48.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 173 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 173
Topic #: 1
[All Certified Data Engineer Professional Questions]

Review the following error traceback:

```
---------------------------------------------------------------------------
AnalysisException                         Traceback (most recent call last)
<command-3293767849433948> in <module>
----> 1 display(df.select(3*"heartrate"))

/databricks/spark/python/pyspark/sql/dataframe.py in select(self, *cols)
   1690         [Row(name='Alice', age=12), Row(name='Bob', age=15)]
   1691         """
-> 1692         jdf = self._jdf.select(self._jcols(*cols))
   1693         return DataFrame(jdf, self.sql_ctx)
   1694

/databricks/spark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py in __call__(self, *args)
   1302
   1303         answer = self.gateway_client.send_command(command)
-> 1304         return_value = get_return_value(
   1305             answer, self.gateway_client, self.target_id, self.name)
   1306

/databricks/spark/python/pyspark/sql/utils.py in deco(*a, **kw)
    121                 # Hide where the exception came from that shows a non-Pythonic
    122                 # JVM exception message.
--> 123                 raise converted from None
    124             else:
    125                 raise
```

```
   1691         """
-> 1692         jdf = self._jdf.select(self._jcols(*cols))
   1693         return DataFrame(jdf, self.sql_ctx)
   1694

/databricks/spark/python/lib/py4j-0.10.9-src.zip/py4j/java_gateway.py in __call__(self, *args)
   1302
   1303         answer = self.gateway_client.send_command(command)
-> 1304         return_value = get_return_value(
   1305             answer, self.gateway_client, self.target_id, self.name)
   1306

/databricks/spark/python/pyspark/sql/utils.py in deco(*a, **kw)
    121                 # Hide where the exception came from that shows a non-Pythonic
    122                 # JVM exception message.
--> 123                 raise converted from None
    124             else:
    125                 raise

AnalysisException: cannot resolve '`heartrateheartrateheartrate`' given input columns:
[spark_catalog.database.table.device_id, spark_catalog.database.table.heartrate,
spark_catalog.database.table.mrn, spark_catalog.database.table.time];
'Project ['heartrateheartrateheartrate]
+- SubqueryAlias spark_catalog.database.table
   +- Relation[device_id#75L,heartrate#76,mrn#77L,time#78] parquet
```

Which statement describes the error being raised?

A. There is a syntax error because the heartrate column is not correctly identified as a column.

B. There is no column in the table named heartrateheartrateheartrate

C. There is a type error because a column object cannot be multiplied.

D. There is a type error because a DataFrame object cannot be multiplied.

Hide Answer

m79590530 6 months, 1 week ago

Selected Answer: B

The final error clearly states that such column name can not be resolved in the source dataframe schema/structure

upvoted 1 times

49.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 170 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 170
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

Hide Answer

Suggested Answer: *D*

Community vote distribution

m79590530 6 months, 1 week ago

Selected Answer: D

Cluster lifecycle events are visible in the Cluster Event Log

upvoted 1 times

imatheushenrique 11 months ago

Its possible to see the metricks of compute with Ganglia, but the question is about a timeline so D, Cluster Event Log seems correct.

upvoted 3 times

50.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 165
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data governance team is reviewing code used for deleting records for compliance with GDPR. The following logic has been implemented to propagate delete requests from the user_lookup table to the user_aggregates table.

```
(spark.read
  .format("delta")
  .option("readChangeData", True)
  .option("startingTimestamp", '2021-08-22 00:00:00')
  .option("endingTimestamp", '2021-08-29 00:00:00')
  .table("user_lookup")
  .createOrReplaceTempView("changes"))

spark.sql("""
  DELETE FROM user_aggregates
  WHERE user_id IN (
    SELECT user_id
    FROM changes
    WHERE _change_type='delete'
    )
""")
```

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

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

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

C. No; the change data feed only tracks inserts and updates, not deleted records.

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

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

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

C. No; the change data feed only tracks inserts and updates, not deleted records.

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

Hide Answer

Suggested Answer: *B*

Community vote distribution

m79590530 6 months, 1 week ago

Selected Answer: B

Default Delta Lake time travel retention is 7 days so records deleted are still accessible via previous table versions up to 7 days later unless somebody changes this default setting to less days and runs VACUUM on the table earlier.

upvoted 1 times

imatheushenrique 11 months ago

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

upvoted 3 times

51.

Question #: 161
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer wants to join a stream of advertisement impressions (when an ad was shown) with another stream of user clicks on advertisements to correlate when impressions led to monetizable clicks.

In the code below, Impressions is a streaming DataFrame with a watermark ("event_time", "10 minutes")

```
.groupBy(
window("event_time", "5 minutes"),
"id")
.count()
).     withWatermark("event_time", 2 hours)
impressions.join(clicks, expr("clickAdId = impressionAdId"), "inner")
```

The data engineer notices the query slowing down significantly.

Which solution would improve the performance?

A. Joining on event time constraint: clickTime >= impressionTime AND clickTime <= impressionTime interval 1 hour

B. Joining on event time constraint: clickTime + 3 hours < impressionTime - 2 hours

C. Joining on event time constraint: clickTime == impressionTime using a leftOuter join

D. Joining on event time constraint: clickTime >= impressionTime - interval 3 hours and removing watermarks

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

Submit

m79590530 6 months, 1 week ago

Selected Answer: A

Answer A is the only possible logically.
B configures clickTime to be earlier than impressionTime
C says that clickTime should be the same as impressionTime with all clicks left joined to impressions
D wants to remove Watermarks which will lead to memory leaks and depletion for both streams staging/aggregation purposes by Spark

upvoted 2 times

52.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 153
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team is configuring environments for development, testing, and production before beginning migration on a new data pipeline. The team requires extensive testing on both the code and data resulting from code execution, and the team wants to develop and test against data as similar to production data as possible.

A junior data engineer suggests that production data can be mounted to the development and testing environments, allowing pre-production code to execute against production data. Because all users have admin privileges in the development environment, the junior data engineer has offered to configure permissions and mount this data for the team.

Which statement captures best practices for this situation?

A. All development, testing, and production code and data should exist in a single, unified workspace; creating separate environments for testing and development complicates administrative overhead.

B. In environments where interactive code will be executed, production data should only be accessible with read permissions; creating isolated databases for each environment further reduces risks.

C. Because access to production data will always be verified using passthrough credentials, it is safe to mount data to any Databricks development environment.

D. Because Delta Lake versions all data and supports time travel, it is not possible for user error or malicious actors to permanently delete production data; as such, it is generally safe to mount production data anywhere.

Hide Answer

Suggested Answer: B

m79590530 6 months, 1 week ago

Selected Answer: B

Production data should be maximum secured against intentional and unintentional modifications by developers or workspace/UC admins. So setting it up with read only access and in different catalog or schema/database per environment is best approach.

upvoted 2 times

53.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 144
Topic #: 1
[All Certified Data Engineer Professional Questions]

Each configuration below is identical to the extent that each cluster has 400 GB total of RAM, 160 total cores and only one Executor per VM.

Given an extremely long-running job for which completion must be guaranteed, which cluster configuration will be able to guarantee completion of the job in light of one or more VM failures?

A. • Total VMs: 8
• 50 GB per Executor
• 20 Cores / Executor

B. • Total VMs: 16
• 25 GB per Executor
• 10 Cores / Executor

C. • Total VMs: 1
• 400 GB per Executor
• 160 Cores/Executor

D. • Total VMs: 4
• 100 GB per Executor
• 40 Cores / Executor

Hide Answer

m79590530 6 months, 1 week ago

Selected Answer: B

Distributing work across more Workers/Executors will have better guarantee in case of 1 or more of them fail

upvoted 1 times

54.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 127
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data science team has requested assistance in accelerating queries on free-form text from user reviews. The data is currently stored in Parquet with the below schema:

item_id INT, user_id INT, review_id INT, rating FLOAT, review STRING

The review column contains the full text of the review left by the user. Specifically, the data science team is looking to identify if any of 30 key words exist in this field.

