89 # Which of the following describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.

B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.

C. CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.

D. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.

E. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

89 # Which of the following describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.

B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.

C. CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.

D. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.

E. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

90 # Which of the following queries is performing a streaming hop from raw data to a Bronze table?

A.
```
(spark.table("sales")
    .groupBy("store")
    .agg(sum("sales"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .table("newSales")
)
```

B.
```
(spark.table("sales")
    .filter(col("units") > 0)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

C.
```
(spark.table("sales")
    .withColumn("avgPrice", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

D.
```
(spark.read.load(rawSalesLocation)
    .write
    .mode("append")
    .table("newSales")
)
```

E.
```
(spark.readStream.load(rawSalesLocation)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```


sthithapragnasya@gmail.com

90 # Which of the following queries is performing a streaming hop from raw data to a Bronze table?

A.
```
(spark.table("sales")
    .groupBy("store")
    .agg(sum("sales"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .table("newSales")
)
```

B.
```
(spark.table("sales")
    .filter(col("units") > 0)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

C.
```
(spark.table("sales")
    .withColumn("avgPrice", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```

D.
```
(spark.read.load(rawSalesLocation)
    .write
    .mode("append")
    .table("newSales")
)
```

E.
```
(spark.readStream.load(rawSalesLocation)
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("append")
    .table("newSales")
)
```


sthithapragnasya@gmail.com

91 # Which data lakehouse feature results in improved data quality over a traditional data lake?

A. A data lakehouse stores data in open formats.

B. A data lakehouse allows the use of SQL queries to examine data.

C. A data lakehouse provides storage solutions for structured and unstructured data.

D. A data lakehouse supports ACID-compliant transactions.

91 # Which data lakehouse feature results in improved data quality over a traditional data lake?

A. A data lakehouse stores data in open formats.

B. A data lakehouse allows the use of SQL queries to examine data.

C. A data lakehouse provides storage solutions for structured and unstructured data.

D. A data lakehouse supports ACID-compliant transactions.

92 # In which scenario will a data team want to utilize cluster pools?

A. An automated report needs to be version-controlled across multiple collaborators.

B. An automated report needs to be runnable by all stakeholders.

C. An automated report needs to be refreshed as quickly as possible.

D. An automated report needs to be made reproducible.

92 # In which scenario will a data team want to utilize cluster pools?

A. An automated report needs to be version-controlled across multiple collaborators.

B. An automated report needs to be runnable by all stakeholders.

C. An automated report needs to be refreshed as quickly as possible.

D. An automated report needs to be made reproducible.

93 # What is hosted completely in the control plane of the classic Databricks architecture?

A. Worker node

B. Databricks web application

C. Driver node

D. Databricks Filesystem

93 # What is hosted completely in the control plane of the classic Databricks architecture?

A. Worker node

B. Databricks web application

C. Driver node

D. Databricks Filesystem

94 # A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.

What is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

A. Databricks Repos allows users to revert to previous versions of a notebook

B. Databricks Repos is wholly housed within the Databricks Data Intelligence Platform

C. Databricks Repos provides the ability to comment on specific changes

D. Databricks Repos supports the use of multiple branches

94 # A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos.

What is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

A. Databricks Repos allows users to revert to previous versions of a notebook

B. Databricks Repos is wholly housed within the Databricks Data Intelligence Platform

C. Databricks Repos provides the ability to comment on specific changes

D. Databricks Repos supports the use of multiple branches

95 # What is a benefit of the Databricks Lakehouse Architecture embracing open source technologies?

A. Avoiding vendor lock-in

B. Simplified governance

C. Ability to scale workloads

D. Cloud-specific integrations

95 # What is a benefit of the Databricks Lakehouse Architecture embracing open source technologies?

A. Avoiding vendor lock-in

B. Simplified governance

C. Ability to scale workloads

D. Cloud-specific integrations

96 # A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.

In which location can the data engineer review their permissions on the table?

A. Jobs

B. Dashboards

C. Catalog Explorer

D. Repos

96 # A data engineer needs to use a Delta table as part of a data pipeline, but they do not know if they have the appropriate permissions.

In which location can the data engineer review their permissions on the table?

A. Jobs

B. Dashboards

C. Catalog Explorer

D. Repos

97 # A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.

Which Git operation does the data engineer need to run to accomplish this task?

A. Clone

B. Pull

C. Merge

D. Push

97 # A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.

Which Git operation does the data engineer need to run to accomplish this task?

A. Clone

B. Pull

C. Merge

D. Push

98 # Which file format is used for storing Delta Lake Table?

A. CSV

B. Parquet

C. JSON

D. Delta

98 # Which file format is used for storing Delta Lake Table?

A. CSV

B. Parquet

C. JSON

D. Delta

99 # A data architect has determined that a table of the following format is necessary:

employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
…	…	…

Which code block is used by SQL DDL command to create an empty Delta table in the above format regardless of whether a table already exists with this name?

A. CREATE OR REPLACE TABLE table_name (employeeId STRING, startDate DATE, avgRating FLOAT)

B. CREATE OR REPLACE TABLE table_name WITH COLUMNS (employeeId STRING, startDate DATE, avgRating FLOAT) USING DELTA

C. CREATE TABLE IF NOT EXISTS table_name (employeeId STRING, startDate DATE, avgRating FLOAT)

D. CREATE TABLE table_name AS SELECT employeeId STRING, startDate DATE, avgRating FLOAT

99 # A data architect has determined that a table of the following format is necessary:

employeeId	startDate	avgRating
a1	2009-01-06	5.5
a2	2018-11-21	7.1
…	…	…

Which code block is used by SQL DDL command to create an empty Delta table in the above format regardless of whether a table already exists with this name?

A. CREATE OR REPLACE TABLE table_name (employeeId STRING, startDate DATE, avgRating FLOAT)

B. CREATE OR REPLACE TABLE table_name WITH COLUMNS (employeeId STRING, startDate DATE, avgRating FLOAT) USING DELTA

C. CREATE TABLE IF NOT EXISTS table_name (employeeId STRING, startDate DATE, avgRating FLOAT)

D. CREATE TABLE table_name AS SELECT employeeId STRING, startDate DATE, avgRating FLOAT

100 # A data engineer has been given a new record of data:

id STRING = 'a1'
rank INTEGER = 6
rating FLOAT = 9.4

Which SQL commands can be used to append the new record to an existing Delta table my_table?

A. INSERT INTO my_table VALUES ('a1', 6, 9.4)

B. INSERT VALUES ('a1', 6, 9.4) INTO my_table

C. UPDATE my_table VALUES ('a1', 6, 9.4)

D. UPDATE VALUES ('a1', 6, 9.4) my_table

100 # A data engineer has been given a new record of data:

id STRING = 'a1'
rank INTEGER = 6
rating FLOAT = 9.4

Which SQL commands can be used to append the new record to an existing Delta table my_table?

A. INSERT INTO my_table VALUES ('a1', 6, 9.4)

B. INSERT VALUES ('a1', 6, 9.4) INTO my_table

C. UPDATE my_table VALUES ('a1', 6, 9.4)

D. UPDATE VALUES ('a1', 6, 9.4) my_table

101 # A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.

Which keyword can be used to compact the small files?

A. OPTIMIZE

B. VACUUM

C. COMPACTION

D. REPARTITION

101 # A data engineer has realized that the data files associated with a Delta table are incredibly small. They want to compact the small files to form larger files to improve performance.

Which keyword can be used to compact the small files?

A. OPTIMIZE

B. VACUUM

C. COMPACTION

D. REPARTITION

102 # A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.

Which of the following data entities should the data engineer create?

A. Table

B. Function

C. View

D. Temporary view

102 # A data engineer wants to create a data entity from a couple of tables. The data entity must be used by other data engineers in other sessions. It also must be saved to a physical location.

Which of the following data entities should the data engineer create?

A. Table

B. Function

C. View

D. Temporary view

103 # A data engineer runs a statement every day to copy the previous day's sales into the table transactions. Each day's sales are in their own file in the location "/transactions/raw".

Today, the data engineer runs the following command to complete this task:

```
COPY INTO transactions
FROM "/transactions/raw"
FILEFORMAT = PARQUET;
```

After running the command today, the data engineer notices that the number of records in table transactions has not changed.

What explains why the statement might not have copied any new records into the table?

A. The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.

B. The COPY INTO statement requires the table to be refreshed to view the copied rows.

C. The previous day's file has already been copied into the table.

D. The PARQUET file format does not support COPY INTO.

103 # A data engineer runs a statement every day to copy the previous day's sales into the table transactions. Each day's sales are in their own file in the location "/transactions/raw".

Today, the data engineer runs the following command to complete this task:

```
COPY INTO transactions
FROM "/transactions/raw"
FILEFORMAT = PARQUET;
```

After running the command today, the data engineer notices that the number of records in table transactions has not changed.

What explains why the statement might not have copied any new records into the table?

A. The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.

B. The COPY INTO statement requires the table to be refreshed to view the copied rows.

C. The previous day's file has already been copied into the table.

D. The PARQUET file format does not support COPY INTO.

104 # Which command can be used to write data into a Delta table while avoiding the writing of duplicate records?

A. DROP

B. INSERT

C. MERGE

D. APPEND

104 # Which command can be used to write data into a Delta table while avoiding the writing of duplicate records?

A. DROP

B. INSERT

C. MERGE

D. APPEND

105 # A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which command could the data engineering team use to access sales in PySpark?

A. SELECT * FROM sales

B. spark.table("sales")

C. spark.sql("sales")

D. spark.delta.table("sales")

105 # A data analyst has created a Delta table sales that is used by the entire data analysis team. They want help from the data engineering team to implement a series of tests to ensure the data is clean. However, the data engineering team uses Python for its tests rather than SQL.

Which command could the data engineering team use to access sales in PySpark?

A. SELECT * FROM sales

B. spark.table("sales")

C. spark.sql("sales")

D. spark.delta.table("sales")

106 # A data engineer has created a new database using the following command:

```
CREATE DATABASE IF NOT EXISTS customer360;
```

In which location will the customer360 database be located?

A. dbfs:/user/hive/database/customer360

B. dbfs:/user/hive/warehouse

C. dbfs:/user/hive/customer360

D. dbfs:/user/hive/database

106 # A data engineer has created a new database using the following command:

CREATE DATABASE IF NOT EXISTS customer360;

In which location will the customer360 database be located?

A. dbfs:/user/hive/database/customer360

B. dbfs:/user/hive/warehouse

C. dbfs:/user/hive/customer360

D. dbfs:/user/hive/database

107 # A data engineer is attempting to drop a Spark SQL table my_table and runs the following command:

DROP TABLE IF EXISTS my_table;

After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.

What is the reason behind the deletion of all these files?

A. The table was managed

B. The table's data was smaller than 10 GB

C. The table did not have a location

D. The table was external

107 # A data engineer is attempting to drop a Spark SQL table my_table and runs the following command:

DROP TABLE IF EXISTS my_table;

After running this command, the engineer notices that the data files and metadata files have been deleted from the file system.

What is the reason behind the deletion of all these files?

A. The table was managed

B. The table's data was smaller than 10 GB

C. The table did not have a location

D. The table was external

108 # A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.

They run the following command:

```
CREATE TABLE new_table

_____
OPTIONS (
  header = "true",
  delimiter = "|"
)
LOCATION "path/to/csv"
```

Which of the following lines of code fills in the above blank to successfully complete the task?

A. FROM "path/to/csv"

B. USING CSV

C. FROM CSV

D. USING DELTA

108 # A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.

They run the following command:

```
CREATE TABLE new_table

_____
OPTIONS (
  header = "true",
  delimiter = "|"
)
LOCATION "path/to/csv"
```

Which of the following lines of code fills in the above blank to successfully complete the task?

A. FROM "path/to/csv"

B. USING CSV

C. FROM CSV

D. USING DELTA

109 # What is a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

A. Parquet files can be partitioned

B. Parquet files will become Delta tables

C. Parquet files have a well-defined schema

D. Parquet files have the ability to be optimized

109 # What is a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

A. Parquet files can be partitioned

B. Parquet files will become Delta tables

C. Parquet files have a well-defined schema

D. Parquet files have the ability to be optimized

110 # Which SQL keyword can be used to convert a table from a long format to a wide format?

A. TRANSFORM

B. PIVOT

C. SUM

D. CONVERT

110 # Which SQL keyword can be used to convert a table from a long format to a wide format?

A. TRANSFORM

B. PIVOT

C. SUM

D. CONVERT

111 # A data engineer has a Python variable table_name that they would like to use in a SQL query. They want to construct a Python code block that will run the query using table_name.

They have the following incomplete code block:

____(f"SELECT customer_id, spend FROM {table_name}")

What can be used to fill in the blank to successfully complete the task?

A. spark.delta.sql

B. spark.sql

C. spark.table

D. dbutils.sql

111 # A data engineer has a Python variable table_name that they would like to use in a SQL query. They want to construct a Python code block that will run the query using table_name.

They have the following incomplete code block:

____(f"SELECT customer_id, spend FROM {table_name}")

What can be used to fill in the blank to successfully complete the task?

A. spark.delta.sql

B. spark.sql

C. spark.table

D. dbutils.sql

112 # A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

sales

customer_id	spend	units
a1	28.94	7
a3	874.1223	
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

The data engineer runs the following query to join these tables together:

```
SELECT
sales.customer id, sales.spend,
favorite_stores.store_id
FROM sales
LEFT JOIN favorite stores
ON sales.customer id = favorite_stores.customer_id;
```

A.

```
customer_idspendstore_id
a1          28.94s1
a2          NULL s1
a4          8.99 s2
```

B.

```
customer_idspendstore_id
a1          28.94 s1
a4          8.99  s2
```

C.

```
customer_idspend store_id
a1          28.94 s1
a3          874.12NULL
a4          8.99  s2
```

D.

```
customer_idspend store_id
a1          28.94 s1
a2          NULL  s1
a3          874.12NULL
a4          8.99  s2
```

112 # A data engineer is working with two tables. Each of these tables is displayed below in its entirety.

sales

customer_id	spend	units
a1	28.94	7
a3	874.1223	
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

The data engineer runs the following query to join these tables together:

```
SELECT
sales.customer id, sales.spend,
favorite_stores.store_id
FROM sales
LEFT JOIN favorite stores
ON sales.customer id = favorite_stores.customer_id;
```

A.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a4	8.99	s2

B.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2

C.

customer_id	spend	store_id
a1	28.94	s1
a3	874.12	NULL
a4	8.99	s2

D.

customer_id	spend	store_id
a1	28.94	s1
a2	NULL	s1
a3	874.12	NULL
a4	8.99	s2

113 # A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.

Which code block successfully completes this task?

A.

```
SELECT store_id, employees,
FILTER (employees, i -> i. years_exp > 5) AS exp_employees
FROM stores;
```

B.

```
SELECT store_id, employees,
FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```

C.

```
SELECT store_id, employees,
FILTER (employees, years_exp > 5) AS exp_employees
FROM stores;
```

D.

```
SELECT store_id, employees,
CASE WHEN employees.years_exp > 5 THEN employees ELSE NULL END AS exp_employees
FROM stores;
```

113 # A data engineer needs to apply custom logic to identify employees with more than 5 years of experience in array column employees in table stores. The custom logic should create a new column exp_employees that is an array of all of the employees with more than 5 years of experience for each row. In order to apply this custom logic at scale, the data engineer wants to use the FILTER higher-order function.

Which code block successfully completes this task?

A.

```
SELECT store_id, employees,
FILTER (employees, i -> i. years_exp > 5) AS exp_employees
FROM stores;
```

B.

```
SELECT store_id, employees,
FILTER (exp_employees, i -> i.years_exp > 5) AS exp_employees
FROM stores;
```

C.

```
SELECT store_id, employees,
FILTER (employees, years_exp > 5) AS exp_employees
FROM stores;
```

D.

```
SELECT store_id, employees,
CASE WHEN employees.years_exp > 5 THEN employees ELSE NULL END AS exp_employees
FROM stores;
```

114 # A data engineer that is new to using Python needs to create a Python function to add two integers together and return the sum?

Which code block can the data engineer use to complete this task?

A.

```
function add_integers (x, y):
        return x + y
```

B.

```
def add_integers (x, y):
        print(x + y)
```

C.

```
def add_integers (x, y):
        x + y
```

D.

```
def add integers (x, y):
        return x + y
```

114 # A data engineer that is new to using Python needs to create a Python function to add two integers together and return the sum?

Which code block can the data engineer use to complete this task?

A.

```
function add_integers (x, y):
        return x + y
```

B.

```
def add_integers (x, y):
        print(x + y)
```

C.

```
def add_integers (x, y):
        x + y
```

D.

```
def add integers (x, y):
        return x + y
```

115 # A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table.

The code block used by the data engineer is below:

```
(spark.table("sales")
    .withColumn("avg_price", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .______
    .table("new_sales")
)
```

Which line of code should the data engineer use to fill in the blank if the data engineer only wants the query to execute a micro-batch to process data every 5 seconds?

A. trigger("5 seconds")

B. trigger(continuous="5 seconds")

C. trigger(once="5 seconds")

D. trigger(processingTime="5 seconds")

115 # A data engineer has configured a Structured Streaming job to read from a table, manipulate the data, and then perform a streaming write into a new table.

The code block used by the data engineer is below:

```
(spark.table("sales")
    .withColumn("avg_price", col("sales") / col("units"))
    .writeStream
    .option("checkpointLocation", checkpointPath)
    .outputMode("complete")
    .______
    .table("new_sales")
)
```

Which line of code should the data engineer use to fill in the blank if the data engineer only wants the query to execute a micro-batch to process data every 5 seconds?

A. trigger("5 seconds")

B. trigger(continuous="5 seconds")

C. trigger(once="5 seconds")

D. trigger(processingTime="5 seconds")

116 # A data engineer is maintaining a data pipeline. Upon data ingestion, the data engineer notices that the source data is starting to have a lower level of quality. The data engineer would like to automate the process of monitoring the quality level.

Which of the following tools can the data engineer use to solve this problem?

A. Auto Loader

B. Unity Catalog

C. Delta Lake

D. Delta Live Tables

116 # A data engineer is maintaining a data pipeline. Upon data ingestion, the data engineer notices that the source data is starting to have a lower level of quality. The data engineer would like to automate the process of monitoring the quality level.

Which of the following tools can the data engineer use to solve this problem?

A. Auto Loader

B. Unity Catalog

C. Delta Lake

D. Delta Live Tables

117 # A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.

Which approach can the data engineer take to identify the table that is dropping the records?

A. They can set up separate expectations for each table when developing their DLT pipeline.

B. They can navigate to the DLT pipeline page, click on the “Error” button, and review the present errors.

C. They can set up DLT to notify them via email when records are dropped.

D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.

117 # A data engineer has three tables in a Delta Live Tables (DLT) pipeline. They have configured the pipeline to drop invalid records at each table. They notice that some data is being dropped due to quality concerns at some point in the DLT pipeline. They would like to determine at which table in their pipeline the data is being dropped.

Which approach can the data engineer take to identify the table that is dropping the records?

A. They can set up separate expectations for each table when developing their DLT pipeline.

B. They can navigate to the DLT pipeline page, click on the “Error” button, and review the present errors.

C. They can set up DLT to notify them via email when records are dropped.

D. They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.

118 # What is used by Spark to record the offset range of the data being processed in each trigger in order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing?

A. Checkpointing and Write-ahead Logs

B. Replayable Sources and Idempotent Sinks

C. Write-ahead Logs and Idempotent Sinks

D. Checkpointing and Idempotent Sinks

118 # What is used by Spark to record the offset range of the data being processed in each trigger in order for Structured Streaming to reliably track the exact progress of the processing so that it can handle any kind of failure by restarting and/or reprocessing?

