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.

|  |  |
| --- | --- |
|  | **Big Data Course** |
| **Team Name** | DHT |
| **Team Leader/**  **Members** | * Trần Văn Tiến (Team leader) * Trương Thị Thuỳ Dung * Đặng Xuân Huyền |
| **Project Title** | **Airline Dataset Analysis using Hadoop, Hive, Spark and Tableau** |
| **Goal** |  |
| Dự án này là trình bày việc triển khai xử lý dữ liệu hãng hàng không bằng cách sử dụng các công nghệ nguồn mở như Hadoop, Hive, Spark, HDFS và Tableau. Dự án nhằm xử lý và phân tích khối lượng lớn dữ liệu liên quan đến hãng hàng không một cách hiệu quả và hiệu suất cao. Hadoop và HDFS cung cấp một khung lưu trữ và xử lý phân tán, cho phép xử lý các bộ dữ liệu lớn. Hive và Spark phục vụ như các ngôn ngữ truy vấn và công cụ xử lý dữ liệu đơn giản hóa các tác vụ thao tác dữ liệu. Ngoài ra, Tableau cung cấp các giải pháp trực quan hóa dữ liệu để trình bày kết quả phân tích một cách trực quan và dễ hiểu. Bằng cách sử dụng các công nghệ này, dự án nhằm trình bày cách các hãng hàng không có thể tận dụng các giải pháp dữ liệu lớn để trích xuất các thông tin có giá trị, tối ưu hóa hoạt động và cải thiện hiệu suất tổng thể trong ngành. | |
| **Abstract** |  |
| Dự án này nhằm khai thác tiềm năng của phân tích dữ liệu lớn trong ngành hàng không bằng cách sử dụng các công nghệ mã nguồn mở như Hadoop, Spark, Hive và Tableau. Với sự gia tăng liên tục của khối lượng dữ liệu từ các hoạt động hàng không, việc áp dụng các công cụ phân tích dữ liệu hiện đại là cần thiết để giúp các hãng hàng không đưa ra quyết định thông minh hơn, cải thiện hiệu suất và nâng cao trải nghiệm khách hàng. Bằng cách triển khai một quy trình ETL (Extract, Transform, Load) với Spark, dữ liệu từ các nguồn khác nhau sẽ được xử lý và chuyển đổi để tạo ra những thông tin có giá trị, sau đó được lưu trữ trong Hive và trực quan hóa với Tableau. Dự án này không chỉ tập trung vào việc xử lý dữ liệu hiệu quả mà còn vào việc hiểu rõ và tối ưu hóa quy trình quản lý dữ liệu trong môi trường hàng không. | |
| **Method** |  |
| * Thu thập dữ liệu từ hệ thống đặt vé, lịch trình bay, dữ liệu hành khách và hoạt động. Lưu trữ dữ liệu trong HDFS. * Xây dựng mô hình dữ liệu theo star schema bao gồm 1 bảng chính flight\_facts và 4 Dim * Nghiên cứu Hadoop, Hive, Spark và Tableau để hiểu cách mỗi công cụ hỗ trợ xử lý và phân tích dữ liệu. * Nghiên cứu cách Hive quản lý dữ liệu, sự khác biệt giữa Hive-managed tables và Hive-external tables. * Nghiên cứu cách Spark xử lý dữ liệu song song và phân tán. * Sử dụng các lệnh Unix cơ bản để quản lý hệ thống tệp và thao tác dữ liệu trên HDFS. * Tiền Xử Lý Dữ Liệu: Làm sạch và tiền xử lý dữ liệu với Spark. * ETL Dữ Liệu: Thực hiện quy trình ETL bằng Spark, trích xuất dữ liệu từ Hive, chuyển đổi và lưu vào một database khác trong Hive. * Thực hiện các truy vấn SQL cơ bản để khám phá và hiểu dữ liệu trong Hive. * Sử dụng phân vùng và phân cụm để tối ưu hóa truy vấn và cải thiện hiệu suất lưu trữ. * Sử dụng Tableau để tạo các biểu đồ, báo cáo và dashboard trực quan. * Lập kế hoạch và triển khai các bước xử lý dữ liệu theo trình tự hợp lý. Tối ưu hóa pipeline dựa trên kết quả đánh giá. | |
|  | |

