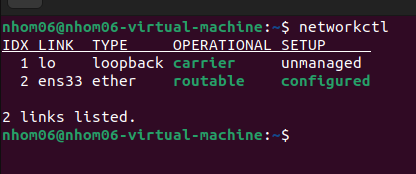
Cấu hình máy ảo:

* Hệ điều hành: Ubuntu 22.04
* Bô nhớ RAM: 4GB
* Số lõi xử lý (CPU): 2
* Bộ nhớ trong: 40GB

1. **Cài đặt Hadoop Eco System**
   1. **Thiết lập IP tĩnh cho master**

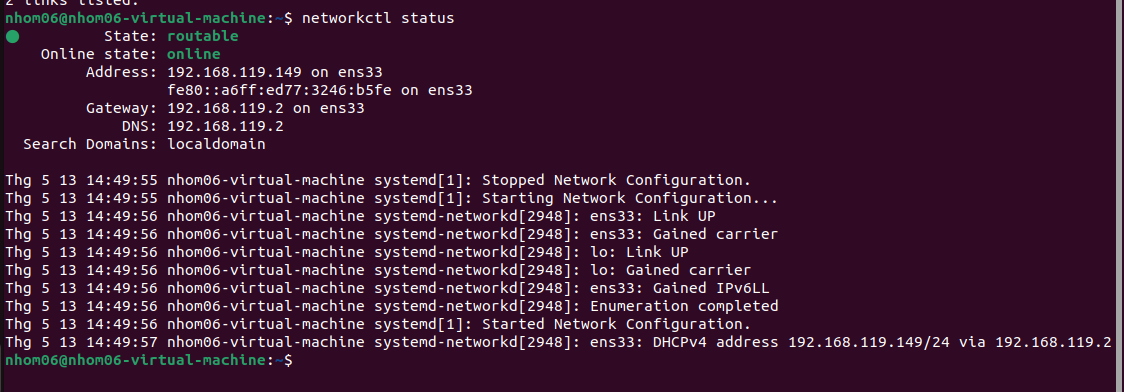
* Kiểm tra các thiết bị mạng

$ networkctl



* In trạng thái của từng địa chỉ IP trên hệ thống

$ networkctl status



Nhấn phím q để thoát thông báo

* Cấu hình IP tĩnh

$ sudo vim /etc/netplan/00-installer-config.yaml

* Thêm vào các nội dung sau:

# This is the network config written by 'subiquity'

network:

ethernets:

ens33:

dhcp4: false

dhcp6: false

addresses: [192.168.119.4/24]

routes:

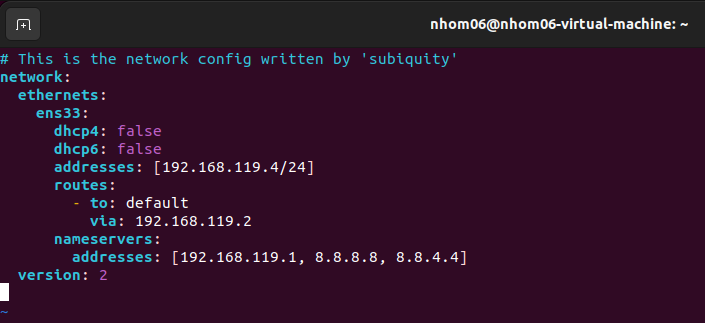
- to: default

via: 192.168.119.2

nameservers:

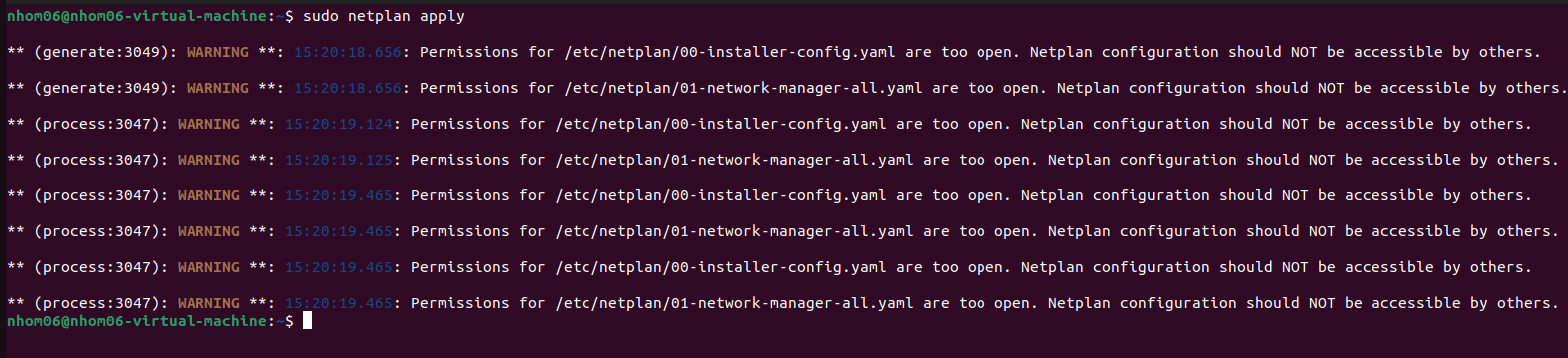
addresses: [192.168.119.1, 8.8.8.8, 8.8.4.4]

version: 2



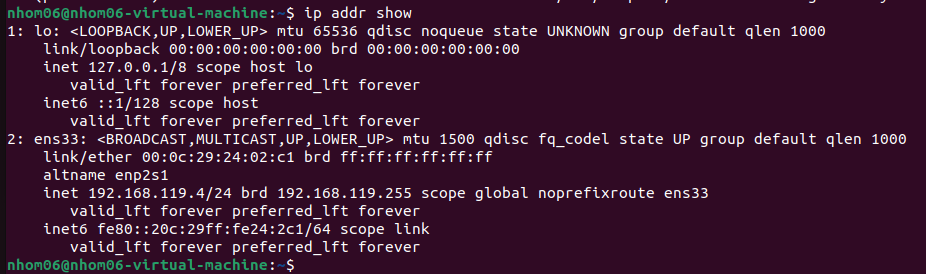
* Lưu file và chạy lệnh sau để lưu cấu hình mới

$ netplan apply



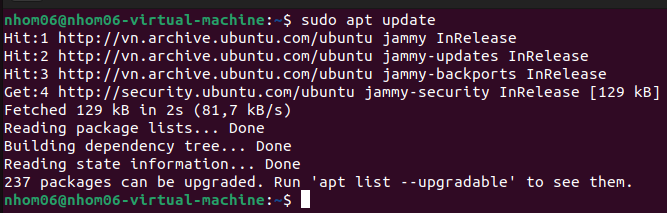
* Hệ thống đã được cấu hình theo IP mới, để kiểm tra chạy 1 trong 2 lệnh sau:

