

VNUHCM-UNIVERSITY OF SCIENCE
FACULTY OF INFORMATION TECHNOLOGY



REPORT
INTRODUCTION TO BIG DATA

< Lab 01 - A Gentle Introduction to Hadoop >

Student: 21127104 - Doan Ngoc Mai
21127129 - Le Nguyen Kieu Oanh
21127229 - Duong Truong Binh
21127616 - Le Phuoc Quang Huy

Lecturer: Le Ngoc Thanh
Nguyen Ngoc Thao

Teaching Assistant: Do Trong Le
Bui Huynh Trung Nam

Class: 21KHDL

Table of Contents

Task Progression	2
Team's Result	2
1 Lab Answer	3
1.1 Setting up Single-node Hadoop Cluster	3
1.2 Huy	9
1.3 Mai	10
1.4 Oanh	17
2 Paper Reading	25
2.1 How do the input keys-values, the intermediate keys-values, and the output keys-values relate?	25
2.2 How does MapReduce deal with node failures?	25
2.3 What is the meaning and implication of locality? What does it use?	26
2.4 Which problem is addressed by introducing a combiner function to the MapReduce model?	27
3 Running a warm-up problem: Word Count	28
3.1 Word Count	28
3.2 Bigrams Count	32
4 Reflection	37
4.1 Challenges and Bugs Encountered	37
4.2 How We Overcame It	37
4.3 Lessons Learned	37
References	39

Task Progression

No. of Task	% Completed
1	100%
2	100%
3	100%
4.1	0%
4.2	0%

Team's Result

ID	Member	Task	% Completed
21127104	Đoàn Ngọc Mai	<ul style="list-style-type: none"> • 1. Setting up Single-node Hadoop Cluster • 2. Paper Reading 	100%
21127129	Lê Nguyễn Kiều Oanh	<ul style="list-style-type: none"> • 1. Setting up Single-node Hadoop Cluster • 3 Running a warm-up problem: Word Count 	100%
21127229	Dương Trường Bình	<ul style="list-style-type: none"> • 1. Setting up Single-node Hadoop Cluster • 3 Running a warm-up problem: Word Count 	100%
21127616	Lê Phước Quang Huy	<ul style="list-style-type: none"> • 1. Setting up Single-node Hadoop Cluster • 2. Paper Reading 	100%

1 Lab Answer

1.1 Setting up Single-node Hadoop Cluster

a Binh

- Create a new user and add it to the *sudo* group

```
Thg 6 25 17:18
duongtruongbinh@Laptop-Acer: ~
duongtruongbinh@Laptop-Acer: $ sudo adduser dtbinh_21127229
[sudo] password for duongtruongbinh:
Adding user `dtbinh_21127229' ...
Adding new group `dtbinh_21127229' (1001) ...
Adding new user `dtbinh_21127229' (1001) with group `dtbinh_21127229' ...
Creating home directory `/home/dtbinh_21127229' ...
Copying files from `/etc/skel' ...
New password:
BAD PASSWORD: The password is a palindrome
Retype new password:
Sorry, passwords do not match.
New password:
BAD PASSWORD: The password is a palindrome
Retype new password:
passwd: password updated successfully
Changing the user information for dtbinh_21127229
Enter the new value, or press ENTER for the default
  Full Name []: 21127229
  Room Number []:
  Work Phone []:
  Home Phone []:
  Other []
Is the information correct? [Y/n] y
duongtruongbinh@Laptop-Acer: $
```

```
Thg 6 25 17:21
dtbinh_21127229@Laptop-Acer: ~
Creating home directory `/home/dtbinh_21127229' ...
Copying files from `/etc/skel' ...
New password:
BAD PASSWORD: The password is a palindrome
Retype new password:
Sorry, passwords do not match.
New password:
BAD PASSWORD: The password is a palindrome
Retype new password:
passwd: password updated successfully
Changing the user information for dtbinh_21127229
Enter the new value, or press ENTER for the default
  Full Name []: 21127229
  Room Number []:
  Work Phone []:
  Home Phone []:
  Other []
Is the information correct? [Y/n] y
duongtruongbinh@Laptop-Acer: $ sudo usermod -aG sudo dtbinh_21127229
duongtruongbinh@Laptop-Acer: $ su - dtbinh_21127229
Password:
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.

dtbinh_21127229@Laptop-Acer: $ sudo su - dtbinh_21127229
[sudo] password for dtbinh_21127229:
[dtbinh_21127229@Laptop-Acer: ~]
```

- Change the shell prompt to show student ID and timestamp by modifying *.bashrc* file

```
Thg 6 25 17:30
dtbinh_21127229@Laptop-Acer: ~
GNU nano 6.2          .bashrc *
# ~/.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
if [ "$USER" == "dtbinh_21127229" ]; then
  PS1='21127229 [\e[1;34m\]\w\[\e[0m\]$ '
fi
export PROMPT_COMMAND="echo -n \$[\$(date +\%Y-\%m-\%d\ \%H:\%M:\%S)\ ]\$ "
```

File Name to Write: .bashrc
 ^D Help M-D DOS Format M-A Append M-B Backup File
 ^C Cancel M-M Mac Format M-P Prepend M-T Browse

```
Thg 6 25 17:31:24
dtbinh_21127229@Laptop-Acer: ~
dtbinh_21127229@Laptop-Acer: $ nano .bashrc
dtbinh_21127229@Laptop-Acer: $ source .bashrc
[2024-06-25 17:31:24] 21127229 -$
```

- Install Java and verify the installation.

```
[2024-06-25 22:33:31] 21127229 $ sudo apt install openjdk-8-jdk -y
[sudo] password for dtbinh_21127229:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni libi
ce-dev libpthread-stubs0-dev libsm-dev libxi-dev libxau-dev libxcb1-dev
  libxdmp-dev libxt-dev openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless xiiproto-dev xorg
-sgml-doctools xtrans-dev
Suggested packages:
  default-jre libice-doc libsm-doc libxcb-doc libxt-doc openjdk-8-demo openjdk-8-source vts
  ualvm fonts-nanum fonts-ipafont-gothic fonts-ipafont-mincho
  fonts-wqy-microhei fonts-wqy-zenhei
The following NEW packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni libi
ce-dev libpthread-stubs0-dev libsm-dev libxi-dev libxau-dev libxcb1-dev
  libxdmp-dev libxt-dev openjdk-8-jdk openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless xi
proto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 20 newly installed, 0 to remove and 4 not upgraded.
Need to get 47.9 MB of archives.
After this operation, 163 MB of additional disk space will be used.
Get:1 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 java-common all 0.72build2 [6.782 B]
Get:2 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openjdk-8-jre-headless amd64 8u
412-ga-1-22.94.1 [30.8 MB]
Get:3 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 ca-certificates-java all 20190909ub
untu1.2 [12.1 kB]
Get:4 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-extra all 2.37-2build1 [2.041
kB]
Get:5 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java all 0.38.0-5build1 [53
1 kB]
Get:6 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java-jni amd64 0.38.0-5build
1 [49.0 kB]
Get:7 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 xorg-sgml-doctools all 1:1.11-1.1 [10.9 kB]
Get:8 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 xiiproto-dev all 2021.5-1 [604 kB]
Get:9 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libice-dev amd64 2:1.0.10-1build2 [51.4 kB]
Get:10 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libpthread-stubs0-dev amd64 0.4-1build2 [5
```

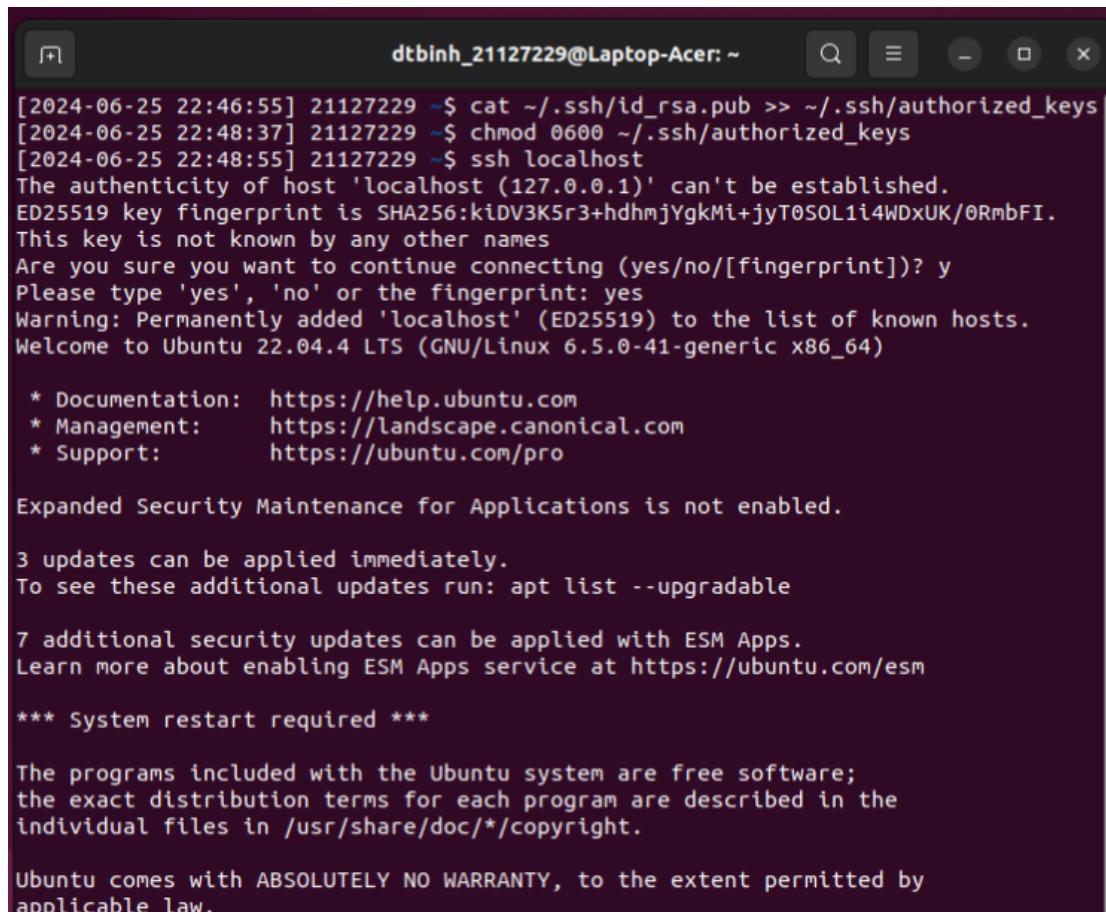
```
[2024-06-25 22:37:49] 21127229 $ java -version
openjdk version "1.8.0_412"
OpenJDK Runtime Environment (build 1.8.0_412-8u412-ga-1-22.04.1-b08)
OpenJDK 64-Bit Server VM (build 25.412-b08, mixed mode)
[2024-06-25 22:37:52] 21127229 $ javac -version
javac 1.8.0_412
[2024-06-25 22:37:56] 21127229 $
```

- Install the *OpenSSH* server and client. Next, generate the public and private keys.

```
[2024-06-25 22:38:39] 21127229 $ sudo apt install openssh-server openssh-client -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openssh-client is already the newest version (1:8.9p1-3ubuntu0.7).
openssh-client set to manually installed.
The following additional packages will be installed:
  ncurses-term openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere ssh-askpass
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
0 upgraded, 4 newly installed, 0 to remove and 4 not upgraded.
Need to get 752 kB of archives.
After this operation, 6.050 kB of additional disk space will be used.
Get:1 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-sftp-server amd64
  1:8.9p1-3ubuntu0.7 [38.9 kB]
Get:2 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-server amd64 1:8.
9p1-3ubuntu0.7 [435 kB]
Get:3 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 ncurses-term all 6.3-2ubu
ntu0.1 [267 kB]
Get:4 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 ssh-import-id all 5.11-0ubuntu1 [10.1 kB]
Fetched 752 kB in 1s (513 kB/s)
Preconfiguring packages ...
Selecting previously unselected package openssh-sftp-server.
(Reading database ... 201631 files and directories currently installed.)
Preparing to unpack .../openssh-sftp-server_1%3a8.9p1-3ubuntu0.7_amd64.deb ...
Unpacking openssh-sftp-server (1:8.9p1-3ubuntu0.7) ...
Selecting previously unselected package openssh-server.
Preparing to unpack .../openssh-server_1%3a8.9p1-3ubuntu0.7_amd64.deb ...
```

```
[2024-06-25 22:43:09] 21127229 $ ssh-keygen -t rsa -P '' -f ~/ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/dtbinh_21127229/.ssh'.
Your identification has been saved in /home/dtbinh_21127229/.ssh/id_rsa
Your public key has been saved in /home/dtbinh_21127229/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:VnGhtnvIir+wf6nEI7p6WrVJgWvcFDr5hRKF2sO3DlE dtbinh_21127229@Laptop-Acer
The key's randomart image is:
+---[RSA 3072]----+
| .+o o. |
| .+=* |
| +* .B . |
| ..=X.+ |
| S+*. . |
| o.o.= o |
| .OB * .. |
| ...O oo |
| +oO...=+ |
+---[SHA256]----+
[2024-06-25 22:43:55] 21127229 $
```

- Add the public key to *authorized_keys*. Use *chmod* to change the file permissions of *authorized_keys*. Finally, verify the SSH configuration.



```
[2024-06-25 22:46:55] 21127229 ~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
[2024-06-25 22:48:37] 21127229 ~$ chmod 0600 ~/.ssh/authorized_keys
[2024-06-25 22:48:55] 21127229 ~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:kiDV3K5r3+hdhmjYgkMi+jyT0SOL1i4WDxUK/0RmbFI.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? y
Please type 'yes', 'no' or the fingerprint: yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 6.5.0-41-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

3 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

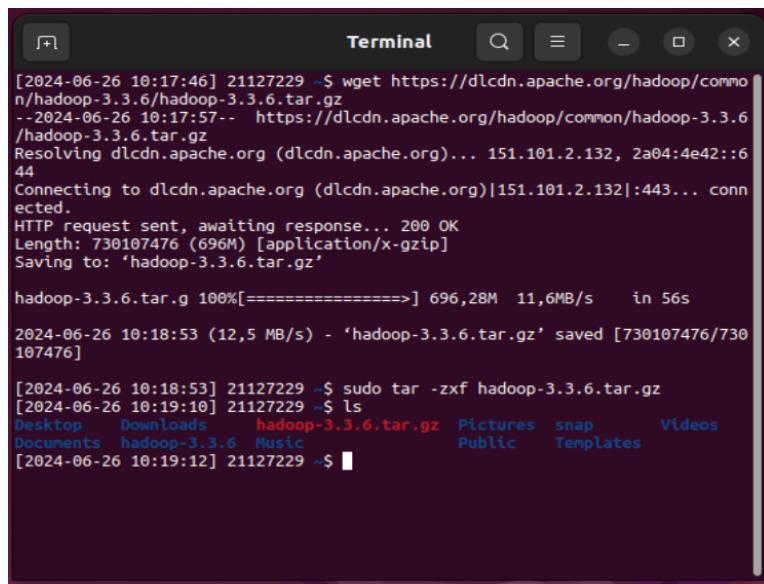
7 additional security updates can be applied with ESM Apps.
Learn more about enabling ESM Apps service at https://ubuntu.com/esm

*** System restart required ***

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.
```

- Install and extract Hadoop version 3.3.6. Create the directories *data/datanode* and *data/namenode*.

