Banking Domain Capstone Project

Project Overview:

This project simulates banking transactions to analyze various performance and optimization techniques in Spark.

---

1. Spark Memory and Configuration Management

- Analyze memory consumption during data loading and transformations.
- Optimize memory configuration to balance execution and storage memory.

---

2. Performance Management in Spark

- Partition, cache, and persist datasets efficiently.
- Use broadcast variables and Spark UI for performance tuning.

---

3. Utilizing Spark 3.x Features

- Implement Adaptive Query Execution (AQE).
- Optimize queries with dynamic partition pruning and GPU acceleration.

---

4. Debugging and Troubleshooting

- Identify and resolve performance bottlenecks.
- Use Spark logs and Web UI for debugging.

Tasks to Perform

1. **Data Generation:** Run the provided Python script to generate 10,000 banking transactions.
2. **Memory Optimization:** Analyze and configure memory settings.
3. **Performance Tuning:** Implement partitioning, caching, and broadcast joins.
4. **Spark 3.x Techniques:** Apply AQE and GPU acceleration.
5. **Troubleshooting:** Debug performance issues using logs and UI.

Banking Domain Capstone Project

Submission Instructions

1. Save the final notebook as **banking_spark_project.ipynb**
2. Submit the file to the **Lumen** platform under the **Capstone Projects** section.
3. Include a **README** detailing the steps performed and findings.

Dataset to be used:

https://github.com/manojkumarsingh77/Fractal_PySpark3Levels/blob/16e294a13bd1112168d704ead3200e7fc3d23538/Data/Level3/CapstoneData/banking_transactions.csv