DATABRICKS-CERTIFIED-PR OFESSIONAL-DATA-ENGINEER Q&As

Databricks Certified Professional Data Engineer Exam

Pass Databricks DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER Exam with 100% Guarantee

Free Download Real Questions & Answers **PDF** and **VCE** file from:

https://www.pass4itsure.com/databricks-certified-professional-data-engineer.html

100% Passing Guarantee
100% Money Back Assurance

Following Questions and Answers are all new published by Databricks Official Exam Center

QUESTION 1

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 fieldpk_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.

Correct Answer: E

Explanation: This is the correct answer because it meets the requirements of maintaining a full record of all values that have ever been valid in the source system and recreating the current table state with only the most recent value for each record. The code ingests all log information into a bronze table, which preserves the raw CDC data as it is. Then, it uses merge into to perform an upsert operation on a silver table, which means it will insert new records or update or delete existing records based on the change type and the pk_id columns. This way, the silver table will alwaysreflect the current state of the source table, while the bronze table will keep the history of all changes. Verified References: [Databricks Certified Data Engineer Professional], under "Delta Lake" section; Databricks Documentation, under "Upsert into a table using merge" section.

QUESTION 2

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

A. sbt

B. CRAN

C. CRAM

D. nom

E. Wheels

F. jars

Correct Answer: D

QUESTION 3

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 abronzetable created with the propertydelta.enableChangeDataFeed = true. They plan to execute the following code as a daily job:

```
from pyspark.sql.functions import col

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

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

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

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

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.

Correct Answer: C

Explanation: This is the correct answer because it describes the execution and results of running the above query multiple times. The query uses the readChanges function to read all change events from a bronze table that has enabled change data feed. The readChanges function takes two arguments: version and options. The version argument specifies which version of the table to read changes from, and can be either a specific version number or -1 to indicate all versions. The options argument specifies additional options for reading changes, such as whether to include deletes or not. In this case, the query reads all changes from all versions of the bronze table and filters out delete events by setting includeDeletes to false. Then, it uses write.format("delta").mode("overwrite") to overwrite a target table using the entire history of inserted or updated records, giving the desired result of a Type 1 table representing all values that have

ever been valid for all rows in the bronze table. Verified References: [Databricks Certified Data Engineer Professional], under "Delta Lake" section; Databricks Documentation, under "Read changes in batch queries" section.

QUESTION 4

The data engineering team has configured a Databricks SQL query and alert to monitor the values in a Delta Lake table. Therecent_sensor_recordingstable contains an identifyingsensor_idalongside thetimestampandtemperaturefor 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 whenmean (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. Therecent_sensor_recordingstable 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

Correct Answer: E

Explanation: This is the correct answer because the query is using a GROUP BY clause on the sensor_id column, which means it will calculate the mean temperature for each sensor separately. The alert will trigger when the mean temperature for any sensor is greater than 120, which means at least one sensor had an average temperature above 120 for three consecutive minutes. The alert will stop when the mean temperature for all sensors drops below 120. Verified References: [Databricks Certified Data Engineer Professional], under "SQL Analytics" section; Databricks Documentation, under "Alerts" section.

QUESTION 5

A junior data engineer has been asked to develop a streaming data pipeline with a grouped aggregation using DataFramedf. 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 DataFramedfhas the following schema:

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

Code block:

```
df.withWatermark("event_time", "10 minutes")
    .groupBy(
        _____,
        "device_id"
    )
    .agg(
        avg("temp").alias("avg_temp"),
        avg("humidity").alias("avg_humidity")
    )
    .writeStream
    .format("delta")
    .saveAsTable("sensor_avg")
```

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

A. to_interval("event_time", "5 minutes").alias("time")

B. window("event_time", "5 minutes").alias("time")

C. "event_time"

D. window("event_time", "10 minutes").alias("time")

E. lag("event_time", "10 minutes").alias("time")

Correct Answer: B

Explanation: This is the correct answer because the window function is used to group streaming data by time intervals. The window function takes two arguments: a time column and a window duration. The window duration specifies how long each window is, and must be a multiple of 1second. In this case, the window duration is "5 minutes", which means each window will cover a non-overlapping five-minute interval. The window function also returns a struct column with two fields: start and end, which represent the start and end time of each window. The alias function is used to rename the struct column as "time". Verified References: [Databricks Certified Data Engineer Professional], under "Structured Streaming" section; Databricks Documentation, under "WINDOW" section.

QUESTION 6

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.

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

Correct Answer: D

Explanation: This is the correct answer because it characterizes the general programming model used by Spark Structured Streaming, which is to treat a live data stream as a table that is being continuously appended. This leads to a new stream processing model that is very similar to a batch processing model, where users can express their streaming computation using the same Dataset/DataFrame API as they would use for static data. The Spark SQL engine will take care of running the streaming query incrementally and continuously and updating the final result as streaming data continues to arrive. Verified References: [Databricks Certified Data Engineer Professional], under "Structured Streaming" section; Databricks Documentation, under "Overview" section.

QUESTION 7

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.

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

Correct Answer: A

Explanation: This is the correct answer because it is true of Delta Lake and the Lakehouse. Delta Lake uses Parquet as the underlying storage format for data files. Parquet is a columnar format that compresses data by column rather than by row. This means that Parquet can achieve high compression ratios for columns that have low cardinality or high repetition of values, such as integers, booleans, or dates. However, for columns that have high cardinality or low repetition of values, such as strings, Parquet cannot compress data very well. Therefore, strings will only be compressed when a character is repeated multiple times within a row. Verified References:[Databricks Certified Data Engineer Professional], under "Delta Lake" section; Databricks Documentation, under "Delta Lake core features - Schema enforcement and evolution" section.

QUESTION 8

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")

Correct Answer: D

Explanation: This is the correct way to get a parameter passed to a notebook by the Databricks Jobs API. The dbutils.notebooks.getParam method returns the value of a parameter passed to a notebook as a string. If no parameter with that name is passed, it returns None by default. You can also specify a default value as a second argument. Verified References: Databricks Certified Data Engineer Professional, under "Databricks Tooling" section; Databricks Documentation, under "Pass parameters to a notebook" section.

QUESTION 9

A junior data engineer has configured a workload that posts the following JSON to the Databricks REST API endpoint2.0/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.

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.

Correct Answer: C

Explanation: This is the correct answer because the JSON posted to the Databricks REST API endpoint 2.0/jobs/create defines a new job with a name, an existing cluster id, and a notebook task. However, it does not specify any schedule or trigger for the job execution. Therefore, three new jobs with the same name and configuration will be created in the workspace, but none of them will be executed until they are manually triggered or scheduled. Verified References:

[Databricks Certified Data Engineer Professional], under "Monitoring and Logging" section; [Databricks Documentation], under "Jobs API - Create" section.

QUESTION 10

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

E. Can Run

Correct Answer: D

Explanation: This is the correct answer because it is the maximum notebook permissions that can be granted to the user without allowing accidental changes to production code or data. Notebook permissions are used to control access to notebooks in Databricks workspaces. There are four types of notebook permissions: Can Manage, Can Edit, Can Run, and Can Read. Can Manage allows full control over the notebook, including editing, running, deleting, exporting, and changing permissions. Can Edit allows modifying and running the notebook, but not changing permissions or deleting it. Can Run allows executing commands in an existing cluster attached to the notebook, but not modifying or exporting it. Can Read allows viewing the notebook content, but not running or modifying it. In this case, granting Can Read permission to the user will allow them to review the production logic in the notebook without allowing them to makeany changes to it or run any commands that may affect production data. Verified References: [Databricks Certified Data Engineer Professional], under "Databricks Workspace" section; Databricks Documentation, under "Notebook permissions" section.