Cài đặt Hadoop Ecosystem – Phần 1

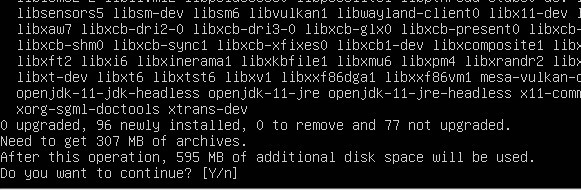
* Ubuntu Server 22.04
* Hadoop 3.3.4
* cài  hadoop 3.3.6
* cài OpenJDK
* (dấu # là dấu nháy)
* cấu hình SSH
* Lưu đã chỉnh: ESC : wq
* bỏ lưu: ESC q!
* Login với vai trò root (pass: root) để thực hiện những công việc sau

# 1. Cài đặt OpenJDK

# apt update

# reboot

# apt install default-jdk

 Chọn Y, nhấn Enter

# 2. Cài đặt SSH

# apt-get install ssh # apt install openssh-server

# reboot

2.1 Cấu hình SSH

# 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

# 3. Tạo user hadoop

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

# adduser hadoopminhchau

# 4. Cài đặt Hadoop 3.3.4

* Chuyển qua hadoopuser

# su hadoopminhchau

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

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

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

* 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

# 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

# Thiết lập Single Node Cluster 6. Standalone Operation



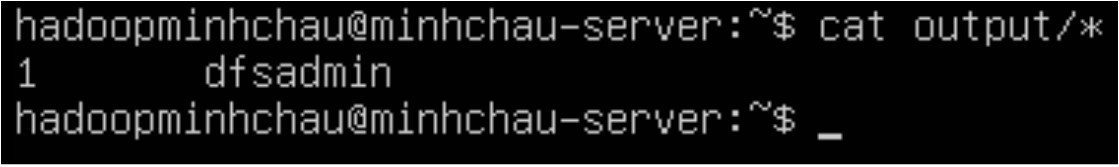
Mặc định, Hadoop được cấu hình chạy ở chế độ không phân tán như là một tiến trình đơn Java.

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

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

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

# cat /home/hadoopminhchau/.ssh/id\_rsa.pub >>

/home/hadoopminhchau/.ssh/authorized\_keys

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

* 1. Cấu hình file core-site.xml

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

<configuration>

<property>

<name>fs.defaultFS</name>

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

</property>

</configuration>

* 1. Cấu hình file hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

</configuration>

* 1. Format hệ thống (chạy 1 lần duy nhất)