|  |  |
| --- | --- |
| **Data** |  |
| Dữ liệu hàng không sẽ được thu thập từ các nguồn dữ liệu công khai và từ các hãng hàng không.  Dữ liệu bao gồm 4 file csv:   * airports.csv: Tập dữ liệu này chứa thông tin về các sận bay * carriers.csv: Tập dữ liệu này chứa thông tin về các hãng hàng không * detailed\_data.csv: Tập dữ liệu này chứa thông tin chi tiết về các chuyến bay * plane-data.csv: Tập dữ liệu này chứa thông tin về các chuyến bay | |
| **Expected  Outcome** |  |
| * Hiểu tổng quan về data source * Hiểu data model * Hiểu tầm quan trọng của từng dịch vụ một cách chi tiết * Hiểu được kiến trúc và các loại dữ liệu của Hive * Hiểu được kiến trúc và các loại dữ liệu của Spark * Hiểu các lệnh cơ bản của Unix * Data processing với Spark * EDA cơ bản với Hive * Hive partitioning and clustering * Hiểu với cách ETL của Spark * Trực quan hóa dữ liệu với tableau * Điều phối công việc của pipeline | |
| **Role by  Member** |  |
| * Cả nhóm   + Tìm hiểu tổng quan về đề tài   + Tìm kiếm dữ liệu và xây dựng mô hình dữ liệu   + Xây dựng Action plan cơ bản * Thành viên 1: Trương Thị Thuỳ Dung   + Tìm hiểu về Hadoop   + Tìm kiếm dữ liệu và xây dựng mô hình dữ liệu   + Trực quan hoá dữ liệu   + Thực hiện power point báo cáo final report   + Thực hiện power point báo cáo action plan * Thành viên 2: Trần Văn Tiến   + Tìm hiểu về Spark   + Data processing với Spark   + Hiểu với cách ETL của Spark   + Trực quan hoá dữ liệu   + Thực hiện word báo cáo Action plan hoàn chỉnh * Thành viên 3: Đặng Xuân Huyền   + Tìm hiểu về Hive, Tableau * EDA cơ bản với Hive * Hive partitioning and clustering * Thực hiện word báo cáo final report | |
|  | |

|  |  |
| --- | --- |
| **Schedule  Summary** |  |
| **Ngày 1 - 5: Giai Đoạn Lập Kế Hoạch:**  **Ngày 1:**   * Giới thiệu tổng quan về dự án Capstone. * Lập đội nhóm và phân công nhiệm vụ.   **Ngày 2:**   * Xác định các yêu cầu và mục tiêu cụ thể của dự án. * Thảo luận và điền mẫu action plan sơ bộ.   **Ngày 3:**   * Thực hiện Design Thinking: Empathize, Define, Ideate. * Lập kế hoạch chi tiết cho từng thành viên.   **Ngày 4:**   * Thiết lập môi trường làm việc: cài đặt và cấu hình Hadoop, Hive, Spark. * Thực hành các lệnh cơ bản của Unix.   **Ngày 5:** Hoàn thiện action plan chi tiết.  **Ngày 6 - 10: Thu Thập và Hiểu Dữ Liệu**  **Ngày 6:**   * Thu thập dữ liệu từ các nguồn khác nhau: hệ thống đặt vé, lịch trình bay, dữ liệu hành khách và dữ liệu hoạt động. * Lưu trữ dữ liệu vào HDFS.   **Ngày 7:**   * Xây dựng và hiểu rõ mô hình dữ liệu (star schema). * Xác định các bảng, quan hệ giữa các bảng và các trường dữ liệu chính.   **Ngày 8:** Nghiên cứu và hiểu rõ về các dịch vụ và công cụ sẽ được sử dụng: Hadoop, Hive, Spark, Tableau.  **Ngày 9:** Nghiên cứu kiến trúc của Hive, bao gồm Hive-managed tables và Hive-external tables.  **Ngày 10:** Nghiên cứu kiến trúc của Spark và các loại dữ liệu Spark hỗ trợ.  **Ngày 11 - 15: Tiền Xử Lý và ETL Dữ Liệu**  **Ngày 11:**   * Sử dụng Spark để làm sạch và tiền xử lý dữ liệu. * Loại bỏ các giá trị thiếu, chuẩn hóa dữ liệu và chuyển đổi định dạng dữ liệu.   **Ngày 12:** Thực hiện các bước ETL: trích xuất dữ liệu thô từ Hive.  **Ngày 13:**   * Tải dữ liệu đã xử lý vào một cơ sở dữ liệu khác trong Hive. * Kiểm tra và đảm bảo dữ liệu đã được lưu trữ đúng cách.   **Ngày 14:** Sử dụng Hive để thực hiện các truy vấn SQL cơ bản nhằm khám phá và hiểu rõ dữ liệu (EDA cơ bản).  **Ngày 15:** Sử dụng các kỹ thuật phân vùng và phân cụm trong Hive để  tối ưu hóa truy vấn và cải thiện hiệu suất lưu trữ.  **Ngày 16 - 20: Trực Quan Hóa và Thực hiện báo cáo**  **Ngày 16:**Sử dụng Tableau để tạo các biểu đồ, báo cáo và dashboard  trực quan từ kết quả phân tích.  **Ngày 17:** Thực hiện điều phối các công việc của pipeline: lập kế hoạch và triển khai các bước xử lý dữ liệu theo trình tự hợp lý.  Ngày 18: Báo cáo dự án Capstone.  **Ngày 19:** Chuẩn bị bài thuyết trình cuối cùng và đánh giá lại toàn bộ quá trình thực hiện dự án  **Ngày 20:**Báo cáo dự án. | |
| **Comment &  Assessment** |  |
| <Comment and assessment **by the instructor.**> | |