

Lab01: Set Up a Hadoop

3.2.1 Multi-Node Cluster on Ubuntu (2 Nodes)

I. Trên toàn bộ máy

1st Step: Cài đặt SSH, PDSH

- Cài SSH

```
sudo apt install ssh
```

Nếu hỏi phải nhập password. Hãy nhập nó

```
torres@torres-VirtualBox:~$ sudo apt install ssh
[sudo] password for torres:
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  ncurses-term openssh-client openssh-server openssh-sftp-server
  ssh-import-id
Suggested packages:
  keychain libpam-ssh monkeysphere ssh-askpass molly-guard rssh
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
The following packages will be upgraded:
  openssh-client
1 upgraded, 5 newly installed, 0 to remove and 478 not upgraded.
Need to get 1256 kB of archives.
After this operation, 5422 kB of additional disk space will be used.
Do you want to continue? [Y/n] █
```

- Cài PDSH

```
sudo apt install pdsh
```

Như trước, nhãy nhập thông tin khi được yêu cầu

```
torres@torres-VirtualBox:~$ sudo apt install pdsh
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  genders libgenders0
Suggested packages:
  rdist
The following NEW packages will be installed:
  genders libgenders0 pdsh
0 upgraded, 3 newly installed, 0 to remove and 478 not upgraded.
Need to get 170 kB of archives.
After this operation, 479 kB of additional disk space will be used.
Do you want to continue? [Y/n] ■
```

- Mở file .bashrc:

```
Sudo nano .bashrc
```

- Ở cuối file, thêm dòng

```
export PDSH_RCMD_TYPE=ssh
```

```
torres@torres-VirtualBox: ~
File Edit View Search Terminal Help
GNU nano 2.9.3 .bashrc Modified
alias alert='notify-send --urgency=low -i "$(($? = 0) && echo terminal || ec$"
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
    if [ -f /usr/share/bash-completion/bash_completion ]; then
        . /usr/share/bash-completion/bash_completion
    elif [ -f /etc/bash_completion ]; then
        . /etc/bash_completion
    fi
fi

export PDSH_RCMD_TYPE=ssh

^G Get Help      ^O Write Out     ^W Where Is      ^K Cut Text      ^J Justify
^X Exit          ^R Read File     ^\ Replace       ^U Uncut Text    ^T To Spell
```

- Để lưu file:

Ctrl X

Y

Enter

- Config SSH. Tạo 1 khóa bằng lệnh:

ssh-keygen -t rsa -P ""

Bấm Enter mỗi khi cần

```
torres@torres-VirtualBox:~$ ssh-keygen -t rsa -P ""  
Generating public/private rsa key pair.  
Enter file in which to save the key (/home/torres/.ssh/id_rsa):  
Your identification has been saved in /home/torres/.ssh/id_rsa.  
Your public key has been saved in /home/torres/.ssh/id_rsa.pub.  
The key fingerprint is:  
SHA256:4mA5iJJkH2jfUJz3ZThFq+41knubh5kvR07cHyA1os torres@torres-VirtualBox  
The key's randomart image is:  
+---[RSA 2048]---+  
| .o . . o+o |  
| o . = . .oo |  
| . =o+ o . |  
| ...ooo. o = |  
| o . =o. S B = |  
| . o o . O + * |  
| . o E B . |  
| . . .+ + |  
| . . .++ . |  
+---[SHA256]---+  
torres@torres-VirtualBox:~$
```

- Copy public key cho authorized_keys với lệnh:

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

```
torres@torres-VirtualBox:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys  
torres@torres-VirtualBox:~$
```

- Check SSH bằng việc connect đến localhost

```
ssh localhost
```

Nhập Yes và bấm Enter mỗi khi cần

```
torres@torres-VirtualBox: ~
File Edit View Search Terminal Help
torres@torres-VirtualBox:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:3e40r2W2P+iNTu397mca4JKsVFFVnJViEXz7WQKGe3Y.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 18.04.2 LTS (GNU/Linux 4.18.0-15-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

484 packages can be updated.
250 updates are security updates.

Your Hardware Enablement Stack (HWE) is supported until April 2023.

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/*copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

torres@torres-VirtualBox:~$
```

Cài đặt thành công cài SSH, PDSH

2nd Step: Cài đặt Java 8

- Để cài java 8. Nhập lệnh

```
sudo apt install openjdk-8-jdk
```

- Như những phần trước, bấm mật khẩu hay trả lời mỗi khi cần

```
torres@torres-VirtualBox: ~
File Edit View Search Terminal Help
Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

torres@torres-VirtualBox:~$ sudo apt install openjdk-8-jdk
[sudo] password for torres:
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
libice-dev libpthread-stubs0-dev libsm-dev libx11-6 libx11-dev libx11-doc
libxau-dev libxcb1 libxcb1-dev libxdmcp-dev libxt-dev
openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless
x11proto-core-dev x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
libice-doc libsm-doc libxcb-doc libxt-doc openjdk-8-demo openjdk-8-source
visualvm icedtea-8-plugin fonts-ipafont-gothic fonts-ipafont-mincho
fonts-wqy-microhei fonts-wqy-zanhei
The following NEW packages will be installed:
libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libx11-doc libxau-dev
libxcb1-dev libxdmcp-dev libxt-dev openjdk-8-jdk openjdk-8-jdk-headless
openjdk-8-jre openjdk-8-jre-headless x11proto-core-dev x11proto-dev
xorg-sgml-doctools xtrans-dev
The following packages will be upgraded:
libx11-6 libxcb1
2 upgraded, 17 newly installed, 0 to remove and 476 not upgraded.
Need to get 41,0 MB/41,6 MB of archives.
After this operation, 159 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```

- Check Java version

```
java -version
```

```
torres@torres-VirtualBox:~$ java -version
openjdk version "1.8.0_242"
OpenJDK Runtime Environment (build 1.8.0_242-8u242-b08-0ubuntu3-18.04-b08)
OpenJDK 64-Bit Server VM (build 25.242-b08, mixed mode)
torres@torres-VirtualBox:~$
```

3rd Step: Tải Hadoop

- Tải hadoop bằng lệnh sau:

```
sudo wget -P ~  
https://mirrors.sonic.net/apache/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
```

```
torres@torres-VirtualBox:~$ sudo wget -P ~ https://mirrors.sonic.net/apache/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz  
--2020-02-02 21:11:42-- https://mirrors.sonic.net/apache/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz  
Resolving mirrors.sonic.net (mirrors.sonic.net)... 157.131.0.16, 2001:5a8:601:2:157:131:0:16  
Connecting to mirrors.sonic.net (mirrors.sonic.net)|157.131.0.16|:443... connected.  
HTTP request sent, awaiting response... 200 OK  
Length: 359196911 (343M) [application/x-gzip]  
Saving to: '/home/torres/hadoop-3.2.1.tar.gz'  
  
hadoop-3.2.1.tar.gz 0%[          ] 744,00K 313KB/s
```

Note: Trong vài trường hợp, link trên sẽ bị chết và bạn không tải được. Lên google gõ hadoop-3.2.1.tar.gz tìm đến link dưới, tải nó về và di chuyển nó từ mục download vào thư mục home



Name	Last modified	Size	Description
 Parent Directory		-	
 hadoop-3.2.1-src.tar.gz	2019-09-23 05:16	30M	
 hadoop-3.2.1-src.tar.gz.asc	2019-09-23 05:16	819	
 hadoop-3.2.1-src.tar.gz.mds	2019-09-24 13:32	1.1K	
 hadoop-3.2.1-src.tar.gz.sha512	2019-09-23 05:16	195	
 hadoop-3.2.1.tar.gz	2019-09-23 05:16	343M	
 hadoop-3.2.1.tar.gz.asc	2019-09-23 05:16	819	
 hadoop-3.2.1.tar.gz.mds	2019-09-24 13:32	958	
 hadoop-3.2.1.tar.gz.sha512	2019-09-23 05:16	191	

- Giải nén hadoop-3.2.1.tar.gz file:

```
tar xzf hadoop-3.2.1.tar.gz
```

```
torres@torres-VirtualBox:~$ tar xzf hadoop-3.2.1.tar.gz
```

- Đổi tên hadoop-3.2.1 thành hadoop:

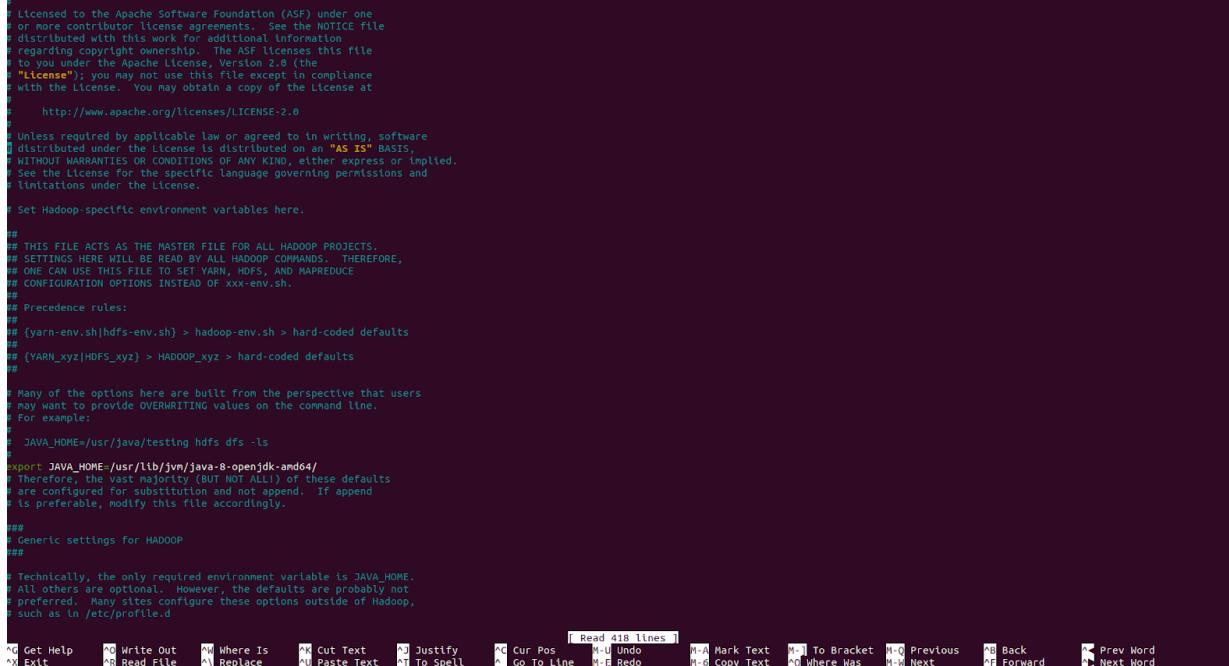
```
mv hadoop-3.2.1 hadoop
```

- Mở file hadoop-env.sh và sửa JAVA_HOME:

```
Sudo nano ~/hadoop/etc/hadoop/hadoop-env.sh
```

- Thêm dòng này vào:

```
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
```



```
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements. See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership. The ASF licenses this file
# to you under the Apache License, Version 2.0 (the
# "License"); you may not use this file except in compliance
# with the License. You may obtain a copy of the License at
#
#     http://www.apache.org/licenses/LICENSE-2.0
#
# Unless required by applicable law or agreed to in writing, software
# distributed under the License is distributed on an "AS IS" BASIS,
# WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
# See the License for the specific language governing permissions and
# limitations under the License.

# Set Hadoop-specific environment variables here.

## THIS FILE ACTS AS THE MASTER FILE FOR ALL HADOOP PROJECTS.
## SETTINGS HERE WILL BE READ BY ALL HADOOP COMMANDS. THEREFORE,
## ONE CAN USE THIS FILE TO SET YARN, HDFS, AND MAPREDUCE
## CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.

## Precedence rules:
## (yarn-env.sh|hdfs-env.sh) > hadoop-env.sh > hard-coded defaults
## [YARN_xyz|HDFS_xyz] > HADOOP_xyz > hard-coded defaults

## Many of the options here are built from the perspective that users
## may want to provide OVERWRITING values on the command line.
## For example:
##   $ JAVA_HOME=/usr/java/testing hdfs dfs -ls
##   export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/
## Therefore, the vast majority (BUT NOT ALL!) of these defaults
## are configured for substitution and not append. If append
## is preferable, modify this file accordingly.

## Generic settings for HADOOP
## Technically, the only required environment variable is JAVA_HOME.
## All others are optional. However, the defaults are probably not
## preferred. Many sites configure these options outside of Hadoop,
## such as in /etc/profile.d.

[ Read 418 lines ]
```

Lưu lại file

- Chuyển hadoop folder đến /usr/local/hadoop:

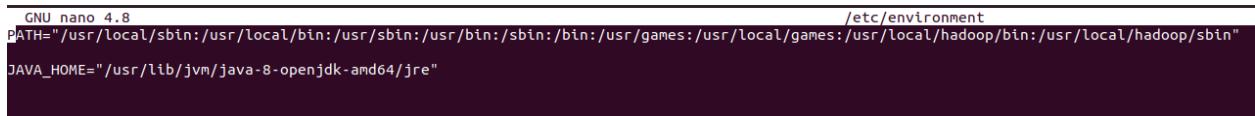
```
sudo mv hadoop /usr/local/hadoop
```

- Mở file environment:

```
sudo nano /etc/environment
```

- Thay thế dòng cũ bằng:

```
PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/
sbin:/bin:/usr/games:/usr/local/games:/usr/local/hadoop/b
in:/usr/local/hadoop/sbin"JAVA_HOME="/usr/lib/jvm/java-8-
openjdk-amd64/jre"
```



```
GNU nano 4.8
PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/usr/local/hadoop/bin:/usr/local/hadoop/sbin"
JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre"
```

Lưu lại file

4th Step: Tạo 1 user hadoopuser mới trên các máy

- Gõ lệnh:

```
sudo adduser hadoopuser
```

//Điền password là 1 cho dễ :D, còn lại cứ bấm Enter

```
torres@torres-VirtualBox:~$ sudo adduser hadoopuser
Adding user 'hadoopuser' ...
Adding new group 'hadoopuser' (1001) ...
Adding new user 'hadoopuser' (1001) with group 'hadoopuser' ...
Creating home directory '/home/hadoopuser' ...
Copying files from '/etc/skel' ...
Enter new UNIX password: 1
Retype new UNIX password: 1
No password supplied
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hadoopuser
Enter the new value, or press ENTER for the default
    Full Name []:
    Room Number []:      Bấm ENTER
    Work Phone []:
    Home Phone []:
    Other []:
Is the information correct? [Y/n] y
torres@torres-VirtualBox:~$
```

- Cấp quyền cho nó, bấm lần lượt các lệnh:

```
sudo usermod -aG hadoopuser hadoopuser
```

```
sudo chown hadoopuser:root -R /usr/local/hadoop/
```

```
sudo chmod g+rwx -R /usr/local/hadoop/
```

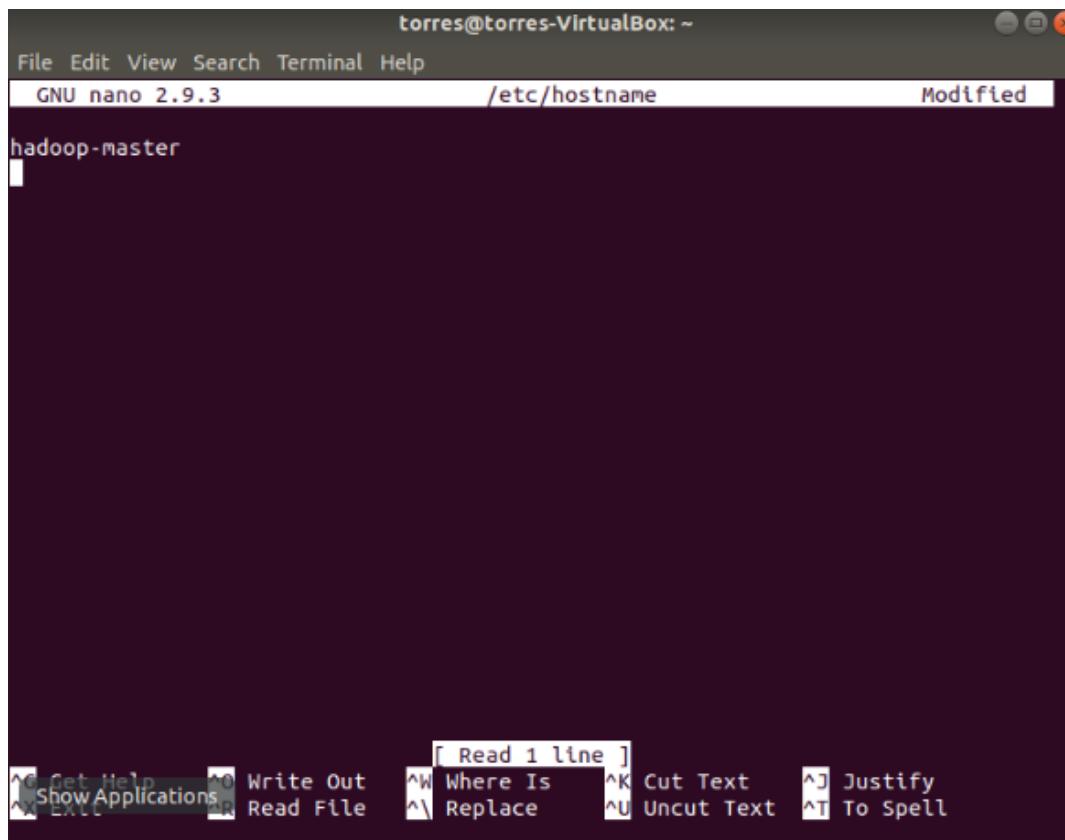
```
sudo adduser hadoopuser sudo
```

```
torres@torres-VirtualBox:~$ sudo usermod -aG hadoopuser hadoopuser
torres@torres-VirtualBox:~$ sudo chown hadoopuser:root -R /usr/local/hadoop
torres@torres-VirtualBox:~$ sudo chmod g+rwx -R /usr/local/hadoop
torres@torres-VirtualBox:~$ sudo adduser hadoopuser sudo
Adding user 'hadoopuser' to group 'sudo' ...
Adding user hadoopuser to group sudo
Done.
torres@torres-VirtualBox:~$
```

- Thay đổi tên máy:

```
sudo nano /etc/hostname
```

- Chọn 1 máy là master, sửa file đó thành



- Chọn 1 máy là slave1, sửa file đó thành

```
torres@torres-VirtualBox: ~
File Edit View Search Terminal Help
GNU nano 2.9.3          /etc/hostname           Modified
hadoop-slave1
```

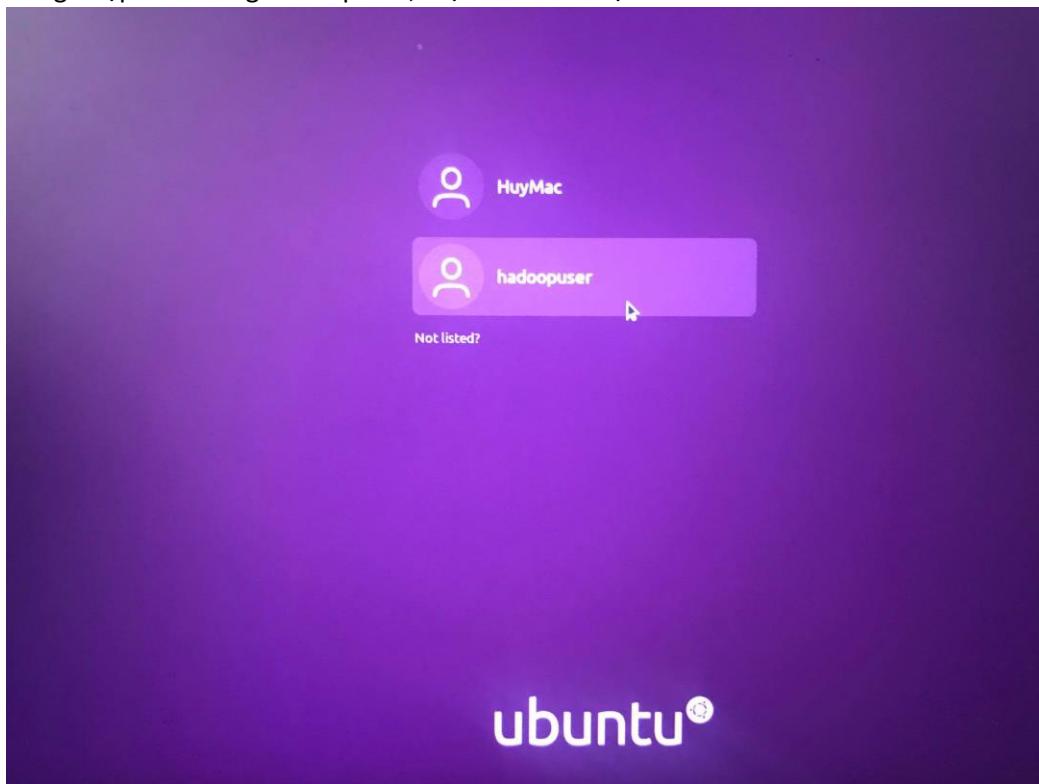
- Chọn 1 máy là slave2, sửa file đó thành

```
torres@torres-VirtualBox: ~
File Edit View Search Terminal Help
GNU nano 2.9.3          /etc/hostname           Modified
hadoop-slave2
```

Lưu lại.

- Gõ lệnh `sudo reboot` để restart máy.

- Đăng nhập vào thằng hadoopuser, mật khẩu vừa đặt là 1



- Kiểm tra địa chỉ IP từng máy bằng lệnh

Ifconfig

```
hadoopuser@hadoop-master:~$ ifconfig
enp4s0: flags=4099<UP,BROADCAST,MULTICAST> mtu 1500
    ether 00:e0:4c:68:00:37 txqueuelen 1000 (Ethernet)
      RX packets 0 bytes 0 (0.0 B)
      RX errors 0 dropped 0 overruns 0 frame 0
      TX packets 0 bytes 0 (0.0 B)
      TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING> mtu 65536
    inet 127.0.0.1 netmask 255.0.0.0
      inet6 ::1 prefixlen 128 scopeid 0x10<host>
        loop txqueuelen 1000 (Local Loopback)
          RX packets 4778 bytes 456014 (456.0 KB)
          RX errors 0 dropped 0 overruns 0 frame 0
          TX packets 4778 bytes 456014 (456.0 KB)
          TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

wlp5s0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST> mtu 1500
    inet 192.168.100.3 netmask 255.255.255.0 broadcast 192.168.100.255
      inet6 2001:ee0:4141:ae12::8 prefixlen 128 scopeid 0x0<global>
      inet6 fe80::9924:1ea9:ebec:ae3 prefixlen 64 scopeid 0x20<link>
      inet6 2001:ee0:4141:ae12:dd9:921f:90d6:82fe prefixlen 64 scopeid 0x0<global>
        ether 08:d4:0c:7b:16:c8 txqueuelen 1000 (Ethernet)
          RX packets 24543 bytes 13447078 (13.4 MB)
          RX errors 0 dropped 0 overruns 0 frame 0
          TX packets 19139 bytes 6207170 (6.2 MB)
          TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0

hadoopuser@hadoop-master:~$
```

- Kiểm tra địa chỉ IP các máy hadoop-master, hadoop-slave1, hadoop-slave2. TH mình:

192.168.100.3 hadoop-master
192.168.100.14 hadoop-slave1
192.168.100.15 hadoop-slave2

- Mở hosts file trên tất cả máy, thêm 3 dòng trên