$ ip addr show

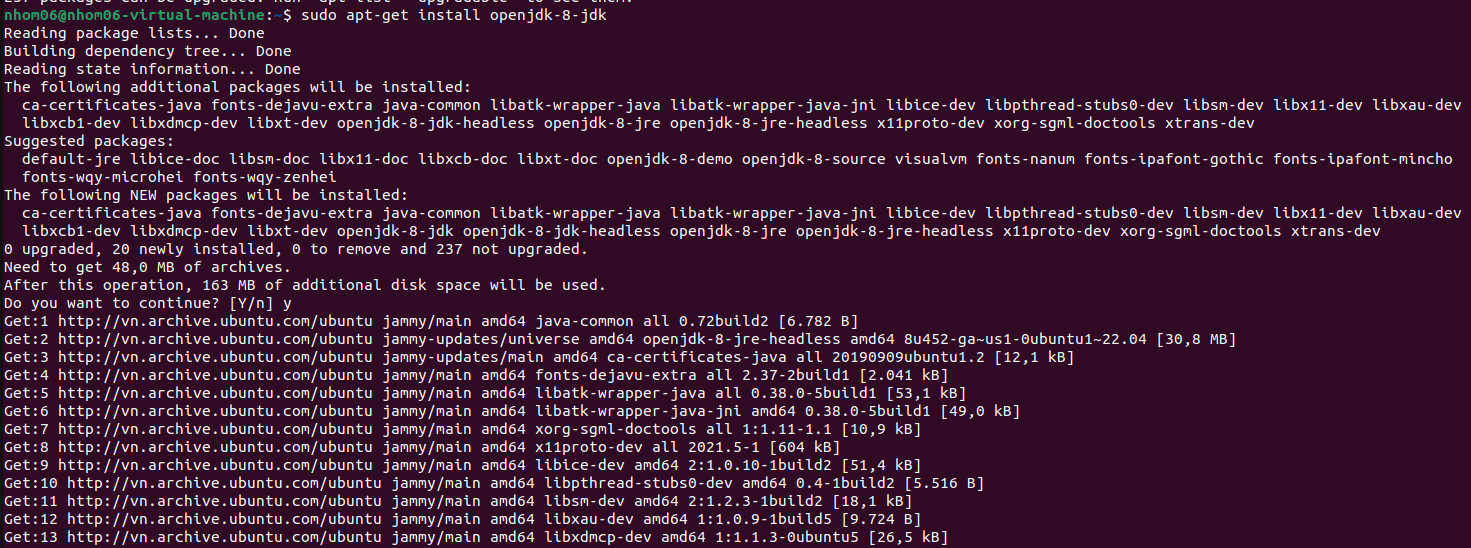


# **Cài đặt OpenJDK (JDK 8)**

$ sudo apt update

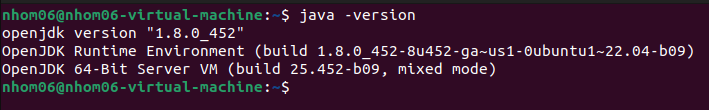


$ sudo apt-get install openjdk-8-jdk

****

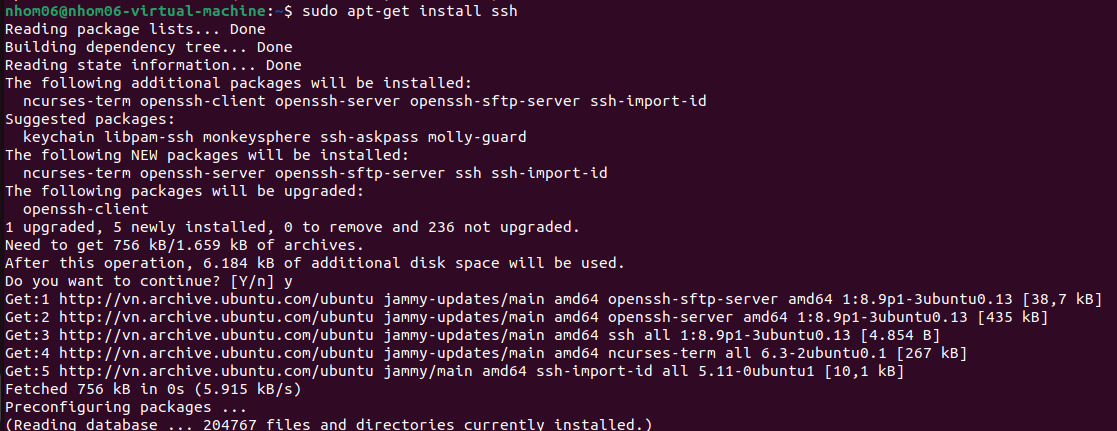
**- Kiểm tra phiên bản JDK**

**$** java -version

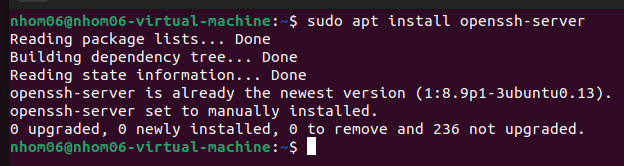
****

# **Cài đặt SSH**

$ sudo apt-get install ssh



$ sudo apt install openssh-server



* **Cấu hình SSH**

$ sudo vim /etc/ssh/sshd\_config

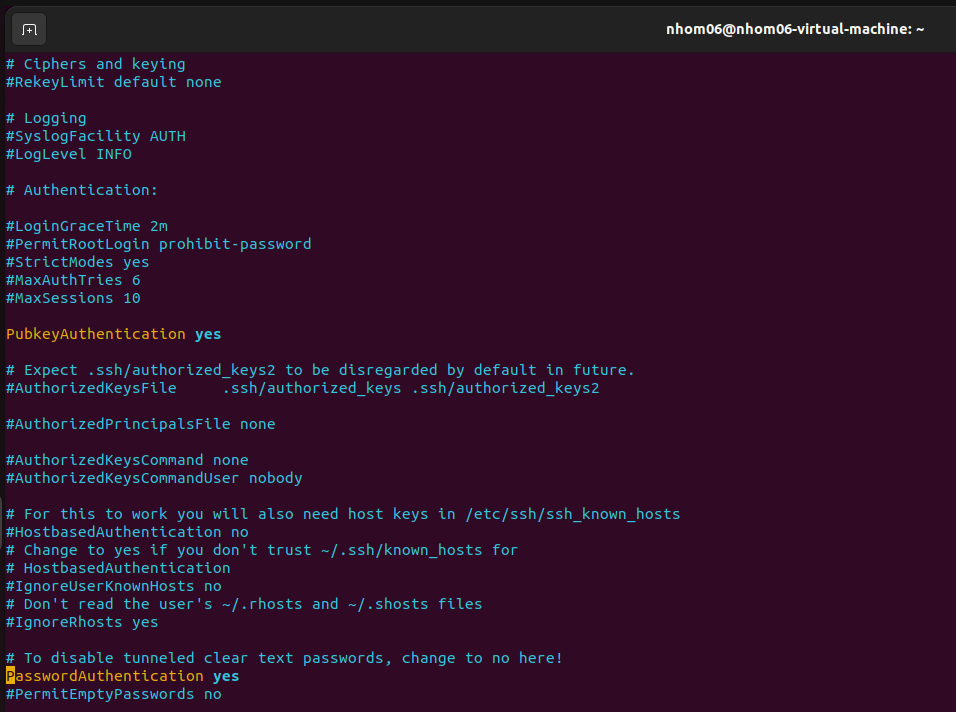
- Tìm đoạn $ PubkeyAuthentication yes. Bỏ dấu $ phía trước thành

... PubkeyAuthentication yes ...

- Tìm đoạn $ PasswordAuthentication yes. Bỏ dấu $ phía trước thành

... PasswordAuthentication yes ...

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



- Khởi động lại SSH

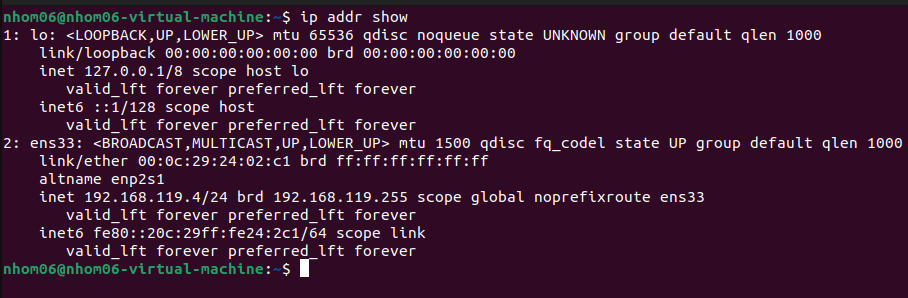
$ service sshd restart



* **Cấu hình host/hostname**

**Kiểm tra ip của các máy master, slave**

$ ip addr show

****

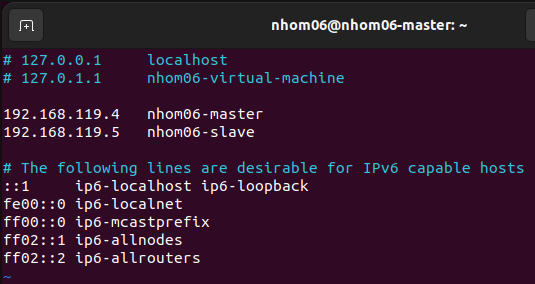
* **Cấu hình host**

$ sudo vim /etc/hosts

* Nhấn phím I để chuyển sang chế độ insert, bổ sung thêm 1 master và 1 slave:

192.168.119.4 nhom06-master

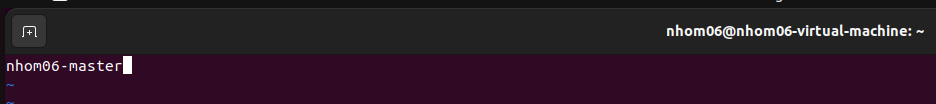
192.168.119.5 nhom06-slave



* **Cài đặt hostname cho master (thực hiện trên máy master)**

$ sudo vim /etc/hostname

* Trong file này sẽ xuất hiện hostname mặc định của máy, xóa đi và đổi thành nhom06-master



