1 year ago

Selected Answer: A

from the provided list, this fits best.
In reality partition size/number can be influenced my many settings

upvoted 1 times

11.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 44 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 44
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data governance team is reviewing code used for deleting records for compliance with GDPR. They note the following logic is used to delete records from the Delta Lake table named users.

```
DELETE FROM users
WHERE user_id IN
  (SELECT user_id FROM delete_requests)
```

Assuming that user_id is a unique identifying key and that delete_requests contains all users that have requested deletion, which statement describes whether successfully executing the above logic guarantees that the records to be deleted are no longer accessible and why?

A. Yes; Delta Lake ACID guarantees provide assurance that the DELETE command succeeded fully and permanently purged these records.

B. No; the Delta cache may return records from previous versions of the table until the cluster is restarted.

C. Yes; the Delta cache immediately updates to reflect the latest data files recorded to disk.

D. No; the Delta Lake DELETE command only provides ACID guarantees when combined with the MERGE INTO command.

E. No; files containing deleted records may still be accessible with time travel until a VACUUM command is used to remove invalidated data files. Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (100%)

12.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 105 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 105
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data science team has created and logged a production model using MLflow. The model accepts a list of column names and returns a new column of type DOUBLE.

The following code correctly imports the production model, loads the customers table containing the customer_id key column into a DataFrame, and defines the feature columns needed for the model.

```
model = mlflow.pyfunc.spark_udf(spark, model_uri="models:/churn/prod")
df = spark.table("customers")
columns = ["account_age", "time_since_last_seen", "app_rating"]
```

Which code block will output a DataFrame with the schema "customer_id LONG, predictions DOUBLE"?

A. df.map(lambda x:model(x[columns])).select("customer_id, predictions")
B. df.select("customer_id", model(*columns).alias("predictions")) Most Voted
C. model.predict(df, columns)
D. df.select("customer_id", pandas_udf(model, columns).alias("predictions"))
E. df.apply(model, columns).select("customer_id, predictions")

Hide Answer

aragorn_brego Highly Voted 11 months, 1 week ago
Selected Answer: B
This code block applies the Spark UDF created from the MLflow model to the DataFrame df by selecting the existing customer_id column and the new column produced by the model, which is aliased to predictions. The model(*columns) part is where the UDF is applied to the columns specified in the columns list, and alias("predictions") is used to name the output column of the model's predictions. This will result in a DataFrame with the desired schema: "customer_id LONG, predictions DOUBLE".
upvoted 7 times

divingbell17 Highly Voted 10 months ago
Selected Answer: B
B is correct. It's a spark udf not pandas
upvoted 6 times

60ties Most Recent 11 months, 2 weeks ago
I think it is B
upvoted 2 times

13.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 162 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 162
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior data engineer has manually configured a series of jobs using the Databricks Jobs UI. Upon reviewing their work, the engineer realizes that they are listed as the "Owner" for each job. They attempt to transfer "Owner" privileges to the "DevOps" group, but cannot successfully accomplish this task.

Which statement explains what is preventing this privilege transfer?

A. Databricks jobs must have exactly one owner; "Owner" privileges cannot be assigned to a group. Most Voted
B. The creator of a Databricks job will always have "Owner" privileges; this configuration cannot be changed.
C. Only workspace administrators can grant "Owner" privileges to a group.
D. A user can only transfer job ownership to a group if they are also a member of that group.

Hide Answer

Suggested Answer: *A*

Community vote distribution
A (100%)

by imatheushenrique at *June 1, 2024, 4:03 a.m.*

03355a2 Highly Voted 10 months ago
Selected Answer: A
This is the correct answer for this question in a past Databricks version, however now you can indeed add a group as a owner to a job.
upvoted 5 times

imatheushenrique Most Recent 11 months ago
A. Databricks jobs must have exactly one owner; "Owner" privileges cannot be assigned to a group.
It's only possivel that a databricks JOB has an owner, not a group.
upvoted 1 times

14.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 158 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 158
Topic #: 1
[All Certified Data Engineer Professional Questions]

A CHECK constraint has been successfully added to the Delta table named activity_details using the following logic:

```
ALTER TABLE activity_details
ADD CONSTRAINT valid_coordinates
CHECK (
  latitude >= -90 AND
  latitude <= 90 AND
  longitude >= -180 AND
  longitude <= 180);
```

A batch job is attempting to insert new records to the table, including a record where latitude = 45.50 and longitude = 212.67.

Which statement describes the outcome of this batch insert?

A. The write will insert all records except those that violate the table constraints; the violating records will be reported in a warning log.
B. The write will fail completely because of the constraint violation and no records will be inserted into the target table. Most Voted
C. The write will insert all records except those that violate the table constraints; the violating records will be recorded to a quarantine table.
D. The write will include all records in the target table; any violations will be indicated in the boolean column named valid_coordinates.

Hide Answer

Suggested Answer: *B*

15.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 141 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 141
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement characterizes the general programming model used by Spark Structured Streaming?

A. Structured Streaming leverages the parallel processing of GPUs to achieve highly parallel data throughput.

B. Structured Streaming is implemented as a messaging bus and is derived from Apache Kafka.

C. Structured Streaming relies on a distributed network of nodes that hold incremental state values for cached stages.

D. Structured Streaming models new data arriving in a data stream as new rows appended to an unbounded table. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

by vexor3 at *July 20, 2024, 10:46 a.m.*

16.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 129 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 129
Topic #: 1
[All Certified Data Engineer Professional Questions]

Assuming that the Databricks CLI has been installed and configured correctly, which Databricks CLI command can be used to upload a custom Python Wheel to object storage mounted with the DBFS for use with a production job?

