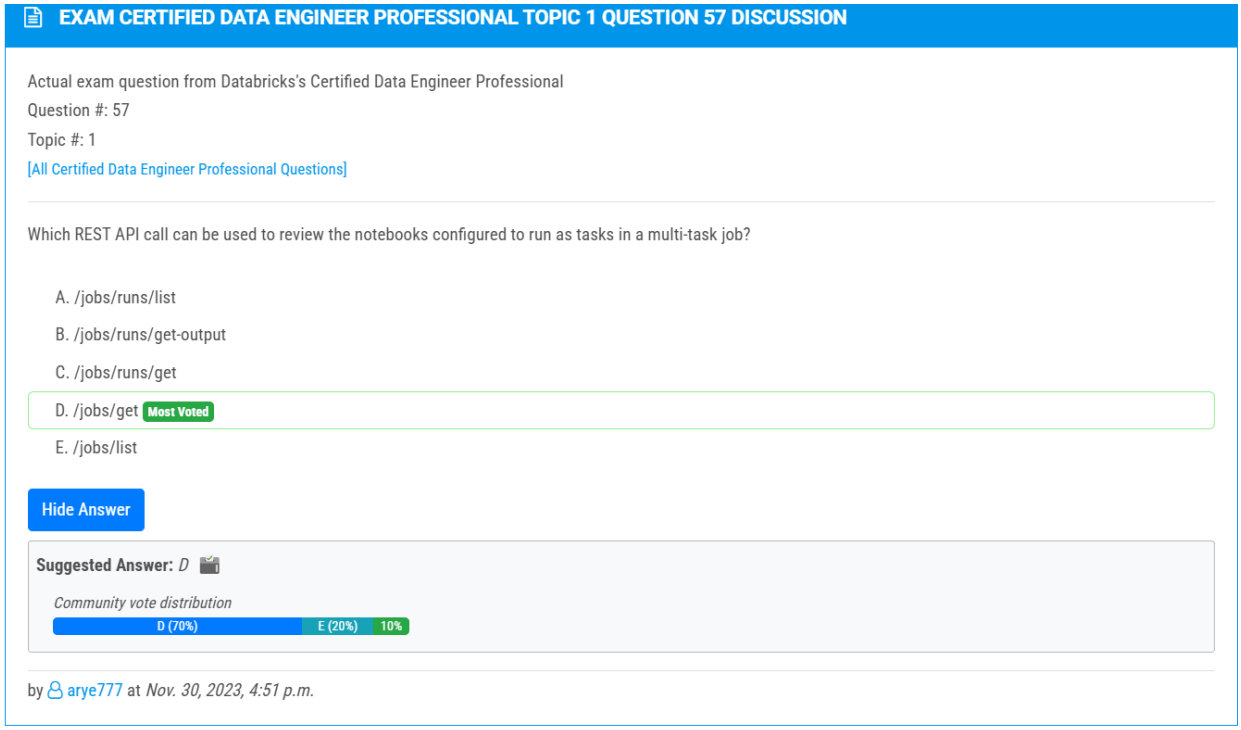
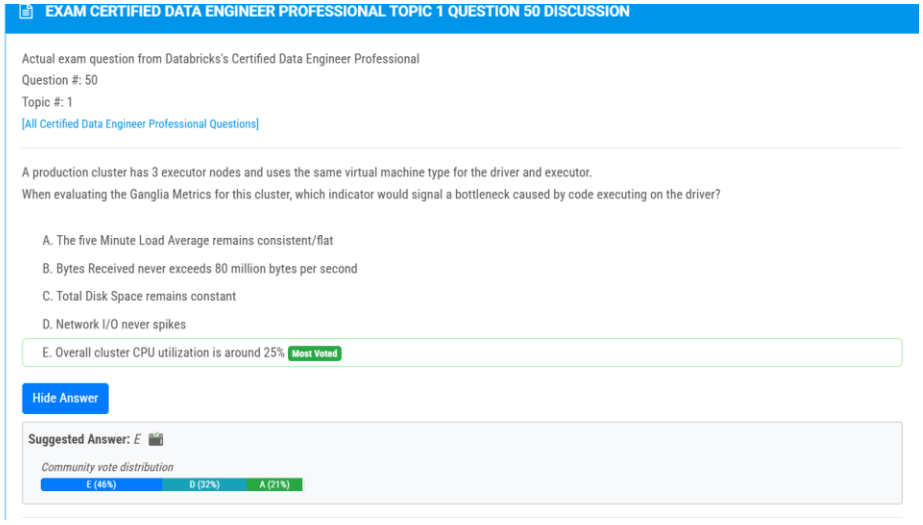
1.



https://docs.databricks.com/api/workspace/jobs/get
responses/settings/tasks/notebook_task/notebook_path

The correct answer is E. /jobs/list, not C. /jobs/runs/get. Here's why: /jobs/list: Provides a list of all jobs in the workspace along with their configurations, including task details like the notebooks assigned to each task. This makes it the best choice for reviewing notebooks configured as tasks in a multi-task job. /jobs/get: Can also be used if the goal is to review the tasks (and notebooks) of a specific job. However, the question does not limit the scope to a single job.

2.



BrianNguyen95 Highly Voted 1 year, 8 months ago

Option E: In a Spark cluster, the driver node is responsible for managing the execution of the Spark application, including scheduling tasks, managing the execution plan, and interacting with the cluster manager. If the overall cluster CPU utilization is low (e.g., around 25%), it may indicate that the driver node is not utilizing the available resources effectively and might be a bottleneck.

upvoted 19 times

fe3b2fc 8 months, 1 week ago

A bottleneck occurs when resources are over utilized not underutilized, so that explanation doesn't make too much sense. CPU utilization would be at 100% and you wouldn't see spike in I/O if the driver was the issue. Conversely if the I/O was spiked and CPU utilization was at 25% , then network could be the issue. D is the only logical answer in this case.

upvoted 3 times

benni_ale 6 months ago

i like this more

upvoted 2 times

guillesd 1 year, 2 months ago

Overall CPU utilization can be misleading. The 25% utilization could be caused by the workload not requiring more than that rather than everything being executed in the driver node.

upvoted 2 times

JoG1221 Most Recent 6 days, 10 hours ago

Selected Answer: A

Option E is valid and insightful, Option A is more targeted when you're specifically trying to detect a bottleneck on the driver.

