viettel

TẬP ĐOÀN CÔNG NGHIỆP – VIỄN THÔNG QUÂN ĐỘI

BÁO CÁO ASSIGNMENT

TRẦN NGỌC BẢO
tnbao120603@gmail.com.

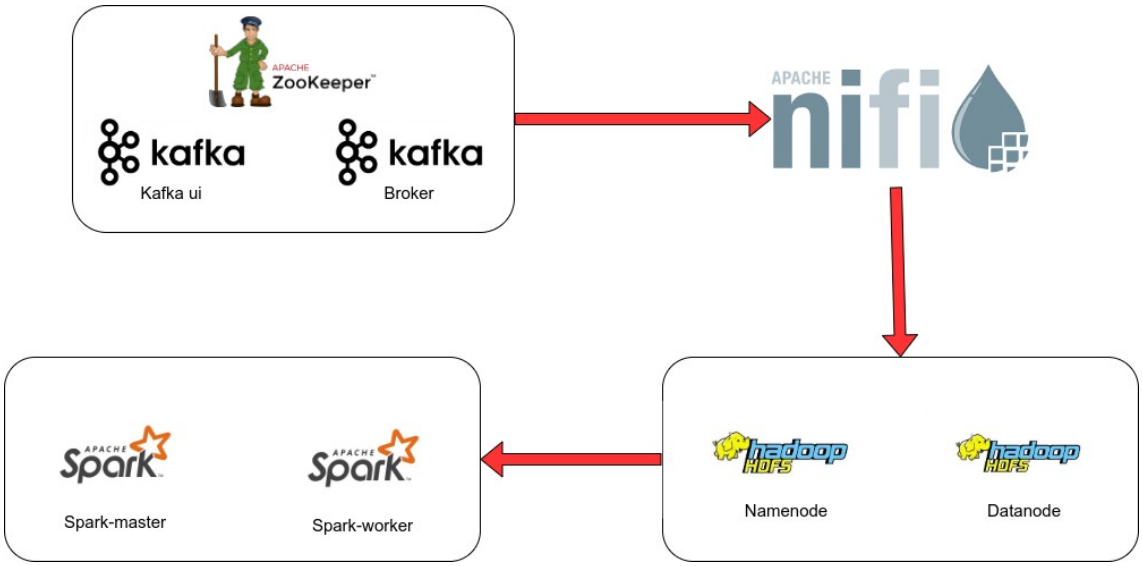
Chương trình Viettel Digital Talent 2024
Lĩnh vực: Data Engineering

