1 # 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
of the following commands 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 CREATE ON DATABASE team TO customers;
E. GRANT USAGE ON DATABASE customers TO team;

2 # A new data engineering team team. has been assigned to an ELT project. The new data engineering
team will need full privileges on the database customers to fully manage the project. Which of the
following commands can be used to grant full permissions on the database to the new data engineering team?
A. GRANT USAGE ON DATABASE customers TO team;
B. GRANT ALL PRIVILEGES ON DATABASE team TO customers;
C. GRANT SELECT PRIVILEGES ON DATABASE customers TO teams;
D. GRANT SELECT CREATE MODIFY USAGE PRIVILEGES ON DATABASE customers TO team;
E. GRANT ALL PRIVILEGES ON DATABASE custorners TO team;

3 # 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 of the following actions 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 be single-node
D. They can use clusters that are from a cluster pool
E. They can configure the clusters to autoscale for larger data sizes

4 # 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 of the following approaches 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. There is no way to determine why a Job task is running slowly.
E. They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

5 # 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 of the following approaches 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. ↳
E. They can set up an Alert without notifications.

6 # 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 of the following
approaches 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 reduce the cluster size of the SQL endpoint.
E. They can ensure the dashboard's SQL endpoint is not one of the included query's SQL Endpoint.

7 # 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 of
the following approaches 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.
E. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability
Optimized."

8 # 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 of the following approaches 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 cannot ensure the query does not cost the organization money beyond the first week of the project's
release.
D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.
E. They can set the query's refresh schedule to end on a certain date in the query scheduler.

9 # 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 of the following approaches 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.
E. They can clone the existing task to a new Job and then edit it to run the new notebook.

10 # 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 of the following approaches 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 cannot determine which table is dropping the records.
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.
E. They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.

11 # Which of the following Structured Streaming queries is performing a hop from a Silver table to a
Gold table?

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

12 # A data engineer is designing a data pipeline. The source system generates files in a shared
directory that is also used by other processes. As a result, the files should be kept as is and will accumulate in the directory. The data engineer needs to identify which files are new since the previous
run in the pipeline, and set up the pipeline to only ingest those new files with each run. Which of the
following tools can the data engineer use to solve this problem?
A. Unity Catalog
B. Delta Lake
C. Databricks SQL
D. Data Explorer
E. Auto Loader

13 # 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 are dropped from the target dataset and loaded into a quarantine table.
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.
E. Records that violate the expectation cause the job to fail.

14 # 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 cade 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")
```

If the data engineer only wants the query to execute a micro-batch to process data every 5 seconds,
which of the following lines of code should the data engineer use to fill in the blank?
• A. trigger("5 seconds")
• B. trigger()
C. trigger(once="5 seconds")
• D. trigger(processing Time="5 seconds")
E. trigger(continuous="5 seconds")

15 # Which of the following tools is used by Auto Loader process data incrementally?
A. Checkpointing
B. Spark Structured Streaming
C. Data Explorer
D. Unity Catalog
E. Databricks SQL

16 # Which of the following describes the relationship between Bronze tables and raw data?
A. Bronze tables contain less data than raw data files.
B. Bronze tables contain more truthful data than raw data.
C. Bronze tables contain aggregates while raw data is unaggregated.
D. Bronze tables contain a less refined view of data than raw data.
E. Bronze tables contain raw data with a schema applied.

17 # Which of the following 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 more data than Silver tables.
E. Gold tables are more likely to contain truthful data than Silver tables.

18 # 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, which of the following two approaches
is used by Spark to record the offset range of the data being processed in each trigger?
A. Checkpointing and Write-ahead Logs
B. Structured Streaming cannot record the offset range of the data being processed in each trigger.
C. Replayable Sources and Idempotent Sinks
D. Write-ahead Logs and Idempotent Sinks
E. Checkpointing and Idempotent Sinks

19 # 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. Assuming previously unprocessed data exists
and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?
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 persist without any processing. The compute resources
will persist but go unused.
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.
E. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for
additional testing.

20 # 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. Unity Catalog
B. Data Explorer
C. Delta Lake
D. Delta Live Tables
E. Auto Loader

21 # 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. Database
B. Function
C. View

D. Temporary view
E. Table

22 # A data engineer is attempting to drop a Spark SQL table my_table. The data engineer wants to
delete all table metadata and data. They run the following command:
DROP TABLE IF EXISTS my_table - While the object no longer appears when they run SHOW TABLES,
the data files still exist. Which of the following describes why the data files still exist and the metadata
files were deleted?
A. The table's data was larger than 10 GB
B. The table's data was smaller than 10 GB
C. The table was external
D. The table did not have a location
E. The table was managed

23 # A data engineer only wants to execute the final block of a Python program if the Python variable day_of_week is equal to 1 and the Python variable review_period is True. Which of the following control flow statements should the data engineer use to begin this "conditionally executed code block?

- if day_of_week = 1 and review_period:
- if day_of_
  _week = 1 and review

_period = "True":

- if day_of_week == 1 and review_period == "True":
- if day_of_week == 1 and review_period:
- if day_of_week= 1 & review_period: = "True":

24 # A data engineering team has two tables. The first table march_transactions is a collection of all retail transactions in the month of March. The second table april_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables. Which of the following commands should be run to create a new table all_transactions that contains all records

from march_transactions and april_transactions without duplicate records?

A. CREATE TABLE all_transactions AS

SELECT * FROM march_transactions

INNERJOIN SELECT * FROM april _transactions;

B. CREATE TABLE all transactions AS

SELECT * FROM march transactions

UNION SELECT * FROM april transactions:|

C. CREATE TABLE a!!_transactions AS

SELECT * FROM march_transactions

OUTER JOIN SELECT * FROM april_transactions; CREATE TABLE all_transactions AS

SELECT * FROM march_transactions

INTERSECT SELECT * from april _transactions;

E. CREATE TABLE all_transactions AS

SELECT * FROM march_transactions

MERGE SELECT * FROM april_transactions;

25 # A data engineer needs to create a table in Databricks using data from their organization's existing SQLite database. They run the following command:
CREATE TABLE jabc_customer360
USING
OPTIONS (
url "jdbc:sqlite:/customers.db",
dbtable "customer360"
Which of the following lines of code fills in the above blank to successfully complete the task?

- org.apache.spark.sql.jdbc
- autoloader
- DELTA
- sqlite
- org.apache.spark.sql.sqlite

26 # 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. Which of the following describes why the statement might not have copied any new records into the table?

- The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.
- The names of the files to be copied were not included with the FILES keyword.
- The previous day's file has already been copied into the table.
- The PARQUET file format does not support COPY INTO.
- The COPY INTO statement requires the table to be refreshed to view the copied rows.

27 # A data analyst has a series of queries in a SQL program. The data analyst wants this program to run every day. They only want the final query in the program to run on Sundays. They ask for help from the data engineering team to complete this task. Which of the following approaches could be used by the data engineering team to complete this task?

- They could submit a feature request with Databricks to add this functionality.
- They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.
- They could only run the entire program on Sundays.
- They could automatically restrict access to the source table in the final query so that it is only accessible on

Sundays.
E. They could redesign the data model to separate the data used in the final query into a new table.

28 # A data engineer needs to apply custom logic to string column city in table stores for a specific use case. In order to apply this custom logic at scale, the data engineer wants to create a SQL user-
defined function (UDF). Which of the following code blocks creates this SQL UDF?
Answer
CREATE FUNCTION combine_nyc (city STRING)
RETURNS STRING
RETURN CASE
WHEN city = "brooklyn" THEN "new york"
ELSE city
END;

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

- DROP
- IGNORE
- MERGE
- APPEND
- INSERT

30 # Which of the following benefits is provided by the array functions from Spark SQL?