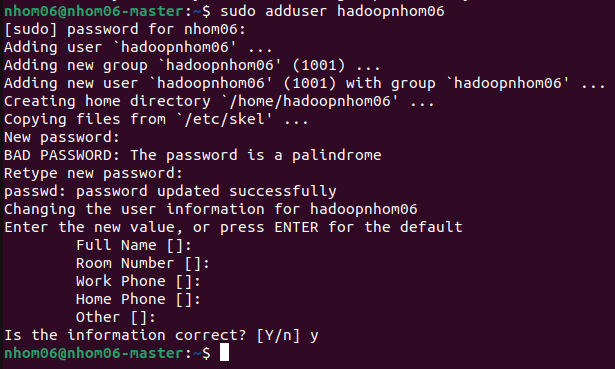
* Restart máy

$ reboot

* 1. **Tạo user hadoop**

- Tạo user hadoopnhom06 để quản lý các permission cho đơn giản

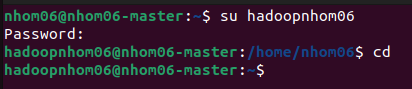
$ sudo adduser hadoopnhom06



# **Cài đặt Hadoop 3.3.6**

- Chuyển qua hadoopuser

$ su hadoopnhom06



- Chuyển qua thư mục /home/hadoopnhom06 để download file:

$ wget <https://archive.apache.org/dist/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz>



- Giải nén file

$ tar -xzf hadoop-3.3.6.tar.gz



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

$ mv hadoop-3.3.6 hadoop

****

* 1. **Cấu hình các thông số cho Hadoop**
* **File .bashrc**

$ sudo vim ~/.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/hadoopnhom06/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

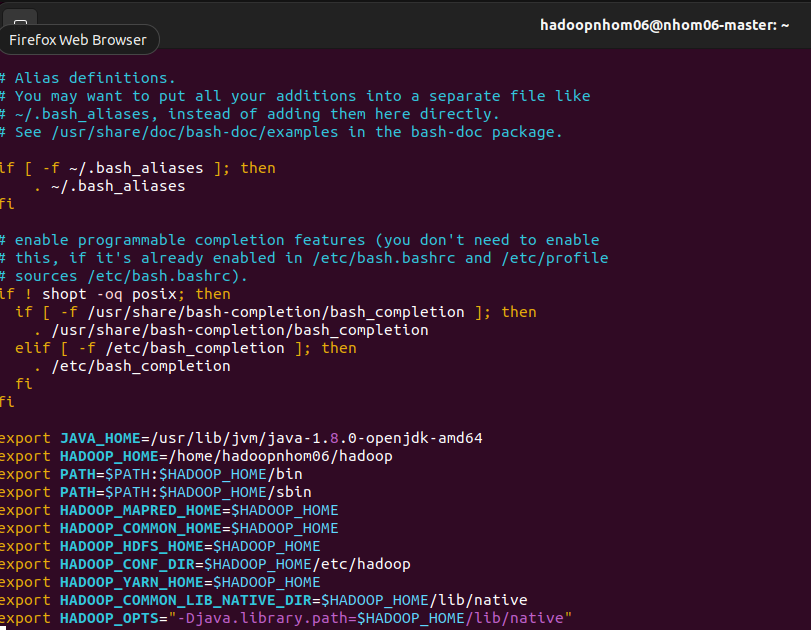
export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

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"



- Nhấn Esc, nhập :wq để lưu và thoát file.

- Soucre file .bashrc

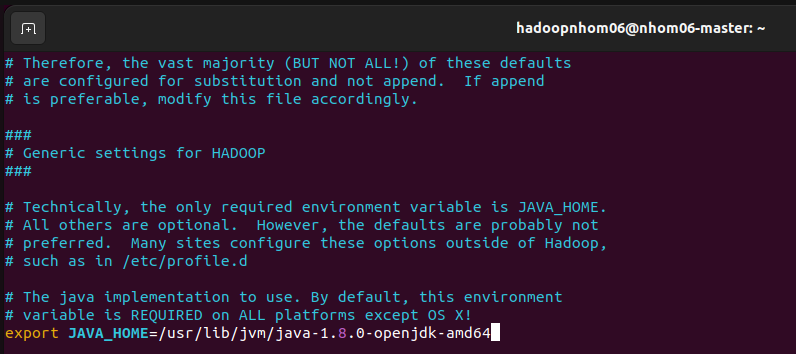
$ source ~/.bashrc

* **File Hadoop-env.sh**

$ vim ~/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



* **File core-site.xml**

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

<configuration>

<property>

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

<value>/home/hadoopnhom06/tmp</value>

<description>Temporary Directory.</description>

</property>

<property>

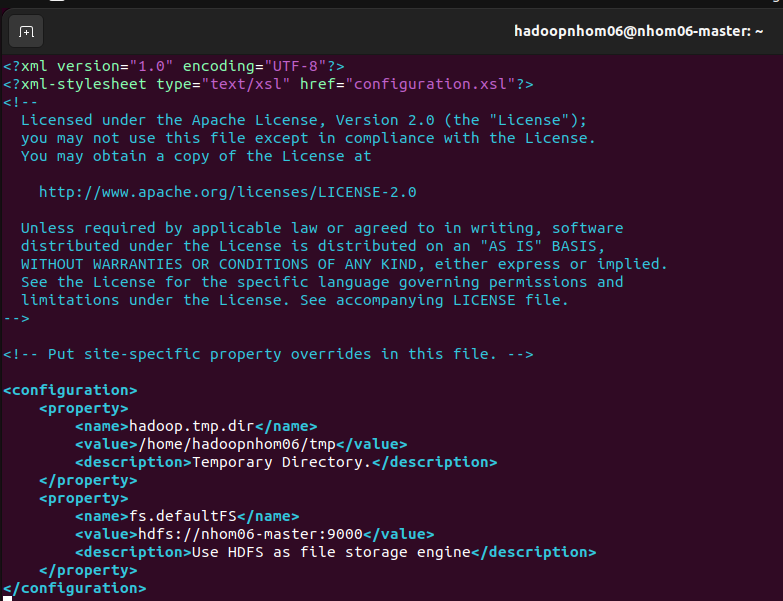
<name>fs.defaultFS</name>

<value>hdfs://nhom06-master:9000</value>

<description>Use HDFS as file storage engine</description>

</property>

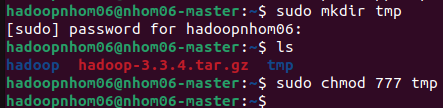
</configuration>



$ sudo mkdir tmp

$ ls

$ sudo chmod 777 tmp



* **File mapred-site (chỉ cấu hình ở master)**

$ cd ~/hadoop/etc/hadoop/

$ vim mapred-site.xml

<configuration>

<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>nhom06-master:9001</value>

<description>The host and port that the MapReduce job tracker runs at. If "local", then jobs are run in-process as a single map and reduce task.</description>

