

# **BÁO CÁO LAB1 LƯU TRỮ XỬ LÝ DỮ LIỆU LỚN : HDFS**

**Nhóm: squad game**

**Thành viên: Lại Ngọc Thăng Long 20183581**

**Nguyễn Đình Dũng 20183506**

**Nguyễn Thành Long 20183586**

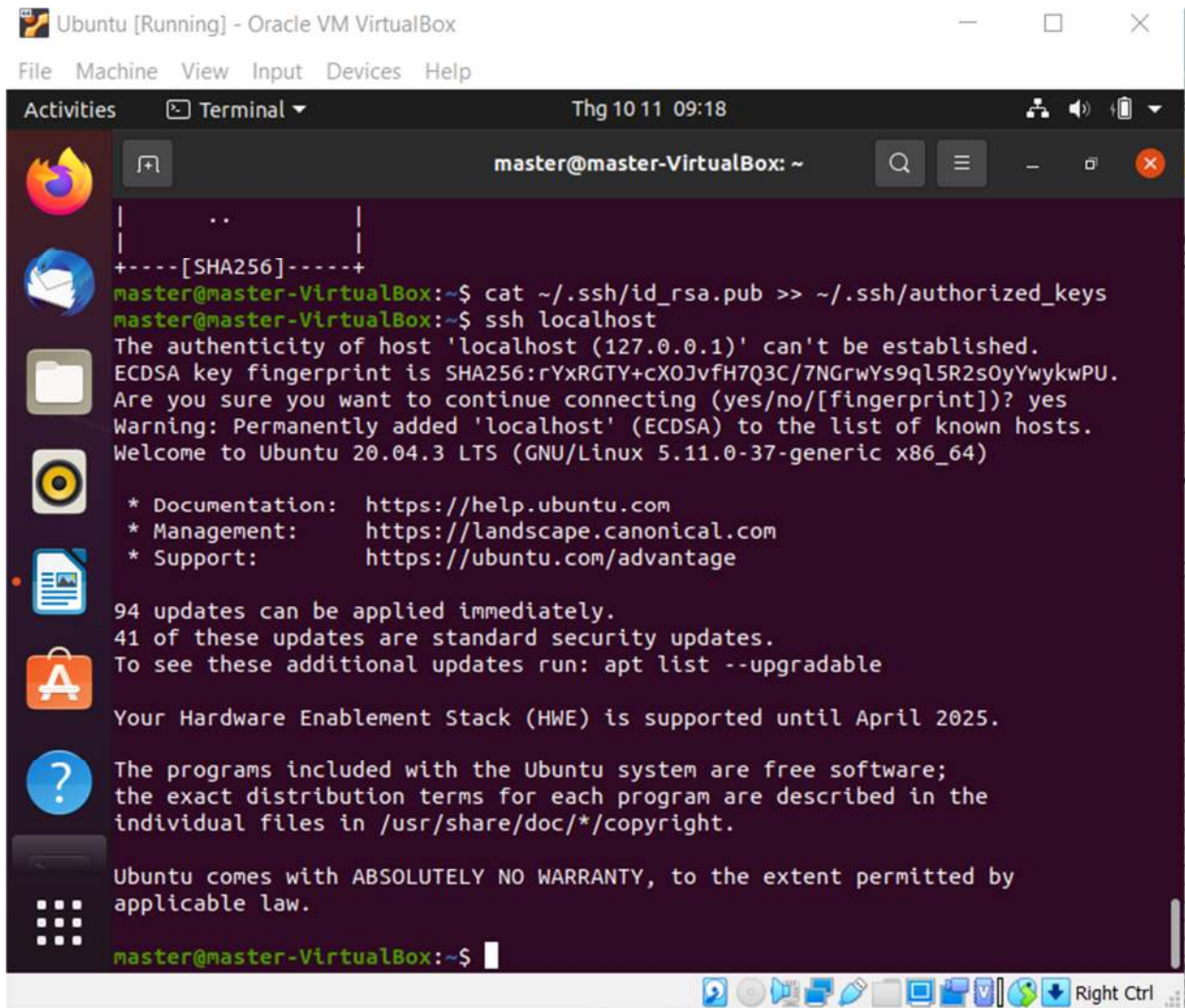
**Nguyễn Khương Duy 20183513**

**BÀI LÀM:**

## **I. Trên toàn bộ máy**

### **1st Step: Cài đặt SSH, PDSH**

Cài đặt thành công SSH, PDSH



The screenshot shows a terminal window titled "Ubuntu [Running] - Oracle VM VirtualBox". The terminal is running on a machine named "master@master-VirtualBox". The user has added their local SSH key to the authorized\_keys file and then attempted to connect to localhost. The terminal output shows the SSH connection process, including the warning about the authenticity of the host and the successful connection to Ubuntu 20.04.3 LTS.

