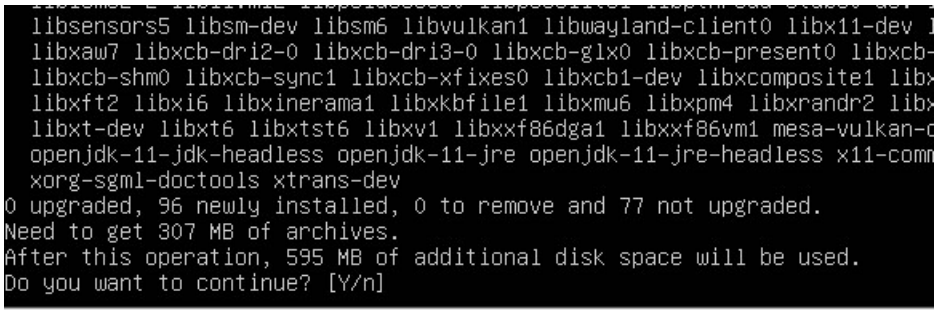
Cài đặt Hadoop Ecosystem – Phần 1

- Ubuntu Server 22.04
- Hadoop 3.3.4
- 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/hadoop-mapreduce-examples-3.3.4.jar grep input output 'dfs[a-z.]+'
# cat output/*
```

```
hadoopminhchau@minhchau-server:~$ cat output/*
1       dfsadmin
hadoopminhchau@minhchau-server:~$ _
```

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
```

7.2 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>
```

7.3 Cấu hình file hdfs-site.xml

```
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
</configuration>
```

7.4 Format hệ thống (chạy 1 lần duy nhất)

```
$ hadoop/bin/hdfs namenode -format
```

7.5 Start NameNode daemon và DataNode daemon

```
$ hadoop/sbin/start-dfs.sh
```

Kiểm tra các daemon đang chạy

```
$ jps
```

```
hadoopminhchau@minhchau-server:~$ jps
2274 NameNode
2645 SecondaryNameNode
2442 DataNode
2796 Jps
hadoopminhchau@minhchau-server:~$
```

Kiểm tra các node còn hoạt động

```
$ ~/hadoop/bin/hdfs dfsadmin -report
```

```
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
Live datanodes (1):

Name: 127.0.0.1:9866 (localhost)
Hostname: minhchau-server
Decommission Status : Normal
Configured Capacity: 10464022528 (9.75 GB)
DFS Used: 24576 (24 KB)
Non DFS Used: 5691629568 (5.30 GB)
DFS Remaining: 4218720256 (3.93 GB)
DFS Used%: 0.00%
DFS Remaining%: 40.32%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Wed Sep 14 01:28:22 UTC 2022
Last Block Report: Wed Sep 14 01:24:01 UTC 2022
Num of Blocks: 0


hadoopminhchau@minhchau-server:~$
```

```
hadoopminhchau@minhchau-server:~$ telnet 127.0.0.1 9866
Trying 127.0.0.1...
Connected to 127.0.0.1.
Escape character is '^]'.
```

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/hadoop-
mapreduce-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/*
```

```
hadoopminhchau@minhchau-server:~$ cat output/*
1       dfsadmin
1       dfs.replication
hadoopminhchau@minhchau-server:~$ _
```

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/hadoop-
mapreduce-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/*
```

```
hadoopminhchau@minhchau-server:~$ cat output/*
1       dfsadmin
1       dfs.replication
hadoopminhchau@minhchau-server:~$ _
```

Khi muốn dừng các daemon thì chạy lệnh sau

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