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

sales

customer_id	spend	units
a1	28.94	7
a3	874.12	23
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

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

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

Which of the following will be returned by the above query?

- A.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2

- B.

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

- C.

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

- D.

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

Ans:

Answer*C*

2. Which of the following benefits is provided by the array functions from Spark SQL?

- A. An ability to work with data in a variety of types at once
- B. An ability to work with data within certain partitions and windows
- C. An ability to work with time-related data in specified intervals
- D. An ability to work with complex, nested data ingested from JSON files

  Ans

Answer*D*

3. Which of the following is hosted completely in the control plane of the classic Databricks architecture?

- A. Worker node
- B. JDBC data source
- C. Databricks web application
- D. Databricks Filesystem
- E. Driver node

Answer*C*

C. Databricks web application In the classic Databricks architecture, the control plane includes components like the Databricks web application, the Databricks REST API, and the Databricks Workspace. These components are responsible for managing and controlling the Databricks environment, including cluster provisioning, notebook management, access control, and job scheduling. The other options, such as worker nodes, JDBC data sources, Databricks Filesystem (DBFS), and driver nodes, are typically part of the data plane or the execution environment, which is separate from the control plane. Worker nodes are responsible for executing tasks and computations, JDBC data sources are used to connect to external databases, DBFS is a distributed file system for data storage, and driver nodes are responsible for coordinating the execution of Spark jobs.

4. Which of the following benefits of using the Databricks Lakehouse Platform is provided by Delta Lake?

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

AnswerD

5. Which of the following describes the storage organization of a Delta table?

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

AnswerC

6. 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?

- A. SELECT * FROM my_table WHERE age > 25;
- B. UPDATE my_table WHERE age > 25;
- C. DELETE FROM my_table WHERE age > 25;
- D. UPDATE my_table WHERE age <= 25;
- E. DELETE FROM my_table WHERE age <= 25;

Answer*C*

7. Which tool is used by Auto Loader to process data incrementally?

- A. Checkpointing
- B. Spark Structured Streaming
- C. Databricks SQL
- D. Unity Catalog

Answer*B*

8. Which of the following commands will return the number of null values in the member_id column?

- A. SELECT count(member_id) FROM my_table;
- B. SELECT count(member_id) - count_null(member_id) FROM my_table;
- C. SELECT count_if(member_id IS NULL) FROM my_table;
- D. SELECT null(member_id) FROM my_table;

AnswerC

9. Which of the following data lakehouse features results in improved data quality over a traditional data lake?

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

Answer*B*

10. 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?

- A. Spark SQL Table
- B. View
- C. Delta Table
- D. Temporary view

AnswerD

11. 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?

- A. Databricks account representative
- B. This transfer is not possible
- C. Workspace administrator
- D. New lead data engineer
- E. Original data engineer

Answer*C*

12. 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?

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

Answer*E*

13. Which of the following commands will return the location of database customer360?

- A. DESCRIBE LOCATION customer360;
- B. DROP DATABASE customer360;
- C. DESCRIBE DATABASE customer360;
- D. ALTER DATABASE customer360 SET DBPROPERTIES ('location' = '/user'};
- E. USE DATABASE customer360;

AnswerC

14. 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?

- A. There is no way to indicate whether a table contains PII.
- B. "COMMENT PII"
- C. TBLPROPERTIES PII
- D. COMMENT "Contains PII"
- E. PII

AnswerD

15.What is stored in the Databricks customer's cloud account?

- A. Databricks web application
- B. Cluster management metadata
- C. Notebooks
- D. Data

Answer*D*

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

- A. DROP
- B. IGNORE
- C. MERGE
- D. APPEND
- E. INSERT

AnswerC

17. 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. Auto Loader

Answer*D*

18. 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?

- A. They could submit a feature request with Databricks to add this functionality.
- B. They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query.
- C. They could only run the entire program on Sundays.
- D. 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.

Answer*B*

19. 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?

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

Answer*C*

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

- A. When the location of the data needs to be changed
- B. When the target table is an external table
- C. When the source is not a Delta table
- D. When the target table cannot contain duplicate records

AnswerD

21. 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 jdbc_customer360
USING _____
OPTIONS (
  url "jdbc:sqlite:/customers.db",
  dbtable "customer360"
)
```

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

- A. org.apache.spark.sql.jdbc
- B. autoloader
- C. org.apache.spark.sql.sqlite
- D. sqlite

Answer*A*

22. 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?

- A. Review the Permissions tab in the table's page in Data Explorer

- B. There is no way to identify the owner of the table
- C. Review the Owner field in the table's page in Data Explorer
- D. Review the Owner field in the table's page in the cloud storage solution

Answer*C*

23. 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

Answer*C*

24. 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

Answer*E*

25.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

Answer*D*

26. 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?

- A. The pipeline will need to be written entirely in Python
- B. The pipeline can have different notebook sources in SQL & Python
- C. The pipeline will need to be written entirely in SQL
- D. The pipeline will need to use a batch source in place of a streaming source

Answer*B*

27. 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?

- A. pyspark.sql.types.DateType
- B. datetime
- C. pyspark.sql.types.TimestampType
- D. Cron syntax

Answer*D*

28. 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?

- A. SELECT * FROM sales
- B. spark.delta.table
- C. spark.sql
- D. spark.table

Answer*C*

29. 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?

- A. Both teams would respond more quickly to ad-hoc requests
- B. Both teams would use the same source of truth for their work
- C. Both teams would reorganize to report to the same department

- D. Both teams would be able to collaborate on projects in real-time

Answer*B*

30. Which Structured Streaming query 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")
)
```

- B.

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

- C.

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

- D.

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

Answer*D* ;.

31. 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")

Answer*D*

32. 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.

Answer*C*

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

- A. CREATE STREAMING LIVE TABLE should be used when the subsequent step in the DLT pipeline is static.
- B. CREATE STREAMING LIVE TABLE should be used when data needs to be processed incrementally.
- C. CREATE STREAMING LIVE TABLE is redundant for DLT and it does not need to be used.
- D. CREATE STREAMING LIVE TABLE should be used when data needs to be processed through complicated aggregations.
- E. CREATE STREAMING LIVE TABLE should be used when the previous step in the DLT pipeline is static.

AnswerB

34. 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?

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

Answer C

35. How can Git operations must be performed outside of Databricks Repos?

A. Commit

B. Pull

C. Merge

D. Clone

AnswerC

36. 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.

AnswerD

37. 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.

AnswerB

38. 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.

AnswerE

39. 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;

AnswerB

40. 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.