A junior data engineer suggests converting this data to Delta Lake will improve query performance.

Which response to the junior data engineer's suggestion is correct?

A. Delta Lake statistics are not optimized for free text fields with high cardinality.
B. Delta Lake statistics are only collected on the first 4 columns in a table.
C. ZORDER ON review will need to be run to see performance gains.
D. The Delta log creates a term matrix for free text fields to support selective filtering.

Hide Answer

Suggested Answer: *A*

Community vote distribution

m79590530 6 months, 1 week ago

Selected Answer: A

Delta Lake optimizations are not well suited for long TIMESTAMP or STRING fields and can not provide good indexing, data skipping or statistics logging for them.

upvoted 1 times

55.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 53 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 53
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which distribution does Databricks support for installing custom Python code packages?

A. sbt
B. CRANC. npm
D. Wheels **Most Voted**
E. jars

Hide Answer

Suggested Answer: *D*

Community vote distribution
D (100%)

by alexvno at *Dec. 18, 2023, 8:29 a.m.*

benni_ale 6 months, 1 week ago
Selected Answer: D
I think D is correct
upvoted 1 times

hal2401me 1 year, 2 months ago
Selected Answer: D
https://learn.microsoft.com/en-us/azure/databricks/workflows/jobs/how-to/use-python-wheels-in-workflows
upvoted 4 times

sodere 1 year, 4 months ago
Selected Answer: D
https://learn.microsoft.com/en-us/azure/databricks/workflows/jobs/how-to/use-python-wheels-in-workflows
upvoted 1 times

alexvno 1 year, 4 months ago
Selected Answer: D
Wheels should be ok
upvoted 2 times

56.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 92 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 92
Topic #: 1
[All Certified Data Engineer Professional Questions]

In order to prevent accidental commits to production data, a senior data engineer has instituted a policy that all development work will reference clones of Delta Lake tables. After testing both DEEP and SHALLOW CLONE, development tables are created using SHALLOW CLONE.

A few weeks after initial table creation, the cloned versions of several tables implemented as Type 1 Slowly Changing Dimension (SCD) stop working. The transaction logs for the source tables show that VACUUM was run the day before.

Which statement describes why the cloned tables are no longer working?

A. Because Type 1 changes overwrite existing records, Delta Lake cannot guarantee data consistency for cloned tables.

B. Running VACUUM automatically invalidates any shallow clones of a table; DEEP CLONE should always be used when a cloned table will be repeatedly queried.

C. Tables created with SHALLOW CLONE are automatically deleted after their default retention threshold of 7 days.

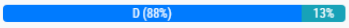
D. The metadata created by the CLONE operation is referencing data files that were purged as invalid by the VACUUM command. Most Voted

E. The data files compacted by VACUUM are not tracked by the cloned metadata; running REFRESH on the cloned table will pull in recent changes.

Hide Answer

Suggested Answer: *D*

Community vote distribution



alexvno Highly Voted 1 year, 4 months ago

Selected Answer: D

Shallow clone: only duplicates the metadata of the table being cloned; the data files of the table itself are not copied. These clones are cheaper to create but are not self-contained and depend on the source from which they were cloned as the source of data. If the files in the source that the clone depends on are removed, for example with VACUUM, a shallow clone may become unusable. Therefore, shallow clones are typically used for short-lived use cases such as testing and experimentation.

upvoted 8 times

benni_ale Most Recent 6 months, 2 weeks ago

I was not sure whether B or D but somehow I think that running VACUUM comand does not invalidate SHALLOW CLONEs . I mean its just that the data referenced by the clone is no longer present. It can still happen that a SHALLOW CLONE is working even after a VACUUM command run on the cloned table (origin) . So B is not completely correct.

upvoted 2 times

vctrhugo 1 year, 2 months ago

Selected Answer: D

In Delta Lake, the VACUUM command deletes data files that are no longer referenced by a Delta table and are older than the retention threshold. When a table is cloned using SHALLOW CLONE, the clone references the same data files as the original table but creates a new transaction log. If VACUUM is run on the original table, it can delete data files that are still being referenced by the cloned table's metadata, causing the cloned table to stop working. This is because the VACUUM command doesn't know about the cloned table's references to the data files. Therefore, it's important to be cautious when running VACUUM on tables that have clones.

upvoted 3 times

57.

Question #: 20

Topic #: 1

[All Certified Data Engineer Professional Questions]

A data architect has designed a system in which two Structured Streaming jobs will concurrently write to a single bronze Delta table. Each job is subscribing to a different topic from an Apache Kafka source, but they will write data with the same schema. To keep the directory structure simple, a data engineer has decided to nest a checkpoint directory to be shared by both streams.

The proposed directory structure is displayed below:

```
./bronze
├── _checkpoint
├── _delta_log
├── year_week=2020_01
├── year_week=2020_02
└── ...
```

Which statement describes whether this checkpoint directory structure is valid for the given scenario and why?

A. No; Delta Lake manages streaming checkpoints in the transaction log.

B. Yes; both of the streams can share a single checkpoint directory.

C. No; only one stream can write to a Delta Lake table.

D. Yes; Delta Lake supports infinite concurrent writers.

E. No; each of the streams needs to have its own checkpoint directory. Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (88%) 13%

thxsgod Highly Voted 1 year, 7 months ago

Selected Answer: E

Correct, E.

Source:
https://docs.databricks.com/en/optimizations/isolation-level.html#:~:text=If%20a%20streaming%20query%20using%20the%20same%20checkpoint%20location%20is%20started%20multiple%20times%20concurrently%20and%20tries%20to%20write%20to%20the%20Delta%20table%20at%20the%20same%20time.%20You%20should%20never%20have%20two%20streaming%20queries%20use%20the%20same%20checkpoint%20location%20and%20run%20at%20the%20same%20time.

upvoted 11 times

benni_ale Most Recent 6 months, 2 weeks ago

Selected Answer: E

E is the correct

upvoted 1 times

imatheushenrique 11 months ago

E. No; each of the streams needs to have its own checkpoint directory.
The checkpoint directory is 1 to 1

upvoted 2 times

svik 11 months, 3 weeks ago

Selected Answer: B

It is not clear from the question that year_week=2020_01 and year_week=2020_02 are used by stream 1 and stream 2 respectively. If they use the common parent checkpoint directory with individual sub folders for checkpointing, that should work fine. In that case the answer should be B

upvoted 2 times

Kill9 10 months, 1 week ago

That are table partitions. They are not used to build checkpoint adress. The adress finish at /bronze

upvoted 1 times

58.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 157 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 157
Topic #: 1
[All Certified Data Engineer Professional Questions]

A small company based in the United States has recently contracted a consulting firm in India to implement several new data engineering pipelines to power artificial intelligence applications. All the company's data is stored in regional cloud storage in the United States.

The workspace administrator at the company is uncertain about where the Databricks workspace used by the contractors should be deployed.

Assuming that all data governance considerations are accounted for, which statement accurately informs this decision?

A. Databricks runs HDFS on cloud volume storage; as such, cloud virtual machines must be deployed in the region where the data is stored.

B. Databricks workspaces do not rely on any regional infrastructure; as such, the decision should be made based upon what is most convenient for the workspace administrator.