```
..
+----[SHA256]-----+
master@master-VirtualBox:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
master@master-VirtualBox:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:rYxRGTY+cX0JvfH7Q3C/7NGrWys9ql5R2s0yYwykwPU.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.11.0-37-generic x86_64)

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

94 updates can be applied immediately.
41 of these updates are standard security updates.
To see these additional updates run: apt list --upgradable

Your Hardware Enablement Stack (HWE) is supported until April 2025.

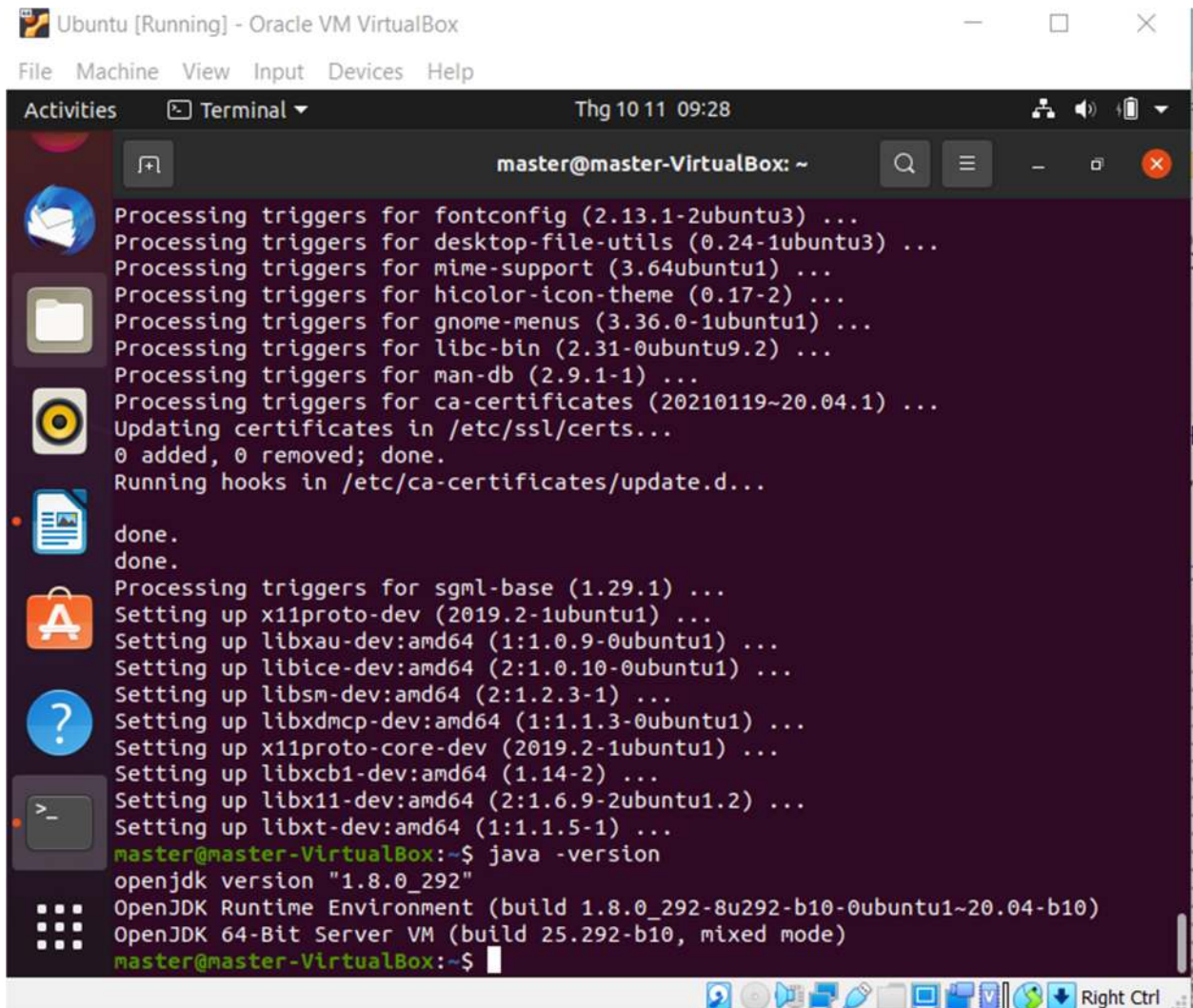
The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.

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

master@master-VirtualBox:~$
```

