Which of the following commands can a data engineer use to compact small data files of a Delta table into larger ones ?

ZORDER BY

COMPACT

VACUUM

OPTIMIZE

A data engineer is trying to use Delta time travel to rollback a table to a previous version, but the data engineer received an error that the data files are no longer present.

Which of the following commands was run on the table that caused deleting the data files?

VACUUM

OPTIMIZE

ZORDER BY

DEEP CLONE

DELETE

In Delta Lake tables, which of the following is the primary format for the data files?

Delta

Parquet

JSON

Hive-specific format

Both, Parquet and JSON

Which of the following locations hosts the Databricks web application ?

Data plane

Control plane

Databricks Filesystem

Databricks-managed cluster

Customer Cloud Account

In Databricks Repos, which of the following operations a data engineer can use to update the local version of a repo from its remote Git repository ?

Clone

Commit

Merge

Push

Pull

In Databricks Repos, which of the following operations a data engineer can use to update the local version of a repo from its remote Git repository ?

Clone

Commit

Merge

Push

Pull

According to the Databricks Lakehouse architecture, which of the following is located in the customer's cloud account?

Databricks web application

Notebooks

Repos

Cluster virtual machines

Workflows

Which of the following best describes Databricks Lakehouse?

- **Single, flexible, high-performance system that supports data, analytics, and machine learning workloads.**
- **Reliable data management system with transactional guarantees for organization's structured data.**
- **Platform that helps reduce the costs of storing organization's open-format data files in the cloud.**
- **Platform for developing increasingly complex machine learning workloads using a simple, SQL-based solution.**
- **Platform that scales data lake workloads for organizations without investing on-premises hardware.**

If the default notebook language is SQL, which of the following options a data engineer can use to run a Python code in this SQL Notebook?

- **They need first to import the python module in a cell**
- **This is not possible! They need to change the default language of the notebook to Python**
- **Databricks detects cells language automatically, so they can write Python syntax in any cell**
- **They can add `%language` magic command at the start of a cell to force language detection.**
- **They can add `%python` at the start of a cell.**

Which of the following tasks is not supported by Databricks Repos, and must be performed in your Git provider ?

Clone, push to, or pull from a remote Git repository.
Create and manage branches for development work.
Create notebooks, and edit notebooks and other files.
Visually compare differences upon commit.
Delete branches

Which of the following statements is **Not** true about Delta Lake ?

Delta Lake provides ACID transaction guarantees
Delta Lake provides scalable data and metadata handling
Delta Lake provides audit history and time travel
Delta Lake builds upon standard data formats: Parquet + XML
Delta Lake supports unified streaming and batch data processing

How long is the default retention period of the VACUUM command ?

0 days
7 days
30 days
90 days
365 days

The data engineering team has a Delta table called **employees** that contains the employees personal information including their gross salaries.

Which of the following code blocks will keep in the table only the employees having a salary greater than 3000 ?

DELETE FROM employees WHERE salary > 3000;

SELECT CASE WHEN salary <= 3000 THEN DELETE ELSE UPDATE END FROM employees;

UPDATE employees WHERE salary > 3000 WHEN MATCHED SELECT;

UPDATE employees WHERE salary <= 3000 WHEN MATCHED DELETE;

Your answer is correct
DELETE FROM employees WHERE salary <= 3000;

A data engineer wants to create a relational object by pulling data from two tables. The relational object must be used by other data engineers in other sessions on the same cluster only. In order to save on storage costs, the date engineer wants to avoid copying and storing physical data.

Which of the following relational objects should the data engineer create?

Temporary view
External table
Managed table
Global Temporary view
View

A data engineer has developed a code block to completely reprocess data based on the following if-condition in Python:

```python
if process_mode = "init" and not is_table_exist:
   print("Start processing ...")
```

This if-condition is returning an invalid syntax error.

Which of the following changes should be made to the code block to fix this error ?

```python
if process_mode = "init" & not is_table_exist:
    print("Start processing ...")
```

```python
if process_mode = "init" and not is_table_exist = True:
    print("Start processing ...")
```

```python
if process_mode = "init" and is_table_exist = False:
    print("Start processing ...")
```

```python
if (process_mode = "init") and (not is_table_exist):
    print("Start processing ...")
```

Correct answer

```python
if process_mode == "init" and not is_table_exist:
    print("Start processing ...")
```

Fill in the below blank to successfully create a table in Databricks using data from an existing PostgreSQL database:

```sql
CREATE TABLE employees
  USING ____________
  OPTIONS (
    url "jdbc:postgresql:dbserver",
    dbtable "employees"
  )
```

org.apache.spark.sql.jdbc

postgresql

DELTA

dbserver

cloudfiles

Which of the following commands can a data engineer use to create a new table along with a comment ?

Correct answer

```
CREATE TABLE payments
COMMENT "This table contains sensitive information"
AS SELECT * FROM bank_transactions
```

Your answer is incorrect

```
CREATE TABLE payments
COMMENT("This table contains sensitive information")
AS SELECT * FROM bank_transactions
```

```
CREATE TABLE payments
AS SELECT * FROM bank_transactions
COMMENT "This table contains sensitive information"
```

```
CREATE TABLE payments
AS SELECT * FROM bank_transactions
COMMENT("This table contains sensitive information")
```

```
COMMENT("This table contains sensitive information")
CREATE TABLE payments
AS SELECT * FROM bank_transactions
```

A junior data engineer usually uses `INSERT INTO` command to write data into a Delta table. A senior data engineer suggested using another command that avoids writing of duplicate records.

Which of the following commands is the one suggested by the senior data engineer ?

MERGE INTO
APPLY CHANGES INTO
UPDATE
COPY INTO
INSERT OR OVERWRITE

A data engineer is designing a Delta Live Tables pipeline. The source system generates files containing changes captured in the source data. Each change event has metadata indicating whether the specified record was inserted, updated, or deleted. In addition to a timestamp column indicating the order in which the changes happened. The data engineer needs to update a target table based on these change events.

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

MERGE INTO

Correct answer

APPLY CHANGES INTO
UPDATE
COPY INTO
cloud_files

In PySpark, which of the following commands can you use to query the Delta table **employees** created in Spark SQL?

pyspark.sql.read(SELECT * FROM employees)

spark.sql("employees")

spark.format(“sql”).read("employees")

spark.table("employees")

Spark SQL tables can not be accessed from PySpark

Which of the following code blocks can a data engineer use to create a user defined function (UDF) ?

CREATE FUNCTION plus_one(value INTEGER)

RETURN value +1

CREATE UDF plus_one(value INTEGER)

RETURNS INTEGER

RETURN value +1;

CREATE UDF plus_one(value INTEGER)

RETURN value +1;

Correct answer
CREATE FUNCTION plus_one(value INTEGER)

RETURNS INTEGER

RETURN value +1;

Your answer is incorrect
CREATE FUNCTION plus_one(value INTEGER)

RETURNS INTEGER

value +1;

When dropping a Delta table, which of the following explains why only the table's metadata will be deleted, while the data files will be kept in the storage ?

The table is deep cloned
Correct answer
The table is external
The user running the command has no permission to delete the data files
Your answer is incorrect
The table is managed
Delta prevents deleting files less than retention threshold, just to ensure that no long-running operations are still referencing any of the files to be deleted

Given the two tables **students_course_1** and **students_course_2**. Which of the following commands can a data engineer use to get all the students from the above two tables without duplicate records ?

```
SELECT * FROM students_course_1
CROSS JOIN
SELECT * FROM students_course_2
```

Your answer is correct

```
SELECT * FROM students_course_1
UNION
SELECT * FROM students_course_2
```

```
SELECT * FROM students_course_1
INTERSECT
SELECT * FROM students_course_2
```

```
SELECT * FROM students_course_1
OUTER JOIN
SELECT * FROM students_course_2
```

```
SELECT * FROM students_course_1
INNER JOIN
SELECT * FROM students_course_2
```

Given the following command:

```
CREATE DATABASE IF NOT EXISTS hr_db ;
```

In which of the following locations will the **hr_db** database be located?