A. configure

B. fs Most Voted

C. workspace

D. libraries

Hide Answer

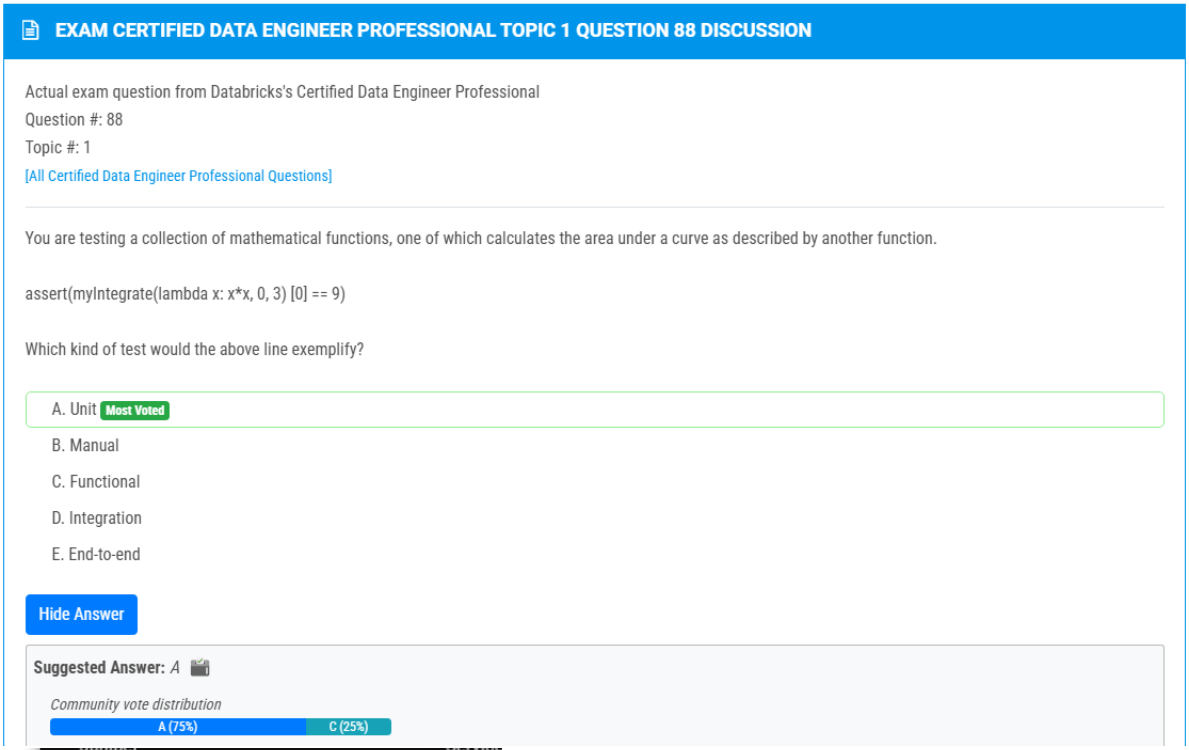
Suggested Answer: *B*

Community vote distribution

B (100%)

by vexor3 at *July 20, 2024, 9:31 a.m.*

17.



upvoted 3 times

barnac1es 1 year, 1 month ago

Selected Answer: C

I think it should be Functional Test

upvoted 3 times

HelixAbdu 9 months ago

There are 3 testing types:
Unit testing
Integration testing
And end to end testing

upvoted 5 times

vctrhugo 1 year, 2 months ago

Selected Answer: A

A. Unit

upvoted 3 times

divingbell17 1 year, 3 months ago

Selected Answer: A

A is correct

upvoted 3 times

18.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 125
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data science team has created and logged a production model using MLflow. The model accepts a list of column names and returns a new column of type DOUBLE.

The following code correctly imports the production model, loads the customers table containing the customer_id key column into a DataFrame, and defines the feature columns needed for the model.

```
model = mlflow.pyfunc.spark_udf (spark,
model_uri="models:/churn/prod")

df = spark.table("customers")

columns = ["account_age", "time_since_last_seen", "app_rating"]
```

Which code block will output a DataFrame with the schema "customer_id LONG, predictions DOUBLE"?

A. df.map(lambda x:model(x[columns])).select("customer_id, predictions")

B. df.select("customer_id",
model(*columns).alias("predictions")) Most Voted

C. model.predict(df, columns)

D. df.apply(model, columns).select("customer_id, predictions")

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

Freyr 11 months ago

Selected Answer: B

Correct Answer: B
This option uses select to specify columns from the DataFrame and applies the model to the specified columns (columns). The output of the model is aliased as "predictions", which ensures the output DataFrame will have the column names "customer_id" and "predictions" with appropriate data types assuming the model returns a double type. This syntax aligns with PySpark's DataFrame transformations and is a typical way to apply a machine learning model to specific columns in Databricks.

upvoted 3 times

19.

Question #: 124
Topic #: 1
[All Certified Data Engineer Professional Questions]

The security team is exploring whether or not the Databricks secrets module can be leveraged for connecting to an external database.

After testing the code with all Python variables being defined with strings, they upload the password to the secrets module and configure the correct permissions for the currently active user. They then modify their code to the following (leaving all other variables unchanged).

```
password = dbutils.secrets.get(scope="db_creds", key="jdbc_password")

print(password)

df = (spark
  .read
  .format("jdbc")
  .option("url", connection)
  .option("dbtable", tablename)
  .option("user", username)
  .option("password", password)
  )
```

Which statement describes what will happen when the above code is executed?

A. The connection to the external table will succeed; the string "REDACTED" will be printed. Most Voted

B. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the encoded password will be saved to DBFS.

C. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the password will be printed in plain text.

D. The connection to the external table will succeed; the string value of password will be printed in plain text.

Hide Answer

Suggested Answer: A upvoted 3 times

Deb9753 10 months, 3 weeks ago

Answer A : When using Databricks secrets, the actual value of the secret is typically protected from being displayed in plain text. Databricks automatically redacts secret values when they are printed in the notebook. So, when you use the print(password) statement, the output will not show the actual password but will instead show [REDACTED].

upvoted 1 times

imatheushenrique 10 months, 3 weeks ago

A. A. The connection to the external table will succeed; the string "REDACTED" will be printed.

upvoted 1 times

20.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 80 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 80
Topic #: 1
[All Certified Data Engineer Professional Questions]

The marketing team is looking to share data in an aggregate table with the sales organization, but the field names used by the teams do not match, and a number of marketing-specific fields have not been approved for the sales org.

Which of the following solutions addresses the situation while emphasizing simplicity?

A. Create a view on the marketing table selecting only those fields approved for the sales team; alias the names of any fields that should be standardized to the sales naming conventions. **Most Voted**

B. Create a new table with the required schema and use Delta Lake's DEEP CLONE functionality to sync up changes committed to one table to the corresponding table.

C. Use a CTAS statement to create a derivative table from the marketing table; configure a production job to propagate changes.

D. Add a parallel table write to the current production pipeline, updating a new sales table that varies as required from the marketing table.

E. Instruct the marketing team to download results as a CSV and email them to the sales organization.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

vctrhugo 1 year, 2 months ago

Selected Answer: A

Creating a view is a simple and efficient way to provide access to a subset of data from a table. In this case, the view can be configured to include only the fields that have been approved for the sales team. Additionally, any fields that need to be renamed to match the sales team's naming conventions can be aliased in the view. This approach does not require the creation of additional tables or the configuration of jobs to sync data, making it a relatively straightforward solution. However, it's important to note that views do not physically store data, so any changes to the underlying marketing table will be reflected in the view. This means that the sales team will always have access to the most up-to-date approved data.

upvoted 4 times

21.

Question #: 104
Topic #: 1
[All Certified Data Engineer Professional Questions]

The Databricks CLI is used to trigger a run of an existing job by passing the job_id parameter. The response that the job run request has been submitted successfully includes a field run_id.

Which statement describes what the number alongside this field represents?