## 2nd Step: Cài đặt J□v□ 8

Cài thành công Java



Ubuntu [Running] - Oracle VM VirtualBox

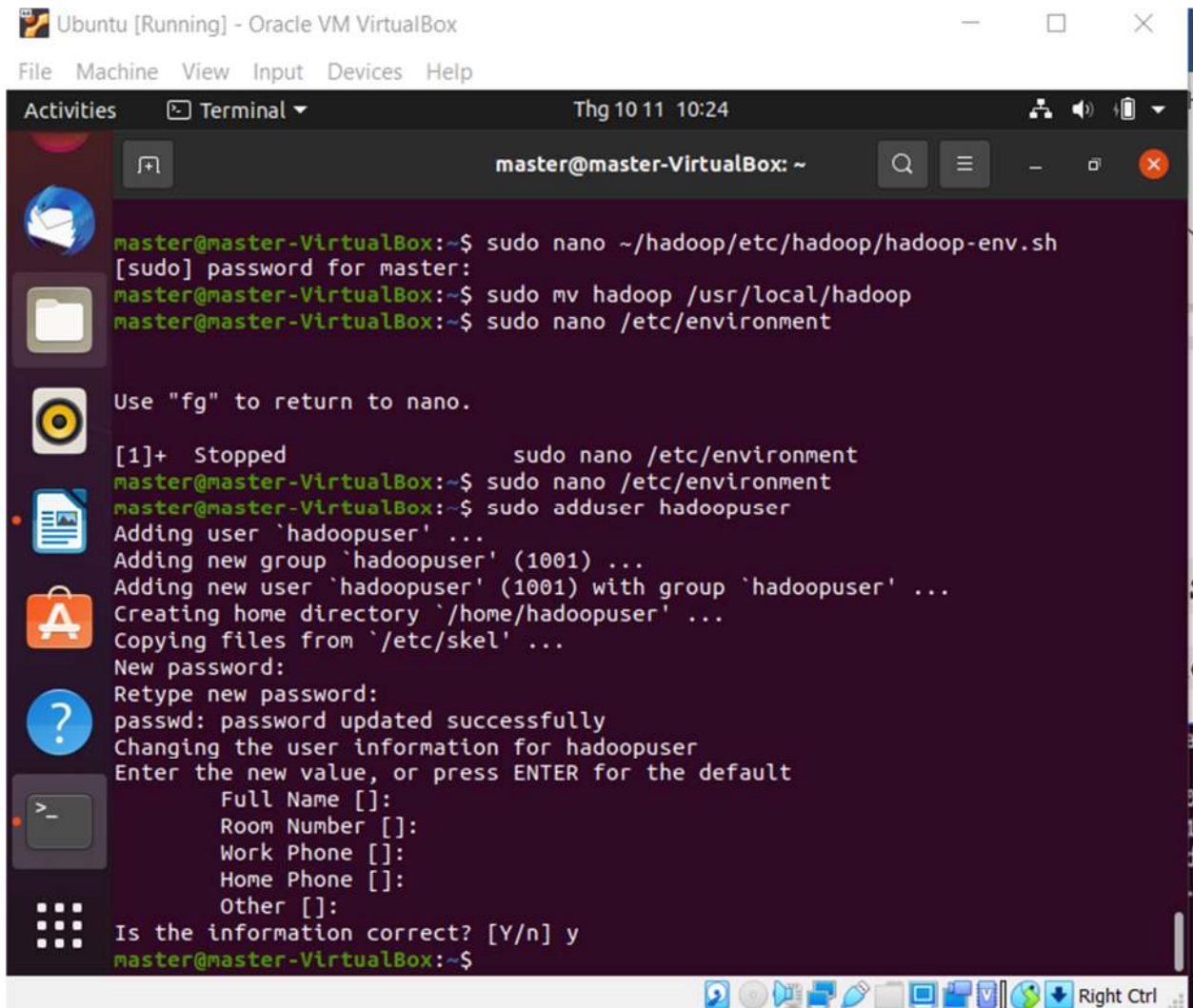
File Machine View Input Devices Help

Activities Terminal Thg 10 11 09:28

```
master@master-VirtualBox: ~  
Processing triggers for fontconfig (2.13.1-2ubuntu3) ...  
Processing triggers for desktop-file-utils (0.24-1ubuntu3) ...  
Processing triggers for mime-support (3.64ubuntu1) ...  
Processing triggers for hicolor-icon-theme (0.17-2) ...  
Processing triggers for gnome-menus (3.36.0-1ubuntu1) ...  
Processing triggers for libc-bin (2.31-0ubuntu9.2) ...  
Processing triggers for man-db (2.9.1-1) ...  
Processing triggers for ca-certificates (20210119~20.04.1) ...  
Updating certificates in /etc/ssl/certs...  
0 added, 0 removed; done.  
Running hooks in /etc/ca-certificates/update.d...  
done.  
done.  
Processing triggers for sgml-base (1.29.1) ...  
Setting up x11proto-dev (2019.2-1ubuntu1) ...  
Setting up libxau-dev:amd64 (1:1.0.9-0ubuntu1) ...  
Setting up libice-dev:amd64 (2:1.0.10-0ubuntu1) ...  
Setting up libsm-dev:amd64 (2:1.2.3-1) ...  
Setting up libxdmcp-dev:amd64 (1:1.1.3-0ubuntu1) ...  
Setting up x11proto-core-dev (2019.2-1ubuntu1) ...  
Setting up libxcb1-dev:amd64 (1.14-2) ...  
Setting up libx11-dev:amd64 (2:1.6.9-2ubuntu1.2) ...  
Setting up libxt-dev:amd64 (1:1.1.5-1) ...  
master@master-VirtualBox:~$ java -version  
openjdk version "1.8.0_292"  
OpenJDK Runtime Environment (build 1.8.0_292-8u292-b10-0ubuntu1~20.04-b10)  
OpenJDK 64-Bit Server VM (build 25.292-b10, mixed mode)  
master@master-VirtualBox:~$
```