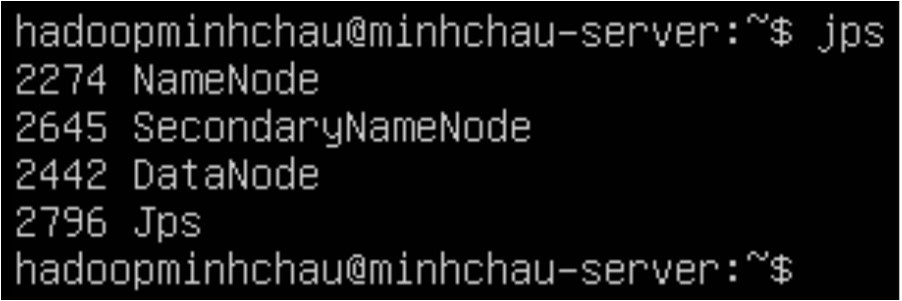
$ hadoop/bin/hdfs namenode -format

* 1. Start NameNode daemon và DataNode daemon

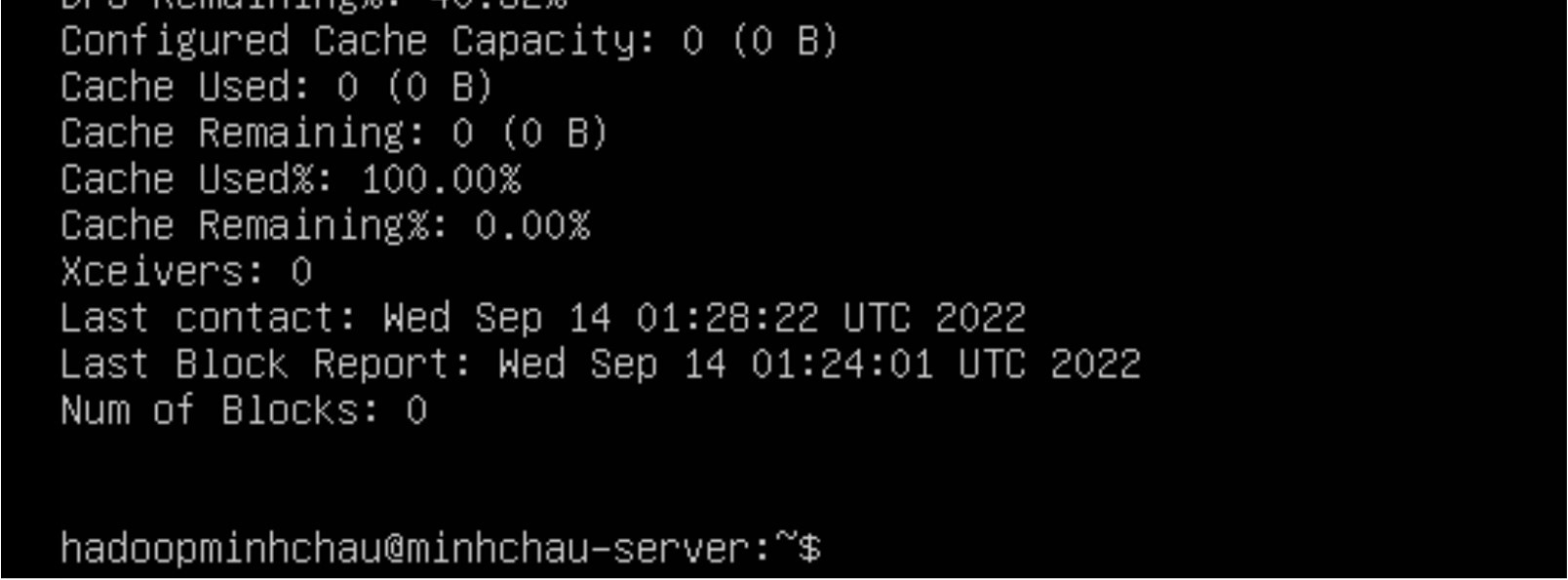
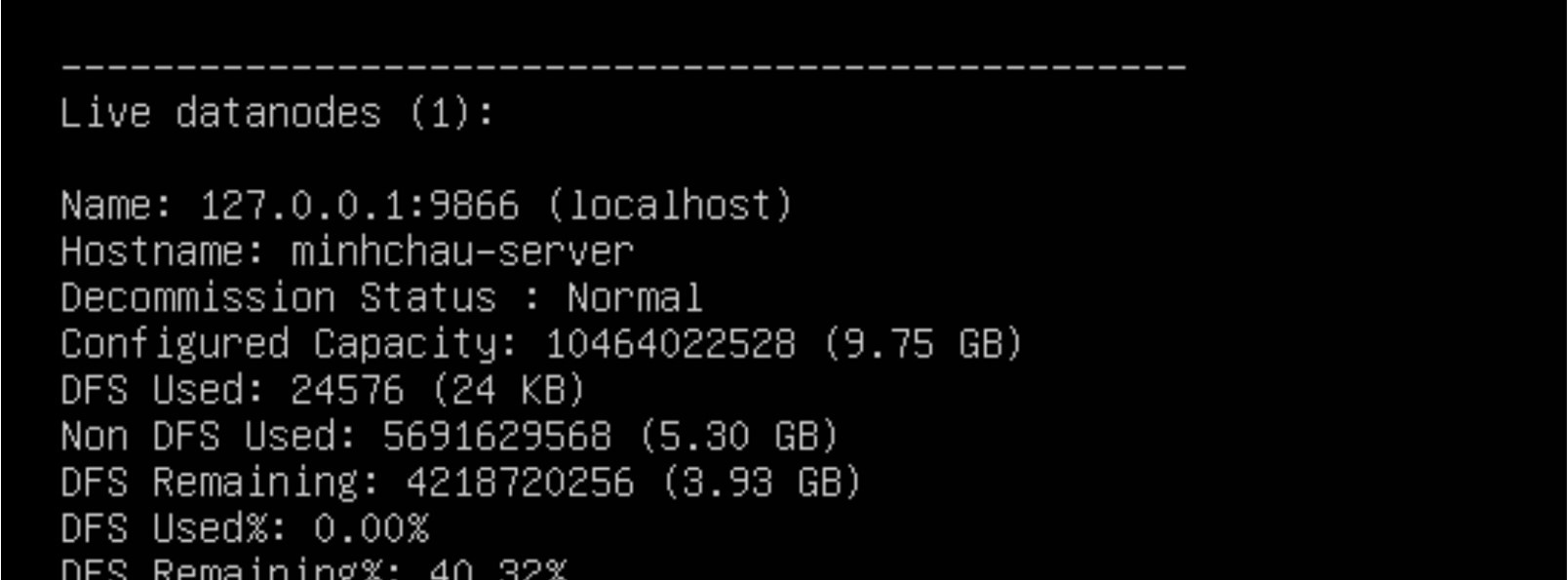
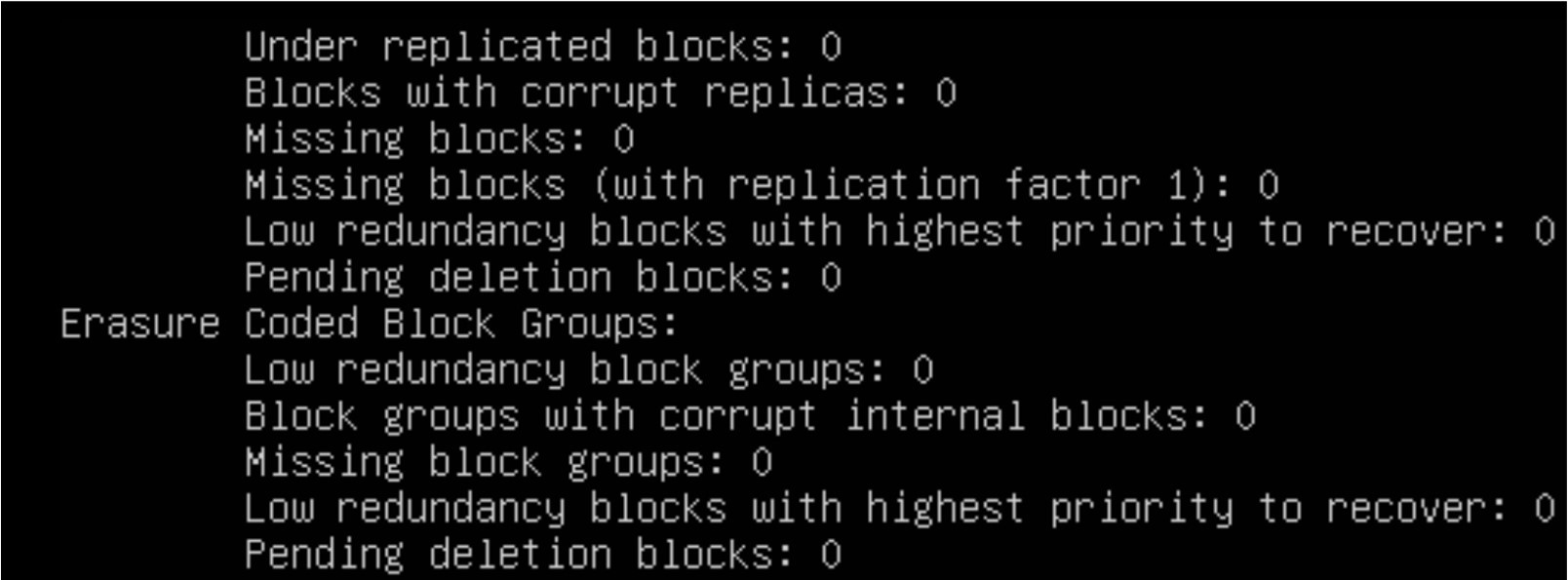
$ hadoop/sbin/start-dfs.sh