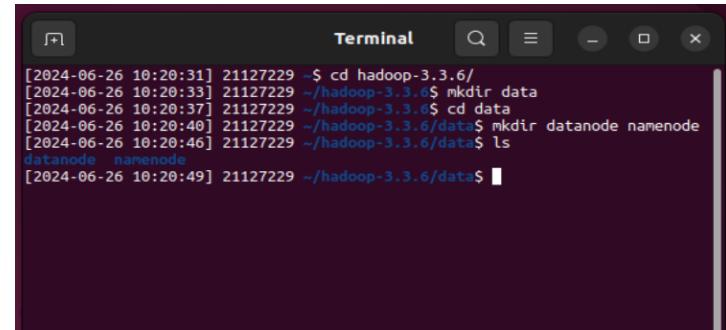


```
[2024-06-26 10:17:46] 21127229 ~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
--2024-06-26 10:17:57-- https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::64
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 730107476 (696M) [application/x-gzip]
Saving to: 'hadoop-3.3.6.tar.gz'

hadoop-3.3.6.tar.g 100%[=====] 696,28M 11,6MB/s   in 56s

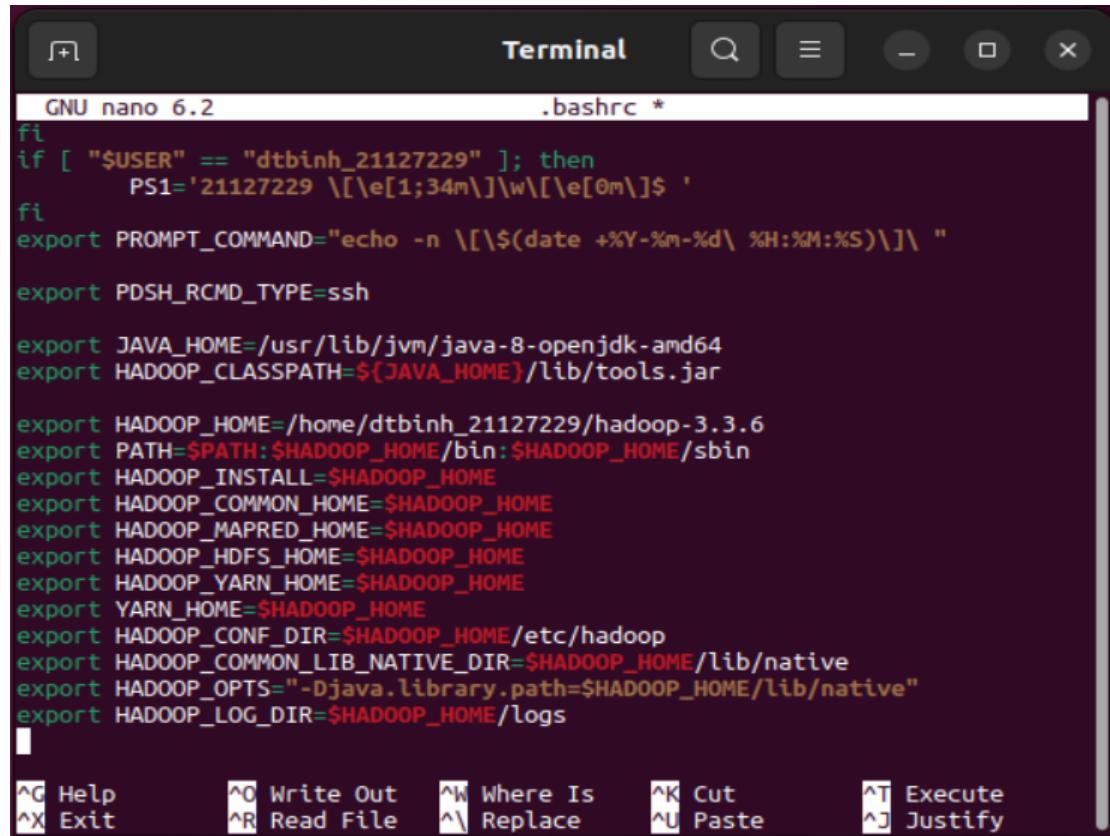
2024-06-26 10:18:53 (12,5 MB/s) - 'hadoop-3.3.6.tar.gz' saved [730107476/730107476]

[2024-06-26 10:18:53] 21127229 ~$ sudo tar -zxf hadoop-3.3.6.tar.gz
[2024-06-26 10:19:10] 21127229 ~$ ls
Desktop  Downloads  hadoop-3.3.6.tar.gz  Pictures  snap      Videos
Documents  hadoop-3.3.6  Music          Public    Templates
[2024-06-26 10:19:12] 21127229 ~$
```



```
[2024-06-26 10:20:31] 21127229 ~$ cd hadoop-3.3.6/
[2024-06-26 10:20:33] 21127229 ~/hadoop-3.3.6$ mkdir data
[2024-06-26 10:20:37] 21127229 ~/hadoop-3.3.6$ cd data
[2024-06-26 10:20:40] 21127229 ~/hadoop-3.3.6$ mkdir datanode namenode
[2024-06-26 10:20:46] 21127229 ~/hadoop-3.3.6$ ls
datanode  namenode
[2024-06-26 10:20:49] 21127229 ~/hadoop-3.3.6$
```

- Modify the *.bashrc* file to configure the Hadoop environment variables and use the command *source ./bashrc* to apply the changes.



```

GNU nano 6.2          .bashrc *

if [ "$USER" == "dtbinh_21127229" ]; then
    PS1='21127229 \[\e[1;34m\]\w\[\e[0m\]$ '
fi
export PROMPT_COMMAND="echo -n \[$(date +%Y-%m-%d\ %H:%M:%S)\]\\" "
export PDSH_RCMD_TYPE=ssh

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar

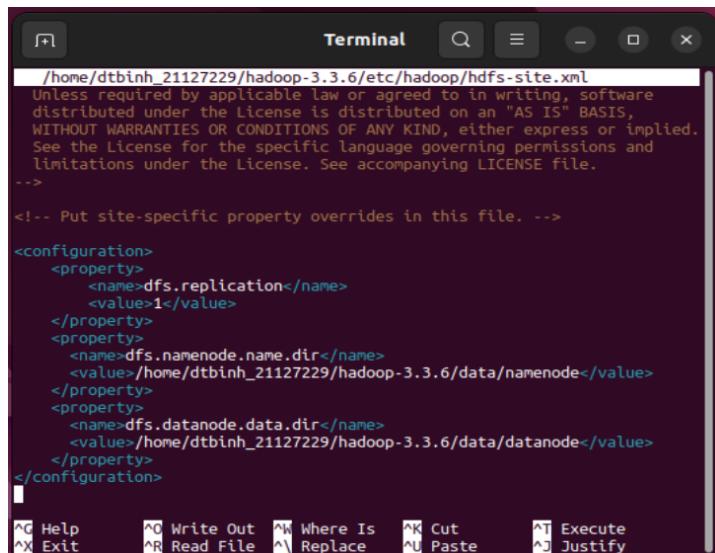
export HADOOP_HOME=/home/dtbinh_21127229/hadoop-3.3.6
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_LOG_DIR=$HADOOP_HOME/logs

```

Toolbar:

- Help
- Write Out
- Where Is
- Cut
- Execute
- Exit
- Read File
- Replace
- Paste
- Justify

- Configure Java environment variables by modifying the files *hadoop-env.sh*, *core-site.xml*, *hdfs-site.xml*, *mapred-site.xml*, and *yarn-site.xml*.



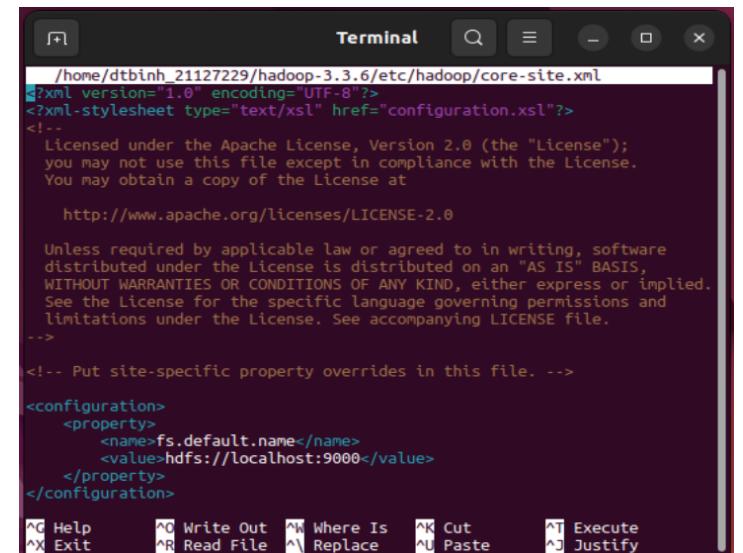
```

<!-- Put site-specific property overrides in this file. -->
<configuration>
    <property>
        <name>dfs.replication</name>
        <value>1</value>
    </property>
    <property>
        <name>dfs.namenode.name.dir</name>
        <value>/home/dtbinh_21127229/hadoop-3.3.6/data/namenode</value>
    </property>
    <property>
        <name>dfs.datanode.data.dir</name>
        <value>/home/dtbinh_21127229/hadoop-3.3.6/data/datanode</value>
    </property>
</configuration>

```

Toolbar:

- Help
- Write Out
- Where Is
- Cut
- Execute
- Exit
- Read File
- Replace
- Paste
- Justify



```

<!-- Put site-specific property overrides in this file. -->
<configuration>
    <property>
        <name>fs.default.name</name>
        <value>hdfs://localhost:9000</value>
    </property>
</configuration>

```

Toolbar:

- Help
- Write Out
- Where Is
- Cut
- Execute
- Exit
- Read File
- Replace
- Paste
- Justify

The image shows two terminal windows side-by-side. The left window displays the contents of `/home/dtbinh_21127229/hadoop-3.3.6/etc/hadoop/mapred-site.xml`. It includes the Apache License 2.0 header and XML configuration for the MapReduce framework, specifying the name as `yarn` and the classpath as `$SHADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$SHADOOP_MAPRED_HOME/lib/*`. The right window displays the contents of `/home/dtbinh_21127229/hadoop-3.3.6/etc/hadoop/yarn-site.xml`. It contains site-specific YARN configuration properties, such as `yarn.nodemanager.aux-services` set to `mapreduce_shuffle`, `yarn.resourcemanager.hostname` set to `127.0.0.1`, and `yarn.acl.enable` set to `0`.

```

/home/dtbinh_21127229/hadoop-3.3.6/etc/hadoop/mapred-site.xml
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>mapreduce.application.classpath</name>
    <value>$SHADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$SHADOOP_MAPRED_HOME/lib/*</value>
  </property>
</configuration>

^G Help      ^O Write Out  ^W Where Is  ^K Cut      ^T Execute
^X Exit      ^R Read File  ^M Replace   ^U Paste    ^J Justify

-----[Terminal]----- /home/dtbinh_21127229/hadoop-3.3.6/etc/hadoop/yarn-site.xml

<!-- Site specific YARN configuration properties -->
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>127.0.0.1</value>
  </property>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>
    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH,</value>
  </property>
</configuration>

^G Help      ^O Write Out  ^W Where Is  ^K Cut      ^T Execute
^X Exit      ^R Read File  ^M Replace   ^U Paste    ^J Justify

```

The image shows a terminal window displaying the contents of `/home/dtbinh_21127229/hadoop-3.3.6/etc/hadoop/hadoop-env.sh`. The script provides configuration options for YARN, HDFS, and MapReduce. It defines precedence rules where `yarn-env.sh|dfs-env.sh` takes precedence over `hadoop-env.sh`, which in turn takes precedence over hard-coded defaults. It also handles environment variable substitution and provides examples for setting `JAVA_HOME` and `HADOOP_HOME`.