```
sudo nano /etc/hosts
```

```
GNU nano 4.8          /etc/hosts
127.0.0.1      localhost
127.0.1.1      huymac-Inspiron-7559

192.168.100.3  hadoop-master
192.168.100.14 hadoop-slave1
192.168.100.15 hadoop-slave2

# The following lines are desirable for IPv6 capable hosts
::1      ip6-localhost ip6-loopback
fe00::0  ip6-localnet
ff00::0  ip6-mcastprefix
ff02::1  ip6-allnodes
ff02::2  ip6-allrouters

[ Read 13 lines ]
^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace   ^U Paste Text ^T To Spell  ^_ Go To Line
```

Lưu lại file, nên thẻ sudo reboot cả 3 máy 1 lần nữa và đăng nhập lại vào hadoopuser

II. Trên máy hadoop-master

1st Step: Dùng SSH để kết nối từ hadoop-master tới hadoop-slave1, hadoop-slave2

- Sinh khóa rsa

```
ssh-keygen -t rsa
```

```
hadoopuser@hadoop-master:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoopuser/.ssh/id_rsa):
Created directory '/home/hadoopuser/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoopuser/.ssh/id_rsa.
Your public key has been saved in /home/hadoopuser/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:wti4V4CRMVGcbBEIbw/G0Rnpf/HoT9T1FE65iCZXS4 hadoopuser@hadoop-master
The key's randomart image is:
+---[RSA 2048]----+
| ..*0+=o o... .o|
| .o*=. o = o o |
| +.+ o = E = o|
| . * o o o = .o|
| o + S . .o |
| . o ...o|
| . . .+|
| . .+|
| .o|
+---[SHA256]----+
hadoopuser@hadoop-master:~$
```

- Copy khóa tới các máy

```
ssh-copy-id hadoopuser@hadoop-master
```

```
hadoopuser@hadoop-master:~$ ssh-copy-id hadoopuser@hadoop-master
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/hadoopuser/.ssh/id_rsa.pub"
The authenticity of host 'hadoop-master (192.168.205.7)' can't be established.
ECDSA key fingerprint is SHA256:Nsjcx3SrmwVnSNcWxvlYIajjHRdwaET+RGeLkVyoHI4.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter
out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompted
now it is to install the new keys
hadoopuser@hadoop-master's password:

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'hadoopuser@hadoop-master'"
and check to make sure that only the key(s) you wanted were added.

hadoopuser@hadoop-master:~$
```

```
ssh-copy-id hadoopuser@hadoop-slave1
```

```
hadoopuser@hadoop-master:~$ ssh-copy-id hadoopuser@hadoop-slave1
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/hadoopuser/.ssh/id_rsa.pub"
The authenticity of host 'hadoop-slave1 (192.168.205.8)' can't be established.
ECDSA key fingerprint is SHA256:Nsjcx3SrmwVnSNcWxvlYIajjHRdwaET+RGeLkVyoHI4.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter
out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are promp
ted now it is to install the new keys
hadoopuser@hadoop-slave1's password:

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'hadoopuser@hadoop-slave1'"
and check to make sure that only the key(s) you wanted were added.

hadoopuser@hadoop-master:~$
```

```
ssh-copy-id hadoopuser@hadoop-slave2
```

```
hadoopuser@hadoop-master:~$ ssh-copy-id hadoopuser@hadoop-slave2
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/hadoopuser/.ssh/id_rsa.pub"
The authenticity of host 'hadoop-slave2 (192.168.205.9)' can't be established.
ECDSA key fingerprint is SHA256:Nsjcx3SrmwVnSNcWxvlYIajjHRdwaET+RGeLkVyoHI4.
Are you sure you want to continue connecting (yes/no)? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter
out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are promp
ted now it is to install the new keys
hadoopuser@hadoop-slave2's password:

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'hadoopuser@hadoop-slave2'"
and check to make sure that only the key(s) you wanted were added.

hadoopuser@hadoop-master:~$
```

2nd Step: Config các file

- Sửa file `core-site.xml`:

```
sudo nano /usr/local/hadoop/etc/hadoop/core-site.xml
```

```
hadoopuser@hadoop-master:~$ sudo nano /usr/local/hadoop/etc/hadoop/core-site.xml
[sudo] password for hadoopuser:
```

Thêm các dòng sau:

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://hadoop-master:9000</value>
</property>
</configuration>
```

```
GNU nano 4.8      /usr/local/hadoop/etc/hadoop/core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://hadoop-master:9000</value>
</property>
</configuration>
```

[Read 24 lines]

^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify ^C Cur Pos
^X Exit ^R Read File ^\ Replace ^U Paste Text ^T To Spell ^_ Go To Line

- Sửa file **hdfs-site.xml**:

```
sudo nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml
```

```
hadoopuser@hadoop-master:~$ sudo nano /usr/local/hadoop/etc/hadoop/hdfs-site.xml
```

Thêm các dòng sau:

```
<configuration>
<property>
<name>dfs.namenode.name.dir</name><value>/usr/local/hadoop/data/nameNode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name><value>/usr/local/hadoop/data/dataNode</value>
</property>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
</configuration>
```

```
hadoopuser@hadoop-master: ~
GNU nano 4.8      /usr/local/hadoop/etc/hadoop/hdfs-site.xml      Modified
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>dfs.namenode.name.dir</name><value>/usr/local/hadoop/data/nameNode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name><value>/usr/local/hadoop/data/dataNode</value>
</property>
<property>
<name>dfs.replication</name>
<value>2</value>
</property>
</configuration>

[ Read 30 lines ]
^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos
^X Exit      ^R Read File  ^\ Replace   ^U Paste Text ^T To Spell  ^_ Go To Line
```

- Sửa file **workers**:

```
sudo nano /usr/local/hadoop/etc/hadoop/workers
```

```
hadoopuser@hadoop-master:~$ sudo nano /usr/local/hadoop/etc/hadoop/workers
```

Thêm các dòng sau:

```
hadoop-slave1
hadoop-slave2
```

```
hadoopuser@hadoop-master: ~
File Edit View Search Terminal Help
GNU nano 2.9.3      /usr/local/hadoop/etc/hadoop/workers      Modified
hadoop-slave1
hadoop-slave2

^G Get Help      ^O Write Out    ^W Where Is      ^K Cut Text    ^J Justify
^X Exit          ^R Read File   ^\ Replace       ^U Uncut Text  ^T To Spell
```

3rd Step:

- Gửi các file đã config sang các máy slave

```
scp /usr/local/hadoop/etc/hadoop/* hadoop-
slave1:/usr/local/hadoop/etc/hadoop/
```

```

hadoopuser@hadoop-master:~$ scp /usr/local/hadoop/etc/hadoop/* hadoop-slave1:/u
sr/local/hadoop/etc/hadoop
capacity-scheduler.xml          100%  8260    293.2KB/s  00:00
configuration.xsl               100% 1335     791.3KB/s  00:00
container-executor.cfg          100% 1940      1.1MB/s  00:00
core-site.xml                   100%   864    776.1KB/s  00:00
hadoop-env.cmd                 100% 3999      1.9MB/s  00:00
hadoop-env.sh                   100%  16KB    2.6MB/s  00:00
hadoop-metrics2.properties     100% 3321      1.4MB/s  00:00
hadoop-policy.xml              100%   11KB  293.3KB/s  00:00
hadoop-user-functions.sh.example 100% 3414      1.2MB/s  00:00
hdfs-site.xml                  100% 1051    201.6KB/s  00:00
httpfs-env.sh                  100% 1484     71.0KB/s  00:00
httpfs-log4j.properties        100% 1657    975.4KB/s  00:00
httpfs-signature.secret        100%   21      12.7KB/s  00:00
httpfs-site.xml                100%   620    351.4KB/s  00:00
kms-acls.xml                   100% 3518      1.8MB/s  00:00
kms-env.sh                      100% 1351    850.0KB/s  00:00
kms-log4j.properties           100% 1860    990.0KB/s  00:00
kms-site.xml                    100%   682    429.3KB/s  00:00
log4j.properties                100%   13KB   11.7MB/s  00:00
mapred-env.cmd                 100%   951    646.1KB/s  00:00
mapred-env.sh                   100% 1764      1.2MB/s  00:00
mapred-queues.xml.template     100% 4113      2.6MB/s  00:00
mapred-site.xml                 100%   758    817.1KB/s  00:00
/usr/local/hadoop/etc/hadoop/shellprofile.d: not a regular file
ssl-client.xml.example          100% 2316     57.9KB/s  00:00