C. Cross-region reads and writes can incur significant costs and latency; whenever possible, compute should be deployed in the same region the data is stored.
Most Voted

D. Databricks notebooks send all executable code from the user's browser to virtual machines over the open internet; whenever possible, choosing a workspace region near the end users is the most secure.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

59.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 109 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 109
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Delta Lake table representing metadata about content posts from users has the following schema:

user_id LONG, post_text STRING, post_id STRING, longitude FLOAT, latitude FLOAT, post_time TIMESTAMP, date DATE

Based on the above schema, which column is a good candidate for partitioning the Delta Table?

A. post_time
B. latitude
C. post_id
D. user_id
E. date **Most Voted**

Hide Answer

Suggested Answer: *E*

Community vote distribution

benni_ale 6 months ago

Selected Answer: E

Date is usually best candidate for time series data without further specifications

upvoted 1 times

vctrhugo 1 year, 2 months ago

Selected Answer: E

Partitioning a Delta Lake table on the date column is a common practice. This is because partitioning by date can significantly improve query performance when dealing with time-series data. It allows for efficient filtering of data based on time periods, which is a common requirement in many analytics workloads. Partitioning by date also helps manage the size of your partitions, as each partition will contain only the data for a specific date. This can lead to more efficient reads and writes, and can also make it easier to manage and maintain your data.

upvoted 4 times

60.

Actual exam question from Databricks's Certified Data Engineer Professional

Question #: 40

Topic #: 1

[All Certified Data Engineer Professional Questions]

The view updates represents an incremental batch of all newly ingested data to be inserted or updated in the customers table.

The following logic is used to process these records.

```
MERGE INTO customers
USING (
  SELECT updates.customer_id as merge_ey, updates.*
  FROM updates

  UNION ALL

  SELECT NULL as merge_key, updates.*
  FROM updates JOIN customers
  ON updates.customer_id = customers.customer_id
  WHERE customers.current = true AND updates.address <> customers.address
) staged_updates
ON customers.customer_id = mergeKey
WHEN MATCHED AND customers.current = true AND customers.address <> staged_updates.address THEN
  UPDATE SET current = false, end_date = staged_updates.effective_date
WHEN NOT MATCHED THEN
  INSERT(customer_id, address, current, effective_date, end_date)
  VALUES(staged_updates.customer_id, staged_updates.address, true, staged_updates.effective_date,
null)
```

Which statement describes this implementation?

A. The customers table is implemented as a Type 3 table; old values are maintained as a new column alongside the current value.

B. The customers table is implemented as a Type 2 table; old values are maintained but marked as no longer current and new values are inserted

```
  FROM updates

  UNION ALL

  SELECT NULL as merge_key, updates.*
  FROM updates JOIN customers
  ON updates.customer_id = customers.customer_id
  WHERE customers.current = true AND updates.address <> customers.address
) staged_updates
ON customers.customer_id = mergeKey
WHEN MATCHED AND customers.current = true AND customers.address <> staged_updates.address THEN
  UPDATE SET current = false, end_date = staged_updates.effective_date
WHEN NOT MATCHED THEN
  INSERT(customer_id, address, current, effective_date, end_date)
  VALUES(staged_updates.customer_id, staged_updates.address, true, staged_updates.effective_date,
null)
```

Which statement describes this implementation?

A. The customers table is implemented as a Type 3 table; old values are maintained as a new column alongside the current value.

B. The customers table is implemented as a Type 2 table; old values are maintained but marked as no longer current and new values are inserted. **Most Voted**

C. The customers table is implemented as a Type 0 table; all writes are append only with no changes to existing values.

D. The customers table is implemented as a Type 1 table; old values are overwritten by new values and no history is maintained.

E. The customers table is implemented as a Type 2 table; old values are overwritten and new customers are appended.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

Submit

Tayari 6 months ago

B is correct

upvoted 1 times

imatheushenrique 10 months, 4 weeks ago

B. The customers table is implemented as a Type 2 table; old values are maintained but marked as no longer current and new values are inserted.

A Type 1 table does not track changes in dimensional attributes - the new value overwrites the existing value. Here, we do not preserve historical changes in data.

A Type 2 Table tracks change over time by creating new rows for each change. A new dimension record is inserted with a high-end date or one with NULL. The previous record is "closed" with an end date. This approach maintains a complete history of changes and allows for as-was reporting use cases.

A data warehousing method called Slowly Changing Dimension (SCD) Type 3 is used to track both the old and new values while managing historical changes in data over time. To reflect the historical and present values of an attribute, SCD Type 3 keeps two extra columns in the dimension table.

upvoted 3 times

spaceexplorer 1 year, 3 months ago

Selected Answer: B

61.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 119 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 119
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement regarding Spark configuration on the Databricks platform is true?

A. The Databricks REST API can be used to modify the Spark configuration properties for an interactive cluster without interrupting jobs currently running on the cluster.

B. Spark configurations set within a notebook will affect all SparkSessions attached to the same interactive cluster.

C. When the same Spark configuration property is set for an interactive cluster and a notebook attached to that cluster, the notebook setting will always be ignored.

D. Spark configuration properties set for an interactive cluster with the Clusters UI will impact all notebooks attached to that cluster. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

by vexor3 at *July 20, 2024, 8:58 a.m.*

62.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 55
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 statement describes a main benefit that offset 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 Most Voted

D. Yields faster deployment and execution times

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

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

by alexvno at *Dec. 18, 2023, 8:47 a.m.*

alexvno Highly Voted 1 year, 4 months ago

Selected Answer: C

Unit tests are small, isolated tests that are used to check specific parts of the code, such as functions or classes

upvoted 5 times

nedlo Most Recent 6 months ago

Selected Answer: C

D is integration tests (how they relate to each other how connect), E is E2E test, C is "testing individually" which is only one fittign definition of unittest

upvoted 2 times

nedlo 6 months ago

i mean E is integration test B is E2E test

upvoted 1 times

jmjm21 10 months, 2 weeks ago

Selected Answer: C

Answer is C.

upvoted 1 times

63.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 147
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior data engineer seeks to leverage Delta Lake's Change Data Feed functionality to create a Type 1 table representing all of the values that have ever been valid for all rows in a bronze table created with the property delta.enableChangeDataFeed = true. They plan to execute the following code as a daily job:

```
from pyspark.sql.functions import col

(spark.read.format("delta")
  .option("readChangeFeed", "true")
  .option("startingVersion", 0)
  .table("bronze")
  .filter(col("_change_type").isin(["update_postimage", "insert"]))
  .write
  .mode("append")
  .table("bronze_history_type1")
 )
```

Which statement describes the execution and results of running the above query multiple times?

A. Each time the job is executed, newly updated records will be merged into the target table, overwriting previous values with the same primary keys.

B. Each time the job is executed, the entire available history of inserted or updated records will be appended to the target table, resulting in many duplicate entries.
Most Voted

C. Each time the job is executed, only those records that have been inserted or updated since the last execution will be appended to the target table, giving the desired result.

D. Each time the job is executed, the differences between the original and current versions are calculated; this may result in duplicate entries for some records.

Hide Answer

benni_ale 5 months, 3 weeks ago
Selected Answer: B
B seems ok
upvoted 1 times

m79590530 6 months, 1 week ago
Selected Answer: B
Since the code is using version 0 for the CDF-enabled table every time it is executed all the historical changes being insert of update for the table will be 'append'-ed to the target table since this is the write command option provided.
upvoted 1 times

Adrifersilva 7 months, 1 week ago
B. This bad effect (many duplicates) happens because the code reads from the starting version 0, appending all changes since the beginning.
upvoted 2 times

64.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 146 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 146
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer is configuring a pipeline that will potentially see late-arriving, duplicate records.

In addition to de-duplicating records within the batch, which of the following approaches allows the data engineer to deduplicate data against previously processed records as it is inserted into a Delta table?

A. Rely on Delta Lake schema enforcement to prevent duplicate records.
B. VACUUM the Delta table after each batch completes.
C. Perform an insert-only merge with a matching condition on a unique key. Most Voted
D. Perform a full outer join on a unique key and overwrite existing data.

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

by m79590530 at *Oct. 20, 2024, 4:54 p.m.*

benni_ale 5 months, 3 weeks ago
Selected Answer: C
C seems logical
upvoted 1 times

m79590530 6 months, 1 week ago
Selected Answer: C
From all the provided options Answer C is the only meaningful and possible one. Also MERGE INTO ... WHEN NOT MATCHED INSERT *; is a standard solution for adding/appending non-existing records (by key) to the target table withOUT duplicating.
upvoted 2 times

65.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 143 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 143
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Spark job is taking longer than expected. Using the Spark UI, a data engineer notes that the Min, Median, and Max Durations for tasks in a particular stage show the minimum and median time to complete a task as roughly the same, but the max duration for a task to be roughly 100 times as long as the minimum.

Which situation is causing increased duration of the overall job?

A. Task queueing resulting from improper thread pool assignment.
B. Spill resulting from attached volume storage being too small.
C. Network latency due to some cluster nodes being in different regions from the source data
D. Skew caused by more data being assigned to a subset of spark-partitions. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution
D (100%)

by vexor3 at *July 20, 2024, 10:47 a.m.*

66.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 140 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 140
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. 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
D. Optimized writes use logical partitions instead of directory partitions; because partition boundaries are only represented in metadata, fewer small files are written.

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

upvoted 1 times

Farid77 6 months, 3 weeks ago

B is correct.
C is wrong OPTIMIZE is a separate process from write.

upvoted 1 times

vexor3 9 months, 1 week ago

Selected Answer: C

C is correct

upvoted 2 times

only_vimal 8 months, 3 weeks ago

Please provide your input to Questions 144,145,146,147,149 also. Thanks in advance

upvoted 1 times

67.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 154 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 154
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data architect has mandated that all tables in the Lakehouse should be configured as external Delta Lake tables.

Which approach will ensure that this requirement is met?

A. Whenever a database is being created, make sure that the LOCATION keyword is used.
B. When the workspace is being configured, make sure that external cloud object storage has been mounted.
C. Whenever a table is being created, make sure that the LOCATION keyword is used.
D. When tables are created, make sure that the UNMANAGED keyword is used in the CREATE TABLE statement.

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

benni_ale 5 months, 3 weeks ago

Selected Answer: C

Location keyword in CTAS statment in only way to create External tables

upvoted 1 times

m79590530 6 months, 1 week ago

Selected Answer: C

CREATE-ing a TABLE with LOCATION key word makes it EXTERNAL TABLE. By CREATE-ing the database/schema with the LOCATION key word we can have specific locations for the schemas/databases containing MANAGED tables when these tables inside these schemas/databases are CREATE-d withOUT the LOCATION key word.
This approach allows for configuring MANAGED TABLES at specific locations by fully leveraging Databricks Lakehouse automatic optimizations and performance tunning for them.

upvoted 2 times

benni_ale 5 months, 3 weeks ago

U are a dragon

upvoted 1 times

68.

Question #: 135

Topic #: 1

[All Certified Data Engineer Professional Questions]

A view is registered with the following code:

```
CREATE VIEW recent_orders AS (
  SELECT a.user_id, a.email, b.order_id, b.order_date
  FROM
    (SELECT user_id, email
    FROM users) a
    INNER JOIN
    (SELECT user_id, order_id, order_date
    FROM orders
    WHERE order_date >= (current_date() - 7)) b
    ON a.user_id = b.user_id
)
```

Both users and orders are Delta Lake tables.

Which statement describes the results of querying recent_orders?

A. All logic will execute when the view is defined and store the result of joining tables to the DBFS; this stored data will be returned when the view is queried.

B. Results will be computed and cached when the view is defined; these cached results will incrementally update as new records are inserted into source tables.

C. All logic will execute at query time and return the result of joining the valid versions of the source tables at the time the query finishes.

D. All logic will execute at query time and return the result of joining the valid versions of the source tables at the time the query began. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (89%) 11%

Freyr Highly Voted 11 months ago

Selected Answer: D

Correct Answer: D

Correct because this option correctly describes the behavior of SQL views in Databricks. The view's query is executed against the current state of the data in the source tables at the moment the query begins. This means that any changes to the data that are committed while the query is running will not be reflected in the results of the query currently executing.

upvoted 5 times

benni_ale Most Recent 5 months, 3 weeks ago

69.

Question #: 132
Topic #: 1
[All Certified Data Engineer Professional Questions]

An hourly batch job is configured to ingest data files from a cloud object storage container where each batch represent all records produced by the source system in a given hour. The batch job to process these records into the Lakehouse is sufficiently delayed to ensure no late-arriving data is missed. The user_id field represents a unique key for the data, which has the following schema:

user_id BIGINT, username STRING, user_utc STRING, user_region STRING, last_login BIGINT, auto_pay BOOLEAN, last_updated BIGINT

New records are all ingested into a table named account_history which maintains a full record of all data in the same schema as the source. The next table in the system is named account_current and is implemented as a Type 1 table representing the most recent value for each unique user_id.

Which implementation can be used to efficiently update the described account_current table as part of each hourly batch job assuming there are millions of user accounts and tens of thousands of records processed hourly?

A. Filter records in account_history using the last_updated field and the most recent hour processed, making sure to deduplicate on username; write a merge statement to update or insert the most recent value for each username.

B. Use Auto Loader to subscribe to new files in the account_history directory; configure a Structured Streaming trigger available job to batch update newly detected files into the account_current table.

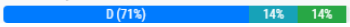
C. Overwrite the account_current table with each batch using the results of a query against the account_history table grouping by user_id and filtering for the max value of last_updated.

D. Filter records in account_history using the last_updated field and the most recent hour processed, as well as the max last_login by user_id write a merge statement to update or insert the most recent value for each user_id. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution



shaojunni 7 months ago

Selected Answer: B

A, D both wrong. They only take data from the latest update. It is too narrow. Same user_id can have several updates within an hour to update different fields. So use auto loader to apply all the updates within an hour is the only correct answer.

upvoted 1 times

fe3b2fc 8 months, 1 week ago

Selected Answer: A

Answer is A. You're meeting all the requirements with less overhead. It's only updating on the most recent record, so duplicates are handled.

Answer D is too much overhead. They're doing a full table scan for all records, which as the question stated, is millions of records.

upvoted 1 times

Onobhas01 7 months, 3 weeks ago

User Id would be a better column to merge into with, username might not be distinct

upvoted 1 times

Freyr 11 months ago

Selected Answer: D

Correct Answer: D

Similar to Option A, but specifically designed around the user_id, which is the primary key. This approach ensures that the account_current is always up-to-date with the latest information per user based on the primary key, reducing the risk of duplicate information and ensuring the integrity of the data with respect to the unique identifier.

upvoted 3 times

70.

Question #: 84
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data architect has decided that once data has been ingested from external sources into the Databricks Lakehouse, table access controls will be leveraged to manage permissions for all production tables and views.

The following logic was executed to grant privileges for interactive queries on a production database to the core engineering group.

GRANT USAGE ON DATABASE prod TO eng;
GRANT SELECT ON DATABASE prod TO eng;

Assuming these are the only privileges that have been granted to the eng group and that these users are not workspace administrators, which statement describes their privileges?

A. Group members have full permissions on the prod database and can also assign permissions to other users or groups.

B. Group members are able to list all tables in the prod database but are not able to see the results of any queries on those tables.

C. Group members are able to query and modify all tables and views in the prod database, but cannot create new tables or views.

D. Group members are able to query all tables and views in the prod database, but cannot create or edit anything in the database. Most Voted

E. Group members are able to create, query, and modify all tables and views in the prod database, but cannot define custom functions.

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

vctrhugo 1 year, 2 months ago

Selected Answer: D

The GRANT statements provided in the logic grant the USAGE privilege, allowing the group members to see the existence of the database, and the SELECT privilege, allowing them to query tables and views. However, they do not have permissions to create or edit anything in the database. Therefore, the correct description is that group members can query all tables and views in the prod database but cannot create or edit any objects in the database.

upvoted 1 times

divingbell17 1 year, 3 months ago

Selected Answer: D

D is correct assuming unity catalog is not enabled

upvoted 1 times

aragorn_brego 1 year, 5 months ago

Selected Answer: D

The GRANT USAGE ON DATABASE statement gives the eng group the ability to access the prod database. This means they can enter the database context and list the tables. The GRANT SELECT ON DATABASE statement additionally grants them permission to perform SELECT queries on all existing tables and views within the prod database. However, these privileges do not include creating new tables or views, modifying existing tables, or assigning permissions to other users or groups.

upvoted 3 times

Dileepvikram 1 year, 5 months ago

D is answer

upvoted 4 times

71.

Question #: 89
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Databricks job has been configured with 3 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.

If task A fails during a scheduled run, which statement describes the results of this run?

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

B. Tasks B and C will attempt to run as configured; any changes made in task A will be rolled back due to task failure.

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

D. Tasks B and C will be skipped; some logic expressed in task A may have been committed before task failure. Most Voted

E. Tasks B and C will be skipped; task A will not commit any changes because of stage failure.

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

mouad_attaqi Highly Voted 6 months ago

Selected Answer: D

D is correct, taks B and C will definitely be skipped, since Task A is notebook, the ACID logic is at cell level, some logic might be executed before failing cell

upvoted 6 times

aragorn_brego Most Recent 5 months, 1 week ago

Selected Answer: D

In Databricks job execution, if a task that other tasks depend on fails, the dependent tasks will not be executed. Since Tasks B and C depend on the successful completion of Task A, they will be skipped if Task A fails. However, if Task A performs any operations that commit changes before the failure occurs (such as writing to a Delta table), those changes remain and are not automatically rolled back unless the logic within Task A specifically includes rollback mechanisms for partial failures.

upvoted 4 times

Dileepvikram 5 months, 2 weeks ago

D is the answer

upvoted 3 times

sturcu 6 months ago

Selected Answer: D

Some ops in task A may have fished before fail

upvoted 3 times

72.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 172 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 172
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineer is using Spark's MEMORY_ONLY storage level.

Which indicators should the data engineer look for in the Spark UI's Storage tab to signal that a cached table is not performing optimally?

A. On Heap Memory Usage is within 75% of Off Heap Memory Usage

B. The RDD Block Name includes the "*" annotation signaling a failure to cache

C. Size on Disk is > 0 **Most Voted**

D. The number of Cached Partitions > the number of Spark Partitions

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (86%) 14%

03355a2 10 months ago
The RDD answer is incorrect for this question due to the fact that while this indicates a failure to cache, it is more specific to identifying individual blocks that failed to cache rather than providing a general signal of a suboptimal performance for the entire cached table.
upvoted 1 times

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

Freyr 11 months ago
Selected Answer: B
Correct Answer: B
Option B, is the most correct and relevant choice for an indicator that a cached table is not performing optimally in a MEMORY_ONLY scenario. If an RDD block includes a "?" annotation, it strongly suggests issues with caching, which would directly impact the performance and expected behavior of MEMORY_ONLY caching. This indication points to a failure to cache the data entirely in memory, which is what MEMORY_ONLY intends to do.

Option C, could also be a relevant indicator in general caching scenarios (e.g., MEMORY_AND_DISK), but it contradicts the MEMORY_ONLY setting directly. Therefore, Option B is chosen based on the specific storage level described.
upvoted 1 times

Freyr 10 months, 3 weeks ago
THE CORRECT ANSWER IS: C
PLEASE IGNORE MY PREVIOUS ANSWER.

Long story short, B is correct in the context of non-functional requirement, but the question is based in functional requirement, and sorry for the confusion.
upvoted 3 times

73.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 182
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team has configured a Databricks SQL query and alert to monitor the values in a Delta Lake table. The recent_sensor_recordings table contains an identifying sensor_id alongside the timestamp and temperature for the most recent 5 minutes of recordings.

The below query is used to create the alert:

```
SELECT MEAN(temperature), MAX(temperature), MIN(temperature)
FROM recent_sensor_recordings
GROUP BY sensor_id
```

The query is set to refresh each minute and always completes in less than 10 seconds. The alert is set to trigger when mean (temperature) > 120. Notifications are triggered to be sent at most every 1 minute.

If this alert raises notifications for 3 consecutive minutes and then stops, which statement must be true?

A. The total average temperature across all sensors exceeded 120 on three consecutive executions of the query
B. The average temperature recordings for at least one sensor exceeded 120 on three consecutive executions of the query Most Voted
C. The source query failed to update properly for three consecutive minutes and then restarted
D. The maximum temperature recording for at least one sensor exceeded 120 on three consecutive executions of the query

Hide Answer

Thameur01 4 months, 3 weeks ago

Selected Answer: B

B, because avg temp is calculated by sensor_id and not total

upvoted 1 times

74.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 77
Topic #: 1
[All Certified Data Engineer Professional Questions]

In order to facilitate near real-time workloads, a data engineer is creating a helper function to leverage the schema detection and evolution functionality of Databricks Auto Loader. The desired function will automatically detect the schema of the source directly, incrementally process JSON files as they arrive in a source directory, and automatically evolve the schema of the table when new fields are detected.

The function is displayed below with a blank:

```
def auto_load_json(source_path: str,
                   checkpoint_path: str,
                   target_table_path: str):
    (spark.readStream
        .format("cloudFiles")
        .option("cloudFiles.format", "json")
        .option("cloudFiles.schemaLocation", checkpoint_path)
        .load(source_path)
        ______
    )
```

Which response correctly fills in the blank to meet the specified requirements?

A.
```
.writeStream
.option("mergeSchema", True)
```

)

Which response correctly fills in the blank to meet the specified requirements?

A.
```
.writeStream
.option("mergeSchema", True)
.start(target_table_path)
```

B.
```
.writeStream
.option("checkpointLocation", checkpoint_path)
.option("mergeSchema", True)
.trigger(once=True)
.start(target_table_path)
```

C.
```
.write
.option("checkpointLocation", checkpoint_path)
.option("mergeSchema", True)
.outputMode("append")
.save(target_table_path)
```

D.
```
.write
.option("mergeSchema", True)
.mode("append")
.save(target_table_path)
```

E. Most Voted
```
.writeStream
.option("checkpointLocation", checkpoint_path)
.option("mergeSchema", True)
.start(target_table_path)
```

Hide Answer

benni_ale 4 months, 3 weeks ago

Selected Answer: E

Evolve Schema = mergeSchema option is needed ; Incrementally = checkpointing is needed; Real-Time = WriteStream with default trigger . The only option that catches all of these is E

upvoted 2 times

35fd6dd 8 months, 2 weeks ago

Selected Answer: E

write is not for spark streaming

upvoted 2 times

Freyr 11 months ago

Selected Answer: E

Reference: https://docs.databricks.com/en/ingestion/auto-loader/schema.html

writeStream: Ensures real-time streaming write capabilities, which is essential f
or near real-time workloads.
checkpointLocation: Necessary for fault tolerance and tracking progress.
mergeSchema: Ensures automatic schema evolution, allowing new columns to be detected and added to the target table.