A. The job_id and number of times the job has been run are concatenated and returned.

B. The total number of jobs that have been run in the workspace.

C. The number of times the job definition has been run in this workspace.

D. The job_id is returned in this field.

E. The globally unique ID of the newly triggered run. **Most Voted**

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (100%)

vctrhugo 8 months, 3 weeks ago

Selected Answer: E

The number alongside the "run_id" field represents the globally unique identifier assigned to the newly triggered run of the job. Each run of a job in Databricks is assigned a unique run_id, allowing you to track and reference that specific execution of the job.

upvoted 4 times

Def21 9 months, 1 week ago

Selected Answer: E

Verified from Databricks UI

upvoted 1 times

22.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 107 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional

Question #: 107

Topic #: 1

[All Certified Data Engineer Professional Questions]

Which statement describes Delta Lake optimized writes?

A. Before a Jobs cluster terminates, OPTIMIZE is executed on all tables modified during the most recent job.

B. An asynchronous job runs after the write completes to detect if files could be further compacted; if yes, an OPTIMIZE job is executed toward a default of 1 GB.

C. Data is queued in a messaging bus instead of committing data directly to memory; all data is committed from the messaging bus in one batch once the job is complete.

D. Optimized writes use logical partitions instead of directory partitions; because partition boundaries are only represented in metadata, fewer small files are written.

E. A shuffle occurs prior to writing to try to group similar data together resulting in fewer files instead of each executor writing multiple files based on directory partitions. Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (100%)

vctrhugo 8 months, 3 weeks ago

Selected Answer: E

Optimized writes improve file size as data is written and benefit subsequent reads on the table.

Optimized writes are most effective for partitioned tables, as they reduce the number of small files written to each partition. Writing fewer large files is more efficient than writing many small files, but you might still see an increase in write latency because data is shuffled before being written.

https://learn.microsoft.com/en-us/azure/databricks/delta/tune-file-size#--optimized-writes-for-delta-lake-on-azure-databricks

upvoted 1 times

lexaneon 9 months, 3 weeks ago

Selected Answer: E

https://docs.databricks.com/en/delta/tune-file-size.html#optimized-writes

upvoted 3 times

alexvno 10 months, 1 week ago

Selected Answer: E

Optimized writes are most effective for partitioned tables, as they reduce the number of small files written to each partition. Writing fewer large files is more efficient than writing many small files, but you might still see an increase in write latency because data is shuffled before being writte

upvoted 3 times

23.

Question #: 81

Topic #: 1

[All Certified Data Engineer Professional Questions]

A CHECK constraint has been successfully added to the Delta table named activity_details using the following logic:

```
ALTER TABLE activity_details
ADD CONSTRAINT valid_coordinates
CHECK (
  latitude >= -90 AND
  latitude <= 90 AND
  longitude >= -180 AND
  longitude <= 180);
```

A batch job is attempting to insert new records to the table, including a record where latitude = 45.50 and longitude = 212.67.

Which statement describes the outcome of this batch insert?

A. The write will fail when the violating record is reached; any records previously processed will be recorded to the target table.

B. The write will fail completely because of the constraint violation and no records will be inserted into the target table. Most Voted

C. The write will insert all records except those that violate the table constraints; the violating records will be recorded to a quarantine table.

D. The write will include all records in the target table; any violations will be indicated in the boolean column named valid_coordinates.

E. The write will insert all records except those that violate the table constraints; the violating records will be reported in a warning log.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

aragorn_brego Highly Voted 11 months, 1 week ago

Selected Answer: B

In systems that support atomic transactions, such as Delta Lake, when a batch operation encounters a record that violates a CHECK constraint, the entire operation fails, and no records are inserted, including those that do not violate the constraint. This is to ensure the atomicity of the transaction, meaning that either all the changes are committed, or none are, maintaining data integrity. The record with a longitude of 212.67 violates the constraint because longitude values must be between -180 and 180 degrees.

upvoted 5 times

vctrhugo Highly Voted 8 months, 3 weeks ago

Selected Answer: B

In Delta Lake, when a batch job attempts to insert records into a table that has a CHECK constraint, if any record violates the constraint, the entire write operation fails. This is because Delta Lake enforces strong transactional guarantees, which means that either all changes in a transaction are saved, or none are.

upvoted 5 times

24.

Question #: 86
Topic #: 1
[All Certified Data Engineer Professional Questions]

When evaluating the Ganglia Metrics for a given cluster with 3 executor nodes, which indicator would signal proper utilization of the VM's resources?

A. The five Minute Load Average remains consistent/flat

B. Bytes Received never exceeds 80 million bytes per second

C. Network I/O never spikes

D. Total Disk Space remains constant

E. CPU Utilization is around 75% Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (100%)

sturcu Highly Voted 11 months, 4 weeks ago

Selected Answer: E

I would look at max CPU utilization and max Memory usage.
Having 75% CPU usage would signify we have a proper utilization of cpu resources

upvoted 7 times

vctrhugo Most Recent 8 months, 3 weeks ago

Selected Answer: E

Proper utilization of VM resources, especially in a distributed computing environment like Spark, often involves efficient usage of CPU resources. A CPU utilization around 75% indicates that the CPU is being utilized without being fully saturated, allowing room for additional processing without causing excessive contention.

upvoted 2 times

alexvno 10 months, 1 week ago

Selected Answer: E

75% good

upvoted 1 times

aragorn_brego 11 months, 1 week ago

Selected Answer: E

An average CPU utilization around 75% is a good indicator of proper utilization of the VM's resources in a distributed computing environment. It suggests that the CPUs are being actively used for computation without being maxed out, which could indicate a bottleneck. It leaves some headroom to handle additional load without causing excessive queuing or delays.

upvoted 3 times

25.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 87
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which of the following technologies can be used to identify key areas of text when parsing Spark Driver log4j output?

A. Regex Most Voted

B. Julia

C. pyspsark.ml.feature

D. Scala Datasets

E. C++

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (89%) 11%

vctrhugo 8 months, 3 weeks ago

Selected Answer: A

It allows us to define patterns that match the structure of the log entries and capture relevant data.

upvoted 3 times

aragorn_brego 11 months, 1 week ago

Selected Answer: A

Regular expressions (regex) can be used to identify and extract patterns from text data, which makes them very useful for parsing log files like the Spark Driver's log4j output. By defining specific regex patterns, you can search for error messages, timestamps, specific log levels, or any other text that follows a particular format within the log files.

upvoted 4 times

sturcu 12 months ago

Selected Answer: A

Regex to extract text

upvoted 3 times

sturcu 12 months ago

Selected Answer: E

Regex to extract text. C++ makes no sense in this context

26.

Question #: 96

Topic #: 1

[All Certified Data Engineer Professional Questions]

A junior data engineer is migrating a workload from a relational database system to the Databricks Lakehouse. The source system uses a star schema, leveraging foreign key constraints and multi-table inserts to validate records on write.

