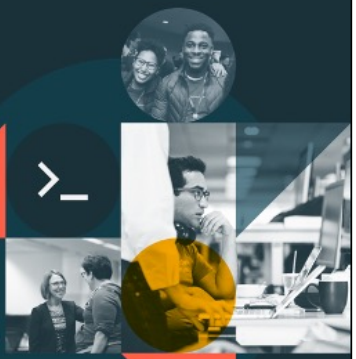
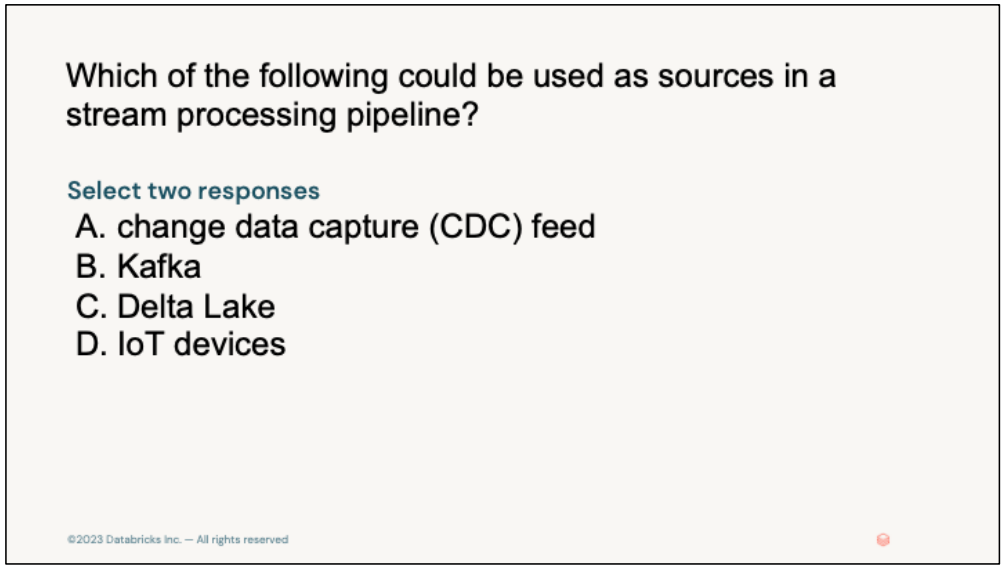
databricks

Incremental Processing with Structured Streaming



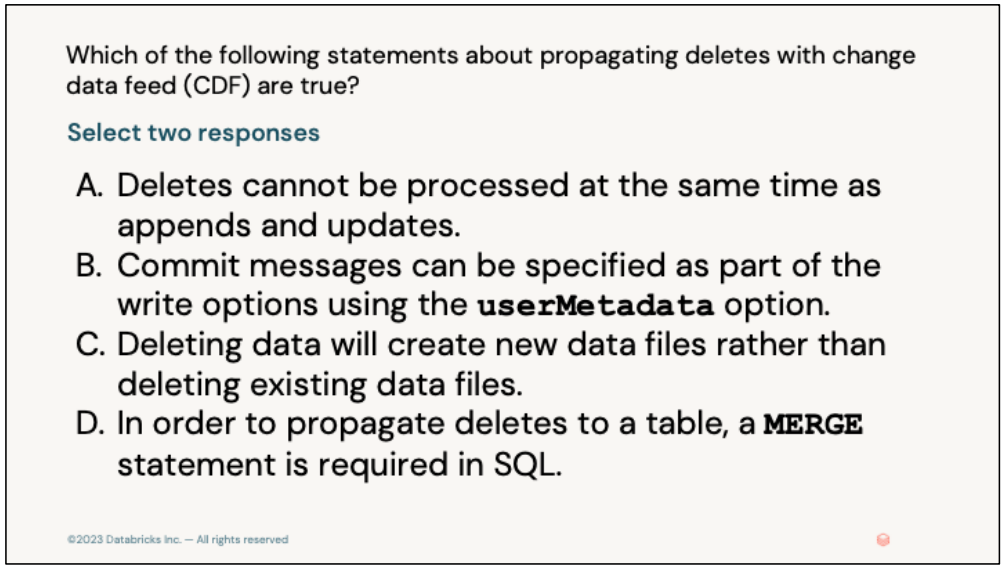
Databricks Academy
2023

Knowledge Check



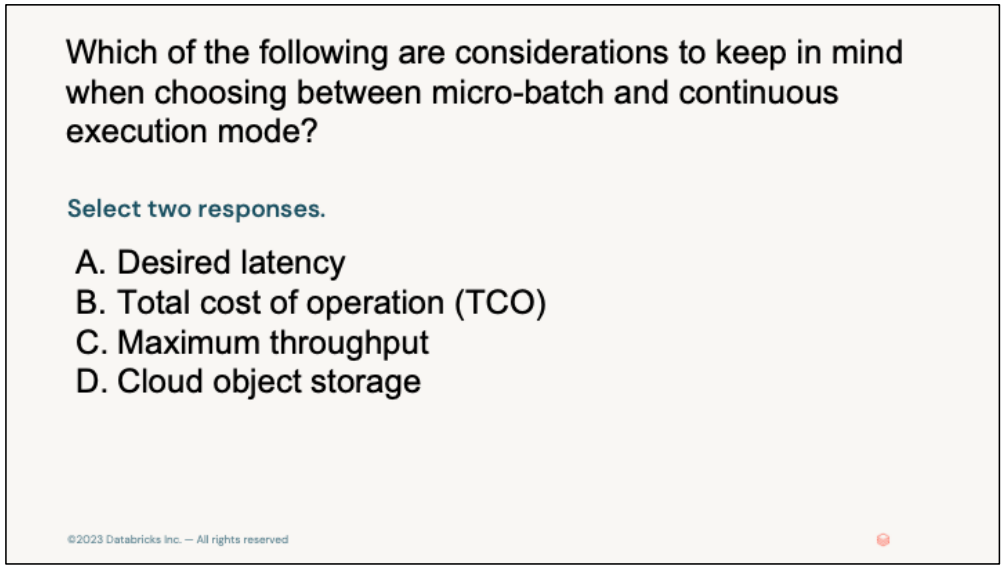
Which of the following could be used as sources in a stream processing pipeline?
Select two responses.