Why Option 'C ' is incorrect?
Uses write instead of writeStream, which is for batch processing, making it inappropriate for real-time streaming.

Why Option 'B ' is incorrect?
Although it includes checkpointLocation and mergeSchema, the addition of trigger(once=True) is not necessary in this context, and it is better suited for batch-like processing.

Reference: https://docs.databricks.com/en/ingestion/auto-loader/schema.html

upvoted 4 times

vikram12apr 1 year, 1 month ago

Selected Answer: E

streamRead & StreamWrite shares the schema using checkpoint location
so cloudFiles.schemaLocation needs to be same for checkpointLocation so that we dont need to specify it manually
also mergeSchema True make sure if any new column detected , it will be added in the target table

https://docs.databricks.com/en/ingestion/auto-loader/schema.html

75.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 191
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior data engineer has configured a workload that posts the following JSON to the Databricks REST API endpoint 2.0/jobs/create.

```
{
  "name": "Ingest new data",
  "existing_cluster_id": "6015-954420-peace720",
  "notebook_task": {
    "notebook_path": "/Prod/ingest.py"
  }
}
```

Assuming that all configurations and referenced resources are available, which statement describes the result of executing this workload three times?

A. The logic defined in the referenced notebook will be executed three times on the referenced existing all purpose cluster.

B. The logic defined in the referenced notebook will be executed three times on new clusters with the configurations of the provided cluster ID.

C. Three new jobs named "Ingest new data" will be defined in the workspace, but no jobs will be executed. Most Voted

D. One new job named "Ingest new data" will be defined in the workspace, but it will not be executed.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

Ayomidetolu_A 4 months, 3 weeks ago

Selected Answer: C

C is the correct answer

upvoted 1 times

divyapsingh 4 months, 3 weeks ago

Selected Answer: C

C is the answer as 3 times call will create three jobs with same name but with different job id.

upvoted 2 times

temple1305 4 months, 3 weeks ago

Selected Answer: C

C correct, 3 jobs created, no executions

upvoted 1 times

76.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 196 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 196
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data architect has designed a system in which two Structured Streaming jobs will concurrently write to a single bronze Delta table. Each job is subscribing to a different topic from an Apache Kafka source, but they will write data with the same schema. To keep the directory structure simple, a data engineer has decided to nest a checkpoint directory to be shared by both streams.

The proposed directory structure is displayed below:

Which statement describes whether this checkpoint directory structure is valid for the given scenario and why?

A. No; Delta Lake manages streaming checkpoints in the transaction log.

B. Yes; both of the streams can share a single checkpoint directory.

C. No; only one stream can write to a Delta Lake table.

D. No; each of the streams needs to have its own checkpoint directory. Most Voted

Hide Answer

divyapsingh 4 months, 3 weeks ago

Selected Answer: D

D is the answer. two steams can not have same checkpoint directory.

upvoted 1 times

77.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 83 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 83
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 of the following solutions 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. Because the value field is stored as binary data, this information is not considered PII and no special precautions should be taken.

D. Separate object storage containers should be specified based on the partition field, allowing isolation at the storage level.

E. Data should be partitioned by the topic field, allowing ACLs and delete statements to leverage partition boundaries. Most Voted

Hide Answer

Suggested Answer: *E*

mouad_attaqi Highly Voted 1 year, 6 months ago

Selected Answer: E

I think answer E is correct, as by default partitionning by a column will create a separate folder for each subset data linked to the partition

upvoted 13 times

benni_ale Most Recent 4 months, 3 weeks ago

Selected Answer: E

Partitioning by topic field let delete queries leverage partioning boundaries

upvoted 2 times

benni_ale 6 months, 1 week ago

Selected Answer: E

E E E E E

upvoted 1 times

ojudz08 1 year, 2 months ago

Selected Answer: D

i think it's best to isolate the storage to avoid mistakenly deleting tables in the same storage so I go with D

upvoted 1 times

78.