Answer*C*

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

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

Answer*B*

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

- A. A key-value pair configuration
- B. At least one notebook library to be executed
- C. A path to cloud storage location for the written data
- D. A location of a target database for the written data

Answer*B*

43. 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

Answer*D*

44. 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 customers TO team;

AnswerE

45. 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;

Answer*E*

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?

- A. Merge
- B. Push
- C. Pull
- D. Commit
- E. Clone

Answer*C*

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

- A. Cloud-specific integrations
- B. Simplified governance
- C. Ability to scale storage
- D. Ability to scale workloads
- E. Avoiding vendor lock-in

Answer*E*

48. 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?

- A. if day_of_week = 1 and review_period:
- B. if day_of_week = 1 and review_period = "True":
- C. if day_of_week = 1 & review_period: = "True":
- D. if day_of_week == 1 and review_period:

  Answer*D*

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

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

Answer*A*

50. 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 raw data with a schema applied.
- D. Bronze tables contain a less refined view of data than raw dat

Answer*C*

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?

- A. REDUCE
- B. OPTIMIZE
- C. COMPACTION
- D. REPARTITION
- E. VACUUM

AnswerB

52. 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?

- A. The VACUUM command was run on the table
- B. The TIME TRAVEL command was run on the table
- C. The DELETE HISTORY command was run on the table
- D. The OPTIMIZE command was nun on the table

Answer*A*

53. 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.
```
CREATE FUNCTION combine_nyc(city STRING)
RETURNS STRING
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```
- B.
```
CREATE UDF combine_nyc(city STRING)
RETURNS STRING
CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```
- C.
```
CREATE FUNCTION combine_nyc(city STRING)
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```
- D.
```
CREATE UDF combine_nyc(city STRING)
RETURNS STRING
RETURN CASE
  WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

Answer*A*

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

- A. None of these
- B. Data lake
- C. Data warehouse
- D. All of these
- E. Data lakehouse

AnswerE

55. 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?

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

AnswerC

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?

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

Answer*D*

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

- A. TRANSFORM
- B. PIVOT
- C. SUM

- D. CONVERT
- E. WHERE

AnswerB

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?

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

AnswerC

59. The Delta transaction log for the 'students' tables is shown using the 'DESCRIBE HISTORY students' command. A Data Engineer needs to query the table as it existed before the UPDATE operation listed in the log.

Which command should the Data Engineer use to achieve this? (Choose two.)

	version	timestamp	operation
1	8	2024–04–22T14:33:31.000	OPTIMIZE
2	7	2024–04–22T14:33:16.000	MERGE
3	6	2024–04–22T14:33:06.000	DELETE
4	5	2024–04–22T14:32:58.000	UPDATE
5	4	2024–04–22T14:32:47.000	WRITE
6	3	2024–04–22T14:32:44.000	WRITE
7	2	2024–04–22T14:32:23.000	WRITE
8	1	2024–04–22T14:32:20.000	WRITE
9	0	2024–04–22T14:31:49.000	CREATE TABLE

- A. SELECT * FROM students@v4
- B. SELECT * FROM students TIMESTAMP AS OF '2024-04-22T 14:32:47.000+00:00'

- C. SELECT * FROM students FROM HISTORY VERSION AS OF 3
- D. SELECT * FROM students VERSION AS OF 5
- E. SELECT * FROM students TIMESTAMP AS OF '2024-04-22T 14:32:58.000+00:00'

Answer *AB*

60. Which method should a Data Engineer apply to ensure Workflows are being triggered on schedule?

- A. Scheduled Workflows require an always-running cluster, which is more expensive but reduces processing latency.
- B. Scheduled Workflows process data as it arrives at configured sources.
- C. Scheduled Workflows can reduce resource consumption and expense since the cluster runs only long enough to execute the pipeline.
- D. Scheduled Workflows run continuously until manually stopped.

Answer C

61. A data engineer needs to access the view created by the sales team, using a shared cluster. The data engineer has been provided usage permissions on the catalog and schema. In order to access the view created by sales team.

What are the minimum permissions the data engineer would require in addition?

- A. Needs SELECT permission on the VIEW and the underlying TABLE.
- B. Needs SELECT permission only on the VIEW
- C. Needs ALL PRIVILEGES on the VIEW
- D. Needs ALL PRIVILEGES at the SCHEMA level

Answer *A*

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?

- A.

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

- B.

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

- C.

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

- D.

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

- E.

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

Answer A

63. Identify the impact of ON VIOLATION DROP ROW and ON VIOLATION FAIL UPDATE for a constraint violation.A data engineer has created an ETL pipeline using Delta Live table to manage their company travel reimbursement detail, they want to ensure that the if the location details has not been provided by the employee, the pipeline needs to be terminated.
How can the scenario be implemented?

A. CONSTRAINT valid_location EXPECT (location = NULL)

B. CONSTRAINT valid_location EXPECT (location != NULL) ON VIOLATION FAIL UPDATE

C. CONSTRAINT valid_location EXPECT (location != NULL) ON DROP ROW

D. CONSTRAINT valid_location EXPECT (location != NULL) ON VIOLATION FAIL

Answer B

64. Identify a scenario to use an external table.
A Data Engineer needs to create a parquet bronze table and wants to ensure that it gets stored in a specific path in an external location.Which table can be created in this scenario?

- A. An external table where the location is pointing to specific path in external location.
- B. An external table where the schema has managed location pointing to specific path in external location.
- C. A managed table where the catalog has managed location pointing to specific path in external location.
- D. A managed table where the location is pointing to specific path in external location.

Answer *A*

65. Data engineer and data analysts 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?

- A. The pipeline can have different notebook sources in SQL & Python

- B. The pipeline will need to be written entirely in SQL
- C. The pipeline will need to use a batch source in place of a streaming source
- D. The pipeline will need to be written entirely in Python

AnswerA

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 + y
```
- 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
```

Answer*D*

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

The code block used by the data engineer is below:

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

The data engineer only wants the query to process all of the available data in as many batches as required.

Which line of code should the data engineer use to fill in the blank?

- A. trigger(availableNow=True)
- B. trigger(processingTime= "once")
- C. trigger(continuous= "once")
- D. trigger(once=True)

Answer*A*

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

sales

customer_id	spend	units
a1	28.94	7
a3	874.12	23
a4	8.99	1

favorite_stores

customer_id	store_id
a1	s1
a2	s1
a4	s2

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

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

Which of the following will be returned by the above query?

- A.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2

- B.

customer_id	spend	units	store_id
a1	28.94	7	s1
a4	8.99	1	s2

- C.

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

- D.

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

- E.

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

Answer*C*

69. What can be used to simplify and unify siloed data architectures that are specialized for specific use cases?

- A. Delta Lake
- B. Data lake
- C. Data warehouse
- D. Data lakehouse

Answer*D*

70. In a healthcare provider organization using Delta Lake to store electronic health records (EHRs), a data analyst needs to analyze a snapshot of the patient_records table from two weeks ago before some recent data corrections were applied.What approach should the Data Engineer take to allow the analyst to query that specific prior version?

