**CÀI ĐẶT HADOOP CLUSTER**

# 1. Setup máy ảo

Đổi tên mật khẩu cho root

|  |
| --- |
| sudo passwd |



Sau đó khởi động lại server

|  |
| --- |
| reboot |



## Java

### 1.1.1 Tải java

|  |
| --- |
| apt update |
| reboot |
| apt install openjdk-8-jdk |



## SSH

### 1.2.1 Tải SSH

|  |
| --- |
| apt-get install ssh |
| apt install openssh-server |
| reboot |

A screen shot of a computer

Description automatically generated



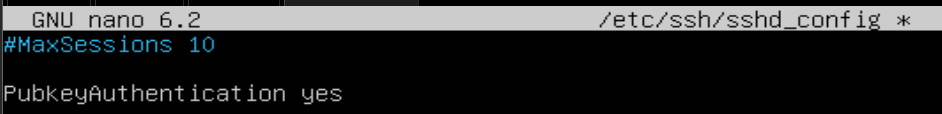


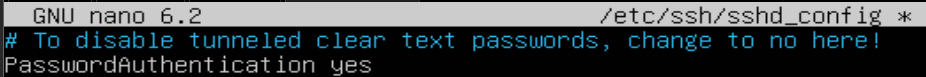
### 1.2.1 Cấu hình SSH

|  |
| --- |
| nano /etc/ssh/sshd\_config |

Tìm đoạn # PubkeyAuthentication yes và # PasswordAuthentication yes.

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





Khởi động lại SSH

|  |
| --- |
| service sshd restart |



# 2. Hadoop

## 2.1 Tạo user hadoopngocphung

|  |
| --- |
| adduser hadoopngocphung |

A screenshot of a computer program

Description automatically generated

## 2.2 Cài đặt Hadoop 3.3.6

|  |
| --- |
| su hadoopngocphung |
| wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz |
| tar -xzf hadoop-3.3.6.tar.gz |
| mv hadoop-3.3.6 hadoop |

A screen shot of a computer

Description automatically generated

## 2.3 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 |

A screen shot of a computer

Description automatically generated

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

|  |
| --- |
| cp hadoop/etc/hadoop/\*.xml input |

A screen shot of a computer

Description automatically generated

|  |
| --- |
| hadoop/bin/hadoop jar hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.6.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"



## 2.4 Cài đặt ssh key

- Tạo ssh key

|  |
| --- |
| ssh-keygen -t rsa -P "" |

**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"



# 3. Cấu hình các file cho Hadoop

## 3.1 File .bashrc

|  |
| --- |
| nano ~/.bashrc |

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

|  |
| --- |
| export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-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 computer screen shot of a program

Description automatically generated

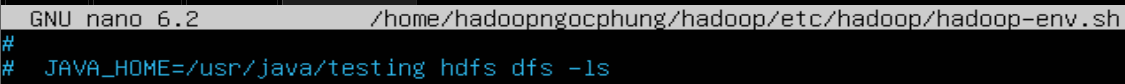
|  |
| --- |
| source ~/.bashrc |

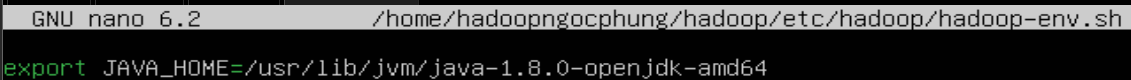
## 3.2 File hadoop-env.sh

|  |
| --- |
| nano ~/hadoop/etc/hadoop/hadoop-env.sh |

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

|  |
| --- |
| export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64 |





## 3.3 File core-site.xml

|  |
| --- |
| nano ~/hadoop/etc/hadoop/core-site.xml |

- Cấu hình lại thông tin như sau:

|  |
| --- |
| <configuration>  <property>  <name>hadoop.tmp.dir </name>  <value>/home/hadoopngocphung/tmp</value>  </property>  <property>  <name>fs.defaultFS</name>  <value>hdfs://ngocphung-master:9000</value>  </property>  </configuration> |

A computer screen with yellow text

Description automatically generated

## 3.4 File mapred-site (chỉ cấu hình ở master)

|  |
| --- |
| nano ~/hadoop/etc/hadoop/mapred-site.xml |
| <property>  <name>mapreduce.application.classpath</name> <value>$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/\*:$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/lib/\*</value>  </property>  <property>  <name>mapreduce.jobtracker.address</name>  <value>ngocphung-master:9001</value>  </property>  <property>  <name>mapreduce.framework.name</name>  <value>yarn</value>  </property>  <property>  <name>yarn.app.mapreduce.am.env</name>  <value>HADOOP\_MAPRED\_HOME=/home/hadoopngocphung/hadoop</value>  </property>  <property>  <name>mapreduce.map.env</name>  <value>HADOOP\_MAPRED\_HOME=/home/hadoopngocphung/hadoop</value>  </property>  <property>  <name>mapreduce.reduce.env</name>  <value>HADOOP\_MAPRED\_HOME=/home/hadoopngocphung/hadoop</value>  </property> |

A screenshot of a computer screen

Description automatically generated

## 3.5 File hdfs-site.xml

|  |
| --- |
| nano ~/hadoop/etc/hadoop/hdfs-site.xml |

- Chỉnh sửa lại thông tin cấu hình như sau:

|  |
| --- |
| <property>  <name>dfs.replication</name>  <value>2</value>  </property>  <property>  <name>dfs.namenode.name.dir</name> <value>/home/hadoopngocphung/hadoop/hadoop\_data/hdfs/namenode</value>  </property>  <property>  <name>dfs.datanode.data.dir</name> <value>/home/hadoopngocphung/hadoop/hadoop\_data/hdfs/datanode</value>  </property> |

A screenshot of a computer

Description automatically generated

## 3.6 File yarn-site.xml

|  |
| --- |
| nano ~/hadoop/etc/hadoop/yarn-site.xml |

Chỉnh sửa lại thông tin cấu hình như sau:

|  |
| --- |
| <property>  <name>yarn.nodemanager.aux-services</name>  <value>mapreduce\_shuffle</value>  </property>  <property>  <name>yarn.nodemanager.env-whitelist</name>  <value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PREPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_MAPRED\_HOME </value>  </property>  <property>  <name>yarn.resourcemanager.scheduler.address</name>  <value>ngocphung-master:9002</value>  </property>  <property>  <name>yarn.resourcemanager.address</name>  <value>ngocphung-master:9003</value>  </property>  <property>  <name>yarn.resourcemanager.webapp.address</name>  <value>ngocphung-master:9004</value>  </property>  <property>  <name>yarn.resourcemanager.resource-tracker.address</name>  <value>ngocphung-master:9005</value>  </property>  <property>  <name>yarn.resourcemanager.admin.address</name>  <value>ngocphung-master:9006</value>  </property> |

A screenshot of a computer program

Description automatically generated

# 4. Cấu hình network

## 4.1 Cấu hình cho MASTER

Cài đặt phần mềm Open vSwitch trên hệ thống

|  |
| --- |
| apt-get install openvswitch-switch |

### 4.1.1 Cài đặt hostname cho master

|  |
| --- |
| vim /etc/hostname |

- Trong file này sẽ xuất hiện hostname mặc định của máy, xóa đi và đổi thành ngocphung-master



### 4.1.2 Thiết lập IP host tĩnh cho MASTER

|  |
| --- |
| networkctl |

**A screen shot of a black background

Description automatically generated**

|  |
| --- |
| networkctl status |

**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

|  |
| --- |
| nano /etc/netplan/00-installer-config.yaml |
| sudo chmod 600 /etc/netplan/00-installer-config.yaml |
| network:  ethernets:  ens33:  dhcp4: false  dhcp6: false  addresses: [192.168.19.1/24]  routes:  - to: default  via: 192.168.19.2  nameservers:  addresses: [192.168.19.1, 8.8.8.8, 8.8.4.4]  version: 2 |

A screenshot of a computer

Description automatically generated

|  |
| --- |
| netplan apply |



- Hệ thống đã được cấu hình theo IP mới, để kiểm tra chạy 1 trong 2 lệnh sau

|  |
| --- |
| networkctl status |

A screenshot of a computer

Description automatically generated

### 4.1.3 Cấu hình host

|  |
| --- |
| vim /etc/hosts |

Điền nội dung, mỗi máy slave là một dòng

|  |
| --- |
| 192.168.19.1 ngocphung-master  192.168.19.3 ngocphung-slave1 |

A black and white screen with blue text

Description automatically generated

### 4.1.4 Liệt kê ra các máy slaves (chỉ thực hiện ở máy master)

|  |
| --- |
| su hadoopngocphung |
| nano ~/hadoop/etc/hadoop/workers |

Điền tên các máy slave1

****

****

## 4.2 Cấu hình cho SLAVE1

### 4.2.1 Tạo máy ngocphung-slave1

- Tắt máy Master.