A. Checkpointing and Write-ahead Logs

B. Replayable Sources and Idempotent Sinks

C. Write-ahead Logs and Idempotent Sinks

D. Checkpointing and Idempotent Sinks

119 # What describes the relationship between Gold tables and Silver tables?

A. Gold tables are more likely to contain aggregations than Silver tables.

B. Gold tables are more likely to contain valuable data than Silver tables.

C. Gold tables are more likely to contain a less refined view of data than Silver tables.

D. Gold tables are more likely to contain truthful data than Silver tables.

119 # What describes the relationship between Gold tables and Silver tables?

A. Gold tables are more likely to contain aggregations than Silver tables.

B. Gold tables are more likely to contain valuable data than Silver tables.

C. Gold tables are more likely to contain a less refined view of data than Silver tables.

D. Gold tables are more likely to contain truthful data than Silver tables.

120 # What describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.

B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.

C. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.

D. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

120 # What describes when to use the CREATE STREAMING LIVE TABLE (formerly CREATE INCREMENTAL LIVE TABLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live Tables (DLT) tables using SQL?

A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.

B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.

C. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.

D. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

121 # A Delta Live Table pipeline includes two datasets defined using STREAMING LIVE TABLE. Three datasets are defined against Delta Lake table sources using LIVE TABLE.

The table is configured to run in Production mode using the Continuous Pipeline Mode.

What is the expected outcome after clicking Start to update the pipeline assuming previously unprocessed data exists and all definitions are valid?

A. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

B. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

C. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped.

D. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.

121 # A Delta Live Table pipeline includes two datasets defined using STREAMING LIVE TABLE. Three datasets are defined against Delta Lake table sources using LIVE TABLE.

The table is configured to run in Production mode using the Continuous Pipeline Mode.

What is the expected outcome after clicking Start to update the pipeline assuming previously unprocessed data exists and all definitions are valid?

A. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

B. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

C. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped.

D. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.

122 # Which type of workloads are compatible with Auto Loader?

A. Streaming workloads

B. Machine learning workloads

C. Serverless workloads

D. Batch workloads

122 # Which type of workloads are compatible with Auto Loader?

A. Streaming workloads

B. Machine learning workloads

C. Serverless workloads

D. Batch workloads

123 # A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.

Why has Auto Loader inferred all of the columns to be of the string type?

A. Auto Loader cannot infer the schema of ingested data

B. JSON data is a text-based format

C. Auto Loader only works with string data

D. All of the fields had at least one null value

123 # A data engineer has developed a data pipeline to ingest data from a JSON source using Auto Loader, but the engineer has not provided any type inference or schema hints in their pipeline. Upon reviewing the data, the data engineer has noticed that all of the columns in the target table are of the string type despite some of the fields only including float or boolean values.

Why has Auto Loader inferred all of the columns to be of the string type?

A. Auto Loader cannot infer the schema of ingested data

B. JSON data is a text-based format

C. Auto Loader only works with string data

D. All of the fields had at least one null value

124 # Which statement regarding the relationship between Silver tables and Bronze tables is always true?

A. Silver tables contain a less refined, less clean view of data than Bronze data.

B. Silver tables contain aggregates while Bronze data is unaggregated.

C. Silver tables contain more data than Bronze tables.

D. Silver tables contain less data than Bronze tables.

124 # Which statement regarding the relationship between Silver tables and Bronze tables is always true?

A. Silver tables contain a less refined, less clean view of data than Bronze data.

B. Silver tables contain aggregates while Bronze data is unaggregated.

C. Silver tables contain more data than Bronze tables.

D. Silver tables contain less data than Bronze tables.

125 # Which query is performing a streaming hop from raw data to a Bronze table?

A.
```
(spark.table("sales")
      .groupBy("store")
      .agg(sum("sales"))
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("complete")
      .table("newSales")
)
```

B.
```
(spark.read.load(rawSalesLocation)
      .write
      .mode("append")
      .table("newSales")
)
```

C.
```
(spark.table("sales")
      .withColumn("avgPrice", col("sales") / col("units"))
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("append")
      .table("newSales")
)
```

D.
```
(spark.readStream.load(rawSalesLocation)
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("append")
      .table("newSales")
)
```

125 # Which query is performing a streaming hop from raw data to a Bronze table?

A.
```
(spark.table("sales")
      .groupBy("store")
      .agg(sum("sales"))
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("complete")
      .table("newSales")
)
```

B.
```
(spark.read.load(rawSalesLocation)
      .write
      .mode("append")
      .table("newSales")
)
```

C.
```
(spark.table("sales")
      .withColumn("avgPrice", col("sales") / col("units"))
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("append")
      .table("newSales")
)
```

D.
```
(spark.readStream.load(rawSalesLocation)
      .writeStream
      .option("checkpointLocation", checkpointPath)
      .outputMode("append")
      .table("newSales")
)
```

126 # A dataset has been defined using Delta Live Tables and includes an expectations clause:

CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW

What is the expected behavior when a batch of data containing data that violates these constraints is processed?

A. Records that violate the expectation cause the job to fail.

B. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.

C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.

D. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.

126 # A dataset has been defined using Delta Live Tables and includes an expectations clause:

CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION DROP ROW

What is the expected behavior when a batch of data containing data that violates these constraints is processed?

A. Records that violate the expectation cause the job to fail.

B. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.

C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.

D. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.

127 # A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start.

Which action can the data engineer perform to improve the start up time for the clusters used for the Job?

A. They can use endpoints available in Databricks SQL

B. They can use jobs clusters instead of all-purpose clusters

C. They can configure the clusters to autoscale for larger data sizes

D. They can use clusters that are from a cluster pool

127 # A data engineer has a Job with multiple tasks that runs nightly. Each of the tasks runs slowly because the clusters take a long time to start.

Which action can the data engineer perform to improve the start up time for the clusters used for the Job?

A. They can use endpoints available in Databricks SQL

B. They can use jobs clusters instead of all-purpose clusters

C. They can configure the clusters to autoscale for larger data sizes

D. They can use clusters that are from a cluster pool

128 # A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.

Which approach can the data engineer use to set up the new task?

A. They can clone the existing task in the existing Job and update it to run the new notebook.

B. They can create a new task in the existing Job and then add it as a dependency of the original task.

C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.

D. They can create a new job from scratch and add both tasks to run concurrently.

128 # A data engineer has a single-task Job that runs each morning before they begin working. After identifying an upstream data issue, they need to set up another task to run a new notebook prior to the original task.

Which approach can the data engineer use to set up the new task?

A. They can clone the existing task in the existing Job and update it to run the new notebook.

B. They can create a new task in the existing Job and then add it as a dependency of the original task.

C. They can create a new task in the existing Job and then add the original task as a dependency of the new task.

D. They can create a new job from scratch and add both tasks to run concurrently.

129 # A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.

Which approach can the tech lead use to identify why the notebook is running slowly as part of the Job?

A. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.

B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.

C. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.

D. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

129 # A single Job runs two notebooks as two separate tasks. A data engineer has noticed that one of the notebooks is running slowly in the Job's current run. The data engineer asks a tech lead for help in identifying why this might be the case.

Which approach can the tech lead use to identify why the notebook is running slowly as part of the Job?

A. They can navigate to the Runs tab in the Jobs UI to immediately review the processing notebook.

B. They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.

C. They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.

D. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

130 # A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.

Which approach can the data engineering team use to improve the latency of the team's queries?

A. They can increase the cluster size of the SQL endpoint.

B. They can increase the maximum bound of the SQL endpoint's scaling range.

C. They can turn on the Auto Stop feature for the SQL endpoint.

D. They can turn on the Serverless feature for the SQL endpoint.

130 # A data analysis team has noticed that their Databricks SQL queries are running too slowly when connected to their always-on SQL endpoint. They claim that this issue is present when many members of the team are running small queries simultaneously. They ask the data engineering team for help. The data engineering team notices that each of the team's queries uses the same SQL endpoint.

