SAMSUNG

Big Data Course

Capstone Project Action Plan

For students (instructor's review required)

Together for Tomorrow!
Enabling People
Education for Future Generations

SAMSUNG

Course	**Big Data Course**
Team Name	Group 01
Team Leader/ Members	Đỗ Kiến Hưng, Nguyễn Văn Quang Duy
Project Title	Real-Time Big Data Processing with PySpark: NYC Taxi Trip Analysis

Goal

- Thiết kế và triển khai một pipeline xử lý dữ liệu lớn **end-to-end**, có khả năng mở rộng, chịu lỗi và có thể bảo trì bằng Apache PySpark.
- Áp dụng thành thạo **kiến trúc Medallion (Bronze → Silver → Gold)** để đảm bảo chất lượng và khả năng quản lý dữ liệu.
- Tích hợp **Spark Structured Streaming** để xử lý dữ liệu gần thời gian thực, mô phỏng một hệ thống giám sát vận hành liên tục.
- Xây dựng một lớp dữ liệu **Gold Layer** chứa các bảng tổng hợp, sẵn sàng cho phân tích (analysis-ready), tập trung vào việc xác định các "điểm nóng" (hotspots) và các chỉ số hiệu suất kinh doanh (KPIs).

Abstract

Dự án này tập trung vào việc xây dựng một hệ thống xử lý dữ liệu lớn, kết hợp cả xử lý theo lô (batch) và theo luồng (streaming) để phân tích dữ liệu taxi tại NYC. Sử dụng PySpark làm framework chính, chúng tôi sẽ triển khai một pipeline ETL hoàn chỉnh theo **kiến trúc Medallion**. Dữ liệu thô từ nguồn sẽ được nạp vào lớp Bronze, sau đó được làm sạch và làm giàu tại lớp Silver, và cuối cùng được tổng hợp thành các insight kinh doanh tại lớp Gold. Một thành phần xử lý luồng sẽ bổ sung cho pipeline, cho phép hệ thống cập nhật các phân tích với dữ liệu mới nhất. Sản phẩm cốt lõi của dự án là một pipeline dữ liệu mạnh mẽ, có cấu trúc và hiệu quả, thể hiện năng lực xây dựng các giải pháp data engineering từ đầu đến cuối.

Method

Dự án được cấu trúc theo 5 giai đoạn chính, tuân thủ kiến trúc Medallion và vòng đời của một dự án dữ liệu:

1. **Giai đoạn 1: Ingestion & Lớp Bronze (Dữ liệu Thô):** Sử dụng spark.read để nạp dữ liệu thô, hợp nhất schema, bổ sung metadata và lưu trữ dưới định dạng Delta Lake đã được phân vùng.
2. **Giai đoạn 2: Cleansing & Lớp Silver (Dữ liệu Sạch):** Áp dụng các quy tắc xác thực, làm sạch và thực hiện kỹ thuật đặc trưng (feature engineering) để tạo ra một nguồn dữ liệu đáng tin cậy.
3. **Giai đoạn 3: Aggregation & Lớp Gold (Dữ liệu Tổng hợp):** Sử dụng các phép toán groupBy và agg để xây dựng các bảng dữ liệu tổng hợp theo các chiều phân tích nghiệp vụ.
4. **Giai đoạn 4: Real-time Layer (Xử lý Luồng):** Sử dụng Spark Structured Streaming để đọc dữ liệu mô phỏng, thực hiện các phép tổng hợp trên cửa sổ thời gian (windowing) và ghi kết quả.
5. **Giai đoạn 5: Presentation Layer (Trình bày & Phân tích):** Sử dụng Jupyter Notebook để kết nối vào lớp Gold, thực hiện các truy vấn cuối cùng và trực quan hóa các insight bằng Matplotlib/Plotly.

Data

- **Nguồn chính (Batch):** Dữ liệu Yellow Taxi Trip Records (định dạng Parquet) từ website của NYC TLC. Dự án sẽ tập trung xử lý dữ liệu **tháng 1 năm 2023** (~3 triệu bản ghi) cho phần phân tích và dashboard.
- **Nguồn mô phỏng (Streaming):** Sử dụng dữ liệu **tháng 1 năm 2024** (~3 triệu bản ghi), được chia thành các file nhỏ để mô phỏng một luồng dữ liệu đầu vào liên tục cho pipeline streaming.

Expected Outcome

1. **Một ứng dụng PySpark có thể thực thi (.py script):** Bao gồm các module cho từng lớp và một script chính (main_pipeline.py) để điều phối toàn bộ pipeline.
2. **Cấu trúc thư mục dữ liệu:** Các thư mục Bronze, Silver, Gold được tổ chức rõ ràng trên hệ thống file, chứa dữ liệu Delta Lake đã được xử lý và phân vùng.
3. **Bảng dữ liệu "Gold":** Ít nhất 4 bảng dữ liệu tổng hợp (hourly, location, payment, vendor) đã được tối ưu hóa cho truy vấn.
4. **Một ứng dụng Spark Streaming:** Có khả năng chạy độc lập để xử lý dữ liệu mới.
5. **Một Jupyter Notebook báo cáo:** Trình bày các bước phân tích và trực quan hóa các insight chính.
6. **Tài liệu kỹ thuật:** Bao gồm README.md chi tiết và báo cáo cuối kỳ.

Role by Member

- **Đỗ Kiến Hưng (Lead Engineer / Backend Data Processor):** Chịu trách nhiệm thiết kế và triển khai các lớp xử lý dữ liệu phức tạp (Silver, Gold, Streaming) và framework Data Quality.
- **Nguyễn Văn Quang Duy (Pipeline Orchestrator / Analytics Engineer):** Chịu trách nhiệm nạp dữ liệu đầu vào (Bronze), điều phối pipeline (main), xuất dữ liệu và thực hiện toàn bộ khâu phân tích, trực quan hóa.

Schedule Summary

Schedule Summary (Tóm tắt Lịch trình):

Dự án được thực hiện trong 8 tuần, chia thành các giai đoạn chính như sau:

- Tuần 1-2: Setup & Data Exploration:
  - Thiết lập môi trường, tải dữ liệu, khám phá sơ bộ.
  - Hoàn thiện thiết kế kiến trúc chi tiết và kế hoạch hành động (WBS).
- Tuần 3-4: Bronze & Silver Layer Development:
  - Triển khai mã nguồn cho lớp Bronze và Silver.
  - Xây dựng framework kiểm tra chất lượng dữ liệu.
- Tuần 5-6: Gold Layer & Analytics Development:
  - Triển khai mã nguồn cho lớp Gold.
  - Xây dựng notebook phân tích và các biểu đồ trực quan hóa ban đầu.
- Tuần 7-8: Streaming, Integration & Finalization:
  - Triển khai pipeline streaming.
  - Tích hợp toàn bộ hệ thống, kiểm thử end-to-end.
  - Hoàn thiện báo cáo, tài liệu và chuẩn bị bài thuyết trình cuối kỳ.

(Chi tiết hơn sẽ được trình bày trong file Work Breakdown Structure - WBS)

Comment & Assessment

‹Comment and assessment by the instructor.›