A. Truncate the table to remove all data, then reload the data from two weeks ago into the truncated table for the analyst to query.

B. Identify the version number corresponding to two weeks ago from the Delta transaction log, share that version number with the analyst to query using VERSION AS OF syntax, or export that version to a new Delta table for the analyst to query.

C. Restore the table to the version from two weeks ago using the RESTORE command, and have the analyst query the restored table.

D. Use the VACUUM command to remove all versions of the table older than two weeks, then the analyst can query the remaining version.

Answer*B*

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?

- A. There was a type mismatch between the specific schema and the inferred schema
- B. JSON data is a text-based format

- C. Auto Loader only works with string data
- D. All of the fields had at least one null value
- E. Auto Loader cannot infer the schema of ingested data

Answer*B*

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?

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

AnswerE

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

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

AnswerD

74. Which two components function in the DB platform architecture's control plane? (Choose two.)

- A. Virtual Machines
- B. Compute Orchestration
- C. Serverless Compute
- D. Compute
- E. Unity Catalog

Answer*BE*

75. Identify how the count_if function and the count where x is null can be used

Consider a table random_values with below data.

What would be the output of below query?
select count_if(col > 1) as count_a. count(*) as count_b.count(col1) as count_c from random_values col1
0
1
2

NULL -
2
3

- A. 3 6 5
- B. 4 6 5
- C. 3 6 6
- D. 4 6 6

Answer *A*

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

- A. Streaming workloads
- B. Machine learning workloads
- C. Serverless workloads
- D. Batch workloads
- E. Dashboard workloads

Answer*A*

77. Differentiate between all-purpose clusters and jobs clusters.

A data engineering team has created a python notebook to load data from cloud storage, this job has been tested and now needs to be scheduled in production.

Which would be the best cluster to be used in this case?

- A. All purpose cluster
- B. Any Unity Catalog-enabled cluster
- C. Jobs Cluster
- D. Serverless SQL warehouse

Answer*C*

78. A data engineer is using the following code block as part of a batch ingestion pipeline to read from a composable table:

```
transactions_df = (spark.read
    .schema(schema)
    .format("delta")
    .table("transactions")
)
```

Which of the following changes needs to be made so this code block will work when the transactions table is a stream source?

- A. Replace predict with a stream-friendly prediction function
- B. Replace schema(schema) with option ("maxFilesPerTrigger", 1)
- C. Replace "transactions" with the path to the location of the Delta table
- D. Replace format("delta") with format("stream")
- E. Replace spark.read with spark.readStream

Answer*E*

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

- A.

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

- B.

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

- C.

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

- D.

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

- E.

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

AnswerE

80. 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?

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

AnswerB

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

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

Answer*D*

82. 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."
- B. They can turn on the Auto Stop feature for the SQL endpoint.
- C. They can increase the cluster size of the SQL endpoint.
- D. They can turn on the Serverless feature for the SQL endpoint.

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

AnswerD

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

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

- A. GRANT ALL PRIVILEGES ON TABLE sales TO team;
- B. GRANT SELECT CREATE MODIFY ON TABLE sales TO team;
- C. GRANT SELECT ON TABLE sales TO team;
- D. GRANT ALL PRIVILEGES ON TABLE team TO sales;

Answer*A*

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

Answer*B*

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

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

- A. GRANT VIEW ON CATALOG customers TO team;
- B. GRANT CREATE ON DATABASE customers TO team;
- C. GRANT USAGE ON CATALOG team TO customers;

- D. GRANT USAGE ON DATABASE customers TO team;

Answer*D*

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

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

- A. They can set a limit to the number of DBUs that are consumed by the SQL Endpoint.
- B. They can set the query's refresh schedule to end after a certain number of refreshes.
- C. They can set the query's refresh schedule to end on a certain date in the query scheduler.
- D. They can set a limit to the number of individuals that are able to manage the query's refresh schedule.

Answer*C*

87. 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?

- 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 one-time notifications.
- D. They can set up an Alert with a new webhook alert destination.
- E. They can set up an Alert without notifications.

Answer*D*

88. 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?

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

AnswerA

89. Which two conditions are applicable for governance in Databricks Unity Catalog? (Choose two.)

- A. You can have more than 1 metastore within a databricks account console but only 1 per region.
- B. Both catalog and schema must have a managed location in Unity Catalog provided metastore is not associated with a location
- C. You can have multiple catalogs within metastore and 1 catalog can be associated with multiple metastore
- D. If catalog is not associated with location, it’s mandatory to associate schema with managed locations
- E. If metastore is not associated with location, it’s mandatory to associate catalog with managed locations

Answer*AD*

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

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

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

Answer*C*

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

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

Answer*D*

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

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

Answer*C*

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

- A. Worker node
- B. Databricks web application
- C. Driver node

- D. Databricks Filesystem

Answer*B*

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

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

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

Answer*D*

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

- A. Avoiding vendor lock-in
- B. Simplified governance
- C. Ability to scale workloads
- D. Cloud-specific integrations

AnswerA

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

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

- A. Jobs
- B. Dashboards
- C. Catalog Explorer
- D. Repos

AnswerC

97. A data engineer is running code in a Databricks Repo that is cloned from a central Git repository. A colleague of the data engineer informs them that changes have been made and

synced to the central Git repository. The data engineer now needs to sync their Databricks Repo to get the changes from the central Git repository.

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

- A. Clone
- B. Pull
- C. Merge
- D. Push

AnswerB

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

- A. CSV
- B. Parquet
- C. JSON
- D. Delta

Answer*B*

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

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

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

- A. INSERT INTO my_table VALUES ('a1', 6, 9.4)
- B. INSERT VALUES ('a1', 6, 9.4) INTO my_table
- C. UPDATE my_table VALUES ('a1', 6, 9.4)
- D. UPDATE VALUES ('a1', 6, 9.4) my_table

Answer*A*

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

Which keyword can be used to compact the small files?

- A. OPTIMIZE
- B. VACUUM
- C. COMPACTION
- D. REPARTITION

Answer*A*

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

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

- A. Table
- B. Function
- C. View
- D. Temporary view

Answer*A*

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

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

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

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

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

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

Answer C

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

- A. DROP
- B. INSERT
- C. MERGE
- D. APPEND

Answer*C*

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

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

- A. SELECT * FROM sales
- B. spark.table("sales")
- C. spark.sql("sales")
- D. spark.delta.table("sales")

Answer*B*

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

CREATE DATABASE IF NOT EXISTS customer360;

In which location will the customer360 database be located?