Which consideration will impact the decisions made by the engineer while migrating this workload?

A. Databricks only allows foreign key constraints on hashed identifiers, which avoid collisions in highly-parallel writes.

B. Databricks supports Spark SQL and JDBC; all logic can be directly migrated from the source system without refactoring.

C. Committing to multiple tables simultaneously requires taking out multiple table locks and can lead to a state of deadlock.

D. All Delta Lake transactions are ACID compliant against a single table, and Databricks does not enforce foreign key constraints. Most Voted

E. Foreign keys must reference a primary key field; multi-table inserts must leverage Delta Lake's upsert functionality.

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

vctrhugo Highly Voted 8 months, 3 weeks ago

Selected Answer: D

In Databricks Delta Lake, transactions are ACID compliant at the table level, meaning that transactions apply to a single table. However, Delta Lake does not enforce foreign key constraints across tables. Therefore, the data engineer needs to be aware that Databricks does not automatically enforce referential integrity between tables through foreign key constraints, and it becomes the responsibility of the data engineer to manage these relationships appropriately.

upvoted 6 times

alexvno Most Recent 10 months, 1 week ago

Selected Answer: D

Primary and foreign keys are informational only and are not enforced.

upvoted 2 times

60ties 11 months, 2 weeks ago

Selected Answer: D

D makes more sense.
Since there are no database-level transactions, locks, or guarantees, and since primary key & foreign key constraints are informational only, there is no guarantee of enforced relations (the start schema) in place will remain in place after migration. This means B cannot be right.

upvoted 1 times

27.

Question #: 71

Topic #: 1

[All Certified Data Engineer Professional Questions]

A junior data engineer has been asked to develop a streaming data pipeline with a grouped aggregation using DataFrame df. The pipeline needs to calculate the average humidity and average temperature for each non-overlapping five-minute interval. Incremental state information should be maintained for 10 minutes for late-arriving data.

Streaming DataFrame df has the following schema:

"device_id INT, event_time TIMESTAMP, temp FLOAT, humidity FLOAT"

Code block:

```
df._____
    .groupBy(
        window("event_time", "5 minutes").alias("time"),
        "device_id"
    )
    .agg(
        avg("temp").alias("avg_temp"),
        avg("humidity").alias("avg_humidity")
    )
    .writeStream
    .format("delta")
    .saveAsTable("sensor_avg")
```

Choose the response that correctly fills in the blank within the code block to complete this task.

A. withWatermark("event_time", "10 minutes")

B. awaitArrival("event_time", "10 minutes")

```
)
.agg(
    avg("temp").alias("avg_temp"),
    avg("humidity").alias("avg_humidity")
)
.writeStream
.format("delta")
.saveAsTable("sensor_avg")
```

Choose the response that correctly fills in the blank within the code block to complete this task.

A. withWatermark("event_time", "10 minutes") **Most Voted**

B. awaitArrival("event_time", "10 minutes")

C. await("event_time + '10 minutes'")

D. slidingWindow("event_time", "10 minutes")

E. delayWrite("event_time", "10 minutes")

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

aragorn_brego Highly Voted 1 year, 5 months ago

Selected Answer: A

To handle late-arriving data in a streaming aggregation, you need to specify a watermark, which tells the streaming query how long to wait for late data. The withWatermark method is used for this purpose in Spark Structured Streaming. It defines the threshold for how late the data can be relative to the latest data that has been seen in the same window.

upvoted 9 times

sturcu Highly Voted 1 year, 6 months ago

Selected Answer: A

withWatermark.
There sliding window is doe through the window function

upvoted 9 times

71dfab9 Most Recent 8 months, 2 weeks ago

Selected Answer: A

The withWatermark method is used in streaming DataFrames when processing real-time data streams. This method helps in managing stateful operations, such as aggregations, by specifying a time column to use for watermarking. Watermarking is a mechanism to handle late data (data that arrives later than expected) by defining a threshold time window beyond which late data is considered too late to be included in aggregations.

The slidingWindow function mentioned in D is not a standard function in Databricks or Apache Spark.

upvoted 1 times

Dileepvikram 1 year, 5 months ago

Answer is A

upvoted 3 times

28.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 175 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 175
Topic #: 1
[All Certified Data Engineer Professional Questions]

What is the first line of a Databricks Python notebook when viewed in a text editor?

A. %python

B. // Databricks notebook source

C. # Databricks notebook source **Most Voted**

D. -- Databricks notebook source

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

minhhnh 8 months ago

Selected Answer: C

The correct answer is:

C. # Databricks notebook source

This is the comment line that appears at the beginning of a Databricks Python notebook when viewed in a text editor.

upvoted 2 times

imatheushenrique 11 months ago

C. # Databricks notebook source
The commentary in the first like will indicate a magic command for a notebook source.

upvoted 2 times

29.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 85 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 85
Topic #: 1
[All Certified Data Engineer Professional Questions]

A distributed team of data analysts share computing resources on an interactive cluster with autoscaling configured. In order to better manage costs and query throughput, the workspace administrator is hoping to evaluate whether cluster upscaling is caused by many concurrent users or resource-intensive queries.

In which location can one review the timeline for cluster resizing events?

A. Workspace audit logs
B. Driver's log file
C. Ganglia
D. Cluster Event Log Most Voted
E. Executor's log file

Hide Answer

Suggested Answer: *D*

Curious76 8 months ago
Selected Answer: D
The Cluster Event Log provides detailed information about various events affecting the cluster throughout its lifecycle, including cluster creation, restarts, termination, and resizing events. It displays the timestamp, event type (e.g., "CLUSTER_RESIZED"), and relevant details for each event, allowing the administrator to review the timeline for cluster scaling behavior and identify potential patterns related to user activity or resource-intensive queries.
upvoted 3 times

vctrhugo 8 months, 3 weeks ago
Selected Answer: D
The timeline for cluster resizing events can be reviewed in the Cluster Event Log. This log provides information about cluster scaling events, including when the cluster is scaled up or down. You can access this information to understand the reasons behind autoscaling events and whether they are triggered by many concurrent users or resource-intensive queries.
upvoted 1 times

alexvno 10 months, 1 week ago
Selected Answer: D
Cluster event log
upvoted 2 times

aragorn_brego 11 months, 1 week ago
Selected Answer: D
The Cluster Event Log in Databricks will show the timeline for cluster resizing events, including details about when and why a cluster was resized (scaled up or down). This log would help the workspace administrator determine the causes of cluster scaling, whether due to many concurrent users submitting jobs or a few users running resource-intensive queries.

less suitable:
C. Ganglia provides metrics on system-level performance, such as CPU and memory usage, but does not log specific cluster scaling events.
upvoted 2 times

30.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 65 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 65
Topic #: 1
[All Certified Data Engineer Professional Questions]

Two of the most common data locations on Databricks are the DBFS root storage and external object storage mounted with dbutils.fs.mount().