Correct answer

dbfs:/user/hive/warehouse

dbfs:/user/hive/db_hr

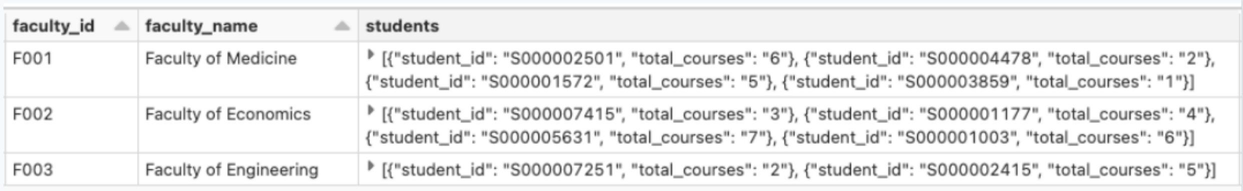
dbfs:/user/hive/databases/db_hr.db

dbfs:/user/hive/databases

Your answer is incorrect

dbfs:/user/hive

Given the following table **faculties**



faculty_id	faculty_name	students
F001	Faculty of Medicine	▸ [{"student_id": "S000002501", "total_courses": "6"}, {"student_id": "S000004478", "total_courses": "2"}, {"student_id": "S000001572", "total_courses": "5"}, {"student_id": "S000003859", "total_courses": "1"}]
F002	Faculty of Economics	▸ [{"student_id": "S000007415", "total_courses": "3"}, {"student_id": "S000001177", "total_courses": "4"}, {"student_id": "S000005631", "total_courses": "7"}, {"student_id": "S000001003", "total_courses": "6"}]
F003	Faculty of Engineering	▸ [{"student_id": "S000007251", "total_courses": "2"}, {"student_id": "S000002415", "total_courses": "5"}]

Fill in the below blank to get the students enrolled in less than 3 courses from the array column **students**

```
SELECT
  faculty_id,
  students,
  ___________ AS few_courses_students
FROM faculties
```

TRANSFORM (students, total_courses < 3)

TRANSFORM (students, i -> i.total_courses < 3)

FILTER (students, total_courses < 3)

Correct answer

FILTER (students, i -> i.total_courses < 3)

CASE WHEN students.total_courses < 3 THEN students
ELSE NULL

END

Given the following Structured Streaming query:

```
(spark.table("orders")
        .withColumn("total_after_tax", col("total")+col("tax"))
    .writeStream
        .option("checkpointLocation", checkpointPath)
        .outputMode("append")
         .______________
        .table("new_orders")
)
```

Fill in the blank to make the query executes a micro-batch to process data every 2 minutes

trigger(once=”2 minutes”)

Your answer is correct
trigger(processingTime=”2 minutes")

processingTime(”2 minutes")

trigger(”2 minutes")

trigger()

Which of the following is used by Auto Loader to load data incrementally?

DEEP CLONE
Multi-hop architecture
COPY INTO
Correct answer
Spark Structured Streaming
Databricks SQL

Which of the following statements best describes Auto Loader ?

- **Auto loader allows applying Change Data Capture (CDC) feed to update tables based on changes captured in source data.**

Your answer is correct

- **Auto loader monitors a source location, in which files accumulate, to identify and ingest only new arriving files with each command run. While the files that have already been ingested in previous runs are skipped.**
- **Auto loader allows cloning a source Delta table to a target destination at a specific version.**
- **Auto loader defines data quality expectations on the contents of a dataset, and reports the records that violate these expectations in metrics.**
- **Auto loader enables efficient insert, update, deletes, and rollback capabilities by adding a storage layer that provides better data reliability to data lakes.**

A data engineer has defined the following data quality constraint in a Delta Live Tables pipeline:

```
CONSTRAINT

CONSTRAINT valid_id EXPECT (id IS NOT NULL) _____________
```

Fill in the above blank so records violating this constraint will be added to the target table, and reported in metrics

ON VIOLATION ADD ROW
ON VIOLATION FAIL UPDATE
ON VIOLATION SUCCESS UPDATE
ON VIOLATION NULL
Your answer is correct
There is no need to add ON VIOLATION clause. By default, records

The data engineer team has a DLT pipeline that updates all the tables once and then stops. The compute resources of the pipeline continue running to allow for quick testing.
Which of the following best describes the execution modes of this DLT pipeline ?
The DLT pipeline executes in Continuous Pipeline mode under Production mode.
The DLT pipeline executes in Continuous Pipeline mode under Development mode.

Your answer is incorrect
The DLT pipeline executes in Triggered Pipeline mode under Production mode.
Correct answer
The DLT pipeline executes in Triggered Pipeline mode under Development mode.
More information is needed to determine the correct response

Which of the following will utilize Gold tables as their source?

Silver tables
Auto loader
Bronze tables
Correct answer
Dashboards
Streaming jobs

Which of the following code blocks can a data engineer use to query the existing streaming table **events** ?

spark.readStream("events")
spark.read

        .table("events")

Correct answer
spark.readStream

        .table("events")

spark.readStream()

        .table("events")

spark.stream

        .read("events")

In multi-hop architecture, which of the following statements best describes the Bronze layer ?

It maintains data that powers analytics, machine learning, and production applications
Your answer is correct
It maintains raw data ingested from various sources
It represents a filtered, cleaned, and enriched version of data
It provides business-level aggregated version of data

It provides a more refined view of the data.

Given the following Structured Streaming query

```
(spark.readStream
        .format("cloudFiles")
        .option("cloudFiles.format", "json")
        .load(ordersLocation)
     .writeStream
        .option("checkpointLocation", checkpointPath)
        .table("uncleanedOrders")
)
```

Which of the following best describe the purpose of this query in a multi-hop architecture?

The query is performing raw data ingestion into a Bronze table

The query is performing a hop from a Bronze table to a Silver table

The query is performing a hop from Silver table to a Gold table

The query is performing data transfer from a Gold table into a production application

This query is performing data quality controls prior to Silver layer

A data engineer has the following query in a Delta Live Tables pipeline:

```
CREATE LIVE TABLE aggregated_sales
AS
  SELECT store_id, sum(total)
  FROM cleaned_sales
  GROUP BY store_id
```

The pipeline is failing to start due to an error in this query

Which of the following changes should be made to this query to successfully start the DLT pipeline ?

```
CREATE STREAMING TABLE aggregated_sales
```

```
AS
  SELECT store_id, sum(total)
  FROM LIVE.cleaned_sales
  GROUP BY store_id
```

```
CREATE TABLE aggregated_sales
AS
  SELECT store_id, sum(total)
  FROM LIVE.cleaned_sales
  GROUP BY store_id
```

```
CREATE LIVE TABLE aggregated_sales
AS
  SELECT store_id, sum(total)
  FROM LIVE.cleaned_sales
  GROUP BY store_id
```

```
CREATE STREAMING LIVE TABLE aggregated_sales
AS
  SELECT store_id, sum(total)
  FROM cleaned_sales
  GROUP BY store_id
```

```
CREATE STREAMING LIVE TABLE aggregated_sales
AS
  SELECT store_id, sum(total)
  FROM STREAM(cleaned_sales)
  GROUP BY store_id
```

A data engineer has defined the following data quality constraint in a Delta Live Tables pipeline:

```
CONSTRAINT valid_id EXPECT (id IS NOT NULL) _______________
```

Fill in the above blank so records violating this constraint will be dropped, and reported in metrics

ON VIOLATION DROP ROW

ON VIOLATION FAIL UPDATE

ON VIOLATION DELETE ROW

ON VIOLATION DISCARD ROW

There is no need to add ON VIOLATION clause. By default, records violating the constraint will be discarded, and reported as invalid in the event log

Which of the following compute resources is available in Databricks SQL ?

Single-node clusters
Multi-nodes clusters
On-premises clusters
Correct answer
SQL warehouses
SQL engines

Which of the following is the benefit of using the Auto Stop feature of Databricks SQL warehouses ?

Improves the performance of the warehouse by automatically stopping ideal services
Correct answer
Minimizes the total running time of the warehouse
Provides higher security by automatically stopping unused ports of the warehouse
Increases the availability of the warehouse by automatically stopping long-running SQL queries
Databricks SQL does not have Auto Stop feature

Which of the following alert destinations is **Not** supported in Databricks SQL ?
Slack
Webhook
Correct answer
SMS
Microsoft Teams
Email

A data engineering team has a long-running multi-tasks Job. The team members need to be notified when the run of this job completes.

Which of the following approaches can be used to send emails to the team members when the job completes ?

They can use Job API to programmatically send emails according to each task status
Your answer is correct
They can configure email notifications settings in the job page
There is no way to notify users when the job completes
Only Job owner can be configured to be notified when the job completes

They can configure email notifications settings per notebook in the task page

A data engineer wants to increase the cluster size of an existing Databricks SQL warehouse.

Which of the following is the benefit of increasing the cluster size of Databricks SQL warehouses ?

Correct answer
Improves the latency of the queries execution
Speeds up the start up time of the SQL warehouse
Reduces cost since large clusters use Spot instances
The cluster size of SQL warehouses is not configurable. Instead, they can increase the number of clusters
The cluster size can not be changed

Which of the following describes Cron syntax in Databricks Jobs ?

It's an expression to represent the maximum concurrent runs of a job
Your answer is correct
It's an expression to represent complex job schedule that can be defined programmatically
It's an expression to represent the retry policy of a job
It's an expression to describe the email notification events (start, success, failure)
It's an expression to represent the run timeout of a job

The data engineer team has a DLT pipeline that updates all the tables at defined intervals until manually stopped. The compute resources terminate when the pipeline is stopped.

Which of the following best describes the execution modes of this DLT pipeline ?

The DLT pipeline executes in Continuous Pipeline mode under Production mode.

The DLT pipeline executes in Continuous Pipeline mode under Development mode.

The DLT pipeline executes in Triggered Pipeline mode under Production mode.

The DLT pipeline executes in Triggered Pipeline mode under Development mode.

More information is needed to determine the correct response

Which part of the Databricks Platform can a data engineer use to grant permissions on tables to users ?

Data Studio

Cluster event log

Workflows

DBFS

Correct answer

Data Explorer

Which of the following commands can a data engineer use to grant full permissions to the HR team on the table **employees** ?

GRANT FULL PRIVILEGES ON TABLE employees TO hr_team

GRANT FULL PRIVILEGES ON TABLE hr_team TO employees

Correct answer

GRANT ALL PRIVILEGES ON TABLE employees TO hr_team

GRANT ALL PRIVILEGES ON TABLE hr_team TO employees

GRANT SELECT, MODIFY, CREATE, READ_METADATA ON TABLE employees TO hr_team

A data engineer uses the following SQL query:

```
GRANT MODIFY ON TABLE employees TO hr_team
```

Which of the following describes the ability given by the `MODIFY` privilege ?

It gives the ability to add data from the table

It gives the ability to delete data from the table

It gives the ability to modify data in the table

Correct answer

All the above abilities are given by the `MODIFY` privilege

None of these options correctly describe the ability given by the `MODIFY` privilege

One of the foundational technologies provided by the Databricks Lakehouse Platform is an open-source, file-based storage format that brings reliability to data lakes.

Which of the following technologies is being described in the above statement?

Delta Lives Tables (DLT)

Delta Lake

Apache Spark

Unity Catalog

Photon

Which of the following commands can a data engineer use to purge stale data files of a Delta table?

DELETE

GARBAGE COLLECTION

CLEAN

VACUUM

OPTIMIZE

In Databricks Repos, which of the following operations a data engineer can use to save local changes of a repo to its remote repository ?

Create Pull Request

Commit & Pull

Correct answer

Commit & Push

Merge & Push

Merge & Pull

In Delta Lake tables, which of the following is the primary format for the transaction log files?

Delta

Parquet

Correct answer

JSON

Hive-specific format

XML

Which of the following functionalities can be performed in Databricks Repos ?

Create pull requests
Create new remote Git repositories
Delete branches
Create CI/CD pipelines
Correct answer
Pull from a remote Git repository

Which of the following locations completely hosts the customer data ?

Customer's cloud account
Control plane
Databricks account
Databricks-managed cluster
Repos

If the default notebook language is Python, which of the following options a data engineer can use to run SQL commands in this Python Notebook ?

They need first to import the SQL library in a cell
This is not possible! They need to change the default language of the notebook to SQL
Databricks detects cells language automatically, so they can write SQL syntax in any cell
They can add `%language` magic command at the start of a cell to force language detection.
Your answer is correct
They can add `%sql` at the start of a cell.

A junior data engineer uses the built-in Databricks Notebooks versioning for source control. A senior data engineer recommended using Databricks Repos instead.

Which of the following could explain why Databricks Repos is recommended instead of Databricks Notebooks versioning?

Databricks Repos supports creating and managing branches for development work.
Databricks Repos automatically tracks the changes and keeps the history.
Databricks Repos allows users to resolve merge conflicts
Databricks Repos allows users to restore previous versions of a notebook

All of these advantages explain why Databricks Repos is recommended instead of Notebooks versioning

Which of the following services provides a data warehousing experience to its users?

Databricks SQL
Databricks Machine Learning
Data Science and Engineering Workspace
Unity Catalog
Delta Lives Tables (DLT)

A data engineer noticed that there are unused data files in the directory of a Delta table. They executed the VACUUM command on this table; however, only some of those unused data files have been deleted.

Which of the following could explain why only some of the unused data files have been deleted after running the VACUUM command ?

- **The deleted data files were larger than the default size threshold. While the remaining files are smaller than the default size threshold and can not be deleted.**
- **The deleted data files were smaller than the default size threshold. While the remaining files are larger than the default size threshold and can not be deleted.**
- **The deleted data files were older than the default retention threshold. While the remaining files are newer than the default retention threshold and can not be deleted.**
- **The deleted data files were newer than the default retention threshold. While the remaining files are older than the default retention threshold and can not be deleted.**
- **More information is needed to determine the correct answer**

The data engineering team has a Delta table called **products** that contains products' details including the net price.

Which of the following code blocks will apply a 50% discount on all the products where the price is greater than 1000 and save the new price to the table?

UPDATE products SET price = price * 0.5 WHERE price >= 1000;

SELECT price * 0.5 AS new_price FROM products WHERE price > 1000;

MERGE INTO products WHERE price < 1000 WHEN MATCHED UPDATE price = price * 0.5;

Correct answer

UPDATE products SET price = price * 0.5 WHERE price > 1000;

MERGE INTO products WHERE price > 1000 WHEN MATCHED UPDATE price = price * 0.5;

A data engineer wants to create a relational object by pulling data from two tables. The relational object will only be used in the current session. In order to save on storage costs, the date engineer wants to avoid copying and storing physical data.

Which of the following relational objects should the data engineer create?

External table

Correct answer

Temporary view

Managed table

Global Temporary view

View

A data engineer has a database named **db_hr**, and they want to know where this database was created in the underlying storage.

Which of the following commands can the data engineer use to complete this task?

DESCRIBE db_hr
DESCRIBE EXTENDED db_hr
Your answer is correct
DESCRIBE DATABASE db_hr
SELECT location FROM db_hr.db
There is no need for a command since all databases are created under the default hive metastore directory

When dropping a Delta table, which of the following explains why both the table's metadata and the data files will be deleted ?

The table is shallow cloned
The table is external
The user running the command has the necessary permissions to delete the data files
Your answer is correct
The table is managed
The data files are older than the default retention period

Given the following commands:

```
CREATE DATABASE db_hr;

USE db_hr;
CREATE TABLE employees;
```

In which of the following locations will the employees table be located?

dbfs:/user/hive/warehouse
Correct answer
dbfs:/user/hive/warehouse/db_hr.db
dbfs:/user/hive/warehouse/db_hr
dbfs:/user/hive/databases/db_hr.db
More information is needed to determine the correct answer

Which of the following commands a data engineer can use to register the table **orders** from an existing SQLite database ?

```
CREATE TABLE orders
  USING sqlite
  OPTIONS (
    url "jdbc:sqlite:/bookstore.db",
    dbtable "orders"
  )
```

Your answer is correct

```
CREATE TABLE orders
  USING org.apache.spark.sql.jdbc
  OPTIONS (
    url "jdbc:sqlite:/bookstore.db",
    dbtable "orders"
  )
```

```
CREATE TABLE orders
  USING cloudfiles
  OPTIONS (
    url "jdbc:sqlite:/bookstore.db",
    dbtable "orders"
  )
```

```
CREATE TABLE orders
  USING EXTERNAL
  OPTIONS (
    url "jdbc:sqlite:/bookstore.db",
    dbtable "orders"
  )
```

```
CREATE TABLE orders
USING DATABASE
OPTIONS (
    url "jdbc:sqlite:/bookstore.db",
    dbtable "orders"
)
```

Which of the following code blocks can a data engineer use to create a Python function to multiply two integers and return the result?

```
def multiply_numbers(num1, num2):
    print(num1 * num2)
```

```
def fun: multiply_numbers(num1, num2):
    return num1 * num2
```

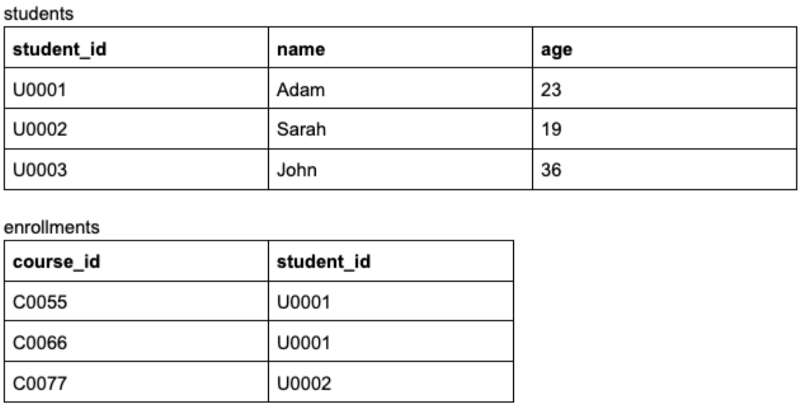
Correct answer

```
def multiply_numbers(num1, num2):
    return num1 * num2
```

```
fun multiply_numbers(num1, num2):
    return num1 * num2
```

```
fun def multiply_numbers(num1, num2):
    return num1 * num2
```

Given the following 2 tables:



students

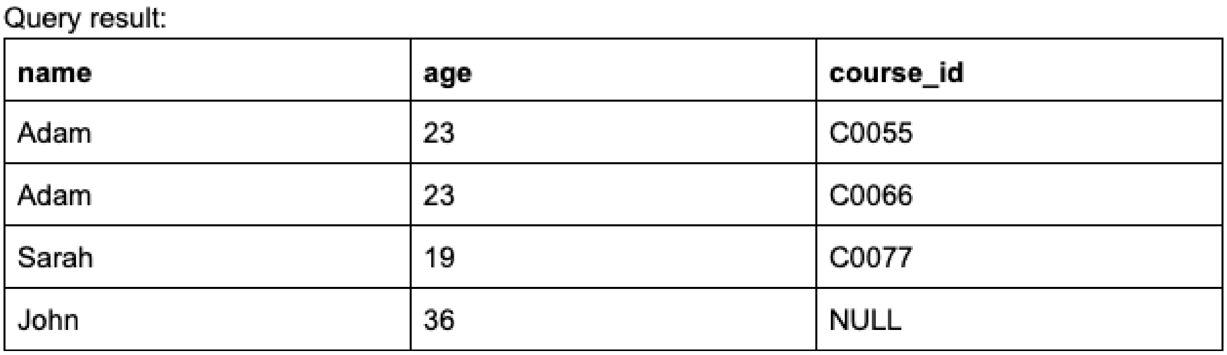
student_id	name	age
U0001	Adam	23
U0002	Sarah	19
U0003	John	36

enrollments

course_id	student_id
C0055	U0001
C0066	U0001
C0077	U0002

Fill in the blank to make the following query returns the below result:

```
SELECT students.name, students.age, enrollments.course_id
FROM students
_____________ enrollments
ON students.student_id = enrollments.student_id
```



Query result:

name	age	course_id
Adam	23	C0055
Adam	23	C0066
Sarah	19	C0077
John	36	NULL

RIGHT JOIN
Correct answer
LEFT JOIN
INNER JOIN
ANTI JOIN
CROSS JOIN

Which of the following SQL keywords can be used to rotate rows of a table by turning row values into multiple columns ?

ROTATE
TRANSFORM
Your answer is correct
PIVOT
GROUP BY
ZORDER BY

Fill in the below blank to get the number of courses incremented by 1 for each student in array column **students**.

```
SELECT
  faculty_id,
  students,
  ___________ AS new_totals
FROM faculties
```

TRANSFORM (students, total_courses + 1)
Correct answer
TRANSFORM (students, i -> i.total_courses + 1)
FILTER (students, total_courses + 1)
Your answer is incorrect
FILTER (students, i -> i.total_courses + 1)
CASE WHEN students.total_courses IS NOT NULL THEN students.total_courses + 1

ELSE NULL

END

Fill in the below blank to successfully create a table using data from CSV files located at **/path/input**

```
CREATE TABLE my_table
(col1 STRING, col2 STRING)
____________
OPTIONS (header = "true",
        delimiter = ";")
LOCATION = "/path/input"
```

FROM CSV
Your answer is correct
USING CSV
USING DELTA
AS
AS CSV

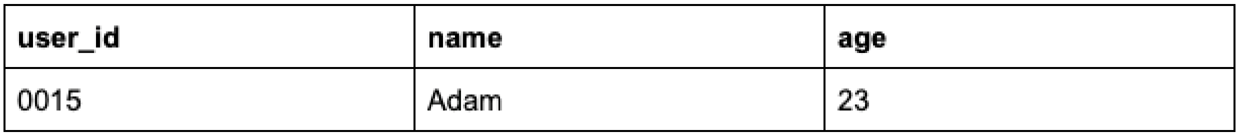
Which of the following statements best describes the usage of `CREATE SCHEMA` command ?

It's used to create a table schema (columns names and datatype)
It's used to create a Hive catalog
It's used to infer and store schema in "cloudFiles.schemaLocation"
Correct answer
It's used to create a database
It's used to merge the schema when writing data into a target table

Which of the following statements is **Not** true about CTAS statements ?

CTAS statements automatically infer schema information from query results
Correct answer
CTAS statements support manual schema declaration
CTAS statements stand for CREATE TABLE _ AS SELECT statement
With CTAS statements, data will be inserted during the table creation
All these statements are Not true about CTAS statements

Which of the following SQL commands will append this new row to the existing Delta table **users**?



user_id	name	age
0015	Adam	23

APPEND INTO users VALUES (“0015”, “Adam”, 23)
INSERT VALUES (“0015”, “Adam”, 23) INTO users
APPEND VALUES (“0015”, “Adam”, 23) INTO users
Your answer is correct
INSERT INTO users VALUES (“0015”, “Adam”, 23)
UPDATE users VALUES (“0015”, “Adam”, 23)

Given the following Structured Streaming query:

```
(spark.table("orders")
        .withColumn("total_after_tax", col("total")+col("tax"))
    .writeStream
        .option("checkpointLocation", checkpointPath)
        .outputMode("append")
        .___________
        .table("new_orders") )
```

Fill in the blank to make the query executes multiple micro-batches to process all available data, then stops the trigger.

trigger(“micro-batches”)
trigger(once=True)
trigger(processingTime=”0 seconds")
trigger(micro-batches=True)
Correct answer
trigger(availableNow=True)

Which of the following techniques allows Auto Loader to track the ingestion progress and store metadata of the discovered files ?

mergeSchema
COPY INTO
Watermarking
Correct answer
Checkpointing
Z-Ordering

A data engineer has defined the following data quality constraint in a Delta Live Tables pipeline:

```
CONSTRAINT valid_id EXPECT (id IS NOT NULL) _____________
```

Fill in the above blank so records violating this constraint cause the pipeline to fail.

ON VIOLATION FAIL
Correct answer
ON VIOLATION FAIL UPDATE
ON VIOLATION DROP ROW
ON VIOLATION FAIL PIPELINE
There is no need to add ON VIOLATION clause. By default, records violating the constraint cause the pipeline to fail.

In multi-hop architecture, which of the following statements best describes the Silver layer tables?

They maintain data that powers analytics, machine learning, and production applications
They maintain raw data ingested from various sources
The table structure in this layer resembles that of the source system table structure with any additional metadata columns like the load time, and input file name.
They provide business-level aggregated version of data
Your answer is correct
They provide a more refined view of raw data, where it's filtered, cleaned, and enriched.

The data engineer team has a DLT pipeline that updates all the tables at defined intervals until manually stopped. The compute resources of the pipeline continue running to allow for quick testing.

Which of the following best describes the execution modes of this DLT pipeline ?

The DLT pipeline executes in Continuous Pipeline mode under Production mode.
Correct answer
The DLT pipeline executes in Continuous Pipeline mode under Development mode.
The DLT pipeline executes in Triggered Pipeline mode under Production mode.
The DLT pipeline executes in Triggered Pipeline mode under Development mode.
More information is needed to determine the correct response

Given the following Structured Streaming query:

```
(spark.readStream
        .table("cleanedOrders")
        .groupBy("productCategory")
        .agg(sum("totalWithTax"))
    .writeStream
        .option("checkpointLocation", checkpointPath)
        .outputMode("complete")
        .table("aggregatedOrders")
)
```

Which of the following best describe the purpose of this query in a multi-hop architecture?

The query is performing raw data ingestion into a Bronze table

The query is performing a hop from a Bronze table to a Silver table

Your answer is correct

The query is performing a hop from Silver layer to a Gold table

The query is performing data transfer from a Gold table into a production application

This query is performing data quality controls prior to Silver layer

Given the following Structured Streaming query:

```
(spark.readStream
        .table("orders")
    .writeStream
        .option("checkpointLocation", checkpointPath)
        .table("Output_Table")
)
```

Which of the following is the trigger Interval for this query ?

Every half second

Every half min

Every half hour

The query will run in batch mode to process all available data at once, then the trigger stops.

More information is needed to determine the correct response

A data engineer has the following query in a Delta Live Tables pipeline

```
CREATE STREAMING LIVE TABLE sales_silver
AS
  SELECT store_id, total + tax AS total_after_tax
  FROM LIVE.sales_bronze
```

The pipeline is failing to start due to an error in this query.

Which of the following changes should be made to this query to successfully start the DLT pipeline ?

```
CREATE LIVE TABLE sales_silver
AS
  SELECT store_id, total + tax AS total_after_tax
  FROM STREAMING(LIVE.sales_bronze)
```

```
CREATE STREAMING TABLE sales_silver
AS
  SELECT store_id, total + tax AS total_after_tax
  FROM STREAM(LIVE.sales_bronze)
```

```
CREATE STREAMING LIVE TABLE sales_silver
AS
  SELECT store_id, total + tax AS total_after_tax
  FROM STREAMING(sales_bronze)
```

```
CREATE STREAMING LIVE TABLE sales_silver
AS
  SELECT store_id, total + tax AS total_after_tax
  FROM STREAMING(LIVE.sales_bronze)
```

Correct answer

```
CREATE STREAMING LIVE TABLE sales_silver
AS
  SELECT store_id, total + tax AS total_after_tax
  FROM STREAM(LIVE.sales_bronze)
```

In multi-hop architecture, which of the following statements best describes the Gold layer tables?

- **They provide a more refined view of the data**
- **They maintain raw data ingested from various sources**
- **The table structure in this layer resembles that of the source system table structure with any additional metadata columns like the load time, and input file name.**
- **They provide business-level aggregations that power analytics, machine learning, and production applications**
- **They represent a filtered, cleaned, and enriched version of data**

The data engineer team has a DLT pipeline that updates all the tables once and then stops. The compute resources of the pipeline terminate when the pipeline is stopped.

Which of the following best describes the execution modes of this DLT pipeline ?

The DLT pipeline executes in Continuous Pipeline mode under Production mode.
The DLT pipeline executes in Continuous Pipeline mode under Development mode.
Correct answer
The DLT pipeline executes in Triggered Pipeline mode under Production mode.
The DLT pipeline executes in Triggered Pipeline mode under Development mode.
More information is needed to determine the correct response

A data engineer needs to determine whether to use Auto Loader or COPY INTO command in order to load input data files incrementally.

In which of the following scenarios should the data engineer use Auto Loader over COPY INTO command ?

Correct answer
If they are going to ingest files in the order of millions or more over time
If they are going to ingest few number of files in the order of thousands
If they are going to load a subset of re-uploaded files
If the data schema is not going to evolve frequently
There is no difference between using Auto Loader and Copy Into command

From which of the following locations can a data engineer set a schedule to automatically refresh a Databricks SQL query ?

From the jobs UI
From the SQL warehouses page in Databricks SQL
From the Alerts page in Databricks SQL
Correct answer
From the query's page in Databricks SQL
There is no way to automatically refresh a query in Databricks SQL. Schedules can be set only for dashboards to refresh their underlying queries.

Databricks provides a declarative ETL framework for building reliable and maintainable data processing pipelines, while maintaining table dependencies and data quality.

Which of the following technologies is being described above?

Delta Live Tables

Delta Lake

Databricks Jobs

Unity Catalog Linage

Databricks SQL

Which of the following services can a data engineer use for orchestration purposes in Databricks platform ?

Delta Live Tables

Cluster Pools

Your answer is correct

Databricks Jobs

Data Explorer

Unity Catalog Linage

A data engineer has a Job with multiple tasks that takes more than 2 hours to complete. In the last run, the final task unexpectedly failed.

Which of the following actions can the data engineer perform to complete this Job Run while minimizing the execution time ?

They can rerun this Job Run to execute all the tasks

Correct answer

They can repair this Job Run so only the failed tasks will be re-executed

They need to delete the failed Run, and start a new Run for the Job

They can keep the failed Run, and simply start a new Run for the Job

They can run the Job in Production mode which automatically retries execution in case of errors

A data engineering team has a multi-tasks Job in production. The team members need to be notified in the case of job failure.

Which of the following approaches can be used to send emails to the team members in the case of job failure ?

They can use Job API to programmatically send emails according to each task status
Your answer is correct
They can configure email notifications settings in the job page
There is no way to notify users in the case of job failure
Only Job owner can be configured to be notified in the case of job failure
They can configure email notifications settings per notebook in the task page

For production jobs, which of the following cluster types is recommended to use?

All-purpose clusters
Production clusters
Correct answer
Job clusters
On-premises clusters
Serverless clusters

In Databricks Jobs, which of the following approaches can a data engineer use to configure a linear dependency between **Task A** and **Task B** ?

They can select the Task A in the Depends On field of the Task B configuration

They can assign Task A an Order number of 1, and assign Task B an Order number of 2
They can visually drag and drop an arrow from Task A to Task B in the Job canvas
They can configure the dependency at the notebook level using the dbutils.jobs utility
Databricks Jobs do not support linear dependency between tasks. This can only be achieved in Delta Live Tables pipelines

Which part of the Databricks Platform can a data engineer use to revoke permissions from users on tables ?

Data Explorer
Cluster event log
Workspace Admin Console
DBFS

There is no way to revoke permissions in Databricks platform. The data engineer needs to clone the table with the updated permissions

A data engineer uses the following SQL query:

```
GRANT USAGE ON DATABASE sales_db TO finance_team
```

Which of the following is the benefit of the **USAGE** privilege ?

Gives read access on the database

Gives full permissions on the entire database

Gives the ability to view database objects and their metadata

Correct answer

No effect! but it's required to perform any action on the database

USAGE privilege is not part of the Databricks governance model

In which of the following locations can a data engineer change the owner of a table?

In DBFS, from the properties tab of the table's data files

In Data Explorer, under the Permissions tab of the table's page

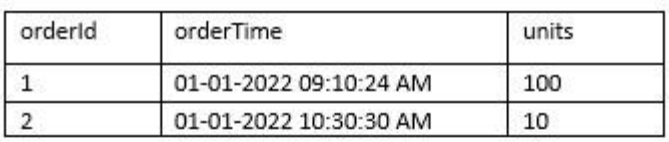
Your answer is correct

In Data Explorer, from the Owner field in the table's page

In Data Explorer, under the Permissions tab of the database's page, since owners are set at database-level

In Data Explorer, from the Owner field in the database's page, since owners are set at database-level

You were asked to create a table that can store the below data, orderTime is a timestamp but the finance team when they query this data normally prefer the orderTime in date format, you would like to create a calculated column that can convert the orderTime column timestamp datatype to date and store it, fill in the blank to complete the DDL.



orderId	orderTime	units
1	01-01-2022 09:10:24 AM	100
2	01-01-2022 10:30:30 AM	10

```
CREATE TABLE orders (
    orderId int,
    orderTime timestamp,
    orderdate date _____________________________________________ ,
    units int)
```

`AS DEFAULT (CAST(orderTime as DATE))`

Correct answer

`GENERATED ALWAYS AS (CAST(orderTime as DATE))`

`GENERATED DEFAULT AS (CAST(orderTime as DATE))`

`AS (CAST(orderTime as DATE))`

Delta lake does not support calculated columns, value should be inserted into the table as part of the ingestion process

The data engineering team noticed that one of the job fails randomly as a result of using spot instances, what feature in Jobs/Tasks can be used to address this issue so the job is more stable when using spot instances?

Use Databrick REST API to monitor and restart the job
Use Jobs runs, active runs UI section to monitor and restart the job
Add second task and add a check condition to rerun the first task if it fails
Restart the job cluster, job automatically restarts
Correct answer
Add a retry policy to the task

What is the main difference between AUTO LOADER and COPY INTO?

COPY INTO supports schema evolution.
AUTO LOADER supports schema evolution.
COPY INTO supports file notification when performing incremental loads.
AUTO LOADER supports reading data from Apache Kafka

Correct answer
AUTO LOADER Supports file notification when performing incremental loads.

Why does AUTO LOADER require schema location?

Schema location is used to store user provided schema
Schema location is used to identify the schema of target table
AUTO LOADER does not require schema location, because its supports Schema evolution
Correct answer
Schema location is used to store schema inferred by AUTO LOADER
Schema location is used to identify the schema of target table and source table

Which of the following statements are incorrect about the lakehouse

Support end-to-end streaming and batch workloads
Supports ACID
Support for diverse data types that can store both structured and unstructured
Supports BI and Machine learning
Correct answer
Storage is coupled with Compute

You are designing a data model that works for both machine learning using images and Batch ETL/ELT workloads. Which of the following features of data lakehouse can help you meet the needs of both workloads?

Data lakehouse requires very little data modeling.
Data lakehouse combines compute and storage for simple governance.
Data lakehouse provides autoscaling for compute clusters.
Correct answer
Data lakehouse can store unstructured data and support ACID transactions.
Data lakehouse fully exists in the cloud.

Which of the following locations in Databricks product architecture hosts jobs/pipelines and queries?
Data plane
Correct answer
Control plane
Databricks Filesystem
JDBC data source
Databricks web application

You are currently working on a notebook that will populate a reporting table for downstream process consumption, this process needs to run on a schedule every hour. what type of cluster are you going to use to set up this job?

Since it's just a single job and we need to run every hour, we can use an all-purpose cluster
Correct answer
The job cluster is best suited for this purpose.
Use Azure VM to read and write delta tables in Python
Use delta live table pipeline to run in continuous mode

Which of the following developer operations in CI/CD flow can be implemented in Databricks Repos?
Merge when code is committed
Pull request and review process
Correct answer
Trigger Databricks Repos API to pull the latest version of code into production folder
Resolve merge conflicts
Delete a branch

You are currently working with the second team and both teams are looking to modify the same notebook, you noticed that the second member is copying the notebooks to the personal folder to edit and replace the collaboration notebook, which notebook feature do you recommend to make the process easier to collaborate.

Databricks notebooks should be copied to a local machine and setup source control locally to version the notebooks
Databricks notebooks support automatic change tracking and versioning
Correct answer
Databricks Notebooks support real-time coauthoring on a single notebook
Databricks notebooks can be exported into dbc archive files and stored in data lake
Databricks notebook can be exported as HTML and imported at a later time

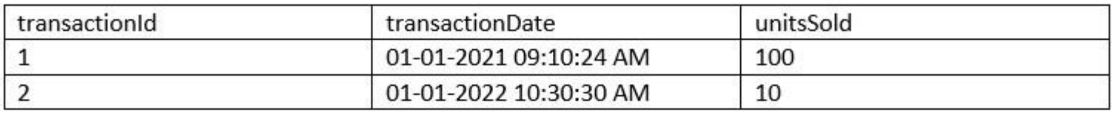
You are currently working on a project that requires the use of SQL and Python in a given notebook, what would be your approach

Create two separate notebooks, one for SQL and the second for Python
Correct answer
A single notebook can support multiple languages, use the magic command to switch between the two.
Use an All-purpose cluster for python, SQL endpoint for SQL
Use job cluster to run python and SQL Endpoint for SQL

Which of the following statements are correct on how Delta Lake implements a lake house?

Delta lake uses a proprietary format to write data, optimized for cloud storage
Using Apache Hadoop on cloud object storage
Delta lake always stores meta data in memory vs storage
Correct answer
Delta lake uses open source, open format, optimized cloud storage and scalable meta data
Delta lake stores data and meta data in computes memory

You were asked to create or overwrite an existing delta table to store the below transaction data.



transactionId	transactionDate	unitsSold
1	01-01-2021 09:10:24 AM	100
2	01-01-2022 10:30:30 AM	10

```
CREATE OR REPLACE DELTA TABLE transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
```

```
CREATE OR REPLACE TABLE IF EXISTS transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
FORMAT DELTA
```

```
CREATE IF EXSITS REPLACE TABLE transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
```

Correct answer

```
CREATE OR REPLACE TABLE transactions (
transactionId int,
transactionDate timestamp,
unitsSold int)
```

if you run the command `VACUUM transactions retain 0 hours`? What is the outcome of this command?

Command will be successful, but no data is removed
Command will fail if you have an active transaction running
Correct answer
Command will fail, you cannot run the command with retentionDurationcheck enabled
Command will be successful, but historical data will be removed

Command runs successful and compacts all of the data in the table

You noticed a colleague is manually copying the data to the backup folder prior to running an update command, incase if the update command did not provide the expected outcome so he can use the backup copy to replace table, which Delta Lake feature would you recommend simplifying the process?

Use time travel feature to refer old data instead of manually copying
Use DEEP CLONE to clone the table prior to update to make a backup copy
Use SHADOW copy of the table as preferred backup choice
Cloud object storage retains previous version of the file
Cloud object storage automatically backups the data

Which one of the following is not a Databricks lakehouse object?

Tables
Views
Database/Schemas
Catalog
Functions
Correct answer
Stored Procedures

What type of table is created when you create delta table with below command?

```
CREATE TABLE transactions USING DELTA LOCATION
"DBFS:/mnt/bronze/transactions"
```

Managed delta table
Correct answer
External table
Managed table
Temp table
Delta Lake table

Which of the following command can be used to drop a managed delta table and the underlying files in the storage?

`DROP TABLE table_name CASCADE`
Correct answer
`DROP TABLE table_name`
Use `DROP TABLE table_name` **command and manually delete files using command** `dbutils.fs.rm("/path",True)`

`DROP TABLE table_name INCLUDE_FILES`
`DROP TABLE table` and run `VACUUM` command

Which of the following is the correct statement for a session scoped temporary view?

Correct answer
Temporary views are lost once the notebook is detached and re-attached
Temporary views stored in memory
Temporary views can be still accessed even if the notebook is detached and attached
Temporary views can be still accessed even if cluster is restarted
Temporary views are created in local_temp database

Which of the following is correct for the global temporary view?
global temporary views cannot be accessed once the notebook is detached and attached
global temporary views can be accessed across many clusters
Correct answer
global temporary views can be still accessed even if the notebook is detached and attached
global temporary views can be still accessed even if the cluster is restarted
global temporary views are created in a database called `temp` database

You are currently working on reloading customer_sales tables using the below query

```
INSERT OVERWRITE customer_sales
SELECT * FROM customers c
INNER JOIN sales_monthly s on s.customer_id = c.customer_id
```

After you ran the above command, the Marketing team quickly wanted to review the old data that was in the table. How does INSERT OVERWRITE impact the data in the `customer_sales` table if you want to see the previous version of the data prior to running the above statement?

Overwrites the data in the table, all historical versions of the data, you can not time travel to previous versions
Correct answer
Overwrites the data in the table but preserves all historical versions of the data, you can time travel to previous versions
Overwrites the current version of the data but clears all historical versions of the data, so you can not time travel to previous versions.
Appends the data to the current version, you can time travel to previous versions

By default, overwrites the data and schema, you cannot perform time travel

Which of the following SQL statement can be used to query a table by eliminating duplicate rows from the query results?

Correct answer

```
SELECT DISTINCT * FROM table_name
```

```
SELECT DISTINCT * FROM table_name HAVING COUNT(*) > 1
```

```
SELECT DISTINCT_ROWS (*) FROM table_name
```

```
SELECT * FROM table_name GROUP BY * HAVING COUNT(*) < 1
```

```
SELECT * FROM table_name GROUP BY * HAVING COUNT(*) > 1
```

Which of the below SQL Statements can be used to create a SQL UDF to convert Celsius to Fahrenheit and vice versa, you need to pass two parameters to this function one, actual temperature, and the second that identifies if its needs to be converted to Fahrenheit or Celcius with a one-word letter F or C?

`select udf_convert(60,'C')` will result in 15.5

`select udf_convert(10,'F')` will result in 50

```
 CREATE UDF FUNCTION udf_convert(temp DOUBLE, measure STRING)
    RETURNS DOUBLE
    RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
            ELSE (temp – 33 ) * 5/9
           END
```

```
CREATE UDF FUNCTION udf_convert(temp DOUBLE, measure STRING)
    RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
            ELSE (temp – 33 ) * 5/9
           END
```

```
CREATE FUNCTION udf_convert(temp DOUBLE, measure STRING)
RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
          ELSE (temp – 33 ) * 5/9
       END
```

Correct answer

```
CREATE FUNCTION udf_convert(temp DOUBLE, measure STRING)
RETURNS DOUBLE
RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
        ELSE (temp – 33 ) * 5/9
       END
```

```
CREATE USER FUNCTION udf_convert(temp DOUBLE, measure STRING)
RETURNS DOUBLE
RETURN CASE WHEN measure == 'F' then (temp * 9/5) + 32
          ELSE (temp – 33 ) * 5/9
  END
```

You are trying to calculate total sales made by all the employees by parsing a complex struct data type that stores employee and sales data, how would you approach this in SQL Table definition,

```
batchId INT, performance ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>,
insertDate TIMESTAMP
```

Sample data of `performance` column

```
[
{ "employeeId":1234
"sales" : 10000},

{ "employeeId":3232
"sales" : 30000}
]
```

Calculate total sales made by all the employees?

Sample data with create table syntax for the data:

```
create or replace table sales as
select 1 as batchId ,
  from_json('[{ "employeeId":1234,"sales" : 10000 },{
"employeeId":3232,"sales" : 30000 }]',
         'ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>') as performance,
  current_timestamp() as insertDate
union all
select 2 as batchId ,
  from_json('[{ "employeeId":1235,"sales" : 10500 },{
"employeeId":3233,"sales" : 32000 }]',
                'ARRAY<STRUCT<employeeId: BIGINT, sales: INT>>') as
performance,
                current_timestamp() as insertDate
```

```
WITH CTE as (SELECT EXPLODE (performance) FROM table_name)
SELECT SUM (performance.sales) FROM CTE
```

```
WITH CTE as (SELECT FLATTEN (performance) FROM table_name)
SELECT SUM (sales) FROM CTE
```

Correct answer

```
select aggregate(flatten(collect_list(performance.sales)), 0, (x, y) -> x + y)
as  total_sales from sales
```

`SELECT SUM(SLICE (performance, sales)) FROM employee`

```
select reduce(flatten(collect_list(performance:sales)), 0, (x, y) -> x + y)
as  total_sales from sales
```

Which of the following statements can be used to test the functionality of code to test number of rows in the table equal to 10 in python?

```
row_count = spark.sql("select count(*) from table").collect()[0][0]
```

`assert (row_count = 10, "Row count did not match")`

`assert if (row_count = 10, "Row count did not match")`

Correct answer

`assert row_count == 10, "Row count did not match"`

`assert if row_count == 10, "Row count did not match"`

`assert row_count = 10, "Row count did not match"`

Question 26Skipped

How do you handle failures gracefully when writing code in Pyspark, fill in the blanks to complete the below statement

```
_____


Spark.read.table("table_name").select("column").write.mode("append").SaveAsTable("new_table_name")

_____

    print(f"query failed")
```

try: failure:

try: catch:

Correct answer

try: except:

try: fail:

try: error:

You are working on a process to query the table based on batch date, and batch date is an input parameter and expected to change every time the program runs, what is the best way to we can parameterize the query to run without manually changing the batch date?

Correct answer

Create a notebook parameter for batch date and assign the value to a python variable and use a spark data frame to filter the data based on the python variable

Create a dynamic view that can calculate the batch date automatically and use the view to query the data

There is no way we can combine python variable and spark code

Manually edit code every time to change the batch date

Store the batch date in the spark configuration and use a spark data frame to filter the data based on the spark configuration.

Which of the following commands results in the successful creation of a view on top of the delta stream(stream on delta table)?

`Spark.read.format("delta").table("sales").createOrReplaceTempView("streaming_vw")`

Correct answer

`Spark.readStream.format("delta").table("sales").createOrReplaceTempView("streaming_vw")`

`Spark.read.format("delta").table("sales").mode("stream").createOrReplaceTempView("streaming_vw")`

`Spark.read.format("delta").table("sales").trigger("stream").createOrReplaceTempView("streaming_vw")`

`Spark.read.format("delta").stream("sales").createOrReplaceTempView("streaming_vw")`

You can not create a view on streaming data source.

