55.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 98 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 98
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

The user_ltv table has the following schema:

email STRING, age INT, ltv INT

The following view definition is executed:

The following view definition is executed:

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

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

SELECT * FROM user_ltv_no_minors

Which statement describes the results returned by this query?

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

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

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

D. All records from all columns will be displayed with the values in user_ltv.

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

Suggested Answer: *A*

Community vote distribution

56.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 106 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 106
Topic #: 1
[All Certified Data Engineer Professional Questions]

A nightly batch job is configured to ingest all data files from a cloud object storage container where records are stored in a nested directory structure YYYY/MM/DD. The data for each date represents all records that were processed by the source system on that date, noting that some records may be delayed as they await moderator approval. Each entry represents a user review of a product and has the following schema:

user_id STRING, review_id BIGINT, product_id BIGINT, review_timestamp TIMESTAMP, review_text STRING

The ingestion job is configured to append all data for the previous date to a target table reviews_raw with an identical schema to the source system. The next step in the pipeline is a batch write to propagate all new records inserted into reviews_raw to a table where data is fully deduplicated, validated, and enriched.

Which solution minimizes the compute costs to propagate this batch of data?

A. Perform a batch read on the reviews_raw table and perform an insert-only merge using the natural composite key user_id, review_id, product_id, review_timestamp.
Most Voted

B. Configure a Structured Streaming read against the reviews_raw table using the trigger once execution mode to process new records as a batch job.

C. Use Delta Lake version history to get the difference between the latest version of reviews_raw and one version prior, then write these records to the next table.

D. Filter all records in the reviews_raw table based on the review_timestamp; batch append those records produced in the last 48 hours.

E. Reprocess all records in reviews_raw and overwrite the next table in the pipeline.

57.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 8 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 8
Topic #: 1
[All Certified Data Engineer Professional Questions]

An upstream source writes Parquet data as hourly batches to directories named with the current date. A nightly batch job runs the following code to ingest all data from the previous day as indicated by the date variable:

```
(spark.read
  .format("parquet")
  .load(f"/mnt/raw_orders/{date}")
  .dropDuplicates(["customer_id", "order_id"])
  .write
  .mode("append")
  .saveAsTable("orders")
)
```

Assume that the fields customer_id and order_id serve as a composite key to uniquely identify each order.
If the upstream system is known to occasionally produce duplicate entries for a single order hours apart, which statement is correct?

A. Each write to the orders table will only contain unique records, and only those records without duplicates in the target table will be written.

B. Each write to the orders table will only contain unique records, but newly written records may have duplicates already present in the target table.

C. Each write to the orders table will only contain unique records; if existing records with the same key are present in the target table, these records will be overwritten.

D. Each write to the orders table will only contain unique records; if existing records with the same key are present in the target table, the operation will fail.

E. Each write to the orders table will run deduplication over the union of new and existing records, ensuring no duplicate records are present.

Assume that the fields customer_id and order_id serve as a composite key to uniquely identify each order.
If the upstream system is known to occasionally produce duplicate entries for a single order hours apart, which statement is correct?

A. Each write to the orders table will only contain unique records, and only those records without duplicates in the target table will be written.

B. Each write to the orders table will only contain unique records, but newly written records may have duplicates already present in the target table. Most Voted

C. Each write to the orders table will only contain unique records; if existing records with the same key are present in the target table, these records will be overwritten.

D. Each write to the orders table will only contain unique records; if existing records with the same key are present in the target table, the operation will fail.

E. Each write to the orders table will run deduplication over the union of new and existing records, ensuring no duplicate records are present.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

58.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 114 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 114
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data team's Structured Streaming job is configured to calculate running aggregates for item sales to update a downstream marketing dashboard. The marketing team has introduced a new promotion, and they would like to add a new field to track the number of times this promotion code is used for each item. A junior data engineer suggests updating the existing query as follows. Note that proposed changes are in bold.

Original query:

```
df.groupBy("item")
    .agg(count("item").alias("total_count"),
         mean("sale_price").alias("avg_price"))
    .writeStream
    .outputMode("complete")
    .option("checkpointLocation", "/item_agg/__checkpoint")
    .start("/item_agg")
```

Proposed query:

```
df.groupBy("item")
    .agg(count("item").alias("total_count"),
         mean("sale_price").alias("avg_price"),
         count("promo_code = 'NEW_MEMBER'").alias("new_member_promo"))
    .writeStream
    .outputMode("complete")
    .option('mergeSchema', 'true')
    .option("checkpointLocation", "/item_agg/__checkpoint")
    .start("/item_agg")
```

Which step must also be completed to put the proposed query into production?