Which of the following statements is correct?

A. DBFS is a file system protocol that allows users to interact with files stored in object storage using syntax and guarantees similar to Unix file systems. Most Voted

B. By default, both the DBFS root and mounted data sources are only accessible to workspace administrators.

C. The DBFS root is the most secure location to store data, because mounted storage volumes must have full public read and write permissions.

D. Neither the DBFS root nor mounted storage can be accessed when using %sh in a Databricks notebook.

E. The DBFS root stores files in ephemeral block volumes attached to the driver, while mounted directories will always persist saved data to external storage between sessions.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

Curious76 8 months ago

Selected Answer: A

A is correct . For E, This statement is partially incorrect. The DBFS root does use ephemeral storage, but not block volumes. Data saved there is lost when the cluster terminates unless explicitly persisted elsewhere. Mounted storage, however, can persist data between sessions depending on the underlying storage service and configuration.

upvoted 4 times

sodere 10 months, 1 week ago

Selected Answer: A

DBFS is a layer on top of cloud storage providers.

upvoted 2 times

aragorn_brego 11 months, 1 week ago

Selected Answer: A

Databricks File System (DBFS) is a layer over a cloud object storage (like AWS S3, Azure Blob Storage, or GCP Cloud Storage) that allows users to interact with data as if they were using a traditional file system. It provides familiar file system semantics and is designed to be consistent with POSIX-like file system behavior, which includes commands and actions similar to those used in Unix and Linux file systems.

upvoted 3 times

31.

Question #: 122

Topic #: 1

[All Certified Data Engineer Professional Questions]

A Delta Lake table was created with the below query:

```
CREATE TABLE prod.sales_by_store
AS (
  SELECT *
  FROM prod.sales a
  INNER JOIN prod.store b
  ON a.store_id = b.store_id
)
```

Consider the following query:

DROP TABLE prod.sales_by_store -

If this statement is executed by a workspace admin, which result will occur?

A. Data will be marked as deleted but still recoverable with Time Travel.

B. The table will be removed from the catalog but the data will remain in storage.

C. The table will be removed from the catalog and the data will be deleted. **Most Voted**

D. An error will occur because Delta Lake prevents the deletion of production data.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

Freyr 11 months ago

Selected Answer: C

Correct Answer: C

No location provided in the table. So, it is a managed table. This will result in deleting the table meta data as well as table data.

upvoted 1 times

32.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 42
Topic #: 1
[All Certified Data Engineer Professional Questions]

A table named user_ltv is being used to create a view that will be used by data analysts on various teams. Users in the workspace are configured into groups, which are used for setting up data access using ACLs.
The user_ltv table has the following schema:
email STRING, age INT, ltv INT
The following view definition is executed:

```
CREATE VIEW email_ltv AS
SELECT
CASE WHEN
  is_member('marketing') THEN email
  ELSE 'REDACTED'
END AS email,
ltv
FROM user_ltv
```

An analyst who is not a member of the marketing group executes the following query:

SELECT * FROM email_ltv -
Which statement describes the results returned by this query?

A. Three columns will be returned, but one column will be named "REDACTED" and contain only null values.

B. Only the email and ltv columns will be returned; the email column will contain all null values.

used for setting up data access using ACLs.
The user_ltv table has the following schema:
email STRING, age INT, ltv INT
The following view definition is executed:

```
CREATE VIEW email_ltv AS
SELECT
CASE WHEN
  is_member('marketing') THEN email
  ELSE 'REDACTED'
END AS email,
ltv
FROM user_ltv
```

An analyst who is not a member of the marketing group executes the following query:

SELECT * FROM email_ltv -
Which statement describes the results returned by this query?

A. Three columns will be returned, but one column will be named "REDACTED" and contain only null values.

B. Only the email and ltv columns will be returned; the email column will contain all null values.

C. The email and ltv columns will be returned with the values in user_ltv.

D. The email.age, and ltv columns will be returned with the values in user_ltv.

E. Only the email and ltv columns will be returned; the email column will contain the string "REDACTED" in each row. Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (100%)

Submit

AndreFR 8 months, 1 week ago

Selected Answer: E

A, D incorrect because 2 columns email & ltv are returned.
B incorrect because email will not always contain null values (unless email is null)
The user is not a member of "marketing", so 3 is the correct answer. If the user were a member of "marketing" group, correct answer would have been C

upvoted 2 times

Isio05 10 months, 3 weeks ago

Selected Answer: E

33.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 38
Topic #: 1
[All Certified Data Engineer Professional Questions]

The downstream consumers of a Delta Lake table have been complaining about data quality issues impacting performance in their applications. Specifically, they have complained that invalid latitude and longitude values in the activity_details table have been breaking their ability to use other geolocation processes.

A junior engineer has written the following code to add CHECK constraints to the Delta Lake table:

```
ALTER TABLE activity_details
ADD CONSTRAINT valid_coordinates
CHECK (
  latitude >= -90 AND
  latitude <= 90 AND
  longitude >= -180 AND
  longitude <= 180);
```

A senior engineer has confirmed the above logic is correct and the valid ranges for latitude and longitude are provided, but the code fails when executed.

Which statement explains the cause of this failure?

A. Because another team uses this table to support a frequently running application, two-phase locking is preventing the operation from committing.

B. The activity_details table already exists; CHECK constraints can only be added during initial table creation.

C. The activity_details table already contains records that violate the constraints; all existing data must pass CHECK constraints in order to add them to an existing table. Most Voted

D. The activity_details table already contains records; CHECK constraints can only be added prior to inserting values into a table.

E. The current table schema does not contain the field valid_coordinates; schema evolution will need to be enabled before altering the table to add a constraint.

Hide Answer

8605246 Highly Voted 1 year, 8 months ago

incorrect the correct option is C, with constraints, if added to an existing table the existing data in the table must be consistent with the constraint otherwise it fails
https://docs.databricks.com/en/sql/language-manual/sql-ref-syntax-ddl-alter-table.html#add-constraint

upvoted 13 times

AndreFR Most Recent 8 months, 1 week ago

Selected Answer: C

```
-- CREATE TABLE
create table test_constraint (t1 varchar(2), n1 int);

-- ADD VALUE
insert into test_constraint values ('v3', 3);

-- ADD CONSTAINT VIOLATED BY CURRENT DATA
-- should throw error : 1 row in spark_catalog.default.test_constaint violate the new CHECK constraint (n1 < 3)

alter table test_constraint add constraint valid_n1 check (n1 < 3);

-- ADD CONSTAINT NOT VIOLATED BY CURRENT DATA (no error)
alter table test_constraint add constraint valid_n1 check (n1 < 100);
```

upvoted 1 times

34.

