DATABRICKS-CERTIFIED-PROFESSIONAL-DATA-ENGINEER$^{Q\&As}$

Databricks Certified Professional Data Engineer Exam

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

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

QUESTION 1

A table in the Lakehouse namedcustomer_churn_paramsis used in churn prediction by the machine learning team. The table contains information about customers derived from a number of upstream sources. Currently, the data engineering

team populates this table nightly by overwriting the table with the current valid values derived from upstream data sources.

The churn prediction model used by the ML team is fairly stable in production. The team is only interested in making predictions on records that have changed in the past 24 hours.

Which approach would simplify the identification of these changed records?

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

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

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

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

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

Correct Answer: B

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 an existing cluster id and a notebook task, but also specifies a new cluster spec with some

configurations. According to the documentation, if both an existing cluster id and a new cluster spec are provided, then a new cluster will be created for each run of the job with those configurations, and then terminated after completion.

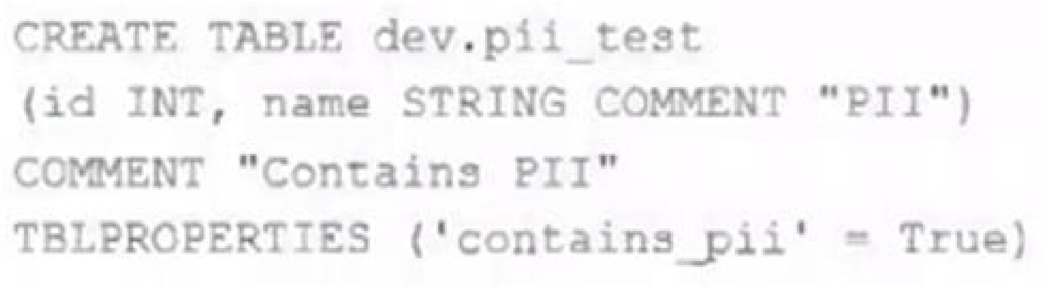
Therefore, the logic defined in the referenced notebook will be executed three times on new clusters with those configurations. Verified References:

[Databricks Certified Data Engineer Professional], under "Monitoring and Logging" section; Databricks Documentation, under "JobsClusterSpecNewCluster" section.

QUESTION 2

The data governance team has instituted a requirement that all tables containing Personal Identifiable Information (PH) must be clearly annotated. This includes adding column comments, table comments, and setting the custom table property"contains_pii" = true.

The following SQL DDL statement is executed to create a new table:



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

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

A. DESCRIBE EXTENDED dev.pii test

B. DESCRIBE DETAIL dev.pii test

C. SHOW TBLPROPERTIES dev.pii test

D. DESCRIBE HISTORY dev.pii test

E. SHOW TABLES dev

Correct Answer: A

Explanation: This is the correct answer because it allows manual confirmation that these three requirements have been met. The requirements are that all tables containing Personal Identifiable Information (PII) must be clearly annotated, which includes adding column comments, table comments, and setting the custom table property "contains_pii" = true. The DESCRIBE EXTENDED command is used to display detailed information about a table, such as its schema, location, properties, and comments. By using this command on the dev.pii_test table, one can verify that the table has been created with the correct column comments, table comment, and custom table property as specified in the SQL DDL statement. Verified References: [Databricks Certified Data Engineer Professional], under "Lakehouse" section; Databricks Documentation, under "DESCRIBE EXTENDED" section.

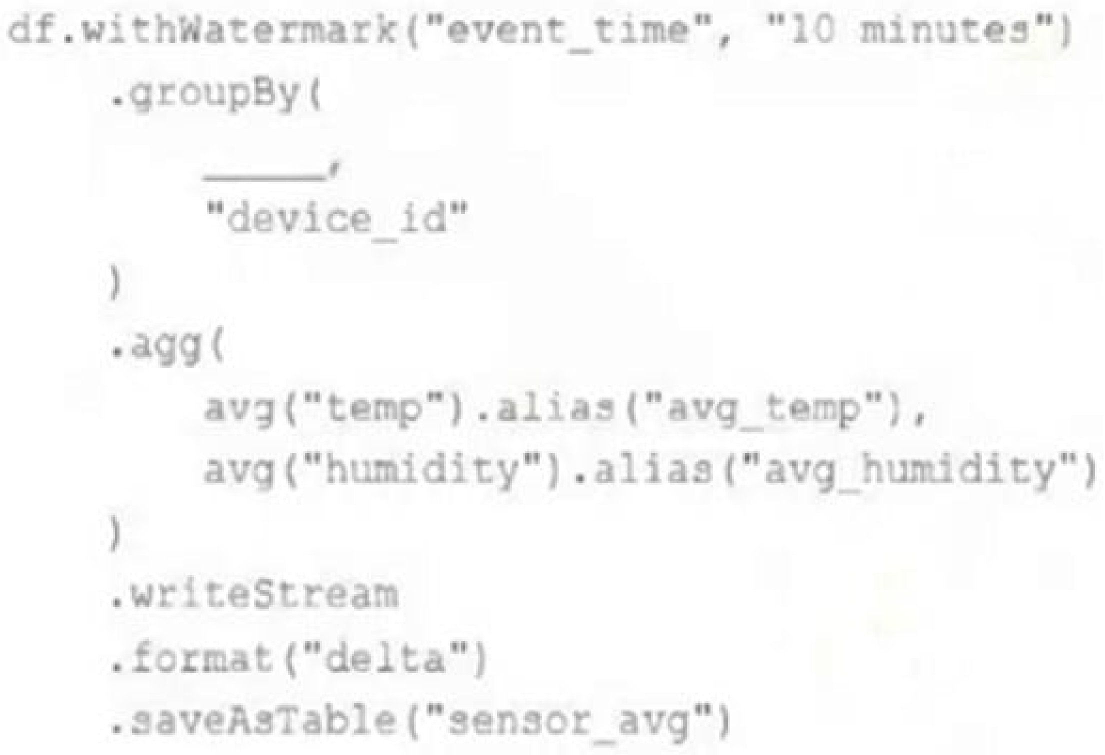
QUESTION 3

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")

