Task 2 – Solution Guidance

This document outlines two approaches to complete Task 2. Team members are encouraged to attempt the task independently before using this guide as a reference.

First Approach

1. Start by creating an account on the Databricks platform.
2. Your SQL script should be similar to the one located in the SQL folder of the repository, specifically the file named databricks_analysis.sql.
3. If you want to try the previously mentioned script, make sure to create a new schema in the Catalog workspace called j&j; and import the CSV files as new tables.
4. Create a new Notebook in your main Workspace. You may create subfolders if desired. Ensure that SQL is selected as the execution language, then copy and paste the SQL query into a notebook cell.
5. Run a cluster for your Notebook, then execute the code cell. The resulting table should be displayed.
6. Optional: You may create a Databricks job to schedule this task to run manually or according to a defined schedule.

Second Approach

1. To view the results of your SQL query without installing a database server or importing the three CSV files as tables, Spark SQL can be used.
2. To implement this approach, create two files similar to analysis.sql (located in the SQL folder) and run_spark_sql.py (located in the PySpark folder of the repository).
3. The analysis.sql script is largely the same as databricks_analysis.sql, with only minor differences. The run_spark_sql.py file contains simple, well-documented Python code. Questions are always welcome.
4. After placing both files on your local machine, open PowerShell or Command Prompt on Windows and run the spark-submit command followed by the name of your Python script, then the paths to your CSV data files and SQL script, respectively. The resulting table should be displayed within a few milliseconds.

Note: Ensure that PySpark is configured correctly, including all required dependencies such as Java, Spark, Python, and Hadoop. For setup assistance, refer to the PySpark_Windows_Setup_Guide PDF file.

Thank you for your patience, and welcome onboard!