- An ability to work with data in a variety of types at once
- An ability to work with data within certain partitions and windows"
- An ability to work with time-related data in specified intervals
- An ability to work with complex, nested data ingested from JSON files
- An ability to work with an array of tables for procedural automation

31 # A data engineer wants to create a new table containing the names of customers that live in France. They have written the following command:
CREATE TABLE customersInFrance
AS
SELECT id, firstName,
lastName,
FROM
customerLocations
WHERE country = 'FRANCE' ;
A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).
Which of the following lines of code fills in the above blank to successtully complete the task?

- There is no way to indicate whether a table contains PII.
- "COMMENT PII"
- TBLPROPERTIES PII
- COMMENT "Contains PII"

32 # Which of the following commands will return the location of database customersou?

A. DESCRIBE LOCATION customer360;

B. DROP DATABASE customer360;

C. DESCRIBE DATABASE customer360;

D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' =
'/user');

E. USE DATABASE customer

33 # 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 of the following commands could the data engineering team use to access sales in PySpark?

- SELECT * FROM sales
- There is no way to share data between PySpark and SQL.
- spark.sql("sales")
- spark.delta.table("sales")
- spark.table("sales")

34 # A data engineer has left the organization. The data team needs to transfer ownership of the data engineer's Delta tables to a new data engineer. The new data engineer is the lead engineer on the data team. Assuming the original data engineer no longer has access, which of the following individuals must be the one to transfer ownership of the Delta tables in Data Explorer?

- Databricks account representative
- This transfer is not possible
- Workspace administrator
- New lead data engineer
- Original data engineer

35 # A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos. Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

- Databricks Repos automatically saves development progress

- Databricks Repos supports the use of multiple branches
- Databricks Repos allows users to revert to previous versions of a notebook
- Databricks Repos provides the ability to comment on specific changes
- Databricks Repos is wholly housed within the Databricks Lakehouse Platform

36 # Which of the following datalakehouse features results in improved data quality over a traditional data lake?

- A data lakehouse provides storage solutions for structured and unstructured data.
- A data lakehouse supports ACID-compliant transactions.
- A data lakehouse allows the use of SQL queries to examine data.
- A data lakehouse stores data in open formats.
- A data lakehouse enables machine learning and artificial Intelligence workloads.

37 # Which of the following Git operations must be performed outside of Databricks Repos?

- Commit
- Pull
- Push I
- Clone
- Merge

38 # A data engineer has realized that they made a mistake when making a daily update to a table.
They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted. Which of the following explains why the data files are no longer present?

- The VACUUM command was run on the table
- The TIME TRAVEL command was run on the table
- The DELETE HISTORY command was run on the table
- The OPTIMIZE command was nun on the table
- The HISTORY command was run on the table

39 # Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my_table and save the updated table?

- SELECT * FROM my_table WHERE age > 25;
- UPDATE my_table WHERE age > 25;
- DELETE FROM my_table WHERE age > 25;
- UPDATE my_table WHERE age <= 25;
- DELETE FROM my_table WHERE age <= 25;

40 # Which of the following describes the storage organization of a Delta table?

- Delta tables are stored in a single file that contains data, history, metadata, and other attributes.
- Delta tables store their data in a single file and all metadata in a collection of files in a separate location.
- Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.
- Delta tables are stored in a collection of files that contain only the data stored within the table.
- Delta tables are stored in a single file that contains only the data stored within the table.

41 # Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

- The ability to manipulate the same data using a variety of languages
- The ability to collaborate in real time on a single notebook

- The ability to set up alerts for query failures
- The ability to support batch and streaming workloads
- The ability to distribute complex data operations

42 # Which of the following is hosted completely in the control plane of the classic Databricks architecture?

- Worker node
- JDBC data source
- Databricks web application
- Databricks Filesystem
- Driver node

43 # Which of the following describes a scenario in which a data team will want to utilize cluster pools?

- An automated report needs to be refreshed as quickly as possible.
- An automated report needs to be made reproducible.
- An automated report needs to be tested to identify errors.
- An automated report needs to be version-controlled across multiple collaborators.
- An automated report needs to be runnable by all stakeholders.

44 # A data organization leader is upset about the data analysis team's reports being different from the data engineering team's reports. The leader believes the siloed nature of their organization's data engineering and data analysis architectures is to blame. Which of the following describes how a data lakehouse could alleviate this issue?

- Both teams would autoscale their work as data size evolves
- Both teams would use the same source of truth for their work
- Both teams would reorganize to report to the same department
- Both teams would be able to collaborate on projects in real-time
- Both teams would respond more quickly to ad-hoc requests

45 # 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 of the following commands can be used to grant full permissions on the database to the new data engineering team?

- GRANT ALL PRIVILEGES ON TABLE sales TO team;
- GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
- GRANT SELECT ON TABLE sales TO team;
- GRANT USAGE ON TABLE sales TO team;
- GRANT ALL PRIVILEGES ON TABLE team TO sales;

46 # 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 of the following Git operations does the data engineer need to run to accomplish this task?

- Merge
- Push
- Pull
- Commit
- Clone

47 # Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

- Cloud-specific integrations
- Simplified governance
- Ability to scale storage
- Ability to scale workloads
- Avoiding vendor lock-in

48 # 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 of the following locations can the data engineer review their permissions on the table?

- Databricks Filesystem
- Jobs
- Dashboards
- Repos
- Data Explorer

49 # Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?

- When they are working interactively with a small amount of data
- When they are running automated reports to be refreshed as quickly as possible
- When they are working with SQL within Databricks SQL
- When they are concerned about the ability to automatically scale with larger data
- When they are manually running reports with a large amount of data

50 # A data engineer has been given a new record of data:
id STRING = 'a1'
rank INTEGER = 6
rating FLOAT = 9.4
Which of the following SQL commands can be used to append the new record to an existing Delta table my_table?

- INSERT INTO my_table VALUES ('a1', 6, 9.4)
- my_table UNION VALUES ('a1', 6, 9.4)
- INSERT VALUES ('a1', 6, 9.4) INTO my _table
- UPDATE my _table VALUES ('a1', 6, 9.4)
- UPDATE VALUES ('a1', 6, 9.4) my_table

51 # 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 of the following keywords can be used to compact the small files?

- REDUCE
- OPTIMIZE
- COMPACTION
- REPARTITION
- VACUUM

52 # In which of the following file formats is data from Delta Lake tables primarily stored?

- Delta
- CSV
- Parqueti
- JSON

°>
E. A proprietary, optimized format specific to Databricks

53 # Which of the following is stored in the Databricks Customer's cloud account?

- Databricks web application
- Cluster management metadata
- Repos
- Data
- Notebooks

54 # Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

- None of these
- Data lake
- Data warehouse
- All of these
- Data lakehouse

55 # A data architect has determined that a table of the following format is necessary:
Which of the following code blocks uses SQL DDL commands to
* table in the above format regardless of
whether a table already exists with this name?
T
C.

employeeIdstartDate avgRating

a 2009-01-065.5
1

a2 2018-1

Answer

CREATE OR REPLACE TABLE table_name employeeId STRING, startDate DATE, avgRating FLOAT

56 # A data engineer has a Python notebook in Databricks, but they need to use SQL to accomplish a specific task within a cell. They still want all of the other cells to use Python without making any changes to those cells.
Which of the following describes how the data engineer can use SQL within a cell of their Python notebook?

- It is not possible to use SQL in a Python notebook
- They can attach the cell to a SQL endpoint rather than a Databricks cluster
- They can simply write SQL syntax in the cell
- They can add %sql to the first line of the cell
- They can change the default language of the notebook to SQL

57 # Which of the following SQL keywords can be used to convert a table from a long format to a wide format?

- TRANSFORM
- PIVOT
- SUM
- CONVERT
- WHERE

58 # Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

- Parquet files can be partitioned
- CREATE TABLE AS SELECT statements cannot be used on files
- Parquet files have a well-defined schema
- Parquet files have the ability to be optimized
- Parquet files will become Delta tables

59 # A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data.
Which of the following relational objects should the data engineer create?