- A. dbfs:/user/hive/database/customer360
- B. dbfs:/user/hive/warehouse
- C. dbfs:/user/hive/customer360
- D. dbfs:/user/hive/database

Answer*B*

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

DROP TABLE IF EXISTS my_table;

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

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

- A. The table was managed
- B. The table's data was smaller than 10 GB
- C. The table did not have a location
- D. The table was external

Answer*A*

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

They run the following command:

```
CREATE TABLE new_table

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

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

- A. FROM "path/to/csv"
- B. USING CSV
- C. FROM CSV
- D. USING DELTA

Answer*B*

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

- A. Parquet files can be partitioned
- B. Parquet files will become Delta tables
- C. Parquet files have a well-defined schema
- D. Parquet files have the ability to be optimized

Answer*C*

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

- A. TRANSFORM
- B. PIVOT

- C. SUM
- D. CONVERT

Correct Answer:B

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

They have the following incomplete code block:

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

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

- A. spark.delta.sql
- B. spark.sql
- C. spark.table
- D. dbutils.sql

AnswerB

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

```
sales
customer_idspend units
a1          28.94 7
a3          874.1223
a4          8.99  1

favorite_stores


customer_idstore_id
a1          s1
a2          s1
a4          s2
```

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

```
SELECT
    sales.customer_id,
    sales.spend,
    favorite_stores.store_id
FROM sales
LEFT JOIN favorite_stores
ON sales.customer_id = favorite_stores.customer_id;
```

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

- A.

- B.

customer_id	spend	store_id
a1	28.94	s1
a4	8.99	s2

- C.

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

- D.

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

AnswerC

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

Which code block successfully completes this task?

- **A.**

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

- **B.**

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

- **C.**

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

- **D.**

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

Answer*A*

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

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

- **A.**

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

- **B.**

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

- **C.**

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

- **D.**

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

Answer*D*

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

The code block used by the data engineer is below:

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

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

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

AnswerD

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

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

- A. Auto Loader
- B. Unity Catalog
- C. Delta Lake

- D. Delta Live Tables

Answer*D*

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

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

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

AnswerD

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

- A. Checkpointing and Write-ahead Logs
- B. Replayable Sources and Idempotent Sinks
- C. Write-ahead Logs and Idempotent Sinks
- D. Checkpointing and Idempotent Sinks

Answer*D*

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

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

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

Answer*A*

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

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

Answer*B*

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

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

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

- A. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will persist to allow for additional testing.
- B. All datasets will be updated once and the pipeline will shut down. The compute resources will persist to allow for additional testing.
- C. All datasets will be updated at set intervals until the pipeline is shut down. The compute resources will be deployed for the update and terminated when the pipeline is stopped.
- D. All datasets will be updated once and the pipeline will shut down. The compute resources will be terminated.

Answer*C*

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

- A. Streaming workloads
- B. Machine learning workloads
- C. Serverless workloads
- D. Batch workloads

Answer*A*

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

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

- A. Auto Loader cannot infer the schema of ingested data
- B. JSON data is a text-based format
- C. Auto Loader only works with string data
- D. All of the fields had at least one null value

- **Answer***B*

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

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

Answer*D*

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

- A.

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

- B.

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

- C.

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

- D.

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

AnswerD

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

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

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

- A. Records that violate the expectation cause the job to fail.
- B. Records that violate the expectation are added to the target dataset and flagged as invalid in a field added to the target dataset.
- C. Records that violate the expectation are dropped from the target dataset and recorded as invalid in the event log.
- D. Records that violate the expectation are added to the target dataset and recorded as invalid in the event log.

Answer*C*

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

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

- A. They can use endpoints available in Databricks SQL
- B. They can use jobs clusters instead of all-purpose clusters
- C. They can configure the clusters to autoscale for larger data sizes
- D. They can use clusters that are from a cluster pool

AnswerD

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

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

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

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

AnswerB

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

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

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

Answer*C*

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

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

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

Answer*B*

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

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

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

Answer*C*

132. A company uses Delta Sharing to collaborate with partners across different cloud providers and geographic regions.

What will result in additional costs due to cross-region or egress fees?

- A. Sharing data within the same cloud provider and region
- B. Transferring data via Delta Sharing across clouds and across different geographic regions
- C. Accessing Delta Sharing data using a VPN within the same data center
- D. Utilizing Delta Sharing for internal data analytics within a single cloud environment

Answer*B*

133. A data engineer is writing a script that is meant to ingest new data from cloud storage. In the event of the Schema change, the ingestion should fail. It should fail until the changes downstream source can be found and verified as intended changes.

Which command will meet the requirements?

- A. failOnNewColumns

- B. none
- C. rescue
- D. addNewColumns

Answer*A*

134. Which SQL code snippet will correctly demonstrate a Data Definition Language (DDL) operation used to create a table?

- A. CREATE TABLE employees (
  id INT,
  name STRING
  );
- B. DROP TABLE employees;
- C. ALTER TABLE employees ADD COLUMN salary DECIMAL(10,2);
- D. INSERT INTO employees (id, name) VALUES (1 'Alice');

Answer*A*

135. A data engineer is working in a Databricks notebook to design and manage a batch ETL pipeline. The engineer is writing SQL and Python code to clean data, transform it, and join large datasets from different sources. The engineer wants to organize these steps into a structured process that can be run regularly and scheduled as part of a data pipeline.

Which Databricks notebook feature is applicable in the use case?

- A. Real-time streaming support
- B. Collaborative editing
- C. Task workflows and job scheduling
- D. Notebook version control

AnswerC

136. A data engineer needs to develop integration tests for an ETL process and deploy a version-controlled, packaged workflow into production using an external job scheduler.

Which tool should the data engineer use for this job?

- A. Databricks Connect

- B. Databricks Asset Bundles
- C. Databricks Command Line Interface
- D. Databricks Software Development Kit

AnswerB

137. Which Databricks asset bundle format is valid?

- A. resources: jobs: hello-job: name: hello-job tasks: - task_key: hello-task existing_cluster_id: 1234-567890-abcde123 notebook_task: notebook_path: ./hello.py
- B. "resources":{ "jobs":{ "name":"hello-job", "tasks":{ "task_key:"hello-task", "existing_cluster_id":"1234-567890-abcde123", "notebook_task":{ "notebook_path": ".hello.py" } } }
- C. configuration = { "resources":{ "jobs":{ "name":"hello-job", "tasks":{ "task_key:"hello-task", "existing_cluster_id":"1234-567890-abcde123", "notebook_task":{ "notebook_path": ".hello.py" } } } }
- D. resources { jobs { name = "hello-job" tasks{ task_key = "hello-task" existing_cluster_id = "1234-567890-abcde123" notebook_task{ notebook_path = ".hello.py" } } } }

AnswerA

138.A data engineer needs to ingest from both streaming and batch sources for a firm that relies on highly accurate data. Occasionally, some of the data picked up by the sensors that provide a streaming input are outside the expected parameters. If this occurs, the data must be dropped, but the stream should not fail.

Which feature of Delta Live Tables meets this requirement?

- A. Change Data Capture
- B. Error Handling
- C. Monitoring
- D. Expectations

Answer*D*

139. A data engineer has inherited a Databricks pipeline from a previous team. The pipeline is missing SLAs and costs more than the allotted budget. On analysis, it is noted that the cluster is not being fully utilized, and the dataset is getting skewed.

How should the data engineer resolve this issue?

- A. Use coalesce() on the dataset to merge partitions and reduce skew.
- B. Increase the number of executors for the job.
- C. Repartition the dataset to have it be more optimally spread across all nodes.
- D. Increase the executor memory for the job.

AnswerC

140. An organization is looking for an optimized storage layer that supports ACID transactions and schema enforcement.

Which technology should the organization use?

- A. Delta Lake
- B. Unity Catalog
- C. Cloud File Storage
- D. Data lake

Answer*A*

141. What are the transformations typically included in building the Bronze layer?

- A. Include columns Load date/time, process ID
- B. Business rules and transformations
- C. Perform extensive data cleansing
- D. Aggregate data from multiple sources

Answer*A*

142. An organization has data stored across multiple external systems, including MySQL, Amazon Redshift, and Google BigQuery. The data engineer wants to perform analytics without ingesting directly into Databricks, ensuring unified governance and minimizing data duplication.

Which feature of Databricks enables querying these external data sources while maintaining centralized governance?

- A. Delta Lake
- B. Lakehouse Federation
- C. MLflow
- D. Databricks Connect

AnswerB

143. An organization needs to share a dataset stored in its Databricks Unity Catalog with an external partner who uses a different data platform that is not Databricks. The goal is to maintain data security and ensure the partner can access the data efficiently.

.

Which method should the data engineer use to securely share the dataset with the external partner?

- *A. Using Delta Sharing with the open sharing protocol*
- *B. Exporting data as CSV files and emailing them*
- *C. Using a third-party API to access the Delta table*
- *D. Databricks-to-Databricks Sharing*

Answer*A*

144. A data engineer streams customer orders into a Kafka topic (orders_topic) and is currently writing the ingestion script of a DLT pipeline. The data engineer needs to ingest the data from Kafka brokers to DLT using Databricks.What is the correct code for ingesting the data?

- A.

```
import dlt

@dlt.table(
    name = "orders_raw",
)
def orders_raw():
    return (
        spark.readStream
        .format("kafka")
        .option("kafka.bootstrap.servers", "broker:9092")
        .option("subscribe", "orders_topic")
        .option("startingOffsets", "earliest")
        .load()
    )
```

- B.

```
import dlt

@dlt.table(
    name = "orders_raw",
)
def orders_raw():
    return (
        spark.readStream
        .format("cloud_files")
        .option("cloudFiles.format", "json")
        .option("cloudFiles.schemaLocation", "/schema/location")
        .load("kafka://broker:9092/orders_topic")
    )
```

- C.

```
CREATE LIVE TABLE orders_raw
AS SELECT
    CAST(value AS STRING) AS json_data
FROM STREAM kafka.`broker:9092/orders_topic`;
```

```
CREATE STREAMING LIVE TABLE orders_raw
AS SELECT
    value:order_id AS order_id,
    value:customer_id AS customer_id,
    value:amount AS amount,
    value:order_status AS order_status,
    value:order_timestamp AS order_timestamp
FROM cloud_files("kafka://broker:9092/orders_topic", "json");
```

- D.

Answer*C*

145. A global retail company sells products across multiple categories (e.g., Electronics, Clothing) and regions (e.g., North, South, East, West). The sales team has provided the data engineer with a PySpark dataframe named sales_df as below and the team wants the data engineer to analyze the sales data to help them make strategic decisions.

sales_df

product_id	category	sales_amount	region
1	Electronics	100	North
2	Clothing	200	South
3	Electronics	150	North
4	Clothing	300	South
5	Electronics	250	East
6	Clothing	400	West

Calculate the total sales amount for each product category and store the results in a new dataframe called category_sales.

What will generate the expected result of category_sales?

category	total_sales_amount
Electronics	500
Clothing	900

- A. category_sales = sales_df.groupBy("category").agg(sum("sales_amount").alias("total_sales_amount"))
- B. category_sales = sales_df.sum("sales_amount").groupBy("category").alias("total_sales_amount"))
- C. category_sales = sales_df.agg(sum("sales_amount").groupBy("category").alias("total_sales_amount"))
- D. category_sales = sales_df.groupBy("region").agg(sum("sales_amount").alias("total_sales_amount"))

A.

```
import dlt

@dlt.table(
    name = "orders_raw",
)
def orders_raw():
    return (
        spark.readStream
        .format("kafka")
        .option("kafka.bootstrap.servers", "broker:9092")
        .option("subscribe", "orders_topic")
        .option("startingOffsets", "earliest")
        .load()
    )
```

- B.

```
import dlt

@dlt.table(
    name = "orders_raw",
)
def orders_raw():
    return (
        spark.readStream
        .format("cloud_files")
        .option("cloudFiles.format", "json")
        .option("cloudFiles.schemaLocation", "/schema/location")
        .load("kafka://broker:9092/orders_topic")
    )
```

- C.

```
CREATE LIVE TABLE orders_raw
AS SELECT
    CAST(value AS STRING) AS json_data
FROM STREAM kafka.`broker:9092/orders_topic`;
```

- D.

```
CREATE STREAMING LIVE TABLE orders_raw
AS SELECT
    value:order_id AS order_id,
    value:customer_id AS customer_id,
    value:amount AS amount,
    value:order_status AS order_status,
    value:order_timestamp AS order_timestamp
FROM cloud_files("kafka://broker:9092/orders_topic", "json");
```

.

Answer*A*

146. A data engineer is designing an ETL pipeline to process both streaming and batch data from multiple sources. The pipeline must ensure data quality, handle schema evolution, and provide easy maintenance. The team is considering using Delta Live Tables (DLT) in Databricks to achieve these goals. They want to understand the key features and benefits of DLT that make it suitable for this use case.

Why is Delta Live Tables (DLT) an appropriate choice?

- A. Automatic data quality checks, built-in support for schema evolution, and declarative pipeline development
- B. Manual schema enforcement, high operational overhead, and limited scalability
- C. Requires custom code for data quality checks, no support for streaming data, and complex pipeline maintenance
- D. Supports only batch processing, no data versioning, and high infrastructure costs

Answer*A*

147. A data engineer is attempting to write Python and SQL in the same command cell and is running into an error. The engineer thought that it was possible to use a Python variable in a select statement.

Why does the command fail?

- A. Databricks supports language interoperability in the same cell but only between Scala and SQL.

- B. Databricks supports multiple languages but only one per notebook.
- C. Databricks supports one language per cell.
- D. Databricks supports language interoperability but only if a special character is used.

Answer*C*

148. Which compute option should be chosen in a scenario where small-scale ad-hoc Python scripts need to be run at high frequency and should wind down quickly after these queries have finished running?

- A. All-purpose Cluster
- B. Job Cluster
- C. Serverless Compute
- D. SQL Warehouse

Answer*C*

149. A data engineer is working on a personal laptop and needs to perform complex transformations on data stored in a Delta Lake on cloud storage. The engineer decides to use Databricks Connect to interact with Databricks clusters and work in their local IDE.

How does Databricks Connect enable the engineer to develop, test, and debug code seamlessly on their local machine while interacting with Databricks clusters?

- A. By providing a local environment that mimics the Databricks runtime, enabling the engineer to develop, test, and debug code using a specific IDE that is required by Databricks
- B. By providing a local environment that mimics the Databricks runtime, enabling the engineer to develop, test, and debug code only through Databricks' own web interface
- C. By allowing direct execution of Spark jobs from the local machine without needing a network connection
- D. By providing a local environment that mimics the Databricks runtime, enabling the engineer to develop, test, and debug code using their preferred IDE

AnswerD

150. A company sells products across multiple categories (e.g., Electronics, Clothing) and regions. The sales team has provided you with a PySpark dataframe named sales_df as below, and the team wants the data engineer to analyze the sales data to help make strategic decisions.

sales_df

product_id	category	sales_amount	region
1	Electronics	100	North
2	Clothing	200	South
3	Electronics	150	North
4	Clothing	300	South
5	Electronics	250	East
6	Clothing	400	West

Calculate the total sales amount for each region and store the results in a new dataframe called region_sales.

Given the expected result:

region_sales

region	total_sales_amount
North	250
South	500
East	250
West	400

Which code will generate the expected result?

- A. region_sales = sales_df.groupBy("category").sum("sales_amount").alias("total_sales_amount")
- B. region_sales = sales_df.groupBy("region").agg(sum("sales_amount").alias("total_sales_amount"))

- C. region_sales = sales_df.sum("sales_amount").groupBy("region").alias("total_sales_amount")
- D. region_sales = sales_df.agg(sum("sales_amount").groupBy("region").alias("total_sales_amount"))

Answer*B*

151. A Data Engineer is building a simple data pipeline using Delta Live Tables (DLT) in Databricks to ingest customer data. The raw customer data is stored in a cloud storage location in JSON format. The task is to create a DLT pipeline that reads the raw JSON data and writes it into a Delta table for further processing.

Which code snippet will correctly ingest the raw JSON data and create a Delta table using DLT?

- **A.**

```
import dlt

@dlt.table
def raw_customers():
    return spark.read.format("csv").load("s3://my-bucket/raw-customers/")
```

- **B.**

```
import dlt

@dlt.view
def raw_customers():
    return spark.format.json("s3://my-bucket/raw-customers/")
```

- **C.**

```
import dlt

@dlt.table
def raw_customers():
    return spark.read.json("s3://my-bucket/raw-customers/")
```

- **D.**

```
import dlt

@dlt.table
def raw_customers():
    return spark.read.format("parquet").load("s3://my-bucket/raw-customers/")
```

Answer*C*

152.A data engineering team is using Kafka to capture event data and then ingest it into Databricks. The team wants to be able to see these historical events. Medallion architecture is already in place. The team wants to be mindful of costs.

Where should this historical event data be stored?

- A. Gold
- B. Silver
- C. Bronze
- D. Raw layer

Answer*C*

153. What is the maximum output supported by a job cluster to ensure a notebook does not fail?

- A. 25MBs
- B. 10MBs
- C. 30MBs
- D. 15MBs

AnswerB

154. Which two items are characteristics of the Gold Layer? (Choose two.)

- A. Historical lineage
- B. Raw Data
- C. Normalised
- D. De-normalised
- E. Read-optimized

Answer*DE*

155. A data engineer has developed an ETL that produce a Delta managed table with liquid clustering feature activated as output. Several consumers are having issues regarding time delay when reading this table.

How could the Data Engineer be sure about the OPTIMIZE command has been executed explicitly?

- A. Check the system table system.storage.predictive_optimization_operations_history
- B. Use SHOW TABLES EXTENDED to check the partitions columns used
- C. Use DESCRIBE DETAIL table to see the file size and number of files for the table
- D. Use DESCRIBE HISTORY table to check if exists any OPTIMIZE operation

Answer*D*

156. A data engineer is reviewing the documentation on audit logs in Databricks for compliance purposes and needs to understand the format in which audit logs output events.

How are events formatted in Databricks audit logs?

- A. In Databricks, audit logs output events in a JSON format.
- B. In Databricks, audit logs output events in a CSV format.
- C. In Databricks, audit logs output events in an XML format.
- D. In Databricks, audit logs output events in a plain text format.

Answer*A*

157. A Python file is ready to go into production and the client wants to use the cheapest but most efficient type of cluster possible. The workload is quite small, only processing 10GBs of data with only simple joins and no complex aggregations or wide transformations.

Which cluster meets the requirement?

- A. Interactive cluster
- B. Job cluster with spot instances enabled
- C. Job cluster with spot instances disabled
- D. Job cluster with Photon enabled

Answer*B*

158. A data engineer is working on a Databricks project that utilizes cloud storage. The data engineer wants to load several JSON files from containers on a storage account as soon as the file arrives within the storage account.

Which syntax should the data engineer follow to first load the files into a dataframe and check that it is working as expected using Python?

- A. df = spark.read.json("input/path")
- B. df = spark.readStream.format("cloud").option("json").load("/input/path")
- C. df = spark.readStream.format("json".load("input/path")
- D. df = spark.readStream.format("cloudFiles").option("cloudFiles.format", "json").load("/input/path")

AnswerD

159. A data engineer team has decided to implement a new data platform on Databricks and is currently deciding how to store each kind of data on each data layer.

What is the appropriate layer and data pairing for medallion architecture?

- A. Silver Layer - Raw data from deposit account application
- B. Bronze Layer - Summary of cash deposit amount for each country and city
- C. Silver Layer - Cleansed master customer data
- D. Gold Layer - Deduplicated money transfer transaction

Answer*C*

160. A data engineer is processing ingested streaming tables and needs to filter out NULL values in the order_datetime column from the raw streaming table orders_raw and store the results in a new table orders_valid using DLT.

Which code snippet should the data engineer use?

- A.

