Cao Thị Ngọc Phụng

21110276

**BÁO CÁO BÀI TẬP 1 (TUẦN 2)**

# **1. Cài đặt OpenJDK**

# apt update

Lệnh này dùng để cập nhật thông tin về các gói phần mềm từ các nguồn đã cấu hình

A screen shot of a computer

Description automatically generated

# reboot

khởi động lại hệ thống

A screenshot of a computer

Description automatically generated

# apt install default-jdk

sử dụng trình quản lý gói APT (Advanced Package Tool) để tải và cài đặt JDK mặc định từ kho lưu trữ gói của hệ điều hành.

A black background with white text

Description automatically generated

A screenshot of a computer

Description automatically generated

# **2. Cài đặt SSH**

# apt-get install ssh

cài đặt dịch vụ SSH (Secure Shell)

A screen shot of a computer

Description automatically generated

# apt install openssh-server

cài đặt dịch vụ máy chủ SSH (OpenSSH)



# reboot

khởi động lại hệ thống



## **2.1 Cấu hình SSH**

# vim /etc/ssh/sshd\_config

- Tìm đoạn # PubkeyAuthentication yes.

Bỏ dấu # phía trước thành PubkeyAuthentication yes

Cho phép xác thực bằng khóa công khai (public key authentication) khi sử dụng SSH. Điều này thường được sử dụng để đảm bảo mức độ bảo mật cao khi đăng nhập vào máy chủ từ xa

A screen shot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated

- Tìm đoạn # PasswordAuthentication yes.

Bỏ dấu # phía trước thành ... PasswordAuthentication yes ...

Người dùng có thể đăng nhập bằng cách nhập mật khẩu thay vì sử dụng phương thức xác thực không dựa vào mật khẩu như khóa công khai.

A screenshot of a computer

Description automatically generated A screenshot of a computer screen

Description automatically generated

- Sau khi sửa thì nhấn phím ESC, nhập :wq để lưu và thoát khỏi vim.

- Khởi động lại SSH



# **3. Tạo user Hadoop**

- Tạo user hadoopngocphung để quản lý các permission cho đơn giản

# adduser hadoopngocphung

**A screenshot of a computer

Description automatically generated**

# **4. Cài đặt Hadoop 3.3.4**

- Chuyển qua hadoopuser

# su hadoopngocphung



- Chuyển qua thư mục /home/hadoopminhchau để download file:

# wget https://dlcdn.apache.org/hadoop/common/hadoo-3.3.2/hadoop-3.3.2.tar.gz

Hoặc

# wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz

Tuy nhiên sao khi em kiểm tra trên web thì hiện tại chỉ còn nhiêu đây phiên bản

A screenshot of a computer

Description automatically generated

Vì vậy em sẽ tải phiên bản 3.3.6

A screenshot of a computer screen

Description automatically generated

- Giải nén file

# tar -xzf hadoop-3.3.4.tar.gz



- Đổi tên thư mục giải nén thành hadoop cho dễ quản lý

# mv hadoop-3.3.4 hadoop



A screen shot of a computer

Description automatically generated

# **5. Thiết lập JAVA\_HOME**

# vim ~/hadoop/etc/hadoop/hadoop-env.sh

- Tìm đoạn export JAVA\_HOME=... sửa thành như sau:

# export JAVA\_HOME=/usr/lib/jvm/java-1.11.0-openjdk-amd64

Chỉ định đường dẫn đến một phiên bản cụ thể

A screenshot of a computer program

Description automatically generated A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

# **6. Standalone Operation**

Mặc định, Hadoop được cấu hình chạy ở chế độ không phân tán như là một tiến trình đơn Java. Ví dụ sau copy tất cả file .xml trong thư mục /etc/hadoop vào thư mục input, sau đó tìm và hiển thị mọi kết quả phù hợp với biểu thức chính quy đã cho.

# mkdir input

Tạo thư mục input

A screen shot of a computer

Description automatically generated

# cp hadoop/etc/hadoop/\*.xml input

Sao chép tất cả các tệp tin có đuôi là ".xml" từ thư mục "hadoop/etc/hadoop/" và đặt chúng vào thư mục "input".

A screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

# hadoop/bin/hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.4.jar grep input output 'dfs[a-z.]+'

Được sử dụng để chạy một ứng dụng MapReduce trong Apache Hadoop

A screenshot of a computer program

Description automatically generated

# cat output/\*

hiển thị nội dung của tất cả các tệp tin trong thư mục "output"



# **7. Pseudo-Distributed Operation**

Hadoop cũng có thể chạy trên một node đơn ở chế độ giả phân tán, trong đó mỗi daemon Hadoop chạy trên một tiến trình Java riêng biệt.

## **7.1 Cài đặt ssh key**

- Tạo ssh key

# ssh-keygen -t rsa -P ""

- Nhấn Enter để chấp nhận giá trị mặc định

**A computer screen with white text

Description automatically generated**

# cat /home/hadoopngocphung/.ssh/id\_rsa.pub >> /home/hadoopngocphung/.ssh/authorized\_keys

Thêm nội dung của tệp tin "id\_rsa.pub" vào cuối tệp tin "authorized\_keys"



# chmod 600 /home/hadoopngocphung/.ssh/authorized\_keys

Thay đổi quyền truy cập của tệp tin "authorized\_keys" trong thư mục .ssh của người dùng "hadoopngocphung"



## **7.2 Cấu hình file core-site.xml**

$ vim ~/hadoop/etc/hadoop/core-site.xml

A screenshot of a computer

Description automatically generated

## **7.3 Cấu hình file hdfs-site.xml**

$ vim ~/hadoop/etc/hadoop/hdfs-site.xml

A screenshot of a computer

Description automatically generated

## **7.4 Format hệ thống (chạy 1 lần duy nhất)**

$ hadoop/bin/hdfs namenode -format

Định dạng (format) NameNode trong hệ thống tệp phân tán Hadoop Distributed File System (HDFS)