HÀ NỘI, 06/2024

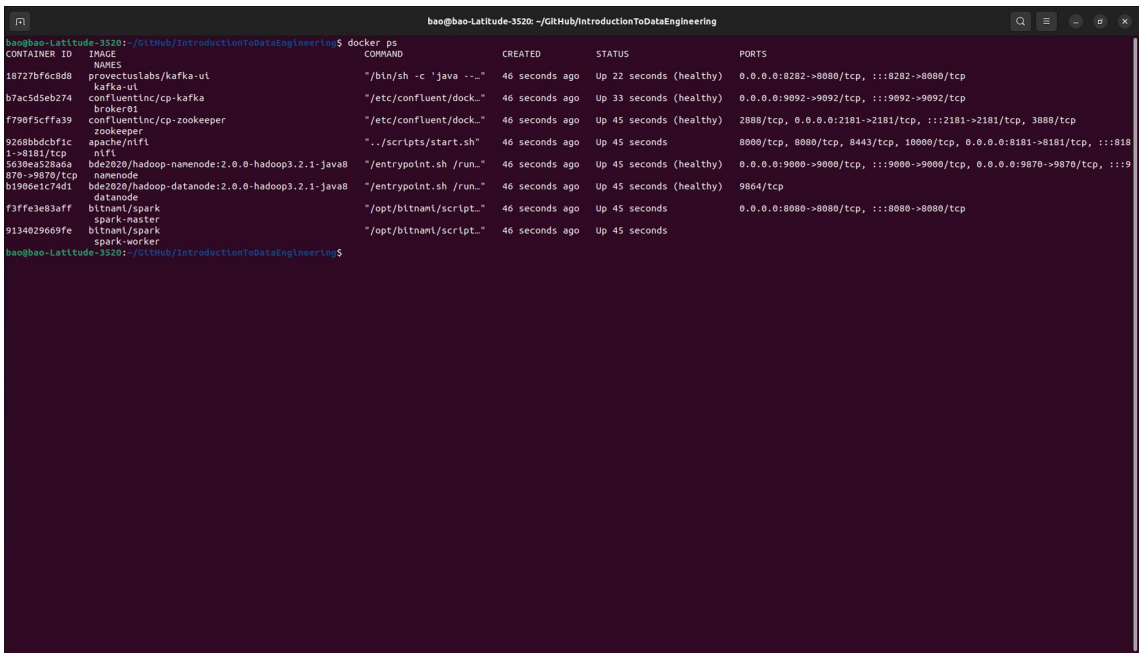
CHƯƠNG 1. Mô tả cách thức triển khai

Triển khai trên môi trường Docker, cụ thể từng phần như sau:

- Data Source: Triển khai một cụm Kafka gồm có 1 Zookeeper để quản lý cụm Kafka, 1 Broker là thành phần chính của cụm để lưu trữ dữ liệu, 1 Kafka-ui để dễ dàng tương tác với cụm trên nền tảng website.
- Data Ingestion: Triển khai một node Nifi để chuyển tiếp dữ liệu từ Kafka tới HDFS của Hadoop.
- Data Storeage: Triển khai một cụm Hadoop gồm có 1 Namenode và 1 DataNode để lưu trữ dữ liệu.
- Data Processing: Triển khai một cụm Spark gồm có 1 Spark-master và 1 Spark-worker để xử lý dữ liệu



Kết quả sau khi khởi động hệ thống bằng Docker compose:

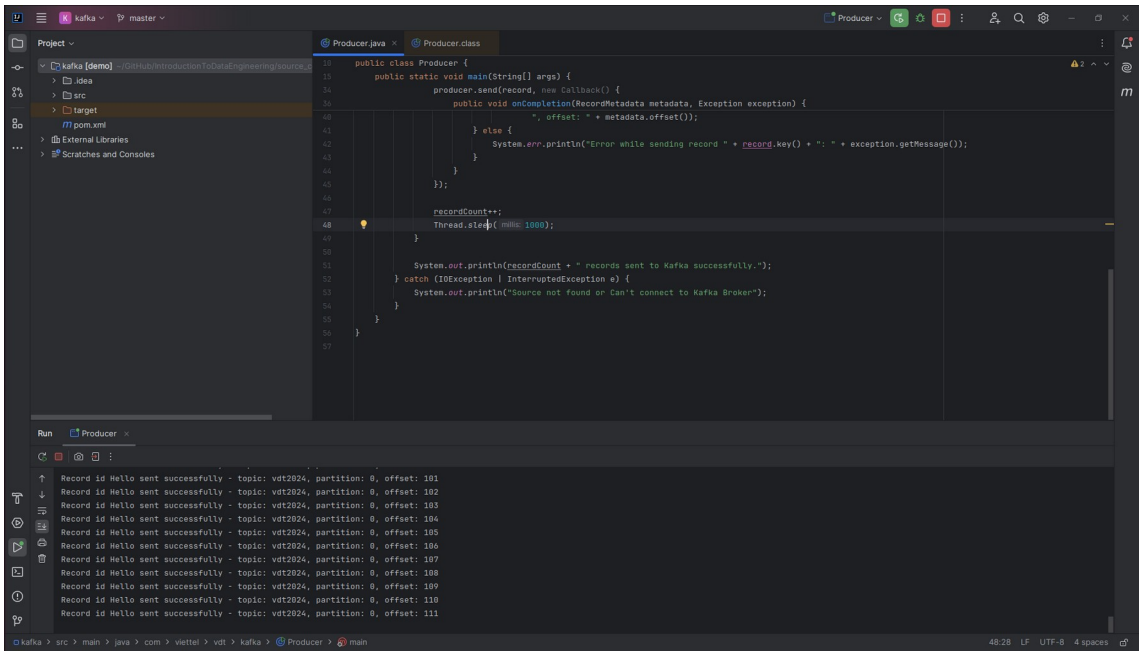


Link github:
https://github.com/Tran-Ngoc-Bao/IntroductionToDataEngineering

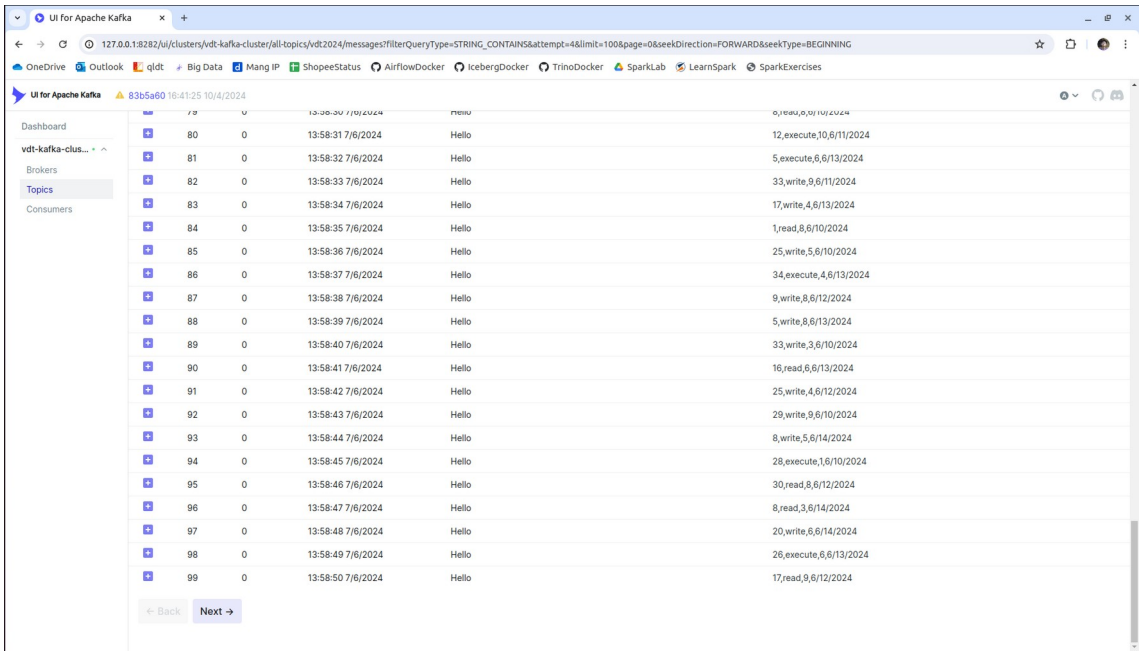
CHƯƠNG 2. THỰC HIỆN CÁC BƯỚC

2.1 Đẩy dữ liệu lên Kafka Topic

Viết một chương trình Java đọc từng dòng từ file log_action.csv và đẩy dữ liệu từng dòng theo định dạng string với dữ liệu mỗi trường cách nhau bởi dấu phẩy. Sau đó đẩy dữ liệu lên Kafka Topic vdt2024 theo từng giây.