- Spark SQL Table
- View
- Database
- Temporary view
- Delta Table

60 # A data analyst has developed a query that runs against Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean.
However, the data engineering team uses Python for its tests rather than SQL.
Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?

- SELECT * FROM sales
- spark.delta.table
- spark.sql
- There is no way to share data between PySpark and SQL.
- spark.table

61 # Which of the following commands will return the number of null values in the member_id column?

- SELECT count(member_id) FROM my_table;
- SELECT count(member_id) - count_null(member_id) FROM my_table;
- SELECT count_if(member_id IS NULL) FROM my_table;
- SELECT null(member_id) FROM my_table;
- SELECT count_null(member_id) FROM my_table;

62 # 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 of the following code blocks successfully completes this task?
Answer
SELECT
store_1d, employees,

FROM stores;
FILTER (employees, i →> i. years_exp > 5) AS exp_employees

63 # 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}")
Which of the following can be used to fill in the blank to successfully complete the task?

- spark.delta.sql
- spark.delta.table
- spark.table
- dbutils.sql
- spark.sql

64 # A data engineer has created a new database using the following command:
CREATE DATABASE IF NOT EXISTS customer360;
In which of the following locations will the customer360 database be located?

- dbfs:/user/hive/database/customer360
- dbfs:/user/hive/warehouse
- dbfs:/user/hive/customer360
- More information is needed to determine the correct response
- dbfs:/user/hive/database

65 # 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.
Which of the following describes why all of these files were deleted?

- The table was managed
- The table's data was smaller than 10 GB
- The table's data was larger than 10 GB
- The table was external
- The table did not have a location

66 # 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 of the following code blocks can the data engineer use to complete this task?
A.
C.
D.
E.
function add
integers (x, Y) :
return x
+ Y
function add_ integers (x, y) :
x+ Y
def add_ integers (x, y) :
print (x + y)
def add_ integers (x, y) :
return x + y

def add_integers (x, Y) :
x + y

67 # In which of the following scenarios should a data engineer use the MERGE INTO command instead of the INSERT INTO command?

- When the location of the data needs to be changed
- When the target table is an external table
- When the source table can be deleted
- When the target table cannot contain duplicate records
- When the source is not a Delta table

68 # A data engineer is working with two tables. Each of these tables is displayed

below in its entirety. sales

customer_id

Le

a3

a4 favorite_stores customer_id

a 1

a2

a4 spend

28.94

874.12

8.99

units

7

23

1 store_id

s1

s1 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;

Answer

| customer_id | spend | store_id |
|---|---|---|
| 리 | 28.94 | s1 |
| | 874.12 | NULL |
| a4 | 8.9 | |

69 # 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?

- None of these lines of code are needed to successfully complete the task
- USING CSV
- FROM CSV
- USING DELTA
- FROM "path/to/csv"

70 # 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:
If the data engineer only wants the query to process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

- processingTime (1)
- triggeravailableNow=True)
- trigger(parallelBatch=True)
- trigger(processingTime="once")

- trigger(continuous="once")

(spark.readStream
• table ("sales")
withColumn ("avg_price", col("sales") / col ("units"))

- writeStream
- option ("checkpointLocation", checkpointFath)
- outputMode ("complete")
- table ("new _sales")

71 # 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.
Which of the following describes why Auto Loader inferred all of the columns to be of the string type?

- There was a type mismatch between the specific schema and the inferred schema
- JSON data is a text-based format
- Auto Loader only works with string data
- All of the fields had at least one null value
- Auto Loader cannot infer the schema of ingested data

72 # 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 Development mode using the Continuous Pipeline Mode.
Assuming previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?

- All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.
- All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist until the pipeline is shut down.
- All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.
- All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.
- All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.

73 # Which of the following data workloads will utilize a Gold table as its source?

- A job that enriches data by parsing its timestamps into a human-readable format
- A job that aggregates uncleaned data to create standard summary statistics
- A job that cleans data by removing malformatted records
- A job that queries aggregated data designed to feed into a dashboard
- A job that ingests raw data from a streaming source into the Lakehouse

74 # Which of the following must be specified when creating a new Delta Live Tables pipeline?

- A key-value pair configuration
- The preferred DBU/hour cost
- A path to cloud storage location for the written data
- A location of a target database for the written data
- At least one notebook library to be executed

75 # A data engineer has joined an existing project and they see the following query in the project repository:
CREATE STREAMING LIVE TABLE loyal_customers AS
SELECT customer_id -
FROM STREAM(LIVE.customers)
WHERE loyalty_level = 'high';
Which of the following describes why the STREAM tunction is included in the query?

- The STREAM function is not needed and will cause an error.
- The table being created is a live table.
- The customers table is a streaming live table.
- The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.
- The data in the customers table has been updated since its last run.

76 # Which of the following describes the type of workloads that are always compatible with Auto Loader?

- Streaming workloads
- Machine learning workloads
- Serverless workloads
- Batch workloads
- Dashboard workloads

77 # A data engineer and data analyst are working together on a data pipeline. The data engineer is working on the raw, bronze, and silver layers of the pipeline using Python, and the data analyst is working on the gold layer of the pipeline using SQL. The raw source of the pipeline is a streaming input.
They now want to migrate their pipeline to use Delta Live Tables.
Which of the following changes will need to be made to the pipeline when migrating to Delta Live Tables?

- None of these changes will need to be made
- The pipeline will need to stop using the medallion-based multi-hop architecture
- The pipeline will need to be written entirely in SQL
- The pipeline will need to use a batch source in place of a streaming source
- The pipeline will need to be written entirely in Python

78 # A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:
Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

- Replace predict with a stream-friendly prediction function
- Replace schema(schema) with option ("maxFilesPerTrigger"
- Replace "transactions" with the path to the location of the Delta table
- Replace format("delta") with format("stream")
- Replace spark.read with spark.readStream

transactions df = (spark. read
. schema (schema)

- format ("delta")
- table ("transactions")

79 # A dataset has been defined using Delta Live Tables and includes an expectations clause:
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION FAIL UPDATE

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

- Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- Records that violate the expectation cause the job to fail.
- Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table.
- Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
- Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.

80 # Which of the following statements regarding the relationship between Silver tables and Bronze tables is always true?

- Silver tables contain a less refined, less clean view of data than Bronze data.
- Silver tables contain aggregates while Bronze data is unaggregated.
- Silver tables contain more data than Bronze tables.
- Silver tables contain a more refined and cleaner view of data than Bronze tables.
- Silver tables contain less data than Bronze tables.

81 # A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.
Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?
A. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."

- They can turn on the Auto Stop feature for the SQL endpoint.
- They can increase the cluster size of the SQL endpoint.
- They can turn on the Serverless feature for the SQL endpoint.
- They can increase the maximum bound of the SQL endpoint's scaling range.

82 # A data engineer has a Job that has a complex run schedule, and they want to transfer that schedule to other Jobs.
Rather than manually selecting each value in the scheduling form in Databricks, which of the following tools can the data engineer use to represent and submit the schedule programmatically?

- pyspark.sql.types.DateType
- datetime
- pyspark.sql.types.TimestampType
- Cron syntax
- There is no way to represent and submit this information programmatically

83 # Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?
A. Manually programm
each cell of the Notebook

- Setting up an Alert in the Job page
- Setting up an Alert in the Notebook
- There is no way to notify the Job owner in the case of Job failure
- MLflow Model Registry Webhooks

84 # An engineering manager uses a Databricks SQL query to monitor ingestion latency for each data source. The manager checks the results of the query every day, but they are manually rerunning the query each day and waiting for the results.

Which of the following approaches can the manager use to ensure the results of the query are updated each day?

- They can schedule the query to refresh every 1 day from the SQL endpoint's page in Databricks SQL.
- They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL.
- They can schedule the query to refresh every 1 day from the query's page in Databricks SQL.
- They can schedule the query to run every 1 day from the Jobs UI.
- They can schedule the query to run every 12 hours from the Jobs UI.