</property>

<property>

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

<value>yarn</value>

<description>The framework for running mapreduce jobs</description>

</property>

<property>

<name>yarn.app.mapreduce.am.env</name>

<value>HADOOP\_MAPRED\_HOME=/home/hadoopnhom06/hadoop</value>

</property>

<property>

<name>mapreduce.map.env</name>

<value>HADOOP\_MAPRED\_HOME=/home/hadoopnhom06/hadoop</value>

</property>

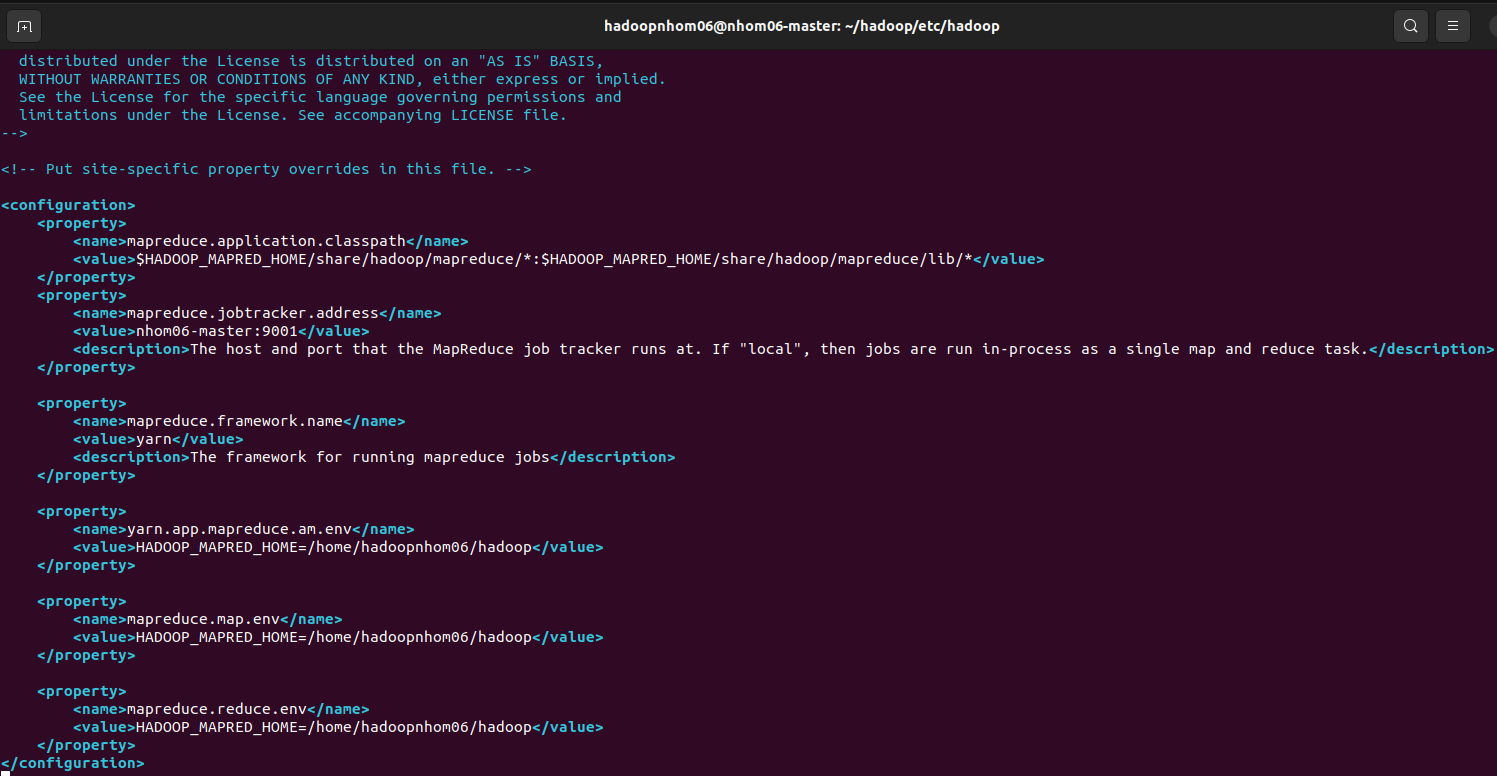
<property>

<name>mapreduce.reduce.env</name>

<value>HADOOP\_MAPRED\_HOME=/home/hadoopnhom06/hadoop</value>

</property>

</configuration>



* **File hdfs-site.xml**

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

<configuration>

<property>

<name>dfs.replication</name>

<value>2</value>

<description>Default block replication. The actual number of replications can be specified when the file is created. The default is used if replication is not specified in create time.</description>

</property>

<property>

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

<value>/home/hadoopnhom06/hadoop/hadoop\_data/hdfs/namenode</value>

<description>Determines where on the local filesystem the DFS name node should store the name table (fsimage). If this is a comma-delimited list of directories then the name table is replicated in all of the directories, for redundancy.</description>

</property>

<property>

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

<value>/home/hadoopnhom06/hadoop/hadoop\_data/hdfs/datanode</value>

<description>Determines where on the local filesystem the DFS data node should store its blocks. If this is a comma-delimited list of directories, then data will be stored in all named directories, typically on different devices. Directories that do not exist are ignored.</description>

</property>

</configuration>



* **File yarn-site.xml**

- Chuyển đến thư mục ~/hadoop/hadoop-yarn-project/hadoop-yarn/conf

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

<configuration>

<!-- Site specific YARN configuration properties -->

<configuration>

<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>nhom06-master:9002</value>

</property>

<property>

<name>yarn.resourcemanager.address</name>

<value>nhom06-master:9003</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.address</name>

<value>nhom06-master:9004</value>

</property>

<property>

<name>yarn.resourcemanager.resourcetracker.address</name>

<value>nhom06-master:9005</value>

</property>

<property>

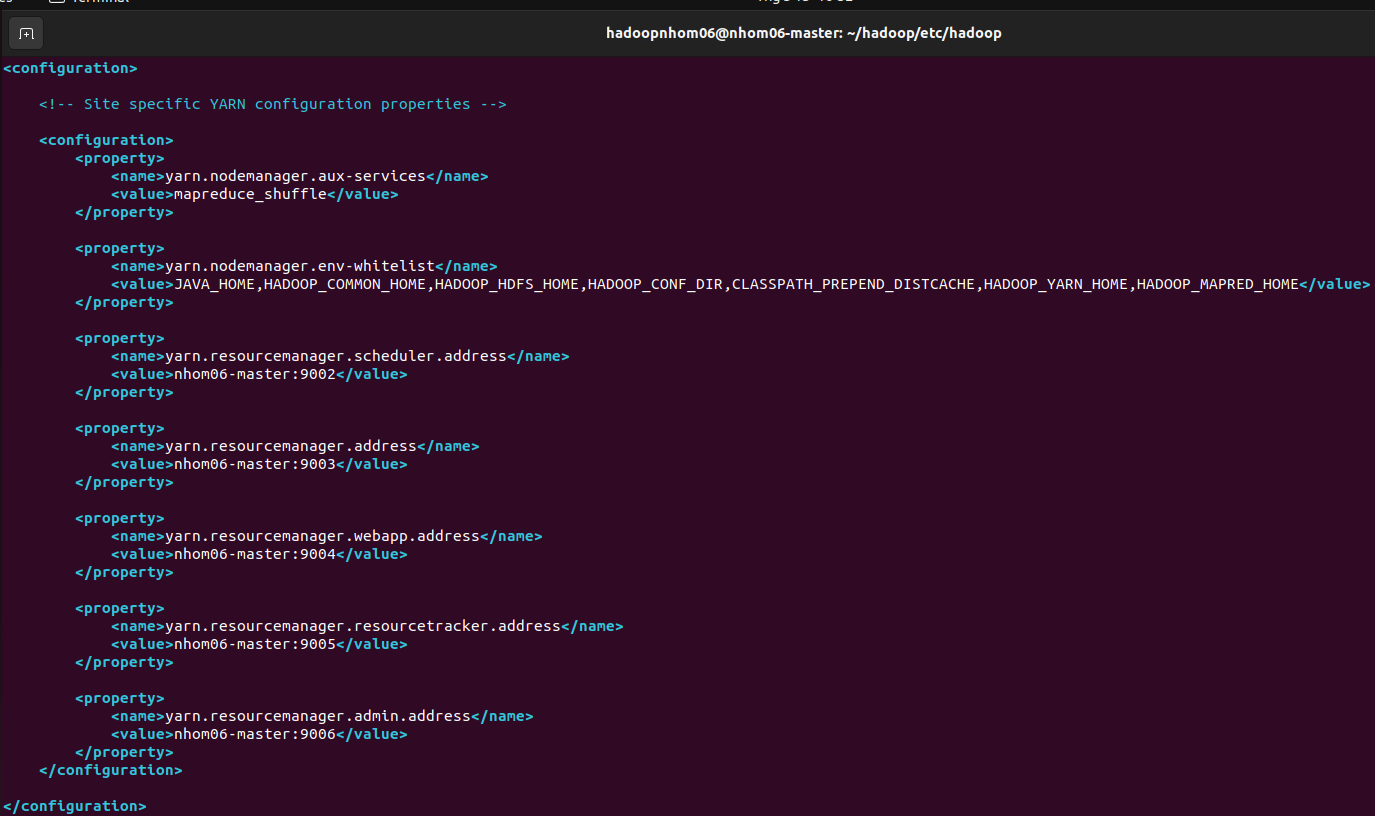
<name>yarn.resourcemanager.admin.address</name>

<value>nhom06-master:9006</value>

</property>

</configuration>

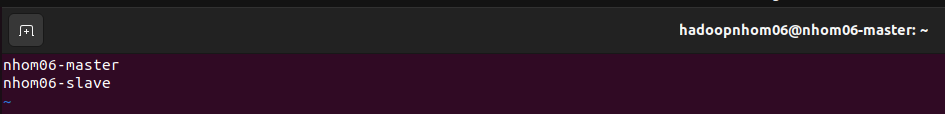
</configuration>



* 1. **Chỉ ra các máy slaves (chỉ cấu hình ở master)**

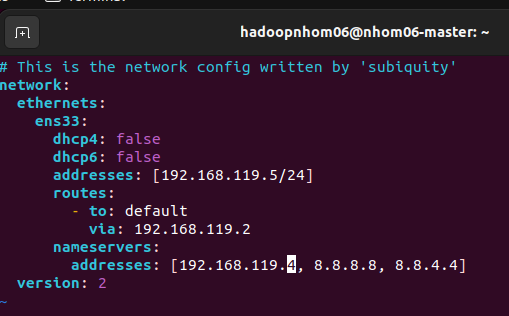
$ vim ~/hadoop/etc/hadoop/workers

Thêm hostname của các máy slave: mỗi máy slave đặt trên 1 dòng

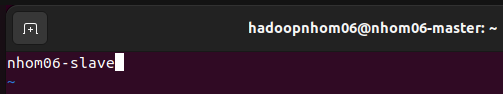


* 1. **Tạo máy nhom06-slave**
* Tắt máy master
* Copy master ra, đổi tên thành slave
* Mở máy slave, chỉnh lại IP tĩnh và các thông số cho phù hợp: hosts, hostname,….

