**PHÂN CÔNG CÔNG VIỆC**

1. **Cài đặt hệ thống Hadoop**

* Phụ trách: *Lưu Quốc Thành*
* Mô tả: Triển khai và cấu hình hệ sinh thái Hadoop trên môi trường làm việc, bao gồm HDFS và các thành phần liên quan phục vụ cho việc xử lý dữ liệu lớn.

1. **Thu thập dữ liệu sử dụng Flume và lưu trữ trên HDFS**

* Phụ trách: *Vũ Bảo Long*
* Mô tả: Thiết lập Apache Flume để thu thập dữ liệu từ nguồn đầu vào và lưu trữ vào hệ thống HDFS nhằm phục vụ cho các bước xử lý tiếp theo.

1. **Xử lý chuỗi dữ liệu bằng Apache Spark và chuyển đổi sang định dạng Parquet/CSV**

* Phụ trách: *Lưu Quốc Thành*
* Mô tả: Sử dụng Apache Spark để làm sạch và xử lý dữ liệu dạng chuỗi, sau đó chuyển đổi sang định dạng Parquet và CSV để tối ưu hóa lưu trữ và phân tích.

1. **Nhập dữ liệu vào Hive và thực hiện các truy vấn phân tích**

* Phụ trách: *Trần Đình Hưng*
* Mô tả: Tạo bảng Hive từ dữ liệu đã xử lý và thực hiện các truy vấn HiveQL để phân tích và trích xuất thông tin hữu ích.

1. **Thực hiện xử lý dữ liệu sử dụng MapReduce**

* Phụ trách: *Vũ Bảo Long, Lưu Quốc Thành, Trần Đình Hưng*
* Mô tả: Xây dựng và thực thi các chương trình MapReduce nhằm khai thác và tổng hợp dữ liệu theo yêu cầu phân tích.

1. **Trực quan hóa dữ liệu bằng Python (Pandas) trên Jupyter Notebook**

* Phụ trách: *Vũ Bảo Long*
* Mô tả: Sử dụng thư viện Pandas và các công cụ trực quan hóa (như Matplotlib, Seaborn) để thể hiện dữ liệu dưới dạng biểu đồ và bảng biểu trên Jupyter Notebook.

**QUY TRÌNH THỰC HIỆN ĐỒ ÁN**

1. **Khởi tạo và cài đặt hệ thống**

* Cài đặt môi trường Hadoop bao gồm HDFS và các thành phần cần thiết phục vụ cho việc lưu trữ và xử lý dữ liệu lớn.
* Kiểm tra cấu hình và đảm bảo hệ thống hoạt động ổn định.

1. **Thu thập và đưa dữ liệu vào HDFS**

* Cấu hình Apache Flume để thu thập dữ liệu từ các nguồn đầu vào (file, log, API,...).
* Chuyển tiếp dữ liệu được thu thập vào hệ thống HDFS nhằm phục vụ cho bước xử lý dữ liệu.

1. **Tiền xử lý và chuyển đổi định dạng dữ liệu**

* Sử dụng Apache Spark để xử lý dữ liệu: làm sạch dữ liệu, phân tích chuỗi, định dạng lại dữ liệu.
* Xuất dữ liệu sau xử lý ra định dạng Parquet và CSV để tiện lợi cho việc lưu trữ, truy vấn và trực quan hóa.

1. **Tạo kho dữ liệu phân tích với Hive**

* Tạo các bảng Hive tương ứng với cấu trúc dữ liệu đã xử lý.
* Nhập dữ liệu từ HDFS vào Hive.
* Thực hiện các truy vấn HiveQL để phân tích dữ liệu.

1. **Xử lý dữ liệu sử dụng MapReduce**

* Viết các chương trình MapReduce để tổng hợp, thống kê hoặc trích xuất các đặc trưng dữ liệu theo yêu cầu cụ thể.
* Kết hợp với kết quả từ Hive để kiểm tra độ chính xác và hiệu quả xử lý.

1. **Phân tích và trực quan hóa dữ liệu**

* Sử dụng Python với thư viện Pandas, Matplotlib, Seaborn trên Jupyter Notebook để trực quan hóa dữ liệu.
* Trình bày các biểu đồ, bảng biểu thể hiện kết quả phân tích từ dữ liệu đã xử lý.

1. **Tổng kết và đánh giá kết quả**

* Đánh giá hiệu quả của các bước xử lý và công nghệ sử dụng.
* Rút ra kết luận, bài học kinh nghiệm và đề xuất hướng phát triển mở rộng trong tương lai.