Question #: 19
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. Events are recorded once per minute per device.
Streaming DataFrame df has the following schema:
"device_id INT, event_time TIMESTAMP, temp FLOAT, humidity FLOAT"
Code block:

```
df.withWatermark("event_time", "10 minutes")
    .groupBy(
        _____,
        "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. to_interval("event_time", "5 minutes").alias("time")
B. window("event_time", "5 minutes").alias("time") Most Voted
C. "event_time"
D. window("event_time", "10 minutes").alias("time")
E. lag("event_time", "10 minutes").alias("time")

Hide Answer

Submit

imatheushenrique 4 months, 4 weeks ago
B. window("event_time", "5 minutes").alias("time")
In Structured Streaming, expressing such windows on event-time is simply performing a special grouping using the window() function. For example, counts over 5 minute tumbling (non-overlapping) windows on the eventTime column in the event is as following.
upvoted 4 times

Jay_98_11 9 months, 2 weeks ago
Selected Answer: B
correct B
upvoted 2 times

79.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 171 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 171
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. CPU Utilization is around 75%
C. Network I/O never spikes
D. Total Disk Space remains constant

Hide Answer

Suggested Answer: *B*

imatheushenrique 4 months, 4 weeks ago
B.
This level of CPU utilization indicates that the cluster is being used without being underutilized.
upvoted 3 times

80.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 74 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 74
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement describes the correct use of pyspark.sql.functions.broadcast?

A. It marks a column as having low enough cardinality to properly map distinct values to available partitions, allowing a broadcast join.

B. It marks a column as small enough to store in memory on all executors, allowing a broadcast join.

C. It caches a copy of the indicated table on attached storage volumes for all active clusters within a Databricks workspace.

D. It marks a DataFrame as small enough to store in memory on all executors, allowing a broadcast join. Most Voted

E. It caches a copy of the indicated table on all nodes in the cluster for use in all future queries during the cluster lifetime.

Hide Answer

Suggested Answer: *D*

Community vote distribution

Freyr 5 months ago

Selected Answer: D

Correct Answer: D. It marks a DataFrame as small enough to store in memory on all executors, allowing a broadcast join.

Reference: https://spark.apache.org/docs/latest/api/python/reference/pyspark.sql/api/pyspark.sql.functions.broadcast.html

upvoted 3 times

aragorn_brego 11 months, 1 week ago

Selected Answer: D

The broadcast function in PySpark is used in the context of joins. When you mark a DataFrame with broadcast, Spark tries to send this DataFrame to all worker nodes so that it can be joined with another DataFrame without shuffling the larger DataFrame across the nodes. This is particularly beneficial when the DataFrame is small enough to fit into the memory of each node. It helps to optimize the join process by reducing the amount of data that needs to be shuffled across the cluster, which can be a very expensive operation in terms of computation and time.

upvoted 3 times

Dileepvikram 11 months, 3 weeks ago

81.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 79 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 79
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data architect has mandated that all tables in the Lakehouse should be configured as external (also known as "unmanaged") Delta Lake tables.

Which approach will ensure that this requirement is met?

A. When a database is being created, make sure that the LOCATION keyword is used.
B. When configuring an external data warehouse for all table storage, leverage Databricks for all ELT.
C. When data is saved to a table, make sure that a full file path is specified alongside the Delta format. Most Voted
D. When tables are created, make sure that the EXTERNAL keyword is used in the CREATE TABLE statement.
E. When the workspace is being configured, make sure that external cloud object storage has been mounted.

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

sturcu Highly Voted 1 year, 6 months ago
Non of the provided.
It should be: When a table is created, make sure LOCATION is provided
upvoted 8 times

vctrhugo Highly Voted 1 year, 2 months ago
Selected Answer: C
In Delta Lake, an external (or unmanaged) table is a table created outside of the data lake but is still accessible from the data lake. The data for external tables is stored in a location specified by the user, not in the default directory of the data lake. When you save data to an external table, you need to specify the full file path where the data will be stored. This makes the table "external" because the data itself is not managed by Delta Lake, only the metadata is. This is why specifying a full file path alongside the Delta format when saving data to a table will ensure that the table is configured as an external Delta Lake table.
upvoted 5 times

Sriramiyer92 Most Recent 4 months, 2 weeks ago
Selected Answer: C
Folks note:
While creating a table - Use of External keyword - Non Mandatory.
Mentioning Location and providing a path - Mandatory.
In option C, it is not mentioned explicitly that Location keyword is used. But since the path is provided.. implies the use of Location keyword indirectly. The devil is in the details. :)
upvoted 3 times

hjy 7 months, 2 weeks ago
'create external table' statement is using in HIVE, so C is correct.

82.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 54 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 54
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which Python variable contains a list of directories to be searched when trying to locate required modules?

A. importlib.resource_path
B. sys.path **Most Voted**
C. os.path
D. pypi.path
E. pylib.source

Hide Answer

Suggested Answer: *B*

Community vote distribution
B (100%)

alexvno Highly Voted 1 year, 4 months ago
Selected Answer: B
sys. path is a built-in variable within the sys module. It contains a list of directories that the interpreter will search in for the required module
upvoted 7 times

Sriramiyer92 Most Recent 4 months, 2 weeks ago
Selected Answer: B
sys.path is a list in Python that contains the directories the interpreter searches for modules when importing them. It is initialized with the default paths when Python starts and can be modified during runtime if needed.
upvoted 1 times

benni_ale 6 months, 1 week ago
Selected Answer: B
sys.path is a built-in variable within the sys module. It contains a list of directories that the interpreter will search in for the required module.
upvoted 1 times

83.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 34 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 34
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data architect has mandated that all tables in the Lakehouse should be configured as external Delta Lake tables.
Which approach will ensure that this requirement is met?

A. Whenever a database is being created, make sure that the LOCATION keyword is used

B. When configuring an external data warehouse for all table storage, leverage Databricks for all ELT.

C. Whenever a table is being created, make sure that the LOCATION keyword is used. **Most Voted**

D. When tables are created, make sure that the EXTERNAL keyword is used in the CREATE TABLE statement.

E. When the workspace is being configured, make sure that external cloud object storage has been mounted.

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (90%) 10%

by chokthewa at *Oct. 21, 2023, 8:17 a.m.*

Sriramiyer92 4 months, 2 weeks ago

Selected Answer: C

Note: External keyword is not mandatory.
Location is mandatory the presence implies, that the table is external

upvoted 2 times

carah 4 months, 2 weeks ago

Selected Answer: C

A. is not correct:
having schema with LOCATION
CREATE SCHEMA my_schema
LOCATION 's3://<bucket-path>/my_schema';

Table Location Scenarios:

Table Without LOCATION:

CREATE TABLE my_schema.my_table (id INT);

The table will be stored in the default warehouse directory (e.g., dbfs:/user/hive/warehouse/), not the schema's LOCATION.

Table With Explicit LOCATION: If you want the table to be stored under the schema's LOCATION, you need to specify the location explicitly:

CREATE TABLE my_schema.my_table (id INT)
LOCATION 's3://<bucket-path>/my_schema/my_table/';

So, if you want all tables under the schema to use the schema's LOCATION, explicitly specify the LOCATION for each table during creation.

upvoted 3 times

84.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 227
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer wants to refactor the following DLT code, which includes multiple table definitions with very similar code.

```
@dlt.table(name=f"t1_dataset")
def t1_dataset():
    return spark.read.table(t1)

@dlt.table(name=f"t2_dataset")
def t2_dataset():
    return spark.read.table(t2)

@dlt.table(name=f"t3_dataset")
def t3_dataset():
    return spark.read.table(t3)

...
```

In an attempt to programmatically create these tables using a parameterized table definition, the data engineer writes the following code.

```
tables = ["t1", "t2", "t3"]

for t in tables:
    @dlt.table(name=f"{t}_dataset")
        def new_table():
```

```
@dlt.table(name=f"t1_dataset")
def t1_dataset():
    return spark.read.table(t1)

@dlt.table(name=f"t2_dataset")
def t2_dataset():
    return spark.read.table(t2)

@dlt.table(name=f"t3_dataset")
def t3_dataset():
    return spark.read.table(t3)

...
```

In an attempt to programmatically create these tables using a parameterized table definition, the data engineer writes the following code.

```
tables = ["t1", "t2", "t3"]

