Certified Data Engineer Professional

1. Topic #: 1
An hourly batch job is configured to ingest data files from a cloud object storage container where each batch represent all records produced by the source system in a given hour. The batch job to process these records into the Lakehouse is sufficiently delayed to ensure no late-arriving data is missed. The user_id field represents a unique key for the data, which has the following schema: user_id BIGINT, username STRING, user_utc STRING, user_region STRING, last_login BIGINT, auto_pay BOOLEAN, last_updated BIGINT
New records are all ingested into a table named account_history which maintains a full record of all data in the same schema as the source. The next table in the system is named account_current and is implemented as a Type 1 table representing the most recent value for each unique user_id.
Assuming there are millions of user accounts and tens of thousands of records processed hourly, which implementation can be used to efficiently update the described account_current table as part of each hourly batch job?

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

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

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

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

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

Answer(s): C

---

2. Question #: 21
Topic #: 1

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

Answer(s): E

---

3. Question #: 140
Topic #: 1
Which statement describes Delta Lake optimized writes?

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

B. B. An asynchronous job runs after the write completes to detect if files could be further compacted; if yes, an OPTIMIZE job is executed toward a default of 1 GB.

C. C. A shuffle occurs prior to writing to try to group similar data together resulting in fewer files instead of each executor writing multiple files based on directory partitions.

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

Answer(s): C

---

4. Question #: 38
Topic #: 1
The downstream consumers of a Delta Lake table have been complaining about data quality issues impacting performance in their applications. Specifically, they have complained that invalid latitude and longitude values in the activity_details table have been breaking their ability to use other geolocation processes.
A junior engineer has written the following code to add CHECK constraints to the Delta Lake table:

A. A senior engineer has confirmed the above logic is correct and the valid ranges for latitude and longitude are provided, but the code fails when executed.

B. Which statement explains the cause of this failure?

Answer(s): C

---

5. Question #: 35
Topic #: 1
To reduce storage and compute costs, the data engineering team has been tasked with curating a series of aggregate tables leveraged by business intelligence dashboards, customer-facing applications, production machine learning models, and ad hoc analytical queries.
The data engineering team has been made aware of new requirements from a customer-facing application, which is the only downstream workload they manage entirely. As a result, an aggregate table used by numerous teams across the organization will need to have a number of fields renamed, and additional fields will also be added.
Which of the solutions addresses the situation while minimally interrupting other teams in the organization without increasing the number of tables that need to be managed?

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

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

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

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

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

Answer(s): B

---

6. Question #: 30
Topic #: 1
A nightly job ingests data into a Delta Lake table using the following code:

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

B. Which code snippet completes this function definition?

C. def new_records():

D. A. return spark.readStream.table(“bronze”)

E. B. return spark.readStream.load(“bronze”)

F. C.

G. D. return spark.read.option(“readChangeFeed”, “true”).table (“bronze”)

H. E.

Answer(s): D

---

7. Question #: 28
Topic #: 1
A junior data engineer seeks to leverage Delta Lake’s Change Data Feed functionality to create a Type 1 table representing all of the values that have ever been valid for all rows in a bronze table created with the property delta.enableChangeDataFeed = true. They plan to execute the following code as a daily job:

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

Answer(s): B

---

8. Question #: 172
Topic #: 1
The data engineer is using Spark's MEMORY_ONLY storage level.

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

Answer(s): C

---

9. Question #: 91
Topic #: 1
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.

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

Answer(s): E

---

10. Question #: 150
Topic #: 1
A nightly job ingests data into a Delta Lake table using the following code:

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

Answer(s): B

---

11. Question #: 131
Topic #: 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 field pk_id.

A. 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.

Answer(s): D

---

12. Question #: 113
Topic #: 1
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.

A. 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.

Answer(s): A

---

13. Question #: 88
Topic #: 1
You are testing a collection of mathematical functions, one of which calculates the area under a curve as described by another function.

A. assert(myIntegrate(lambda x: x*x, 0, 3) [0] == 9)

Answer(s): A

---

14. Question #: 128
Topic #: 1
The data engineering team has configured a job to process customer requests to be forgotten (have their data deleted). All user data that needs to be deleted is stored in Delta Lake tables using default table settings.

A. The team has decided to process all deletions from the previous week as a batch job at 1am each Sunday. The total duration of this job is less than one hour. Every Monday at 3am, a batch job executes

a series of VACUUM commands on all Delta Lake tables throughout the organization.

Answer(s): C

---

15. Question #: 126
Topic #: 1
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.

A. Before executing the code, running SHOW TABLES on the current database indicates the database contains only two tables: geo_lookup and sales.

Answer(s): D

---

16. Question #: 125
Topic #: 1
The data science team has created and logged a production model using MLflow. The model accepts a list of column names and returns a new column of type DOUBLE.

A. The following code correctly imports the production model, loads the customers table containing the customer_id key column into a DataFrame, and defines the feature columns needed for the model.

Answer(s): B

---

17. Question #: 124
Topic #: 1
The security team is exploring whether or not the Databricks secrets module can be leveraged for connecting to an external database.

A. After testing the code with all Python variables being defined with strings, they upload the password to the secrets module and configure the correct permissions for the currently active user. They then modify their code to the following (leaving all other variables unchanged).

Answer(s): A

---

18. Question #: 99
Topic #: 1
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.

A. 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?

Answer(s): B

---

19. Question #: 49
Topic #: 1
A user new to Databricks is trying to troubleshoot long execution times for some pipeline logic they are working on. Presently, the user is executing code cell-by-cell, using display() calls to confirm code is producing the logically correct results as new transformations are added to an operation. To get a measure of average time to execute, the user is running each cell multiple times interactively.
Which of the following adjustments will get a more accurate measure of how code is likely to perform in production?

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

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

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

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

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

Answer(s): B

---

20. Question #: 80
Topic #: 1
The marketing team is looking to share data in an aggregate table with the sales organization, but the field names used by the teams do not match, and a number of marketing-specific fields have not been approved for the sales org.

A. Which of the following solutions addresses the situation while emphasizing simplicity?

Answer(s): A