- **A. change data capture (CDC) feed**
- B. Kafka
- C. Delta Lake
- **D. IoT devices**



Which of the following statements about propagating deletes with change data feed (CDF) are true? Select two responses.

A. Deletes cannot be processed at the same time as appends and updates.
B. Commit messages can be specified as part of the write options using the `userMetadata` option.
C. Deleting data will create new data files rather than deleting existing data files.
D. In order to propagate deletes to a table, a `MERGE` statement is required in SQL.



Which of the following are considerations to keep in mind when choosing between micro-batch and continuous execution mode?

A. Desired latency
B. Total cost of operation (TCO)
C. Maximum throughput
D. Cloud object storage

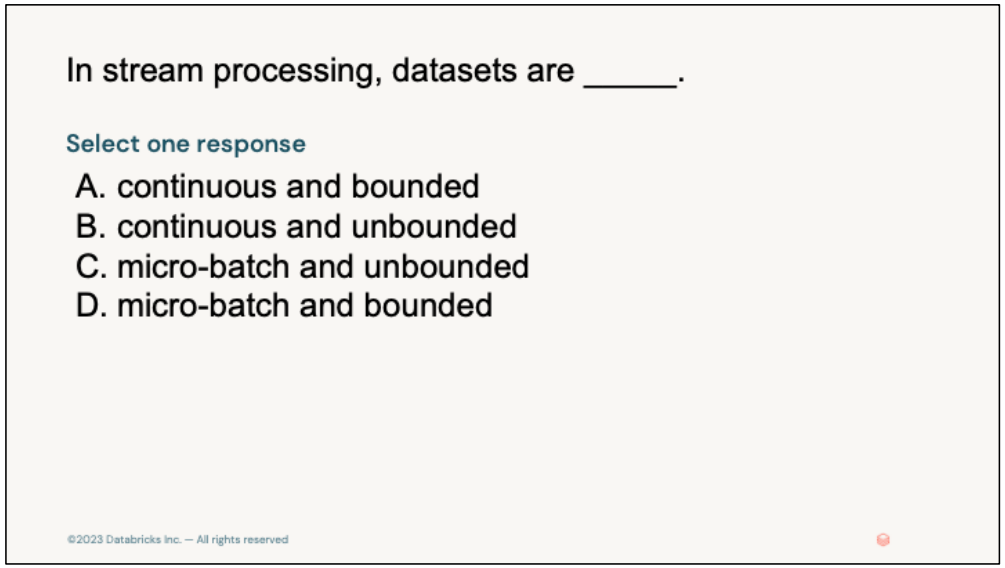
Which of the following functions completes the following code snippet to return a Spark DataFrame in a structured streaming query?

```
spark.readStream.format("kafka")
  .option("kafka.bootstrap.servers",...)
  .option("subscribe", "topic")
  _____
```

Select one response.

A. .load()
B. .print()
C. .return()
D. .merge()

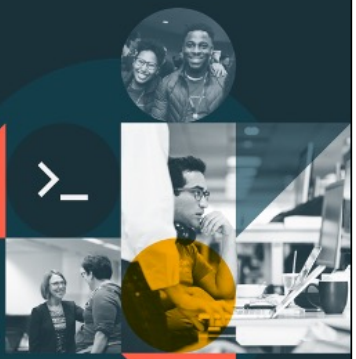
A. .load()
B. .print()
C. .return()
D. .merge()



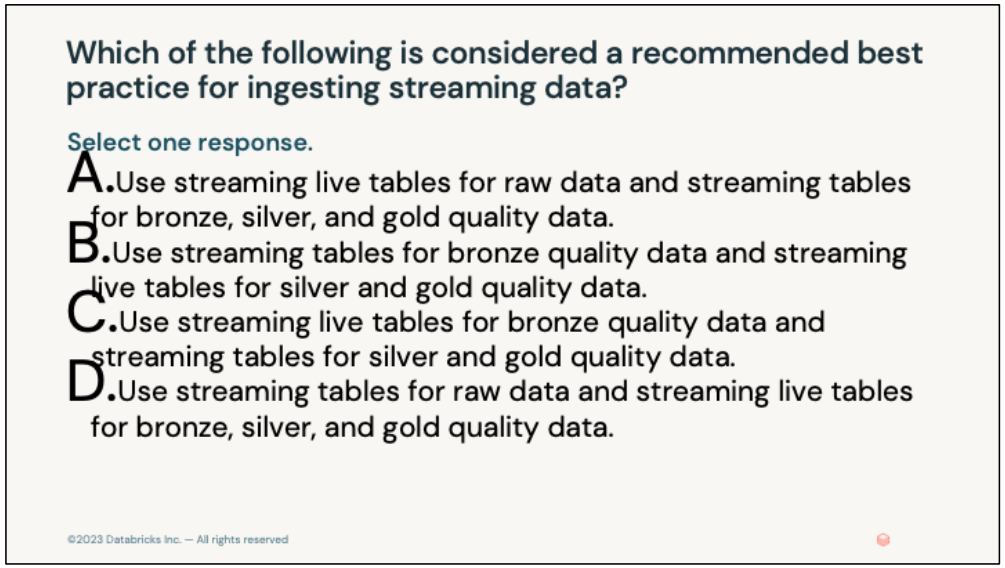
A. continuous and bounded
B. continuous and unbounded
C. micro-batch and unbounded
D. micro-batch and bounded

databricks

Streaming ETL Patterns with DLT



Knowledge Check



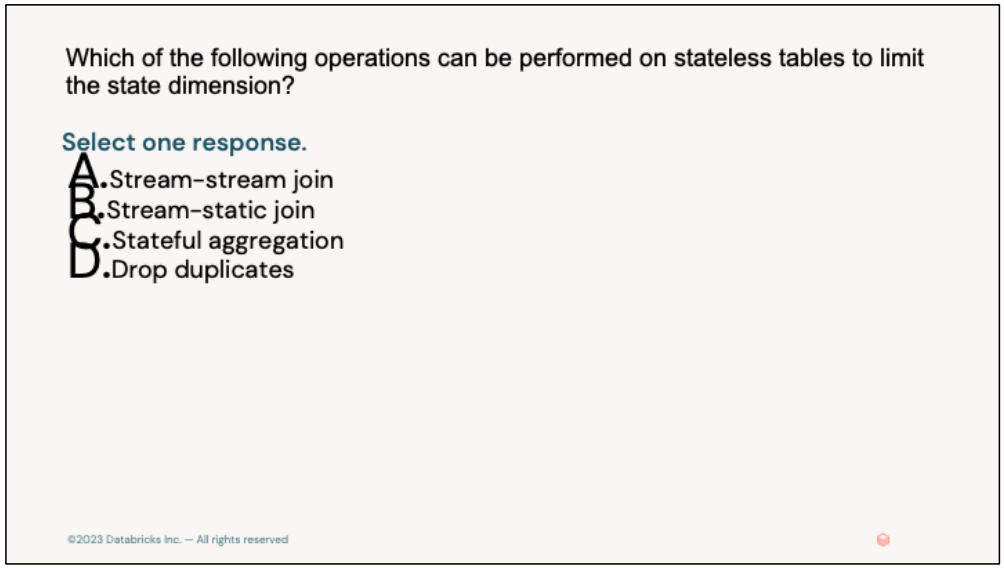
A. Use streaming live tables for raw data and streaming tables for bronze, silver, and gold quality data.
B. Use streaming tables for bronze quality data and streaming live tables for silver and gold quality data.
C. Use streaming live tables for bronze quality data and streaming tables for silver and gold quality data.
D. Use streaming tables for raw data and streaming live tables for bronze, silver, and gold quality data.

A data engineer has data that needs to be updated. However, they need to have access to a recorded history of the information previously stored in the dataset before the update. Which of the following table types should the data engineer use for their data?

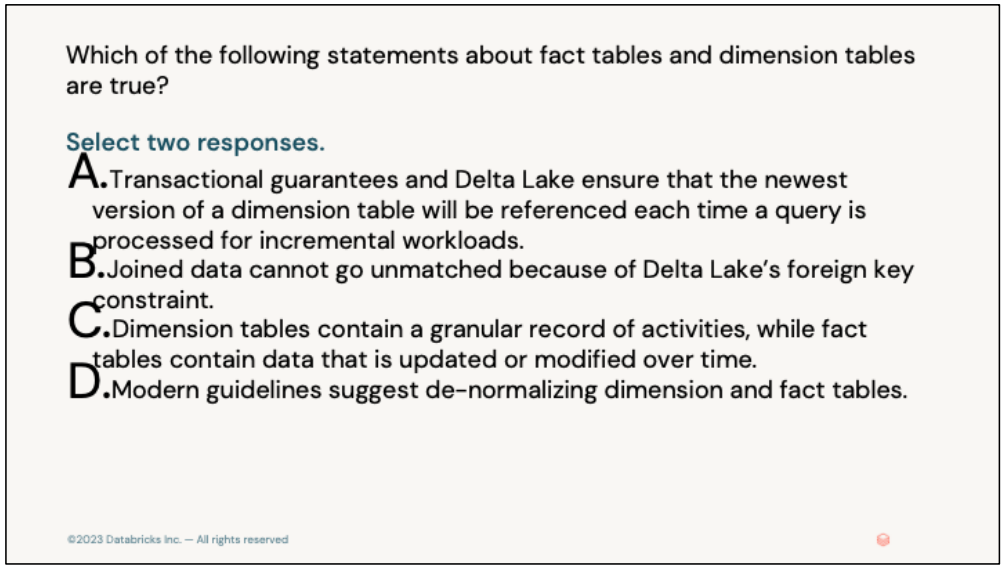
Select one response.

A. Type 0
B. Type 1
C. Type 2
D. Type 1 or Type 2

A. Type 0
B. Type 1
C. Type 2
D. Type 1 or Type 2



A. Stream-stream join
B. Stream-static join
C. Stateful aggregation
D. Drop duplicates



A. Transactional guarantees and Delta Lake ensure that the newest version of a dimension table will be referenced each time a query is processed for incremental workloads.
B. Joined data cannot go unmatched because of Delta Lake's foreign key constraint.
C. Dimension tables contain a granular record of activities, while fact tables contain data that is updated or modified over time.
D. Modern guidelines suggest denormalizing dimension and fact tables.

The following line of code is supposed to create a set of inverted rules for a quarantine table.

quarantine_rules = _____

Which of the following correctly fills in the blank?

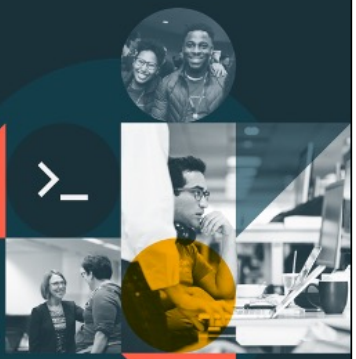
Select one response.

A. `{"invalid_record": f"NOT({' AND '.join(rules.values())})"}`
B. `{"invalid_record": f"&&({' ! '.join(rules.values())})"}`
C. `{"invalid_record": f"NOT({' OR '.join(rules.values())})"}`
D. `{"invalid_record": f"IF({' NULL '.join(rules.values())})"}`

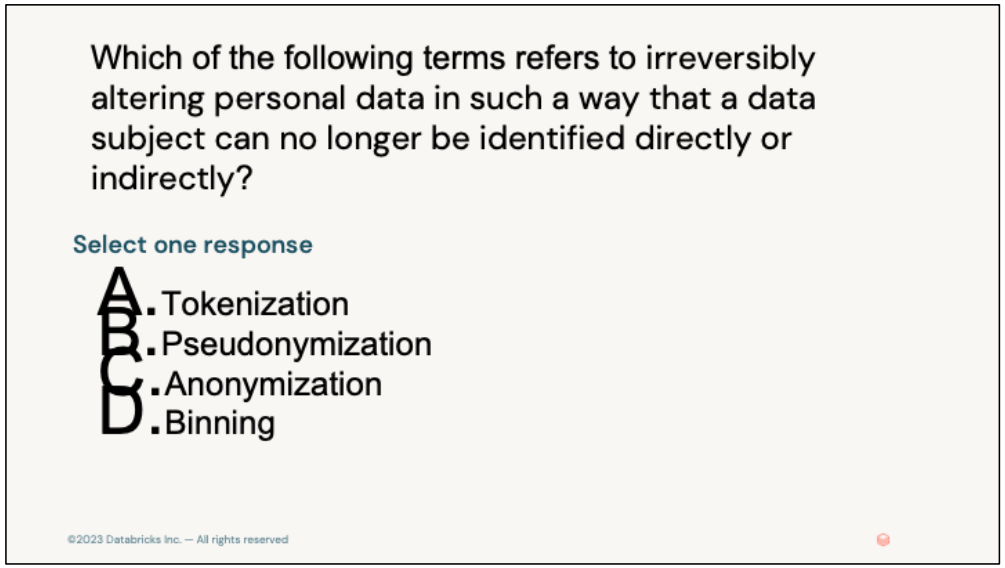
A. {"invalid_record": f"NOT({' AND '.join(rules.values())})"}
B. {"invalid_record": f"&&({' ! '.join(rules.values())})"}
C. {"invalid_record": f"NOT({' OR '.join(rules.values())})"}
D. {"invalid_record": f"IF({' NULL '.join(rules.values())})"}

databricks

Data Privacy Patterns

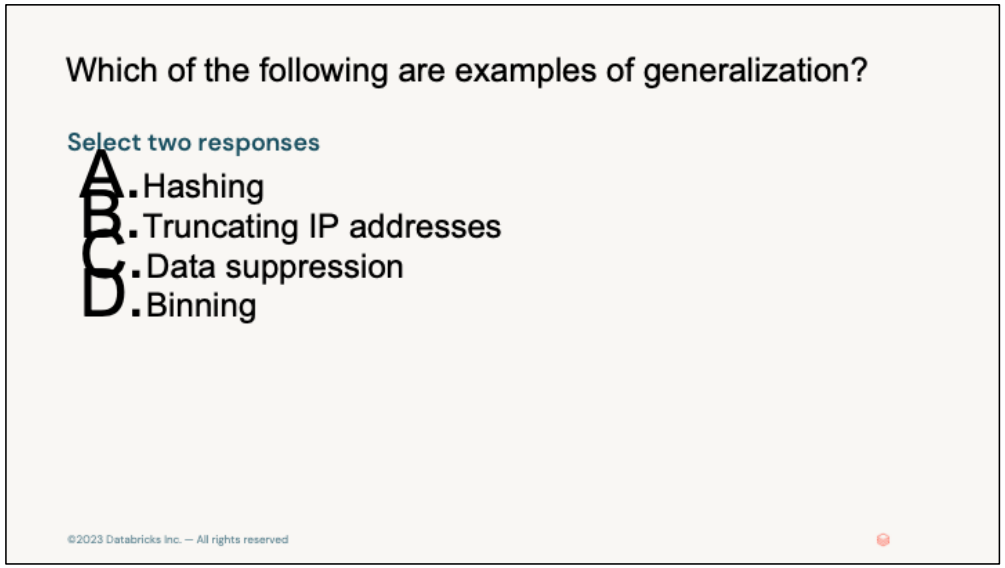


Knowledge Check

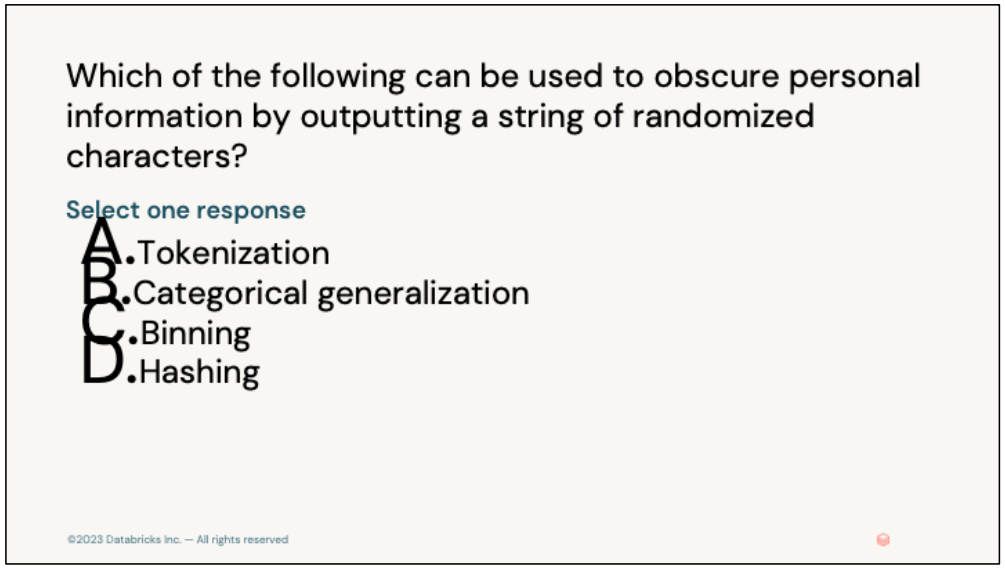


A. Tokenization
B. Pseudonymization
C. Anonymization
D. Binning

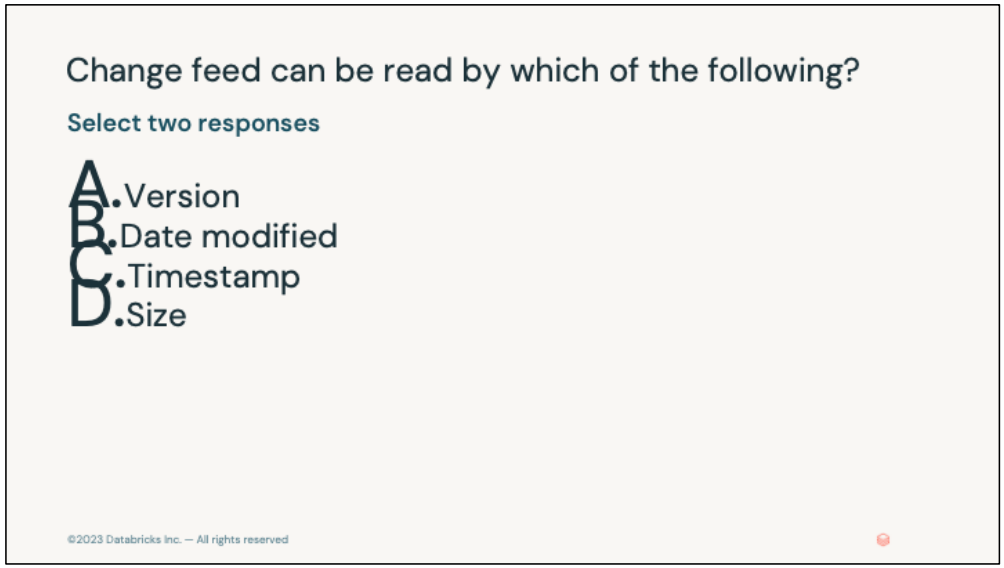
A. HIPAA
B. PCI-DSS
C. GDPR
D. CCPA



A. Hashing
B. Truncating IP addresses
C. Data suppression
D. Binning



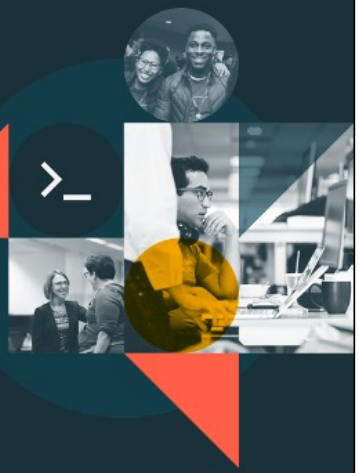
A. Tokenization
B. Categorical generalization
C. Binning
D. Hashing



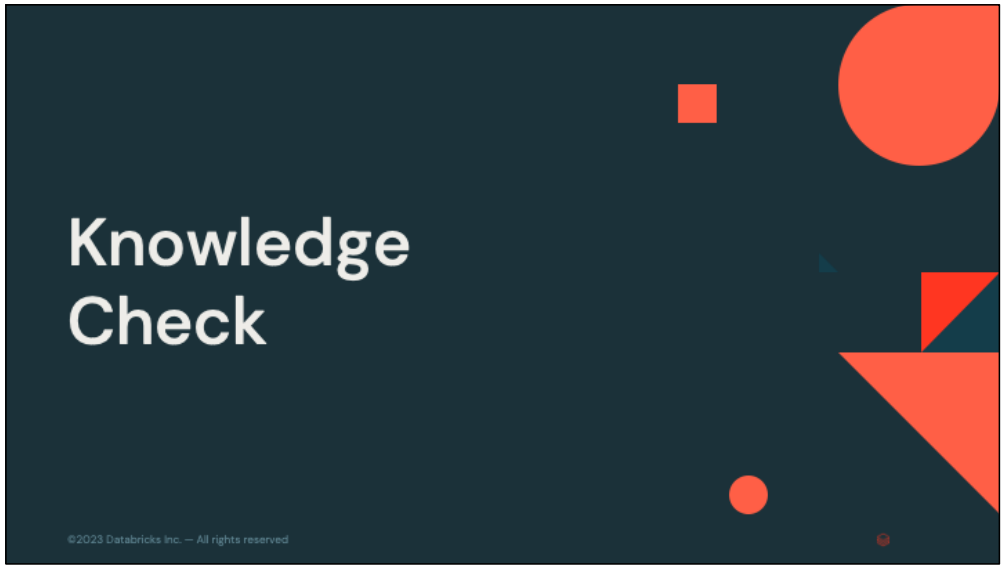
A. **Version**
B. Date modified
C. **Timestamp**
D. Size

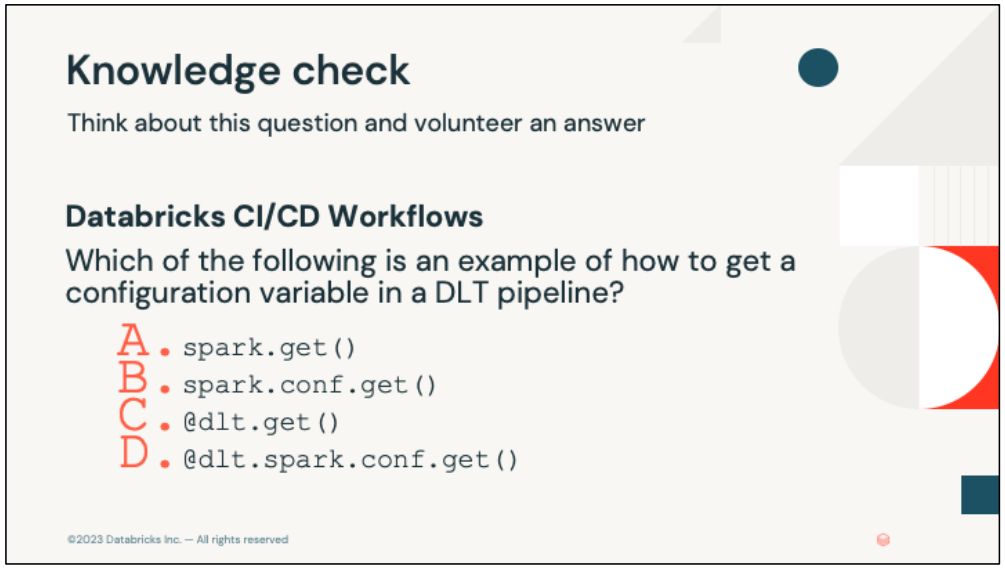
databricks
Academy

SWE Practices for DLT Pipelines

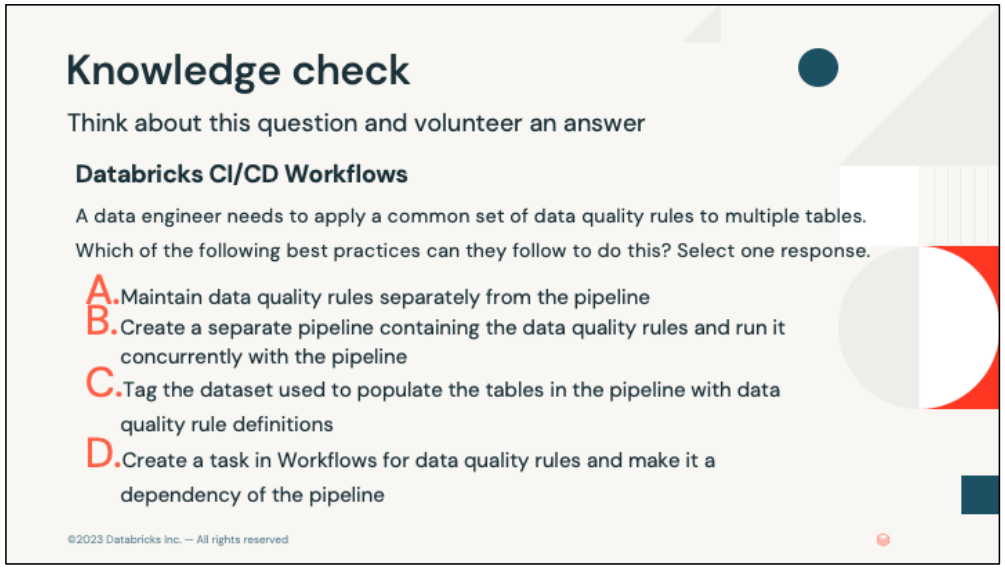


Databricks Academy
2023

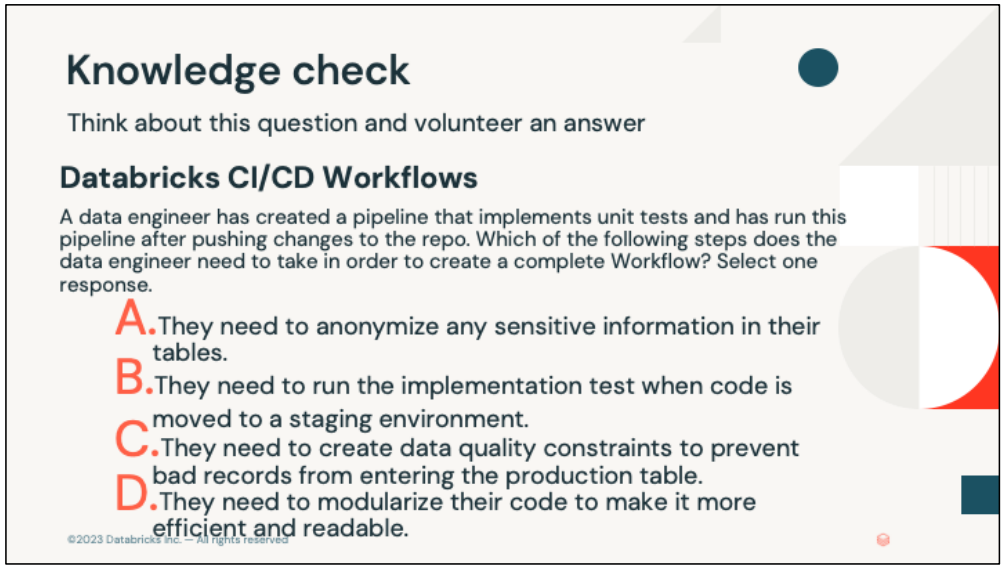

Knowledge
Check
©2023 Databricks Inc. — All rights reserved


Knowledge check
Think about this question and volunteer an answer
Databricks CI/CD Workflows
Which of the following is an example of how to get a configuration variable in a DLT pipeline?
A. spark.get()
B. spark.conf.get()
C. @dlt.get()
D. @dlt.spark.conf.get()
©2023 Databricks Inc. — All rights reserved

B


Knowledge check
Think about this question and volunteer an answer
Databricks CI/CD Workflows
A data engineer needs to apply a common set of data quality rules to multiple tables.
Which of the following best practices can they follow to do this? Select one response.
A. Maintain data quality rules separately from the pipeline
B. Create a separate pipeline containing the data quality rules and run it concurrently with the pipeline
C. Tag the dataset used to populate the tables in the pipeline with data quality rule definitions
D. Create a task in Workflows for data quality rules and make it a dependency of the pipeline
©2023 Databricks Inc. — All rights reserved

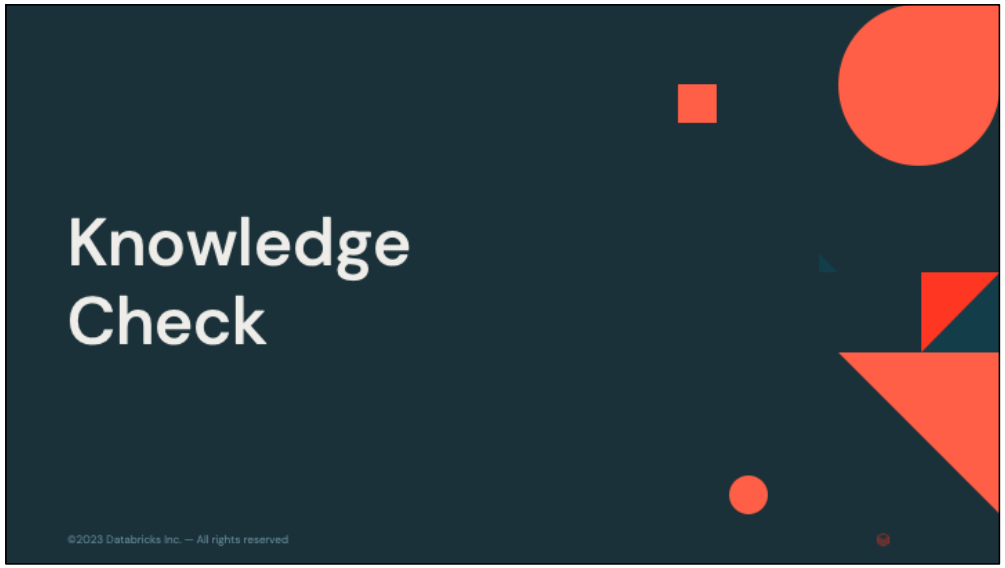
A


Knowledge check
Think about this question and volunteer an answer
Databricks CI/CD Workflows
A data engineer has created a pipeline that implements unit tests and has run this pipeline after pushing changes to the repo. Which of the following steps does the data engineer need to take in order to create a complete Workflow? Select one response.
A. They need to anonymize any sensitive information in their tables.
B. They need to run the implementation test when code is moved to a staging environment.
C. They need to create data quality constraints to prevent bad records from entering the production table.
D. They need to modularize their code to make it more efficient and readable.
©2023 Databricks Inc. — All rights reserved

B

databricks

Automate Production Workflows

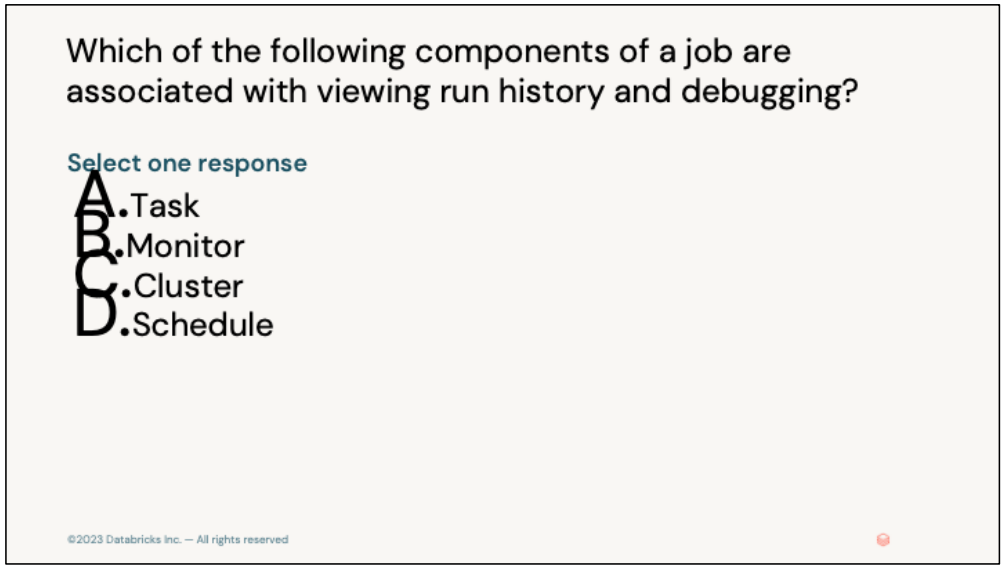

Knowledge
Check
©2023 Databricks Inc. — All rights reserved

A data engineer wants to create a single refined table out of raw data from several streaming sources. Which of the following data ingestion patterns can be used to complete this task?