```
CREATE OR REFRESH STREAMING TABLE orders_valid(
  CONSTRAINT valid_date
  EXPECT (order_datetime IS NOT NULL)
  ON VIOLATION DROP ROW
)
AS SELECT * FROM STREAM orders_raw;
```

- B.

```
CREATE OR REFRESH STREAMING TABLE orders_valid
(
  CONSTRAINT valid_date EXPECT (order_datetime IS NOT NULL)
  ON VIOLATION DROP ROW
)
AS SELECT * FROM orders_raw;
```

- C.

```
CREATE OR REFRESH STREAMING TABLE orders_valid
AS
SELECT *
FROM STREAM(orders_raw)
WHERE order_datetime IS NOT NULL;
```

- D.

```
CREATE OR REPLACE STREAMING TABLE orders_valid
(
  FILTER (order_datetime IS NOT NULL)
)
AS SELECT * FROM STREAM(orders_raw);
```

Answer*A*

161. A data engineer is managing a data pipeline in Databricks, where multiple Delta tables are used for various transformations. The team wants to track how data flows through the pipeline, including identifying dependencies between Delta tables, notebooks, jobs, and dashboards. The data engineer is utilizing the Unity Catalog lineage feature to monitor this process.

How does Unity Catalog's data lineage feature support the visualization of relationships between Delta tables, notebooks, jobs, and dashboards?

- A. Unity Catalog lineage visualizes dependencies between Delta tables, notebooks, and jobs, but does not provide column-level tracing or relationships with dashboards.
- B. Unity Catalog lineage only supports visualizing relationships at the table level and does not extend to notebooks, jobs, or dashboards.
- C. Unity Catalog lineage provides an interactive graph that tracks dependencies between tables and notebooks but excludes any job-related dependencies or dashboard visualizations.
- D. Unity Catalog provides an interactive graph that visualizes the dependencies between Delta tables, notebooks, jobs, and dashboards, while also supporting column-level tracking of data transformations.

Answer*B*

162. A data engineer needs to conduct Exploratory Analysis on data residing in a database that is within the company's custom-defined network in the cloud. The data engineer is using SQL for this task.

Which type of SQL Warehouse will enable the data engineer to process large numbers of queries quickly and cost-effectively?

- A. Serverless compute for notebooks

- B. Pro SQL Warehouse
- C. Classic SQL Warehouse
- D. Serverless SQL Warehouse

Answer*B*

163. A data engineer is configuring Unity Catalog in Databricks and needs to assign a role to a user who should have the ability to grant and revoke privileges on various data objects within a specific schema, but should not have read/write access over the schema or its objects.

Which role should the data engineer assign to this user?

- A. Table Owner
- B. Catalog Owner
- C. Schema Owner
- D. USE catalog/schema privilege on the schema

Answer*C*

164. A data engineer is debugging a Python notebook in Databricks that processes a dataset using PySpark. The notebook fails with an error during a DataFrame transformation. The engineer wants to inspect the state of variables, such as the input DataFrame and intermediate results, to identify where the error occurs.

Which tool should the engineer use to debug the notebook and inspect the values of variables like DataFrames?

- A. Use the Databricks CLI to download and analyze driver logs for detailed error messages
- B. Use the Python Notebook Interactive Debugger to set breakpoints and inspect variable values in real-time
- C. Use the Ganglia UI to monitor cluster resource usage and identify hardware issues
- D. Use the Spark UI to analyze the execution plan and identify stages where the job failed

AnswerB

165. A data engineer wants to create an external table in Databricks that references data stored in an Azure Data Lake Storage (ADLS) location. The goal is to enable Databricks to access and query this external data without moving it into the Databricks-managed storage.

Which step should the data engineer take to successfully create the external table?

- A. Use the CREATE MANAGED TABLE statement and specify the LOCATION clause with the path to the external data.
- B. CREATE UNMANAGED TABLE statement without specifying a LOCATION clause.
- C. Use the CREATE TABLE statement and specify the LOCATION clause with the path to the external data.
- D. CREATE EXTERNAL TABLE statement without specifying a LOCATION clause**.**

Answer*C*

166. A data engineer is developing a small proof of concept in a notebook. When running the entire notebook, the Cluster usage spikes. The data engineer wants to keep the development requirements and get real-time results.

Which Cluster meets these requirements?

- A. All Purpose Cluster with autoscaling
- B. Job Cluster with Photon enabled and autoscaling
- C. Job Cluster with autoscaling enabled
- D. All-Purpose Cluster with a large fixed memory size

Answer*A*

167. A data engineer needs to process SQL queries on a large dataset with fluctuating workloads. The workload requires automatic scaling based on the volume of queries, without the need to manage or provision infrastructure. The solution should be cost-efficient and charge only for the compute resources used during query execution.

Which compute option should the data engineer use?

- A. Databricks SQL Analytics
- B. Databricks Runtime for ML
- C. Databricks Jobs
- D. Serverless SQL Warehouse

Answer*D*

168. What is the functionality of AutoLoader in Databricks?

- A. Auto Loader automatically ingests and processes new files from cloud storage, handling both batch and streaming data with support for schema evolution.
- B. Auto Loader automatically ingests and processes new files from cloud storage, handling batch and streaming data with no support for schema evolution.
- C. Auto Loader automatically ingests and processes new files from cloud storage, handling only streaming data with no support for schema evolution.
- D. Auto Loader automatically ingests and processes new files from cloud storage, handling batch data with support for schema evolution.

AnswerA

169. A company is collaborating with a partner that does not use Databricks but needs access to a large historical dataset stored in Delta format. The data engineer needs to ensure that the partner can access the data securely, without the need for them to set up an account, and with read-only access.

How should the data be shared?

- A. Share the dataset by exporting it to a CSV file and manually transferring the file to the partner's system.
- B. Grant your partner access to your Databricks workspace and assign them full write permissions to the Delta table, enabling them to modify the dataset.
- C. Share the dataset using Unity Catalog, ensuring that both teams have full write access to the data within the same organization.

- D. Share the dataset using Delta Sharing, which allows your partner to access the data using a secure, read-only URL without requiring a Databricks account, ensuring that they cannot modify the data.

AnswerD

170. A data engineer is using the Databricks OPTIMIZE command on a Delta table.

What happens when OPTIMIZE is run twice on the same table with the same data?

- A. It has no effect because it is idempotent.
- B. It changes the number of tuples per file significantly.
- C. It further reduces file sizes by re-clustering the data.
- D. It triggers a full liquid clustering process.

Answer A

171. A data engineer at a company that uses Databricks with Unity Catalog needs to share a collection of tables with an external partner who also uses a Databricks workspace enabled for Unity Catalog. The data engineer decides to use Delta Sharing to accomplish this.

What is the first piece of information the data engineer should request from the external partner to set up Delta Sharing?

- A. The IP address of their Databricks workspace
- B. The name of their Databricks cluster
- C. The sharing identifier of their Unity Catalog metastore
- D. Their Databricks account password

Answer*C*

172. A Databricks workflow fails at the last stage due to an error in a notebook. This workflow runs daily. The data engineer fixes the mistake and wants to rerun the pipeline. This workflow is very costly and time-intensive to run.

Which action should the data engineer do in order to minimise downtime and cost?

- A. Re-run the entire workflow
- B. Repair run
- C. Restart the cluster
- D. Switch to another cluster

AnswerB

173. An organization has implemented a data pipeline in Databricks and needs to ensure it can scale automatically based on varying workloads without manual cluster management. The goal is to meet the company's Service Level Agreements (SLAs), which require high availability and minimal downtime, while Databricks automatically handles resource allocation and optimization.

Which approach fulfills these requirements?

- A. Deploy Job Clusters with fixed configurations, dedicated to specific tasks, without automatic scaling.
- B. Use Spot Instances to allocate resources dynamically while minimizing costs, with potential interruptions.
- C. Use Interactive Clusters in Databricks, adjusting cluster sizes manually based on workload demands.
- D. Use Serverless compute in Databricks to automatically scale and provision resources with minimal manual intervention.

Answer*D*

174. A data engineer has been provided a PySpark DataFrame named df with columns product and revenue. The data engineer needs to compute complex aggregations to determine each product's total revenue, average revenue, and transaction count.

Which code snippet should the data engineer use?

- A.