85 # In which of the following scenarios should a data engineer select a Task in the Depends On field of a new Databricks Job Task?

- When another task needs to be replaced by the new task
- When another task needs to fail before the new task begins
- When another task has the same dependency libraries as the new task
- When another task needs to use as little compute resources as possible
- When another task needs to successfully complete before the new task begins

86 # A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.
Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with $0 in sales is greater than zero?

- They can set up an Alert with a custom template.
- They can set up an Alert with a new email alert destination.
- They can set up an Alert with one-time notifications.
- They can set up an Alert with a new webhook alert destination.
- They can set up an Alert without notifications.

87 # A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that feeds the dashboard is automatically processed using a Databricks Job.
Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- They can turn on the Auto Stop feature for the SQL endpoint.
- They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.
- They can reduce the cluster size of the SQL endpoint.
- They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- They can set up the dashboard's SQL endpoint to be serverless.

88 # A data engineer needs access to a table new_table, but they do not have the correct permissions. They can ask the table owner for permission, but they do not know who the table owner is.
Which of the following approaches can be used to identify the owner of new_table?

- Review the Permissions tab in the table's page in Data Explorer
- All of these options can be used to identify the owner of the table
- Review the Owner field in the table's page in Data Explorer
- Review the Owner field in the table's page in the cloud storage solution

- There is no way to identify the owner of the table

1 # 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 of the following commands can be used to grant the necessary permission on the entire database to the new team?

- GRANT VIEW ON CATALOG customers TO team;
- GRANT CREATE ON DATABASE customers TO team;
- GRANT USAGE ON CATALOG team TO customers;
- GRANT CREATE ON DATABASE team TO customers;
- GRANT USAGE ON DATABASE customers TO team;

2 # A new data engineering team team. has been assigned to an ELT project. The new data engineering team will need full privileges on the database customers to fully manage the project. Which of the following commands can be used to grant full permissions on the database to the new data engineering team?

- GRANT USAGE ON DATABASE customers TO team;
- GRANT ALL PRIVILEGES ON DATABASE team TO customers;
- GRANT SELECT PRIVILEGES ON DATABASE customers TO teams;
- GRANT SELECT CREATE MODIFY USAGE PRIVILEGES ON DATABASE customers TO team;
- GRANT ALL PRIVILEGES ON DATABASE custorners TO team;

3 # 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 of the following actions can the data engineer perform to improve the start up time for the clusters used for the Job?

- They can use endpoints available in Databricks SQL
- They can use jobs clusters instead of all-purpose clusters
- They can configure the clusters to be single-node
- They can use clusters that are from a cluster pool
- They can configure the clusters to autoscale for larger data sizes

4 # 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 of the following approaches can the tech lead use to identify why the notebook is running slowly as part of the Job?

- They can navigate to the Tasks tab in the Jobs UI and click on the active run to review the processing notebook.
- They can navigate to the Runs tab in the Jobs UI and click on the active run to review the processing notebook.
- There is no way to determine why a Job task is running slowly.
- They can navigate to the Tasks tab in the Jobs UI to immediately review the processing notebook.

5 # 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 of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of NULL values reaches 100?

- They can set up an Alert with a custom template.
- They can set up an Alert with a new email alert destination.
- They can set up an Alert with a new webhook alert destination.
- They can set up an Alert with one-time notifications.

- They can set up an Alert without notifications.

6 # 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 of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- They can set up the dashboard's SQL endpoint to be serverless.
- They can turn on the Auto Stop feature for the SQL endpoint.
- They can reduce the cluster size of the SQL endpoint.
- They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.

7 # 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 of the following approaches can the data engineering team use to improve the latency of the team's queries?

- They can increase the cluster size of the SQL endpoint.
- They can increase the maximum bound of the SQL endpoint's scaling range.
- They can turn on the Auto Stop feature for the SQL endpoint.
- They can turn on the Serverless feature for the SQL endpoint.
- They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability

Optimized."
8 # 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 of the following approaches 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?

- They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
- They can set the query's refresh schedule to end after a certain number of refreshes.
- They cannot ensure the query does not cost the organization money beyond the first week of the project's release.
- They can set a limit to the number of individuals that are able to manage the query's refresh schedule.
- They can set the query's refresh schedule to end on a certain date in the query scheduler.

9 # 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 of the following approaches can the data engineer use to set up the new task?

- They can clone the existing task in the existing Job and update it to run the new notebook.
- They can create a new task in the existing Job and then add it as a dependency of the original task.
- They can create a new task in the existing Job and then add the original task as a dependency of the new task.
- They can create a new job from scratch and add both tasks to run concurrently.
- They can clone the existing task to a new Job and then edit it to run the new notebook.

10 # 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 of the following approaches can the data engineer take to identify the table that is dropping the records?

- They can set up separate expectations for each table when developing their DLT pipeline.
- They cannot determine which table is dropping the records.
- They can set up DLT to notify them via email when records are dropped.
- They can navigate to the DLT pipeline page, click on each table, and view the data quality statistics.
- They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.

11 # Which of the following Structured Streaming queries is performing a hop from a Silver table to a Gold table?
A.
(spark. readstream. load (rawsalesLocation)

- writeStream
- option ("checkpointLocation", checkpointPath)
- outputMode ("append" )
- table ("newSales")

D.
(spark. table ("sales")

- withColumn ("avgPrice", col("sales") / col("units"))
- writestream
- option ("checkpointLocation", checkpointPath)
- outputMode ("append" )
- table ("newSales")

B.
(spark. read. load (rawsalesLocation)

- writestream
- option ("checkpointLocation", checkpointPath)
- outputMode ("append" )
- table ("newSales")

e.(spark. table ("sales")

- groupBy ("store")
- agg (sum ("sales") )
- writeStream
- option ("checkpointLocation", checkpointPath)
- outputMode ("complete")
- table ("newSales")

C
(spark. table ("sales")
filter (col ("units") > 0)

- option ("checkpointlocation", checkpointPath)
- outputMode ("append"

12 # A data engineer is designing a data pipeline. The source system generates files in a shared directory that is also used by other processes. As a result, the files should be kept as is and will accumulate in the directory. The data

engineer needs to identify which files are new since the previous run in the pipeline, and set up the pipeline to only ingest those new files with each run. Which of the following tools can the data engineer use to solve this problem?

- Unity Catalog
- Delta Lake
- Databricks SQL
- Data Explorer
- Auto Loader

13 # 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?

- Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table.
- Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.
- Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
- Records that violate the expectation cause the job to fail.

14 # 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 cade 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")

If the data engineer only wants the query to execute a micro-batch to process data every 5 seconds, which of the following lines of code should the data engineer use to fill in the blank?

- A. trigger("5 seconds")
- B. trigger)
- C. trigger(once="5 seconds")
- D. trigger(processingTime="5 seconds")

E. trigger(continuous="5 seconds")
15 # Which of the following tools is used by Auto Loader process data incrementally?

- Checkpointing
- Spark Structured Streaming
- Data Explorer
- Unity Catalog
- Databricks SQL

16 # Which of the following describes the relationship between Bronze tables and raw data?

- Bronze tables contain less data than raw data files.
- Bronze tables contain more truthful data than raw data.
- Bronze tables contain aggregates while raw data is unaggregated.
- Bronze tables contain a less refined view of data than raw data.

- Bronze tables contain raw data with a schema applied.

17 # Which of the following describes the relationship between Gold tables and Silver tables?

- Gold tables are more likely to contain aggregations than Silver tables.
- Gold tables are more likely to contain valuable data than Silver tables.
- Gold tables are more likely to contain a less refined view of data than Silver tables.
- Gold tables are more likely to contain more data than Silver tables.
- Gold tables are more likely to contain truthful data than Silver tables.

18 # 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, which of the following two approaches is used by Spark to record the offset range of the data being processed in each trigger?

- Checkpointing and Write-ahead Logs
- Structured Streaming cannot record the offset range of the data being processed in each trigger.