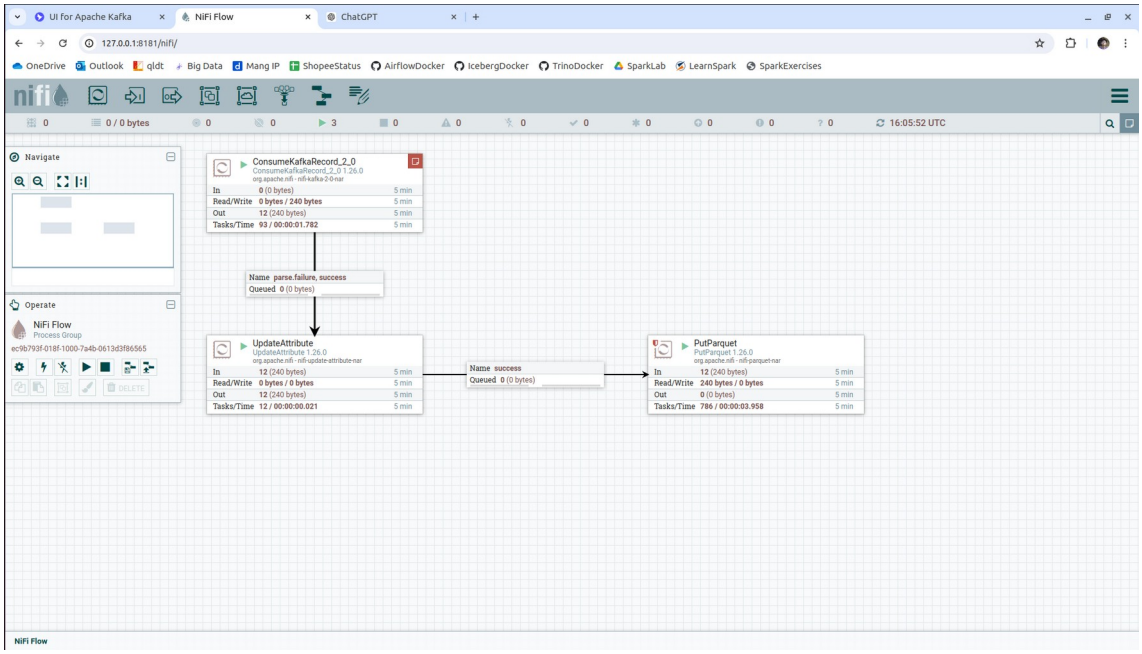
Kết quả thu được hiển thị trên Kafka-ui:



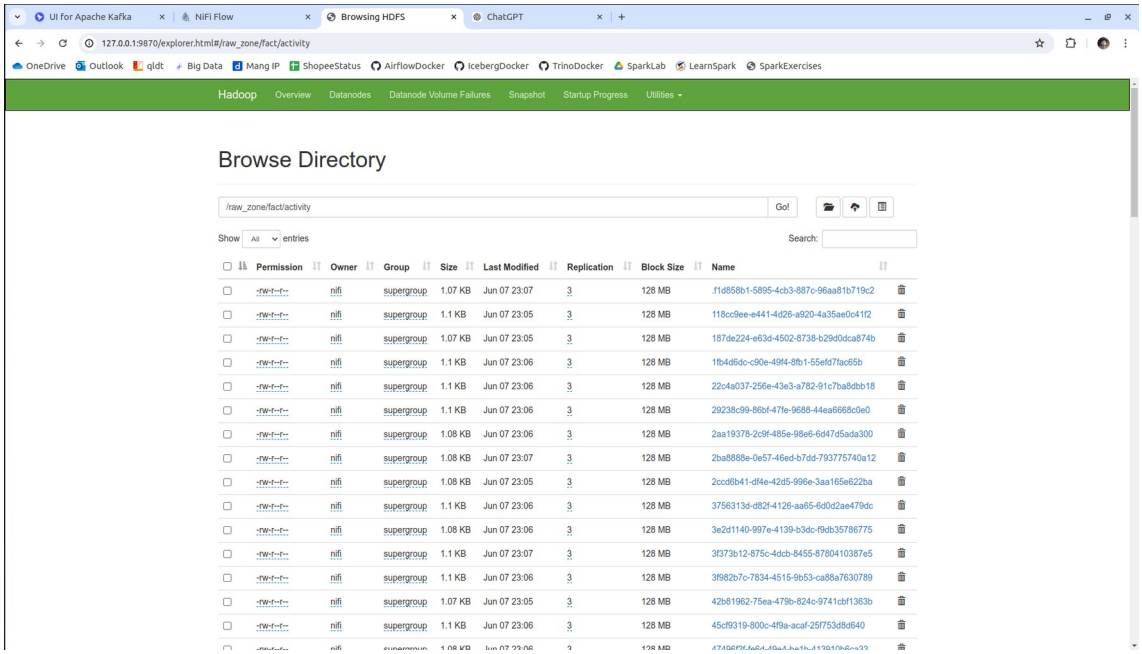
2.2 Triển khai Nifi, kéo dữ liệu từ Kafka Topic vdt2024, xử lý và lưu dữ liệu dưới dạn parquet xuống HDFS