A. Specify a new checkpointLocation **Most Voted**

B. Remove .option('mergeSchema', 'true') from the streaming write

C. Increase the shuffle partitions to account for additional aggregates

D. Run REFRESH TABLE delta.'/item_agg'

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

59.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 39 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 39
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which of the following is true of Delta Lake and the Lakehouse?

A. Because Parquet compresses data row by row. strings will only be compressed when a character is repeated multiple times.

B. Delta Lake automatically collects statistics on the first 32 columns of each table which are leveraged in data skipping based on query filters. **Most Voted**

C. Views in the Lakehouse maintain a valid cache of the most recent versions of source tables at all times.

D. Primary and foreign key constraints can be leveraged to ensure duplicate values are never entered into a dimension table.

E. Z-order can only be applied to numeric values stored in Delta Lake tables.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (89%) 11%

60.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 180
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?

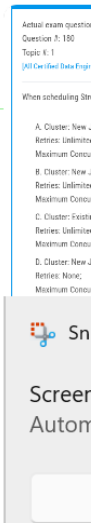
A. Cluster: New Job Cluster;
Retries: Unlimited;
Maximum Concurrent Runs: 1

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

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

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

Hide Answer



Answer: A

61.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 149 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 149
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data pipeline uses Structured Streaming to ingest data from Apache Kafka to Delta Lake. Data is being stored in a bronze table, and includes the Kafka-generated timestamp, key, and value. Three months after the pipeline is deployed, the data engineering team has noticed some latency issues during certain times of the day.

A senior data engineer updates the Delta Table's schema and ingestion logic to include the current timestamp (as recorded by Apache Spark) as well as the Kafka topic and partition. The team plans to use these additional metadata fields to diagnose the transient processing delays.

Which limitation will the team face while diagnosing this problem?

A. New fields will not be computed for historic records.

B. Spark cannot capture the topic and partition fields from a Kafka source.

C. Updating the table schema requires a default value provided for each field added.

D. Updating the table schema will invalidate the Delta transaction log metadata.

Hide Answer

Suggested Answer: *A*

62.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 138
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:

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")
```

Which line of code 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")
C. "event_time"
D. lag("event_time", "10 minutes").alias("time")

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

63.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 136 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 136
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data ingestion task requires a one-TB JSON dataset to be written out to Parquet with a target part-file size of 512 MB. Because Parquet is being used instead of Delta Lake, built-in file-sizing features such as Auto-Optimize & Auto-Compaction cannot be used.

Which strategy will yield the best performance without shuffling data?

A. Set spark.sql.files.maxPartitionBytes to 512 MB, ingest the data, execute the narrow transformations, and then write to parquet. Most Voted

B. Set spark.sql.shuffle.partitions to 2,048 partitions (1TB*1024*1024/512), ingest the data, execute the narrow transformations, optimize the data by sorting it (which automatically repartitions the data), and then write to parquet.

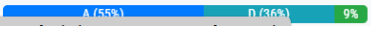
C. Set spark.sql.adaptive.advisoryPartitionSizeInBytes to 512 MB bytes, ingest the data, execute the narrow transformations, coalesce to 2,048 partitions (1TB*1024*1024/512), and then write to parquet.

D. Ingest the data, execute the narrow transformations, repartition to 2,048 partitions (1TB* 1024*1024/512), and then write to parquet.

Hide Answer

Suggested Answer: *A*

Community vote distribution



64.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 134 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 134
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Delta lake table with CDF enabled 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. 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.

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.

Answer: c

65.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 133 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 133
Topic #: 1
[All Certified Data Engineer Professional Questions]

The business intelligence team has a dashboard configured to track various summary metrics for retail stores. This includes total sales for the previous day alongside totals and averages for a variety of time periods. The fields required to populate this dashboard have the following schema:

```
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
```

For demand forecasting, the Lakehouse contains a validated table of all itemized sales updated incrementally in near real-time. This table, named products_per_order, includes the following fields:

```
store_id INT, order_id INT, product_id INT, quantity INT, price FLOAT,
order_timestamp TIMESTAMP
```

Because reporting on long-term sales trends is less volatile, analysts using the new dashboard only require data to be refreshed once daily. Because the dashboard will be queried interactively by many users throughout a normal business day, it should return results quickly and reduce total compute associated with each materialization.

Which solution meets the expectations of the end users while controlling and limiting possible costs?

Which solution meets the expectations of the end users while controlling and limiting possible costs?

A. Populate the dashboard by configuring a nightly batch job to save the required values as a table overwritten with each update.

B. Use Structured Streaming to configure a live dashboard against the products_per_order table within a Databricks notebook.

C. Define a view against the products_per_order table and define the dashboard against this view.

D. Use the Delta Cache to persist the products_per_order table in memory to quickly update the dashboard with each query.

Show Suggested Answer

Answer: A

66.

Topic #: 1

[All Certified Data Engineer Professional Questions]

An upstream source writes Parquet data as hourly batches to directories named with the current date. A nightly batch job runs the following code to ingest all data from the previous day as indicated by the date variable:

```
(spark.read
  .format("parquet")
  .load(f"/mnt/raw_orders/{date}")
  .dropDuplicates(["customer_id", "order_id"])
  .write
  .mode("append")
  .saveAsTable("orders")
)
```

Assume that the fields customer_id and order_id serve as a composite key to uniquely identify each order.

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

A. Each write to the orders table will only contain unique records, and only those records without duplicates in the target table will be written.

B. Each write to the orders table will only contain unique records, but newly written records may have duplicates already present in the target table.

C. Each write to the orders table will only contain unique records; if existing records with the same key are present in the target table, these records will be overwritten.

D. Each write to the orders table will run deduplication over the union of new and existing records, ensuring no duplicate records are present.

Hide Answer

Suggested Answer: B

Answer: B

67.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 76 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 76
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data pipeline uses Structured Streaming to ingest data from Apache Kafka to Delta Lake. Data is being stored in a bronze table, and includes the Kafka-generated timestamp, key, and value. Three months after the pipeline is deployed, the data engineering team has noticed some latency issues during certain times of the day.

A senior data engineer updates the Delta Table's schema and ingestion logic to include the current timestamp (as recorded by Apache Spark) as well as the Kafka topic and partition. The team plans to use these additional metadata fields to diagnose the transient processing delays.

Which limitation will the team face while diagnosing this problem?

A. New fields will not be computed for historic records. Most Voted

B. Spark cannot capture the topic and partition fields from a Kafka source.

C. New fields cannot be added to a production Delta table.

D. Updating the table schema will invalidate the Delta transaction log metadata.

E. Updating the table schema requires a default value provided for each field added.

68.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 142
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

A. spark.sql.files.maxPartitionBytes **Most Voted**

B. spark.sql.autoBroadcastJoinThreshold

C. spark.sql.adaptive.advisoryPartitionSizeInBytes

D. spark.sql.adaptive.coalescePartitions.minPartitionNum

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

69.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 37
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 leverages user workstations as the driver during interactive development; as such, users should always use a workspace deployed in a region they are physically near.

E. 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

70.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 126 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 126
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"
```

What will be the outcome of executing these command cells m order m an interactive notebook?

