Big Data Course

Capstone Project   
Action Plan

For students (instructor’s 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.

|  |  |
| --- | --- |
| Course | Big Data Course |
| Team Name | Group 01 |
| Team Leader/  Members | Kiến Hưng, Tuấn Anh, Văn Hiền, 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 có khả năng mở rộng và chịu lỗi (scalable & fault-tolerant), sử dụng Apache PySpark.  Kiến trúc Medallion (Bronze -> Silver -> Gold)  Spark Structured Streaming  Gold Layer, Analysis-ready, hotspots. | |
| Abstract |  |
| Dự án này xây dựng một hệ thống xử lý dữ liệu lớn, xử lý theo lô (batch) và theo luồng (streaming). Sử dụng tập dữ liệu công khai về các chuyến đi taxi ở New York làm bối cảnh, chúng tôi sẽ xây dựng một pipeline ETL hoàn chỉnh. Pipeline sẽ được thiết kế theo kiến trúc Medallion. 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. | |
| Method |  |
| Giai đoạn 1: Ingestion & Lớp Bronze (Dữ liệu Thô)  Giai đoạn 2: Cleansing & Lớp Silver (Dữ liệu Sạch)  Giai đoạn 3: Aggregation & Lớp Gold (Dữ liệu Tổng hợp cho Nghiệp vụ)  Giai đoạn 4: Real-time Layer (Xử lý Luồng)  Giai đoạn 5: Presentation Layer (Trình bày & Phân tích) | |
| Data |  |
| Nguồn chính (Batch): Dữ liệu Yellow Taxi Trip Records (định dạng Parquet) từ website của NYC TLC.  Nguồn mô phỏng (Streaming): Một tập con của dữ liệu lịch sử sẽ được chia thành các file nhỏ. | |
| Expected  Outcome |  |
| Một ứng dụng PySpark có thể thực thi (.py script)  Cấu trúc thư mục dữ liệu  Bảng dữ liệu "Gold"  Một ứng dụng Spark Streaming  Một Jupyter Notebook báo cáo  Tài liệu kỹ thuật | |
| Role by  Member |  |
| Lead Data Engineer / Project Manager  Data Pipeline Engineer (Batch)  Streaming Data Engineer  Data Quality & Analytics Engineer | |

|  |  |
| --- | --- |
| Schedule  Summary |  |
| <Provide a summarized schedule.>  <Use the provided WBS workbook for more detailed action plan.> | |
| Comment &  Assessment |  |
| <Comment and assessment **by the instructor.**> | |