- Copy Master ra, đổi tên thành Slave1

A screenshot of a black box

Description automatically generated

- Mở máy slave, chỉnh lại IP tĩnh và các thông số cho phù hợp: hosts, hostname…

### 4.2.2 Cài đặt hostname cho slave



|  |
| --- |
| vim /etc/hostname |

- Trong file này sẽ xuất hiện hostname mặc định của máy, xóa đi và đổi thành ngocphung-master

### 4.2.3 Thiết lập IP host tĩnh cho SLAVE1

|  |
| --- |
| nano /etc/netplan/00-installer-config.yaml |
| network:  ethernets:  ens33:  dhcp4: false  dhcp6: false  addresses: [192.168.19.3/24]  routes:  - to: default  via: 192.168.19.2  nameservers:  addresses: [192.168.19.1, 8.8.8.8, 8.8.4.4]  version: 2 |

A screenshot of a computer

Description automatically generated

|  |
| --- |
| netplan apply |



- Hệ thống đã được cấu hình theo IP mới, để kiểm tra chạy 1 trong 2 lệnh sau

|  |
| --- |
| networkctl status |

A screenshot of a computer

Description automatically generated

# 5. Cài đặt ssh key giữa các node

## 5.1 Share key

Thao tác này chỉ thực hiện trên master

- Đăng nhập với hadoopngocphung



- Tạo ssh key

|  |
| --- |
| ssh-keygen -t rsa -P "" |

A screenshot of a computer

Description automatically generated

- Nhấn Enter để chấp nhận giá trị mặc định

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



|  |
| --- |
| chmod 600 /home/hadoopngocphung/.ssh/authorized\_keys |



- Share ssh key giữa master - master

|  |
| --- |
| ssh-copy-id -i ~/.ssh/authorized\_keys ngocphung-master |

- Share ssh key giữa master - slave

|  |
| --- |
| # ssh-copy-id -i ~/.ssh/authorized\_keys ngocphung-slave1 |

A screenshot of a computer program

Description automatically generated

## 5.2 Test kết nối ssh

- Test kết nối tới master

|  |
| --- |
| ssh hadoopngocphung@ngocphung-master |

A screenshot of a computer

Description automatically generated

- Đăng xuất

|  |
| --- |
| logout |

- Test kết nối tới slave

A screenshot of a computer

Description automatically generated

|  |
| --- |
| ssh hadoopngocphung@ngocphung-slave1 |

# 6. Format namenode

Thao tác này chỉ thực hiện trên master và chỉ làm 1 lần.

Cập nhật lại các thông tin cấu hình của master

|  |
| --- |
| hadoop/bin/hdfs namenode -format |



# 7. Tạo test

- Ra thư mục gốc, tạo file test

|  |
| --- |
| nano test.sh |

Chứa các nội dung sau

|  |
| --- |
| #!/bin/bash  # test the hadoop cluster by running wordcount  # create input files  mkdir input  echo "Hello World" >input/file1.txt  echo "Hello Hadoop" >input/file2.txt  # create input directory on HDFS  hadoop fs -mkdir -p input1  # put input files to HDFS  hdfs dfs -put ./input/\* input1  # run wordcount  hadoop jar $HADOOP\_HOME/share/hadoop/mapreduce/sources/hadoop-mapreduce-examples-3.3.6-sources.jar org.apache.hadoop.examples.WordCount input1 output1  # print the input files  echo -e "\ninput file1.txt:"  hdfs dfs -cat input1/file1.txt  echo -e "\ninput file2.txt:"  hdfs dfs -cat input1/file2.txt  # print the output of wordcount  echo -e "\nwordcount output:"  hdfs dfs -cat output1/part-r-00000 |

Cấp quyền và chạy thử

|  |
| --- |
| chmod +x test.sh |
| ./test.sh |

Nếu muốn chạy lại

|  |
| --- |
| rm -rf input |
| hadoop fs -rm -r input1 |
| hadoop fs -rm -r output1 |

A screenshot of a computer screen

Description automatically generatedA screen shot of a computer

Description automatically generated