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"
```

What will be the outcome of executing these command cells m order m 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. Cmd 1 will succeed and Cmd 2 will fail. countries_af will be a Python variable containing a list of strings. Most Voted

Hide Answer

Suggested Answer: *D*

71.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 116 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 116
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. As long as code in the development environment declares USE dev_db at the top of each notebook, there is no possibility of inadvertently committing changes back to production data sources.

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.

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. Most Voted

C. As long as code in the development environment declares USE dev_db at the top of each notebook, there is no possibility of inadvertently committing changes back to production data sources.

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.

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

Hide Answer

Suggested Answer: *B*

Community vote distribution

72.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 99 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 99
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'
    )
""")
```


```
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. Most Voted

C. Yes; the change data feed uses foreign keys to ensure delete consistency throughout the Lakehouse.

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

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

Hide Answer

Suggested Answer: *B*

73.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 97 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 97
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

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

Which piece of information is critical to this decision?

A. Data corruption can occur if a query fails in a partially completed state because Type 2 tables require setting multiple fields in a single update.
B. Shallow clones can be combined with Type 1 tables to accelerate historic queries for long-term versioning.
C. Delta Lake time travel cannot be used to query previous versions of these tables because Type 1 changes modify data files in place.
D. Delta Lake time travel does not scale well in cost or latency to provide a long-term versioning solution. **Most Voted**
E. Delta Lake only supports Type 0 tables; once records are inserted to a Delta Lake table, they cannot be modified.

Hide Answer

74.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 93 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 93
Topic #: 1
[All Certified Data Engineer Professional Questions]

You are performing a join operation to combine values from a static userLookup table with a streaming DataFrame streamingDF.

Which code block attempts to perform an invalid stream-static join?

A. userLookup.join(streamingDF, ["userid"], how="inner")
B. streamingDF.join(userLookup, ["user_id"], how="outer") **Most Voted**
C. streamingDF.join(userLookup, ["user_id"], how="left")
D. streamingDF.join(userLookup, ["userid"], how="inner")
E. userLookup.join(streamingDF, ["user_id"], how="right")

Hide Answer

Suggested Answer: *B*

Community vote distribution

75.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 75 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 75
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

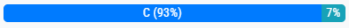
In addition to de-duplicating records within the batch, which of the following approaches allows the data engineer to deduplicate data against previously processed records as it is inserted into a Delta table?

A. Set the configuration delta.deduplicate = true.
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.
E. Rely on Delta Lake schema enforcement to prevent duplicate records.

Hide Answer

Suggested Answer: *C*

Community vote distribution



76.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 110 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 110
Topic #: 1
[All Certified Data Engineer Professional Questions]

A large company seeks to implement a near real-time solution involving hundreds of pipelines with parallel updates of many tables with extremely high volume and high velocity data.

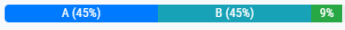
Which of the following solutions would you implement to achieve this requirement?

A. Use Databricks High Concurrency clusters, which leverage optimized cloud storage connections to maximize data throughput. Most Voted
B. Partition ingestion tables by a small time duration to allow for many data files to be written in parallel.
C. Configure Databricks to save all data to attached SSD volumes instead of object storage, increasing file I/O significantly.
D. Isolate Delta Lake tables in their own storage containers to avoid API limits imposed by cloud vendors.
E. Store all tables in a single database to ensure that the Databricks Catalyst Metastore can load balance overall throughput.

Hide Answer

Suggested Answer: *A*

Community vote distribution



77.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 45 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional

Question #: 45

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

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

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

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

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

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

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (68%) E (24%) 8%

78.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 226 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 226
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Data Engineer wants to run unit tests using common Python testing frameworks on Python functions defined across several Databricks notebooks currently used in production.

How can the data engineer run unit tests against functions that work with data in production?

A. Define and import unit test functions from a separate Databricks notebook
B. Define and unit test functions using Files in Repos
C. Run unit tests against non-production data that closely mirrors production Most Voted
D. Define unit tests and functions within the same notebook

Hide Answer

Suggested Answer: *B*

Community vote distribution

79.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 225 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 225
Topic #: 1
[All Certified Data Engineer Professional Questions]

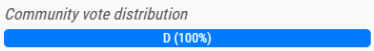
Which REST API call can be used to review the notebooks configured to run as tasks in a multi-task job?

A. /jobs/runs/list
B. /jobs/list
C. /jobs/runs/get
D. /jobs/get Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution



80.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 130 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional

Question #: 130

Topic #: 1

[All Certified Data Engineer Professional Questions]

The following table consists of items found in user carts within an e-commerce website.

```
Carts (id LONG, items ARRAY<STRUCT<id: LONG, count: INT>>)
id  items                                          email
1001[{id: "DESK65", count: 1}]                     "u1@domain.com"
1002[{id: "KYBD45", count: 1}, {id: "M27", count: 2}]"u2@domain.com"
1003[{id: "M27", count: 1}]                        "u3@domain.com"
```

The following MERGE statement is used to update this table using an updates view, with schema evolution enabled on this table.

```
MERGE INTO carts c
USING updates u
ON c.id = u.id
WHEN MATCHED
  THEN UPDATE SET *
```

How would the following update be handled?

The following MERGE statement is used to update this table using an updates view, with schema evolution enabled on this table.

```
MERGE INTO carts c
USING updates u
ON c.id = u.id
WHEN MATCHED
  THEN UPDATE SET *
```

How would the following update be handled?

```
(new nested field, missing existing column)
id  items
1001[{id: "DESK65", count: 2, coupon: "BOGO50"}]
```

A. The update throws an error because changes to existing columns in the target schema are not supported.

B. The new nested Field is added to the target schema, and dynamically read as NULL for existing unmatched records.

C. The update is moved to a separate "rescued" column because it is missing a column expected in the target schema.

D. The new nested field is added to the target schema, and files underlying existing records are updated to include NULL values for the new field.

Show Suggested Answer

Answer: B

81.

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 198
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement describes the default execution mode for Databricks Auto Loader?

A. Cloud vendor-specific queue storage and notification services are configured to track newly arriving files; new files are incrementally and idempotently loaded into the target Delta Lake table.

B. New files are identified by listing the input directory; the target table is materialized by directly querying all valid files in the source directory.

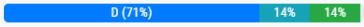
C. Webhooks trigger a Databricks job to run anytime new data arrives in a source directory; new data are automatically merged into target tables using rules inferred from the data.

D. New files are identified by listing the input directory; new files are incrementally and idempotently loaded into the target Delta Lake table. Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution



82.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 194 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 194
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data engineer is performing a join operation to combine values from a static userLookup table with a streaming DataFrame streamingDF.

Which code block attempts to perform an invalid stream-static join?

A. userLookup.join(streamingDF, ["user_id"], how="right")

B. streamingDF.join(userLookup, ["user_id"], how="inner")

C. userLookup.join(streamingDF, ["user_id"), how="inner")

D. userLookup.join(streamingDF, ["user_id"], how="left") Most Voted

Hide Answer

Suggested Answer: *D*

Community vote distribution

83.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 193 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 193
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?

Both users and orders are Delta Lake tables.

Which statement describes the results of querying recent_orders?

A. The versions of each source table will be stored in the table transaction log; query results will be saved to DBFS with each query.

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. 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 (80%) B (20%)

84.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 137 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 137
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

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

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. Most Voted

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.

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (83%) A (17%)

85.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 120 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 120
Topic #: 1
[All Certified Data Engineer Professional Questions]

The business reporting team requires that data for their dashboards be updated every hour. The total processing time for the pipeline that extracts, transforms, and loads the data for their pipeline runs in 10 minutes.

Assuming normal operating conditions, which configuration will meet their service-level agreement requirements with the lowest cost?

A. Configure a job that executes every time new data lands in a given directory

B. Schedule a job to execute the pipeline once an hour on a new job cluster

C. Schedule a Structured Streaming job with a trigger interval of 60 minutes

D. Schedule a job to execute the pipeline once an hour on a dedicated interactive cluster

Hide Answer

Suggested Answer: *B*

Community vote distribution

B (100%)

86.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 118 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 118
Topic #: 1
[All Certified Data Engineer Professional Questions]

A member of the data engineering team has submitted a short notebook that they wish to schedule as part of a larger data pipeline. Assume that the commands provided below produce the logically correct results when run as presented.

```
Cmd 1
rawDF = spark.table("raw_data")
Cmd 2
rawDF.printSchema()
Cmd 3
flattenedDF = rawDF.select("*", "values.*")
Cmd 4
finalDF = flattenedDF.drop("values")
```