upvoted 1 times

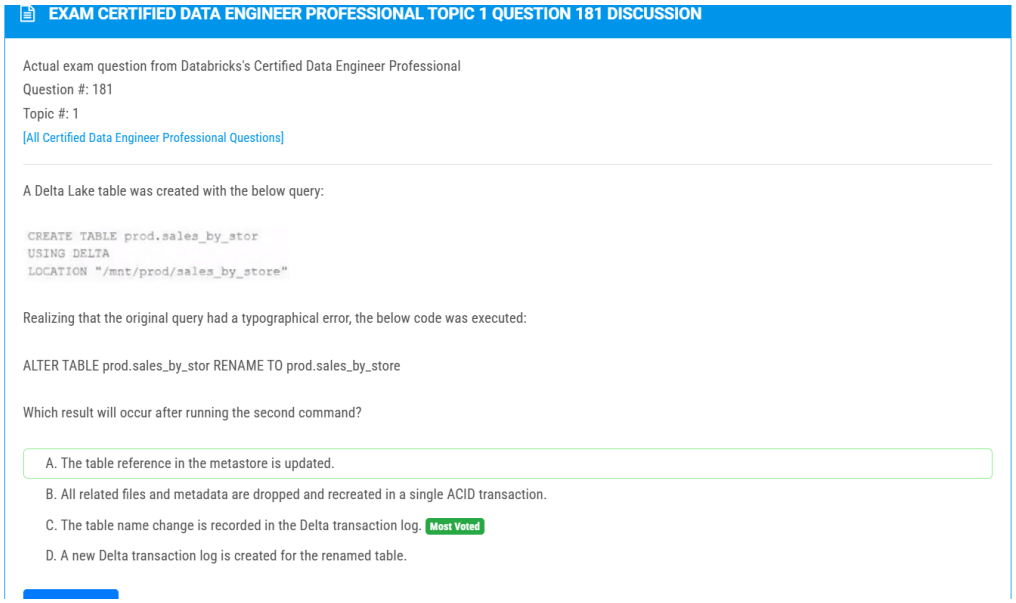
Tedet 1 month, 3 weeks ago

Selected Answer: A

When you see the "Five Minute Load Average" remain consistent or flat, it could indicate that the driver is under heavy load and is struggling to keep up with the workload. In the case of a Spark cluster, if the driver is handling too much work, it can become a bottleneck and prevent the overall job from progressing efficiently.

upvoted 2 times

3.



Stalker200 1 week, 1 day ago

Selected Answer: A

The Hive Metastore (or Unity Catalog) updates the logical table name. The data and files stay untouched.

upvoted 1 times

lakime 1 month, 3 weeks ago

Selected Answer: C

while the metastore is updated, the key mechanism for tracking changes in Delta Lake is the transaction log.

upvoted 1 times

A and C 50% each

4.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 117 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 117
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer, User A, has promoted a pipeline to production by using the REST API to programmatically create several jobs. A DevOps engineer, User B, has configured an external orchestration tool to trigger job runs through the REST API. Both users authorized the REST API calls using their personal access tokens.

A workspace admin, User C, inherits responsibility for managing this pipeline. User C uses the Databricks Jobs UI to take "Owner" privileges of each job. Jobs continue to be triggered using the credentials and tooling configured by User B.

An application has been configured to collect and parse run information returned by the REST API. Which statement describes the value returned in the creator_user_name field?

A. Once User C takes "Owner" privileges, their email address will appear in this field; prior to this, User A's email address will appear in this field.

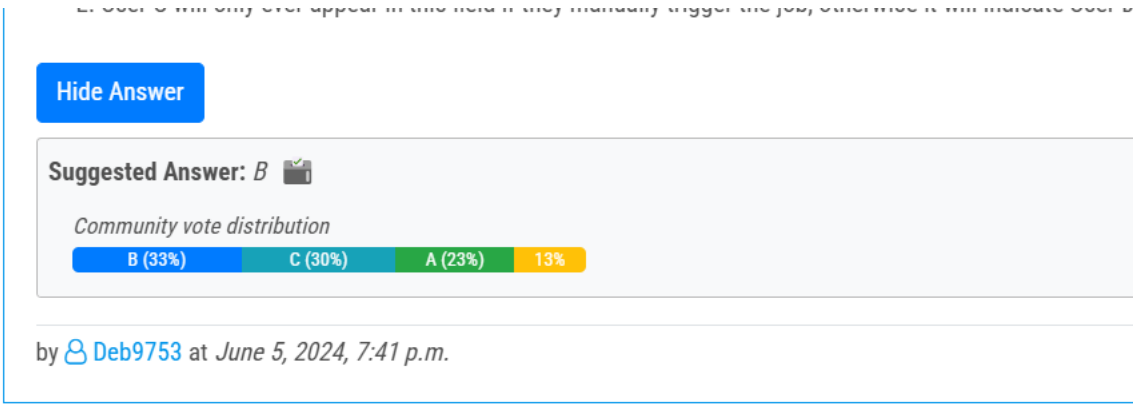
B. User B's email address will always appear in this field, as their credentials are always used to trigger the run.

C. User A's email address will always appear in this field, as they still own the underlying notebooks.

D. Once User C takes "Owner" privileges, their email address will appear in this field; prior to this, User B's email address will appear in this field.

E. User C will only ever appear in this field if they manually trigger the job, otherwise it will indicate User B.

Show Suggested Answer



5.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 52 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional

Question #: 52

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
```

```
-> 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. The code executed was PySpark but was executed in a Scala notebook.

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.

Which statement describes the error being raised?

A. The code executed was PySpark but was executed in a Scala notebook.

B. There is no column in the table named heartrateheartrateheartrate **Most Voted**

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.

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

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (73%) E (23%) 5%

by CertPeople at *Sept. 12, 2023, 8:47 a.m.*

6.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 18 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 18
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement regarding stream-static joins and static Delta tables is correct?

A. Each microbatch of a stream-static join will use the most recent version of the static Delta table as of each microbatch. **Most Voted**

B. Each microbatch of a stream-static join will use the most recent version of the static Delta table as of the job's initialization.

C. The checkpoint directory will be used to track state information for the unique keys present in the join.

D. Stream-static joins cannot use static Delta tables because of consistency issues.

E. The checkpoint directory will be used to track updates to the static Delta table.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (80%) B (20%)

by BrianNguyen95 at *Aug. 17, 2023, 2:05 p.m.*

7.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 17 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 17
Topic #: 1
[All Certified Data Engineer Professional Questions]

A production workload incrementally applies updates from an external Change Data Capture feed to a Delta Lake table as an always-on Structured Stream job. When data was initially migrated for this table, OPTIMIZE was executed and most data files were resized to 1 GB. Auto Optimize and Auto Compaction were both turned on for the streaming production job. Recent review of data files shows that most data files are under 64 MB, although each partition in the table contains at least 1 GB of data and the total table size is over 10 TB.
Which of the following likely explains these smaller file sizes?