Question #: 66

Topic #: 1

[All Certified Data Engineer Professional Questions]

The following code has been migrated to a Databricks notebook from a legacy workload:

```
%sh
git clone https://github.com/foo/data_loader;
python ./data_loader/run.py;
mv ./output /dbfs/mnt/new_data
```

The code executes successfully and provides the logically correct results, however, it takes over 20 minutes to extract and load around 1 GB of data.

Which statement is a possible explanation for this behavior?

A. %sh triggers a cluster restart to collect and install Git. Most of the latency is related to cluster startup time.

B. Instead of cloning, the code should use %sh pip install so that the Python code can get executed in parallel across all nodes in a cluster.

C. %sh does not distribute file moving operations; the final line of code should be updated to use %fs instead.

D. Python will always execute slower than Scala on Databricks. The run.py script should be refactored to Scala.

E. %sh executes shell code on the driver node. The code does not take advantage of the worker nodes or Databricks optimized Spark. Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (100%)

aragorn_brego Highly Voted 1 year, 5 months ago

Selected Answer: E

When using %sh in a Databricks notebook, the commands are executed in a shell environment on the driver node. This means that only the resources of the driver node are used, and the execution does not leverage the distributed computing capabilities of the worker nodes in the Spark cluster. This can result in slower performance, especially for data-intensive tasks, compared to an approach that distributes the workload across all nodes in the cluster using Spark.

upvoted 9 times

robodog Most Recent 8 months, 1 week ago

Selected Answer: E

Option E correct

upvoted 1 times

Freyr 11 months ago

Selected Answer: E

Option E: Correct. The %sh magic command in Databricks runs shell commands on the driver node only. This means the operations within %sh do not leverage the distributed nature of the Databricks cluster. Consequently, the Git clone, Python script execution, and file move operations are all performed on a single node (the driver), which explains why it takes a long time to process and move 1 GB of data. This approach does not utilize the parallel processing capabilities of the worker nodes or the optimization features of Databricks Spark.

Option C: Incorrect. %sh does not inherently distribute any operations, but the issue here is broader than just file moving operations. Using %fs for file operations is a best practice, but it does not resolve the inefficiency of running all commands on the driver node.

upvoted 2 times

Dileepvikram 1 year, 5 months ago

E is the answer as the command is ran in the driver node and other nodes in the cluster are not used

upvoted 2 times

sturcu 1 year, 6 months ago

Selected Answer: E

%sh run Bash commands on the driver node of the cluster.
https://www.databricks.com/blog/2020/08/31/introducing-the-databricks-web-terminal.html

upvoted 3 times

35.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 56 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 56
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement describes integration testing?

A. Validates interactions between subsystems of your application **Most Voted**
B. Requires an automated testing framework
C. Requires manual intervention
D. Validates an application use case
E. Validates behavior of individual elements of your application

Hide Answer

Suggested Answer: *A*

Community vote distribution
A (100%)

Submit

robodog 8 months, 1 week ago
Selected Answer: A
Answer is A
upvoted 2 times

alexvno 1 year, 4 months ago
Selected Answer: A
Integration testing is a type of software testing where components of the software are gradually integrated and then tested as a unified group
upvoted 4 times

36.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 82 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 82
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior data engineer has manually configured a series of jobs using the Databricks Jobs UI. Upon reviewing their work, the engineer realizes that they are listed as the "Owner" for each job. They attempt to transfer "Owner" privileges to the "DevOps" group, but cannot successfully accomplish this task.

Which statement explains what is preventing this privilege transfer?

A. Databricks jobs must have exactly one owner; "Owner" privileges cannot be assigned to a group. Most Voted

B. The creator of a Databricks job will always have "Owner" privileges; this configuration cannot be changed.

C. Other than the default "admins" group, only individual users can be granted privileges on jobs.

D. A user can only transfer job ownership to a group if they are also a member of that group.

E. Only workspace administrators can grant "Owner" privileges to a group.

Hide Answer

Suggested Answer: *A*

hal2401me 7 months, 3 weeks ago

Selected Answer: A

did a test. "group cannot be owner" is displayed.

upvoted 3 times

vctrhugo 8 months, 3 weeks ago

Selected Answer: A

In Databricks, each job must have exactly one owner, which is typically the user who created the job. This "Owner" privilege allows the user to perform any action on the job, including modifying its settings or deleting it. However, this privilege cannot be assigned to a group. If you want to allow multiple users or a group of users to manage a job, you can use ACLs (Access Control Lists) to grant them the necessary permissions. But the "Owner" privilege will still remain with the individual user who created the job.

upvoted 1 times

sturcu 1 year ago

Selected Answer: A

Correct
A job cannot have more than one owner. A job cannot have a group as an owner

upvoted 4 times

37.

Question #: 78
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team maintains the following code:

```
import pyspark.sql.functions as F

(spark.table("silver_customer_sales")
  .groupBy("customer_id")
  .agg(
    F.min("sale_date").alias("first_transaction_date"),
    F.max("sale_date").alias("last_transaction_date"),
    F.mean("sale_total").alias("average_sales"),
    F.countDistinct("order_id").alias("total_orders"),
    F.sum("sale_total").alias("lifetime_value")
  ).write
  .mode("overwrite")
  .table("gold_customer_lifetime_sales_summary")
)
```

Assuming that this code produces logically correct results and the data in the source table has been de-duplicated and validated, which statement describes what will occur when this code is executed?

A. The silver_customer_sales table will be overwritten by aggregated values calculated from all records in the gold_customer_lifetime_sales_summary table as a batch job.

B. A batch job will update the gold_customer_lifetime_sales_summary table, replacing only those rows that have different values than the current version of the table, using customer_id as the primary key.

C. The gold_customer_lifetime_sales_summary table will be overwritten by aggregated values calculated from all records in the silver_customer_sales table as a batch job.

)

Assuming that this code produces logically correct results and the data in the source table has been de-duplicated and validated, which statement describes what will occur when this code is executed?

A. The silver_customer_sales table will be overwritten by aggregated values calculated from all records in the gold_customer_lifetime_sales_summary table as a batch job.

B. A batch job will update the gold_customer_lifetime_sales_summary table, replacing only those rows that have different values than the current version of the table, using customer_id as the primary key.

C. The gold_customer_lifetime_sales_summary table will be overwritten by aggregated values calculated from all records in the silver_customer_sales table as a batch job. Most Voted

D. An incremental job will leverage running information in the state store to update aggregate values in the gold_customer_lifetime_sales_summary table.

E. An incremental job will detect if new rows have been written to the silver_customer_sales table; if new rows are detected, all aggregates will be recalculated and used to overwrite the gold_customer_lifetime_sales_summary table.