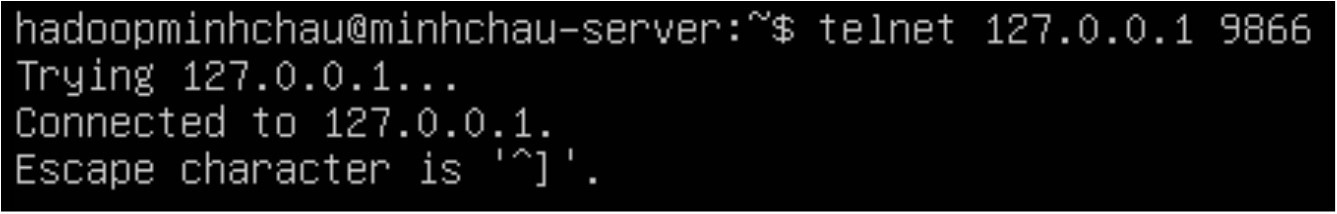
Kiểm tra các daemon đang chạy

$ jps

 Kiểm tra các node còn hoạt động

$ ~/hadoop/bin/hdfs dfsadmin -report





Ko làm

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

Copy các file .xml vào hệ thống file phân tán

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

/user/hadoopminhchau/input

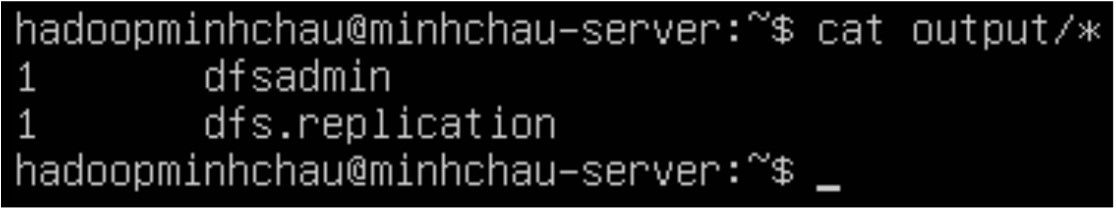
Hiển thị kết quả phù hợp với biểu thức chính quy

$ ~/hadoop/bin/hadoop jar hadoop/share/hadoop/mapreduce/hadoopmapreduce-examples-3.3.4.jar grep input output ‘dfs[a-z.]+’

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

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

$ cat output/\*



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/hadoopminhchau/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"

* Soucre file .bashrc

# source ~/.bashrc

7.7 Cấu hình file mapred-site.xml

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

<configuration>

<property>

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

<value>yarn</value>

</property>

<property>

<name>mapreduce.application.classpath</name>

<value>$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/\*:$HADOO

P\_MAPRED\_HOME/share/hadoop/mapreduce/lib/\*</value>

</property>

</configuration>

7.8 Cấu hình file yarn-site.xml

<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,HADOO

P\_CONF\_DIR,CLASSPATH\_PREPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HA DOOP\_MAPRED\_HOME</value>

</property>

</configuration>

7.9 Chạy lại các daemon

$ hadoop/sbin/start-dfs.sh

$ hadoop/sbin/start-yarn.sh

Nếu báo lỗi cannot connect… thì chạy lệnh format namenode lần nữa

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

Copy các file .xml vào hệ thống file phân tán

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

/user/hadoopminhchau/input

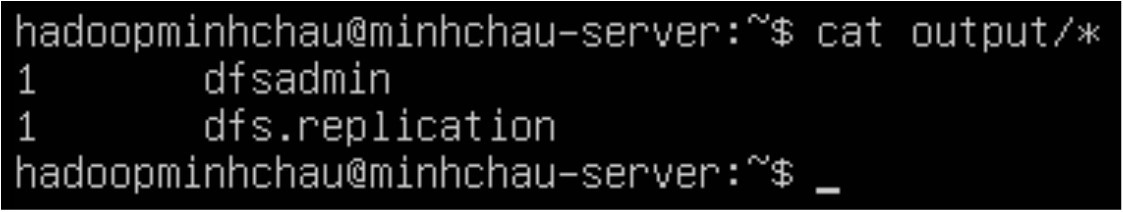
Hiển thị kết quả phù hợp với biểu thức chính quy

$ ~/hadoop/bin/hadoop jar hadoop/share/hadoop/mapreduce/hadoopmapreduce-examples-3.3.4.jar grep input output ‘dfs[a-z.]+’

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

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

$ cat output/\*



Khi muốn dừng các daemon thì chạy lệnh sau

$ hadoop/sbin/stop-dfs.sh $ hadoop/sbin/stop-yarn.sh