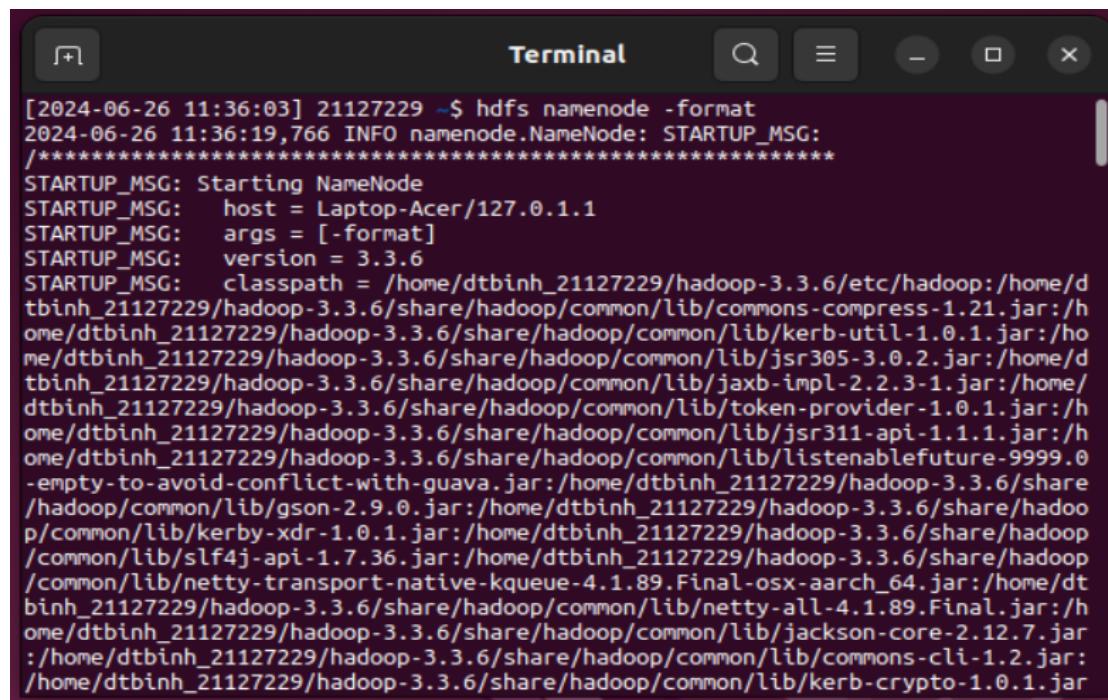
```

/home/dtbinh_21127229/hadoop-3.3.6/etc/hadoop/hadoop-env.sh
## ONE CAN USE THIS FILE TO SET YARN, HDFS, AND MAPREDUCE
## CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.
##
## Precedence rules:
##
## {yarn-env.sh|dfs-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
####

^G Help      ^O Write Out  ^W Where Is  ^K Cut      ^T Execute
^X Exit      ^R Read File  ^M Replace   ^U Paste    ^J Justify

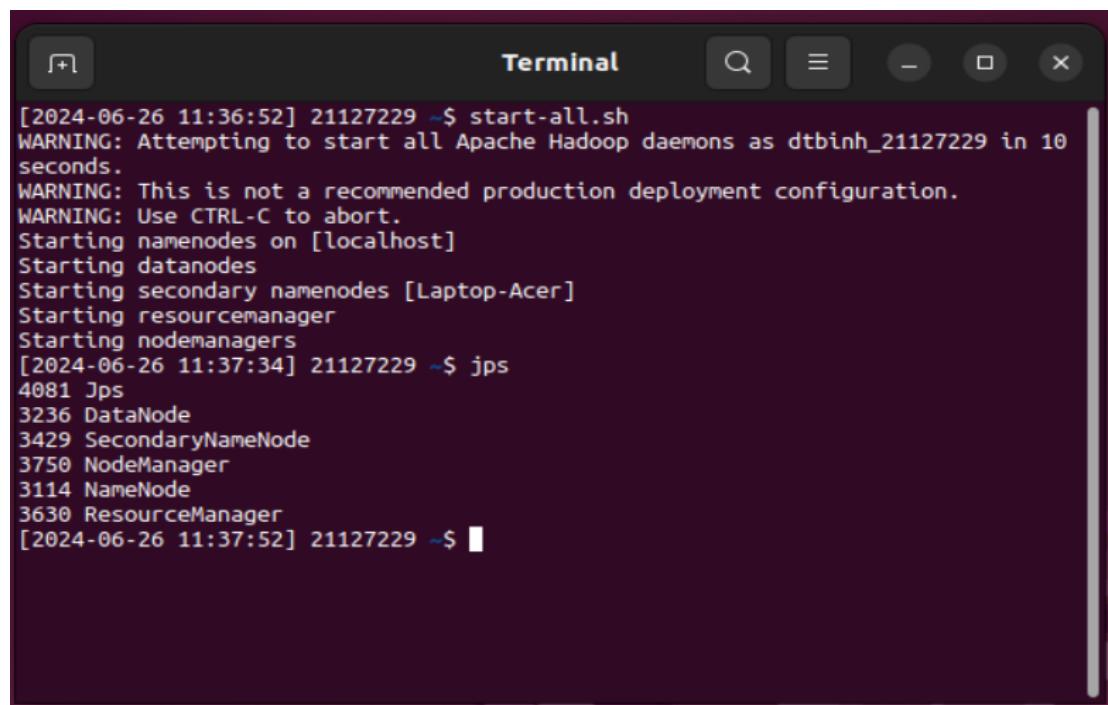
```

- Validate the Hadoop configuration and format the NameNode.



```
[2024-06-26 11:36:03] 21127229 ~$ hdfs namenode -format
2024-06-26 11:36:19,766 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = Laptop-Acer/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.6
STARTUP_MSG: classpath = /home/dtbinh_21127229/hadoop-3.3.6/etc/hadoop:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/commons-compress-1.21.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/kerb-util-1.0.1.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/jsr305-3.0.2.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/jaxb-impl-2.2.3-1.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/token-provider-1.0.1.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/jsr311-api-1.1.1.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/gson-2.9.0.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/kerby-xdr-1.0.1.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/slf4j-api-1.7.36.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/netty-transport-native-kqueue-4.1.89.Final-osx-aarch_64.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/netty-all-4.1.89.Final.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/jackson-core-2.12.7.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/commons-cli-1.2.jar:/home/dtbinh_21127229/hadoop-3.3.6/share/hadoop/common/lib/kerb-crypto-1.0.1.jar
```

- Start the Hadoop cluster



```
[2024-06-26 11:36:52] 21127229 ~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as dtbinh_21127229 in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [Laptop-Acer]
Starting resourcemanager
Starting nodemanagers
[2024-06-26 11:37:34] 21127229 ~$ jps
4081 Jps
3236 DataNode
3429 SecondaryNameNode
3750 NodeManager
3114 NameNode
3630 ResourceManager
[2024-06-26 11:37:52] 21127229 ~$ █
```

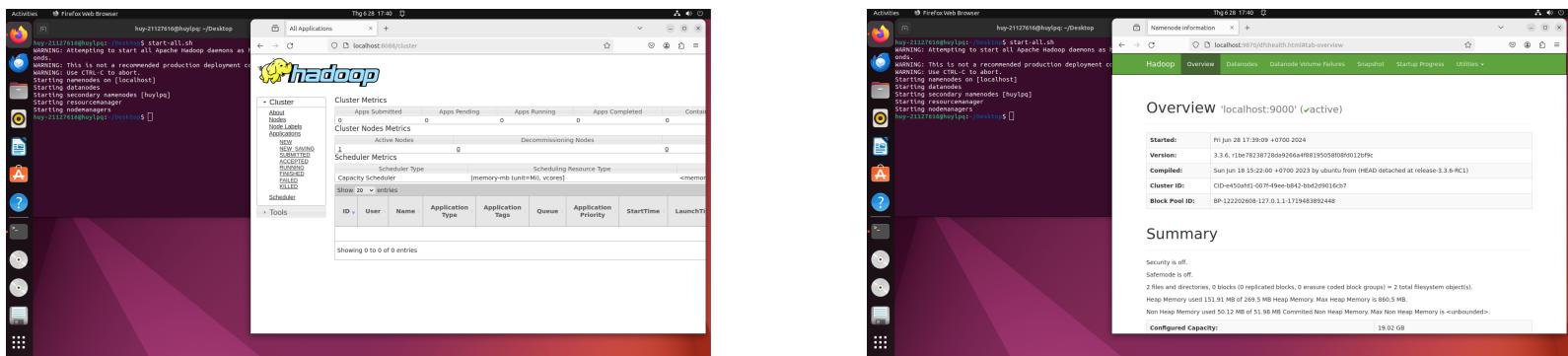
- View the result using the Hadoop Web Interface at *localhost:9870* and *localhost:8088*.

1.2 Huy

- The installation steps are the same for all group members.
- Start Hadoop Cluster

```
huy-21127616@huylpq:~$ start-all.sh
*****SHUTDOWN_MSG: Shutting down NameNode at huylpq/127.0.1.1*****
huy-21127616@huylpq:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as huy-21127616 in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [huylpq]
huylpq: Warning: Permanently added 'huylpq' (ED25519) to the list of known hosts .
Starting resourcemanager
Starting nodemanagers
huy-21127616@huylpq:~$ hdfs dfs -mkdir /input
huy-21127616@huylpq:~$ jps
54128 NameNode
54657 ResourceManager
54774 NodeManager
56040 Jps
54250 DataNode
54461 SecondaryNameNode
huy-21127616@huylpq:~$
```

- Finally, checking the result using the Hadoop Web Interface at *localhost:8088* and *localhost:9870*.



1.3 Mai

- Install Java and make sure that a suitable version of Java installed.

```
doanngomai_21127104@21127104:~$ sudo apt install openjdk-8-jdk -y
[sudo] password for doanngomai_21127104:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libkmpc-1.0 libkmpcbackend-1.0d1
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev
  libxdmcp-dev libxt-dev openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
  default-jre libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-8-demo openjdk-8-source vtslumv fonts-nanum fonts-ipafont-gothic fonts-ipafont-mncho
  fonts-wqy-microhei fonts-wqy-zhenhei
The following NEW packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev
  libxdmcp-dev libxt-dev openjdk-8-jdk openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 20 newly installed, 0 to remove and 5 not upgraded.
Need to get 47,9 MB of archives.
After this operation, 163 MB of additional disk space will be used.
Get:1 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 java-common all 0.72build2 [6.782 B]
Get:2 http://vn.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 openjdk-8-jre-headless all 8u412+ga-1-22.04.1 [30,8 MB]
Get:3 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java-jni amd64 2.37.2-2ubuntu1.2 [12,1 kB]
Get:4 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-extra all 2.37-2buil1 [2,041 kB]
Get:5 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java all 0.38.0-5buil1 [53,1 kB]
Get:6 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java-jni amd64 0.38.0-5buil1 [49,0 kB]
Get:7 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 xorg-sgml-doctools all 1:1.11.1-1 [10,9 kB]
Get:8 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 x11proto-dev all 2021.5-1 [604 kB]
Get:9 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libice-dev amd64 2:1.0.10-1buil2 [51,4 kB]
Get:10 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libpthread-stubs0-dev amd64 0.2.2-2.2-1buil2 [19,1 kB]
Get:11 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libsm-dev amd64 0.2.2-2.2-1buil2 [19,1 kB]
Get:12 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libxau-dev amd64 1:1.0.9-1buil2 [9,724 kB]
Get:13 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libxdmcp-dev amd64 1:1.1.3-0ubuntu5 [26,5 kB]
Get:14 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 xtrans-dev all 1:4.0.1 [68,9 kB]
Get:15 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libxcb1-dev amd64 1.14-3ubuntu3 [86,5 kB]
Get:16 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 libx11-dev amd64 2:1.7.5-1ubuntu0.3 [744 kB]
Get:17 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libxt-dev amd64 1:1.2.1-1 [396 kB]
Get:18 http://vn.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 openjdk-8-jre amd64 8u412+ga-1-22.04.1 [75,3 kB]
```

```
doanngomai_21127104@21127104:~$ java -version
OpenJDK Runtime Environment (build 1.8.0_412-8u412+ga-1-22.04.1-b08)
OpenJDK 64-Bit Server VM (build 25.412-b08, mixed mode)
[2024-06-26 20:26:01] doanngomai_21127104$ javac -version
javac 1.8.0_412
[2024-06-26 20:26:30] doanngomai_21127104$
```

- Install OpenSSH, create an SSH key pair, add the public key to the authorized keys in the

SSH directory, adjust the user permissions using the chmod command, and verify the SSH connection to localhost.

```
dngocmai@21127104:~$ java -version
openjdk version "1.8.0_412"
OpenJDK Runtime Environment (build 1.8.0_412-8u412-ga-1-22.04.1-b08)
OpenJDK 64-Bit Server VM (build 25.412-b08, mixed mode)

dngocmai@21127104:~$ which javac
/usr/bin/javac

dngocmai@21127104:~$ ^C
dngocmai@21127104:~$ readlink -f /usr/bin/javac
/usr/lib/jvm/java-8-openjdk-amd64/bin/javac
dngocmai@21127104:~$ sudo apt install openssh-server openssh-client -y
[sudo] password for dngocmai:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openssh-client is already the newest version (1:8.9p1-3ubuntu0.7).
openssh-server is already the newest version (1:8.9p1-3ubuntu0.7).
The following packages were automatically installed and are no longer required:
libwpe-1.0-1 libwpebackend-fdo-1.0-1
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 5 not upgraded.

dngocmai@21127104:~$
```

```
The key's randomart image is:
+---[RSA 3072]---+
...+o+....|
... o ... o |
o ... + . . |
. . + o . . |
. + S o+o+| |
o + E...+*+o| |
o ..oo+o| |
o ..ooo| |
. .+| |

dngocmai@21127104:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
dngocmai@21127104:~$ chmod 0600 ~/.ssh/authorized_keys
dngocmai@21127104:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
It's address in the host key is somekey (ED25519) but I have seen it before.
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 6.5.0-41-generic x86_64)

 * Documentation: https://Help.ubuntu.com
 * Management: https://landscape.canonical.com
 * Support: https://Ubuntu.com/pro
Expanded Security Maintenance for Applications is not enabled.
0 updates can be applied immediately.

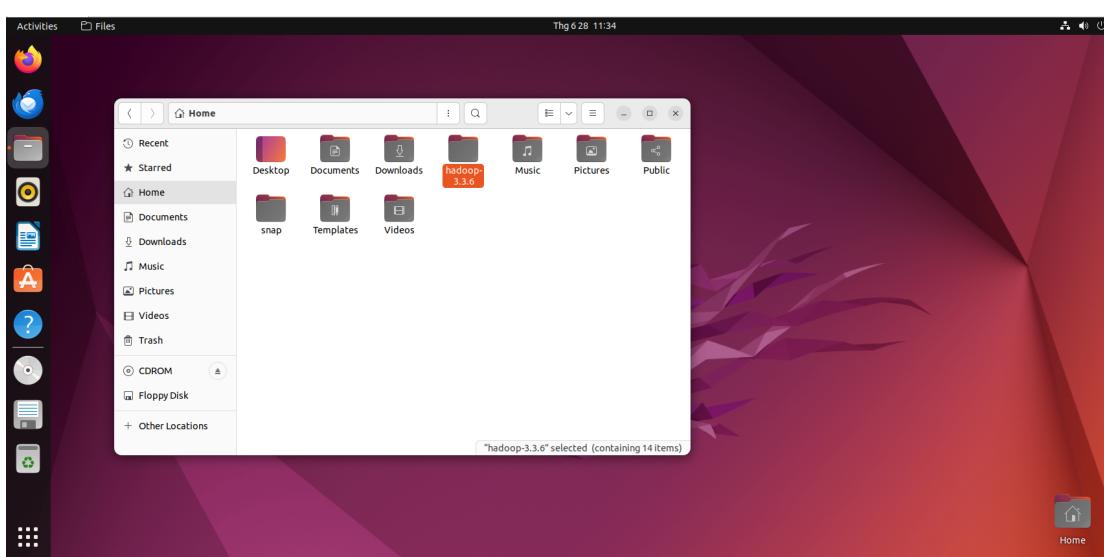
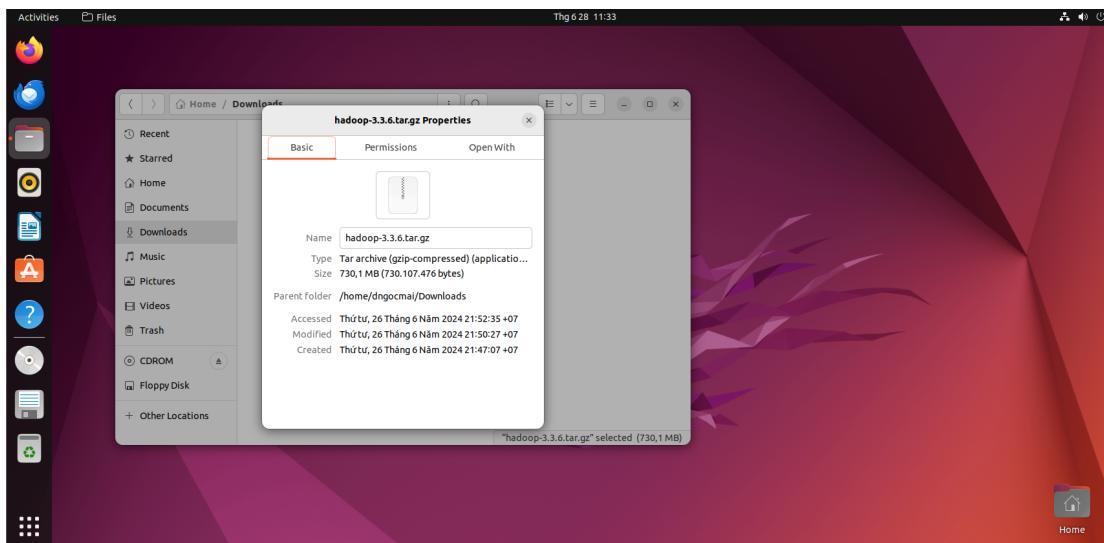
Enable ESM Apps to receive additional future security updates.
See https://Ubuntu.com/esm or run: sudo pro status

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.

dngocmai@21127104:~$
```

- Download Hadoop 3.3.6 using Firefox and extract it to the home directory.