**3rd Step: Tải Hadoop**

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



```
master@master-VirtualBox:~$ sudo nano ~/hadoop/etc/hadoop/hadoop-env.sh
[sudo] password for master:
master@master-VirtualBox:~$ sudo mv hadoop /usr/local/hadoop
master@master-VirtualBox:~$ sudo nano /etc/environment

Use "fg" to return to nano.

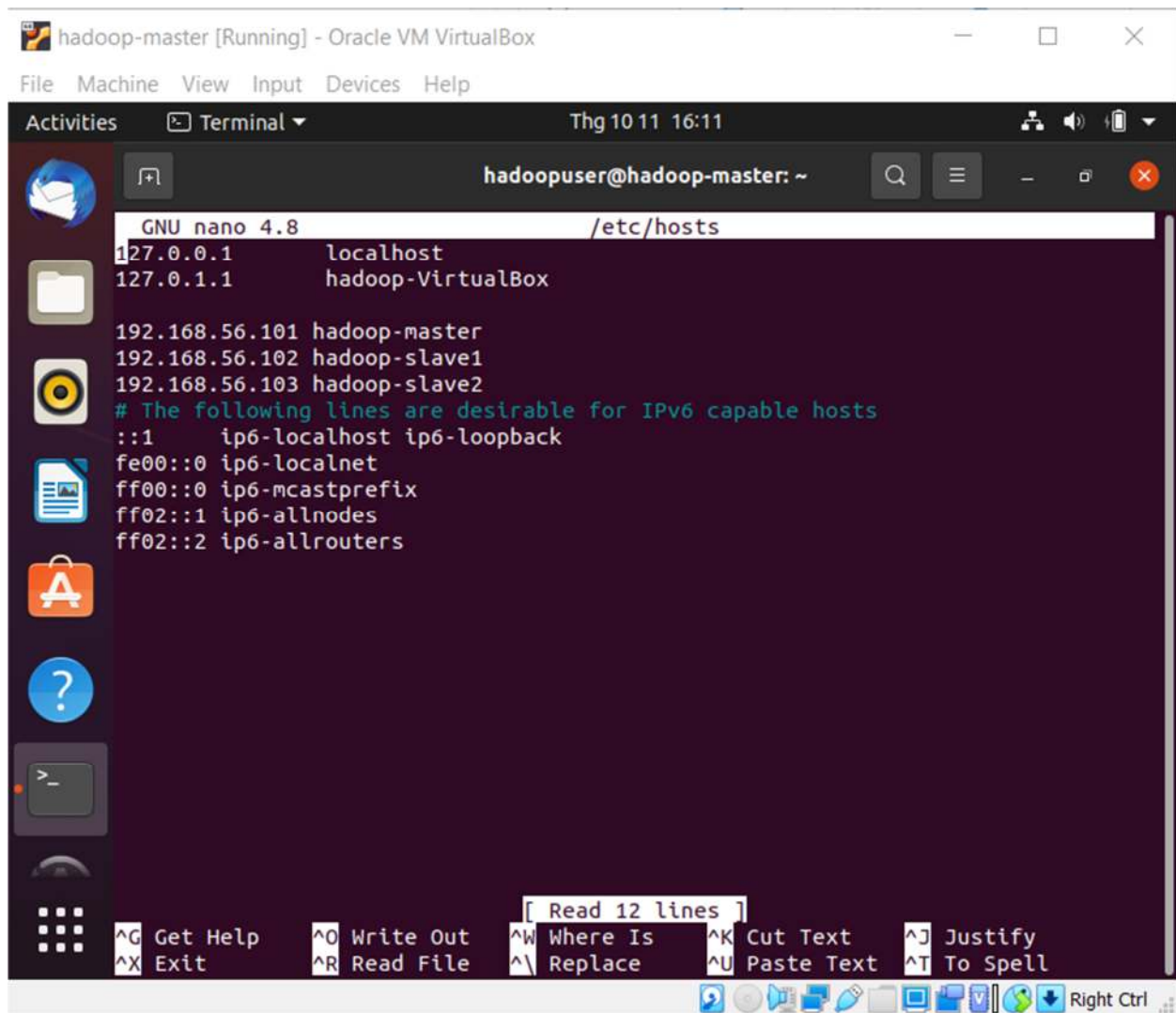
[1]+  Stopped                  sudo nano /etc/environment
master@master-VirtualBox:~$ sudo nano /etc/environment
master@master-VirtualBox:~$ sudo adduser hadoopuser
Adding user `hadoopuser' ...
Adding new group `hadoopuser' (1001) ...
Adding new user `hadoopuser' (1001) with group `hadoopuser' ...
Creating home directory `/home/hadoopuser' ...
Copying files from `/etc/skel' ...
New password:
Retype new 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 []:
Work Phone []:
Home Phone []:
Other []:
Is the information correct? [Y/n] y
master@master-VirtualBox:~$
```

Setup network

<https://www.youtube.com/watch?v=ADmgVT4ovak>

Cấu hình các node





```
GNU nano 4.8 /etc/hosts
127.0.0.1    localhost
127.0.1.1    hadoop-VirtualBox

192.168.56.101 hadoop-master
192.168.56.102 hadoop-slave1
192.168.56.103 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
```

## II. Trên máy hadoop-master

**1st Step: Dùng SSH để kết nối từ h□doop-m□ster tới h□doop-sl□ve1, h□doop-sl□ve2**

**2nd Step: Config các file**

Cấu hình 2 bản sao:

The screenshot shows a terminal window titled 'hadoop-master [Running] - Oracle VM VirtualBox'. The terminal is running the 'nano' text editor, editing the file '/usr/local/hadoop/etc/hadoop/hdfs-site.xml'. The file content includes the Apache License 2.0 text and a configuration block for HDFS properties. The properties being configured are:

- `dfs.namenode.name.dir` set to `/usr/local/hadoop/data/nameNode`
- `dfs.datanode.data.dir` set to `/usr/local/hadoop/data/dataNode`
- `dfs.replication` set to `2`

The terminal window also shows a sidebar with application icons and a bottom status bar with keyboard shortcuts like '^G Get Help', '^X Exit', '^O Write Out', '^R Read File', '^W Where Is', '^\_ Replace', '^K Cut Text', '^U Paste Text', '^J Justify', and '^I To Spell'.

### 3rd Step:

- Gửi các file đã config sang các máy slave

### 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:~$ /usr/local/hadoop/sbin/start-dfs.sh
Starting namenodes on [hadoop-master]
Starting datanodes
Starting secondary namenodes [hadoop-master]
hadoopuser@hadoop-master:~$ jps
```

- Trên máy master, kiểm tra:

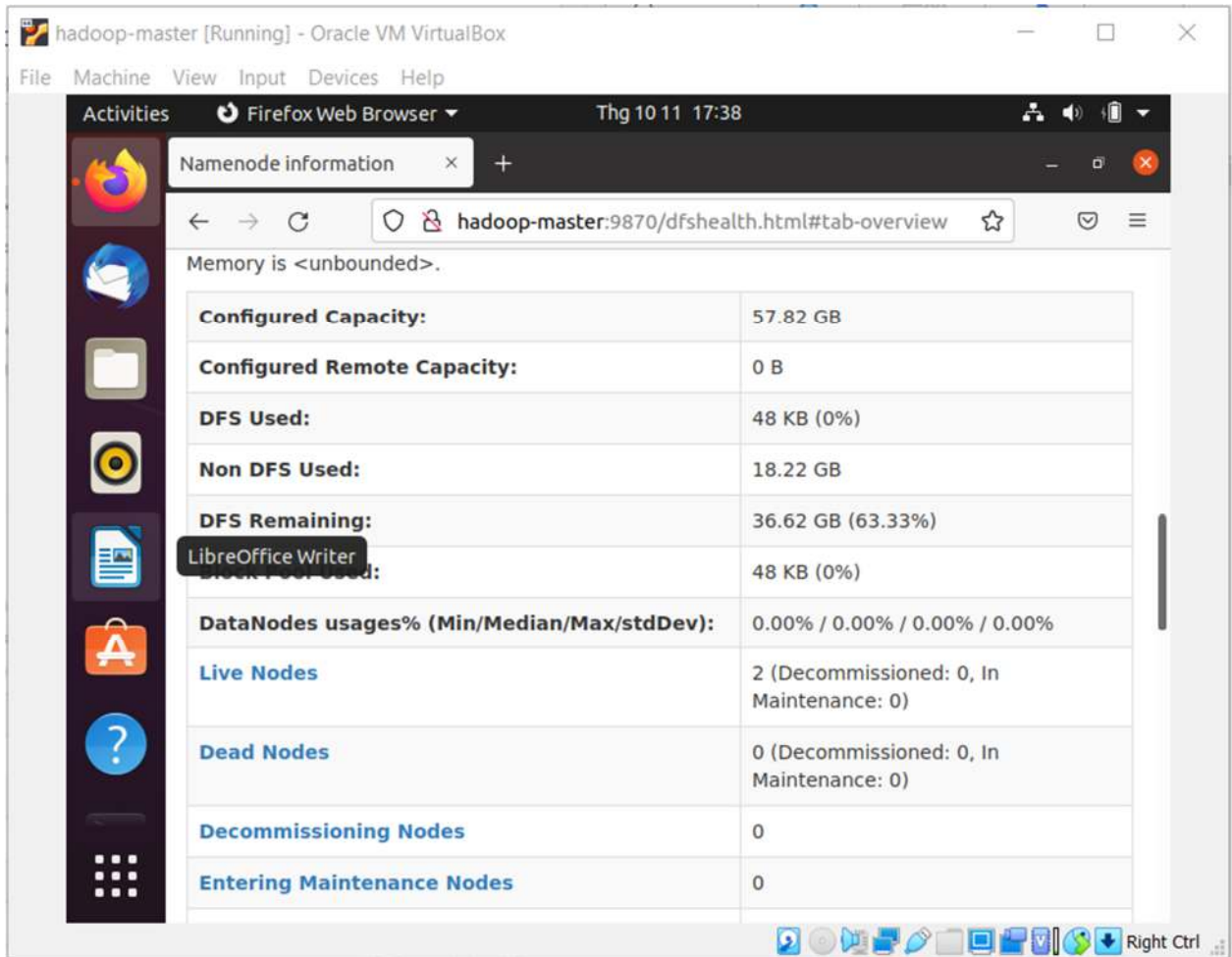
```
hadoopuser@hadoop-master:~$ jps
4520 NameNode
4843 Jps
4749 SecondaryNameNode
hadoopuser@hadoop-master:~$
```

- Trên máy slave1 và slave2

