**BÁO CÁO ASSIGNMENT**

**TRẦN PHÚC MẠNH LINH**

STT: 21

**Data Engineering**

**VDT2024**

Source Code: [megaheart/viettel-assignment-21-linh-tpm-1 (github.com)](https://github.com/megaheart/viettel-assignment-21-linh-tpm-1)

Sinh viên thực hiện

Ký và ghi rõ họ tên

**MỤC LỤC**

[CHƯƠNG 1. CÁCH THỨC TRIỂN KHAI 1](#_Toc168442141)

[1.1 Tổng quan các thành phần 1](#_Toc168442142)

[1.2 Kafka Producer 1](#_Toc168442143)

[1.3 Cấu hình Kafka 2](#_Toc168442144)

[1.4 Cấu hình Nifi 3](#_Toc168442145)

[1.5 Cấu hình HDFS 4](#_Toc168442146)

[1.6 Thiết lập Spark Notebook 6](#_Toc168442147)

[CHƯƠNG 2. THỰC HIỆN XỬ LÝ 7](#_Toc168442148)

[2.1 Khởi chạy Kafka Producer 7](#_Toc168442149)

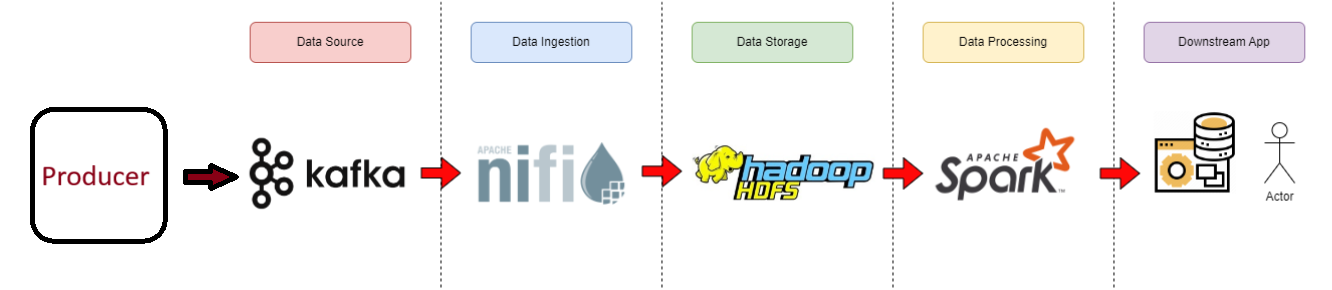
[2.2 Lấy dữ liệu từ Kafka lưu vào HDFS 7](#_Toc168442150)

[2.3 Lưu trữ file “danh\_sach\_sv\_de.csv” xuống HDFS 8](#_Toc168442151)

[2.4 Xử lý dữ liệu lưu trữ dưới HDFS, sử dụng Apache Spark 9](#_Toc168442152)

# CÁCH THỨC TRIỂN KHAI

## Tổng quan các thành phần



Hệ thống sẽ bao gồm các thành phần

* **Kafka Producer**: Cung cấp các bản ghi từ log\_action.csv đẩy lên Kafka topic *vdt2024*
* **Kafka Broker**: Trung gian điều phối việc pub/sub của các publisher/consumer và lưu trữ các topic, trong đó có topic *vdt2024*
* **Nifi Flow**: Phụ trách việc lấy dữ liệu từ topic *vdt2024* và lưu lại vào trong HDFS với đường dẫn */raw\_zone/fact/activity* với format parquet.
* **HDFS**: Lưu trữ các action và file *danh\_sach\_sv\_de.csv*
* **Spark**: Xử lý dữ liệu lưu dưới HDFS và xuất ra output là tổng số file được tương tác hàng ngày theo mỗi loại activity mà mỗi sinh viên thực hiện.