```
from pyspark.sql import functions as F
aggregated_df = df.groupBy("product").agg( F.sum("revenue").alias("total_revenue"),
F.avg("revenue").alias("avg_revenue"), F.count("*").alias("transaction_count") )
```

- B

```
aggregated_df = df.groupby("product").agg(
    "sum(revenue)",
    "avg(revenue)",
    "count(revenue)"
)
```

- C.

```
from pyspark.sql import functions as F aggregated_df = df.select("product",

"revenue").groupBy("product").agg( F.sum("revenue"), F.mean("revenue") )
```

- D.

```
aggregated_df =
df.groupBy("product").agg( {"revenue": "sum", "revenue": "avg", "revenue": "count"} )
```

Answer*A*

175. A Databricks single-task workflow fails at the last task due to an error in a notebook. The data engineer fixes the mistake in the notebook.

What should the data engineer do to rerun the workflow?

- A. Repair the task
- B. Rerun the pipeline
- C. Restart the cluster
- D. Switch the cluster

AnswerA

176. A data engineer needs to provide access to a group named manufacturing-team. The team needs privileges to create tables in the quality schema.

Which set of SQL commands will grant a group named manufacturing-team to create tables in a schema named production with the parent catalog named manufacturing with the least privileges?

- A. GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE SCHEMA ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE CATALOG ON CATALOG manufacturing TO manufacturing-team;
- B. GRANT USE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE SCHEMA ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE CATALOG ON CATALOG manufacturing TO manufacturing-team;
- C. GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT CREATE SCHEMA ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT CREATE CATALOG ON CATALOG manufacturing TO manufacturing-team;
- D. GRANT CREATE TABLE ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT CREATE SCHEMA ON SCHEMA manufacturing.quality TO manufacturing-team; GRANT USE CATALOG ON CATALOG manufacturing TO manufacturing-team;

AnswerA

177. A data engineer has written a function in a Databricks Notebook to calculate the population of bacteria in a given medium.

```
def calculate_population_of_bacteria (intital_population, expo
    return future_population = intitial_population ** exponten
```

Analysts use this function in the notebook and sometimes provide input arguments of the wrong data type, which can cause errors during execution.

Which Databricks feature will help the data engineer quickly identify if an incorrect data type has been provided as input?

- A. The Spark User interface has a debug tab that contains the variables that are used in this session.
- B. The Databricks debugger enables breakpoints that will raise an error if the wrong data type is submitted.
- C. The Databricks debugger enables the use of a variable explorer to see at a glance the value of the variables.

- D. The Data Engineer should add print statements to find out what the variable is.

Answer*B*

178. A data engineer is inspecting an ETL pipeline based on a Pyspark job that consistently encounters performance bottlenecks. Based on developer feedback, the data engineer assumes the job is low on compute resources. To pinpoint the issue, the data engineer observes the Spark UI and finds out the job has a high CPU time vs Task time.

Which course of action should the data engineer take?

- A. High CPU time vs Task time means an under-utilized cluster. The data engineer may need to repartition data to spread the jobs more evenly throughout the cluster.
- B. High CPU time vs Task time means efficient use of cluster and no change needed
- C. High CPU time vs Task time means a CPU over-utilized job. The data engineer may need to consider executor and core tuning or resizing the cluster
- D. High CPU time vs Task time means over-utilized memory and the need to increase parallelism

AnswerC

179. A data engineer needs to parse only png files in a directory that contains files with different suffixes.

Which code should the data engineer use to achieve this task?

- A. df = spark.readStream.format("cloudFiles") \ .option("cloudFiles.format", "binaryFile") \
  .append("/*.png")
- B. df = spark.readstream. format("cloudFiles") \
  .option("cloudFiles.format", "binaryFile") \
  .option("pathGlobfilter", "*.png") \
  .load()
- C. df = spark.readStream.format("cloudFiles") \ .option("cloudFiles.format", "binaryFile") \
  .option("pathGlobfilter", "*.png") \ .append()

- D. df = spark.readstream.format("cloudFiles") \ .option("cloudFiles.format", "binaryFile") \
  .load("/*.png")

Answer*B*

180. Which languages are supported by Serverless compute clusters? (Choose two.)

- A. SQL
- B. Python
- C. R
- D. Scala
- E. Java

Answer*AB*

181. A data engineer is developing an ETL process based on Spark SQL. The execution fails. The data engineer checks the Spark UI and can see the ERRORS as follows:

"java.lang.OutofMemoryError: Java heap space"

Which two corrective actions should the data engineer perform to resolve this issue? (Choose two.)

- A. Narrow the filters in order to collect less data in the query
- B. Upsize the worker nodes and activate autoshuffle partitions
- C. Upsize the driver node and deactivate autoshuffle partitions
- D. Cache the dataset in order to boost the query performance
- E. Fix the shuffle partitions to 50 to ensure the allocation

Answer*AB*