**$** sudovim /etc/netplan/00-installer-config.yaml



$ sudo vim /etc/hostname



$ reboot

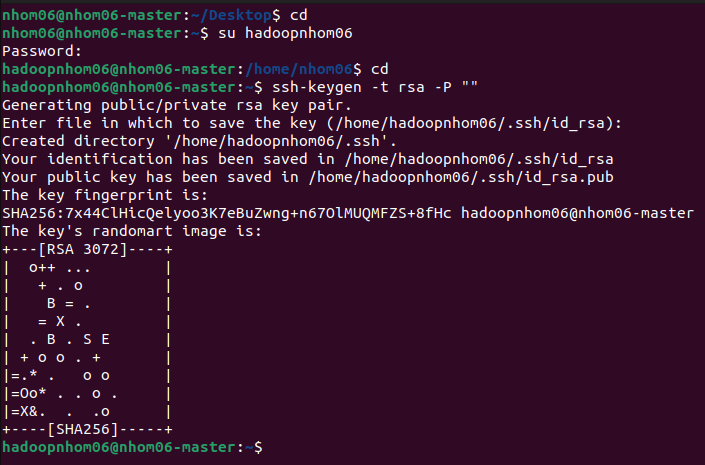
* 1. **Cài đặt ssh key giữa các node**

( Thao tác trên máy master)

* Khởi động máy master, đăng nhập Hadoop user
* Tạo ssh key

$ ssh-keygen -t rsa -P ""

* Nhấn Enter để chấp nhận giá trị mặc định

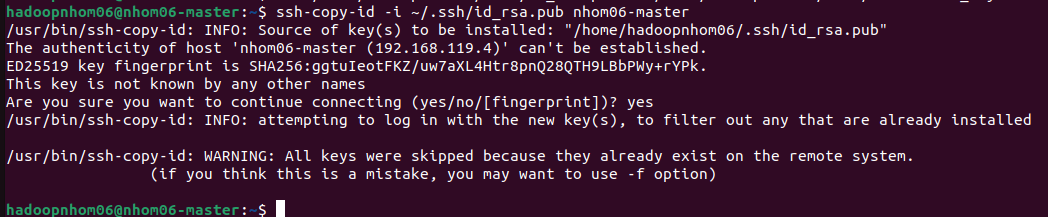


$ cat /home/hadoopnhom06/.ssh/id\_rsa.pub >> /home/hadoopnhom06/.ssh/authorized\_keys

$ sudo chmod 600 /home/hadoopnhom06/.ssh/authorized\_keys

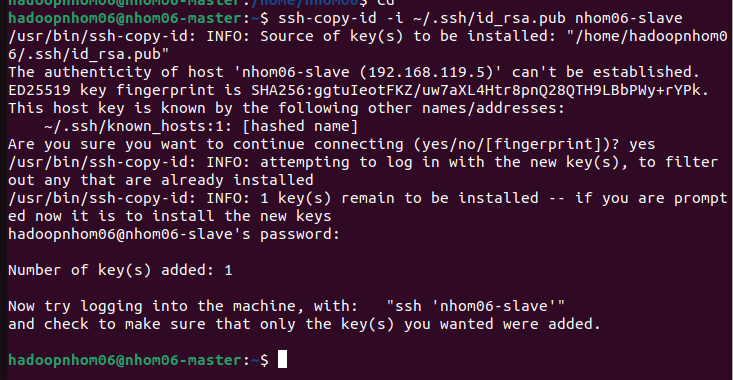
* Share ssh key giữa master – master

$ ssh-copy-id -i ~/.ssh/id\_rsa.pub nhom06-master



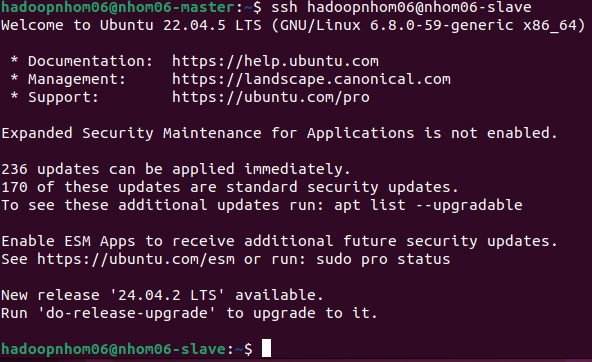
* Share ssh key giữa master – slave

$ ssh-copy-id -i ~/.ssh/id\_rsa.pub nhom06-slave



* 1. **Test kết nối ssh**
* Test kết nối tới master

$ ssh hadoopnhom06@nhom06-slave

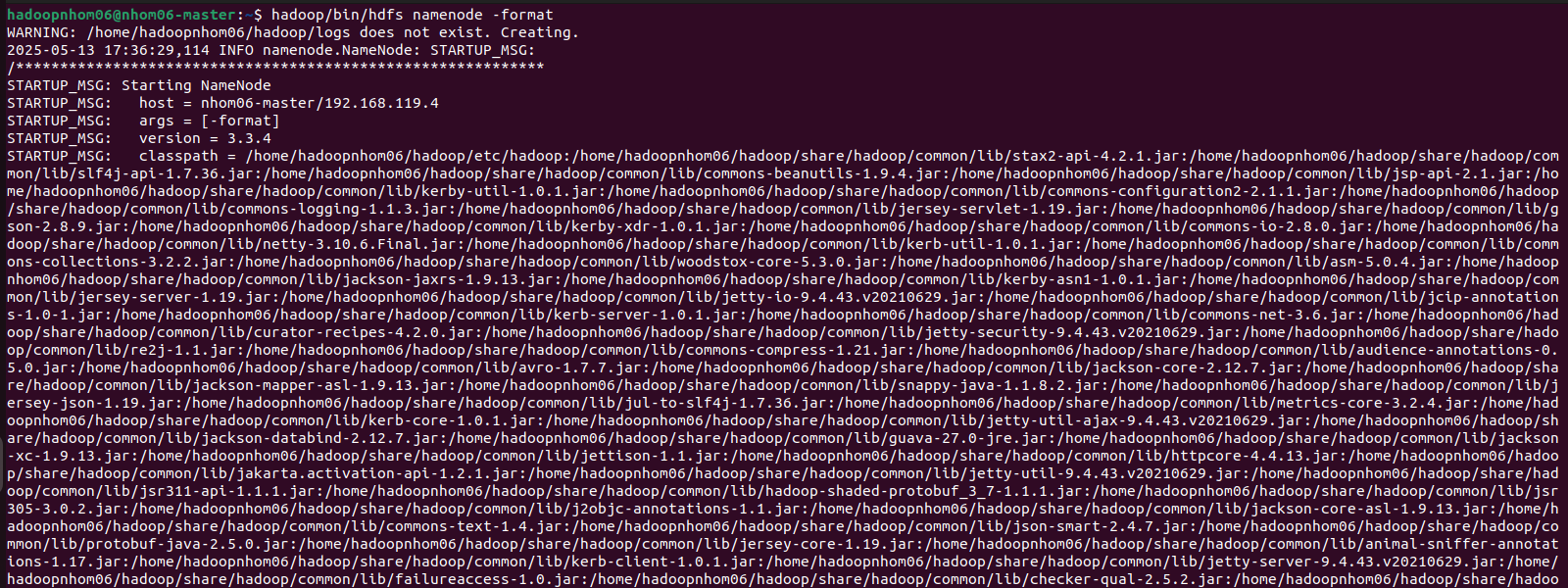


* Đăng xuất

$ logout

* 1. **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



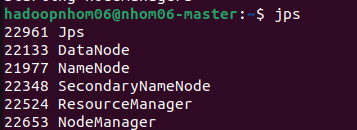
* 1. **Kiểm tra xem mọi thứ đã ổn**
* Trên master chúng ta chạy lệnh sau để khởi dộng các thành phần có trong Hadoop