```

```

scp /usr/local/hadoop/etc/hadoop/* hadoop-
slave2:/usr/local/hadoop/etc/hadoop/

```

```

hadoopuser@hadoop-master:~$ scp /usr/local/hadoop/etc/hadoop/* hadoop-slave2:/u
sr/local/hadoop/etc/hadoop/
capacity-scheduler.xml          100%  8260     1.7MB/s  00:00
configuration.xsl               100% 1335    128.3KB/s  00:00
container-executor.cfg          100% 1940     1.9MB/s  00:00
core-site.xml                   100%   864    737.1KB/s  00:00
hadoop-env.cmd                 100% 3999     3.2MB/s  00:00
hadoop-env.sh                   100%  16KB    1.3MB/s  00:00
hadoop-metrics2.properties     100% 3321     2.9MB/s  00:00
hadoop-policy.xml              100%   11KB  1.4MB/s  00:00
hadoop-user-functions.sh.example 100% 3414     2.7MB/s  00:00
hdfs-site.xml                  100% 1051    732.5KB/s  00:00
httpfs-env.sh                  100% 1484     1.1MB/s  00:00
httpfs-log4j.properties        100% 1657     1.5MB/s  00:00
httpfs-signature.secret        100%   21      15.4KB/s  00:00
httpfs-site.xml                100%   620    529.2KB/s  00:00
kms-acls.xml                   100% 3518     2.7MB/s  00:00
kms-env.sh                      100% 1351    987.0KB/s  00:00
kms-log4j.properties           100% 1860     1.8MB/s  00:00
kms-site.xml                    100%   682    302.6KB/s  00:00
log4j.properties                100%   13KB   11.4MB/s  00:00
mapred-env.cmd                 100%   951    902.9KB/s  00:00
mapred-env.sh                   100% 1764     1.6MB/s  00:00
mapred-queues.xml.template     100% 4113     3.8MB/s  00:00
mapred-site.xml                 100%   758    799.8KB/s  00:00
/usr/local/hadoop/etc/hadoop/shellprofile.d: not a regular file
ssl-client.xml.example          100% 2316    228.1KB/s  00:00

```

4th Step: Format HDFS file system

- Format HDFS file system:

```
source /etc/environment  
hdfs namenode -format
```

```
hadoopuser@hadoop-master:~$ source /etc/environment  
hadoopuser@hadoop-master:~$ hdfs namenode -format
```

5th Step: Start HDFS

- Start HDFS

```
start-dfs.sh
```

```
hadoopuser@hadoop-master:~$ start-dfs.sh  
Starting namenodes on [hadoop-master]  
Starting datanodes  
hadoop-slave1: WARNING: /usr/local/hadoop/logs does not exist. Creating.  
hadoop-slave2: WARNING: /usr/local/hadoop/logs does not exist. Creating.  
Starting secondary namenodes [hadoop-master]
```

- Trên máy master, kiểm tra:

```
hadoopuser@hadoop-master:~$ jps  
4138 Jps  
3771 NameNode  
4014 SecondaryNameNode  
hadoopuser@hadoop-master:~$
```

- Trên máy slave1 và slave2

```
hadoopuser@hadoop-slave1:~$ jps  
1808 DataNode  
2024 Jps  
hadoopuser@hadoop-slave1:~$
```

```

hadoopuser@hadoop-slave2:~$ jps
1814 DataNode
2031 Jps
hadoopuser@hadoop-slave2:~$
```

- Check trên web hadoop-master:9870



Overview 'node0:9000' (active)

Started:	Wed Oct 11 12:03:32 +0200 2017
Version:	2.8.1, r20fe5304904fc2f5a18053c389e43cd26f7a70fe
Compiled:	Fri Jun 02 08:14:00 +0200 2017 by vinodkv from branch-2.8.1-private
Cluster ID:	CD-4dbcd40a-870d-4d30-b382-db0aff4f6ab2
Block Pool ID:	BP-1692617712-192.168.56.10-1507716172624

Summary

Security is off.
 Safemode is off.
 7 files and directories, 3 blocks ≈ 10 total filesystem object(s).
 Heap Memory used 36.96 MB of 55.91 MB Heap Memory. Max Heap Memory is 966.69 MB.
 Non Heap Memory used 42.08 MB of 43.06 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	14.16 GB
DFS Used:	1.22 MB (0.01%)
Non DFS Used:	7.2 GB
DFS Remaining:	6.21 GB (43.82%)
Block Pool Used:	1.22 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.01% / 0.01% / 0.00%
Live Nodes	2 (Decommissioned: 0)
Dead Nodes	0 (Decommissioned: 0)
Decommissioning Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0

The screenshot shows a Mozilla Firefox browser window titled "Namenode Information - Mozilla Firefox". The address bar displays "hadoop-master:9870/dfshealth.html#tab". The main content area is titled "In operation". It features a table with the following data:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used
✓ hadoop-slave1:9866 (192.168.205.8:9866)	http://hadoop-slave1:9864	1s	9m	19.56 GB	0	24 KB (0%)
✓ hadoop-slave2:9866 (192.168.205.9:9866)	http://hadoop-slave2:9864	2s	9m	19.56 GB	0	24 KB (0%)

Below the table, a message says "Showing 1 to 2 of 2 entries" with buttons for "Previous", "1", and "Next".

III. Đẩy dữ liệu lên cụm HDFS

Làm trên máy master

Cách 1: Bằng câu lệnh

- Tạo home directory. Tên /data

```
hdfs dfs -mkdir -p /data
```

- Tạo thư mục books

```
hdfs dfs -mkdir books
```

- Tải 1 số file text trên mạng về máy để đẩy lên

```
cd /home/hadoopuser
```

```
wget -O alice.txt
```

```
https://www.gutenberg.org/files/11/11-0.txt
```

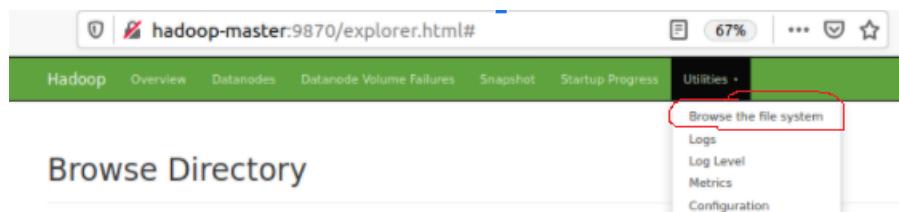
```
wget -O holmes.txt  
https://www.gutenberg.org/files/1661/1661-0.txt
```

```
wget -O frankenstein.txt  
https://www.gutenberg.org/files/84/84-0.txt
```

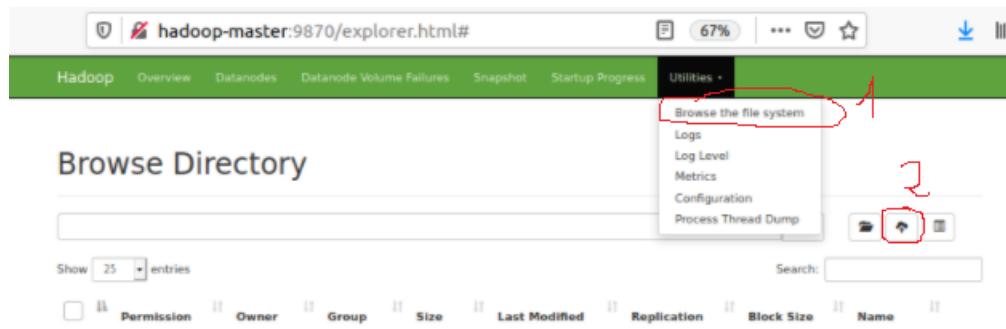
- Up lên cụm HDFS

```
hdfs dfs -put alice.txt holmes.txt frankenstein.txt  
books
```