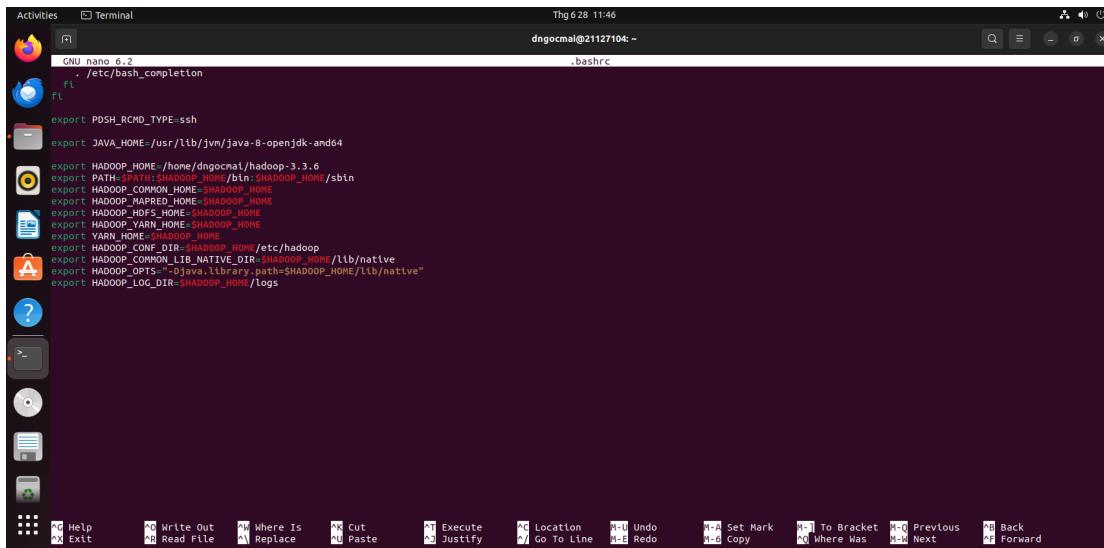
- Command to configure file *bashrc*, *core-site.xml*, *hadoop-env.sh*, *hdfs-site.xml*, *yard-site.xml*

```

dngocmai@21127104: ~$ sudo nano .bashrc
[sudo] password for dngocmai:
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/hadoop-env.sh
dngocmai@21127104: ~$ echo SHADOOP_HOME
dngocmai@21127104: ~$ source ~/.bashrc
dngocmai@21127104: ~$ do nano SHADOOP_HOME/etc/hadoop/hadoop-env.sh
bash: do: command not found
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/hadoop-env.sh
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/core-site.xml
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/core-site.xml
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/hdfs-site.xml
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/mapred-site.xml
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/yarn-site.xml
dngocmai@21127104: ~$ sudo nano SHADOOP_HOME/etc/hadoop/yarn-site.xml
dngocmai@21127104: ~$ 

```

- Configure file *bashrc*



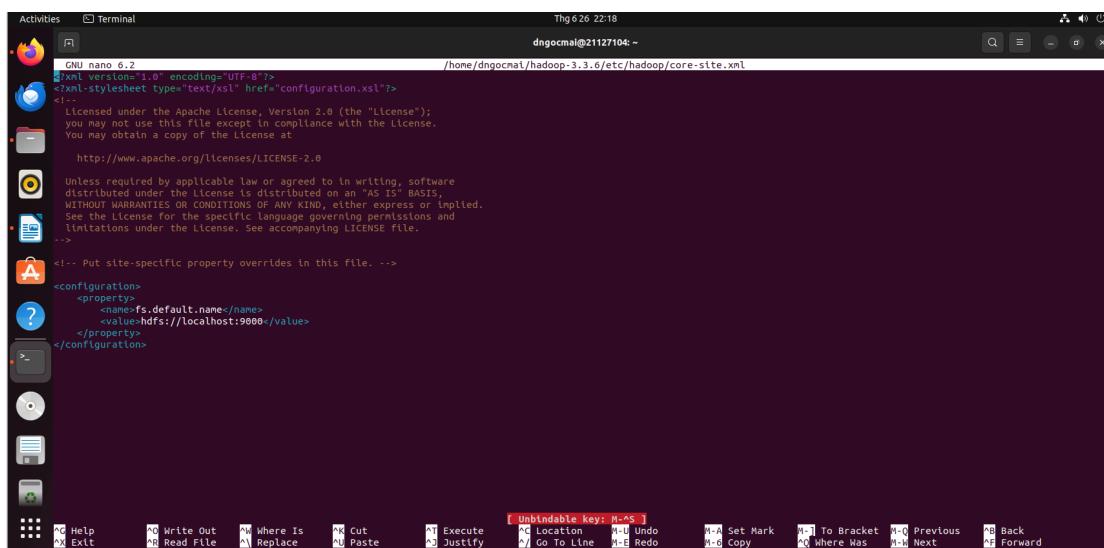
```

Activities Terminal Thg 6 28 11:46
dngocmai@21127104: ~ .bashrc

export PSSH_RCMD_TYPE=ssh
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/home/dngocmai/hadoop-3.3.6
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_LOG_DIR=$HADOOP_HOME/logs

```

- Configure file *core-site.xml*



```

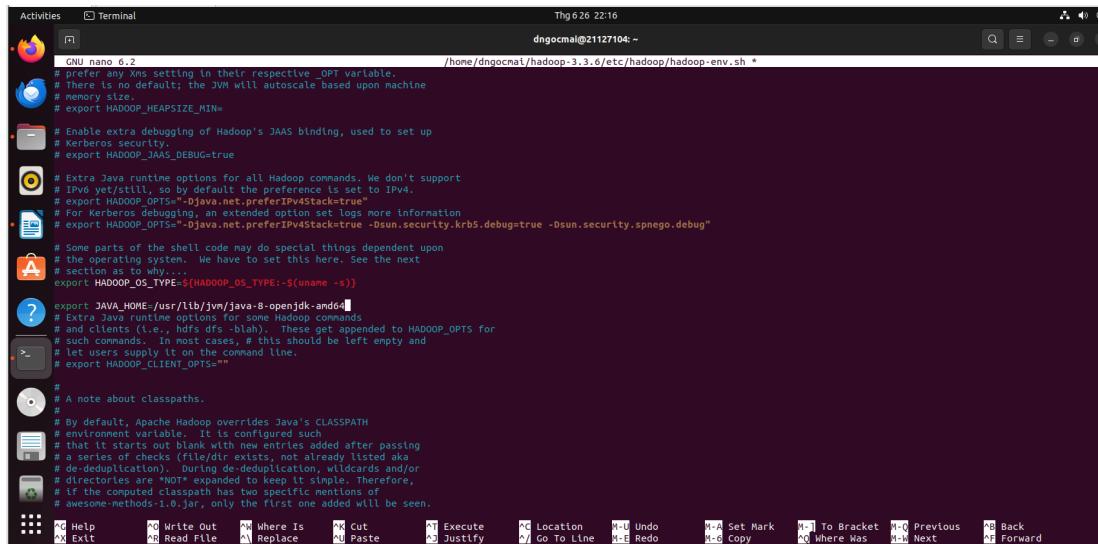
Activities Terminal Thg 6 26 22:18
dngocmai@21127104: ~ /home/dngocmai/hadoop-3.3.6/etc/hadoop/core-site.xml

<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE configuration SYSTEM "text/xml">
<!-- 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.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
</configuration>

```

- Configure file *Hadoop-env.sh*



```

GNU nano 6.2                               /home/dngocmai/hadoop-3.3.6/etc/hadoop/hadoop-env.sh *
# By default, Hadoop sets their respective _OPT variables.
# There is no default; the JVM will autoscale based upon machine
# memory size.
# export HADOOP_HEAPSIZE_MIN=1024

# Enable extra debugging of Hadoop's JAAS binding, used to set up
# Kerberos security.
# export HADOOP_JAAS_DEBUG=true

# Extra Java runtime options for all Hadoop commands. We don't support
# IPv6 yet/still, so by default the preference is set to IPv4.
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true"
# For Kerberos Debugging, an extended option set logs more information
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true -Dsun.security.krb5.debug=true -Dsun.security.spnego.debug"

# Some parts of the shell code may do special things dependent upon
# the operating system. We have to set this here. See the next
# section as to why...
# export HADOOP_OS_TYPE=$(uname -s)

# Extra Java runtime options for some Hadoop commands
# and clients (i.e. hdfs dfs -blah). These get appended to HADOOP_OPTS for
# sub commands. In most cases, # this should be left empty and
# let users supply it on the command line.
# export HADOOP_CLIENT_OPTS=""

# A note about classpaths.
# By default, Apache Hadoop overrides Java's CLASSPATH
# environment variable. It is configured such
# that it starts out blank with new entries added after passing
# a series of checks (file/dtr exists, not already listed aka
# de-duplication) during de-duplication, afterwards and/or
# after for each directory. No PATH expansion is done. Therefore,
# if the computed classpath has two specific mentions of
# awesome-methods-1.0.jar, only the first one added will be seen.

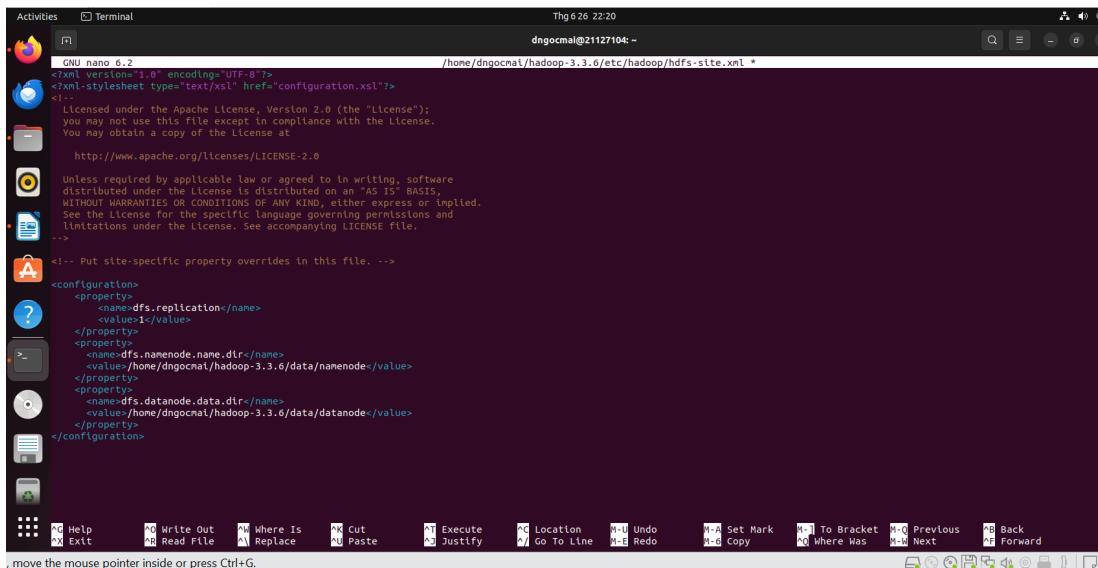
JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
# Extra Java runtime options for all Hadoop commands
# and clients (i.e. hdfs dfs -blah). These get appended to HADOOP_OPTS for
# sub commands. In most cases, # this should be left empty and
# let users supply it on the command line.
# export HADOOP_CLIENT_OPTS=""

# By default, Apache Hadoop overrides Java's CLASSPATH
# environment variable. It is configured such
# that it starts out blank with new entries added after passing
# a series of checks (file/dtr exists, not already listed aka
# de-duplication) during de-duplication, afterwards and/or
# after for each directory. No PATH expansion is done. Therefore,
# if the computed classpath has two specific mentions of
# awesome-methods-1.0.jar, only the first one added will be seen.

M-U Undo      M-A Set Mark   M-J To Bracket  M-Q Previous  AB Back
M-C Location  M-E Redo      M-Q Where Was    M-W Next     M-B Forward
M-Z Cut       M-Z Copy      M-Z Paste       M-Z Select    M-Z Select
M-Z Replace   M-Z Read File M-Z Write Out   M-Z Help     M-Z Exit

```

- Configure file *hdfs-site.xml*



```

GNU nano 6.2                               /home/dngocmai/hadoop-3.3.6/etc/hadoop/hdfs-site.xml *
<?xml version='1.0' encoding='UTF-8'?>
<xmllistener 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.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/dngocmai/hadoop-3.3.6/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/dngocmai/hadoop-3.3.6/data/datanode</value>
  </property>
</configuration>

M-U Undo      M-A Set Mark   M-J To Bracket  M-Q Previous  AB Back
M-C Location  M-E Redo      M-Q Where Was    M-W Next     M-B Forward
M-Z Cut       M-Z Copy      M-Z Paste       M-Z Select    M-Z Select
M-Z Replace   M-Z Read File M-Z Write Out   M-Z Help     M-Z Exit