A screenshot of a computer

Description automatically generated

## **7.5 Start NameNode daemon và DataNode daemon**

$ hadoop/sbin/start-dfs.sh

Bắt đầu các thành phần của hệ thống tệp phân tán Hadoop Distributed File System (HDFS)

**Kiểm tra các daemon đang chạy**

A screenshot of a computer screen

Description automatically generated

$ jps

Hiển thị danh sách các tiến trình Java hiện đang chạy trên máy tính

A black screen with white text

Description automatically generated

**Kiểm tra các node còn hoạt động**

$ ~/hadoop/bin/hdfs dfsadmin -report  
Để xem báo cáo về trạng thái của Hadoop Distributed File System (HDFS) từ một thành phần quản trị hệ thống.

A screenshot of a computer

Description automatically generated

$ telnet 127.0.0.1 9866

Thử nghiệm kết nối Telnet tới máy chủ đang lắng nghe trên cổng 9866 trên địa chỉ IP 127.0.0.1

A screen shot of a computer

Description automatically generated

**Chạy lại ví dụ trên, tạo thư mục trong HDFS để thực thi tác vụ MapReduce**

$ ~/hadoop/bin/hdfs dfs -mkdir /user

$ ~/hadoop/bin/hdfs dfs -mkdir /user/hadoopngocphung

$ ~/hadoop/bin/hdfs dfs -mkdir /user/hadoopngocphung/input



**Copy các file .xml vào hệ thống file phân tán**

$ ~/hadoop/bin/hdfs dfs -put hadoop/etc/hadoop/\*.xml /user/hadoopngocphung/input

Sao chép các tệp tin có đuôi là ".xml" từ thư mục "hadoop/etc/hadoop/" vào thư mục "/user/hadoopngocphung/input

****

**Hiển thị kết quả phù hợp với biểu thức chính quy**

$ ~/hadoop/bin/hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar grep input output ‘dfs[a-z.]+’

Chạy một ứng dụng MapReduce trong Apache Hadoop

A screenshot of a computer program

Description automatically generated

**Copy kết quả từ hệ thống file phân tán ra thư mục bên ngoài**

$ ~/hadoop/bin/hdfs dfs -get output/ output

Sao chép dữ liệu từ thư mục "output" trong Hadoop Distributed File System (HDFS) vào thư mục cục bộ có tên là "output"

$ cat output/\*

A screen shot of a computer

Description automatically generated

**Khi muốn dừng các daemon thì chạy lệnh sau**

$ hadoop/sbin/stop-dfs.sh

YARN trên Single Node

- Chúng ta có thể chạy các tác vụ MapReduce trên YARN ở chế độ giả phân tán bằng cách thiết lập vài thông số và chạy thêm các daemon ResourceManager và NodeManager để quản lý tài nguyên của Cluster.

## **7.6 Cấu hình file .bashrc**

$ vim ~/.bashrc

- Thêm vào cuối file .bashrc nội dung như sau:

export JAVA\_HOME=/usr/lib/jvm/java-1.11.0-openjdk-amd64

export HADOOP\_HOME=/home/hadoopngocphung/hadoop

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

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

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop

export HADOOP\_YARN\_HOME=$HADOOP\_HOME

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

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

A screenshot of a computer program

Description automatically generated

**- Soucre file .bashrc**

# source ~/.bashrc

## **7.7 Cấu hình file mapred-site.xml**

$ vim hadoop/etc/hadoop/mapred-site.xml

A screenshot of a computer program

Description automatically generated

## **7.8 Cấu hình file yarn-site.xml**

A screenshot of a computer

Description automatically generated

## **7.9 Chạy lại các daemon**

$ hadoop/sbin/start-dfs.sh

A screen shot of a computer

Description automatically generated

$ hadoop/sbin/start-yarn.sh

A screenshot of a computer

Description automatically generated

A black screen with white text

Description automatically generated

## **7.10 Chạy lại ví dụ demo trên, xóa các thư mục và tập tin đã tạo lúc trước (nếu có)**

$ rm -rf output



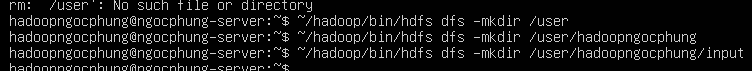
$ ~/hadoop/bin/hdfs dfs -rm -r /user

**Tạo lại các thư mục cần thiết**

$ ~/hadoop/bin/hdfs dfs -mkdir /user

$ ~/hadoop/bin/hdfs dfs -mkdir /user/hadoopngocphung

$ ~/hadoop/bin/hdfs dfs -mkdir /user/hadoopngocphung/input



**Copy các file .xml vào hệ thống file phân tán**

$ ~/hadoop/bin/hdfs dfs -put hadoop/etc/hadoop/\*.xml /user/hadoopngocphung/input



**Hiển thị kết quả phù hợp với biểu thức chính quy**

$ ~/hadoop/bin/hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.jar grep input output ‘dfs[a-z.]+’

A computer screen shot of a computer program

Description automatically generated

**Copy kết quả từ hệ thống file phân tán ra thư mục bên ngoài**

$ ~/hadoop/bin/hdfs dfs -get output/ output

$ cat output/\*

A screen shot of a computer

Description automatically generated

**Khi muốn dừng các daemon thì chạy lệnh sau**

$ hadoop/sbin/stop-dfs.sh

$ hadoop/sbin/stop-yarn.sh

A screenshot of a computer program

Description automatically generated