- Check trên web



Cách 2: Up bằng web



Lab02: Set Up a MapReduce in Hadoop 3.2.1 Multi-Node Cluster on Ubuntu (2 Nodes)

I. Trên 3 máy master, slave1, slave2

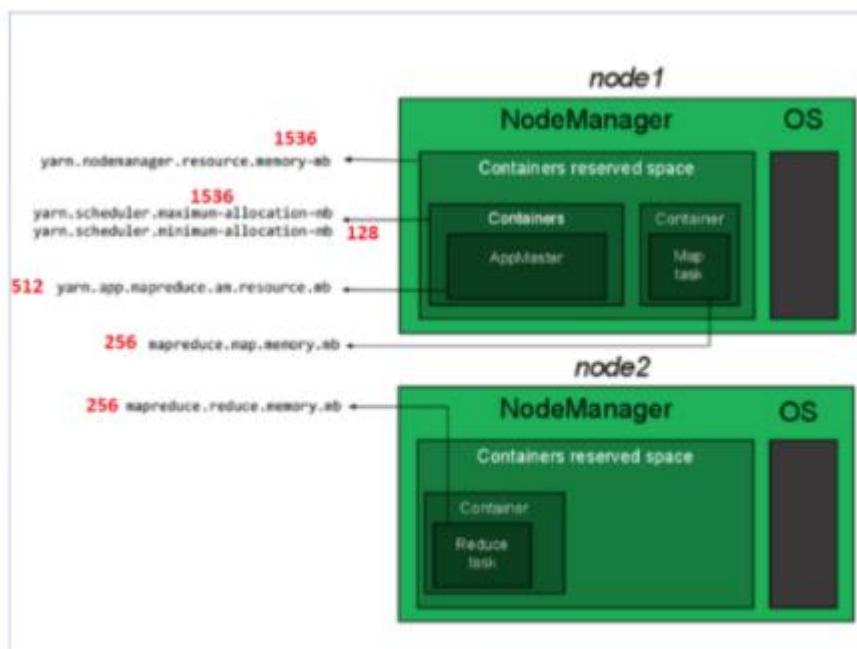


Figure 8: Sample config for 2GB Nodes (will change in next lab)

1st Step: Configure yarn

```
export HADOOP_HOME="/usr/local/hadoop"
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
```

```
hadoopuser@hadoop-master:~$ export HADOOP_HOME="/usr/local/hadoop"
hadoopuser@hadoop-master:~$ export HADOOP_COMMON_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$ export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
hadoopuser@hadoop-master:~$ export HADOOP_HDFS_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$ export HADOOP_MAPRED_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$ export HADOOP_YARN_HOME=$HADOOP_HOME
hadoopuser@hadoop-master:~$
```

2nd Step: Configure mapred-site.xml

```
sudo nano /usr/local/hadoop/etc/hadoop/mapred-site.xml
```

```
<configuration>
<property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
</property>
<property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
</property>
<property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
</property>
<property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
```

```
    </property>

<property>
    <name>yarn.app.mapreduce.am.resource.mb</name>
    <value>512</value>
</property>

<property>
    <name>mapreduce.map.memory.mb</name>
    <value>256</value>
</property>

<property>
    <name>mapreduce.reduce.memory.mb</name>
    <value>256</value>
</property>
</configuration>
```

GNU nano 4.8 /usr/local/hadoop/etc/hadoop/mapred-site.xml

```
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
 Licensed under the Apache License, Version 2.0 (the "License");
 you may not use this file except in compliance with the License.
 You may obtain a copy of the License at

 http://www.apache.org/licenses/LICENSE-2.0

 Unless required by applicable law or agreed to in writing, software
 distributed under the License is distributed on an "AS IS" BASIS,
 WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
 See the License for the specific language governing permissions and
 limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
</property>
<property>
    <name>yarn.app.mapreduce.am.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
</property>
<property>
    <name>mapreduce.map.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
</property>
<property>
    <name>mapreduce.reduce.env</name>
    <value>HADOOP_MAPRED_HOME=$HADOOP_HOME</value>
</property>

<property>
    <name>yarn.app.mapreduce.am.resource.mb</name>
    <value>512</value>
</property>

<property>
    <name>mapreduce.map.memory.mb</name>
    <value>256</value>
</property>

<property>
    <name>mapreduce.reduce.memory.mb</name>
    <value>256</value>
</property>
</configuration>
```

^G Get Help **^O** Write Out **^W** Where Is **^K** Cut Text **^J** Justify **^C** Cur Pos
^X Exit **^R** Read File **^L** Replace **^U** Paste Text **^T** To Spell **^** Go To Line

3rd Step: Config yarn-site.xml

```
sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.xml
```

```
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
    <name>yarn.acl.enable</name>
    <value>0</value>
</property>
<property>
    <name>yarn.resourcemanager.hostname</name>
    <value>hadoop-master</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>

<property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>1536</value>
</property>

<property>
```

```
        <name>yarn.scheduler.maximum-allocation-
mb</name>
        <value>1536</value>
    </property>

<property>
        <name>yarn.scheduler.minimum-allocation-
mb</name>
        <value>128</value>
    </property>

<property>
        <name>yarn.nodemanager.vmem-check-enabled</name>
        <value>false</value>
    </property>
</configuration>
```

```
GNU nano 4.8      /usr/local/hadoop/etc/hadoop/yarn-site.xml
<?xml version="1.0"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<configuration>

<!-- Site specific YARN configuration properties -->
<property>
    <name>yarn.acl.enable</name>
    <value>0</value>
</property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>hadoop-master</value>
</property>
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>

<property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>1536</value>
</property>

<property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>1536</value>
</property>

<property>
    <name>yarn.scheduler.minimum-allocation-mb</name>
    <value>128</value>
</property>

<property>
    <name>yarn.nodemanager.vmem-check-enabled</name>
    <value>false</value>
</property>
</configuration>

[ Read 50 lines ]
^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos
^X Exit      ^R Read File  ^Y Replace  ^U Paste Text^T To Spell  ^L Go To Line
```