Xây dựng một luồng xử lý dữ liệu đơn giản bằng Nifi, gồm có các Processer sau:

- ConsumeKafkaRecord_2_0: Lấy dữ liệu real time từ Kafka và chuyển tiếp đi.
- UpdateAttribute: Giúp chuyển định dạng dữ liệu thô từ Kafka sang định dạng file parquet, rồi sau đó chuyển tiếp đi.
- PutParquet: Lưu trữ file parquet vào HDFS được cấu hình trong processer.

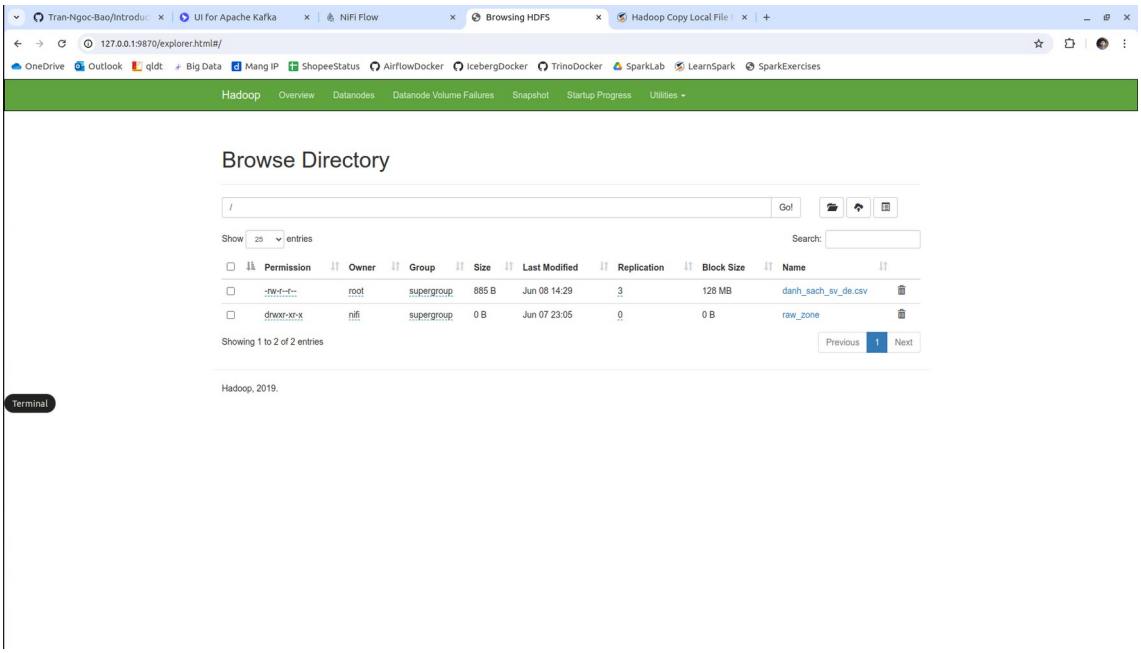


Kết quả thu được trên Hadoop:



2.3 Lưu trữ file danh_sach_sv_de.csv xuống HDFS

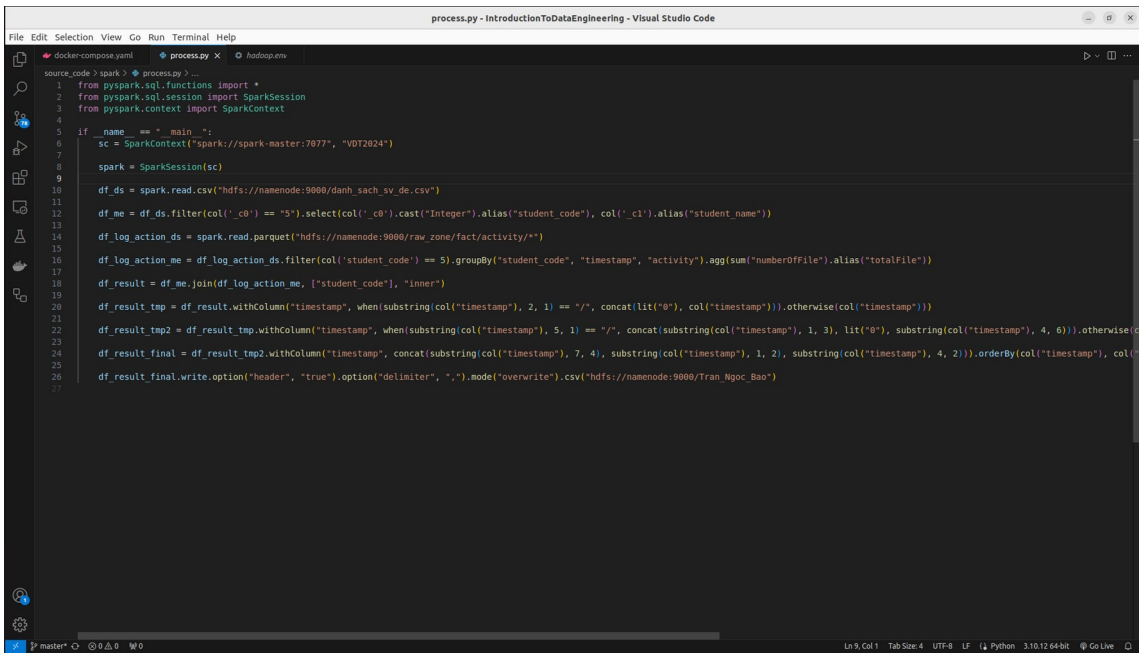
Đẩy trực tiếp file danh_sach_sc_de.csv từ local vào HDFS bằng câu lệnh trong terminal. Kết quả thu được:



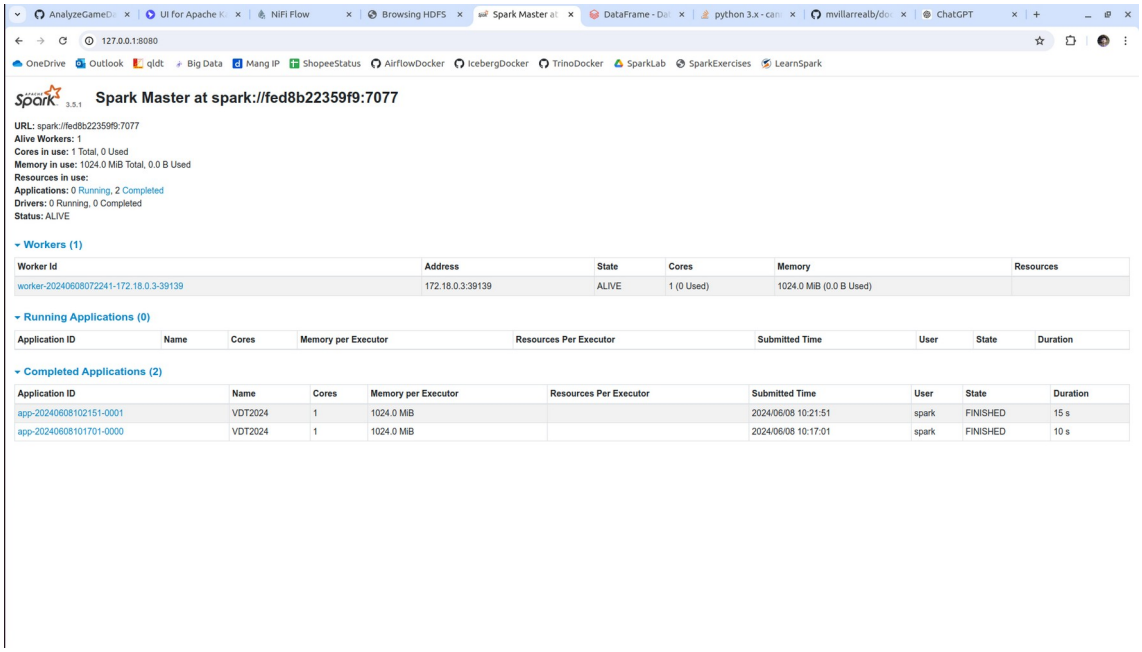
2.4 Sử dụng Apache Spark xử lý dữ liệu lưu trữ dưới HDFS

