Task 4 – Unit Testing Data Pipelines

-- Task 4 Steps --

-- Please try to solve the task first without using this file as a guidance --

1) Use the same Databricks account used in the previous tasks (Task 2 and 3).

2) Two testing scripts were created in the Tests folder in our repository to test our data processing functions for cleaning, transforming and building.

3) It is not a must that your code/solution is the same as our scripts as long as it fulfills the main logic which is unit testing our data processing logic/data pipeline to ensure reliability and robustness.

4) pytest framework were used.

5) Connect your Github Repository to your Databricks account to access the testing scripts or you can do it any other way as long as Databricks can access the files when it runs pytest.

6) A runner script is found in PySpark folder in our repository called run_spark_unit_tests.py were it is simply used in a Databricks Notebook to for unit testing our pipeline.

7) Take care that the testing scripts imports the functions in pipeline script so this must be handled. (You will notice a __init_.py file is present in each folder to mark it as a python package)

8) Create a new cluster in your Notebook, run and attach it, and paste your test runner script code in a new cell then run it and the expected results should be two passed unit tests (No Failures As They Test That The Functions Are Working Correctly)

Important Note: Make sure to create a Cell first to install pytest using pip then restart the Python kernel, for example:

```
# To install pytest and restart Python kernel

%pip install -q pytest

dbutils.library.restartPython()
```

Thank You For Your Patience and Welcome Onboard !!