for t in tables:
    @dlt.table(name=f"{t}_dataset")
         def new_table():
             return spark.read.table(t)
```

The pipeline runs an update with this refactored code, but generates a different DAG showing incorrect configuration values for these tables.

How can the data engineer fix this?

A. Wrap the for loop inside another table definition, using generalized names and properties to replace with those from the inner table definition.

B. Convert the list of configuration values to a dictionary of table settings, using table names as keys.

C. Move the table definition into a separate function, and make calls to this function using different input parameters inside the for loop.

D. Load the configuration values for these tables from a separate file, located at a path provided by a pipeline parameter.

Thameur01 4 months, 2 weeks ago

Selected Answer: C

here is a correct implementation:
def create_table(t):
@dlt.table(name=f"{t}_dataset")
def table_definition():
return spark.read.table(t)

tables = ["t1", "t2", "t3"]
for t in tables:
create_table(t)

upvoted 2 times

benni_ale 4 months, 2 weeks ago

Selected Answer: C

Problem seems to be the fact that the new_table function has no parameter and the t variable won't be recognized as a variable.. However i have not tested it as DLT is not available in free membership. :(

upvoted 2 times

85.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 111 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 111
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which describes 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. Use b in a notebook cell

C. Use %pip install in a notebook cell **Most Voted**

D. Use %sh pip install in a notebook cell

E. Install libraries from PyPI using the cluster UI

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

by 60ties at *Nov. 15, 2023, 9:12 p.m.*

JamesWright 4 months, 2 weeks ago
C is correct
upvoted 3 times

aragorn_brego 5 months, 1 week ago
Selected Answer: C
In Databricks notebooks, you can use the %pip install command in a notebook cell to install a Python package. This will install the package on all nodes in the currently active cluster at the notebook level. It is a feature provided by Databricks to facilitate the installation of Python libraries for the notebook environment specifically.
upvoted 3 times

60ties 5 months, 1 week ago
Selected Answer: C
C is correct
upvoted 3 times

86.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 189 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 189
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Delta table of weather records is partitioned by date and has the below schema:

date DATE, device_id INT, temp FLOAT, latitude FLOAT, longitude FLOAT

To find all the records from within the Arctic Circle, you execute a query with the below filter:

latitude > 66.3

Which statement describes how the Delta engine identifies which files to load?

A. All records are cached to an operational database and then the filter is applied

B. The Parquet file footers are scanned for min and max statistics for the latitude column

C. The Hive metastore is scanned for min and max statistics for the latitude column

D. The Delta log is scanned for min and max statistics for the latitude column Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

Thameur01 4 months, 2 weeks ago

Selected Answer: D

As per the documentation, I understand that the table statistics can be fetched through the delta log (eg min, max, count) in order to not read the underlying data of a delta table. This is the case for numerical types, and timestamp is supposed to be supported.

upvoted 2 times

temple1305 4 months, 3 weeks ago

Selected Answer: D

Delta Table's log consist statistics for columns

upvoted 2 times

87.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 221 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 221
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer needs to capture pipeline settings from an existing setting in the workspace, and use them to create and version a JSON file to create a new pipeline.

Which command should the data engineer enter in a web terminal configured with the Databricks CLI?

A. Use list pipelines to get the specs for all pipelines; get the pipeline spec from the returned results; parse and use this to create a pipeline
B. Stop the existing pipeline; use the returned settings in a reset command
C. Use the get command to capture the settings for the existing pipeline; remove the pipeline_id and rename the pipeline; use this in a create command **Most Voted**
D. Use the clone command to create a copy of an existing pipeline; use the get JSON command to get the pipeline definition; save this to git

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

benni_ale 4 months, 2 weeks ago

Selected Answer: C

I say C from common logical sense, however il have not properly tested it... I just don't see any problems with that

upvoted 2 times

88.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 101 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 101
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which indicators would you look for in the Spark UI's Storage tab to signal that a cached table is not performing optimally? Assume you are using Spark's MEMORY_ONLY storage level.

A. Size on Disk is < Size in Memory
B. The RDD Block Name includes the "*" annotation signaling a failure to cache
C. Size on Disk is > 0 **Most Voted**
D. The number of Cached Partitions > the number of Spark Partitions
E. On Heap Memory Usage is within 75% of Off Heap Memory Usage

Hide Answer

Suggested Answer: *C*

Community vote distribution
C (100%)

vctrhugo Highly Voted 1 year, 2 months ago

Selected Answer: C

C. Size on Disk is > 0

When using Spark's MEMORY_ONLY storage level, the ideal scenario is that the data is fully cached in memory, and the Size on Disk should be 0 (indicating that the data is not spilled to disk). If the Size on Disk is greater than 0, it suggests that some data has been spilled to disk, which can lead to degraded performance as reading from disk is slower than reading from memory.

upvoted 7 times

benni_ale Most Recent 4 months, 3 weeks ago

89.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 9 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 9
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior member of the data engineering team is exploring the language interoperability of Databricks notebooks. The intended outcome of the below code is to register a view of all sales that occurred in countries on the continent of Africa that appear in the geo_lookup table.
Before executing the code, running SHOW TABLES on the current database indicates the database contains only two tables: geo_lookup and sales.

Cmd 1

```
%python
countries_af = [x[0] for x in
spark.table("geo_lookup").filter("continent='AF'").select("country").collect()]
```

Cmd 2

```
%sql
CREATE VIEW sales_af AS
  SELECT *
  FROM sales
  WHERE city IN countries_af
  AND CONTINENT = "AF"
```

Which statement correctly describes the outcome of executing these command cells in order in an interactive notebook?

A. Both commands will succeed. Executing show tables will show that countries_af and sales_af have been registered as views.

B. Cmd 1 will succeed. Cmd 2 will search all accessible databases for a table or view named countries_af: if this entity exists, Cmd 2 will succeed.

C. Cmd 1 will succeed and Cmd 2 will fail. countries_af will be a Python variable representing a PySpark DataFrame.

D. Both commands will fail. No new variables, tables, or views will be created.

E. Cmd 1 will succeed and Cmd 2 will fail. countries_af will be a Python variable containing a list of strings. **Most Voted**

aragorn_brego Highly Voted 1 year, 5 months ago

Selected Answer: E

Cmd 1 is a PySpark command that collects the list of countries from the 'geo_lookup' table where the continent is Africa ('AF'). This command will execute successfully, resulting in countries_af being a list of country names (strings) in Python's local memory.

Cmd 2 is an SQL command intended to create a view named 'sales_af' from the 'sales' table, filtered by the cities in the countries_af list. However, this will fail because the countries_af variable exists in the Python environment and is not recognized in the SQL context. SQL does not have access to Python variables directly; they are two separate execution contexts within a Databricks notebook. There is no table or view named countries_af that SQL can reference; it is merely a Python list variable.

The other options are incorrect because they either assume cross-contextual operation between Python and SQL within a Databricks notebook (which is not possible in the way described in the commands), or they do not correctly interpret the outcome of running the commands.

upvoted 12 times

freely 4 months, 2 weeks ago

I mean without sepecifying the catalog and the schema in a unity catalog context ? this will only succeed if the table is in the defaut catalog and schema

upvoted 1 times

benni_ale Most Recent 6 months, 3 weeks ago

Selected Answer: E

E , the collect method outputs strings so the python variable bill be a list of string which should not be called as a spark table as in cmd 2

upvoted 1 times

imatheushenrique 10 months, 3 weeks ago

E. Cmd 1 will succeed and Cmd 2 will fail. countries_af will be a Python variable containing a list of strings.

upvoted 1 times

90.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 36 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 36
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Delta Lake table representing metadata about content posts from users has the following schema: user_id LONG, post_text STRING, post_id STRING, longitude FLOAT, latitude FLOAT, post_time TIMESTAMP, date DATE
This table is partitioned by the date column. A query is run with the following filter: longitude < 20 & longitude > -20
Which statement describes how data will be filtered?

A. Statistics in the Delta Log will be used to identify partitions that might Include files in the filtered range.

B. No file skipping will occur because the optimizer does not know the relationship between the partition column and the longitude.

C. The Delta Engine will use row-level statistics in the transaction log to identify the flies that meet the filter criteria.

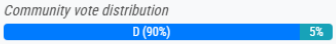
D. Statistics in the Delta Log will be used to identify data files that might include records in the filtered range. Most Voted

E. The Delta Engine will scan the parquet file footers to identify each row that meets the filter criteria.

Hide Answer

Suggested Answer: *D*

Community vote distribution



Enduresoul Highly Voted 1 year, 5 months ago

Selected Answer: D

D is correct. A partition can include multiple files. And the statistics are collected for each file.

upvoted 12 times

AlejandroU Most Recent 4 months, 2 weeks ago

Selected Answer: B

Answer B. Single Comparison Filter (e.g., latitude > 66.3): File skipping is highly efficient because Delta can use min/max statistics to directly eliminate files that don't meet the condition.
Range Filters (e.g., longitude < 20 AND longitude > -20): File skipping is still possible but less efficient, because Delta has to evaluate whether any records in the file might meet the condition, even if the min and max values of the column in the file overlap with the filter range.
So in summary, file skipping works best with single comparisons like latitude > 66.3 but is less effective with range filters like longitude < 20 AND longitude > -20.

upvoted 1 times

Sriramiyer92 4 months, 2 weeks ago

Selected Answer: D

Do not get confused between option c and d. Given answer is correct.

upvoted 1 times

hebied 4 months, 4 weeks ago

Selected Answer: D

D is more suitable

upvoted 1 times

AndreFR 8 months, 1 week ago

Selected Answer: D

Min and max values of each parquet file are stored in Delta Logs
Delta data skipping automatically collects the stats (min, max, etc.) for the first 32 columns for each underlying Parquet file when you write data into a Delta table.
Databricks takes advantage of this information (minimum and maximum values) at query time to skip unnecessary files in order to speed up the queries.
https://www.databricks.com/discover/pages/optimize-data-workloads-guide#delta-data

upvoted 2 times