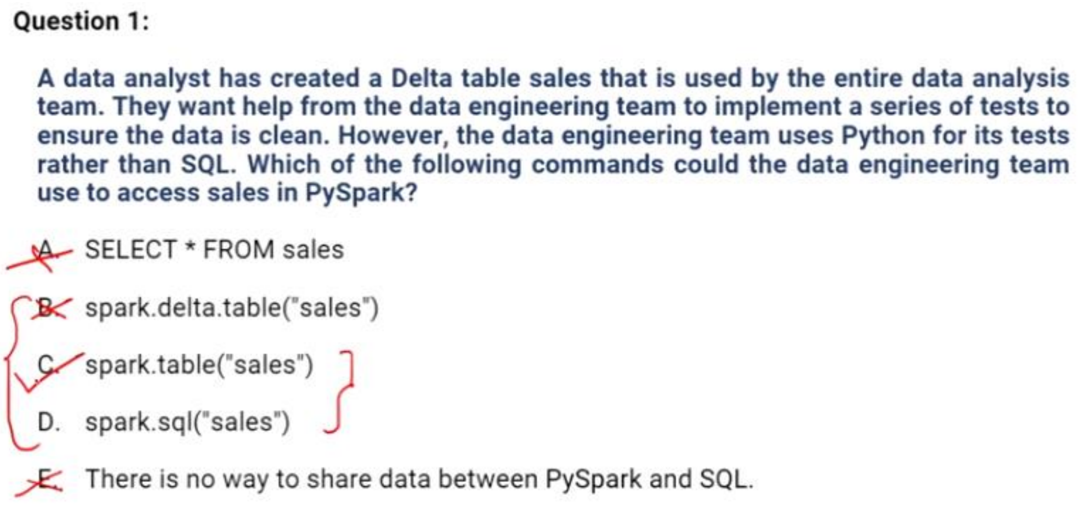
EXAM PREPARATION PART-2

Wednesday, April 24, 2024 12:11 PM

ELT with Spark Sql and Python - highest weightage

1. Select, insert, merge, delete, drop commands
2. Personally identifiable Information (PII) or Comments
3. User-defined functions (UDF)
4. Arrays
5. Jdbc
6. Set operations



Question 1:

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

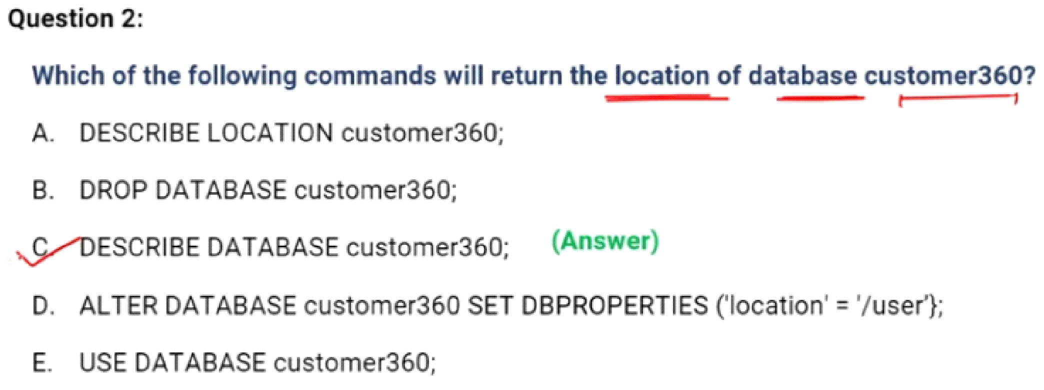
B. spark.delta.table("sales")

[✓ C.] spark.table("sales")

D. spark.sql("sales")

E. There is no way to share data between PySpark and SQL.

Syntax is spark.table(3 level name space)



Question 2:

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

A. DESCRIBE LOCATION customer360;

B. DROP DATABASE customer360;

[✓ C.] DESCRIBE DATABASE customer360; **(Answer)**

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

E. USE DATABASE customer360;

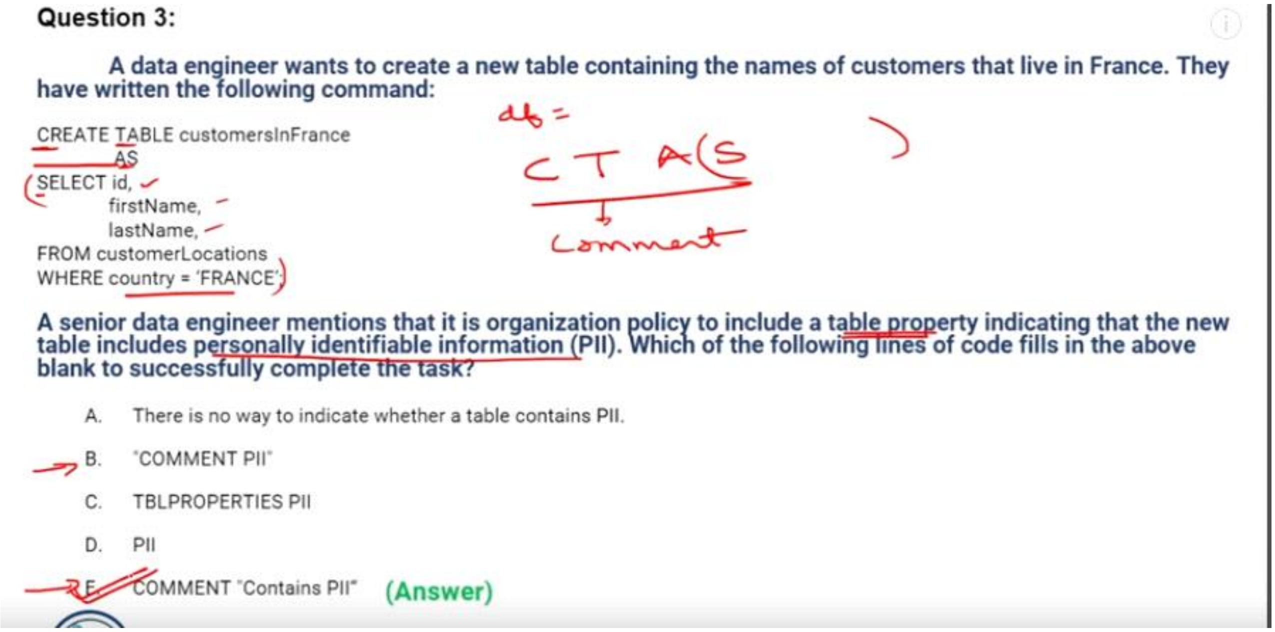
Describe extended table_name

Describe detail table_name

---- These two are used to get the details of tables including location

But if we want to get the details of database including location we have to use the below command

Describe database db_name



Question 3:

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';
```

df =

CTAS

comment

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.

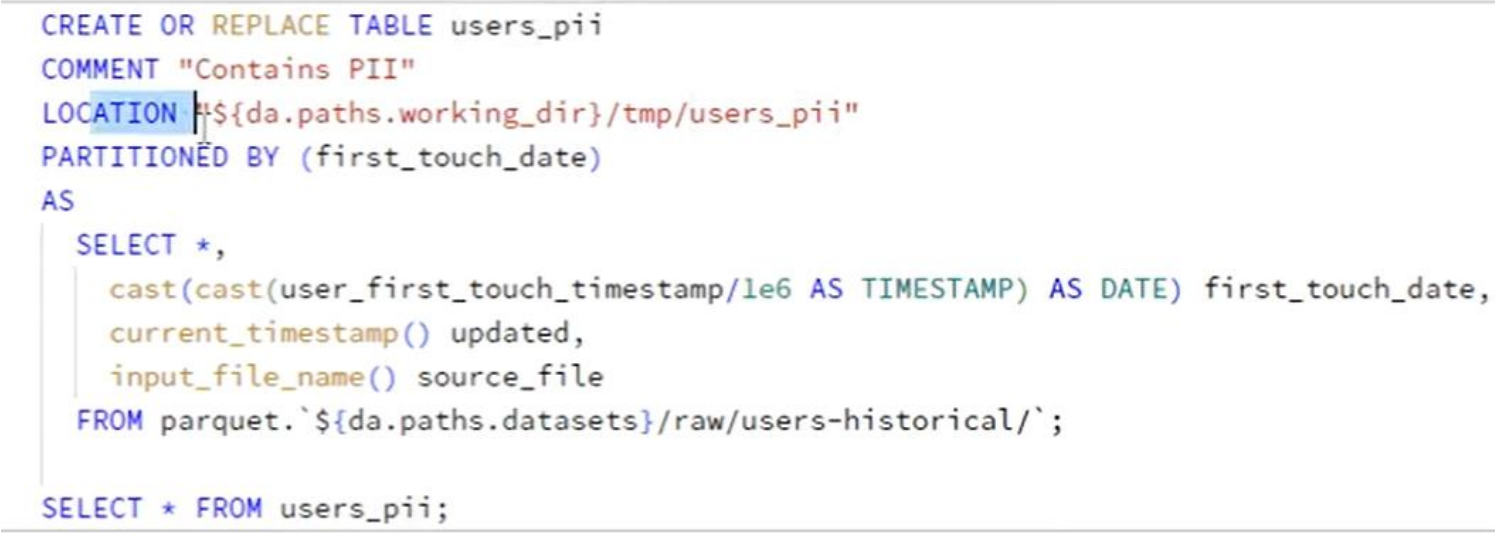
B. "COMMENT PII"

C. TBLPROPERTIES PII

D. PII

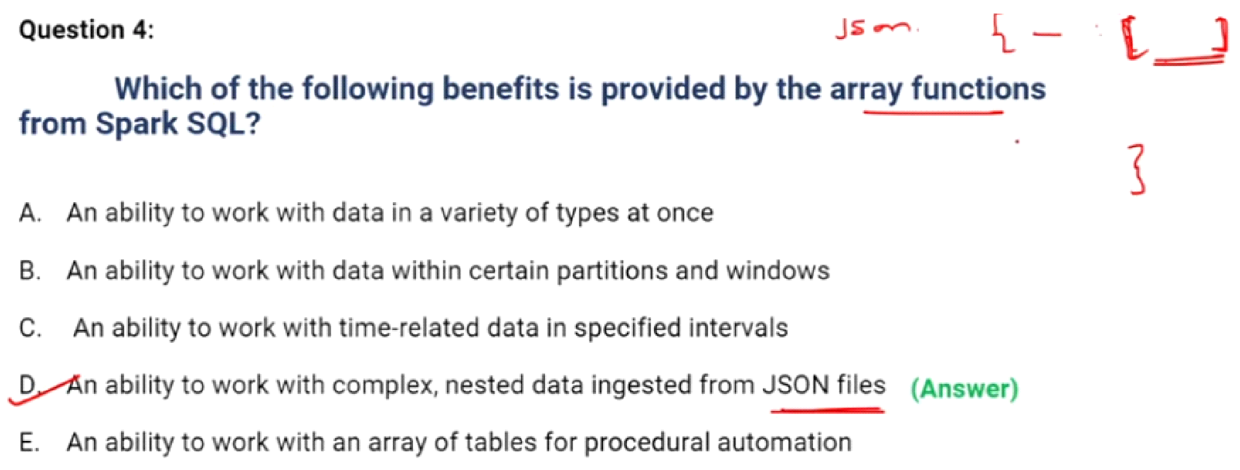
E. COMMENT "Contains PII" **(Answer)**

Comments in sql (comment is optional table property)