A. Databricks has autotuned to a smaller target file size to reduce duration of MERGE operations Most Voted
B. Z-order indices calculated on the table are preventing file compaction
C. Bloom filter indices calculated on the table are preventing file compaction
D. Databricks has autotuned to a smaller target file size based on the overall size of data in the table
E. Databricks has autotuned to a smaller target file size based on the amount of data in each partition

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (73%) E (24%) 3%

8.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 16 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 16
Topic #: 1
[All Certified Data Engineer Professional Questions]

A table is registered with the following code:

```
CREATE TABLE 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 at query time and return the result of joining the valid versions of the source tables at the time the query finishes.
B. All logic will execute when the table is defined and store the result of joining tables to the DBFS; this stored data will be returned when the table is queried.
C. Results will be computed and cached when the table is defined; these cached results will incrementally update as new records are inserted into source tables.
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.
E. The versions of each source table will be stored in the table transaction log; query results will be saved to DBFS with each query.

A. Databricks has autotuned to a smaller target file size to reduce duration of MERGE operations **Most Voted**

B. Z-order indices calculated on the table are preventing file compaction

C. Bloom filter indices calculated on the table are preventing file compaction

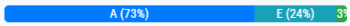
D. Databricks has autotuned to a smaller target file size based on the overall size of data in the table

E. Databricks has autotuned to a smaller target file size based on the amount of data in each partition

Hide Answer

Suggested Answer: *A*

Community vote distribution



9.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 15 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 15
Topic #: 1
[All Certified Data Engineer Professional Questions]

A table in the Lakehouse named customer_churn_params is used in churn prediction by the machine learning team. The table contains information about customers derived from a number of upstream sources. Currently, the data engineering team populates this table nightly by overwriting the table with the current valid values derived from upstream data sources.
The churn prediction model used by the ML team is fairly stable in production. The team is only interested in making predictions on records that have changed in the past 24 hours.
Which approach would simplify the identification of these changed records?

A. Apply the churn model to all rows in the customer_churn_params table, but implement logic to perform an upsert into the predictions table that ignores rows where predictions have not changed.

B. Convert the batch job to a Structured Streaming job using the complete output mode; configure a Structured Streaming job to read from the customer_churn_params table and incrementally predict against the churn model.

C. Calculate the difference between the previous model predictions and the current customer_churn_params on a key identifying unique customers before making new predictions; only make predictions on those customers not in the previous predictions.

D. Modify the overwrite logic to include a field populated by calling spark.sql.functions.current_timestamp() as data are being written; use this field to identify records written on a particular date.

E. Replace the current overwrite logic with a merge statement to modify only those records that have changed; write logic to make predictions on the changed records identified by the change data feed.

A. Apply the churn model to all rows in the customer_churn_params table, but implement logic to perform an upsert into the predictions table that ignores rows where predictions have not changed.

B. Convert the batch job to a Structured Streaming job using the complete output mode; configure a Structured Streaming job to read from the customer_churn_params table and incrementally predict against the churn model.

C. Calculate the difference between the previous model predictions and the current customer_churn_params on a key identifying unique customers before making new predictions; only make predictions on those customers not in the previous predictions.

D. Modify the overwrite logic to include a field populated by calling spark.sql.functions.current_timestamp() as data are being written; use this field to identify records written on a particular date.

E. Replace the current overwrite logic with a merge statement to modify only those records that have changed; write logic to make predictions on the changed records identified by the change data feed. **Most Voted**

Hide Answer

Suggested Answer: *E*

Community vote distribution

10.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 14 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 14
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.
Assuming there are millions of user accounts and tens of thousands of records processed hourly, which implementation can be used to efficiently update the described account_current table as part of each hourly batch job?

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

B. 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.

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

D. Use Delta Lake version history to get the difference between the latest version of account_history and one version prior, then write these records to account_current.

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

B. 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.

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

D. Use Delta Lake version history to get the difference between the latest version of account_history and one version prior, then write these records to account_current.

E. 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.

Hide Answer

Suggested Answer: *C*

Community vote distribution

11.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 10 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 10
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. All records are cached to attached storage and then the filter is applied

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

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

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (86%) 14%

12.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 219 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 219
Topic #: 1
[All Certified Data Engineer Professional Questions]

A user new to Databricks is trying to troubleshoot long execution times for some pipeline logic they are working on. Presently, the user is executing code cell-by-cell, using display() calls to confirm code is producing the logically correct results as new transformations are added to an operation. To get a measure of average time to execute, the user is running each cell multiple times interactively.

Which of the following adjustments will get a more accurate measure of how code is likely to perform in production?

A. The Jobs UI should be leveraged to occasionally run the notebook as a job and track execution time during incremental code development because Photon can only be enabled on clusters launched for scheduled jobs.

B. The only way to meaningfully troubleshoot code execution times in development notebooks is to use production-sized data and production-sized clusters with Run All execution. Most Voted

C. Production code development should only be done using an IDE; executing code against a local build of open source Spark and Delta Lake will provide the most accurate benchmarks for how code will perform in production.

D. Calling display() forces a job to trigger, while many transformations will only add to the logical query plan; because of caching, repeated execution of the same logic does not provide meaningful results.

Hide Answer

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (50%) D (50%)

13.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 187 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 187
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?

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("append")
  .option("checkpointPath", "/_checkpoints/churn_preds")
  .table("churn_preds")
)
```

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

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (71%) D (29%)

14.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 148 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 148
Topic #: 1
[All Certified Data Engineer Professional Questions]

A DLT pipeline includes the following streaming tables:

• raw_iot ingests raw device measurement data from a heart rate tracking device.
• bpm_stats incrementally computes user statistics based on BPM measurements from raw_iot.

How can the data engineer configure this pipeline to be able to retain manually deleted or updated records in the raw_iot table, while recomputing the downstream table bpm_stats table when a pipeline update is run?

A. Set the pipelines.reset.allowed property to false on raw_iot

B. Set the skipChangeCommits flag to true on raw_iot **Most Voted**

C. Set the pipelines.reset.allowed property to false on bpm_stats

D. Set the skipChangeCommits flag to true on bpm_stats

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (54%) A (46%)

15.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 64 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 64
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?

```
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. Nothing will occur until a COMMIT command is executed.

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.

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

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

Show Suggested Answer

Suggested Answer: *C*

Community vote distribution

16.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 29 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 29
Topic #: 1
[All Certified Data Engineer Professional Questions]

A new data engineer notices that a critical field was omitted from an application that writes its Kafka source to Delta Lake. This happened even though the critical field was in the Kafka source. That field was further missing from data written to dependent, long-term storage. The retention threshold on the Kafka service is seven days. The pipeline has been in production for three months.
Which describes how Delta Lake can help to avoid data loss of this nature in the future?

A. The Delta log and Structured Streaming checkpoints record the full history of the Kafka producer.

B. Delta Lake schema evolution can retroactively calculate the correct value for newly added fields, as long as the data was in the original source.

C. Delta Lake automatically checks that all fields present in the source data are included in the ingestion layer.

D. Data can never be permanently dropped or deleted from Delta Lake, so data loss is not possible under any circumstance.

E. Ingesting all raw data and metadata from Kafka to a bronze Delta table creates a permanent, replayable history of the data state. Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution

E (92%) 8%

17.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 11 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 11
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 Delta Lake time travel provides full access to the entire history of a table, deleted records can always be recreated by users with full admin privileges.

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.

E. 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

Hide Answer

Suggested Answer: *E*

Suggested Answer: *E*

Community vote distribution

E (61%) A (39%)

18.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 166 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 166
Topic #: 1
[All Certified Data Engineer Professional Questions]

An external object storage container has been mounted to the location /mnt/finance_eda_bucket.

The following logic was executed to create a database for the finance team:

```
CREATE DATABASE finance_eda_db
LOCATION '/mnt/finance_eda_bucket';
GRANT USAGE ON DATABASE finance_eda_db TO finance;
GRANT CREATE ON DATABASE finance_eda_db TO finance;
```

After the database was successfully created and permissions configured, a member of the finance team runs the following code:

```
CREATE TABLE finance_eda_db.tx_sales AS
SELECT *
FROM sales
WHERE state = "TX";
```

If all users on the finance team are members of the finance group, which statement describes how the tx_sales table will be created?

A. A logical table will persist the query plan to the Hive Metastore in the Databricks control plane.

B. An external table will be created in the storage container mounted to /mnt/finance_eda_bucket.

C. A managed table will be created in the DBFS root storage container.

D. An managed table will be created in the storage container mounted to /mnt/finance_eda_bucket. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (90%) 10%

19.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 131 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 131
Topic #: 1
[All Certified Data Engineer Professional Questions]

An upstream system is emitting change data capture (CDC) logs that are being written to a cloud object storage directory. Each record in the log indicates the change type (insert, update, or delete) and the values for each field after the change. The source table has a primary key identified by the field pk_id.

For auditing purposes, the data governance team wishes to maintain a full record of all values that have ever been valid in the source system. For analytical purposes, only the most recent value for each record needs to be recorded. The Databricks job to ingest these records occurs once per hour, but each individual record may have changed multiple times over the course of an hour.

Which solution meets these requirements?

A. Iterate through an ordered set of changes to the table, applying each in turn to create the current state of the table, (insert, update, delete), timestamp of change, and the values.

B. Use merge into to insert, update, or delete the most recent entry for each pk_id into a table, then propagate all changes throughout the system.

C. Deduplicate records in each batch by pk_id and overwrite the target table.

D. Use Delta Lake's change data feed to automatically process CDC data from an external system, propagating all changes to all dependent tables in the Lakehouse.
Most Voted

20.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 67 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 67
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. Most Voted

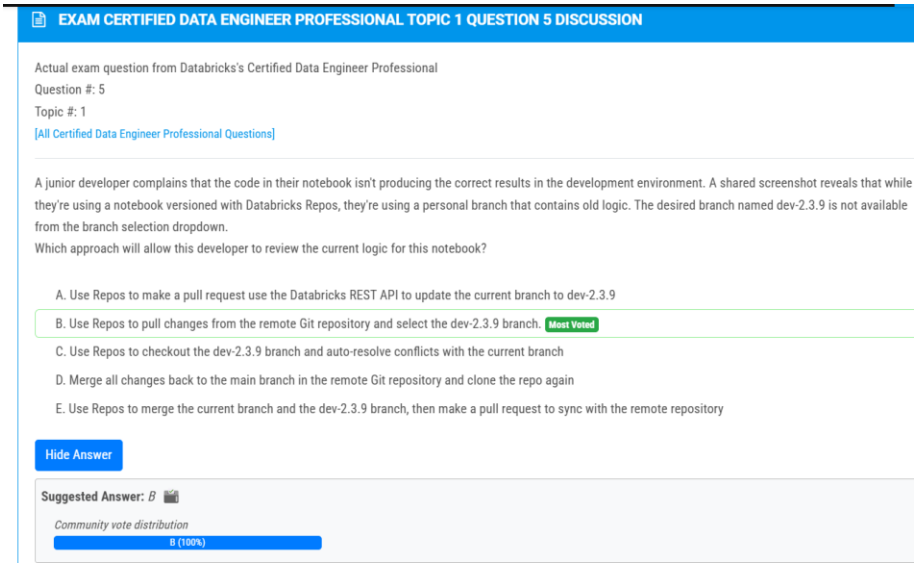
B. Text data cannot be stored with Delta Lake.

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.

E. Delta Lake statistics are only collected on the first 4 columns in a table.

21.



EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 5 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 5
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior developer complains that the code in their notebook isn't producing the correct results in the development environment. A shared screenshot reveals that while they're using a notebook versioned with Databricks Repos, they're using a personal branch that contains old logic. The desired branch named dev-2.3.9 is not available from the branch selection dropdown.
Which approach will allow this developer to review the current logic for this notebook?

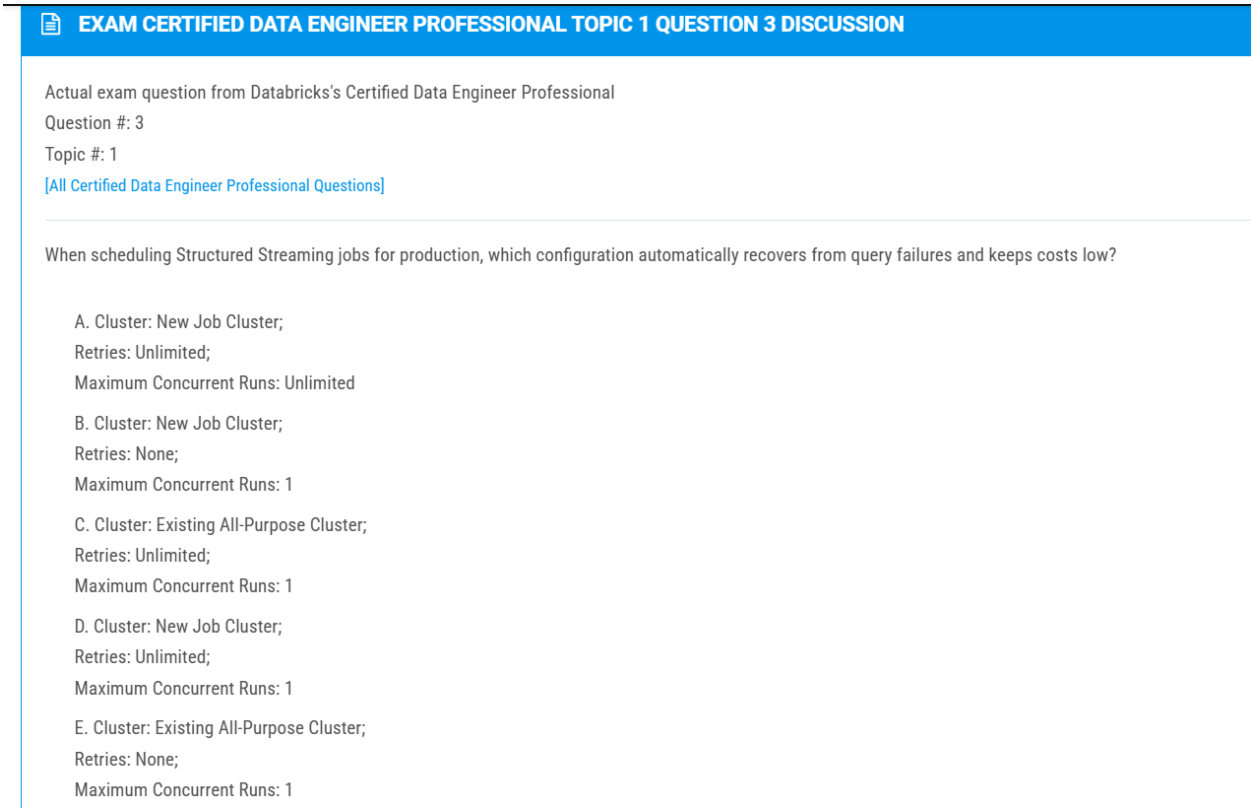
A. Use Repos to make a pull request use the Databricks REST API to update the current branch to dev-2.3.9
B. Use Repos to pull changes from the remote Git repository and select the dev-2.3.9 branch. Most Voted
C. Use Repos to checkout the dev-2.3.9 branch and auto-resolve conflicts with the current branch
D. Merge all changes back to the main branch in the remote Git repository and clone the repo again
E. Use Repos to merge the current branch and the dev-2.3.9 branch, then make a pull request to sync with the remote repository

Hide Answer

Suggested Answer: B

Community vote distribution
B (100%)

22.



EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 3 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 3
Topic #: 1
[All Certified Data Engineer Professional Questions]

When scheduling Structured Streaming jobs for production, which configuration automatically recovers from query failures and keeps costs low?

A. Cluster: New Job Cluster;
Retries: Unlimited;
Maximum Concurrent Runs: Unlimited

B. Cluster: New Job Cluster;
Retries: None;
Maximum Concurrent Runs: 1

C. Cluster: Existing All-Purpose Cluster;
Retries: Unlimited;
Maximum Concurrent Runs: 1

D. Cluster: New Job Cluster;
Retries: Unlimited;
Maximum Concurrent Runs: 1

E. Cluster: Existing All-Purpose Cluster;
Retries: None;
Maximum Concurrent Runs: 1

D. Cluster: New Job Cluster;
Retries: Unlimited;
Maximum Concurrent Runs: 1 **Most Voted**

E. Cluster: Existing All-Purpose Cluster;
Retries: None;
Maximum Concurrent Runs: 1

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

23.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 2 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 2
Topic #: 1
[All Certified Data Engineer Professional Questions]

The Databricks workspace administrator has configured interactive clusters for each of the data engineering groups. To control costs, clusters are set to terminate after 30 minutes of inactivity. Each user should be able to execute workloads against their assigned clusters at any time of the day.
Assuming users have been added to a workspace but not granted any permissions, which of the following describes the minimal permissions a user would need to start and attach to an already configured cluster.

A. "Can Manage" privileges on the required cluster
B. Workspace Admin privileges, cluster creation allowed, "Can Attach To" privileges on the required cluster
C. Cluster creation allowed, "Can Attach To" privileges on the required cluster
D. "Can Restart" privileges on the required cluster **Most Voted**
E. Cluster creation allowed, "Can Restart" privileges on the required cluster

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (78%) C (18%) 5%

24.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 95 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 95
Topic #: 1
[All Certified Data Engineer Professional Questions]

A task orchestrator has been configured to run two hourly tasks. First, an outside system writes Parquet data to a directory mounted at /mnt/raw_orders/. After this data is written, a Databricks job containing the following code is executed:

```
(spark.readStream
  .format("parquet")
  .load("/mnt/raw_orders/")
  .withWatermark("time", "2 hours")
  .dropDuplicates(["customer_id", "order_id"])
  .writeStream
  .trigger(once=True)
  .table("orders")
)
```

Assume that the fields customer_id and order_id serve as a composite key to uniquely identify each order, and that the time field indicates when the record was queued in the source system.

If the upstream system is known to occasionally enqueue duplicate entries for a single order hours apart, which statement is correct?

Assume that the fields customer_id and order_id serve as a composite key to uniquely identify each order, and that the time field indicates when the record was queued in the source system.

If the upstream system is known to occasionally enqueue duplicate entries for a single order hours apart, which statement is correct?

A. Duplicate records enqueued more than 2 hours apart may be retained and the orders table may contain duplicate records with the same customer_id and order_id.
Most Voted

B. All records will be held in the state store for 2 hours before being deduplicated and committed to the orders table.

C. The orders table will contain only the most recent 2 hours of records and no duplicates will be present.

D. Duplicate records arriving more than 2 hours apart will be dropped, but duplicates that arrive in the same batch may both be written to the orders table.

E. The orders table will not contain duplicates, but records arriving more than 2 hours late will be ignored and missing from the table.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (54%) E (46%)

25.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 26 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 26
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 a job with at least one wide transformation, which of the following cluster configurations will result in maximum performance?

A. • Total VMs; 1
• 400 GB per Executor
• 160 Cores / Executor

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

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

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

A. • Total VMs; 1
• 400 GB per Executor
• 160 Cores / Executor

B. • Total VMs: 8
• 50 GB per Executor
• 20 Cores / Executor Most Voted

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

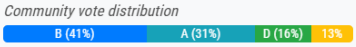
D. • Total VMs: 4
• 100 GB per Executor
• 40 Cores/Executor

E. • Total VMs:2
• 200 GB per Executor
• 80 Cores / Executor

Hide Answer

Suggested Answer: *B*

Community vote distribution



26.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 41 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 41
Topic #: 1
[All Certified Data Engineer Professional Questions]

The DevOps team has configured a production workload as a collection of notebooks scheduled to run daily using the Jobs UI. A new data engineering hire is onboarding to the team and has requested access to one of these notebooks to review the production logic.
What are the maximum notebook permissions that can be granted to the user without allowing accidental changes to production code or data?

A. Can Manage

B. Can Edit

C. No permissions

D. Can Read Most Voted

E. Can Run

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (73%) E (27%)

27.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 30 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 30
Topic #: 1
[All Certified Data Engineer Professional Questions]

A nightly job ingests data into a Delta Lake table using the following code:

```
from pyspark.sql.functions import current_timestamp, input_file_name, col
from pyspark.sql.column import Column

def ingest_daily_batch(time_col: Column, year:int, month:int, day:int):
    (spark.read
        .format("parquet")
        .load(f"/mnt/daily_batch/{year}/{month}/{day}")
        .select("*",
                time_col.alias("ingest_time"),
                inpute_file_name().alias("source_file")
               )
        .write
        .mode("append")
        .saveAsTable("bronze")
    )
```

The next step in the pipeline requires a function that returns an object that can be used to manipulate new records that have not yet been processed to the next table in the pipeline.
Which code snippet completes this function definition?
def new_records():

Which code snippet completes this function definition?
def new_records():

A. return spark.readStream.table("bronze") Most Voted

B. return spark.readStream.load("bronze")

C.
```
return (spark.read
    .table("bronze")
    .filter(col("ingest_time") == current_timestamp())
)
```

D. return spark.read.option("readChangeFeed", "true").table ("bronze")

E.
```
return (spark.read
    .table("bronze")
    .filter(col("source_file") == f"/mnt/daily_batch/{year}/{month}/{day}")
)
```

Hide Answer

Suggested Answer: *A*

Community vote distribution

28.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 60 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 60
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team maintains a table of aggregate statistics through batch nightly updates. This includes total sales for the previous day alongside totals and averages for a variety of time periods including the 7 previous days, year-to-date, and quarter-to-date. This table is named store_saies_summary and the schema is as follows:

```
store_id INT, total_sales_qtd FLOAT, avg_daily_sales_qtd FLOAT,  total_sales_ytd FLOAT,
avg_daily_sales_ytd FLOAT, previous_day_sales FLOAT, total_sales_7d FLOAT, avg_daily_sales_7d
FLOAT, updated TIMESTAMP
```

The table daily_store_sales contains all the information needed to update store_sales_summary. The schema for this table is: store_id INT, sales_date DATE, total_sales FLOAT
If daily_store_sales is implemented as a Type 1 table and the total_sales column might be adjusted after manual data auditing, which approach is the safest to generate accurate reports in the store_sales_summary table?

A. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and overwrite the store_sales_summary table with each Update.

B. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and append new rows nightly to the store_sales_summary table.

C. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and use upsert logic to update results in the store_sales_summary table.

D. Implement the appropriate aggregate logic as a Structured Streaming read against the daily_store_sales table and use upsert logic to update results in the store_sales_summary table.

If daily_store_sales is implemented as a Type 1 table and the total_sales column might be adjusted after manual data auditing, which approach is the safest to generate accurate reports in the store_sales_summary table?

A. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and overwrite the store_sales_summary table with each Update.

B. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and append new rows nightly to the store_sales_summary table.

C. Implement the appropriate aggregate logic as a batch read against the daily_store_sales table and use upsert logic to update results in the store_sales_summary table. Most Voted

D. Implement the appropriate aggregate logic as a Structured Streaming read against the daily_store_sales table and use upsert logic to update results in the store_sales_summary table.

E. Use Structured Streaming to subscribe to the change data feed for daily_store_sales and apply changes to the aggregates in the store_sales_summary table with each update.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (49%) A (43%) 9%

29.

[All Certified Data Engineer Professional Questions]

A user new to Databricks is trying to troubleshoot long execution times for some pipeline logic they are working on. Presently, the user is executing code cell-by-cell, using display() calls to confirm code is producing the logically correct results as new transformations are added to an operation. To get a measure of average time to execute, the user is running each cell multiple times interactively.
Which of the following adjustments will get a more accurate measure of how code is likely to perform in production?

A. Scala is the only language that can be accurately tested using interactive notebooks; because the best performance is achieved by using Scala code compiled to JARs, all PySpark and Spark SQL logic should be refactored.

B. The only way to meaningfully troubleshoot code execution times in development notebooks Is to use production-sized data and production-sized clusters with Run All execution. Most Voted

C. Production code development should only be done using an IDE; executing code against a local build of open source Spark and Delta Lake will provide the most accurate benchmarks for how code will perform in production.

D. Calling display() forces a job to trigger, while many transformations will only add to the logical query plan; because of caching, repeated execution of the same logic does not provide meaningful results.

E. The Jobs UI should be leveraged to occasionally run the notebook as a job and track execution time during incremental code development because Photon can only be enabled on clusters launched for scheduled jobs.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (62%) D (38%)

30.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 1 DISCUSSION

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

An upstream system has been configured to pass the date for a given batch of data to the Databricks Jobs API as a parameter. The notebook to be scheduled will use this parameter to load data with the following code: df = spark.read.format("parquet").load(f"/mnt/source/(date)")
Which code block should be used to create the date Python variable used in the above code block?

A. date = spark.conf.get("date")

B. input_dict = input()
date= input_dict["date"]

C. import sys
date = sys.argv[1]

D. date = dbutils.notebooks.getParam("date")

E. dbutils.widgets.text("date", "null")
date = dbutils.widgets.get("date") Most Voted

Hide Answer

31.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 150 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 150
Topic #: 1
[All Certified Data Engineer Professional Questions]

A nightly job ingests data into a Delta Lake table using the following code:

```
from pyspark.sql.functions import current_timestamp, input_file_name,
from pyspark.sql.column import Column

def ingest_daily_batch(time_col: Column, year:int, month:int, day:int):
    (spark.read
        .format("parquet")
        .load(f"/mnt/daily_batch/{year}/{month}/{day}")
        .select("*",
                time_col.alias("ingest_time"),
                inpute_file_name().alias("source_file")
               )
        .write
        .mode("append")
        .saveAsTable("bronze")
    )
```

The next step in the pipeline requires a function that returns an object that can be used to manipulate new records that have not yet been processed to the next table in the pipeline.

```
        .load(f"/mnt/daily_batch/{year}/{month}/{day}")
        .select("*",
                time_col.alias("ingest_time"),
                inpute_file_name().alias("source_file")
               )
        .write
        .mode("append")
        .saveAsTable("bronze")
    )
```

The next step in the pipeline requires a function that returns an object that can be used to manipulate new records that have not yet been processed to the next the pipeline.

Which code snippet completes this function definition?

def new_records():

A. return spark.readStream.table("bronze") **Most Voted**

B. return spark.read.option("readChangeFeed", "true").table ("bronze")

C.
```
return (spark.read
    .table("bronze")
    .filter(col("ingest_time") == current_timestamp())
)
```

D.
```
return (spark.read
    .table("bronze")
    .filter(col("source_file") == f"/mnt/daily_batch/{year}/{month}/{day}")
)
```

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (42%) B (38%) D (21%)

32.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 48 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 48
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer, User A, has promoted a new pipeline to production by using the REST API to programmatically create several jobs. A DevOps engineer, User B, has configured an external orchestration tool to trigger job runs through the REST API. Both users authorized the REST API calls using their personal access tokens.
Which statement describes the contents of the workspace audit logs concerning these events?

A. Because the REST API was used for job creation and triggering runs, a Service Principal will be automatically used to identify these events.

B. Because User B last configured the jobs, their identity will be associated with both the job creation events and the job run events.

C. Because these events are managed separately, User A will have their identity associated with the job creation events and User B will have their identity associated with the job run events. Most Voted

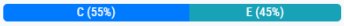
D. Because the REST API was used for job creation and triggering runs, user identity will not be captured in the audit logs.

E. Because User A created the jobs, their identity will be associated with both the job creation events and the job run events.

Hide Answer

Suggested Answer: *C*

Community vote distribution



33.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 212 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 212
Topic #: 1
[All Certified Data Engineer Professional Questions]

A team of data engineers are adding tables to a DLT pipeline that contain repetitive expectations for many of the same data quality checks. One member of the team suggests reusing these data quality rules across all tables defined for this pipeline.

What approach would allow them to do this?

A. Add data quality constraints to tables in this pipeline using an external job with access to pipeline configuration files.

B. Use global Python variables to make expectations visible across DLT notebooks included in the same pipeline.

C. Maintain data quality rules in a separate Databricks notebook that each DLT notebook or file can import as a library.

D. Maintain data quality rules in a Delta table outside of this pipeline's target schema, providing the schema name as a pipeline parameter. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

34.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 218 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 218
Topic #: 1
[All Certified Data Engineer Professional Questions]

A user wants to use DLT expectations to validate that a derived table report contains all records from the source, included in the table validation_copy.

The user attempts and fails to accomplish this by adding an expectation to the report table definition.