```
Cmd 3
flattenedDF = rawDF.select("*", "values.*")
Cmd 4
finalDF = flattenedDF.drop("values")
Cmd 5
display(finalDF)
Cmd 6
finalDF.write.mode("append").saveAsTable("flat_data")
```

Which command should be removed from the notebook before scheduling it as a job?

A. Cmd 2
B. Cmd 3
C. Cmd 4
D. Cmd 5 **Most Voted**

Hide Answer

Suggested Answer: *D*

87.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 113 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 113
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Delta Lake 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.

Immediately after each update succeeds, the data engineering team would like to determine the difference between the new version and the previous version of the table.

Given the current implementation, which method can be used?

A. Execute a query to calculate the difference between the new version and the previous version using Delta Lake's built-in versioning and lime travel functionality. Most Voted

B. Parse the Delta Lake transaction log to identify all newly written data files.

C. Parse the Spark event logs to identify those rows that were updated, inserted, or deleted.

D. Execute DESCRIBE HISTORY customer_churn_params to obtain the full operation metrics for the update, including a log of all records that have been added or modified.

E. Use Delta Lake's change data feed to identify those records that have been updated, inserted, or deleted.

88.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 108 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 108
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement describes the default execution mode for Databricks Auto Loader?

A. Cloud vendor-specific queue storage and notification services are configured to track newly arriving files; the target table is materialized by directly querying all valid files in the source directory.

B. New files are identified by listing the input directory; the target table is materialized by directly querying all valid files in the source directory.

C. Webhooks trigger a Databricks job to run anytime new data arrives in a source directory; new data are automatically merged into target tables using rules inferred from the data.

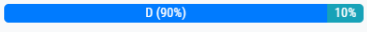
D. New files are identified by listing the input directory; new files are incrementally and idempotently loaded into the target Delta Lake table. Most Voted

E. Cloud vendor-specific queue storage and notification services are configured to track newly arriving files; new files are incrementally and idempotently loaded into the target Delta Lake table.

Hide Answer

Suggested Answer: *D*

Community vote distribution



89.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 103 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 103
Topic #: 1
[All Certified Data Engineer Professional Questions]

Which statement describes a key benefit of an end-to-end test?

A. Makes it easier to automate your test suite

B. Pinpoints errors in the building blocks of your application

C. Provides testing coverage for all code paths and branches

D. Closely simulates real world usage of your application **Most Voted**

This answer is currently the most voted for in the discussion

E. Ensures code is optimized for a real-life workflow

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (86%) 14%

90.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 100 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 100
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

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

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

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

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

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

D. "Manage" permissions should be set on a secret scope containing only those credentials that will be used by a given team.
No additional configuration is necessary as long as all users are configured as administrators in the workspace where secrets have been added.

91.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 70 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 70
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data ingestion task requires a one-TB JSON dataset to be written out to Parquet with a target part-file size of 512 MB. Because Parquet is being used instead of Delta Lake, built-in file-sizing features such as Auto-Optimize & Auto-Compaction cannot be used.

Which strategy will yield the best performance without shuffling data?

A. Set spark.sql.files.maxPartitionBytes to 512 MB, ingest the data, execute the narrow transformations, and then write to parquet. Most Voted

B. Set spark.sql.shuffle.partitions to 2,048 partitions (1TB*1024*1024/512), ingest the data, execute the narrow transformations, optimize the data by sorting it (which automatically repartitions the data), and then write to parquet.

C. Set spark.sql.adaptive.advisoryPartitionSizeInBytes to 512 MB bytes, ingest the data, execute the narrow transformations, coalesce to 2,048 partitions (1TB*1024*1024/512), and then write to parquet.

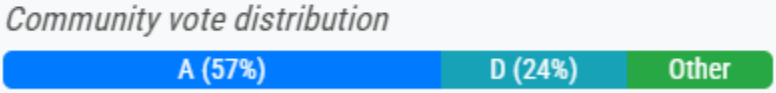
D. Ingest the data, execute the narrow transformations, repartition to 2,048 partitions (1TB* 1024*1024/512), and then write to parquet.

E. Set spark.sql.shuffle.partitions to 512, ingest the data, execute the narrow transformations, and then write to parquet.

Hide Answer

Suggested Answer: *A*

Community vote distribution



92.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 68 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 68
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

A. configure

B. fs Most Voted

C. jobs

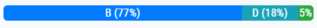
D. libraries

E. workspace

Hide Answer

Suggested Answer: *B*

Community vote distribution



93.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 63 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 63
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Databricks SQL dashboard has been configured to monitor the total number of records present in a collection of Delta Lake tables using the following query pattern:

SELECT COUNT (*) FROM table -

Which of the following describes how results are generated each time the dashboard is updated?

A. The total count of rows is calculated by scanning all data files
B. The total count of rows will be returned from cached results unless REFRESH is run
C. The total count of records is calculated from the Delta transaction logs Most Voted
D. The total count of records is calculated from the parquet file metadata
E. The total count of records is calculated from the Hive metastore

Hide Answer

Suggested Answer: *C*

94.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 62 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 62
Topic #: 1
[All Certified Data Engineer Professional Questions]

The business reporting team requires that data for their dashboards be updated every hour. The total processing time for the pipeline that extracts transforms, and loads the data for their pipeline runs in 10 minutes.

Assuming normal operating conditions, which configuration will meet their service-level agreement requirements with the lowest cost?

A. Manually trigger a job anytime the business reporting team refreshes their dashboards
B. Schedule a job to execute the pipeline once an hour on a new job cluster Most Voted
C. Schedule a Structured Streaming job with a trigger interval of 60 minutes
D. Schedule a job to execute the pipeline once an hour on a dedicated interactive cluster
E. Configure a job that executes every time new data lands in a given directory

Hide Answer

Suggested Answer: *B*

Community vote distribution
B (87%) 13%

95.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 61 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 61
Topic #: 1
[All Certified Data Engineer Professional Questions]

A member of the data engineering team has submitted a short notebook that they wish to schedule as part of a larger data pipeline. Assume that the commands provided below produce the logically correct results when run as presented.

Cmd 1
```
rawDF = spark.table("raw_data")
```

Cmd 2
```
rawDF.printSchema()
```

Cmd 3
```
flattenedDF = rawDF.select("*", "values.*")
```

Cmd 4
```
finalDF = flattenedDF.drop("values")
```

Cmd 5
```
finalDF.explain()
```

Cmd 4
```
finalDF = flattenedDF.drop("values")
```

Cmd 5
```
finalDF.explain()
```

Cmd 6
```
display(finalDF)
```

Cmd 7
```
finalDF.write.mode("append").saveAsTable("flat_data")
```

Which command should be removed from the notebook before scheduling it as a job?

A. Cmd 2
B. Cmd 3
C. Cmd 4
D. Cmd 5
E. Cmd 6 Most Voted

Hide Answer

96.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 59 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 59
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Delta Lake table was created with the below query:

```
CREATE TABLE prod.sales_by_stor
USING DELTA
LOCATION "/mnt/prod/sales_by_store"
```

Realizing that the original query had a typographical error, the below code was executed:
ALTER TABLE prod.sales_by_stor RENAME TO prod.sales_by_store
Which result will occur after running the second command?

A. The table reference in the metastore is updated and no data is changed. Most Voted

B. The table name change is recorded in the Delta transaction log.

C. All related files and metadata are dropped and recreated in a single ACID transaction.

D. The table reference in the metastore is updated and all data files are moved.

E. A new Delta transaction log Is created for the renamed table.

Hide Answer

Suggested Answer: *A*

97.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 58 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 58
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 tasks A and B complete successfully but task C fails during a scheduled run, which statement describes the resulting state?

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

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

C. All logic expressed in the notebook associated with task A will have been successfully completed; tasks B and C will not commit any changes because of stage failure.

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

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

Hide Answer

Suggested Answer: *A*

98.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 91 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 91
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

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

A. Use Repos to checkout all changes and send the git diff log to the team.

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

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

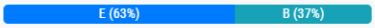
D. Use Repos to merge all differences and make a pull request back to the remote repository.

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

Hide Answer

Suggested Answer: *E*

Community vote distribution



99.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 90 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 90
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. Spark configuration properties can only be set for an interactive cluster by creating a global init script.

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

E. 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.

Hide Answer

Suggested Answer: *D*

Community vote distribution

100.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 33 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 33
Topic #: 1
[All Certified Data Engineer Professional Questions]

The data engineering team is migrating an enterprise system with thousands of tables and views into the Lakehouse. They plan to implement the target architecture using a series of bronze, silver, and gold tables. Bronze tables will almost exclusively be used by production data engineering workloads, while silver tables will be used to support both data engineering and machine learning workloads. Gold tables will largely serve business intelligence and reporting purposes. While personal identifying information (PII) exists in all tiers of data, pseudonymization and anonymization rules are in place for all data at the silver and gold levels.
The organization is interested in reducing security concerns while maximizing the ability to collaborate across diverse teams.
Which statement exemplifies best practices for implementing this system?

A. Isolating tables in separate databases based on data quality tiers allows for easy permissions management through database ACLs and allows physical separation of default storage locations for managed tables.

B. Because databases on Databricks are merely a logical construct, choices around database organization do not impact security or discoverability in the Lakehouse.

C. Storing all production tables in a single database provides a unified view of all data assets available throughout the Lakehouse, simplifying discoverability by granting all users view privileges on this database.

D. Working in the default Databricks database provides the greatest security when working with managed tables, as these will be created in the DBFS root.

E. Because all tables must live in the same storage containers used for the database they're created in, organizations should be prepared to create between dozens and thousands of databases depending on their data isolation requirements.

Show Suggested Answer

Answer: A

101.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 32 DISCUSSION

The data engineering team maintains the following code:

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

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

    itemDF.itemName))

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

    accountDF.city))

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

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

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

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

C. An incremental job will leverage information in the state store to identify unjoined rows in the source tables and write these rows to the enriched_iteinized_orders_by_account table.

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

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

102.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 25 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 25
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

E. Credential validation errors while pulling data from an external system.

Hide Answer

Suggested Answer: *D*

Community vote distribution

D (100%)

103.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 23 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 23
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

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

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

C. Structured Streaming uses specialized hardware and I/O streams to achieve sub-second latency for data transfer.

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

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

Hide Answer

Suggested Answer: *D*

104.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 21 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 21
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Structured Streaming job deployed to production has been experiencing delays during peak hours of the day. At present, during normal execution, each microbatch of data is processed in less than 3 seconds. During peak hours of the day, execution time for each microbatch becomes very inconsistent, sometimes exceeding 30 seconds. The streaming write is currently configured with a trigger interval of 10 seconds.
Holding all other variables constant and assuming records need to be processed in less than 10 seconds, which adjustment will meet the requirement?

A. Decrease the trigger interval to 5 seconds; triggering batches more frequently allows idle executors to begin processing the next batch while longer running tasks from previous batches finish.

B. Increase the trigger interval to 30 seconds; setting the trigger interval near the maximum execution time observed for each batch is always best practice to ensure no records are dropped.

C. The trigger interval cannot be modified without modifying the checkpoint directory; to maintain the current stream state, increase the number of shuffle partitions to maximize parallelism.

D. Use the trigger once option and configure a Databricks job to execute the query every 10 seconds; this ensures all backlogged records are processed with each batch.

E. Decrease the trigger interval to 5 seconds; triggering batches more frequently may prevent records from backing up and large batches from causing spill. Most Voted

Hide Answer

105.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 12 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 12
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. Three new jobs named "Ingest new data" will be defined in the workspace, and they will each run once daily.
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.
E. The logic defined in the referenced notebook will be executed three times on the referenced existing all purpose cluster.

Hide Answer

Suggested Answer: *C*

106.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 69 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 69
Topic #: 1
[All Certified Data Engineer Professional Questions]

The business intelligence team has a dashboard configured to track various summary metrics for retail stores. This includes total sales for the previous day alongside totals and averages for a variety of time periods. The fields required to populate this dashboard have the following schema:

```
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
```

For demand forecasting, the Lakehouse contains a validated table of all itemized sales updated incrementally in near real-time. This table, named products_per_order, includes the following fields:

```
store_id INT, order_id INT, product_id INT, quantity INT, price FLOAT,
order_timestamp TIMESTAMP
```

Because reporting on long-term sales trends is less volatile, analysts using the new dashboard only require data to be refreshed once daily. Because the dashboard will be queried interactively by many users throughout a normal business day, it should return results quickly and reduce total compute associated with each materialization.

Which solution meets the expectations of the end users while controlling and limiting possible costs?

Because reporting on long-term sales trends is less volatile, analysts using the new dashboard only require data to be refreshed once daily. Because the dashboard will be queried interactively by many users throughout a normal business day, it should return results quickly and reduce total compute associated with each materialization.

Which solution meets the expectations of the end users while controlling and limiting possible costs?

A. Populate the dashboard by configuring a nightly batch job to save the required values as a table overwritten with each update. Most Voted

B. Use Structured Streaming to configure a live dashboard against the products_per_order table within a Databricks notebook.

C. Configure a webhook to execute an incremental read against products_per_order each time the dashboard is refreshed.

D. Use the Delta Cache to persist the products_per_order table in memory to quickly update the dashboard with each query.

E. Define a view against the products_per_order table and define the dashboard against this view.

Hide Answer

Suggested Answer: *A*

107.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 121 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 121
Topic #: 1
[All Certified Data Engineer Professional Questions]

A Databricks SQL dashboard has been configured to monitor the total number of records present in a collection of Delta Lake tables using the following query pattern:

SELECT COUNT (*) FROM table -

Which of the following describes how results are generated each time the dashboard is updated?

A. The total count of rows is calculated by scanning all data files

B. The total count of rows will be returned from cached results unless REFRESH is run

C. The total count of records is calculated from the Delta transaction logs Most Voted

D. The total count of records is calculated from the parquet file metadata

Hide Answer

Suggested Answer: *C*

Community vote distribution

C (100%)

108.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 72 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 72
Topic #: 1
[All Certified Data Engineer Professional Questions]

A data team's Structured Streaming job is configured to calculate running aggregates for item sales to update a downstream marketing dashboard. The marketing team has introduced a new promotion, and they would like to add a new field to track the number of times this promotion code is used for each item. A junior data engineer suggests updating the existing query as follows. Note that proposed changes are in bold.

Original query:

```
df.groupBy("item")
    .agg(count("item").alias("total_count"),
         mean("sale_price").alias("avg_price"))
    .writeStream
    .outputMode("complete")
    .option("checkpointLocation", "/item_agg/__checkpoint")
    .start("/item_agg")
```

Proposed query:

```
df.groupBy("item")
    .agg(count("item").alias("total_count"),
         mean("sale_price").alias("avg_price"))
    .writeStream
    .outputMode("complete")
    .option("checkpointLocation", "/item_agg/__checkpoint")
    .start("/item_agg")
```

Proposed query:

.start("/item_agg")

Proposed query:

.start("/item_agg")

Which step must also be completed to put the proposed query into production?

A. Specify a new checkpointLocation **Most Voted**

B. Increase the shuffle partitions to account for additional aggregates

C. Run REFRESH TABLE delta.'/item_agg'

D. Register the data in the "/item_agg" directory to the Hive metastore

E. Remove .option('mergeSchema', 'true') from the streaming write

Hide Answer

Suggested Answer: *A*

Community vote distribution

A (100%)

109.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 73 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 73
Topic #: 1
[All Certified Data Engineer Professional Questions]

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

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

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

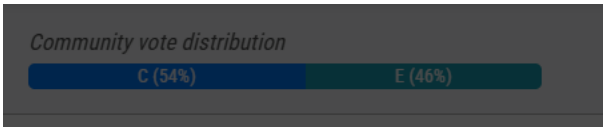
B. Increase the number of shuffle partitions to maximize parallelism, since the trigger interval cannot be modified without modifying the checkpoint directory.

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

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

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

Hide Answer



110.

EXAM CERTIFIED DATA ENGINEER PROFESSIONAL TOPIC 1 QUESTION 186 DISCUSSION

Actual exam question from Databricks's Certified Data Engineer Professional
Question #: 186
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. Cluster creation allowed, "Can Restart" 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

Hide Answer

.