```
CREATE OR REPLACE TABLE users_pii
COMMENT "Contains PII"
LOCATION "${da.paths.working_dir}/tmp/users_pii"
PARTITIONED BY (first_touch_date)
AS
  SELECT *,
    cast(cast(user_first_touch_timestamp/1e6 AS TIMESTAMP) AS DATE) first_touch_date,
    current_timestamp() updated,
    input_file_name() source_file
  FROM parquet.`${da.paths.datasets}/raw/users-historical/`;

SELECT * FROM users_pii;
```



Question 4:

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

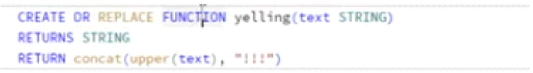
C. An ability to work with time-related data in specified intervals

D. An ability to work with complex, nested data ingested from JSON files (Answer)

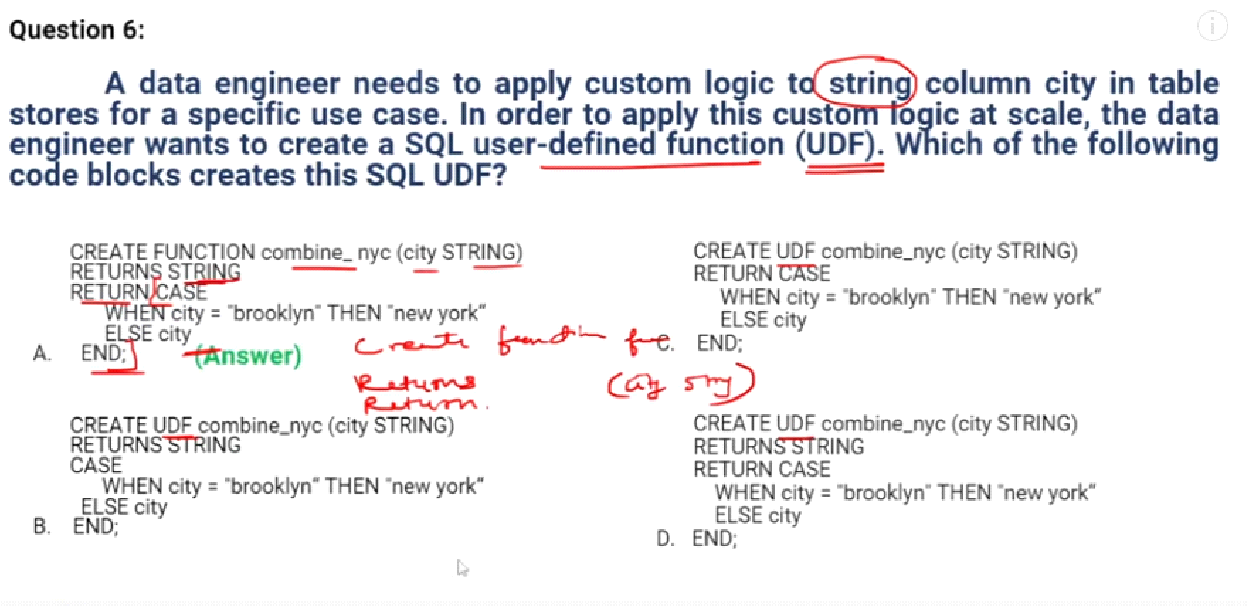
E. An ability to work with an array of tables for procedural automation

SQL UDF's

At minimum, a SQL UDF requires a function name, optional parameters, the type to be returned and custom logic.



```
CREATE OR REPLACE FUNCTION yelling(text STRING)
RETURNS STRING
RETURN concat(upper(text), "!!!")
```



Question 6:

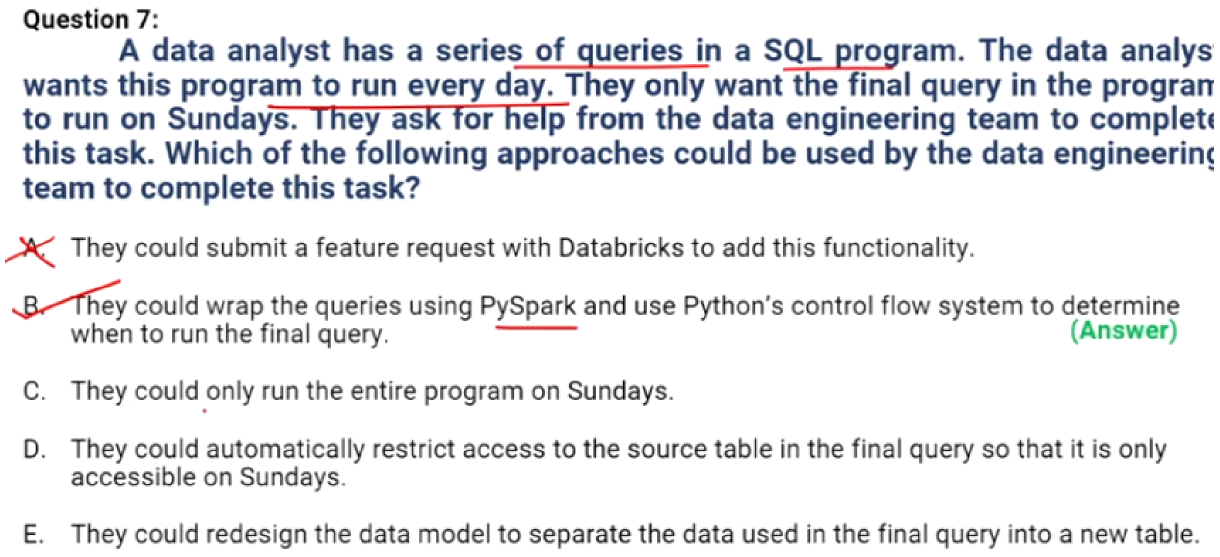
A data engineer needs to apply custom logic to string column city in table stores for a specific use case. In order to apply this custom logic at scale, the data engineer wants to create a SQL user-defined function (UDF). Which of the following code blocks creates this SQL UDF?

A.
```
CREATE FUNCTION combine_ nyc (city STRING)
RETURNS STRING
RETURN CASE
    WHEN city = "brooklyn" THEN "new york"
    ELSE city
END;
```
(Answer)

B.
```
CREATE UDF combine_nyc (city STRING)
RETURNS STRING
CASE
    WHEN city = "brooklyn" THEN "new york"
  ELSE city
END;
```

C.
```
CREATE UDF 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;
```



Question 7:

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

A. They could submit a feature request with Databricks to add this functionality.

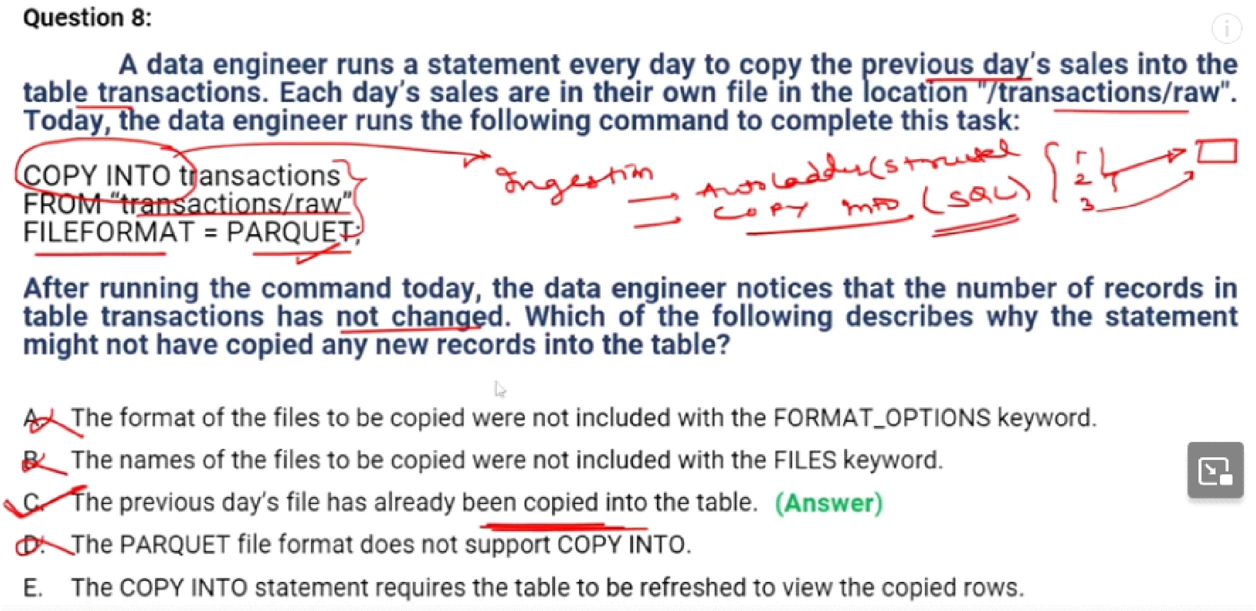
B. They could wrap the queries using PySpark and use Python's control flow system to determine when to run the final query. **(Answer)**

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.

Use the python's control flow



Question 8:

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;
```

Ingestion

=> Autoloader (structured

=> COPY INTO (SQL)

1
2
3

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?

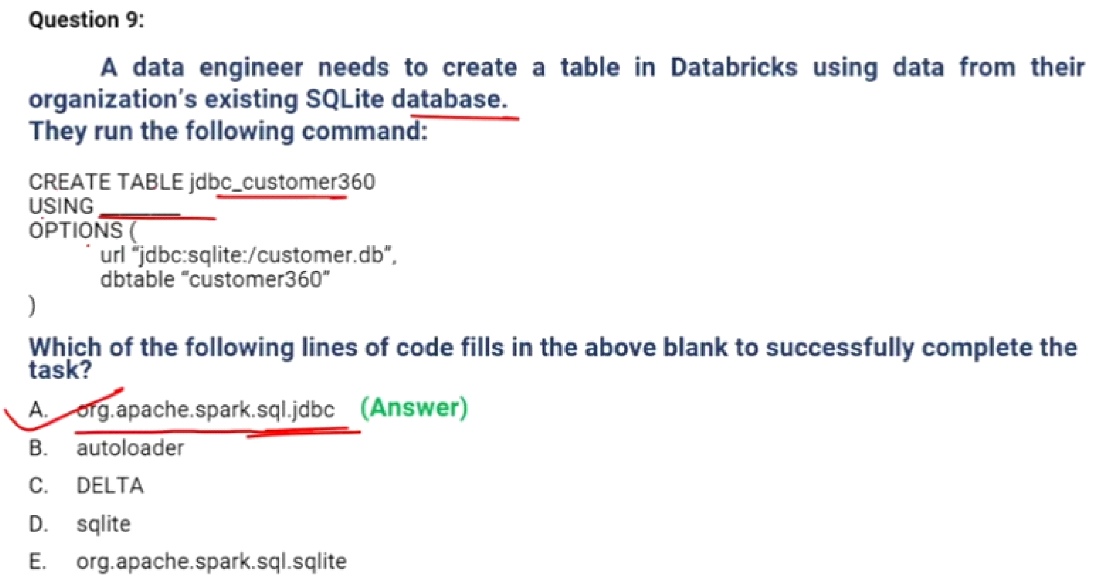
A. The format of the files to be copied were not included with the FORMAT_OPTIONS keyword.

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. **(Answer)**

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.



