Mentor Feedback Response

Dear Learner,

First, great initiative on working with Spark! However, I noticed that you're hardcoding logic into your Spark job. This makes it difficult to test, reuse, or extend in production. A better approach would be to modularize - break down your pipeline into functions like load_data(), clean_data(), and generate_reports(). This improves readability and collaboration.

Also, I hear your concern about SQL being outdated - but that's a common misconception. SQL is still heavily used across Spark SQL, BigQuery, Redshift, and Snowflake. It's declarative, concise, and widely known. For example, grouping customers by sales in SQL takes one line, while PySpark takes several.

Here's a quick 10-day learning plan:

- Day 1&2: Write modular Python functions and refactor your Spark job

- Day 3&4: Learn SQL syntax: SELECT, JOIN, GROUP BY

- Day 5&6: Practice with Spark SQL using your dataset

- Day 7&8: Build a mini ETL project using PySpark + SQL

- Day 9&10: Clean up and optimize your code

You're doing great - with just a few tweaks and SQL practice, you'll build world-class pipelines!

Best regards,
Dhanraj Tiwari