II. Trên máy master

1st Step: Start yarn

```
start-yarn.sh
```

```
hadoopuser@hadoop-master:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

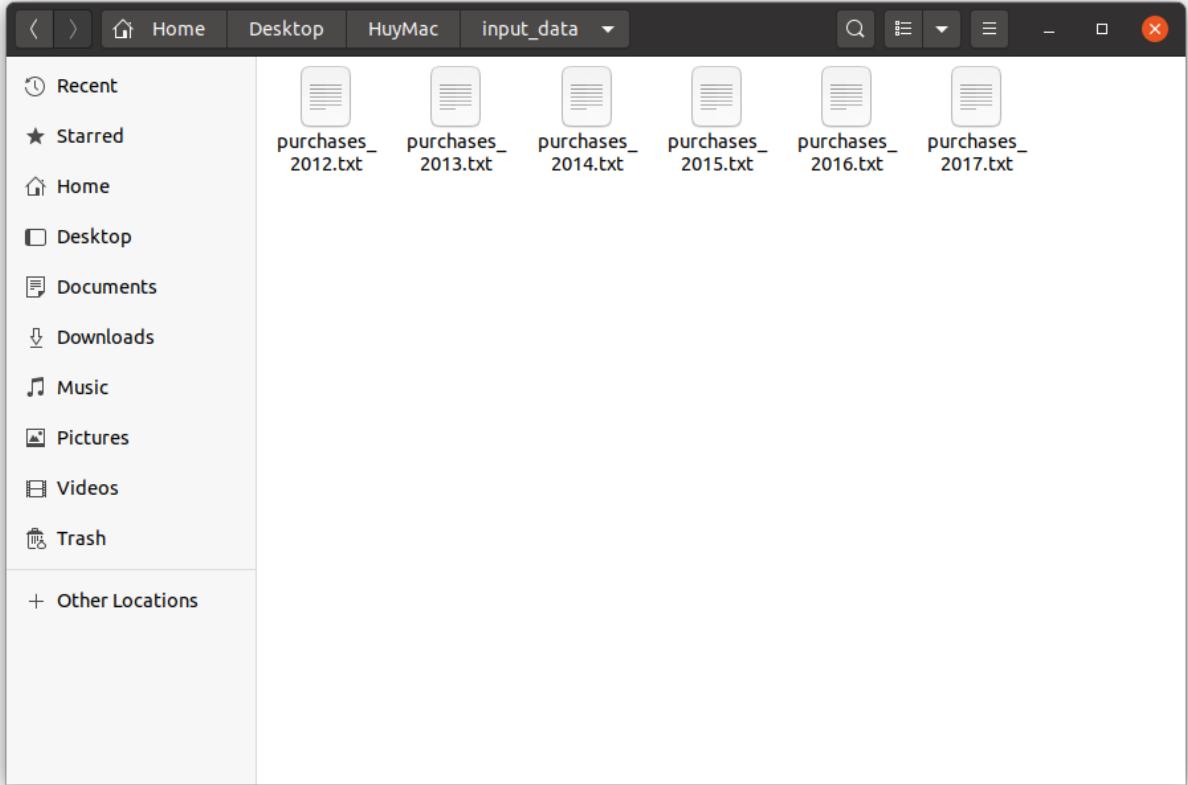
Lên web, truy cập hadoop-master:8088/cluster

The screenshot shows the Hadoop ResourceManager UI running in Mozilla Firefox. The title bar says "All Applications - Mozilla Firefox". The main content area has a "hadoop" logo and the word "hadoop" in blue. On the left, there's a sidebar with a tree view under "Cluster" showing "About", "Nodes", "Node Labels", "Applications" (with sub-options: NEW, NEW_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), and "Scheduler". Below the sidebar is a "Tools" button. The main content area has three sections: "Cluster Metrics" (Shows 0 Apps Submitted, 0 Apps Pending, 0 Apps Running, 0 Apps Completed), "Cluster Nodes Metrics" (Shows 2 Active Nodes, 0 Decommissioning Nodes), and "Scheduler Metrics" (Shows Capacity Scheduler as the Scheduler Type, with Scheduling Resource Type as [memory-mb (unit=Mi), vcores]). A table below shows application details with columns: ID, User, Name, Application Type, Queue, Application Priority, Start Time, and Last Modified. The message "Showing 0 to 0 of 0 entries" is at the bottom.

2nd Step: Tạo MapReduce Job

- Chuẩn bị data: Tài trên https://github.com/MacHuy/HDFS-MultiNode/blob/main/HuyMac/input_data/test.txt

- Hoặc có thể toàn bộ folder HuyMac và move nó để ở Desktop:
<https://github.com/MacHuy/HDFS-MultiNode>

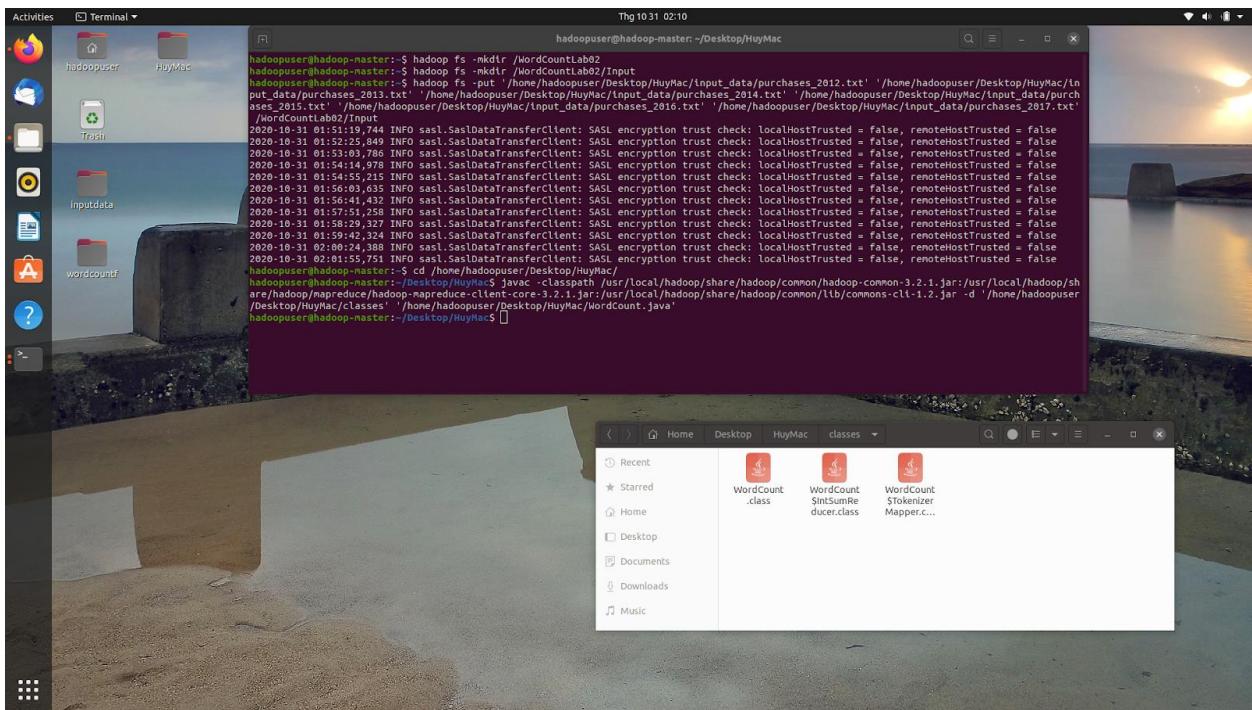


- Updata lên HDFS folder /WordCountLab02/Input
 Thực thi file wordcount.java

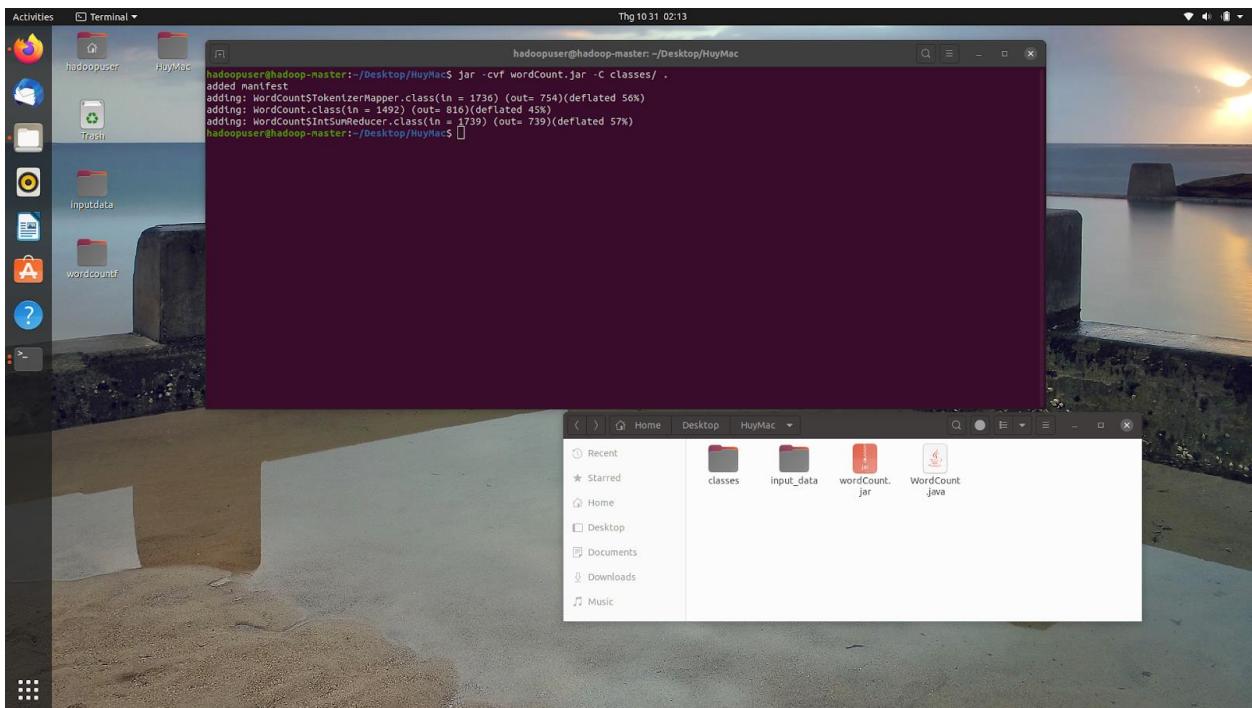
```