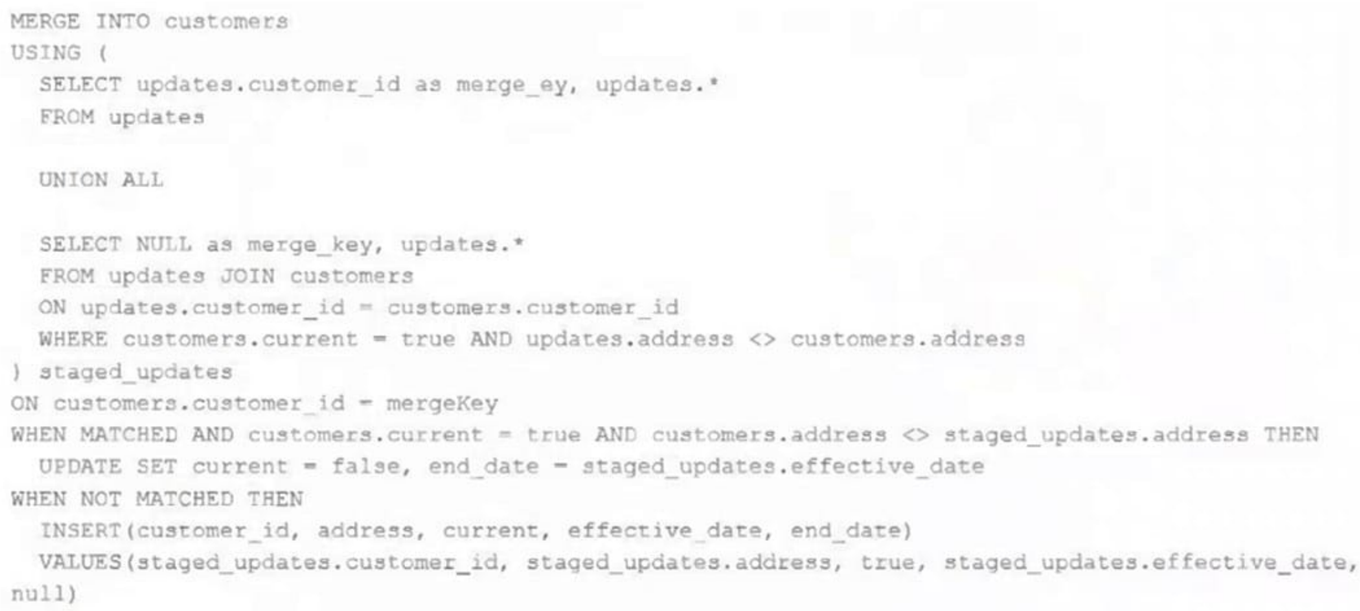
Correct Answer: B

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 4

The viewupdatesrepresents an incremental batch of all newly ingested data to be inserted or updated in thecustomerstable.

The following logic is used to process these records.



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

  UNION ALL

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

Which statement describes this implementation?

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

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

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

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

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

Correct Answer: B

Explanation: The logic uses the MERGE INTO command to merge new records from the view updates into the table customers. The MERGE INTO command takes two arguments:

a target table and a source table or view. The command also specifies a condition to match records between the target and the source, and a set of actions to perform when there is a match or not. In this case, the condition is to match

records by customer_id, which is the primary key of the customers table. The actions are to update the existing record in the target with the new values from the source, and set the current_flag to false to indicate that the record is no longer

current; and to insert a new record in the target with the new values from the source, and set the current_flag to true to indicate that the record is current. This means that old values are maintained but marked as no longer current and new

values are inserted, which is the definition of a Type 2 table. Verified References: [Databricks Certified Data Engineer Professional], under "Delta Lake" section; Databricks Documentation, under "Merge Into (Delta Lake on Databricks)"

section.

QUESTION 5

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

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

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

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

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

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

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

Correct Answer: C

Explanation: The events are that a data engineer, User A, has promoted a new pipeline to production by using the REST API to programmatically create several jobs, and a DevOps engineer, User B, has configured an external orchestration tool to trigger job runs through the REST API. Both users authorized the REST API calls using their personal access tokens. The workspace audit logs are logs that record user activities in a Databricks workspace, such as creating, updating, or deleting objects like clusters, jobs, notebooks, or tables. The workspace audit logs also capture the identity of the user who performed each activity, as well as the time and details of the activity. Because these events are managed separately, User A will have their identity associated with the job creation events and User B will have their identity associated with the job run events in the workspace audit logs. Verified References: [Databricks Certified Data Engineer Professional], under "Databricks Workspace" section; Databricks Documentation, under "Workspace audit logs" section.