## Kafka Producer

Đoạn mã thực hiện Kafka Producer được viết bằng Python yêu cầu thư viện *confluent\_kafka* để chạy, file chương trình nằm ở *kafka\producer\_kafka.py*

A screen shot of a computer program

Description automatically generated

## Cấu hình Kafka

Kafka sẽ gồm 2 container với cấu hình khởi chạy như sau:

A screen shot of a computer

Description automatically generated

A screenshot of a computer program

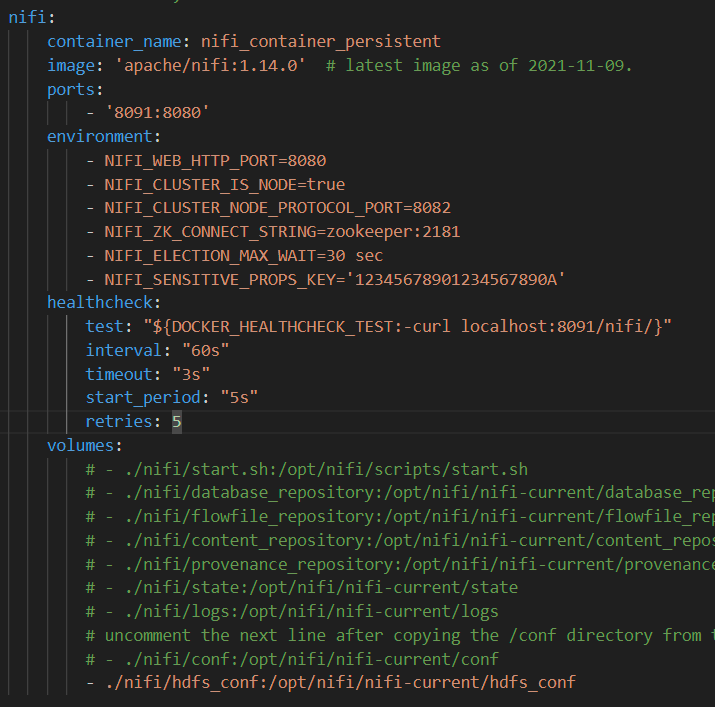
Description automatically generated

Trong đó zookeeper phụ trách quản lý các metadata (các topic, partition, offset và cấu hình cluster), các leader election, tình trạng các broker và đồng bộ hóa producer và consumer. Còn broker01 sẽ phụ trách lưu trữ các topic trên các partition cũng như cho phép các producer/consumer pub/sub dữ liệu.

Broker ở đây được cấu hình để có thể lắng nghe các yêu cầu pub/sub đến domain localhost:9092 hoặc broker01:9093 với giao thức PLAINTEXT. Do chỉ chạy 1 node nên em thiết lập chỉ dùng 1 partition.

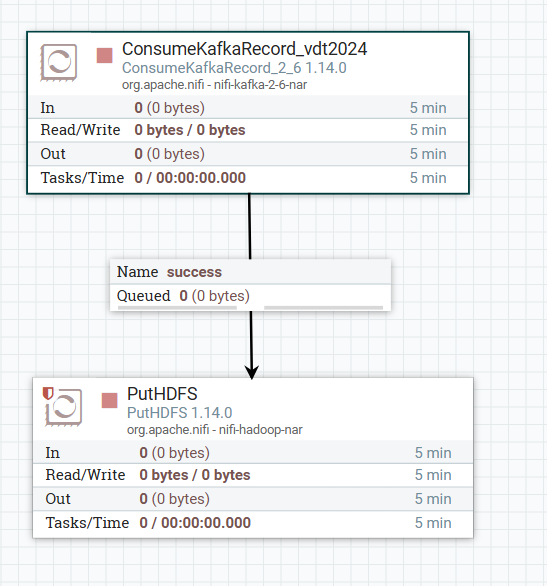
## Cấu hình Nifi

Nifi sẽ được chạy dưới 1 container với thiết lập như sau:



Có thể truy cập trang Web từ http://localhost:8091/nifi/ để chỉnh sửa Flow.

