Big Data Course

Capstone Project   
Final Report

For students (instructor review required)

ⓒ2023 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung Innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Real-Time Big Data Processing with PySpark: NYC Taxi Trip Analysis |

< Date (DD/MM/YY) >

GROUP 01

Đỗ Kiến Hưng

Nguyễn Văn Quang Duy

Nguyễn Văn Hiền

Vũ Đặng Tuấn Anh

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

Trong kỷ nguyên số, dữ liệu được tạo ra với tốc độ chưa từng có, đặt ra những thách thức lớn về khả năng lưu trữ, xử lý và khai thác thông tin. Ngành giao thông vận tải đô thị, đặc biệt tại các siêu đô thị như New York City (NYC), là một trong những nguồn tạo ra dữ liệu khổng lồ và liên tục. Mỗi ngày, hàng triệu chuyến đi taxi được ghi nhận, tạo thành một kho dữ liệu vô giá chứa đựng các thông tin chi tiết về hành vi di chuyển, các quy luật về cung-cầu, và hiệu quả hoạt động của hệ thống giao thông.

Ủy ban Taxi và Limousine (TLC) của NYC công khai bộ dữ liệu này, với kích thước lên đến hàng Terabyte và được cập nhật hàng tháng. Tuy nhiên, việc xử lý một khối lượng dữ liệu lớn như vậy bằng các công cụ truyền thống là bất khả thi. Điều này đòi hỏi phải có một nền tảng xử lý dữ liệu phân tán, mạnh mẽ, có khả năng mở rộng và xử lý hiệu quả cả dữ liệu lịch sử (batch) và dữ liệu thời gian thực (real-time). Apache Spark, với API Python là PySpark, đã nổi lên như một tiêu chuẩn công nghiệp cho các tác vụ này nhờ vào tốc độ xử lý vượt trội, tính linh hoạt và hệ sinh thái phong phú.

Dự án này ra đời trong bối cảnh đó, nhằm mục đích áp dụng các kỹ thuật xử lý dữ liệu lớn hiện đại để giải quyết một bài toán thực tế, chứng minh năng lực xây dựng một hệ thống từ đầu đến cuối, từ dữ liệu thô đến các insight có giá trị.

1.2. Motivation and Objective

Động lực chính của nhóm bắt nguồn từ mong muốn được áp dụng các kiến thức lý thuyết về Big Data đã học vào một dự án thực tế, có quy mô và độ phức tạp cao. Thay vì chỉ thực hiện các phân tích đơn lẻ, chúng tôi muốn xây dựng một pipeline dữ liệu (data pipeline) hoàn chỉnh, một sản phẩm cốt lõi trong công việc của một kỹ sư dữ liệu. Việc lựa chọn dữ liệu taxi NYC không chỉ vì quy mô lớn của nó, mà còn vì tiềm năng khai thác các insight hữu ích cho việc quy hoạch đô thị, tối ưu hóa dịch vụ và hỗ trợ các quyết định kinh doanh dựa trên dữ liệu.

Mục tiêu bao trùm của dự án là thiết kế và triển khai một pipeline xử lý dữ liệu lớn có khả năng mở rộng và chịu lỗi, sử dụng Apache PySpark với kiến trúc Medallion để phân tích dữ liệu chuyến đi taxi tại NYC.

Các mục tiêu cụ thể bao gồm:

* Xây dựng Pipeline theo kiến trúc Medallion (Bronze → Silver → Gold):

Bronze Layer: Nạp dữ liệu thô (raw data) từ nguồn mà không thay đổi.

Silver Layer: Làm sạch, xác thực, chuẩn hóa và làm giàu dữ liệu.

Gold Layer: Tổng hợp dữ liệu thành các bảng phân tích (aggregated tables) sẵn sàng cho việc báo cáo và khai thác.