```

- Configure file *mapred-site.xml*

```

GNU nano 6.2                               /home/dngocmai/hadoop-3.3.6/etc/hadoop/mapred-site.xml
<!--
  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>mapreduce.application.classpath</name>
    <value>$HADOOP_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>

```

- Configure file *yarn-site.xml*

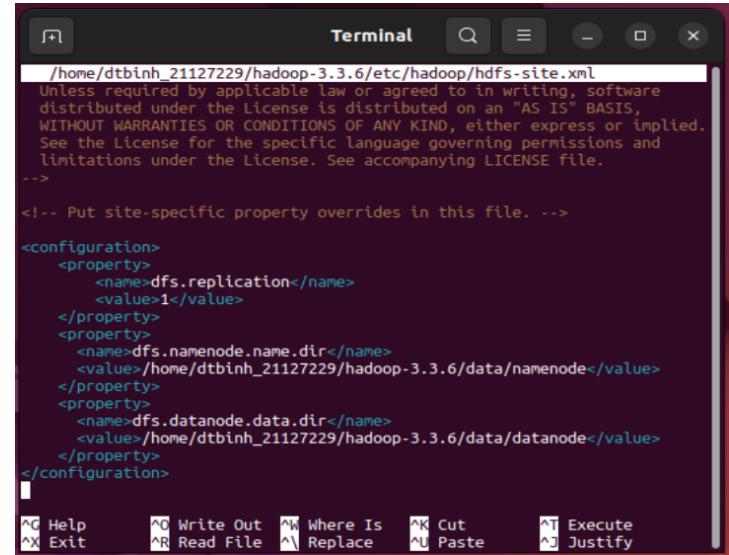
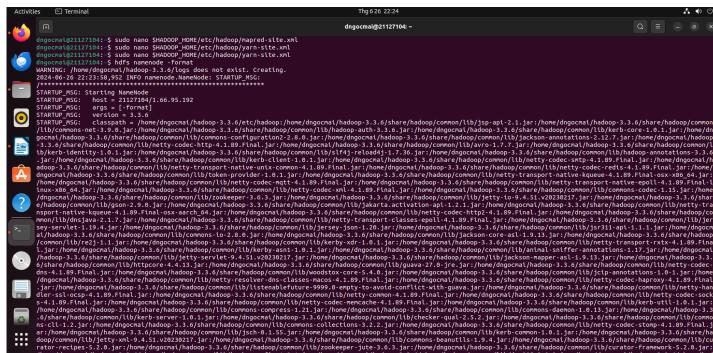
```

GNU nano 6.2                               /home/dngocmai/hadoop-3.3.6/etc/hadoop/yarn-site.xml
<!--
  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

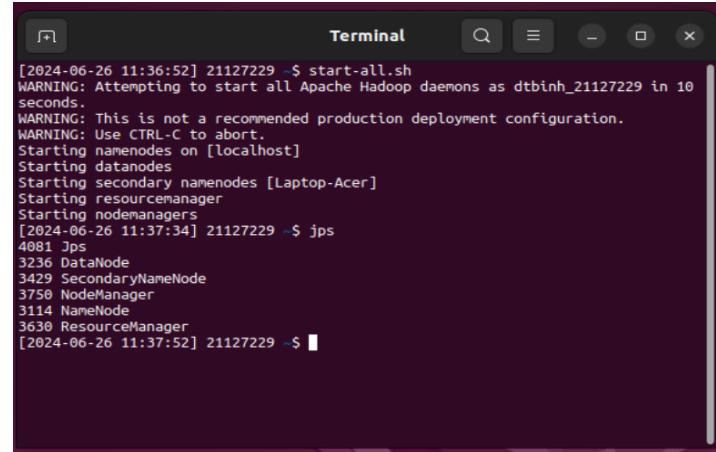
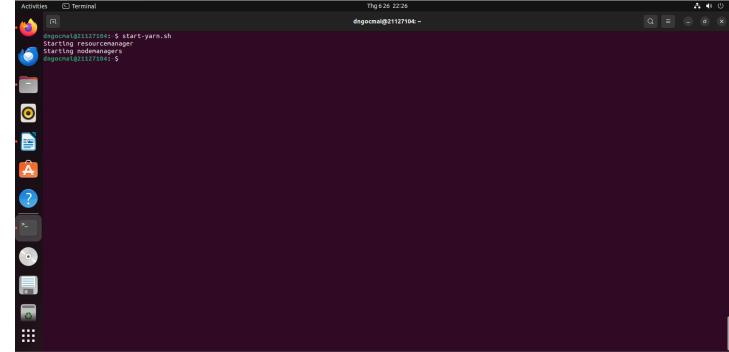
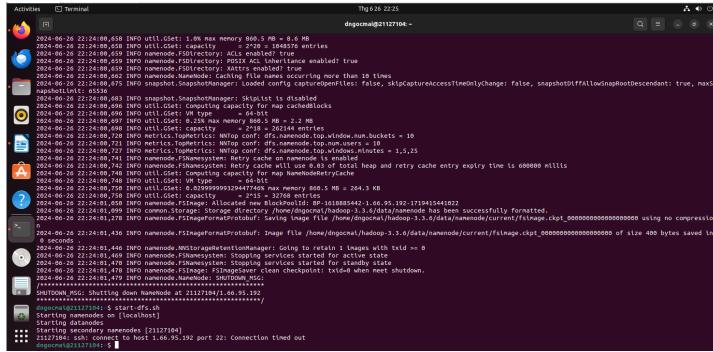
  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>127.0.0.1</value>
  </property>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</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_HOME,PATH,LANG,TZ,HADOOP_MAPRED_HOME</value>
  </property>
</configuration>

```

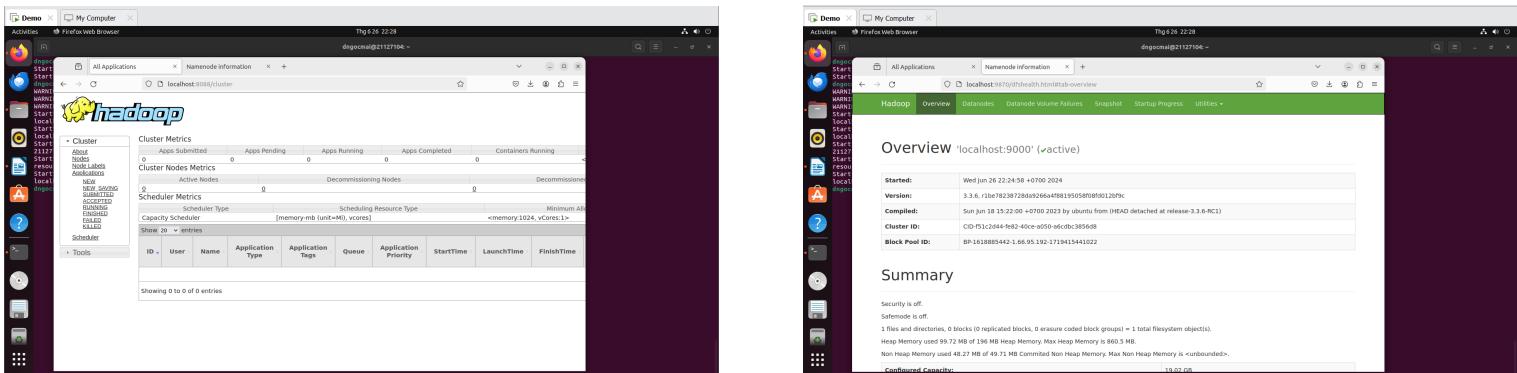
- Format HDFS Name Node



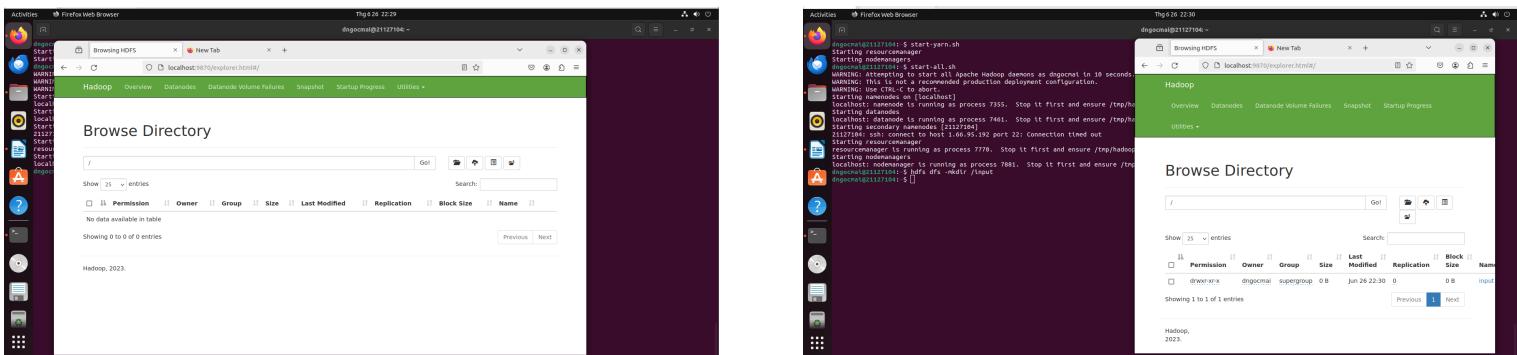
- Start Hadoop Cluster



- Verifying the installation by accessing the Hadoop Web Interface at `localhost:8088` and `localhost:9870`.



- Run a Hadoop MapReduce Job to search for patterns in HDFS and then retrieve results



1.4 Oanh

- Firstly, installing Java

```
lkoanh_21127129@lkoanh-virtual-machine:~$ sudo apt install openjdk-8-jdk -y
[sudo] password for lkoanh_21127129:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni liblclc-dev
  libpthread-stubs0-dev libsm-dev libx11-dev libxcb-dev libxdmcp-dev libxt-dev openjdk-8-jdk-headless
  openjdk-8-jre openjdk-8-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
  default-jre-lbicc-doc libsnx-doc libx11-doc libxcb-doc libxt-doc openjdk-8-demo openjdk-8-source visualvm fonts-nanum
  fonts-ipafont-gothic fonts-ipafont-mincho fonts-wqy-microhei fonts-wqy-zhenhei
The following NEW packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni liblclc-dev
  libpthread-stubs0-dev libsm-dev libx11-dev libxcb-dev libxdmcp-dev libxt-dev openjdk-8-jdk
  openjdk-8-jdk-headless openjdk-8-jre openjdk-8-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 20 newly installed, 0 to remove and 191 not upgraded.
Need to get 47,9 MB of archives.
After this operation, 163 MB of additional disk space will be used.
Get:1 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 java-common all 0.72build2 [6.782 B]
Get:2 http://vn.archive.ubuntu.com/ubuntu jammy-updates/universe amd64 openjdk-8-jre-headless amd64 8u412+ga-1-22.04.1 [30,8 MB]
Get:3 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 ca-certificates-java all 20190909ubuntu1.2 [12,1 kB]
Get:4 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 fonts-dejavu-extra all 2.37-2build1 [2.041 kB]
Get:5 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java all 0.38.0-5build1 [53,1 kB]
Get:6 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libatk-wrapper-java-jni amd64 0.38.0-5build1 [49,0 kB]
Get:7 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 xorg-sgml-doctools all 1:1.11.1-1 [10,9 kB]
Get:8 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 x11proto-dev all 2021.5-1 [600 kB]
Get:9 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 liblclc-dev amd64 2:1.0.10-1build2 [51,4 kB]
Get:10 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libpthread-stubs0-dev amd64 0.4-1build2 [5.516 B]
Get:11 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libsm-dev amd64 2:1.2.3-1build2 [18,1 kB]
Get:12 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libxau-dev amd64 1:1.0.9-1build5 [9.724 B]
Get:13 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 libxdmcp-dev amd64 1:1.1.3-0ubuntu5 [26,5 kB]
Get:14 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 xtrans-dev all 1.4.0-1 [68,9 kB]
```

- Checking Java and Javac version

```

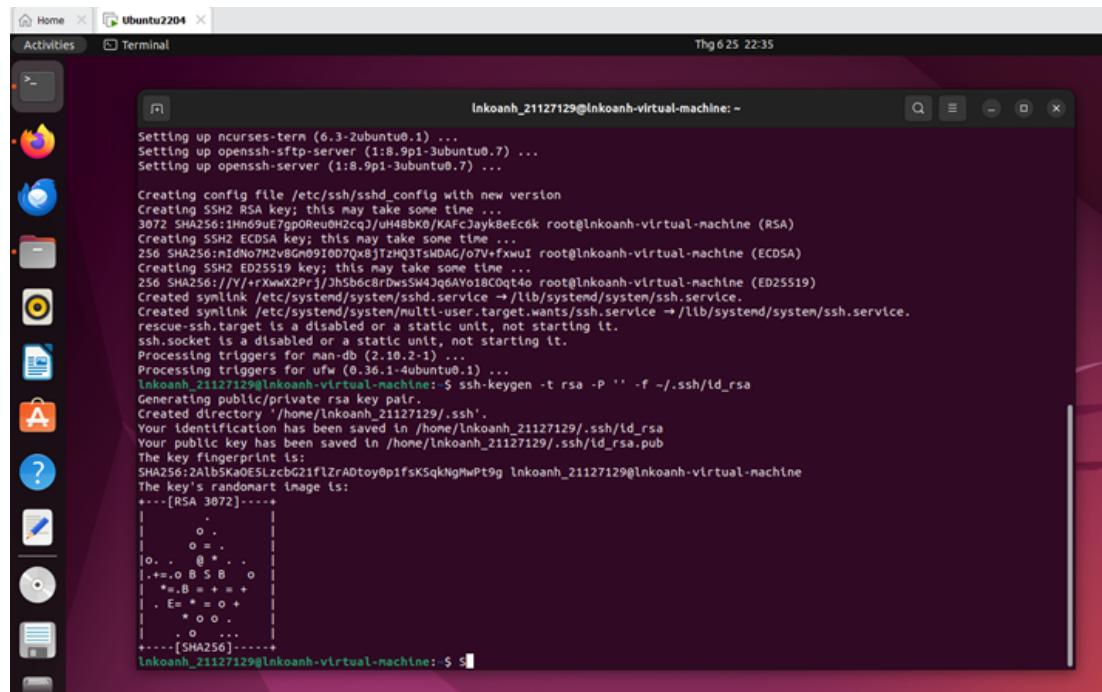
Inkoanh_21127129@lnkoanh-virtual-machine:~$ java -version
openjdk version "1.8.0_412"
OpenJDK Runtime Environment (build 1.8.0_412-8u412-ga-1-22.04.1-b08)
OpenJDK 64-Bit Server VM (build 25.412-b08, mixed mode)
Inkoanh_21127129@lnkoanh-virtual-machine:~$ javac -version
javac 1.8.0_412
Inkoanh_21127129@lnkoanh-virtual-machine:~$ which java
/usr/bin/java
Inkoanh_21127129@lnkoanh-virtual-machine:~$ readlink -f /usr/bin/java
/usr/lib/jvm/java-8-openjdk-amd64/bin/java
Inkoanh_21127129@lnkoanh-virtual-machine:~$
```

- Installing OpenSSH

```

Inkoanh_21127129@lnkoanh-virtual-machine:~$ sudo apt install openssh-server openssh-client -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  curses-term openssh-sftp-server ssh-import-id
Suggested packages:
  keychain libpam-ssh monkeysphere ssh-askpass molly-guard
The following NEW packages will be installed:
  curses-term openssh-server openssh-sftp-server ssh-import-id
The following packages will be upgraded:
  openssh-client
1 upgraded, 4 newly installed, 0 to remove and 190 not upgraded.
Need to get 752 kB or 1.658 kB of archives.
After this operation, 6.050 kB of additional disk space will be used.
Get:1 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-sftp-server amd64 1:8.9p1-3ubuntu0.7 [38,9 kB]
Get:2 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-server amd64 1:8.9p1-3ubuntu0.7 [435 kB]
Get:3 http://vn.archive.ubuntu.com/ubuntu jammy-updates/main amd64 curses-term all 6.3-2ubuntu0.1 [267 kB]
Get:4 http://vn.archive.ubuntu.com/ubuntu jammy/main amd64 ssh-import-id all 5.11-0ubuntu1 [10,1 kB]
Fetched 752 kB in 0s (1.688 kB/s)
Preconfiguring packages ...
(Reading database ... 202233 files and directories currently installed.)
Preparing to unpack .../openssh-client_1k3a8.9p1-3ubuntu0.7_amd64.deb ...
Unpacking openssh-client (1:8.9p1-3ubuntu0.7) over (1:8.9p1-3ubuntu0.6) ...
Selecting previously unselected package openssh-sftp-server.
Preparing to unpack .../openssh-sftp-server_1k3a8.9p1-3ubuntu0.7_amd64.deb ...
Unpacking openssh-sftp-server (1:8.9p1-3ubuntu0.7) ...
Selecting previously unselected package openssh-server.
Preparing to unpack .../openssh-server_1k3a8.9p1-3ubuntu0.7_amd64.deb ...
Unpacking openssh-server (1:8.9p1-3ubuntu0.7) ...
Selecting previously unselected package ncurses-term.
Preparing to unpack .../ncurses-term_6.3-2ubuntu0.1_all.deb ...
Unpacking ncurses-term (6.3-2ubuntu0.1) ...
Selecting previously unselected package ssh-import-id.
Preparing to unpack .../ssh-import-id_5.11-0ubuntu1_all.deb ...
Unpacking ssh-import-id (5.11-0ubuntu1) ...
Setting up openssh-client (1:8.9p1-3ubuntu0.7) ...
```

- Generating SSH Key



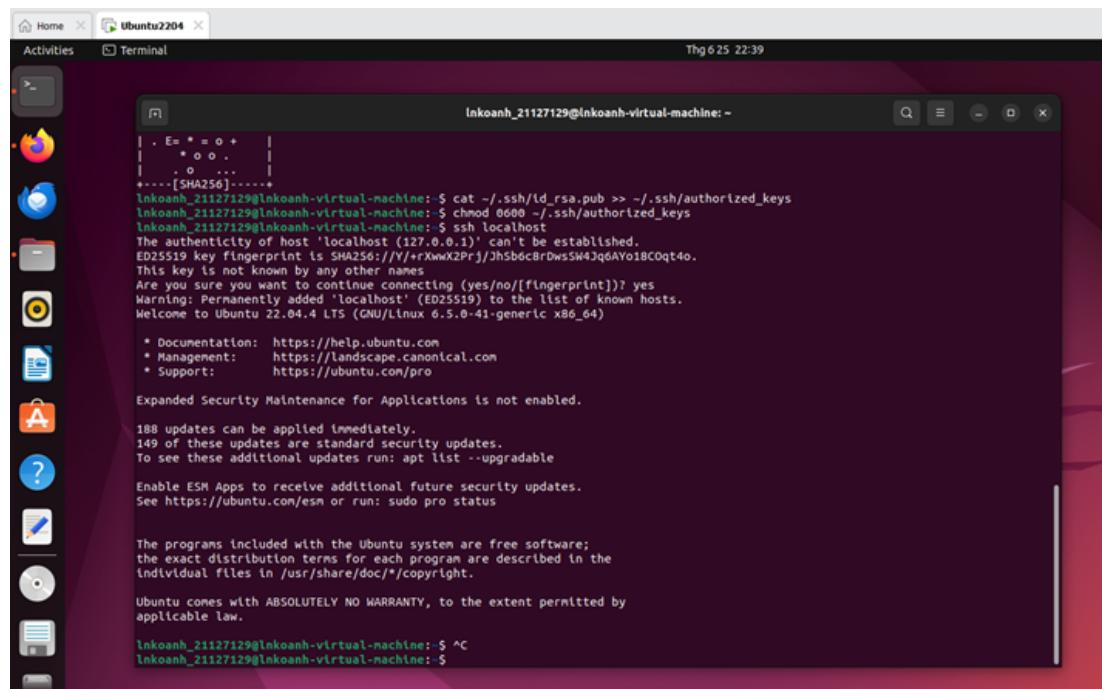
```

Setting up curses-term (6.3-2ubuntu0.1) ...
Setting up openssh-sftp-server (1:8.9p1-3ubuntu0.7) ...
Setting up openssh-server (1:8.9p1-3ubuntu0.7) ...

Creating config file /etc/ssh/sshd_config with new version
Creating SSH2 RSA key; this may take some time ...
3072 SHA256:1Hn69UE7gpOReuOH2cqJUH48bK0/KAFcJaykBeEcK root@lnkoanh-virtual-machine (RSA)
Creating SSH2 ECDSA Key; this may take some time ...
256 SHA256:midNoM2vBnG0910D7QxJzHq3TsWDAG/o7V+fxwUI root@lnkoanh-virtual-machine (ECDSA)
Creating SSH2 ED25519 key; this may take some time ...
256 SHA256://Y+rXnwX2PrlJ3hSboc8r0wsSW4JgqAYo18C0qt4o root@lnkoanh-virtual-machine (ED25519)
Created symlink /etc/systemd/system/multi-user.target.wants/ssh.service → /lib/systemd/system/ssh.service.
Created symlink /etc/systemd/system/ssh.service → /lib/systemd/system/ssh.service.
rescue-ssh.target is a disabled or a static unit, not starting it.
ssh.socket is a disabled or a static unit, not starting it.
Processing triggers for man-db (2.10.2-1) ...
Processing triggers for ufw (0.36.1-4ubuntu0.1) ...
lnkoanh_21127129@lnkoanh-virtual-machine:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/lnkoanh_21127129/.ssh'.
Your identification has been saved in /home/lnkoanh_21127129/.ssh/id_rsa
Your public key has been saved in /home/lnkoanh_21127129/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:2AlbSKaE5LzcBZiflZrADtoy0pfsKSqkNgMwPt9g lnkoanh_21127129@lnkoanh-virtual-machine
The key's randomart image is:
+---[RSA 3072]---+
| . .
| o .
| o = .
| o .. @ * . .
| .+o . B S B o
| *=B = + =
| . E= * = o +
| * o o .
| . o .
+---[SHA256]---+
lnkoanh_21127129@lnkoanh-virtual-machine:~$ s

```

- Store public key as authorized key in SSH directory, set permission for user with chmod command and check SSH localhost



```

lnkoanh_21127129@lnkoanh-virtual-machine:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
lnkoanh_21127129@lnkoanh-virtual-machine:~$ chmod 0600 ~/.ssh/authorized_keys
lnkoanh_21127129@lnkoanh-virtual-machine:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256://Y+rXnwX2PrlJ3hSboc8r0wsSW4JgqAYo18C0qt4o.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 6.5.0-41-generic x86_64)

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

Expanded Security Maintenance for Applications is not enabled.

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

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

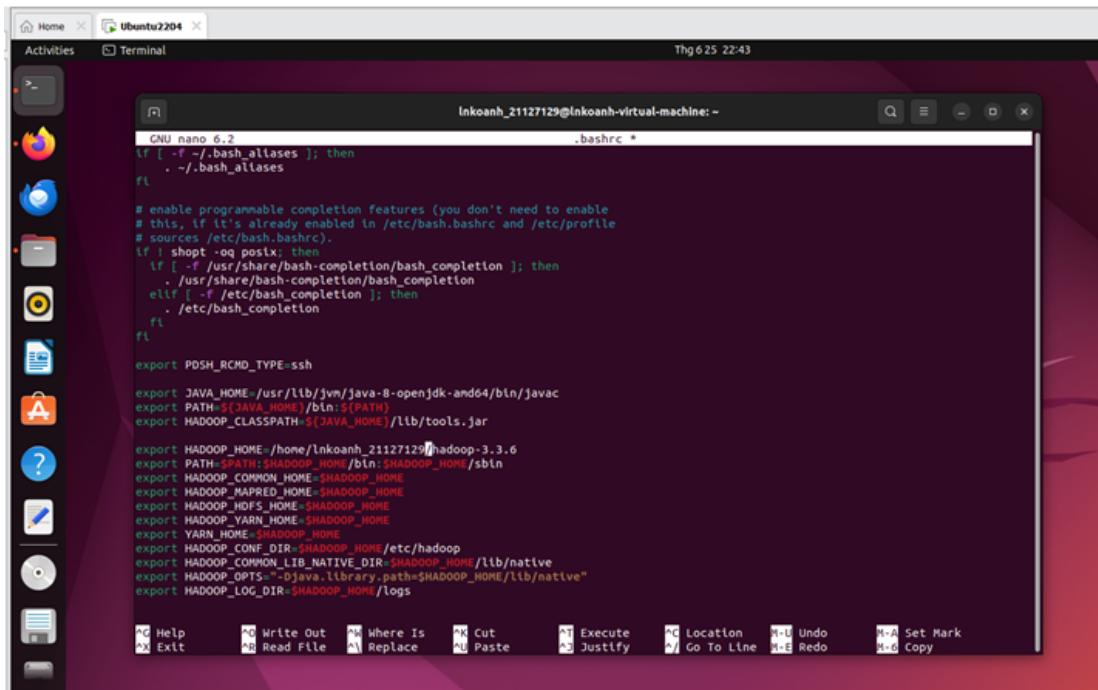
The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

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

lnkoanh_21127129@lnkoanh-virtual-machine:~$ ^C
lnkoanh_21127129@lnkoanh-virtual-machine:~$ 

```

- Download Hadoop 3.3.6 and extract file at Home
- Configure file *bashrc*



```

GNU nano 6.2          .bashrc *
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

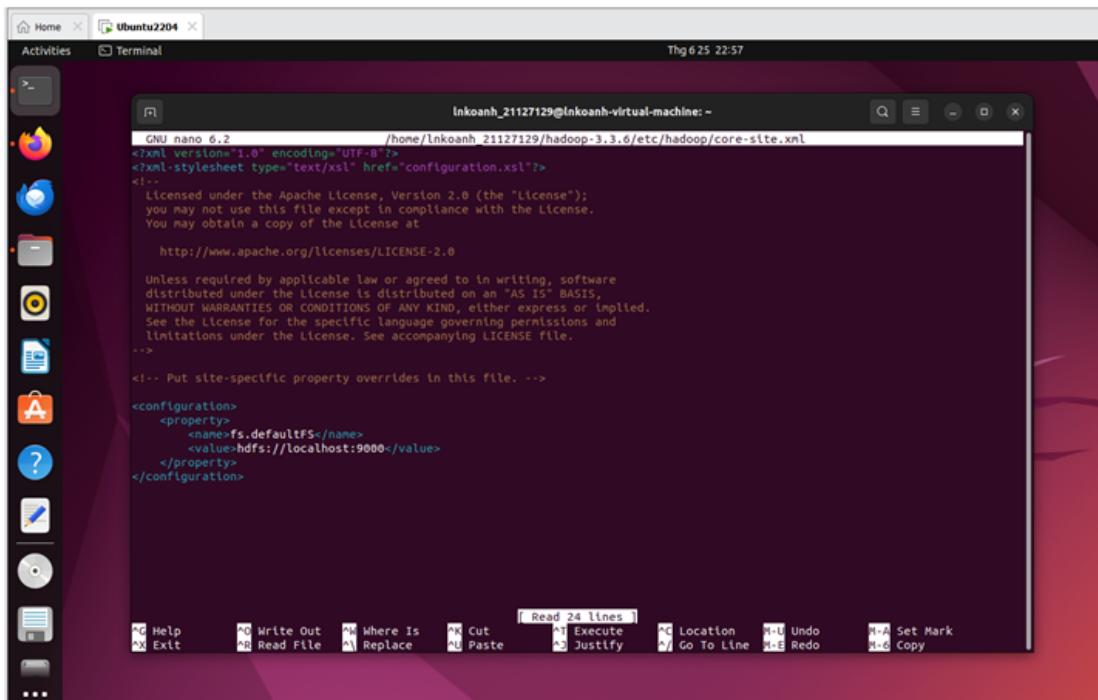
export PDSH_RCMD_TYPE=ssh

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/bin/java
export PATH=$JAVA_HOME/bin:$PATH
export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar

export HADOOP_HOME=/home/lnkoanh_21127129/hadoop-3.3.6
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export HADOOP_LOG_DIR=$HADOOP_HOME/logs

```

- Configure file *core-site.xml*



```

GNU nano 6.2          /home/lnkoanh_21127129/hadoop-3.3.6/etc/hadoop/core-site.xml
<?xml version='1.0' encoding='UTF-8'?>
<xsl: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://localhost:9000</value>
  </property>
</configuration>