Select one response

A. Sequence
B. Filter
C. Fan-out
D. Funnel

D


Which of the following components of a job are associated with viewing run history and debugging?
Select one response
A. Task
B. Monitor
C. Cluster
D. Schedule
©2023 Databricks Inc. — All rights reserved

B

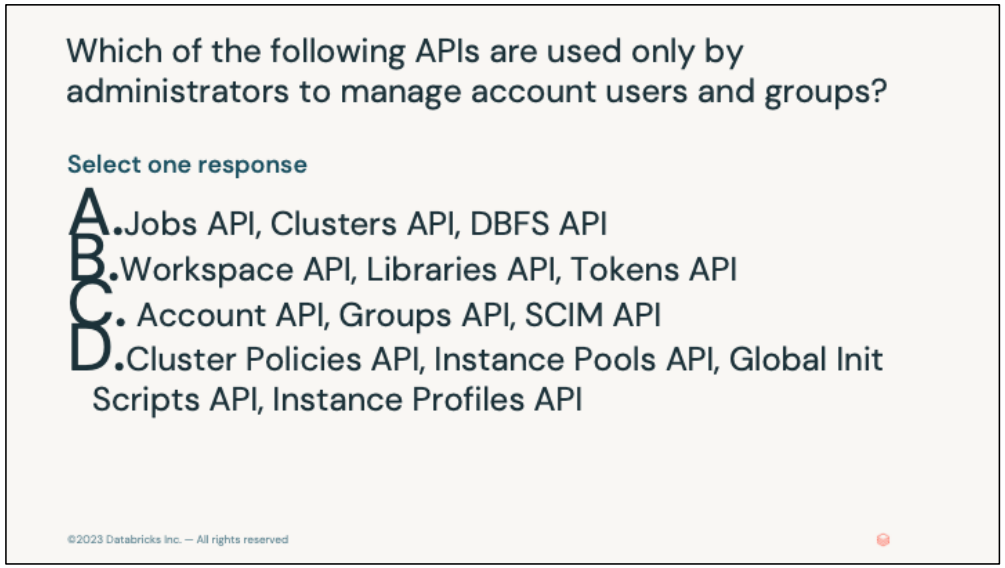
In what order do the following commands need to be run in order to install and configure the Databricks CLI in the Databricks workspace at the location stored in the variable `path`?

1. `databricks configure -token`
2. `pip install databricks-cli`
3. `databricks workspace ls $path`

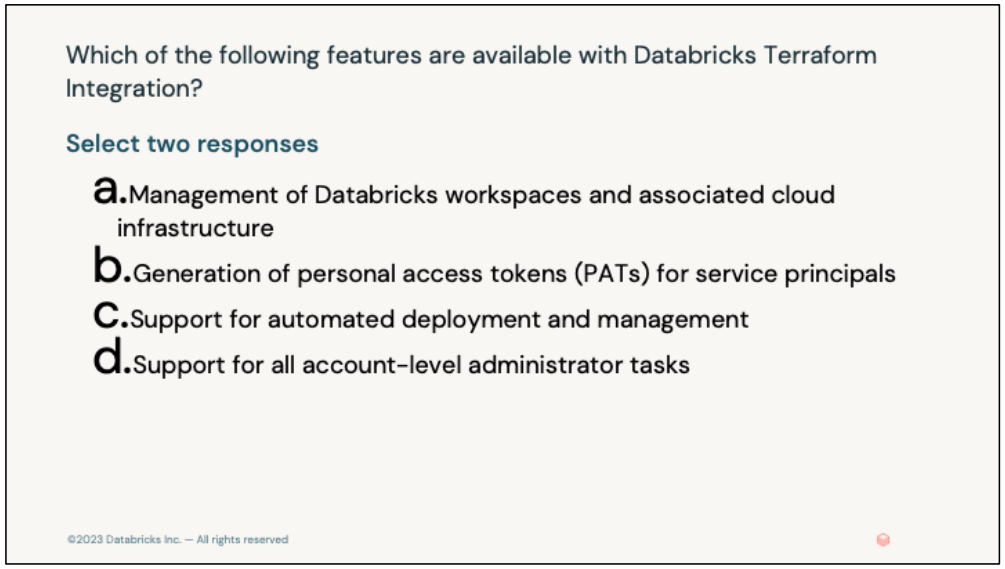
Select one response

A. 3, 2, 1
B. 1, 3, 2
C. 2, 1, 3
D. 2, 3, 1

C


Which of the following APIs are used only by administrators to manage account users and groups?
Select one response
A. Jobs API, Clusters API, DBFS API
B. Workspace API, Libraries API, Tokens API
C. Account API, Groups API, SCIM API
D. Cluster Policies API, Instance Pools API, Global Init Scripts API, Instance Profiles API
©2023 Databricks Inc. — All rights reserved

C


Which of the following features are available with Databricks Terraform Integration?
Select two responses
a. Management of Databricks workspaces and associated cloud infrastructure
b. Generation of personal access tokens (PATs) for service principals
c. Support for automated deployment and management
d. Support for all account-level administrator tasks
©2023 Databricks Inc. — All rights reserved

A,C