```
CREATE LIVE TABLE report (
  CONSTRAINT no_missing_records EXPECT (key IN validation_copy)
)
AS SELECT <...>
```

Which approach would allow using DLT expectations to validate all expected records are present in this table?

A. Define a temporary table that performs a left outer join on validation_copy and report, and define an expectation that no report key values are null Most Voted
B. Define a SQL UDF that performs a left outer join on two tables, and check if this returns null values for report key values in a DLT expectation for the report table
C. Define a view that performs a left outer join on validation_copy and report, and reference this view in DLT expectations for the report table
D. Define a function that performs a left outer join on validation_copy and report, and check against the result in a DLT expectation for the report table

Hide Answer

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (50%) C (50%)

35.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 202 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 202
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior data engineer on your team has implemented the following code block.

```
MERGE INTO events
USING new_events
ON events.event_id = new_events.event_id
WHEN NOT MATCHED
  INSERT *
```

The view new_events contains a batch of records with the same schema as the events Delta table. The event_id field serves as a unique key for this table.

When this query is executed, what will happen with new records that have the same event_id as an existing record?

A. They are merged.
B. They are ignored. Most Voted
C. They are updated.
D. They are inserted.

Hide Answer

36.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 155 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 155
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.

Hide Answer

Suggested Answer: *A*

Community vote distribution

37.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 169 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 169
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer, User A, has promoted a new pipeline to production by using the REST API to programmatically create several jobs. A DevOps engineer, User B, has configured an external orchestration tool to trigger job runs through the REST API. Both users authorized the REST API calls using their personal access tokens.

Which statement describes the contents of the workspace audit logs concerning these events?

A. Because the REST API was used for job creation and triggering runs, a Service Principal will be automatically used to identify these events.

B. Because User A created the jobs, their identity will be associated with both the job creation events and the job run events.

C. Because these events are managed separately, User A will have their identity associated with the job creation events and User B will have their identity associated with the job run events. Most Voted

D. Because the REST API was used for job creation and triggering runs, user identity will not be captured in the audit logs.

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

38.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 145 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 145
Topic #: 1
[All Certified Data Engineer Professional Questions]

A task orchestrator has been configured to run two hourly tasks. First, an outside system writes Parquet data to a directory mounted at /mnt/raw_orders/. After this data is written, a Databricks job containing the following code is executed:

```
(spark.readStream
  .format("parquet")
  .load("/mnt/raw_orders/")
  .withWatermark("time", "2 hours")
  .dropDuplicates(["customer_id", "order_id"])
  .writeStream
  .trigger(once=True)
  .table("orders")
)
```

Assume that the fields customer_id and order_id serve as a composite key to uniquely identify each order, and that the time field indicates when the record was queued in the source system.

Assume that the fields customer_id and order_id serve as a composite key to uniquely identify each order, and that the time field indicates when the record was queued in the source system.

If the upstream system is known to occasionally enqueue duplicate entries for a single order hours apart, which statement is correct?

A. Duplicate records enqueued more than 2 hours apart may be retained and the orders table may contain duplicate records with the same customer_id and order_id. Most Voted

B. All records will be held in the state store for 2 hours before being deduplicated and committed to the orders table.

C. The orders table will contain only the most recent 2 hours of records and no duplicates will be present.

D. The orders table will not contain duplicates, but records arriving more than 2 hours late will be ignored and missing from the table.

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (63%) D (38%)

39.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 220 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 220
Topic #: 1
[All Certified Data Engineer Professional Questions]

Where in the Spark UI can one diagnose a performance problem induced by not leveraging predicate push-down?

A. In the Executor's log file, by grepping for "predicate push-down"

B. In the Stage's Detail screen, in the Completed Stages table, by noting the size of data read from the Input column

C. In the Query Detail screen, by interpreting the Physical Plan Most Voted

D. In the Delta Lake transaction log. by noting the column statistics

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

40.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 115 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 115
Topic #: 1
[All Certified Data Engineer Professional Questions]

When using CLI or REST API to get results from jobs with multiple tasks, which statement correctly describes the response structure?

A. Each run of a job will have a unique job_id; all tasks within this job will have a unique job_id

B. Each run of a job will have a unique job_id; all tasks within this job will have a unique task_id

C. Each run of a job will have a unique orchestration_id; all tasks within this job will have a unique run_id

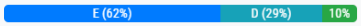
D. Each run of a job will have a unique run_id; all tasks within this job will have a unique task_id

E. Each run of a job will have a unique run_id; all tasks within this job will also have a unique run_id Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution



41.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 159 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 159
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. Foreign keys must reference a primary key field; multi-table inserts must leverage Delta Lake's upsert functionality.

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

Hide Answer

Suggested Answer: *D*

Community vote distribution

42.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 156 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 156
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 scan the parquet file footers to identify each row that meets 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

43.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 51 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 51
Topic #: 1
[All Certified Data Engineer Professional Questions]

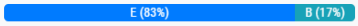
Where in the Spark UI can one diagnose a performance problem induced by not leveraging predicate push-down?

A. In the Executor's log file, by grepping for "predicate push-down"
B. In the Stage's Detail screen, in the Completed Stages table, by noting the size of data read from the Input column
C. In the Storage Detail screen, by noting which RDDs are not stored on disk
D. In the Delta Lake transaction log. by noting the column statistics
E. In the Query Detail screen, by interpreting the Physical Plan Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution



44.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 47 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 47
Topic #: 1
[All Certified Data Engineer Professional Questions]

What statement is true regarding 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

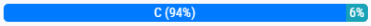
D. It is retained for 60 days, after which logs are archived

E. It is retained for 90 days or until the run-id is re-used through custom run configuration

Hide Answer

Suggested Answer: *C*

Community vote distribution



45.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 46 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 46
Topic #: 1
[All Certified Data Engineer Professional Questions]

Although the Databricks Utilities Secrets module provides tools to store sensitive credentials and avoid accidentally displaying them in plain text users should still be careful with which credentials are stored here and which users have access to using these secrets.
Which statement describes a limitation of Databricks Secrets?

A. Because the SHA256 hash is used to obfuscate stored secrets, reversing this hash will display the value in plain text.

B. Account administrators can see all secrets in plain text by logging on to the Databricks Accounts console.

C. Secrets are stored in an administrators-only table within the Hive Metastore; database administrators have permission to query this table by default.

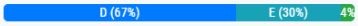
D. Iterating through a stored secret and printing each character will display secret contents in plain text. Most Voted

E. The Databricks REST API can be used to list secrets in plain text if the personal access token has proper credentials.

Hide Answer

Suggested Answer: *D*

Community vote distribution



46.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 43 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 43
Topic #: 1
[All Certified Data Engineer Professional Questions]

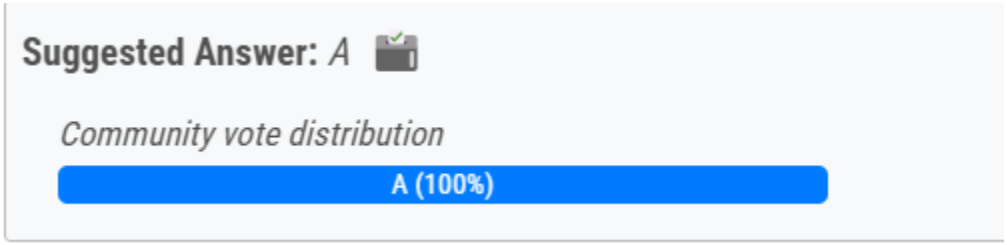
The data governance team has instituted a requirement that all tables containing Personal Identifiable Information (PH) must be clearly annotated. This includes adding column comments, table comments, and setting the custom table property "contains_pii" = true.
The following SQL DDL statement is executed to create a new table:

```
CREATE TABLE dev.pii_test
(id INT, name STRING COMMENT "PII")
COMMENT "Contains PII"
TBLPROPERTIES ('contains_pii' = True)
```

Which command allows manual confirmation that these three requirements have been met?

A. DESCRIBE EXTENDED dev.pii_test
B. DESCRIBE DETAIL dev.pii_test
C. SHOW TBLPROPERTIES dev.pii_test
D. DESCRIBE HISTORY dev.pii_test
E. SHOW TABLES dev

Show Suggested Answer



47.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 31 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 31
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. Human labor in writing code is the largest cost associated with data engineering workloads; as such, automating table declaration logic should be a priority in all migration workloads.

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

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

48.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 28 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 28
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_t...sin(["update_postimage", "insert"]))
  .write
  .mode("append")
  .table("bronze_history_type1")
 )
```

image14

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, the target table will be overwritten using the entire history of inserted or updated records, 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.

E. 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.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (92%) 8%

49.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 13 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 13
Topic #: 1
[All Certified Data Engineer Professional Questions]

An upstream system is emitting change data capture (CDC) logs that are being written to a cloud object storage directory. Each record in the log indicates the change type (insert, update, or delete) and the values for each field after the change. The source table has a primary key identified by the field pk_id.
For auditing purposes, the data governance team wishes to maintain a full record of all values that have ever been valid in the source system. For analytical purposes, only the most recent value for each record needs to be recorded. The Databricks job to ingest these records occurs once per hour, but each individual record may have changed multiple times over the course of an hour.
Which solution meets these requirements?

A. Create a separate history table for each pk_id resolve the current state of the table by running a union all filtering the history tables for the most recent state.

B. Use MERGE INTO to insert, update, or delete the most recent entry for each pk_id into a bronze table, then propagate all changes throughout the system.

C. Iterate through an ordered set of changes to the table, applying each in turn; rely on Delta Lake's versioning ability to create an audit log.

D. Use Delta Lake's change data feed to automatically process CDC data from an external system, propagating all changes to all dependent tables in the Lakehouse.

E. Ingest all log information into a bronze table; use MERGE INTO to insert, update, or delete the most recent entry for each pk_id into a silver table to recreate the current table state. Most Voted

50.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 35 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 35
Topic #: 1
[All Certified Data Engineer Professional Questions]

To reduce storage and compute costs, the data engineering team has been tasked with curating a series of aggregate tables leveraged by business intelligence dashboards, customer-facing applications, production machine learning models, and ad hoc analytical queries.
The data engineering team has been made aware of new requirements from a customer-facing application, which is the only downstream workload they manage entirely. As a result, an aggregate table used by numerous teams across the organization will need to have a number of fields renamed, and additional fields will also be added.
Which of the solutions addresses the situation while minimally interrupting other teams in the organization without increasing the number of tables that need to be managed?

A. Send all users notice that the schema for the table will be changing; include in the communication the logic necessary to revert the new table schema to match historic queries.

B. Configure a new table with all the requisite fields and new names and use this as the source for the customer-facing application; create a view that maintains the original data schema and table name by aliasing select fields from the new table.

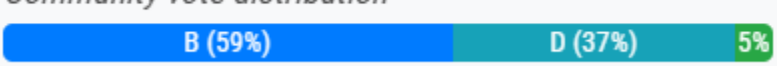
C. Create a new table with the required schema and new fields and use Delta Lake's deep clone functionality to sync up changes committed to one table to the corresponding table.

D. Replace the current table definition with a logical view defined with the query logic currently writing the aggregate table; create a new table to power the customer-facing application.

E. Add a table comment warning all users that the table schema and field names will be changing on a given date; overwrite the table in place to the specifications of the customer-facing application.

Suggested Answer: *B*

Community vote distribution



51.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 22 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 22
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement describes Delta Lake Auto Compaction?

A. 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.

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

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

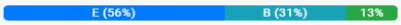
D. Data is queued in a messaging bus instead of committing data directly to memory; all data is committed from the messaging bus in one batch once the job is complete.

E. 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 128 MB. Most Voted

Hide Answer

Suggested Answer: *E*

Community vote distribution



52.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 4 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 4
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 recent_sensor_recordings table was unresponsive for three consecutive runs of the query

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

E. The average temperature recordings for at least one sensor exceeded 120 on three consecutive executions of the query Most Voted

53.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 27 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 27
Topic #: 1
[All Certified Data Engineer Professional Questions]

A junior data engineer on your team has implemented the following code block.

```
MERGE INTO events
USING new_events
ON events.event_id = new_events.event_id
WHEN NOT MATCHED
  INSERT *
```

The view new_events contains a batch of records with the same schema as the events Delta table. The event_id field serves as a unique key for this table.
When this query is executed, what will happen with new records that have the same event_id as an existing record?

A. They are merged.
B. They are ignored. Most Voted
C. They are updated.
D. They are inserted.
E. They are deleted.

Hide Answer

54.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 112 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 112
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

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

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

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

E. • Total VMs: 2
• 200 GB per Executor
• 80 Cores / Executor

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)