Viết một chương trình Pyspark để xử lý dữ liệu với chi tiết mã nguồn như sau:

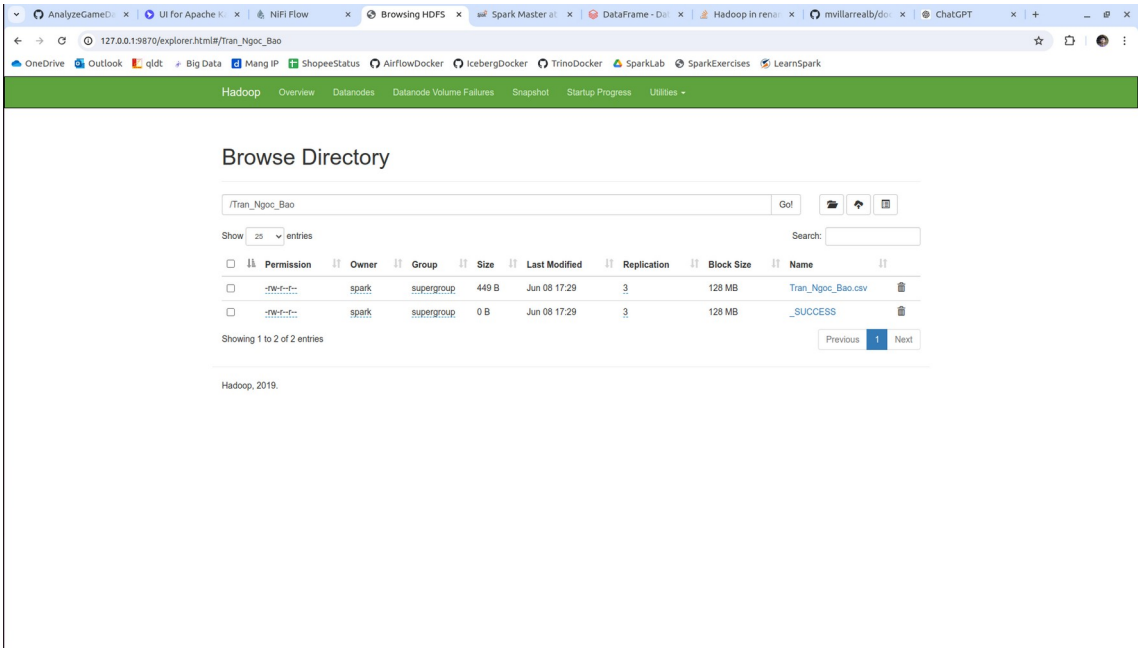
- Đọc file danh_sach_sv_de.csv
- Lọc thông tin cá nhân theo mã số sinh viên và đổi lại header cho dễ nhớ
- Đọc tất cả các file parquet được lưu ở đường dẫn /raw_zone/fact/activity trong HDFS
- Lọc log theo mã số sinh viên và tính tổng số file thực hiện theo từng phân loại
- Join 2 dataframe lại với nhau với khóa mã số sinh viên
- Xử lý dữ liệu thời gian theo đúng định dạng
- Lưu file dưới định dạng csv với các yêu cầu đề ra



Kết quả khi submit file Pyspark lên Spark-master:



Kết quả thu được trên HDFS:



2.5 Kết quả đầu ra

Trường dữ liệu date được sắp xếp theo thời gian từ lâu nhất tới sớm nhất và trường dữ liệu action được sắp xếp theo thứ tự alphabet