- Replayable Sources and Idempotent Sinks
- Write-ahead Logs and Idempotent Sinks
- Checkpointing and Idempotent Sinks

19 # 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. Assuming previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?

- All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.
- All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.
- 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.
- All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.
- All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

20 # 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?

- Unity Catalog
- Data Explorer
- Delta Lake
- Delta Live Tables
- Auto Loader

21 # 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?

- Database
- Function
- View
- Temporary view
- Table

22 # A data engineer is attempting to drop a Spark SQL table my_table. The data engineer wants to delete all table metadata and data. They run the following command:
DROP TABLE IF EXISTS my_table - While the object no longer appears when they run SHOW TABLES, the data files still exist. Which of the following describes why the data files still exist and the metadata files were deleted?

- The table's data was larger than 10 GB
- The table's data was smaller than 10 GB
- The table was external
- The table did not have a location
- The table was managed

23 # A data engineer only wants to execute the final block of a Python program if the Python variable day_of_week is equal to 1 and the Python variable review_period is True. Which of the following control flow statements should the data engineer use to begin this conditionally executed code block?

- if day_of_week = 1 and review_period:
- if day_of
  _week = 1 and review_period = "True":
- if day_of_week== 1 and review_period == "True":
- if day_of_week == 1 and review_period:
- if day_of_week = 1 & review_period: = "True":

24 # A data engineering team has two tables. The first table march_transactions is a collection of all retail transactions in the month of March. The second table april_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables.Which of the following commands should be run to create a new table all_transactions that contains all records

from march_transactions and april_transactions without duplicate records?

A. CREATE TABLE all_ transactions AS

SELECT * FROM march_transactions

INNER JOIN SELECT * FROM april_transactions;

b.CREATE TABLE all_transactions AS

SELECT * FROM march_transactions

UNION SELECT * FROM april_ _transactions;

C.

CREATE TABLE all

_transactions AS

SELECT * FROM march_transactions

OUTER JOIN SELECT * FROM april_transactions;

D. CREATE TABLE all _transactions AS

SELECT * FROM march_transactions

INTERSECT SELECT * from april_transactions;

E. CREATE TABLE all_transactions AS

SELECT * FROM march transactions

MERGE SELECT * FROM

25 # A data engineer needs to create a table in Databricks using data from their organization's existing SQLite database. They run the following command:
CREATE TABLE jabc_customer360
USING
OPTIONS (
url "jdbc:sqlite:/customers.db",
dbtable "customer360"
Which of the following lines of code fills in the above blank to successfully complete the task?

- org.apache.spark.sql.jdbc
- autoloader
- DELTA
- sqlite
- org.apache.spark.sql.sqlite

26 # 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. Which of the following describes why the statement might not have copied any new records into the table?

- The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.
- The names of the files to be copied were not included with the FILES keyword.
- The previous day's file has already been copied into the table.
- The PARQUET file format does not support COPY INTO.
- The COPY INTO statement requires the table to be refreshed to view the copied rows.

27 # A data analyst has a series of queries in a SQL program. The data analyst wants this program to run every day. They only want the final query in the program to run on Sundays. They ask for help from the data engineering team to complete this task. Which of the following approaches could be used by the data engineering team to complete this task?

- They could submit a feature request with Databricks to add this functionality.
- They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.
- They could only run the entire program on Sundays.
- They could automatically restrict access to the source tabie in the final query so that it is only accessible on

Sundays.

E. They could redesign the data model to separate the data used in the final query into a new table.

28 # A data engineer needs to apply custom logic to string column city in table stores for a specific use case. In order to apply this custom logic at scale, the data engineer wants to create a SQL user-

defined function (UDF). Which of the following code blocks creates this SQL UDF?

A.

B.

CREATE FUNCTION combine_nyc (city STRING)

RETURNS STRING

RETURN CASE

WHEN

city = "brooklyn" THEN "new york"

ELSE city

END;

bCREATE UDE combine_ nyc (city STRING)

RETURNS STRING

CASE

WHEN city = "brooklyn" THEN "new york"

ELSE city

END;

D.

CREATE FUNCTION combine_ nyc (city STRING)

RETURN CASE

WHEN city = "brooklyn" THEN "new york"

ELSE city

END;

CREATE UDE combine_nyc (city STRING)

RETURNS STRING

RETURN CASE

WHEN city = "brooklyn" THEN "new york"

ELSE city

END;

C.

CREATE UDE combine_nyc (city STRING)

RETURN CASE

WHEN city = "brooklyn" THEN "new york"

ELSE city

END;

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

- DROP
- IGNORE
- MERGE
- APPEND
- INSERT

30 # Which of the following benefits is provided by the array functions from Spark SQL?

- An ability to work with data in a variety of types at once
- An ability to work with data within certain partitions and windows
- An ability to work with time-related data in specified intervals
- An ability to work with complex, nested data ingested from JSON files
- An ability to work with an array of tables for procedural automation

31 # A data engineer wants to create a new table containing the names of customers that live in France. They have written the following command:
CREATE TABLE customersInFrance
AS
SELECT id, firstName,
lastName,
FROM customerLocations
WHERE country = 'FRANCE' ;
A senior data engineer mentions that it is organization policy to include a table property indicating that the new table includes personally identifiable information (PII).
Which of the following lines of code fills in the above blank to successfully complete the task?

- There is no way to indicate whether a table contains PII.
- "COMMENT PII"
- TBLPROPERTIES PII
- COMMENT "Contains PII"

32 # Which of the following commands will return the location of database customer360?

- DESCRIBE LOCATION customer360;
- DROP DATABASE customer360;
- DESCRIBE DATABASE customer360;
- ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user'};
- USE DATABASE customer360;

33 # 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 of the following commands could the data engineering team use to access sales in PySpark?

- SELECT * FROM sales
- There is no way to share data between PySpark and SQL.
- spark.sql("sales")
- spark.delta.table("sales")
- spark.table("sales")

34 # A data engineer has left the organization. The data team needs to transfer ownership of the data engineer's Delta tables to a new data engineer. The new data engineer is the lead engineer on the data team. Assuming the original data engineer no longer has access, which of the following individuals must be the one to transfer ownership of the Delta tables in Data Explorer?

- Databricks account representative
- This transfer is not possible
- Workspace administrator
- New lead data engineer
- Original data engineer

35 # A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version their project using Databricks Repos. Which of the following is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

- Databricks Repos automatically saves development progress
- Databricks Repos supports the use of multiple branches
- Databricks Repos allows users to revert to previous versions of a notebook
- Databricks Repos provides the ability to comment on specific changes
- Databricks Repos is wholly housed within the Databricks Lakehouse Platform

36 # Which of the following datalakehouse features results in improved data quality over a traditional data lake?

- A data lakehouse provides storage solutions for structured and unstructured data.
- A data lakehouse supports ACID-compliant transactions.
- A data lakehouse allows the use of SQL queries to examine data.
- A data lakehouse stores data in open formats.
- A data lakehouse enables machine learning and artificial Intelligence workloads.

37 # Which of the following Git operations must be performed outside of Databricks Repos?

- Commit
- Pull
- Push
- Clone
- Merge

38 # A data engineer has realized that they made a mistake when making a daily update to a table.
They need to use Delta time travel to restore the table to a version that is 3 days old. However, when the data engineer attempts to time travel to the older version, they are unable to restore the data because the data files have been deleted. Which of the following explains why the data files are no longer present?

- The VACUUM command was run on the table
- The TIME TRAVEL command was run on the table
- The DELETE HISTORY command was run on the table
- The OPTIMIZE command was nun on the table
- The HISTORY command was run on the table

39 # Which of the following code blocks will remove the rows where the value in column age is greater than 25 from the existing Delta table my_table and save the updated table?