**$** start-all.sh (chỉ cần chạy trên master, không cần chạy trên slave)

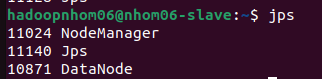
* Kiểm tra các thành phần có chạy đủ bằng lệnh sau

$ jps

* Máy master

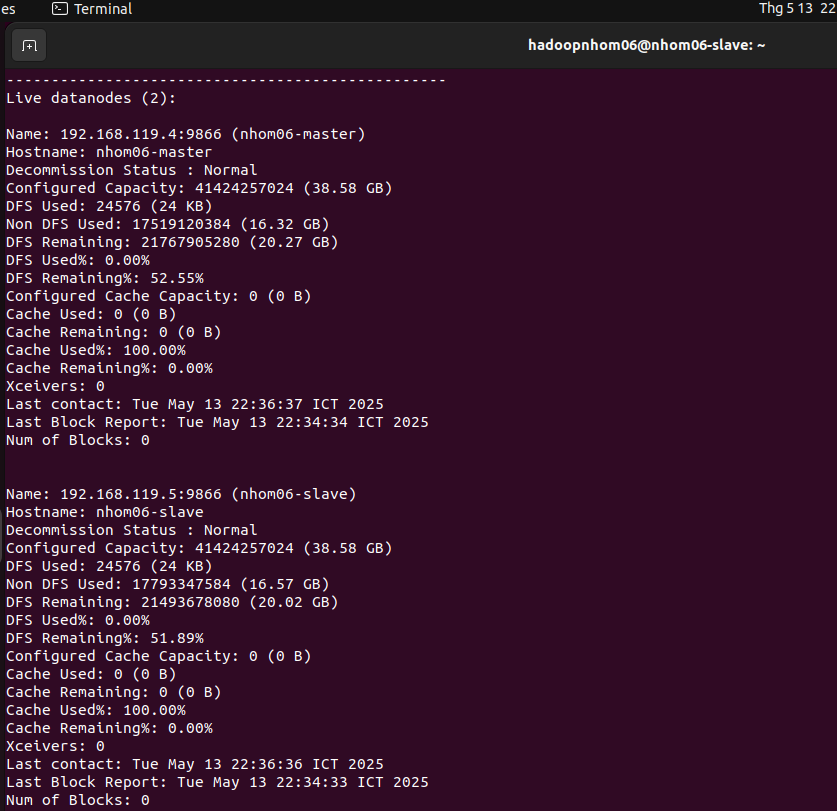


* Máy slave



* Kiểm tra các máy slave còn hoạt động hay không

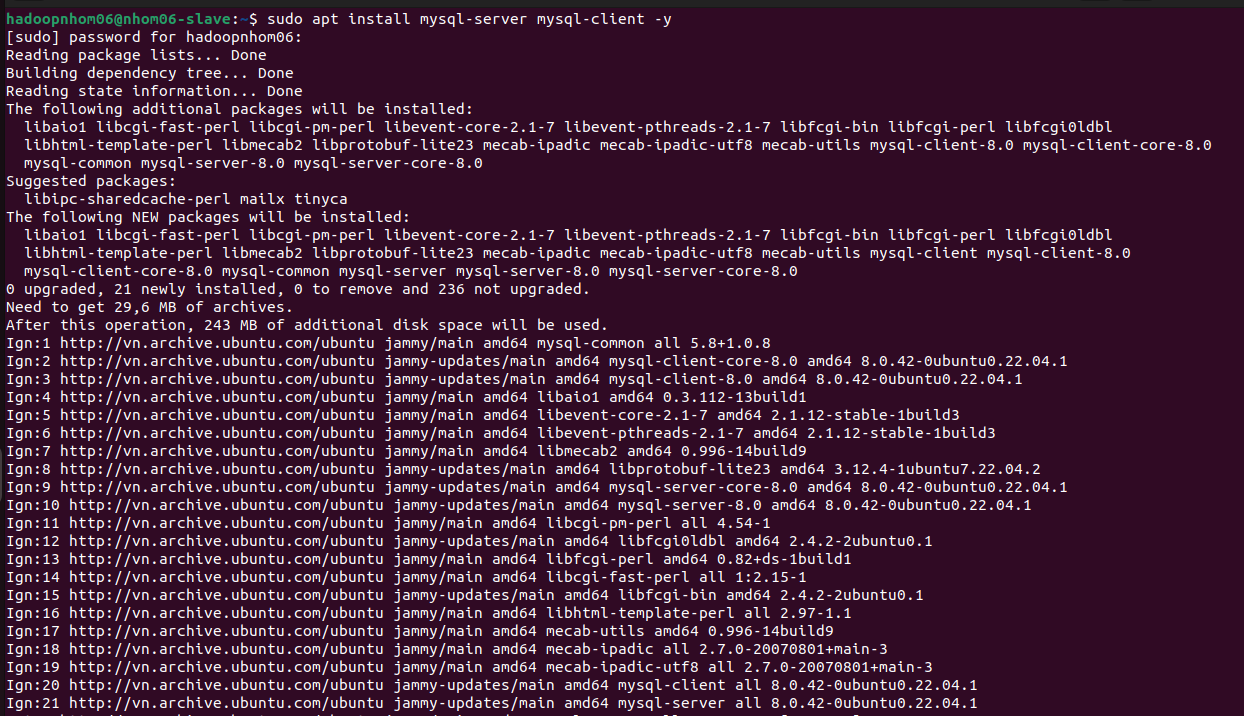
$ hdfs dfsadmin -report



1. **Cài đặt và cấu hình MySQL**

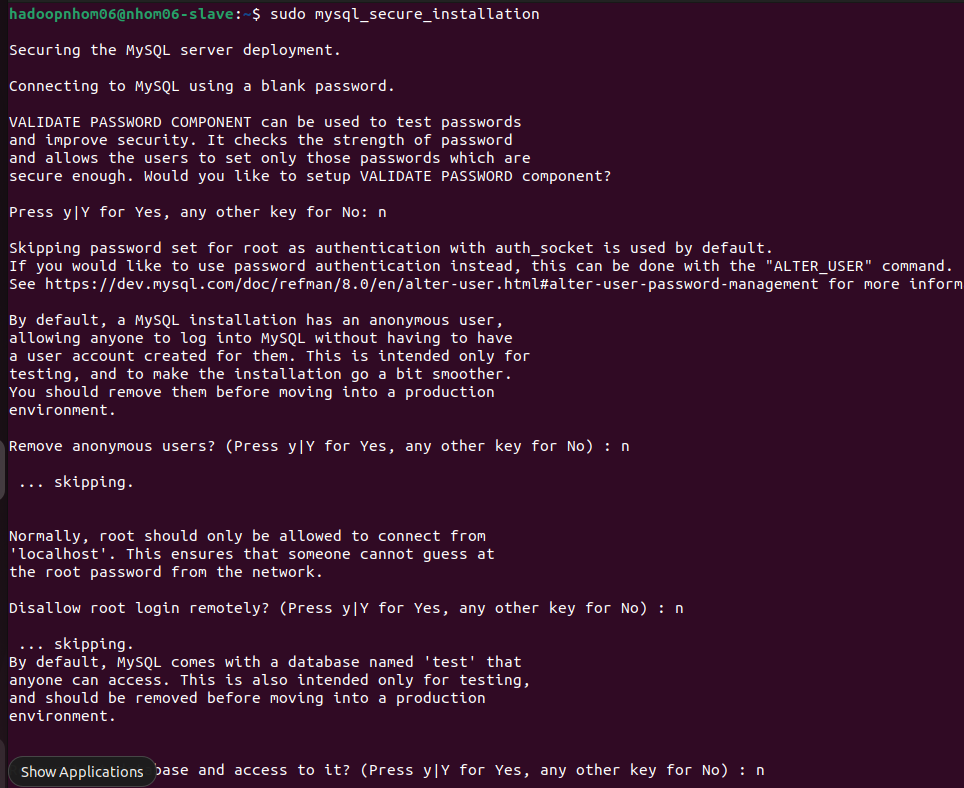
* Cài đặt gói mysql-server

$ sudo apt install mysql-server mysql-client -y



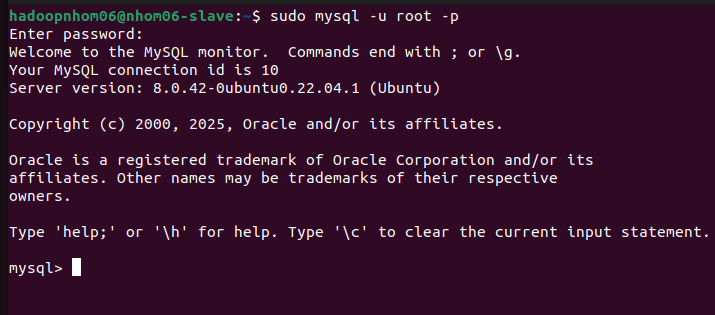
- Để cấu hình cài đặt tùy chọn

$ sudo mysql\_secure\_installation