```
hadoopuser@hadoop-slave1: ~
hadoopuser@hadoop-slave1:~$ jps
3093 Jps
3021 DataNode
hadoopuser@hadoop-slave1:~$
```

```
hadoopuser@hadoop-slave2: ~
hadoopuser@hadoop-slave2:~$ jps
2994 DataNode
3066 Jps
hadoopuser@hadoop-slave2:~$
```

Check trên web:





hadoop-master [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Activities Firefox Web Browser Thg 10 11 17:39

Namenode information x +

hadoop-master:9870/dfshealth.html#tab-datanode

In operation

Show 25 entries

Search:

Node	Http Address	Last contact	Last Block Report	Capacity	Blocks	Block pool used	Ve
✓hadoop-slave1:9866 (192.168.56.102:9866)	<a href="http://hadoop-slave1:9864">http://hadoop-slave1:9864</a>	2s	4m	28.91 GB	0	24 KB (0%)	
✓hadoop-slave2:9866 (192.168.56.103:9866)	<a href="http://hadoop-slave2:9864">http://hadoop-slave2:9864</a>	2s	4m	28.91 GB	0	24 KB (0%)	

Showing 1 to 2 of 2 entries

Previous 1 Next

### III. **Đẩy dữ liệu lên cụm HDFS**

Đẩy 1GB dữ liệu lưu trữ trên HDFS, replication=2

hadoop-master [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Activities Firefox Web Browser Thg 10 11 21:35

Browsing HDFS

hadoop-master:9870/explorer.html#/test

# Directory

Go!

Show 25 entries

Search:

Owner	Group	Size	Last Modified	Replication	Block Size	Name
hadoopuser	supergroup	1000 MB	Oct 11 21:21	2	128 MB	testfile.txt

Showing 1 to 1 of 1 entries

Show Applications Previous 1 Next

Right Ctrl