Question 9:

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:/customer.db",
        dbtable "customer360"
)
```

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

A. org.apache.spark.sql.jdbc **(Answer)**

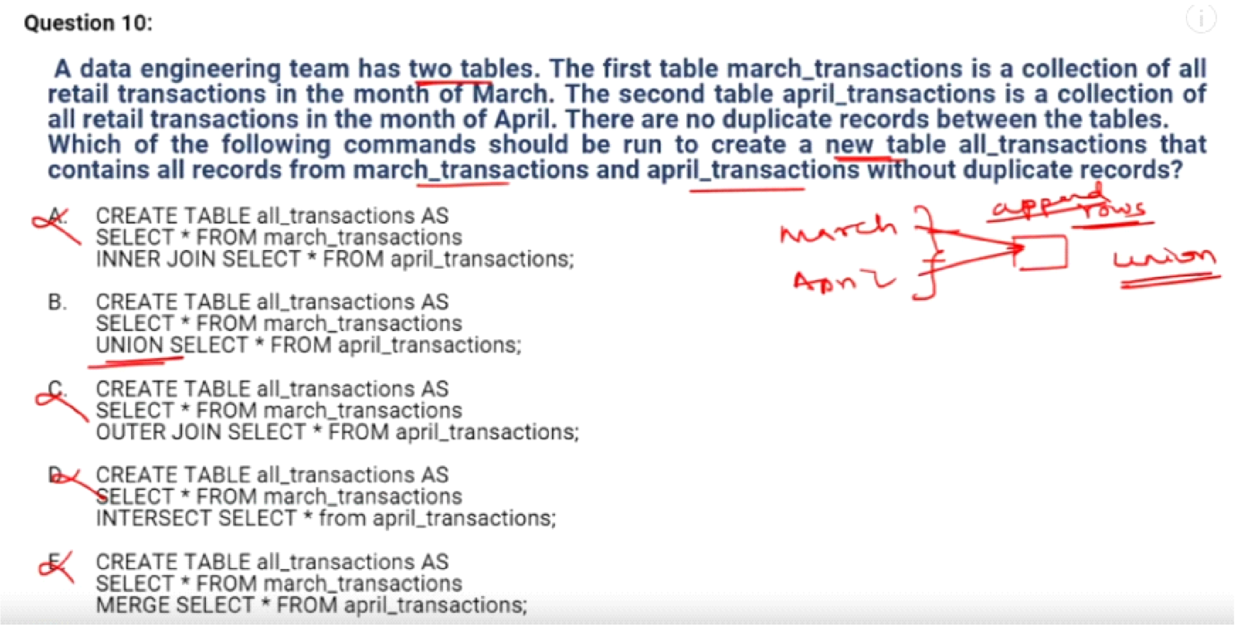
B. autoloader

C. DELTA

D. sqlite

E. org.apache.spark.sql.sqlite

Whenever you are using any sql database we have to use ---> Using jdbc



Question 10:

A data engineering team has two tables. The first table march_transactions is a collection of all retail transactions in the month of March. The second table april_transactions is a collection of all retail transactions in the month of April. There are no duplicate records between the tables. Which of the following commands should be run to create a new table all_transactions that contains all records from march_transactions and april_transactions without duplicate records?

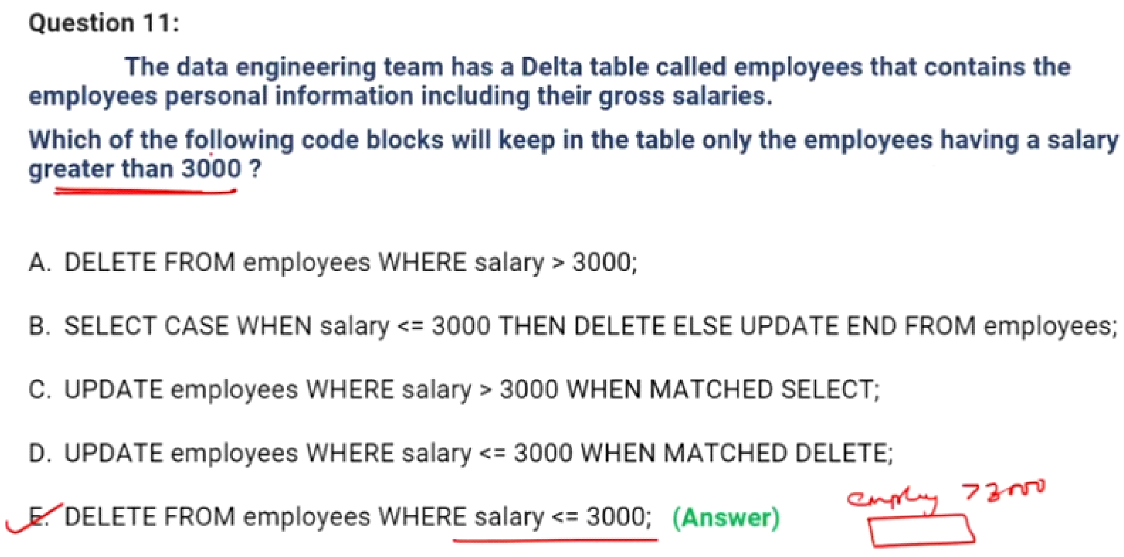
A. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
INNER JOIN SELECT * FROM april_transactions;

B. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
UNION SELECT * FROM april_transactions;

C. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
OUTER JOIN SELECT * FROM april_transactions;

D. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
INTERSECT SELECT * from april_transactions;

E. CREATE TABLE all_transactions AS
SELECT * FROM march_transactions
MERGE SELECT * FROM april_transactions;



Question 11:

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 ?

A. DELETE FROM employees WHERE salary > 3000;

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

C. UPDATE employees WHERE salary > 3000 WHEN MATCHED SELECT;

D. UPDATE employees WHERE salary <= 3000 WHEN MATCHED DELETE;

E. DELETE FROM employees WHERE salary <= 3000; **(Answer)**

Objects are three types

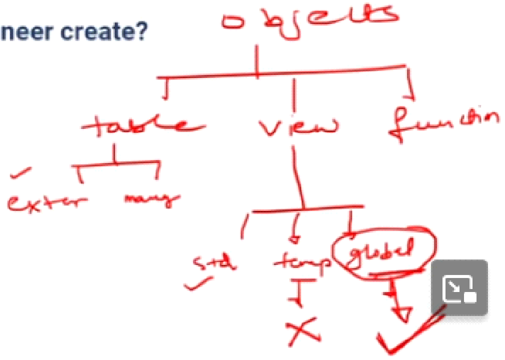