```

- Configure file *Hadoop-env.sh*

```

GNU nano 6.2 /home/lnkoanh_21127129/hadoop-3.3.6/etc/hadoop/hadoop-env.sh
# IPv6 yet/still, so by default the preference is set to IPv4.
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true"
# For Kerberos debugging, an extended option set logs more information
# export HADOOP_OPTS="-Djava.net.preferIPv4Stack=true -Dsun.security.krb5.debug=true -Dsun.security.spnego.debug"

# Some parts of the shell code may do special things dependent upon
# the operating system. We have to set this here. See the next
# section as to why...
export HADOOP_OS_TYPE=$(uname -s)
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64/bin/java

# Extra Java runtime options for some Hadoop commands
# and clients (i.e., hdfs dfs -blah). These get appended to HADOOP_OPTS for
# such commands. In most cases, # this should be left empty and
# let users supply it on the command line.
# export HADOOP_CLIENT_OPTS=""

#
# A note about classpaths.
#
# By default, Apache Hadoop overrides Java's CLASSPATH
# environment variable. It is configured such
# that it starts out blank with new entries added after passing
# a series of checks (file/dir exists, not already listed aka
# de-duplication). During de-duplication, wildcards and/or
# directories are *NOT* expanded to keep it simple. Therefore,
# if the computed classpath has two specific mentions of
# awesome-methods-1.0.jar, only the first one added will be seen.
# If two directories are in the classpath that both contain
# awesome-methods-1.0.jar, then Java will pick up both versions.

# An additional, custom CLASSPATH. Site-wide configs should be
# handled via the shellprofile functionality, utilizing the

```

- Configure file *hdfs-site.xml*

```

GNU nano 6.2 /home/lnkoanh_21127129/hadoop-3.3.6/etc/hadoop/hdfs-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>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/lnkoanh_21127129/hadoop-3.3.6/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/lnkoanh_21127129/hadoop-3.3.6/data/datanode</value>
  </property>
</configuration>

```

- Configure file *mapred-site.xml*

```

GNU nano 6.2                               /home/lnkoanh_21127129@lnkoanh-virtual-machine:-
<?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.
-->

<!-- Put site-specific property overrides in this file. -->
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>mapreduce.application.classpath</name>
    <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>

```

- Configure file *yard-site.xml*

```

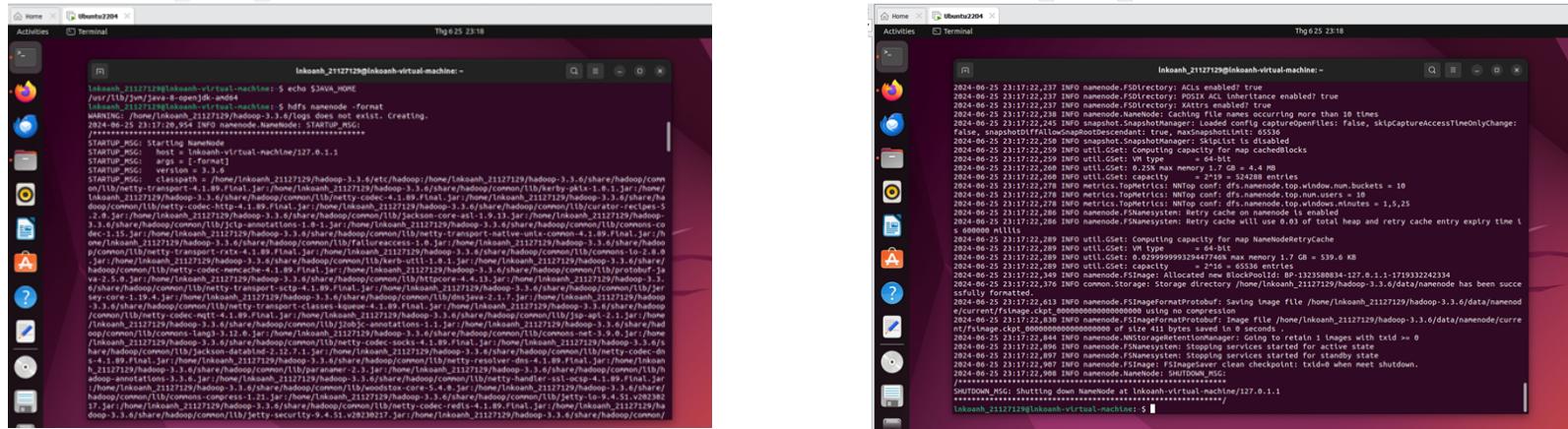
GNU nano 6.2                               /home/lnkoanh_21127129@lnkoanh-virtual-machine:-
<?xml version="1.0"?>
<!-- Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0

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.
-->

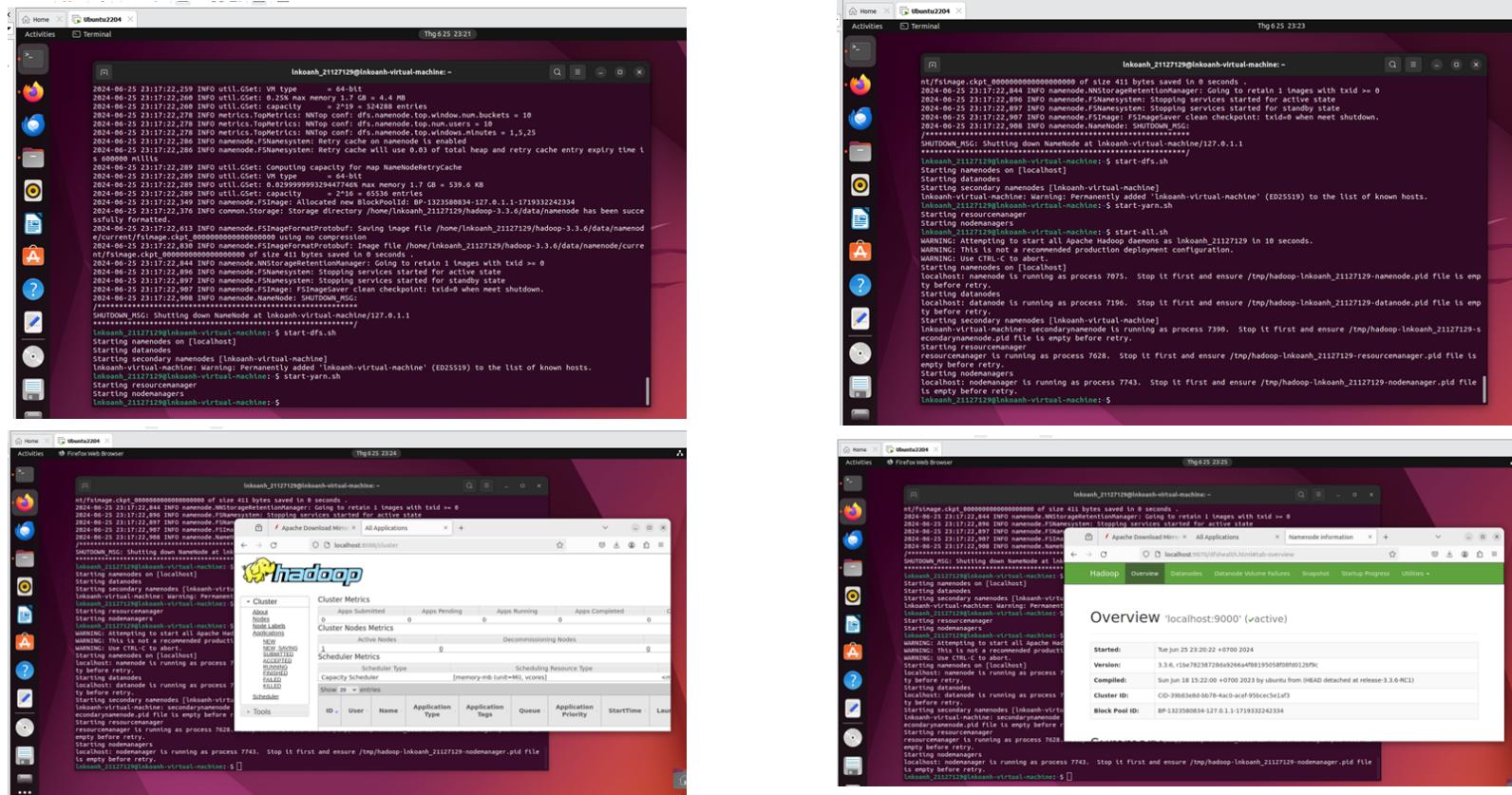
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>
    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_PID_DIR</value>
  </property>
</configuration>

```

- Format HDFS Name Node



- Starting Hadoop Cluster



- Finally, checking the result using the Hadoop Web Interface at *localhost:8088* and *localhost:9870*.

All Applications

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	Start Time	Launch Time	Finish Time	State	Final Status	Running Containers	All
No data available in table													

Overview localhost:9000 (active)

Started:	Tue Jun 25 23:20:22 +0700 2024
Version:	3.3.0, r12e78230728d9206a48819505808580120Phc
Compiled:	Sun Jun 18 15:22:00 +0700 2023 by ubuntu from HEAD detached at release-3.3-PC1
Cluster ID:	CID-39838edbd8b78-4ac0-acf95dcce5e1af3
Block Pool ID:	BP-1323580834-127.0.1.171933242334

Summary

Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 132.12 MB of 300 MB Heap Memory. Max Heap Memory is 1.72 GB.
Non-Heap Memory used 49.56 MB of 51.4 MB Committed Non-Heap Memory. Max Non-Heap Memory is <unbounded>.

Configured Capacity:	28.87 GB
Configured Remote Capacity:	0.0
DFS Used:	24 KB (0%)

2 Paper Reading

2.1 How do the input keys-values, the intermediate keys-values, and the output keys-values relate?

In the MapReduce model, the relationship between the input, intermediate, and output keys-values is as follows:

- **Input Keys-Values:** These are the initial data pairs processed by the Map function. The user writes the Map function to take these input pairs and generate intermediate key-value pairs.
- **Intermediate Keys-Values:** The intermediate key-value pairs generated by the Map function are grouped by key. All values associated with the same intermediate key are then passed to the Reduce function.
- **Output Keys-Values:** The intermediate key-value pairs are then passed to the Reduce phase, where the reduce function processes them to produce the final output key-value pairs. The output keys are typically aggregated or summarized versions of the intermediate keys, and the output values are the final processed data corresponding to these output keys.

2.2 How does MapReduce deal with node failures?

MapReduce deals with node failures through several mechanisms to ensure the reliability and efficiency of its operations.

- **Worker Failure:** The master node periodically pings each worker node to check its status. If a worker node fails to respond within a certain timeframe, it is marked as failed. All map tasks that were completed by the failed worker are reset to their initial idle state, making them eligible for reassignment to other workers. Similarly, any map or reduce tasks in progress on the failed worker are also reset and rescheduled. Completed map tasks must be re-executed because their output is stored on the local disk of the failed machine and becomes inaccessible. In contrast, completed reduce tasks do not need to be re-executed as their output is stored in a global file system.
- **Master Failure:** To handle potential master node failures, the master can periodically save checkpoints of its state. If the master node fails, a new master can be restarted from the last checkpointed state. However, given the low likelihood of master node failure and the

complexity involved, the current implementation of MapReduce opts to abort the computation if the master fails. Clients can detect this condition and choose to retry the MapReduce operation if needed.

- **Backup Tasks:** To mitigate the impact of stragglers—tasks that take unusually long to complete—the master schedules backup executions of the remaining in-progress tasks when the operation is near completion. The task is marked as complete when either the primary or the backup execution finishes. This approach typically increases computational resources by only a small percentage and significantly reduces the time required to complete large MapReduce operations.
- **Semantics in the Presence of Failures:** The system ensures that deterministic map and reduce operations yield the same output as a sequential execution would, even in the presence of failures. This is achieved through atomic commits of map and reduce task outputs.
- **Skipping Bad Records:** MapReduce provides an optional mode of execution to skip records that cause deterministic crashes, allowing the job to progress despite the presence of problematic records. This feature is useful in scenarios where ignoring a few records is acceptable, such as statistical analysis on large datasets.

2.3 What is the meaning and implication of locality? What does it use?

Meaning of Locality: Locality refers to the strategic scheduling of map tasks to run on the machines that already hold the data to be processed, or at least on machines that are close to these data-holding machines. This approach leverages the fact that input data is stored on local disks across a cluster, with each file divided into 64 MB blocks and replicated on multiple machines.

Implication of Locality: The implication of this strategy is a significant reduction in the consumption of network bandwidth. By reading data locally, rather than transferring it over the network, MapReduce operations can run more efficiently, particularly when dealing with large datasets spread across many workers in a cluster. Locality thereby helps in conserving network resources, reducing potential bottlenecks, and improving overall performance of the MapReduce jobs.

Locality uses the structure of the Google File System (GFS) for data storage and replication, and the intelligent scheduling capabilities of the MapReduce master, which takes into account the location of data replicas to optimize task placement.

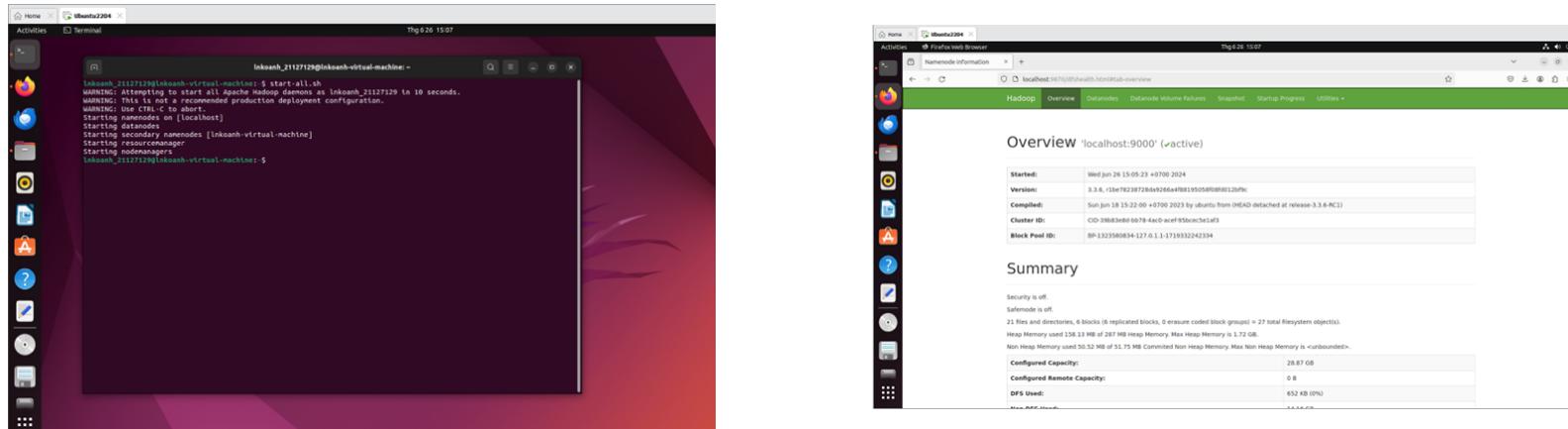
2.4 Which problem is addressed by introducing a combiner function to the MapReduce model?