Nifi Flow sẽ được cấu hình như sau:



Processor **ConsumeKafkaRecord\_vdt2024** sẽ phụ trách lấy dữ liệu từ topic *vdt2024* và chuyển đổi về dạng parquet còn processor **PutHDFS** sẽ phụ trách đẩy dữ liệu parquet xuống HDFS với đường dẫn */raw\_zone/fact/activity*

Có thể import Template của Flow trên từ file *nifi\kafka2hdfs.xml*

## Cấu hình HDFS

HDFS sẽ được chạy trên 1 namenode container và 2 datanode container:



A computer screen shot of a program

Description automatically generated

Trong đó cổng 9000 trên namenode sẽ dùng để kết nối phục vụ cho việc đọc ghi trên HDFS, còn cổng 9870 thì phục vụ cho trang web giúp dễ dàng theo dõi trạng thái HDFS.

Ngoài ra một số cấu hình Hadoop khác được thiết lập trên file *hadoop/hadoop.env*:

A computer screen shot of a computer program

Description automatically generated

Các thiết lập liên quan đến HDFS bao gồm việc cho phép truy cập HDFS qua HTTP (WebHDFS), vô hiệu hóa kiểm tra quyền trên HDFS, và tắt kiểm tra IP/tên máy chủ trong quá trình đăng ký datanode với namenode. Ngoài ra, hệ số nhân bản được thiết lập là 1 (vì chỉ có 2 data node).

## Thiết lập Spark Notebook

Spark Notebook sẽ được chạy trong một container được thiết lập như sau:

A screen shot of a computer

Description automatically generated

Việc này giúp cho Spark được thực thi trong cùng 1 network với các container của HDFS, từ đó có thể kết nối được với HDFS dễ dàng hơn.

File cấu hình của notebook là *spark\jupyter\_notebook\_config.py*. Bên trong file là cấu hình mật khẩu của Spark Notebook

File Notebook nằm ở *spark\process\_data\_spark.ipynb*, để thực hiện file này ta cần kết nối đến Jupiter-server thông qua domain <http://localhost:8888> và mật khẩu là 123456

# THỰC HIỆN XỬ LÝ

## Khởi chạy Kafka Producer

Thực thi file chương trình nằm ở *kafka\producer\_kafka.py*, chương trình sẽ đẩy từng dòng trong file *log\_action.csv* lên kafka topic *vdt2024* theo từng giây.

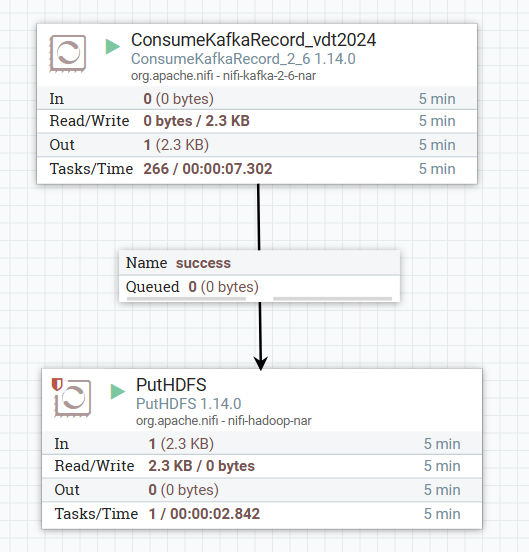
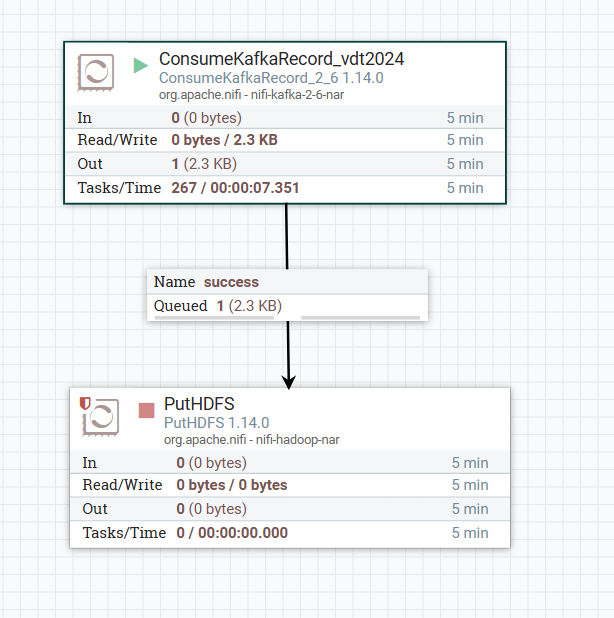
A screenshot of a computer

Description automatically generated

## Lấy dữ liệu từ Kafka lưu vào HDFS

Tiến hành lấy dữ liệu từ topic *vdt2024* và lưu lại vào trong HDFS với đường dẫn */raw\_zone/fact/activity* với format parquet bằng Nifi.

Vào trang Web <http://localhost:8091/nifi/> sau đó khởi chạy ConsumeKafkaRecord\_vdt2024 và sau đó là PutHDFS.



Kết thúc thì file HDFS với đường dẫn */raw\_zone/fact/activity* đã được ghi thành công.

A close up of a computer screen

Description automatically generated

## Lưu trữ file “danh\_sach\_sv\_de.csv” xuống HDFS

Mở file Notebook nằm ở *spark\process\_data\_spark.ipynb*, để thực hiện file này ta cần kết nối đến Jupiter-server container thông qua domain <http://localhost:8888> và mật khẩu là 123456.

Đầu tiên ta tạo một Spark Context và Spark Session

A screen shot of a computer

Description automatically generated

Sau đó ta tạo cấu trúc dữ liệu cho Action và SV\_DE:

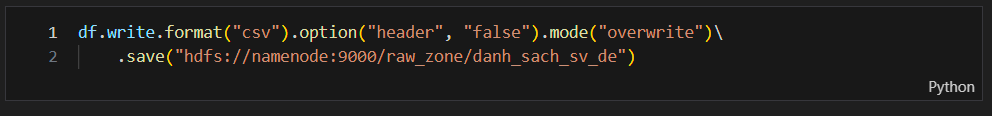
A screen shot of a computer code

Description automatically generated

Tiếp theo ta đọc file *danh\_sach\_sv\_de.csv* rồi sau đó lưu xuống HDFS:

A screen shot of a computer

Description automatically generated



Ta mở [http://localhost:9870/](http://localhost:9870/explorer.html#/raw_zone/danh_sach_sv_de) lên để kiểm tra kết quả. File đã được lưu xuống HDFS.

A screenshot of a computer

Description automatically generated

## Xử lý dữ liệu lưu trữ dưới HDFS, sử dụng Apache Spark

Trong notebook, ta thực thi đoạn lệnh đọc danh sách sinh viên DE từ HDFS:

A screen shot of a computer

Description automatically generated

Tiếp tục đọc */raw\_zone/fact/activity* từ HDFS:

A screen shot of a computer

Description automatically generated

Sau đó tiến hành tính tổng số file được tương tác hàng ngày theo mỗi loại activity

mà sinh viên đó thực hiện. Trong quá trình này thực hiện join với Dataframe của SV DE để lấy họ tên luôn. Quá trình này do chỉ toàn Transformations nên Spark chưa thực hiện Job.

A screen shot of a computer code

Description automatically generated

Tiếp theo ta sắp xếp lại các cột cho đúng thứ tự đề bài yêu cầu và lưu ra file csv kết quả:

A screen shot of a computer code

Description automatically generated

Do lưu bằng Spark Dataframe write nó sẽ không ra một file duy nhất với tên sinh viên như yêu cầu nên em phải lưu bằng thư viện chuẩn của python để đáp ứng yêu cầu của đề bài. Cách lưu bằng Dataframe write:

A screen shot of a computer code

Description automatically generated