- SELECT * FROM my_table WHERE age > 25;
- UPDATE my_table WHERE age > 25;
- DELETE FROM my_table WHERE age > 25;
- UPDATE my_table WHERE age <= 25;
- DELETE FROM my_table WHERE age <= 25;

40 # Which of the following describes the storage organization of a Delta table?

- Delta tables are stored in a single file that contains data, history, metadata, and other attributes.
- Delta tables store their data in a single file and all metadata in a collection of files in a separate location.
- Delta tables are stored in a collection of files that contain data, history, metadata, and other attributes.
- Delta tables are stored in a collection of files that contain only the data stored within the table.
- Delta tables are stored in a single file that contains only the data stored within the table.

41 # Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

- The ability to manipulate the same data using a variety of languages
- The ability to collaborate in real time on a single notebook
- The ability to set up alerts for query failures
- The ability to support batch and streaming workloads
- The ability to distribute complex data operations

42 # Which of the following is hosted completely in the control plane of the classic Databricks architecture?

- Worker node
- JDBC data source
- Databricks web application
- Databricks Filesystem
- Driver node

43 # Which of the following describes a scenario in which a data team will want to utilize cluster pools?

- An automated report needs to be refreshed as quickly as possible.
- An automated report needs to be made reproducible.
- An automated report needs to be tested to identify errors.
- An automated report needs to be version-controlled across multiple collaborators.
- An automated report needs to be runnable by all stakeholders.

44 # A data organization leader is upset about the data analysis team's reports being different from the data engineering team's reports. The leader believes the siloed nature of their organization's data engineering and data analysis architectures is to blame. Which of the following describes how a data lakehouse could alleviate this issue?

- Both teams would autoscale their work as data size evolves
- Both teams would use the same source of truth for their work
- Both teams would reorganize to report to the same department
- Both teams would be able to collaborate on projects in real-time
- Both teams would respond more quickly to ad-hoc requests

45 # 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 of the following commands can be used to grant full permissions on the database to the new data engineering team?

- GRANT ALL PRIVILEGES ON TABLE sales TO team;
- GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
- GRANT SELECT ON TABLE sales TO team;
- GRANT USAGE ON TABLE sales TO team;
- GRANT ALL PRIVILEGES ON TABLE team TO sales;

46 # 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 of the following Git operations does the data engineer need to run to accomplish this task?

- Merge
- Push
- Pull
- Commit
- Clone

47 # Which of the following is a benefit of the Databricks Lakehouse Platform embracing open source technologies?

- Cloud-specific integrations
- Simplified governance
- Ability to scale storage
- Ability to scale workloads
- Avoiding vendor lock-in

48 # 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 of the following locations can the data engineer review their permissions on the table?

- Databricks Filesystem
- Jobs
- Dashboards
- Repos
- Data Explorer

49 # Which of the following describes a scenario in which a data engineer will want to use a single-node cluster?

- When they are working interactively with a small amount of data

- When they are running automated reports to be refreshed as quickly as possible
- When they are working with SQL within Databricks SQL
- When they are concerned about the ability to automatically scale with larger data
- When they are manually running reports with a large amount of data

50 # A data engineer has been given a new record of data:
id STRING = 'a1'
rank INTEGER = 6
rating FLOAT = 9.4
Which of the following SQL commands can be used to append the new record to an existing Delta table my_table?

- INSERT INTO my_table VALUES ('a1', 6, 9.4)
- my_table UNION VALUES ('a1', 6, 9.4)
- INSERT VALUES ('a1', 6, 9.4) INTO my_table
- UPDATE my_table VALUES ('a1', 6, 9.4)
- UPDATE VALUES ('a1', 6, 9.4) my _table

51 # 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 of the following keywords can be used to compact the small files?

- REDUCE
- OPTIMIZE
- COMPACTION
- REPARTITION
- VACUUM

52 # In which of the following file formats is data from Delta Lake tables primarily stored?

- Delta
- CSV
- Parquet
- JSON
- A proprietary, optimized format specific to Databricks

53 # Which of the following is stored in the Databricks customer's cloud account?

- Databricks web application
- Cluster management metadata
- Repos
- Data
- Notebooks

54 # Which of the following can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

- None of these
- Data lake
- Data warehouse
- All of these
- Data lakehouse

55 # A data architect has determined that a table of the following format is necessary: Which of the following code blocks uses SQL DDL commands to create an empty Delta table in the above format regardless of

whether a table already exists with this name?

Answer

CREATE OR REPLACE TABLE table_name
employeeId STRING, startDate DATE,
56 # A data engineer has a Python notebook in Databricks, but they need to use SQL to accomplish a specific task within a cell. They still want all of the other cells to use Python without making any changes to those cells.
Which of the following describes how the data engineer can use SQL within a cell of their Python notebook?

- It is not possible to use SQL in a Python notebook
- They can attach the cell to a SQL endpoint rather than a Databricks cluster
- They can simply write SQL syntax in the cell
- They can add %sql to the first line of the cell
- They can change the default language of the notebook to SQL

57 # Which of the following SQL keywords can be used to convert a table from a long format to a wide format?

- TRANSFORM
- PIVOT
- SUM
- CONVERT
- WHERE

58 # Which of the following describes a benefit of creating an external table from Parquet rather than CSV when using a CREATE TABLE AS SELECT statement?

- Parquet files can be partitioned
- CREATE TABLE AS SELECT statements cannot be used on files
- Parquet files have a well-defined schema
- Parquet files have the ability to be optimized
- Parquet files will become Delta tables

59 # A data engineer wants to create a relational object by pulling data from two tables. The relational object does not need to be used by other data engineers in other sessions. In order to save on storage costs, the data engineer wants to avoid copying and storing physical data.
Which of the following relational objects should the data engineer create?

- Spark SQL Table
- View
- Database
- Temporary view
- Delta Table

60 # A data analyst has developed a query that runs against Delta table. They want help from the data engineering team to implement a series of tests to ensure the data returned by the query is clean.
However, the data engineering team uses Python for its tests rather than SQL.
Which of the following operations could the data engineering team use to run the query and operate with the results in PySpark?

- SELECT * FROM sales
- spark.delta.table
- spark.sq
- There is no way to share data between PySpark and SQL.

- spark.table

61 # Which of the following commands will return the number of null values in the member_id column?

- SELECT count(member_id) FROM my_table;
- SELECT count(member_id) - count_

_null(member_id) FROM my_table;

- SELECT count_if(member_id IS NULL) FROM my_table;
- SELECT null(member_id) FROM my_table;
- SELECT count_null(member_id) FROM my_table;

62 # 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 of the following code blocks successfully completes this task?

Answer

SELECT
store_id, employees,
FROM stores;
FILTER (employees, i → i. years_exp > 5) As exp_employees

63 # 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)")
Which of the following can be used to fill in the blank to successfully complete the task?

- spark.delta.sql
- spark.delta.table
- spark.table
- dbutils.sal
- spark.sql

64 # A data engineer has created a new database using the following command:
CREATE DATABASE IF NOT EXISTS customer360;
In which of the following locations will the customer360 database be located?

- dbfs:/user/hive/database/customer360
- dbfs:/user/hive/warehouse
- dbfs:/user/hive/customer360
- More information is needed to determine the correct response
- dbfs:/user/hive/database

65 # 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.

Which of the following describes why all of these files were deleted?

- The table was managed
- The table's data was smaller than 10 GB
- The table's data was larger than 10 GB
- The table was external
- The table did not have a location

66 # 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 of the following code blocks can the data engineer use to complete this task?

A.

function add_integers(x, Y):

return x

B.

function add_ integers (x, Y) :

X+Y

C.

def add_integers (x, y) :

print (x + y)

D.

def add_integers (x, y) :

return x + y

* E.

def add_ integers (x, y) :

X + Y

67 # In which of the following scenarios should a data engineer use the MERGE INTO command instead of the INSERT INTO command?