1. Table
   - External
   - Managed
2. View
   - Standard Persistent (Normal View)
   - Temp View
   - Global Temp View
3. Function

There are no stored procedures in databricks

Question 12:

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?



A. Temporary view

B. External table

C. Managed table

[✓ D.] Global Temporary view **(Answer)**

E. View

Question 13:

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

```
1.if process_mode = "init" and not is_table_exist:
2.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 ?

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

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

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

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

[✓ E.]
```
if process_mode == "init" and not is_table_exist:
print("Start processing ...")
```

Question 14:

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

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

[✓ A.] **org.apache.spark.sql.jdbc**
B. Postgresql
C. DELTA
D. Dbserver
E. cloudfiles

Question 15:

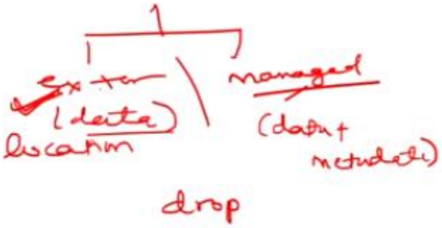
A data engineer only wants to execute the final block of a Python program if the Python variable day_of_week is equal to 1 and the Python variable review_period is True. Which of the following control flow statements should the data engineer use to begin this conditionally executed code block?

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 and review_period == "True":
✓ D. if day_of_week == 1 and review_period: **(Answer)**
E. if day_of_week = 1 & review_period: = "True":

Question 16:

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 **(Answer)**
D. The table did not have a location
E. The table was managed



Question 17: View , table

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)**