Which of the following techniques structured streaming uses to create an end-to-end fault tolerance?

Checkpointing and Water marking

Write ahead logging and water marking

Correct answer

Checkpointing and idempotent sinks

Write ahead logging and idempotent sinks

Stream will failover to available nodes in the cluste

Which of the following two options are supported in identifying the arrival of new files, and incremental data from Cloud object storage using Auto Loader?

Correct answer

Directory listing, File notification

Checking pointing, watermarking

Writing ahead logging, read head logging

File hashing, Dynamic file lookup

Checkpointing and Write ahead logging

Which of the following data workloads will utilize a Bronze table as its destination?

A job that aggregates cleaned data to create standard summary statistics

A job that queries aggregated data to publish key insights into a dashboard

Correct answer

A job that ingests raw data from a streaming source into the Lakehouse

A job that develops a feature set for a machine learning application

A job that enriches data by parsing its timestamps into a human-readable format

Which of the following data workloads will utilize a silver table as its source?
A job that enriches data by parsing its timestamps into a human-readable format
A job that queries aggregated data that already feeds into a dashboard
A job that ingests raw data from a streaming source into the Lakehouse
Correct answer
A job that aggregates cleaned data to create standard summary statistics
A job that cleans data by removing malformatted records

Which of the following data workloads will utilize a gold table as its source?
A job that enriches data by parsing its timestamps into a human-readable format
Correct answer
A job that queries aggregated data that already feeds into a dashboard
A job that ingests raw data from a streaming source into the Lakehouse
A job that aggregates cleaned data to create standard summary statistics
A job that cleans data by removing malformatted records

You are currently asked to work on building a data pipeline, you have noticed that you are currently working with a data source that has a lot of data quality issues and you need to monitor data quality and enforce it as part of the data ingestion process, which of the following tools can be used to address this problem?
AUTO LOADER
Correct answer
DELTA LIVE TABLES
JOBS and TASKS
UNITY Catalog and Data Governance
STRUCTURED STREAMING with MULTI HOP

When building a DLT s pipeline you have two options to create a live tables, what is the main difference between `CREATE STREAMING LIVE TABLE` vs `CREATE LIVE TABLE`?
`CREATE STREAMING LIVE` **table is used in MULTI HOP Architecture**
`CREATE LIVE TABLE` **is used when working with Streaming data sources and Incremental data**
Correct answer
`CREATE STREAMING LIVE TABLE` **is used when working with Streaming data sources and Incremental data**
There is no difference both are the same, `CREATE STRAMING LIVE` **will be deprecated soon**
`CREATE LIVE TABLE` **is used in DELTA LIVE TABLES, CREATE STREAMING LIVE can only used in Structured Streaming applications**

A particular job seems to be performing slower and slower over time, the team thinks this started to happen when a recent production change was implemented, you were asked to take look at the job history and see if we can identify trends and root cause, where in the workspace UI can you perform this analysis?

Correct answer

Under jobs UI select the job you are interested, under runs we can see current active runs and last 60 days historical run

Under jobs UI select the job cluster, under spark UI select the application job logs, then you can access last 60 day historical runs

Under Workspace logs, select job logs and select the job you want to monitor to view the last 60 day historical runs

Under Compute UI, select Job cluster and select the job cluster to see last 60 day historical runs

Historical job runs can only be accessed by REST API

What are the different ways you can schedule a job in Databricks workspace?

Continuous, Incremental

On-Demand runs, File notification from Cloud object storage

Correct answer

Cron, On Demand runs

Cron, File notification from Cloud object storage

Once, Continuous

You have noticed that Databricks SQL queries are running slow, you are asked to look reason why queries are running slow and identify steps to improve the performance, when you looked at the issue you noticed all the queries are running in parallel and using a SQL endpoint(SQL Warehouse) with a single cluster. Which of the following steps can be taken to improve the performance/response times of the queries?

*Please note Databricks recently renamed SQL endpoint to SQL warehouse.

They can turn on the Serverless feature for the SQL endpoint(SQL warehouse).

Correct answer

They can increase the maximum bound of the SQL endpoint(SQL warehouse)'s scaling range

They can increase the warehouse size from 2X-Smal to 4XLarge of the SQL endpoint(SQL warehouse).

They can turn on the Auto Stop feature for the SQL endpoint(SQL warehouse).

They can turn on the Serverless feature for the SQL endpoint(SQL warehouse) and change the Spot Instance Policy to "Reliability Optimized."

You currently working with the marketing team to setup a dashboard for ad campaign analysis, since the team is not sure how often the dashboard should be refreshed they have decided to do a manual refresh on an as needed basis. Which of the following steps can be taken to reduce the overall cost of the compute when the team is not using the compute?

*Please note that Databricks recently change the name of SQL Endpoint to
SQL Warehouses.

They can turn on the Serverless feature for the SQL endpoint(SQL Warehouse).
They can decrease the maximum bound of the SQL endpoint(SQL Warehouse) scaling range.
They can decrease the cluster size of the SQL endpoint(SQL Warehouse).
Correct answer
They can turn on the Auto Stop feature for the SQL endpoint(SQL Warehouse).
They can turn on the Serverless feature for the SQL endpoint(SQL Warehouse) and change the Spot Instance Policy from "Reliability Optimized" to "Cost optimized"

You had worked with the Data analysts team to set up a SQL Endpoint(SQL warehouse) point so they can easily query and analyze data in the gold layer, but once they started consuming the SQL Endpoint(SQL warehouse) you noticed that during the peak hours as the number of users increase you are seeing queries taking longer to finish, which of the following steps can be taken to resolve the issue?

*Please note Databricks recently renamed SQL endpoint to SQL warehouse.

They can turn on the Serverless feature for the SQL endpoint(SQL warehouse).
Correct answer
They can increase the maximum bound of the SQL endpoint(SQL warehouse) 's scaling range.
They can increase the cluster size from 2X-Small to 4X-Large of the SQL endpoint(SQL warehouse) .
They can turn on the Auto Stop feature for the SQL endpoint(SQL warehouse) .
They can turn on the Serverless feature for the SQL endpoint(SQL warehouse) and change the Spot Instance Policy from "Cost optimized" to "Reliability Optimized."

The research team has put together a funnel analysis query to monitor the customer traffic on the e-commerce platform, the query takes about 30 mins to run on a small SQL endpoint cluster with max scaling set to 1 cluster. What steps can be taken to improve the performance of the query?
They can turn on the Serverless feature for the SQL endpoint.
They can increase the maximum bound of the SQL endpoint's scaling range anywhere from between 1 to 100 to review the performance and select the size that meets the required SLA.

Correct answer
They can increase the cluster size anywhere from X small to 3XL to review the performance and select the size that meets the required SLA.
They can turn off the Auto Stop feature for the SQL endpoint to more than 30 mins.
They can turn on the Serverless feature for the SQL endpoint and change the Spot Instance Policy from “Cost optimized” to “Reliability Optimized.”

Unity catalog simplifies managing multiple workspaces, by storing and managing permissions and ACL at _______ level
Workspace
Correct answer
Account
Storage
Data pane
Control pane

Which of the following section in the UI can be used to manage permissions and grants to tables?
User Settings
Admin UI
Workspace admin settings
User access control lists
Correct answer
Data Explorer

Which of the following is not a privilege in the Unity catalog?
SELECT
MODIFY
Correct answer
DELETE
CREATE TABLE
EXECUTE

A team member is leaving the team and he/she is currently the owner of the few tables, instead of transfering the ownership to a user you have decided to transfer the ownership to a group so in the future anyone in the group can manage the permissions rather than a single individual, which of the following commands help you accomplish this?
Correct answer

```
ALTER TABLE table_name OWNER to 'group'
TRANSFER OWNER table_name to 'group'
GRANT OWNER table_name to 'group'
ALTER OWNER ON table_name to 'group'
GRANT OWNER On table_name to 'group'
```