hadoopuser@hadoop-master:~$ hadoop fs -mkdir /WordCountLab02
hadoopuser@hadoop-master:~$ hadoop fs -mkdir /WordCountLab02/Input
hadoopuser@hadoop-master:~$ hadoop fs -put '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2012.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2013.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2014.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2015.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2016.txt' '/home/hadoopuser/Desktop/HuyMac/input_data/purchases_2017.txt' /WordCountLab02/Input
2020-10-31 01:51:19.744 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:52:25.849 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:53:03.786 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:54:14.978 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:54:55.215 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:56:03.635 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:56:41.432 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:57:51.258 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:58:29.327 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 01:59:42.324 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 02:00:24.388 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2020-10-31 02:01:55.751 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
hadoopuser@hadoop-master:~$ cd /home/hadoopuser/Desktop/HuyMac/
hadoopuser@hadoop-master:~/Desktop/HuyMac$ javac -classpath /usr/local/hadoop/share/hadoop/common/hadoop-common-3.2.1.jar:/usr/local/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.2.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-cli-1.2.jar -d '/home/hadoopuser/Desktop/HuyMac/classes' '/home/hadoopuser/Desktop/HuyMac/WordCount.java'
hadoopuser@hadoop-master:~/Desktop/HuyMac$ 
```

- Xuất hiện 3 file class trong folder /classes



- Tạo file jar



- Sử dụng file jar thực thi chạy job mapreduce

- Lên web, check job đang chạy

- #### - Xem kết quả

```
hadoopuser@hadoop-master:~$ hdfs dfs -cat /WordCountTutorial/Output/*
```

```

hadoopuser@hadoop-master: ~/Desktop/HuyMac
Nashville 239124
New 481668
Newark 243462
Norfolk 241386
North 240078
Oakland 238368
Oklahoma 242676
Omaha 241254
Orlando 241170
Orleans 239464
Paso 239292
Paul 240960
Pet 1375332
Petersburg 248558
Philadelphia 244488
Phoenix 241998
Pittsburgh 242148
Plano 241020
Portland 240390
Raleigh 241566
Reno 241524
Richmond 239898
Riverside 239778
Rochester 242730
Rouge 242322
Sacramento 243366
Saint 240960
San 1200120
Santa 241836
Scoville 241038
Seattle 239196
Spokane 241332
Sporting 1379592
Springs 242334
St. 480450
Stockton 239976
Supplies 1375332
Tampa 240816
Toledo 240834
Toys 1379784
Tucson 239220
Tulsa 241482
Vegas 481068
Video 1381422
Virginia 241014
Vista 4963326
Vista 240468
Washington 243018
Wayne 242634
Wichita 242532
Winston-Salem 241248
Women's 1388360
North 241116
York 242184
and 1378002
hadoopuser@hadoop-master: ~/Desktop/HuyMac

```

TH chạy lại job, nhớ đổi sang file output mới, folder output cũ đã có, HDFS ko cho phép overwrite

```

Activities Terminal Flg 10:31 02/27
hadoopuser@hadoop-master: ~/Desktop/HuyMac
Map Input records=24830856
Map output records=167897378
Map output bytes=1939467624
Map output compressed bytes=7735584
Input split bytes=1560
Combine Input records=176928474
Combine Input records=1641156
Reduce Input records=1560
Reduce shuffle bytes=7735584
Reduce Input records=610452
Reduce Output records=52878
Splitted Records=52608
Shuffled Maps =12
Failed Shuffles=0
Map Reduces=12
GC time elapsed (ms)=9988
CPU time spent (ms)=351988
Physical Memory (bytes) snapshot=375527248
Virtual Memory (bytes) snapshot=24765580432
Total committed heap usage (bytes)=2697461768
Peak Map Physical memory (bytes)=39939440
Peak Map Virtual memory (bytes)=191591824
Peak Reduce Physical memory (bytes)=3985248
Peak Reduce Virtual memory (bytes)=1921873152
Shuffle Errors
 0 File Underreplicated
 0 CONNECTION=0
 0 TD_ERROR=0
 0 WRONG_LENGTH=0
 0 UNKNOWN=0
 0 WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=126761150
  File Output Counters
    Bytes Written=578924
hadoopuser@hadoop-master: ~/Desktop/HuyMac$ hadoop fs -ls /home/hadoopuser/Desktop/HuyMac/wordCount.jar' wordCount /WordCountLab02/Input /WordCountLab02/Output2
2028-10-31 02:27:22,722 [main] INFO org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://hadoop-master:9000/WordCountLab02/Output2 already exists
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException: Output directory hdfs://hadoop-master:9000/WordCountLab02/Output2 already exists
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSpecs(FileOutputFormat.java:164)
        at org.apache.hadoop.mapreduce.JobSubmitter.checkSpecs(JobSubmitter.java:277)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitter.java:143)
        at org.apache.hadoop.mapreduce.Job$1.run(Job.java:1570)
        at org.apache.hadoop.mapreduce.Job$1.run(Job.java:1567)
        at java.security.AccessController.doPrivileged(Native Method)
        at java.security.PrivilegedActionInvoker.invoke(PrivilegedActionInvoker.java:62)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1738)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1567)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1588)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
        at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
        at java.lang.reflect.Method.invoke(Method.java:498)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:323)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:236)
hadoopuser@hadoop-master: ~/Desktop/HuyMac$ 

```

Trường hợp chỉ có 1 datanode slave1

The screenshot shows the Hadoop Web UI interface. The title bar indicates it's running on port 8088. The main content area is titled "RUNNING Applications". A table lists the following information:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1604078433559_0004	hadoopuser	GuiaChungTa	MAPREDUCE	default	0	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	RUNNING	UNDEFINED	4	1280	0	0	83.3	83.3	0.0	0.0	ApplicationMaster	0

At the bottom of the table, it says "Showing 1 to 1 of 1 entries".

- Check các job finished

The screenshot shows the Hadoop Web UI interface. The title bar indicates it's running on port 8088. The main content area is titled "FINISHED Applications". A table lists the following information:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
application_1604078433559_0004	hadoopuser	GuiaChungTa	MAPREDUCE	default	0	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	0.0	History	0
application_1604078433559_0003	hadoopuser	GuiaChungTa	MAPREDUCE	default	0	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	0.0	History	0
application_1604078433559_0002	hadoopuser	GuiaChungTa	MAPREDUCE	default	0	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	0.0	History	0
application_1604078433559_0001	hadoopuser	word count	MAPREDUCE	default	0	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	Sat Oct 31 10:12:00 +0700 2020	FINISHED	SUCCEEDED	N/A	N/A	N/A	N/A	0.0	0.0	0.0	0.0	History	0

At the bottom of the table, it says "Showing 1 to 4 of 4 entries".

Toàn bộ về Hadoop 3.2.1 Multi-Node Cluster and Mapreduce Job:

<https://github.com/MacHuy/HDFS-MultiNode> (pdf file)