- When the location of the data needs to be changed
- When the target table is an external table
- When the source table can be deleted
- When the target table cannot contain duplicate records
- When the source is not a Delta table

69 # 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/c.sv"
Which of the following lines of code fills in the above blank to successfully complete the task?

- None of these lines of code are needed to successfully complete the task
- USING CSV
- FROM CSV
- USING DELTA
- FROM "path/to/csv"

70 # 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:
If the data engineer only wants the query to process all of the available data in as many batches as required, which of the following lines of code should the data engineer use to fill in the blank?

- processingTime (1)
- trigger(availableNow=True)
- trigger(parallelBatch=True)
- trigger(processingTime="once")
- trigger(continuous="once")

71 # 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.
Which of the following describes why Auto Loader inferred all of the columns to be of the string type?

- There was a type mismatch between the specific schema and the inferred schema
- JSON data is a text-based format
- Auto Loader only works with string data
- All of the fields had at least one null value
- Auto Loader cannot infer the schema of ingested data

72 # 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 Development mode using the Continuous Pipeline Mode.
Assuming previously unprocessed data exists and all definitions are valid, what is the expected outcome after clicking Start to update the pipeline?

- All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.
- All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist until the pipeline is shut down.
- All datasets will be updated once and the pipeline will persist without any processing. The compute resources will persist but go unused.
- All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.

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

73 # Which of the following data workloads will utilize a Gold table as its source?

- A job that enriches data by parsing its timestamps into a human-readable format
- A job that aggregates uncleaned data to create standard summary statistics
- A job that cleans data by removing malformatted records
- A job that queries aggregated data designed to feed into a dashboard
- A job that ingests raw data from a streaming source into the Lakehouse

74 # Which of the following must be specified when creating a new Delta Live Tables pipeline?

- A key-value pair configuration
- The preferred DBU/hour cost
- A path to cloud storage location for the written data
- A location of a target database for the written data
- At least one notebook library to be executed

75 # A data engineer has joined an existing project and they see the following query in the project repository:
CREATE STREAMING LIVE TABLE loyal_customers AS
SELECT customer_id -
FROM STREAM(LIVE.customers)
WHERE loyalty_level = 'high';
Which of the following describes why the STREAM function is included in the query?

- The STREAM function is not needed and will cause an error.
- The table being created is a live table.
- The customers table is a streaming live table.
- The customers table is a reference to a Structured Streaming query on a PySpark DataFrame.
- The data in the customers table has been updated since its last run.

76 # Which of the following describes the type of workloads that are always compatible with Auto Loader?

- Streaming workloads
- Machine learning workloads
- Serverless workloads
- Batch workloads
- Dashboard workloads

77 # A data engineer and data analyst are working together on a data pipeline. The data engineer is working on the raw, bronze, and silver layers of the pipeline using Python, and the data analyst is working on the gold layer of the pipeline using SQL. The raw source of the pipeline is a streaming input.
They now want to migrate their pipeline to use Delta Live Tables.
Which of the following changes will need to be made to the pipeline when migrating to Delta Live Tables?

- None of these changes will need to be made
- The pipeline will need to stop using the medallion-based multi-hop architecture
- The pipeline will need to be written entirely in SQL
- The pipeline will need to use a batch source in place of a streaming source
- The pipeline will need to be written entirely in Python

78 # A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:
Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

- Replace predict with a stream-friendly prediction function
- Replace schema(schema) with option ("maxFilesPerTrigger"
- Replace "transactions" with the path to the location of the Delta table
- Replace format("delta") with format("stream")
- Replace spark.read with spark.readStream

79 # A dataset has been defined using Delta Live Tables and includes an expectations clause:
CONSTRAINT valid_timestamp EXPECT (timestamp > '2020-01-01') ON VIOLATION FAIL UPDATE
What is the expected behavior when a batch of data containing data that violates these constraints is processed?

- Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- Records that violate the expectation cause the job to fail.
- Records that violate the expectation are dropped from the target dataset and loaded into a quarantine table.
- Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.
- Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.

80 # Which of the following statements regarding the relationship between Silver tables and Bronze tables is always true?

- Silver tables contain a less refined, less clean view of data than Bronze data.
- Silver tables contain aggregates while Bronze data is unaggregated.
- Silver tables contain more data than Bronze tables.
- Silver tables contain a more refined and cleaner view of data than Bronze tables.
- Silver tables contain less data than Bronze tables.

81 # A data engineering team has noticed that their Databricks SQL queries are running too slowly when they are submitted to a non-running SQL endpoint. The data engineering team wants this issue to be resolved.
Which of the following approaches can the team use to reduce the time it takes to return results in this scenario?
A. They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy to "Reliability Optimized."

- They can turn on the Auto Stop feature for the SQL endpoint.
- They can increase the cluster size of the SQL endpoint.
- They can turn on the Serverless feature for the SQL endpoint.
- They can increase the maximum bound of the SQL endpoint's scaling range.

82 # A data engineer has a Job that has a complex run schedule, and they want to transfer that schedule to other Jobs.
Rather than manually selecting each value in the scheduling form in Databricks, which of the following tools can the data engineer use to represent and submit the schedule programmatically?

- pyspark.sql.types.DateType
- datetime
- pyspark.sql.types.TimestampType
- Cron syntax
- There is no way to represent and submit this information programmatically

83 # Which of the following approaches should be used to send the Databricks Job owner an email in the case that the Job fails?

- Manually programming in an alert system in each cell of the Notebook
- Setting up an Alert in the Job page
- Setting up an Alert in the Notebook
- There is no way to notify the Job owner in the case of Job failure
- MLflow Model Registry Webhooks

84 # An engineering manager uses a Databricks SQL query to monitor ingestion latency for each data source. The manager checks the results of the query every day, but they are manually rerunning the query each day and waiting for the results.
Which of the following approaches can the manager use to ensure the results of the query are updated each day?

- They can schedule the query to refresh every 1 day from the SQL endpoint's page in Databricks SQL.
- They can schedule the query to refresh every 12 hours from the SQL endpoint's page in Databricks SQL.
- They can schedule the query to refresh every 1 day from the query's page in Databricks SQL.
- They can schedule the query to run every 1 day from the Jobs UI.
- They can schedule the query to run every 12 hours from the Jobs UI.

85 # In which of the following scenarios should a data engineer select a Task in the Depends On field of a new Databricks Job Task?

- When another task needs to be replaced by the new task
- When another task needs to fail before the new task begins
- When another task has the same dependency libraries as the new task
- When another task needs to use as little compute resources as possible
- When another task needs to successfully complete before the new task begins

86 # A data engineer has been using a Databricks SQL dashboard to monitor the cleanliness of the input data to a data analytics dashboard for a retail use case. The job has a Databricks SQL query that returns the number of store-level records where sales is equal to zero. The data engineer wants their entire team to be notified via a messaging webhook whenever this value is greater than 0.
Which of the following approaches can the data engineer use to notify their entire team via a messaging webhook whenever the number of stores with $0 in sales is greater than zero?

- They can set up an Alert with a custom template.
- They can set up an Alert with a new email alert destination.
- They can set up an Alert with one-time notifications.
- They can set up an Alert with a new webhook alert destination.
- They can set up an Alert without notifications.

87 # A data engineer wants to schedule their Databricks SQL dashboard to refresh every hour, but they only want the associated SQL endpoint to be running when it is necessary. The dashboard has multiple queries on multiple datasets associated with it. The data that feeds the dashboard is automatically processed using a Databricks Job.
Which of the following approaches can the data engineer use to minimize the total running time of the SQL endpoint used in the refresh schedule of their dashboard?

- They can turn on the Auto Stop feature for the SQL endpoint.
- They can ensure the dashboard's SQL endpoint is not one of the included query's SQL endpoint.
- They can reduce the cluster size of the SQL endpoint.
- They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- They can set up the dashboard's SQL endpoint to be serverless.

88 # A data engineer needs access to a table new_table, but they do not have the correct permissions. They can ask the table owner for permission, but they do not know who the table owner is.
Which of the following approaches can be used to identify the owner of new_table?