Hide Answer

Suggested Answer: *C*

Community vote distribution

aragorn_brego Highly Voted 11 months, 1 week ago

Selected Answer: C

The code is performing a batch aggregation operation on the "silver_customer_sales" table grouped by "customer_id". It calculates the first and last transaction dates, the average sales, the total number of distinct orders, and the lifetime value of sales for each customer. The .mode("overwrite") operation specifies that the output table "gold_customer_lifetime_sales_summary" should be overwritten with the result of this aggregation. This means that every time this code runs, it will replace the existing "gold_customer_lifetime_sales_summary" table with a new version that reflects the current state of the "silver_customer_sales" table.

upvoted 7 times

hal2401me Most Recent 7 months, 3 weeks ago

38.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 123 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 123
Topic #: 1
[All Certified Data Engineer Professional Questions]

A developer has successfully configured their credentials for Databricks Repos and cloned a remote Git repository. They do not have privileges to make changes to the main branch, which is the only branch currently visible in their workspace.

Which approach allows this user to share their code updates without the risk of overwriting the work of their teammates?

A. Use Repos to create a new branch, commit all changes, and push changes to the remote Git repository. Most Voted

B. Use Repos to create a fork of the remote repository, commit all changes, and make a pull request on the source repository.

C. Use Repos to pull changes from the remote Git repository; commit and push changes to a branch that appeared as changes were pulled.

D. Use Repos to merge all differences and make a pull request back to the remote repository.

Hide Answer

Suggested Answer: *A*

Community vote distribution
A (100%)

by Ati1362 at *June 24, 2024, 1:03 p.m.*

39.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 7
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data science team has created and logged a production model using MLflow. The following code correctly imports and applies the production model to output the predictions as a new DataFrame named preds with the schema "customer_id LONG, predictions DOUBLE, date DATE".

```
from pyspark.sql.functions import current_date

model = mlflow.pyfunc.spark_udf(spark, model_uri="models:/churn/prod")
df = spark.table("customers")
columns = ["account_age", "time_since_last_seen", "app_rating"]
preds = (df.select(
  "customer_id",
  model(*columns).alias("predictions"),
  current_date().alias("date")
  )
```

The data science team would like predictions saved to a Delta Lake table with the ability to compare all predictions across time. Churn predictions will be made at most once per day.
Which code block accomplishes this task while minimizing potential compute costs?

A. preds.write.mode("append").saveAsTable("churn_preds")

B. preds.write.format("delta").save("/preds/churn_preds")

C.
```
(preds.writeStream
  .outputMode("overwrite")
  .option("checkpointPath", "/_checkpoints/churn_preds")
  .start("/preds/churn_preds")
)
```

D.
```
(preds.write
  .format("delta")
  .mode("overwrite")
  .saveAsTable("churn_preds")
)
```

E.
```
(preds.writeStream
  .outputMode("append")
```

once per day.

Which code block accomplishes this task while minimizing potential compute costs?

A. preds.write.mode("append").saveAsTable("churn_preds") **Most Voted**

B. preds.write.format("delta").save("/preds/churn_preds")

C.
```
(preds.writeStream
  .outputMode("overwrite")
  .option("checkpointPath", "/_checkpoints/churn_preds")
  .start("/preds/churn_preds")
)
```

D.
```
(preds.write
  .format("delta")
  .mode("overwrite")
  .saveAsTable("churn_preds")
)
```

E.
```
(preds.writeStream
  .outputMode("append")
  .option("checkpointPath", "/_checkpoints/churn_preds")
  .table("churn_preds")
)
```

Hide Solution | Discussion 13

Correct Answer: *A*

Community vote distribution

A (100%)

Starvosxant 1 year, 6 months ago

First: the default node Databricks saves tables IS Delta Format. So no reason why you say it wouldn't benefit from Lakehouse features.
Second: the default write mode is Error, means that if you try to write to a location and that already exists there, it will prone a Error. And the question specify that you gonna write once a day.
You better revisit basic topics before continue to the professional level certification, or buy the dump entirely.

upvoted 4 times

Eertyy 1 year, 7 months ago

Here's why:
A. saves the data as a managed table, which may not be efficient for large-scale data or frequent updates. It doesn't utilize Delta Lake capabilities.
C.is used for streaming operations, not batch processing. Also, using "overwrite" as output mode will replace the existing data each time, which is not suitable for keeping historical predictions.
D.is similar to option A but with "overwrite" mode. It will replace the entire table each time, which is not suitable for maintaining a historical record of predictions.

E. is also for streaming operations and not for batch processing. Additionally, it uses the "table" method, which is not typically used for writing batch data into Delta Lake tables.
Option B is suitable for batch processing, writes data in Delta Lake format, and allows you to efficiently maintain a historical record of predictions while minimizing compute costs.

upvoted 3 times

pradyumn9999 1 year, 6 months ago

Its also said they want to compare past values as well, so mode needs to be append. By default is error mode.

upvoted 4 times

40.

Question #6 *Topic 1*

The security team is exploring whether or not the Databricks secrets module can be leveraged for connecting to an external database.
After testing the code with all Python variables being defined with strings, they upload the password to the secrets module and configure the correct permissions for the currently active user. They then modify their code to the following (leaving all other variables unchanged).

```
password = dbutils.secrets.get(scope="db_creds", key="jdbc_password")

print(password)

df = (spark
  .read
  .format("jdbc")
  .option("url", connection)
  .option("dbtable", tablename)
  .option("user", username)
  .option("password", password)
  )
```

Which statement describes what will happen when the above code is executed?

A. The connection to the external table will fail; the string "REDACTED" will be printed.

B. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the encoded password will be saved to DBFS.

C. An interactive input box will appear in the notebook; if the right password is provided, the connection will succeed and the password will be printed in plain text.

D. The connection to the external table will succeed; the string value of password will be printed in plain text.

E. The connection to the external table will succeed; the string "REDACTED" will be printed. Most Voted

Hide Solution | Discussion 11

Correct Answer: *E*

Community vote distribution

E (100%)

akashdesarda 7 months ago

Selected Answer: E

Whatever we read using dbutls.secret module is always printed as '[REDACTED]', but when consumed in code, underlying vales are passed.

upvoted 4 times

imatheushenrique 10 months, 3 weeks ago

E. The connection to the external table will succeed; the string "REDACTED" will be printed.

upvoted 1 times

41.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 94 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 94
Topic #: 1
[All Certified Data Engineer Professional Questions]

Spill occurs as a result of executing various wide transformations. However, diagnosing spill requires one to proactively look for key indicators.

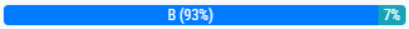
Where in the Spark UI are two of the primary indicators that a partition is spilling to disk?

A. Query's detail screen and Job's detail screen
B. Stage's detail screen and Executor's log files **Most Voted**
C. Driver's and Executor's log files
D. Executor's detail screen and Executor's log files
E. Stage's detail screen and Query's detail screen

Hide Answer

Suggested Answer: *B*

Community vote distribution



Selected Answer: B

In the Spark UI, the Stage's detail screen provides key metrics about each stage of a job, including the amount of data that has been spilled to disk. If you see a high number in the "Spill (Memory)" or "Spill (Disk)" columns, it's an indication that a partition is spilling to disk.