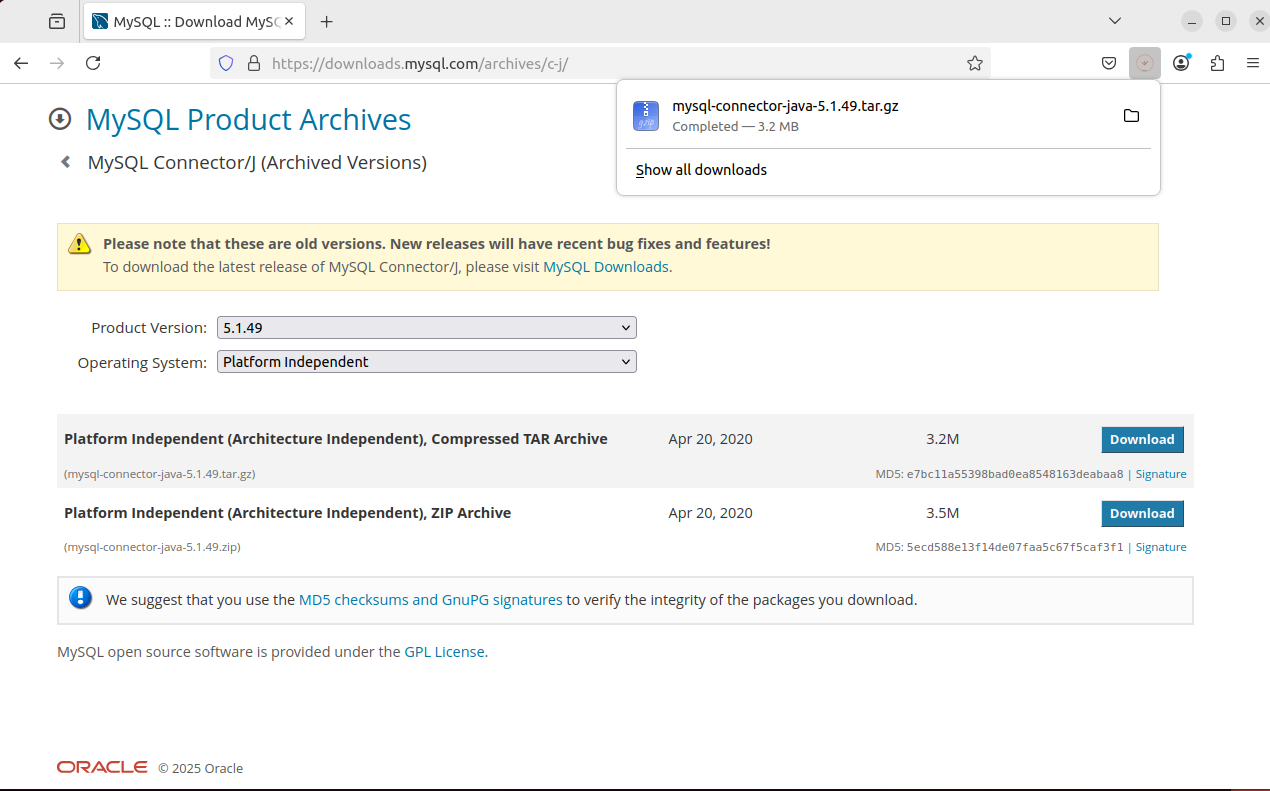
- Vào mysql client

$ sudo mysql -u root -p



- Tải file cấu hình theo link:

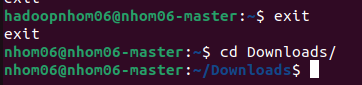
<https://downloads.mysql.com/archives/c-j/>



- Chuyển về tài khoản thường

$ exit

$ cd /home/nhom06/Downloads/



Giải nén file cài đặt

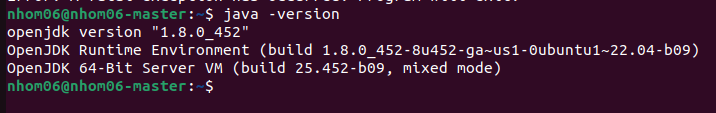
# tar -zxf mysql-connector-java-5.1.49.tar.gz



1. **Cài đặt & cấu hình Apache Sqoop**

Bước 1: kiểm tra cài đặt JAVA

$ java -version

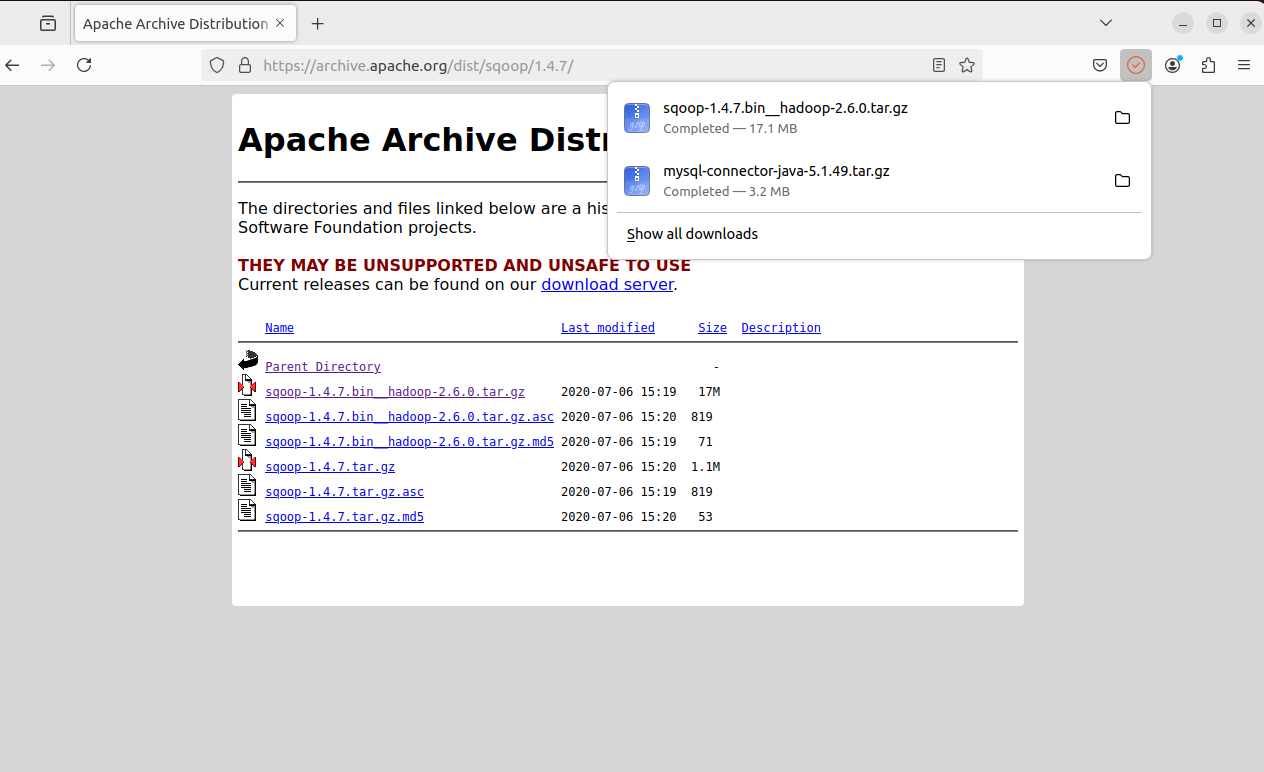


**- Cài đặt SQOOP**

Tìm phiên bản Sqoop mới nhất cần cài đặt qua link: <https://archive.apache.org/dist>