- Review the Permissions tab in the table's page in Data Explorer
- All of these options can be used to identify the owner of the table
- Review the Owner field in the table's page in Data Explorer
- Review the Owner field in the table's page in the cloud storage solution
- There is no way to identify the owner of the table

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?'

- CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
- CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.
- CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.
- CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
- 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 stregming hop from raw data to a Bronze table?

Answer

(spark.readStream. load (rawSalesLocation)
• writeStream

- option ("checkpointlocation", checkpointPath)
- outputMode ("append" )

.table ("newSales")

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

- A data lakehouse stores data in open formats.
- A data lakehouse allows the use of SQL queries to examine data.
- A data lakehouse provides storage solutions for structured and unstructured data.
- A data lakehouse supports ACID-compliant transactions.

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

- An automated report needs to be version-controlled across multiple collaborators.
- An automated report needs to be runnable by all stakeholders.
- An automated report needs to be refreshed as quickly as possible.
- An automated report needs to be made reproducible.

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

- Worker node
- Databricks web application
- Driver node

- Databricks Filesystem

94 # A data engineer needs to determine whether to use the built-in Databricks Notebooks versioning or version theis project using Databricks Repos.
What is an advantage of using Databricks Repos over the Databricks Notebooks versioning?

- Databricks Repos allows users to revert to previous versions of a notebook
- Databricks Repos is wholly housed within the Databricks Data Intelligence Platform
- Databricks Repos provides the ability to comment on specific changes
- Databricks Repos supports the use of multiple branches

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

- Avoiding vendor lock-in
- 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?

- Jobs
- Dashboards
- Catalog Explorer
- 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 theif 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?

- Clone
- Pull
- Merge
- Push

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

- CSV
- Parquet
- JSON
- Delta

99 # A data architect has determined that a table of the following format is necessary:
employeeIdstartDate avgRating
a l a2
2009-01-06
2018-11-21
5.5
7.1
.com

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?

- CREATE OR REPLACE TABLE table_name (employeeId STRING, startDate DATE, avgRating FLOAT)
- 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)
sthithap
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
ail.com

- INSERT INTO my_table VALUES ('al', 6, 9.4)
- INSERT VALUES ('a1', 6, 9.4) INTO my_table
- UPDATE my_table VALUES ('a1', 6, 9.4)
- UPDATE VALUES ('a1', 6, 9.4) my_table

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

- OPTIMIZE
- VACUUM
- COMPACTION
- 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?

- Table
- Function
- View
- Temporary view

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

- The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.
- The COPY INTO statement requires the table to be refreshed to view the copied rows.
- The previous day's file has already been copied into the table.

sthith
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?

- DROP
- INSERT
- MERGE
- 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 gasure 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?

- SELECT * FROM sales
- spark.table("sales")
- spark.sql("sales")
- 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?

- dbfs:/user/hive/database/customer360
- dofs:/user/hive/warehouse
- dbfs:/user/hive/customer360
- 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 dala files and metadata files have been

deleted from the file system.

I.com

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

Answer : The table is managed

108 # A data engineer needs to create a table in Databricks using data from a CSV file at location /path/to/csv.
CREATE TABLE new table
They run the following command:
OPTIONS (

header = "true"
delimiter = "|"
mail.com
LOCATION "path/to/c.sv"

- FROM "path/to/csv"
- USING CSV
- FROM CSV
- USING DELTA

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

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

- Parquet files can be partitioned
- Parquet files will become Delta tables
- Parquet files have a well-defined schema
- Parquet files have the ability to be optimized

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

- TRANSFORM
- PIVOT
- SUM
- 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'SERECT customer_id, spend FROM (table_name)")
@mail.com
What can be used to fill in the blank to successtully complete the task?

- spark.delta.sql
- spark.sql
- spark.table
- dbutils.sql

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

customer_id

a1

a2

a4

store_id

I.com

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 JOy favorite stores

ON sales.customer id = favorite_stores.customer_id;

Answer

customer_idspend store_id

URUAT

28.94 sl

a3

a4

874.12NULL

8.99 52

113 # A data engineer needs to apply custom logic to identify employees with more thaff 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.

C.

gmai

SELECT store_id, employees,

SELECT store_id, employees, FILTER (employees, i - i. years_exp > 5) AS FILTER (employees, years_exp > 5) AS exp_employees exp_employees

FROM stores;

B.

SELECT store_id, employees,

agnas

FROM stores;

D.

SELECT store_id, employees,

CASE WHEN employees.years_exp >

5

THEN FILTER (exp_employees, i -> i.years_exp > 5) AS employees ELSE NULL END AS exp_employees

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?

function add _integers (x, y):

return x + y

B.

def add _integers (x, y):

print(x + y)

retron

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.

ail.
com
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 "S seconds

- trigger(once="5 seconds")
- 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?

- Auto Loader
- Unity Catalog
- Delta Lake
- 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?

- They can set up separate expectations for each table when developing their DLT pipeline.
- They can navigate to the DLT pipeline page, click on the "Error" button, and review the present errors.
- They can set up DLT to notify them via email when records are dropped.
- 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?

- Checkpointing and Write-ahead Logs
- Replayable Sources and Idempotent Sinks
- Write-ahead Logs and Idempotent Sinks
- Checkpointing and Idempotent Sinks

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

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

- 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 TARLE) syntax over the CREATE LIVE TABLE syntax when creating Delta Live
Tables (DLT) tables using SQL?

- CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
- CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.
- CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
- 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?

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

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

122 # Which type of workloads are compatible with Auto Loader?
A. Streaming workloads
< OB. Machine learning workloads

- Serverless workloads
- 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?

- Auto Loader cannot infer the schema of ingested data
- JSON data is a text-based format
- Auto Loader only works with string data
- 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?

- Silver tables contain a less refined, less clean view of data than Bronze data.
- Silver tables contain aggregates while Bronze data is unaggregated.
- Silver tables contain more data than Bronze tables.
- Silver tables contain less data than Bronze tables.

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

Answer

(spark. readstream. load (rawSalestocation)

- writeStream
- option ("checkpointlocation", checkpointPath)
- outputMode ("append" )

. table ("newSales")
CONSTRANT ele inezan EXPECT (inestem, - 2020101) ON SOLA TON DROP ROWD
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.
no

- Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.
- Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.

sthithapram
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?

- They can use endpoints available in Databricks SQL
- They can use jobs clusters instead of all-purpose clusters

C. They can confeure the duster t autocale for anger diasies
gman
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?

- They can clone the existing task in the existing Job and update it to run the new notebook.
- They can create a new task in the existing Job and then add it as a dependency of the original task.
- They can create a new task in the existing Job and then add the original task as a dependency of the new task.

sthithapragna
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?

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

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

sthithapra**
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?

- They can increase the cluster size of the SQL endpoint.
- They can increase the maximum bound of the SQL endpoint's scaling range.
- They can turn on the Auto Stop feature for the SQL endpoint.
- 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?

- They can set up an Alert with a custom template.
- They can set up an Alert with a new email alert destination.
- They can set up an Alert with a new webhook alert destination.
- 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?

- They can ensure the dashboard's SQL endpoint matches each of the queries' SQL endpoints.
- They can set up the dashboard's SQL endpoint to be serverless.
- They can turn on the Auto Stop feature for the SQL endpoint.
- 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?

- They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
- They can set the query's refresh schedule to end after a certain number of refreshes.
- They can set the query's refresh schedule to end on a certain date in the query scheduler.
- 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?

Answer

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.
om
team?

- GRANT ALL PRIVILEGES ON TABLE sales TO team;
- GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
- GRANT SELECT ON TABLE sales TO team;
- GRANT ALL PRIVILEGES ON TABLE team TO sales;

sthithapragnasya@gmar
Which command can be used to grant full permissions on the database to the new data engineering