BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC NHA TRANG

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN MÔN HỌC**

**XỬ LÝ DỮ LIỆU LỚN (INS358)**

**XÂY DỰNG CHƯƠNG TRÌNH TỔNG HỢP CÁC TWEETS CỦA ELON MUSK TRÊN HADOOP/SPARK**

**Sinh viên thực hiện: Trần Duy Phát**

**MSSV: 63135054 Lớp: 63.CNT-3**

**Giảng viên: TS. Nguyễn Đình Hưng**

**Khánh Hòa - 2024**

**KẾT QUẢ ĐÁNH GIÁ ĐỒ ÁN MÔN HỌC**

*Họ và tên sinh viên:*

*MSSV:*

*Lớp:*

|  |  |  |
| --- | --- | --- |
| **Nội dung** | **Trọng số** | **Điểm** |
| **1. Giải quyết vấn đề** | | |
| 1.1. Phân tích bài toán; thu thập, khảo sát và chuẩn bị dữ liệu; thiết kế giải thuật | 20% |  |
| 1.2. Cài đặt, triển khai ứng dụng trên Hadoop | 20% |  |
| 1.3. Cài đặt, triển khai ứng dụng trên Spark | 20% |  |
| **2. Báo cáo bài tập lớn** | | |
| 2.1. Nội dung báo cáo | 20% |  |
| 2.2. Vấn đáp | 20% |  |
| **Điểm trung bình** | |  |

*Giảng viên*

**Lời cam đoan**

Tôi cam đoan đây là công trình do tôi tự thực hiện. Các nội dung nghiên cứu, số liệu và kết quả thực nghiệm là trung thực. Các số liệu, công trình sử dụng của tác giả khác đều được trích dẫn nguồn gốc rõ ràng.

Tất cả phần mềm sử dụng trong đồ án này đều là mã nguồn mở.

Nếu phát hiện có bất kì sự gian lận nào, tôi xin chịu hoàn toàn trách nhiệm.

Trần Duy Phát

Mục lục

[Mục lục 4](#_Toc194164218)

[Danh sách hình 6](#_Toc194164219)

[Chương 1. Giới thiệu 7](#_Toc194164220)

[1.1 Tổng quát về dữ liệu lớn 7](#_Toc194164221)

[1.2 Mục tiêu của đề tài 7](#_Toc194164222)

[1.3 Cấu trúc của đồ án 7](#_Toc194164223)

[Chương 2. Nội dung và phương pháp thực hiện 8](#_Toc194164224)

[2.1 Phân tích bài toàn 8](#_Toc194164225)

[2.2 Thu thập dữ liệu 8](#_Toc194164226)

[2.3 Cài đặt và triển khai ứng dụng trên Hadoop 8](#_Toc194164227)

[2.3.1 Cài đặt Hadoop 8](#_Toc194164228)

[2.3.2 Xây dựng giải thuật 15](#_Toc194164229)

[2.3.3 Lập trình ứng dụng 15](#_Toc194164230)

[2.3.4 Thực thi ứng dụng 17](#_Toc194164231)

[2.4 Cài đặt và triển khai ứng dụng trên Spark 19](#_Toc194164232)

[2.4.1 Cài đặt Spark 19](#_Toc194164233)

[2.4.2 Lập trình ứng dụng 20](#_Toc194164234)

[2.4.3 Thực thi ứng dụng 22](#_Toc194164235)

[Chương 3. Kết luận 24](#_Toc194164236)

[3.1 Đánh giá chung 24](#_Toc194164237)

[3.1.1 Những kết quả đạt được 24](#_Toc194164238)

[3.1.2 Một số hạn chế 24](#_Toc194164239)

[3.2 Hướng phát triển 24](#_Toc194164240)

[Tài liệu tham khảo 25](#_Toc194164241)

Danh sách hình

[Hình 2.1 Cài đặt SSH 9](#_Toc194164242)

[Hình 2.2 Minh hoạ khi kiểm tra SSH 9](#_Toc194164243)

[Hình 2.3 Ảnh minh hoạ Hadoop đã chạy 14](#_Toc194164244)

[Hình 2.4 Hình ảnh khi truy cập Hadoop 14](#_Toc194164245)

[Hình 2.5 Mô tả thuật toán hoạt động 15](#_Toc194164246)

[Hình 2.6 Ảnh mình hoạ khi chạy hoàn thành MapReduce 18](#_Toc194164247)

[Hình 2.7 Kết quả trả về. 18](#_Toc194164248)

[Hình 2.8 Kết quả của thực thi 19](#_Toc194164249)

[Hình 2.9 Ảnh minh hoạ Pyspark đã chạy 20](#_Toc194164250)

[Hình 2.10 Ảnh mình hoạ code tạo python của ứng dụng spark 20](#_Toc194164251)

[Hình 2.11 Ảnh mình hoạ dòng lệnh chạy ứng dụng cho Spark. 22](#_Toc194164252)

[Hình 2.12 Ảnh mình hoạ kết quả về đếm số tweet theo ngày. 23](#_Toc194164253)

[Hình 2.13 Ảnh mình hoạ kết quả về đếm số tweet theo giờ. 23](#_Toc194164254)

[Hình 2.14 Kết quả khung giờ Elon Musk hay đăng tweet nhất. 23](#_Toc194164255)

1. Giới thiệu
   1. Tổng quát về dữ liệu lớn

Dữ liệu lớn (Big Data) đề cập đến tập hợp các dữ liệu với khối lượng khổng lồ, đa dạng và được tạo ra với tốc độ nhanh, vượt xa khả năng xử lý của các hệ thống truyền thống. Việc khai thác và phân tích dữ liệu lớn không chỉ giúp các doanh nghiệp và tổ chức hiểu rõ hơn về xu hướng thị trường, hành vi người tiêu dùng mà còn hỗ trợ tối ưu hóa quy trình sản xuất, nâng cao hiệu quả hoạt động và ra quyết định dựa trên các thông tin chính xác. Công nghệ tiên tiến như trí tuệ nhân tạo và học máy đang được ứng dụng rộng rãi nhằm chuyển đổi dữ liệu thành giá trị thực tiễn, đồng thời tạo ra cơ hội đổi mới sáng tạo trong kỷ nguyên số. Tuy nhiên, sự bùng nổ của dữ liệu lớn cũng đặt ra những thách thức về cơ sở hạ tầng, bảo mật thông tin và quản lý dữ liệu hiệu quả.

* 1. Mục tiêu của đề tài

Đề tài nhằm đạt được các mục tiêu chính sau:

• Nghiên cứu toàn diện khái niệm và những ứng dụng thực tiễn của dữ liệu lớn.

• Khám phá và phân tích các phương pháp, công nghệ cùng các công cụ tiêu biểu được ứng dụng trong lĩnh vực xử lý dữ liệu lớn.

• Áp dụng kiến thức đã học để thiết kế và xây dựng một ứng dụng cơ bản xử lý dữ liệu lớn.

* 1. Cấu trúc của đồ án

Đồ án gồm các phần sau:

• Chương 1: Giới thiệu.

• Chương 1: Nội dung và phương pháp thực hiện.

• Chương 1: Kết luận

1. Nội dung và phương pháp thực hiện
   1. Phân tích bài toàn

Cho bộ dữ liệu chứ các tweets của Elon Musk chứa: id, thời gian tạo (create\_at) và nội dung (text). Yêu cầu bài toàn: Đếm số tweet của từng ngày, đếm số tweet theo từng khung giờ để xác định khung giờ mà Elon Musk thường đăng tweet.

* 1. Thu thập dữ liệu

Trong dự án này tôi sử dụng nguồn dữ liệu:   
<https://github.com/nd-hung/Big-Data/blob/main/datasets/ElonMusk_tweets.csv>

* 1. Cài đặt và triển khai ứng dụng trên Hadoop
     1. Cài đặt Hadoop

**1. Tạo tài khoản quản trị Hadoop**

Để đảm bảo bảo mật, tạo một tài khoản riêng để quản lý Hadoop:

sudo adduser hdoop

Nhập mật khẩu khi được yêu cầu.

Cấp quyền sudo cho tài khoản:

sudo usermod -aG sudo hdoop

Đăng nhập tài khoản Hadoop:

su - hdoop

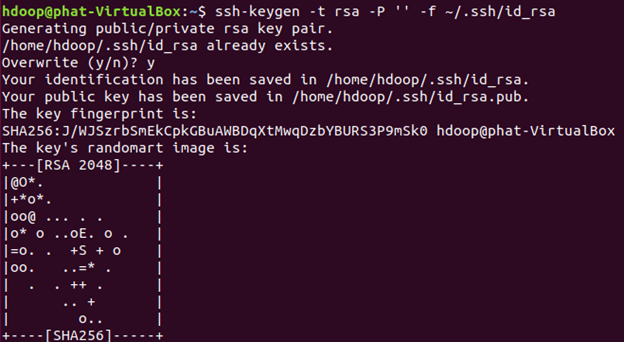
**2. Cài đặt SSH**

SSH giúp truy cập từ xa. Cài đặt SSH bằng lệnh:

sudo apt install openssh-server openssh-client -y

Tạo cặp khóa SSH để đăng nhập không cần mật khẩu:

ssh-keygen -t rsa -P '' -f ~/.ssh/id\_rsa



Cài đặt SSH

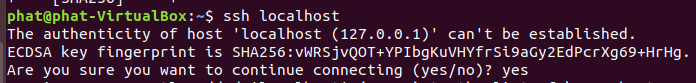
Thêm khóa vào danh sách được ủy quyền:

cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

chmod 0600 ~/.ssh/authorized\_keys

Kiểm tra SSH:

ssh localhost



Minh hoạ khi kiểm tra SSH

Nhập **yes** khi được hỏi.

Lưu và khởi động lại hệ thống. Kiểm tra trạng thái:

cat /proc/sys/net/ipv6/conf/all/disable\_ipv6

Nếu kết quả là 1, IPv6 đã bị tắt.

**4. Cài đặt Java**

Hadoop 3.2.x yêu cầu Java 8:

sudo apt install openjdk-8-jdk -y

**5. Cài đặt Hadoop**

Tải Hadoop 3.2.2:

wget https://archive.apache.org/dist/hadoop/common/hadoop-3.2.2/hadoop-3.2.2.tar.gz

Giải nén:

tar xzf hadoop-3.2.2.tar.gz

**6. Thiết lập biến môi trường**

Mở tệp *~/.bashrc*:

sudo nano ~/.bashrc

Thêm vào cuối:

# Hadoop Environment Variables  
export HADOOP\_HOME=/home/hdoop/hadoop-3.2.2  
export PATH=$PATH:$HADOOP\_HOME/sbin:$HADOOP\_HOME/bin  
export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"  
Djava.library.path=$HADOOP\_HOME/lib/native"

Lưu và áp dụng:

source ~/.bashrc

7. Cấu hình Hadoop

Mở hadoop-env.sh:

sudo nano $HADOOP\_HOME/etc/hadoop/hadoop-env.sh

Thêm:

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64

Thiết lập thư mục tạm:

sudo mkdir -p /app/hadoop/tmp

sudo chown hdoop:hdoop /app/hadoop/tmp

sudo chmod 750 /app/hadoop/tmp

Mở *core-site.xml*:

sudo nano $HADOOP\_HOME/etc/hadoop/core-site.xml

Thêm:

<configuration>

<property>

<name>hadoop.tmp.dir</name>

<value>/app/hadoop/tmp</value>

</property>

<property>

<name>fs.defaultFS</name>

<value>hdfs://localhost:9000</value>

</property>

</configuration>

Mở *hdfs-site.xml*:

sudo nano $HADOOP\_HOME/etc/hadoop/hdfs-site.xml

Thêm:

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

</configuration>

Mở *mapred-site.xml*:

sudo nano $HADOOP\_HOME/etc/hadoop/mapred-site.xml

Thêm:

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

Mở *yarn-site.xml*:

sudo nano $HADOOP\_HOME/etc/hadoop/yarn-site.xml

Thêm:

<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

</configuration>

8. Định dạng HDFS

hdfs namenode -format

9. Khởi động và dừng Hadoop

Khởi động:

start-dfs.sh

start-yarn.sh

Hoặc:

start-all.sh

Dừng Hadoop:

stop-dfs.sh

stop-yarn.sh

Hoặc:

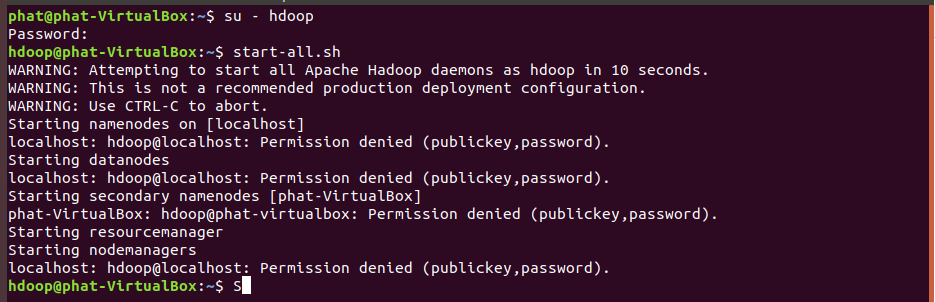
stop-all.sh

10. Truy cập Hadoop qua trình duyệt

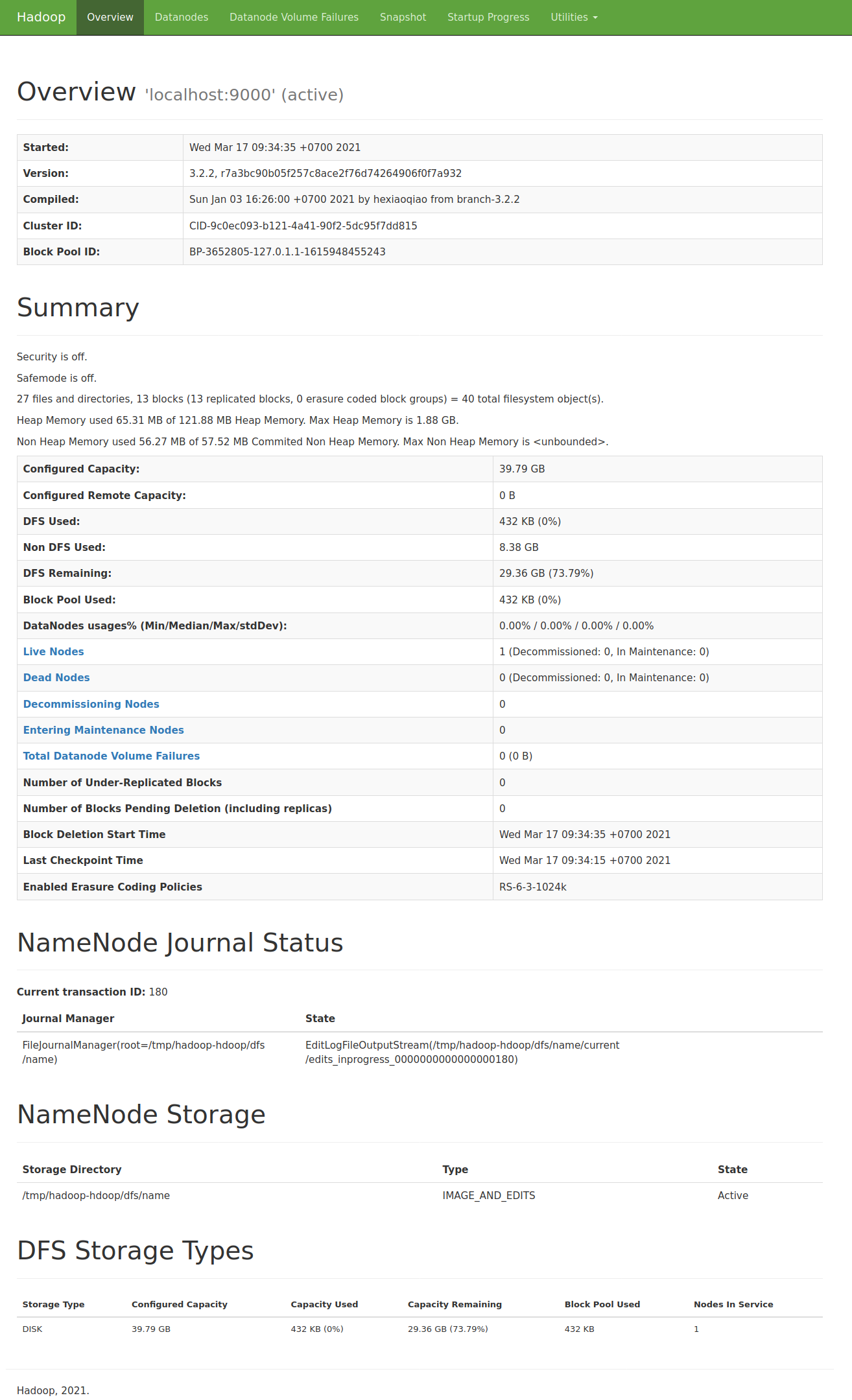
Mở trình duyệt và truy cập:

<http://localhost:9870>

Hoàn tất quá trình cài đặt và cấu hình Hadoop trên Ubuntu.

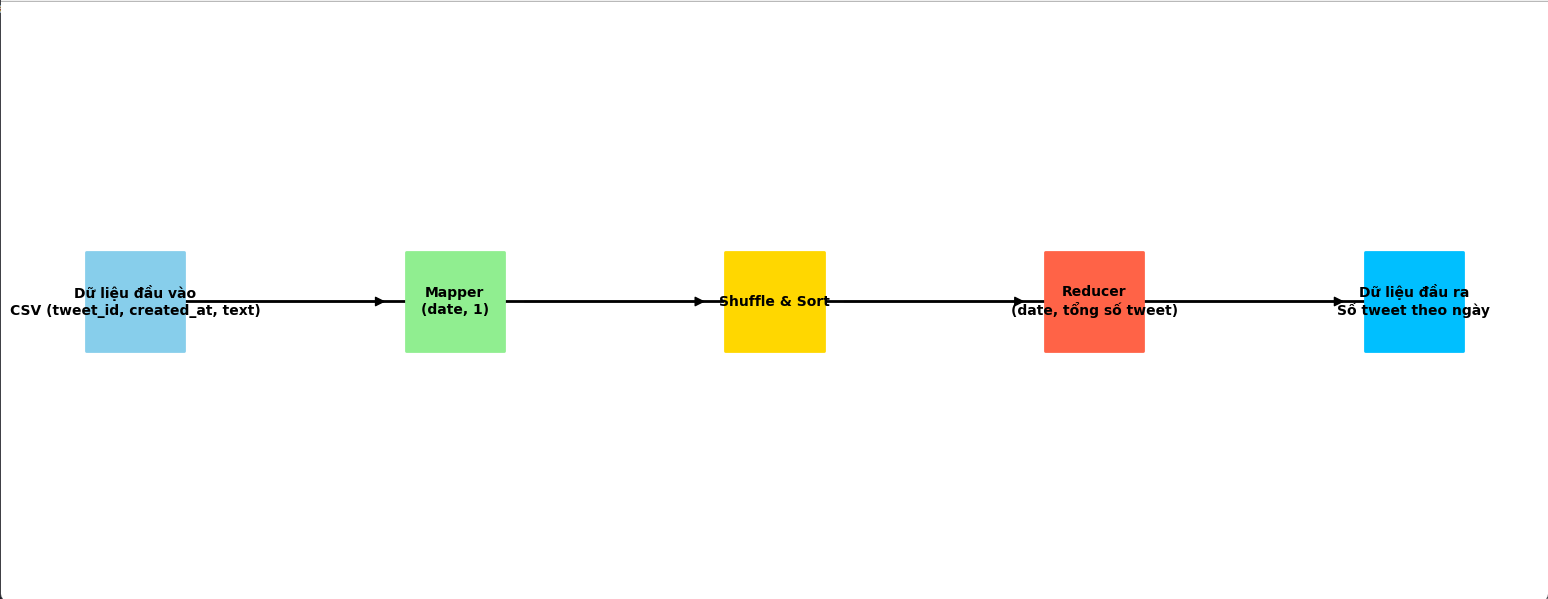


Ảnh minh hoạ Hadoop đã chạy

[](https://github.com/nd-hung/Big-Data/blob/main/Lab1_Hadoop_Installation/figs/hadoop_ui.png)

Hình ảnh khi truy cập Hadoop

* + 1. Xây dựng giải thuật



Mô tả thuật toán hoạt động

* + 1. Lập trình ứng dụng

1. Đăng nhập tài khoản hdoop

Mở terminal và chạy lệnh:

su - hdoop

2. Tạo các file Python cho MapReduce

2.1. Tạo file mapper\_date.py

Mở terminal và nhập lệnh:

nano mapper\_date.py

Dán nội dung sau:

#!/usr/bin/python3

import sys

import csv

for line in sys.stdin:

try:

row = next(csv.reader([line])) # Đọc dòng CSV

tweet\_id, created\_at, text = row

date = created\_at[:10] # Lấy ngày từ created\_at (YYYY-MM-DD HH:MM:SS)

print(f"{date}\t1") # Output: (date, 1)

except Exception:

continue # Bỏ qua dòng lỗi

Lưu file (Ctrl+X, nhấn Y, Enter).

2.2. Tạo file reducer\_date.py

Mở terminal và nhập lệnh:

nano reducer\_date.py

Dán nội dung sau:

#!/usr/bin/python3

import sys

from collections import defaultdict

tweet\_count = defaultdict(int)

for line in sys.stdin:

day, count = line.strip().split("\t")

tweet\_count[day] += int(count)

for day in sorted(tweet\_count):

print(f"{day}\t{tweet\_count[day]}")

Lưu file.

2.3. Cấp quyền thực thi

Chạy lệnh sau:

chmod +x mapper\_date.py reducer\_date.py

3. Upload dữ liệu lên HDFS

Chạy Hadoop:

start-all.sh

Giả sử file dữ liệu ElonMusk\_tweets.csv đã được lưu tại /home/phat/Downloads/tweet. Tải file lên HDFS:

hdfs dfs -mkdir -p /user/hdoop/data

hdfs dfs -copyFromLocal /home/phat/Downloads/tweet /user/hdoop/data

* + 1. Thực thi ứng dụng

1. Kiểm tra file trong HDFS:

hdfs dfs -ls /user/hdoop/data

Xóa output cũ (nếu có):

hdfs dfs -rm -r /user/hdoop/data/tweet\_count\_by\_date

2. Chạy job MapReduce:

hadoop jar $HADOOP\_HOME/share/hadoop/tools/lib/hadoop-streaming-3.2.2.jar \

-file mapper\_date.py -mapper mapper\_date.py \

-file reducer\_date.py -reducer reducer\_date.py \

-input /user/hdoop/data/tweet \

-output /user/hdoop/data/tweet\_count\_by\_date

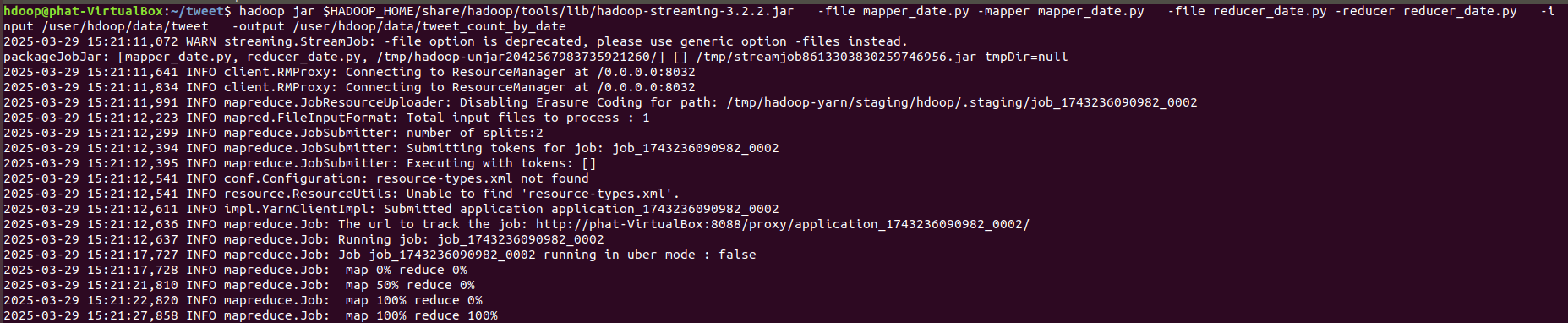
3. Xem kết quả:

hdfs dfs -cat /user/hdoop/data/tweet\_count\_by\_date/part-00000

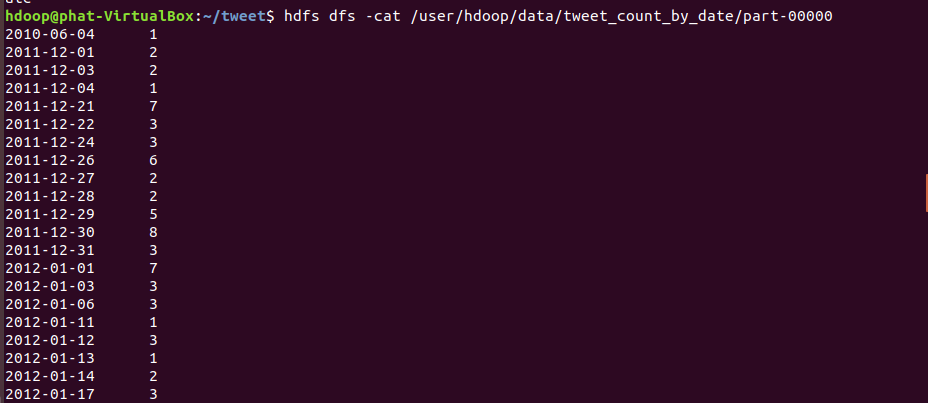
4. Dừng Hadoop

Sau khi hoàn tất công việc, dừng Hadoop:

stop-all.sh



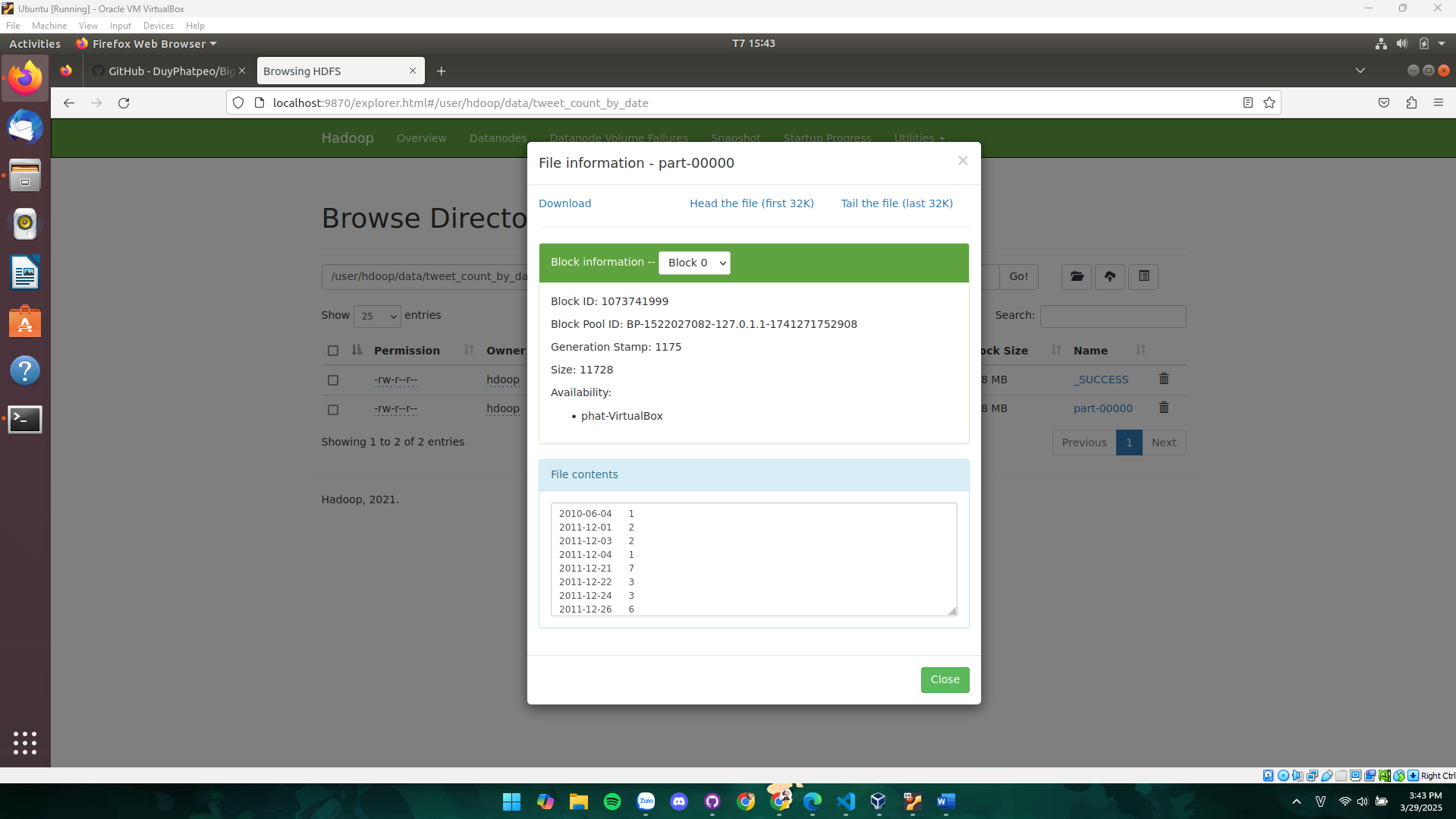
Ảnh mình hoạ khi chạy hoàn thành MapReduce



Kết quả trả về.

Khi truy cập vào Hadoop:

localhost:9870/explorer.html#/user/hdoop/date/tweet\_count\_by\_date



Kết quả của thực thi

* 1. Cài đặt và triển khai ứng dụng trên Spark
     1. Cài đặt Spark

**1. Tải và giải nén Spark**

Tải phiên bản Spark 3.5.5:

wget https://dlcdn.apache.org/spark/spark-3.5.5/spark-3.5.5-bin-hadoop3.tgz

Giải nén:

tar xvf spark-3.5.5-bin-hadoop3.tgz

Di chuyển thư mục Spark:

sudo mv spark-3.5.5-bin-hadoop3 /opt/spark

**2. Thiết lập biến môi trường**

Mở file ~/.bashrc và thêm các dòng sau:

echo "export SPARK\_HOME=/opt/spark" >> ~/.bashrc

echo "export PATH=$SPARK\_HOME/bin:$SPARK\_HOME/sbin:$PATH" >> ~/.bashrc

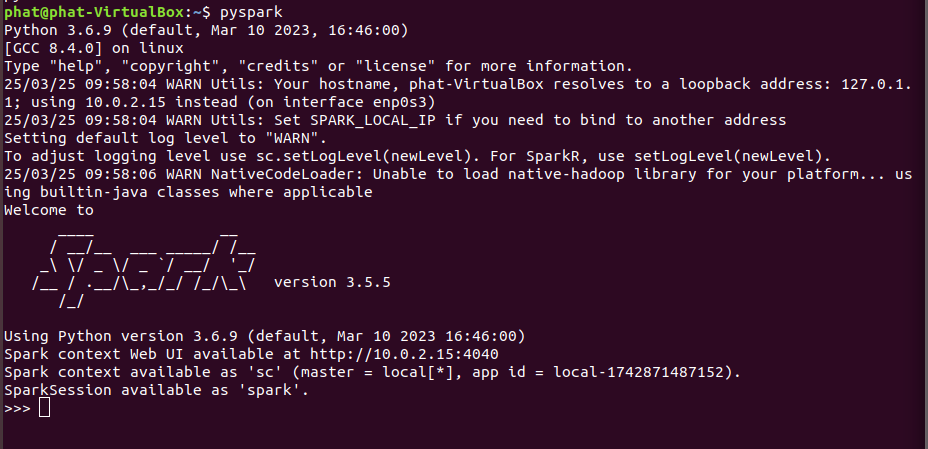
echo "export PYSPARK\_PYTHON=python3" >> ~/.bashrc

Nạp lại cấu hình:

source ~/.bashrc

**3. Khởi động PySpark**

pyspark



Ảnh minh hoạ Pyspark đã chạy

* + 1. Lập trình ứng dụng

**Mở terminal và tạo file Python**

Nhập lệnh sau để mở trình chỉnh sửa nano:

nano tweet\_analysis.py

Sau đó, dán nội dung code vào file và lưu lại (**Ctrl + X → Y → Enter**).



Ảnh mình hoạ code tạo python của ứng dụng spark

Sau đó dán vào file python: :///mnt/data/ElonMusk\_tweets.csv là nơi lưu file dữ liệu

from pyspark.sql import SparkSession

# Khởi tạo SparkSession

spark = SparkSession.builder.appName("TweetAnalysis").getOrCreate()

sc = spark.sparkContext

# Đọc file dữ liệu

rdd = sc.textFile("file:///mnt/data/ElonMusk\_tweets.csv")

# Bỏ dòng tiêu đề

header = rdd.first()

rdd = rdd.filter(lambda line: line != header)

# (a) Đếm số tweet theo ngày

def extract\_date(line):

fields = line.split(",")

if len(fields) < 2:

return None

date = fields[1].split(" ")[0] # Lấy phần YYYY-MM-DD

return (date, 1)

tweet\_by\_date = rdd.map(extract\_date).filter(lambda x: x is not None).reduceByKey(lambda a, b: a + b)

tweet\_by\_date\_sorted = tweet\_by\_date.sortByKey()

tweet\_by\_date\_sorted.coalesce(1).saveAsTextFile("tweet\_count\_by\_date")

# In ra 10 dòng đầu tiên

tweet\_by\_date\_sorted.take(5)

# (b) Đếm số tweet theo khung giờ

def extract\_hour(line):

fields = line.split(",")

if len(fields) < 2:

return None

hour = fields[1].split(" ")[1].split(":")[0] # Lấy giờ (HH)

return (hour, 1)

tweet\_by\_hour = rdd.map(extract\_hour).filter(lambda x: x is not None).reduceByKey(lambda a, b: a + b)

tweet\_by\_hour\_sorted = tweet\_by\_hour.sortByKey()

tweet\_by\_hour\_sorted.coalesce(1).saveAsTextFile("tweet\_count\_by\_hour")

# In ra 10 dòng đầu tiên

tweet\_by\_hour\_sorted.take(5)

# (c) Tìm khung giờ Elon Musk hay đăng tweet nhất

most\_active\_hour = tweet\_by\_hour.max(lambda x: x[1])

print(f"Khung giờ Elon Musk hay đăng tweet nhất: {most\_active\_hour[0]}h với {most\_active\_hour[1]} tweet")

# Dừng SparkSession

spark.stop()

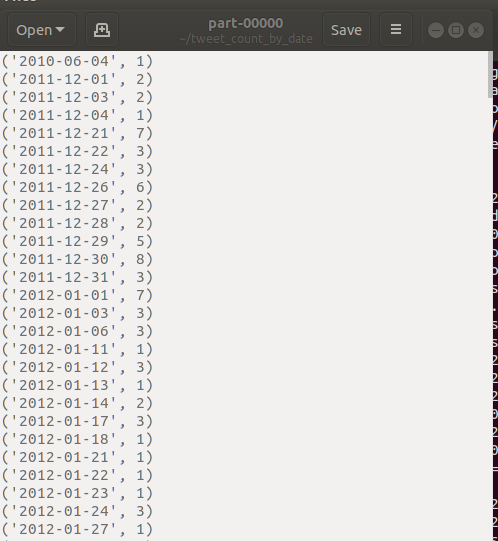
* + 1. Thực thi ứng dụng

Trong terminal, nhập lệnh sau để chạy:

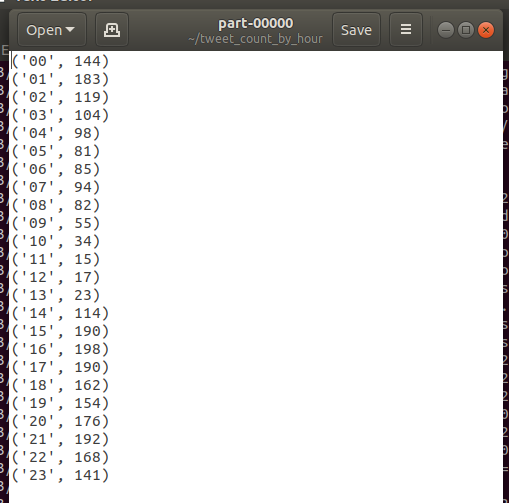
spark-submit tweet\_analysis.py



Ảnh mình hoạ dòng lệnh chạy ứng dụng cho Spark.



Ảnh mình hoạ kết quả về đếm số tweet theo ngày.



Ảnh mình hoạ kết quả về đếm số tweet theo giờ.



Kết quả khung giờ Elon Musk hay đăng tweet nhất.

1. Kết luận
   1. Đánh giá chung
      1. Những kết quả đạt được

* Hệ thống hoạt động ổn định trên Hadoop, hỗ trợ phân tích dữ liệu lớn.
* Ứng dụng Apache Spark giúp cải thiện tốc độ xử lý.
* Kết quả hỗ trợ phân tích xu hướng trên mạng xã hội.
  + 1. Một số hạn chế
* Hiệu suất xử lý dữ liệu lớn chưa được tối ưu.
* Một số lỗi trong dữ liệu chưa được xử lý hoàn toàn.
* Giao diện trực quan còn hạn chế, cần cải thiện.
* Việc triển khai Spark vẫn đang thử nghiệm, cần tối ưu thêm.
  1. Hướng phát triển
* Mở rộng phân tích theo chủ đề, cảm xúc của tweet.
* Xây dựng giao diện trực quan với biểu đồ, dashboard.
* Ứng dụng AI, Machine Learning để phân tích sâu hơn.
* So sánh hiệu suất giữa Hadoop MapReduce và Spark.
* Thử nghiệm Spark Streaming để xử lý dữ liệu thời gian thực.
* Nghiên cứu thêm các công nghệ như Apache Flink, Druid, Data Lake.

Tài liệu tham khảo

[1] Nguyễn Đình Hưng – Tài liệu môn xử lý dữ liệu lớn - 2025