Which approach can the data engineering team use to improve the latency of the team's queries?

A. They can increase the cluster size of the SQL endpoint.

B. They can increase the maximum bound of the SQL endpoint's scaling range.

C. They can turn on the Auto Stop feature for the SQL endpoint.

D. They can turn on the Serverless feature for the SQL endpoint.

131 # A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.

Which approach can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?

A. They can set up an Alert with a custom template.

B. They can set up an Alert with a new email alert destination.

C. They can set up an Alert with a new webhook alert destination.

D. They can set up an Alert with one-time notifications.

131 # A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to an ELT job. The ELT job has its Databricks SQL query that returns the number of input records containing unexpected NULL values. The data engineer wants their entire team to be notified via a messaging webhook whenever this value reaches 100.

Which approach can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?

A. They can set up an Alert with a custom template.

B. They can set up an Alert with a new email alert destination.

C. They can set up an Alert with a new webhook alert destination.

D. They can set up an Alert with one-time notifications.

132 # A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.

Which approach can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

A. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.

B. They can set up the dashboard's SQL endpoint to be serverless.

C. They can turn on the Auto Stop feature for the SQL endpoint.

D. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

132 # A data engineer wants to schedule their Databricks SQL dashboard to refresh once per day, but they only want the associated SQL endpoint to be running when it is necessary.

Which approach can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

A. They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.

B. They can set up the dashboard's SQL endpoint to be serverless.

C. They can turn on the Auto Stop feature for the SQL endpoint.

D. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

133 # An engineering manager wants to monitor the performance of a recent project using a Databricks SQL query. For the first week following the project's release, the manager wants the query results to be updated every minute. However, the manager is concerned that the compute resources used for the query will be left running and cost the organization a lot of money beyond the first week of the project's release.

Which approach can the engineering team use to ensure the query does not cost the organization any money beyond the first week of the project's release?

A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.

B. They can set the query's refresh schedule to end after a certain number of refreshes.

C. They can set the query's refresh schedule to end on a certain date in the query scheduler.

D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.

133 # An engineering manager wants to monitor the performance of a recent project using a Databricks SQL query. For the first week following the project's release, the manager wants the query results to be updated every minute. However, the manager is concerned that the compute resources used for the query will be left running and cost the organization a lot of money beyond the first week of the project's release.

Which approach can the engineering team use to ensure the query does not cost the organization any money beyond the first week of the project's release?

A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.

B. They can set the query's refresh schedule to end after a certain number of refreshes.

C. They can set the query's refresh schedule to end on a certain date in the query scheduler.

D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.

134 # A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.

Which command can be used to grant the necessary permission on the entire database to the new team?

A. GRANT VIEW ON CATALOG customers TO team;

B. GRANT CREATE ON DATABASE customers TO team;

C. GRANT USAGE ON CATALOG team TO customers;

D. GRANT USAGE ON DATABASE customers TO team;

134 # A new data engineering team has been assigned to work on a project. The team will need access to database customers in order to see what tables already exist. The team has its own group team.

Which command can be used to grant the necessary permission on the entire database to the new team?

A. GRANT VIEW ON CATALOG customers TO team;

B. GRANT CREATE ON DATABASE customers TO team;

C. GRANT USAGE ON CATALOG team TO customers;

D. GRANT USAGE ON DATABASE customers TO team;

135 # A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.

Which command can be used to grant full permissions on the database to the new data engineering team?

A. GRANT ALL PRIVILEGES ON TABLE sales TO team;

B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;

C. GRANT SELECT ON TABLE sales TO team;

D. GRANT ALL PRIVILEGES ON TABLE team TO sales;

135 # A new data engineering team team has been assigned to an ELT project. The new data engineering team will need full privileges on the table sales to fully manage the project.

Which command can be used to grant full permissions on the database to the new data engineering team?

A. GRANT ALL PRIVILEGES ON TABLE sales TO team;

B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;

C. GRANT SELECT ON TABLE sales TO team;

D. GRANT ALL PRIVILEGES ON TABLE team TO sales;