* Tích hợp xử lý luồng nhằm triển khai một thành phần xử lý dữ liệu gần thời gian thực (near real-time) bằng Spark Structured Streaming để mô phỏng khả năng cập nhật của hệ thống.
* Đảm bảo chất lượng dữ liệu bằng cách xây dựng một framework kiểm tra chất lượng dữ liệu tự động qua các lớp của pipeline.
* Khai thác Insight từ lớp dữ liệu Gold, thực hiện các phân tích để xác định các chỉ số hiệu suất chính (KPIs) như các "điểm nóng" (hotspots) có nhu cầu cao, các xu hướng về doanh thu và hiệu quả hoạt động.
* Đóng gói ứng dụng để có thể chạy độc lập và trình bày kết quả một cách trực quan, dễ hiểu.

1.3. Members and Role Assignments

Nhóm chúng tôi bao gồm 4 thành viên, với sự phân công vai trò thực hiện dự án:

* Đỗ Kiến Hưng (Team Leader / Data Engineer):

Quản lý chung dự án, chịu trách nhiệm về thiết kế kiến trúc tổng thể của pipeline Medallion. Đảm bảo tiến độ, điều phối công việc và là đầu mối liên lạc chính. Thiết kế cấu trúc các lớp Bronze, Silver, Gold và phát triển script main\_pipeline.py.

* Vũ Đặng Tuấn Anh (Data Pipeline Engineer):

Tập trung phát triển các pipeline xử lý theo lô (batch processing). Viết mã cho các script bronze\_layer.py và silver\_layer.py. Tối ưu hóa hiệu năng của các tác vụ Spark (partitioning, caching) và viết các bài kiểm thử đơn vị (unit tests) cho pipeline batch.

* Nguyễn Văn Quang Duy (Streaming Engineer):

Chịu trách nhiệm thiết kế và triển khai pipeline xử lý dữ liệu thời gian thực (streaming\_pipeline.py). Xử lý các khái niệm của Spark Streaming như windowing, watermarking, và checkpointing. Thiết lập môi trường mô phỏng luồng dữ liệu.

* Nguyễn Văn Hiền (Analytics Engineer):

Chịu trách nhiệm về chất lượng dữ liệu, phân tích và trực quan hóa. Phát triển gold\_layer.py và framework data\_quality.py. Xây dựng các notebook phân tích (03\_analytics\_dashboard.ipynb), tạo báo cáo chất lượng dữ liệu và trực quan hóa các insight cuối cùng.

1.4. Schedule and Milestones

Dự án được thực hiện trong 8 tuần, với các cột mốc quan trọng được xác định:

* Tuần 1-2: Setup & Data Exploration (Thiết lập và khám phá dữ liệu)

Hoàn thành việc thiết lập môi trường lập trình (Python, PySpark, Jupyter). Tải về và khám phá sơ bộ dữ liệu taxi NYC. Hoàn thành thiết kế kiến trúc Medallion và phân công vai trò chi tiết.

* Tuần 3-4: Bronze & Silver Layer Development (Phát triển lớp Bronze & Silver)

Hoàn thành script cho pipeline Bronze (data ingestion) và pipeline Silver (data cleaning & feature engineering). Hoàn thành framework kiểm tra chất lượng dữ liệu cơ bản.

* Tuần 5-6: Gold Layer & Analytics (Phát triển lớp Gold & phân tích)

Hoàn thành script cho pipeline Gold (data aggregation). Xây dựng notebook phân tích đầu tiên để trực quan hóa dữ liệu từ lớp Gold, xác thực các logic nghiệp vụ.

* Tuần 7-8: Streaming & Integration (Phát triển luồng & tích hợp)

Hoàn thành pipeline xử lý luồng. Tích hợp tất cả các thành phần vào script main\_pipeline.py chính. Chuẩn bị tài liệu kỹ thuật, báo cáo cuối kỳ và bài thuyết trình. Hoàn thành live demo.

2. Project Execution

2.1. Simulated Scenario Description

2.2. Datasets Selection and Description

2.3. Data Ingestion Pipeline

2.4. Data Transformation Processing

2.5. Data Query and Insight

3. Results

3.1. Data Ingestion Scripts and Code

3.2. Data Transformation Scripts and Code

3.3. Description and Sample of Transformed Datasets

3.4. Data Visualization of Query Results

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

|  |
| --- |
| <ATTACH A TEAM PICTURE HERE> |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |