Shape, square

Description automatically generated

ĐẠI HỌC BÁCH KHOA HÀ NỘI

**Trường Công Nghệ Thông Tin & Truyền Thông**

----- □ 🕮 □ -----

A picture containing icon

Description automatically generated

**Báo cáo môn Lưu trữ và Xử lý dữ liệu lớn**

***Đề Tài:* Lưu trữ, so sánh và xử lý dữ liệu chứng khoán Hoa Kỳ**

**Giảng viên hướng dẫn: TS. Nguyễn Hữu Đức**

Thực hiện:

|  |  |
| --- | --- |
| Lê Hoàng Long | 20232099M |

*Hà Nội, tháng 2 năm 2024*

[1. Lời tựa 2](#_Toc20123)

[2.Mục tiêu đề tài 3](#_Toc11926)

[3.Nội dung thực hiện 3](#_Toc22158)

[4. Hướng phát triển 10](#_Toc10008)

[Reference 11](#_Toc15256)

# **1. Lời tựa**

Công tác lưu trữ, xử lý dữ liệu lớn là một công tác đặc thù. Vấn đề đầu tiên được đặt ra là an toàn, an ninh và bảo mật dữ liệu. Sau khi đã đáp ứng được các yêu cầu nêu trên, thì có những vấn đề sau được đặt ra

* Làm thế nào để lưu trữ dữ liện lớn một cách hiệu quả nhằm đáp ứng 5 đặc điểm của dữ liệu lớn bao gồm: velocity, variety, value, veracity và volume
* Lầm thế nào để phân tích lượng dữ liệu lớn đủ nhanh, đủ chính xác để đưa ra những chỉ số, những thông tin hay sự thực bị ẩn giấu bên trong tầng tầng lớp lớp các bộ dữ liệu mà thông tin đan xen lẫn lộn vừa liên kết với nhau lại vừa bài trừ lẫn nhau
* Làm thế nào để lưu trữ những thông tin đã được xử lý một cách nhanh gọn, đơn giản và những thông tin đó có thể được truy cập bất kỳ lúc nào, với lượng truy cập không hạn chế nhằm đáp ứng yêu cầu truy xuất của các đơn vị nghiệp vụ
* Làm thế nào để trực quan hóa được thông tin đã được xử lý nhằm giúp dơn giản hóa quá trình tiếp nhận của các đơn vị liên quan

Những yêu cầu trên có thể được đáp ứng phần nào thông qua quy trình xử lý như sau

* Thông tin thu thập sẽ được lưu trữ trên cụm máy chủ Hadoop HDFS với khả năng chịu tải lớn và chịu lỗi cao
* Thông tin sẽ được xử lý bằng Spark nhằm tăng tốc quá trình xử lý, tránh nút thắt cổ chai dữ liệu
* Thông tin đã được lọc sẽ được đẩy vào cụm máy chủ ElasticSearch nhằm đáp ứng yêu cầu lưu trữ và xử lý bằng Machine Learning
* Để trực quan hóa những thông tin thu được sau quá trình xử lý, Kibana được sử dụng

Dữ liệu sử dụng trong báo cáo được lấy từ trang web MarketWatch và là dữ liệu công khai

https://www.marketwatch.com/investing/stock/csv/download-data

## 2.Mục tiêu đề tài

Áp dụng những lý thuyết của môn học vào công việc thực tiễn.

## 3.Nội dung thực hiện

* 1. Cài đặt trên máy chủ Redhat 9
     1. Hadoop
        1. Đường dẫn đến bài viết hướng dẫn trên mạng: <https://tecadmin.net/install-hadoop-centos-8/>
        2. Các bước cụ thể
           1. Tạm thời ngắt hoạt động của SELinux:

sudo setenforce 0

* + - * 1. Cài đặt Java

sudo dnf install java-21-openjdk-devel

* + - * 1. Tạo người dùng hadoop để đáp ứng các yêu cầu bảo mật

sudo useradd hadoop

sudo password hadoop

* + - * 1. Truy cập vào tài khoản người dung hadoop thực hiện những việc sau

Thiết lập SSH ky-based authentication

ssh-keygen -t rsa

ssh-copy-id localhost

Cài đặt hadoop

Lấy file từ máy chủ Apache   
wget [http://apachemirror.wuchna.com/hadoop/common/hadoop-3.5.0/hadoop-3.5.0.tar.gz](http://apachemirror.wuchna.com/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz)

Giải nén file   
tar -xvzf hadoop-3.5.0.tar.gz

Đổi tên thư mục: mv hadoop-3.5.0 hadoop

Bổ sung thông tin vào ~/.bashrc

export JAVA\_HOME=/usr/lib/jvm/jdk/

export HADOOP\_HOME=/home/hadoop/hadoop

export HADOOP\_INSTALL=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export HADOOP\_YARN\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export PATH=$PATH:$HADOOP\_HOME/sbin:$HADOOP\_HOME/bin

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"

Cập nhật thông cho file $HADOOP\_HOME/etc/hadoop/hadoop-env.sh  
với nội dung

export JAVA\_HOME=/usr/lib/jvm/jdk/

Tạo thư mục chứa dữ liệu cho hadoop

mkdir -p ~/hadoopdata/hdfs/namenode  
mkdir -p ~/hadoopdata/hdfs/datanode

Cập nhật thông tin cho file $HADOOP\_HOME/etc/hadoop/core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

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

</property>

</configuration>

Cập nhật nội dung cho file $HADOOP\_HOME/etc/hadoop/hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.name.dir</name>

<value>file:///home/hadoop/hadoopdata/hdfs/namenode</value>

</property>

<property>

<name>dfs.data.dir</name>

<value>file:///home/hadoop/hadoopdata/hdfs/datanode</value>

</property>

</configuration>

Cập nhật nội dung cho file $HADOOP\_HOME/etc/hadoop/mapred-site.xml

<configuration>

<property>

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

<value>mapreduce\_shuffle</value>

</property>

</configuration>

Thực hiện chạy HDFS

hdfs namenode -format

start-dfs.sh

start-yarn.sh

jps

Thiết lập tường lửa

sudo firewall-cmd --permanent --add-port=9870/tcp

sudo firewall-cmd --permanent --add-port=8088/tcp

* + 1. Spark
       1. Tải spark từ trang web <https://spark.apache.org/downloads.html>
       2. Cài đặt notebook để sử dụng pyspark
          1. pip install notebook
       3. Giải nén và thiết lập thông tin trong ~/bashrc như sau

export SPARK\_HOME=/home/oracle/soft/spark-3.5.0-bin-hadoop3

export PATH=$PATH:$SPARK\_HOME/bin:$SPARK\_HOME/sbin

export PYSPARK\_PYTHON=/usr/bin/python3

export PYSPARK\_DRIVER\_PYTHON='jupyter'

export PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook --no-browser --port=8889'

* + 1. Elasticsearch
       1. Tải elasticsearch từ trang web: <https://www.elastic.co/elasticsearch>
       2. Giải nén và thiết lập thông tin trong ~/bashrc như sau

export JAVA\_HOME=/home/oracle/soft/jdk-21.0.2

export PATH=$PATH:JAVA\_HOME/bin

export ES\_HOME=/home/oracle/soft/elasticsearch-8.12.1

export PATH=$PATH:$ES\_HOME/bin

* + 1. Kibana
       1. Tải kibana từ trang web: <https://www.elastic.co/kibana>
       2. Giải nén và thiết lập thông tin trong ~/bashrc như sau

export KB\_HOME=/home/oracle/soft/kibana-8.12.1

export PATH=$PATH:$KB\_HOME/bin

* 1. Xử lý:
     1. Đưa dữ liệu vào hadoop thông qua lệnh

hadoop fs -put stock.csv /

* + 1. Xử lý trên Spark
       1. Tạo phiên làm việc với jupyter notebook trên pyspark thông qua lệnh

pyspark

* + - 1. Truy cập vào đường dẫn <http://localhost:8889>
      2. Thực thi các đoạn mã python sau

# read from HDFS

df = spark.read.csv('hdfs://localhost:9000/stock.csv')

# install the package

!pip install elasticsearch

# write to Elasticsearch

from elasticsearch import Elasticsearch

# CERT\_FINGERPRINT could be generated by using this command

# openssl s\_client -connect localhost:9200 -servername localhost -showcerts </dev/null 2>/dev/null | openssl x509 -fingerprint -sha256 -noout -in /dev/stdin

CERT\_FINGERPRINT = 'AF:83:21:CE:66:CA:87:6A:15:D0:48:99:ED:B3:10:E2:4B:9C:1B:2F:7C:E9:7B:42:00:83:36:39:12:3F:32:80'

# password could be reset and generate by using this command

# elasticsearch-reset-password -u elastic

PASSWORD='pass'

es\_obj = Elasticsearch("https://localhost:9200", \

ssl\_assert\_fingerprint=CERT\_FINGERPRINT, \

basic\_auth=('elastic', PASSWORD))

print(es\_obj.info()) # Verify the connection

from elasticsearch.helpers import bulk

bulk\_data = []

for i, row in idf.iterrows():

bulk\_data.append(

{

"\_index": "stock\_index",

"\_id": i,

"\_source": {

"Date": row["Date"],

"Open": row["Open"],

"High": row["High"],

"Low": row["Low"],

"Close": row["Close"],

"Volume": row["Volume"],

}

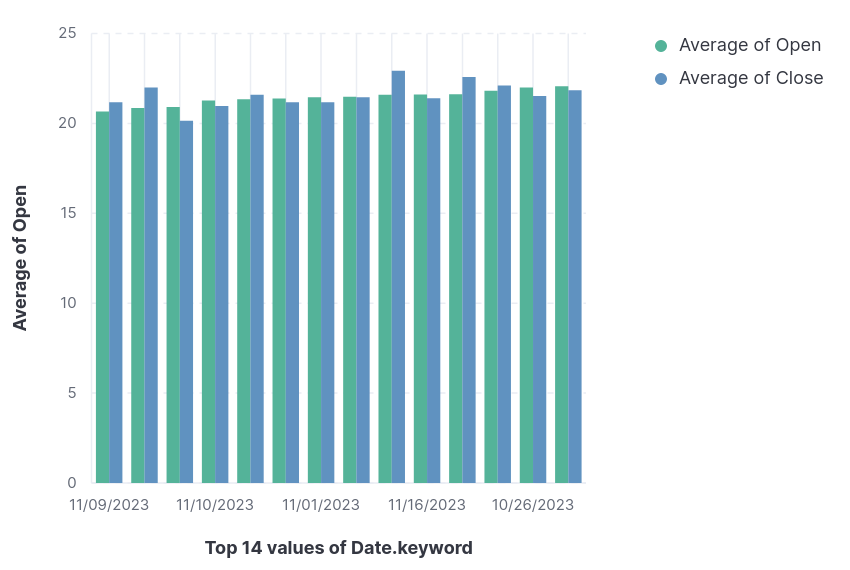
}

)

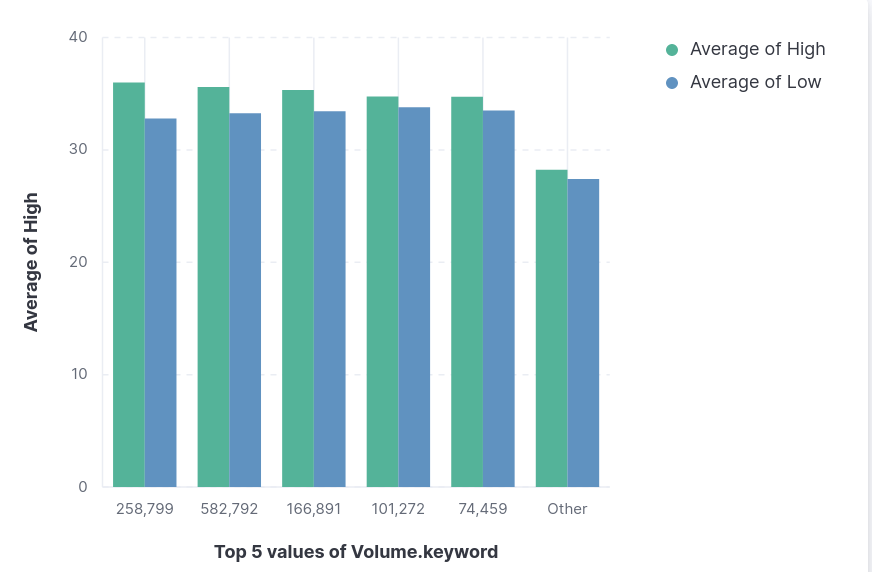
pass

bulk(es\_obj, bulk\_data)

* + 1. Tạo biểu đồ trực quan trên Kibana
       1. Biểu đồ đánh giá 14 ngày có trung bình giá cổ phiếu mở cửa cao nhất trong năm và thấy rằng các ngày này thường có giá trung bình cổ phiếu khi đóng cửa phiên giao dịch không khác biệt nhiều so với khi mở cửa. Điều này cho thấy sự trầm lắng của thị trường vốn



* + - 1. Biểu đồ đánh giá lượng cổ phiếu giao dịch cảu 5 ngày cao nhất trong năm và ta nhận thấy rằng lượng giao dịch của nhóm cố phiếu giá trị cao và nhóm cổ phiếu có giá trị thấp cùng tăng và cùng giảm. Điều đó cho thấy chưa có dấu hiệu đầu cơ lớn trên thị trường vào các mã cổ phiếu nhỏ lẻ



# **4. Hướng phát triển**

Trong tương lai, nhóm dự kiến sẽ tự động hóa quá trình truy xuất dữ liệu, áp dụng phân tích thông qua .ứng dụng học máy trên Kibana nhằm đưa ra những kết luận tốt hơn, định tính hơn dựa vào lượng dữ liệu lớn thu thập

# **Reference**

Source code: <https://github.com/adventurouscricket/stock-bigdata>

Elasticsearch: <https://www.elastic.co>