→ Chọn thư mục sqoop

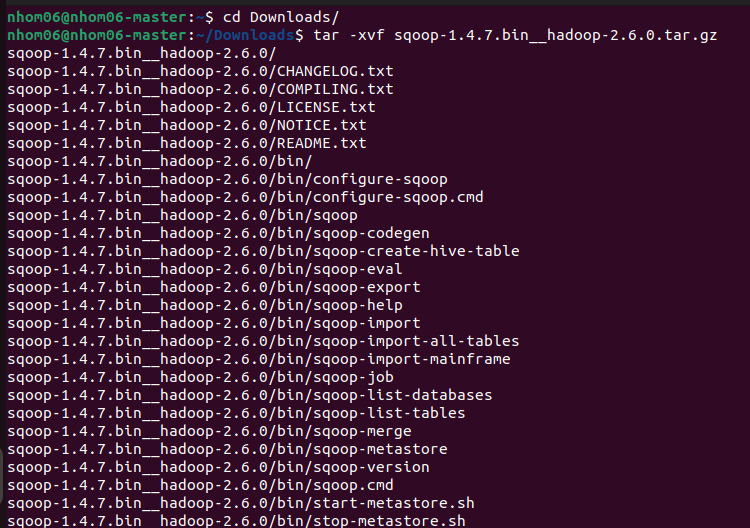
→ Chọn phiên bản 1.4.7



$ cd Downloads/ (chứa sqoop-1.4.7.bin\_\_hadoop-2.6.0.tar.gz)

- Bung và giải nén tập tin

$ tar -xvf sqoop-1.4.7.bin\_\_hadoop-2.6.0.tar.gz

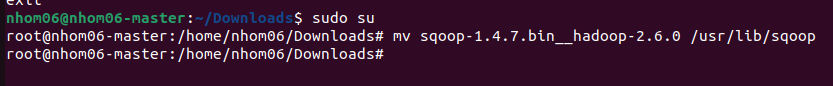


Đăng nhập tài khoản root

$ sudo su

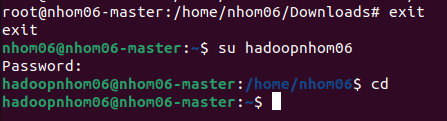
Di chuyển thư mục vào thư mục /usr/lib/sqoop

$ mv sqoop-1.4.7.bin\_\_hadoop-2.6.0 /usr/lib/sqoop



$ exit

$ su hadoopnhom06



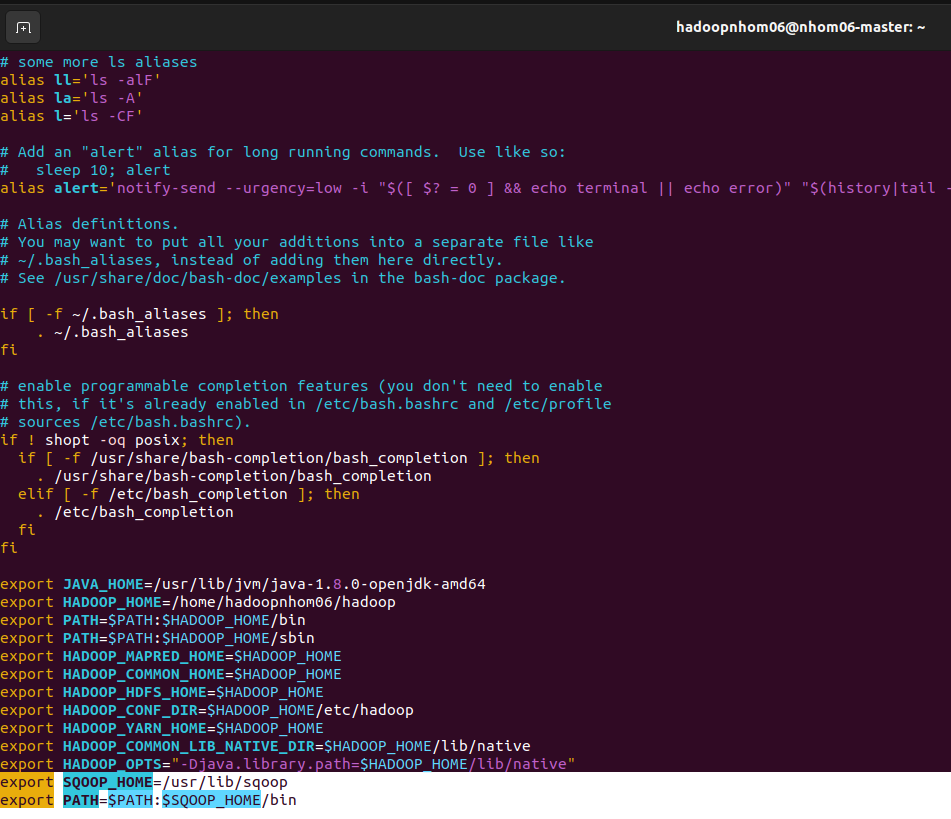
Cấu hình tập tin ~/.bashrc chứa các biến môi trường của hệ thống

$ vi ~/.bashrc

Thêm hai dòng sau vào cuối tập tin

export SQOOP\_HOME=/usr/lib/sqoop

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



Cập nhật tập tin ~/.bashrc

$ source ~/.bashrc

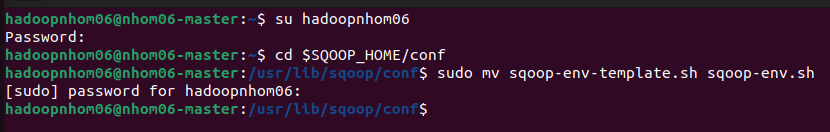
**\* CẤU HÌNH SQOOP**

Cấu hình tập tin sqoop-env.sh trong thư mục $SQOOP\_HOME/conf cho Hadoop

$ su hadoopnhom06 (chuyển về tài khoản hadoop)

$ cd $SQOOP\_HOME/conf

$ sudo mv sqoop-env-template.sh sqoop-env.sh

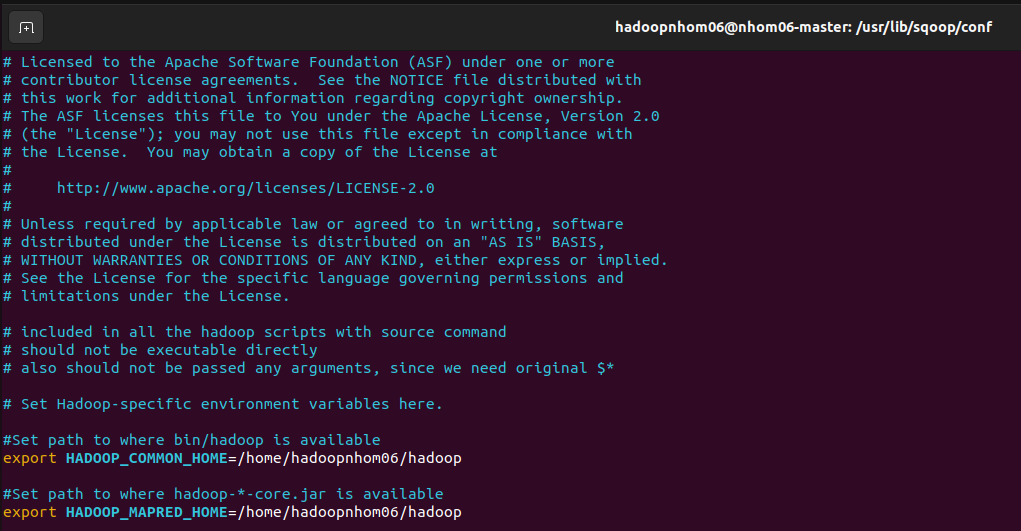


$ sudo vi sqoop-env.sh

Cập nhật các giá trị sau:

export HADOOP\_COMMON\_HOME=/home/hadoopnhom06/hadoop

export HADOOP\_MAPRED\_HOME=/home/hadoopnhom06/hadoop



**\* CẤU HÌNH KẾT NỐI MYSQL**

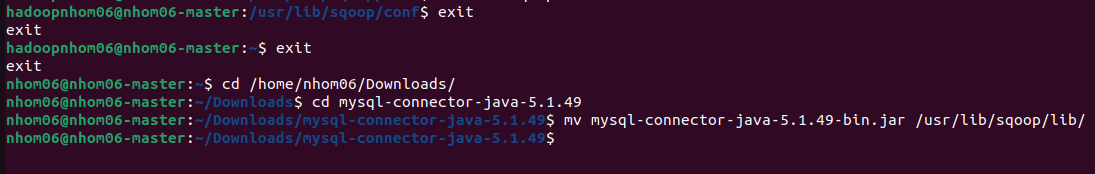
Chuyển về tài khoản thường

$ exit

$ cd /home/nhom06/Downloads/

$ cd mysql-connector-java-5.1.49

$ mv mysql-connector-java-5.1.49-bin.jar /usr/lib/sqoop/lib/



Xác thực sqoop (vào tài khoản hadoop)

$ su hadoopnhom06

$ cd $SQOOP\_HOME/bin

$ sqoop-version