The Executor's log files can also provide valuable information about spill. If a task is spilling a lot of data, you'll see messages in the logs like "Spilling UnsafeExternalSorter to disk" or "Task memory spill". These messages indicate that the task ran out of memory and had to spill data to disk.

upvoted 5 times

42.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 128 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 128
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team has configured a job to process customer requests to be forgotten (have their data deleted). All user data that needs to be deleted is stored in Delta Lake tables using default table settings.

The team has decided to process all deletions from the previous week as a batch job at 1am each Sunday. The total duration of this job is less than one hour. Every Monday at 3am, a batch job executes a series of VACUUM commands on all Delta Lake tables throughout the organization.

The compliance officer has recently learned about Delta Lake's time travel functionality. They are concerned that this might allow continued access to deleted data.

Assuming all delete logic is correctly implemented, which statement correctly addresses this concern?

A. Because the VACUUM command permanently deletes all files containing deleted records, deleted records may be accessible with time travel for around 24 hours.

B. Because the default data retention threshold is 24 hours, data files containing deleted records will be retained until the VACUUM job is run the following day.

C. Because the default data retention threshold is 7 days, data files containing deleted records will be retained until the VACUUM job is run 8 days later. Most Voted

D. Because Delta Lake's delete statements have ACID guarantees, deleted records will be permanently purged from all storage systems as soon as a delete job completes.

Hide Answer

Suggested Answer: *C*

03355a2 10 months ago

Selected Answer: A

Since the team is expecting last week's data to be deleted on Sunday at 1am to 2am. The data will be available for approx 24hrs until the vacuum command is run on Monday at 3am.

upvoted 1 times

cales 6 months, 2 weeks ago

No! By default Vacuum does not remove rows deleted whithin the last 7 days. To do it you should modify the property delta.deletedFileRetentionDuration
https://docs.databricks.com/en/delta/history.html#configure-data-retention-for-time-travel-queries

upvoted 1 times

43.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 179 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 179
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Databricks job has been configured with three tasks, each of which is a Databricks notebook. Task A does not depend on other tasks. Tasks B and C run in parallel, with each having a serial dependency on task A.

What will be the resulting state if tasks A and B complete successfully but task C fails during a scheduled run?

A. All logic expressed in the notebook associated with tasks A and B will have been successfully completed; some operations in task C may have completed successfully.

B. Unless all tasks complete successfully, no changes will be committed to the Lakehouse; because task C failed, all commits will be rolled back automatically.

C. Because all tasks are managed as a dependency graph, no changes will be committed to the Lakehouse until all tasks have successfully been completed.

D. All logic expressed in the notebook associated with tasks A and B will have been successfully completed; any changes made in task C will be rolled back due to task failure.

Hide Answer

Suggested Answer: *A*

Community vote distribution

m79590530 6 months, 1 week ago

Selected Answer: A

Each Notebook or Task consists of multiple commands and actions performed by them. Each action may be on the data in the Delta Lake where ACID transactions take place and fully rollback certain data manipulations if some of them fail but the Notebooks/Tasks in the Job themselves will not completely fail or rollback. Therefore Answer A correctly describes the result considering the dependencies configures between the Notebooks/Tasks as described in the question.

upvoted 1 times

imatheushenrique 11 months ago

A: A. All logic expressed in the notebook associated with tasks A and B will have been successfully completed; some operations in task C may have completed successfully.

Because this type of orchestration indicates a Fan-Out.

upvoted 1 times

44.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 178 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 178
Topic #: 1
[All Certified Data Engineer Professional Questions]

The Databricks CLI is used to trigger a run of an existing job by passing the job_id parameter. The response that the job run request has been submitted successfully includes a field run_id.

Which statement describes what the number alongside this field represents?

A. The job_id and number of times the job has been run are concatenated and returned.
B. The globally unique ID of the newly triggered run.
C. The number of times the job definition has been run in this workspace.
D. The job_id is returned in this field.

Hide Answer

Suggested Answer: *B*

Community vote distribution
B (100%)

m79590530 6 months, 1 week ago

Selected Answer: B

Every job run creates and assigns a globally unique run ID to the job RUN as well as globally unique run ID's for the Tasks RUN's inside the Job.

upvoted 1 times

45.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 177 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 177
Topic #: 1
[All Certified Data Engineer Professional Questions]

What describes integration testing?

A. It validates an application use case.
B. It validates behavior of individual elements of an application,
C. It requires an automated testing framework.
D. It validates interactions between subsystems of your application.

Hide Answer

Suggested Answer: *D*

Community vote distribution
D (100%)

m79590530 6 months, 1 week ago

Selected Answer: D

Interactions and cooperation between the solution and/or interfacing systems/data consumers components, subsystems and interfaces is exactly Integration testing.

upvoted 1 times

imatheushenrique 11 months ago

D. It validates interactions between subsystems of your application.
An integration test is used for different softwares validation components, subsystems, or applications that has dependencies.

upvoted 2 times

46.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 176 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 176
Topic #: 1
[All Certified Data Engineer Professional Questions]

Incorporating unit tests into a PySpark application requires upfront attention to the design of your jobs, or a potentially significant refactoring of existing code.

Which benefit offsets this additional effort?

A. Improves the quality of your data

B. Validates a complete use case of your application

C. Troubleshooting is easier since all steps are isolated and tested individually

D. Ensures that all steps interact correctly to achieve the desired end result

Hide Answer

Suggested Answer: *C*

Community vote distribution

m79590530 6 months, 1 week ago

Selected Answer: C

C is testing each unit of the solution separately. It doesn't necessarily validate the data quality as mentioned in A.
B is more for Business Case scenario testing like end-to-end testing for real life, real data execution.
D is more related to Integration testing.

upvoted 1 times

imatheushenrique 11 months ago

C. Troubleshooting is easier since all steps are isolated and tested individually
The unit tests will ensuree that specific functions and transformations will work as intended.

upvoted 1 times

47.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 174 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 174
Topic #: 1
[All Certified Data Engineer Professional Questions]

What is a method of installing a Python package scoped at the notebook level to all nodes in the currently active cluster?

A. Run source env/bin/activate in a notebook setup script

B. Install libraries from PyPI using the cluster UI

C. Use %pip install in a notebook cell

D. Use %sh pip install in a notebook cell

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

m79590530 6 months, 1 week ago

Selected Answer: C

C is correct as '%sh pip install ...' runs only on the driver node and the Cluster UI PyPi or other library installs are not scoped to a specific notebook only but to all spark sessions in all notebooks on all cluster nodes.

upvoted 1 times

imatheushenrique 11 months ago

Is necessary just run %pip install some_library inside a notebook cell
C.
OBS:
For the last update of a library can be executed %pip install some_library -U

upvoted 1 times