The problem addressed by introducing a combiner function to the MapReduce model is the excessive network traffic caused by the large volume of intermediate data generated during the map phase, particularly when there is significant repetition in the intermediate keys. The combiner function performs partial merging of this data locally on each machine that performs a map task, thereby reducing the amount of data that needs to be transferred over the network. This approach alleviates network congestion and improves the overall performance of MapReduce operations.

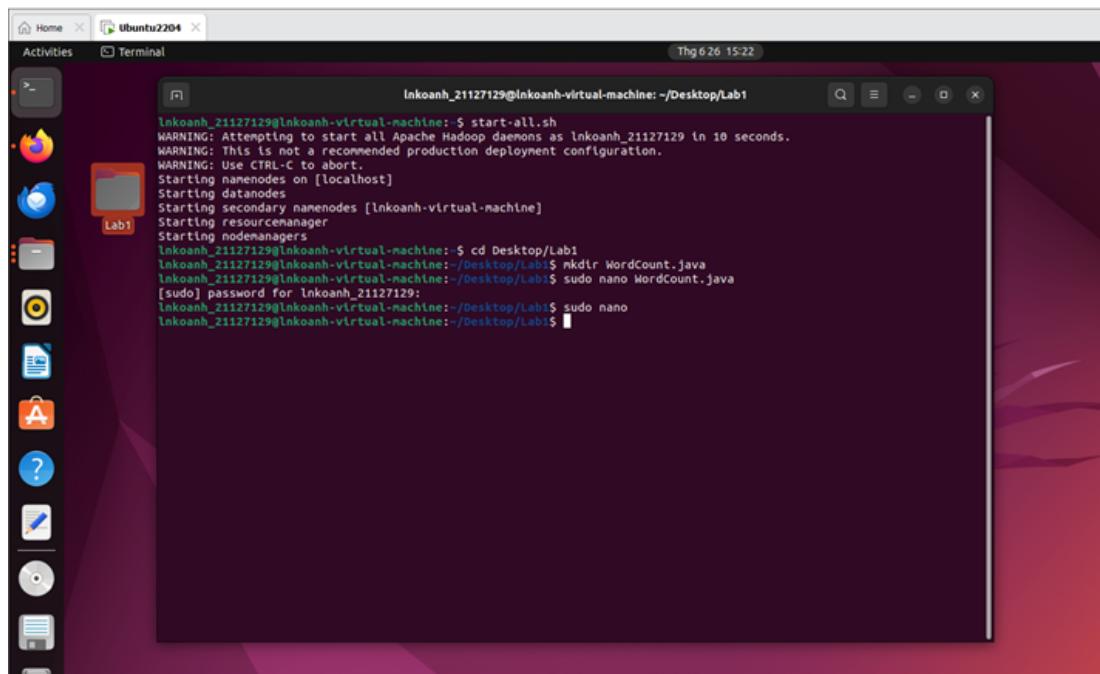
3 Running a warm-up problem: Word Count

3.1 Word Count

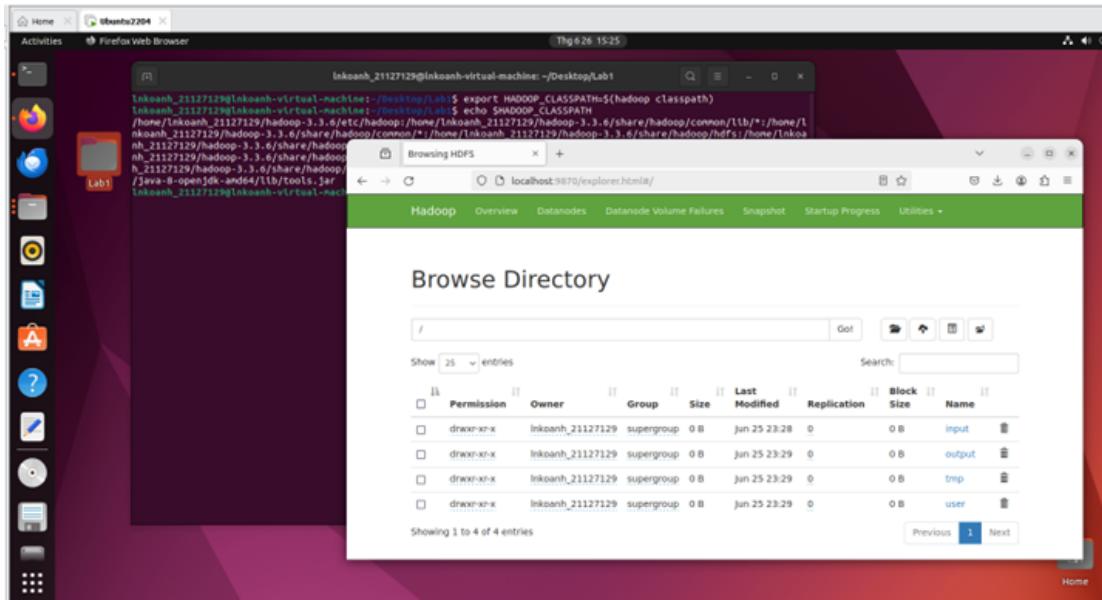
- Firstly, we need to start all the Apache Hadoop Daemons.



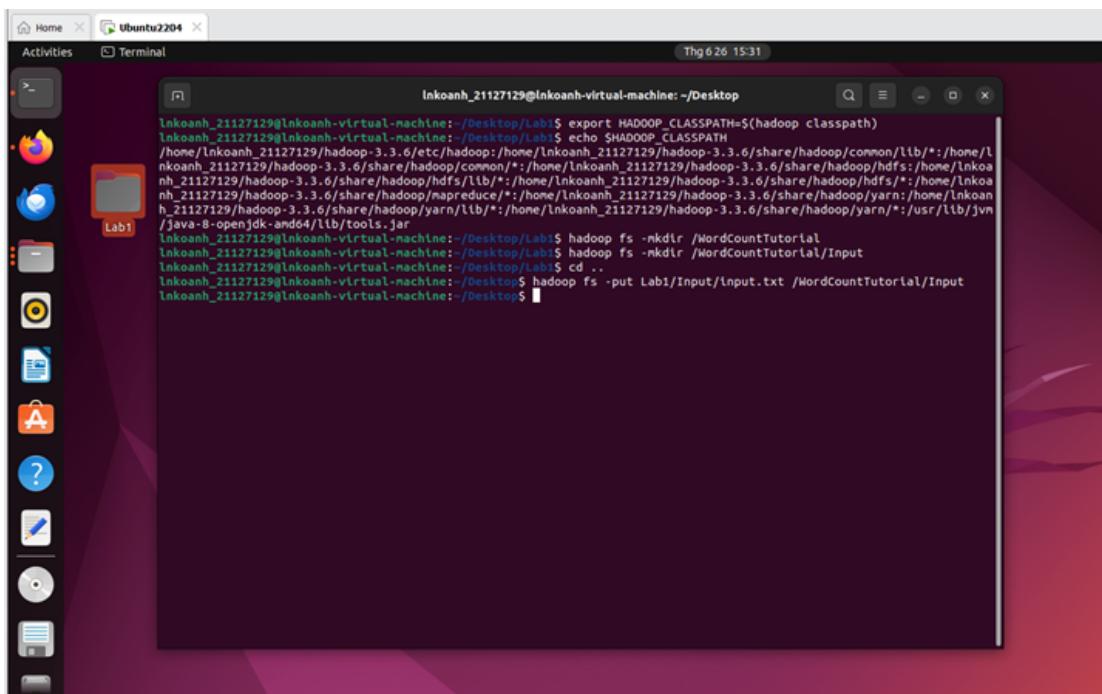
- Create a directory on the Desktop named Lab1 and inside it create two folders, one called *Input* and the other called *tutorial-classes* using GUI normally.



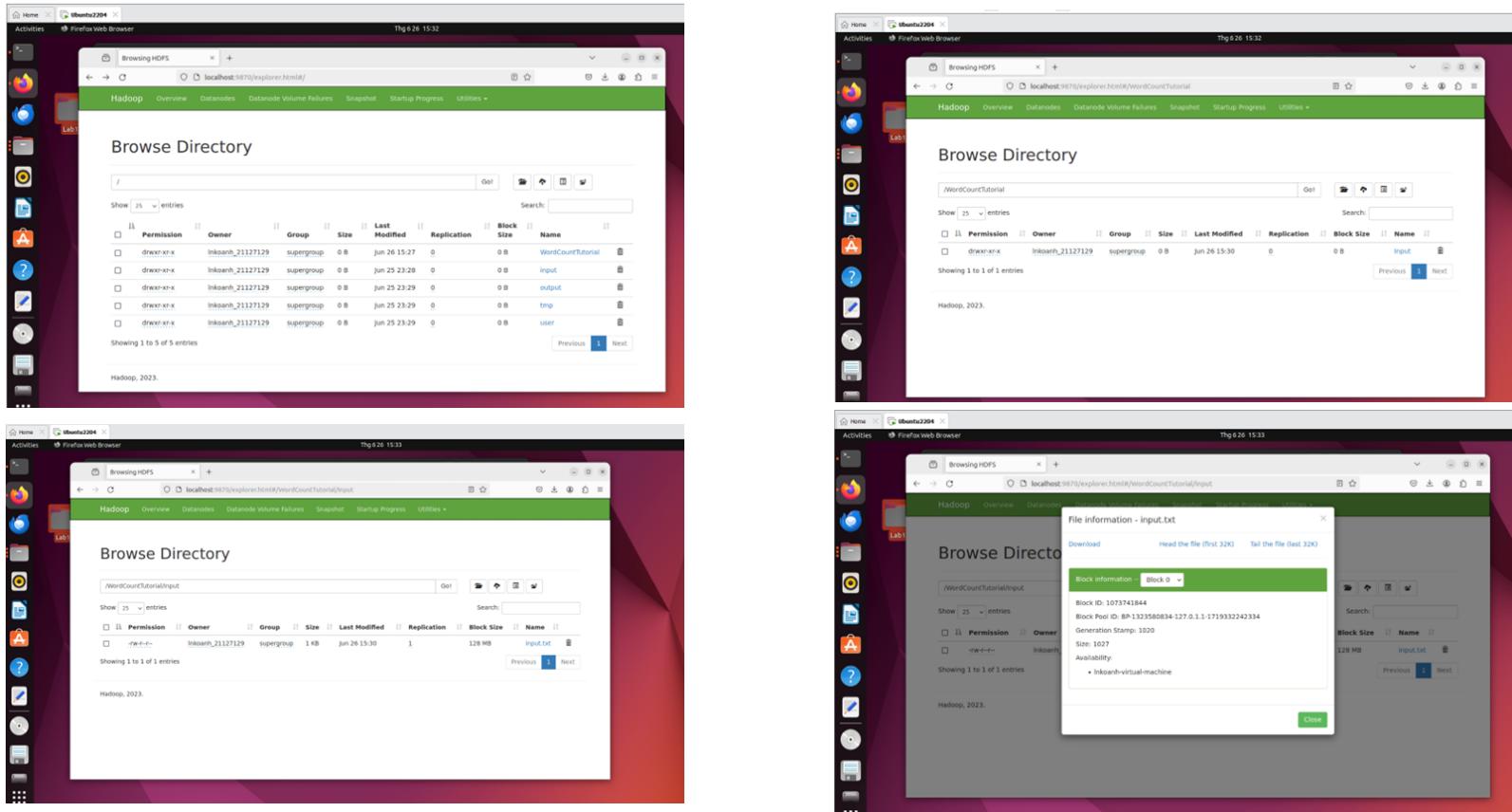
- Let see he Hadoop Web Interface at *localhost:9870* before we upload the file to HDFS.



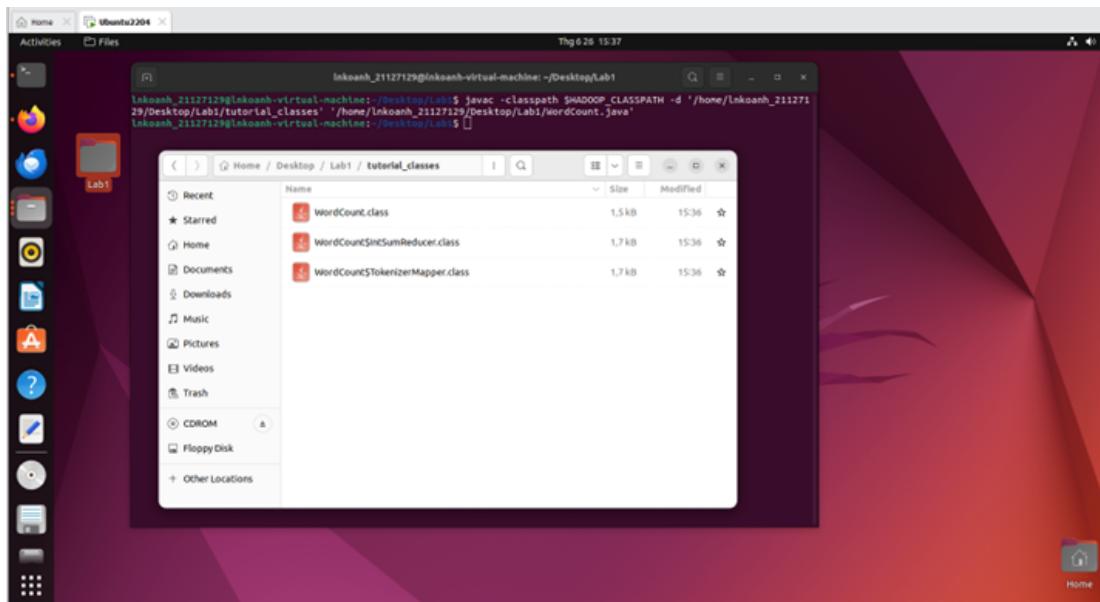
- Add the file attached with this document “WordCount.java” in the directory Lab1 and add the file attached with this document “input.txt” in the directory Lab1/Input using GUI normally.
- Next, we create these directories on HDFS



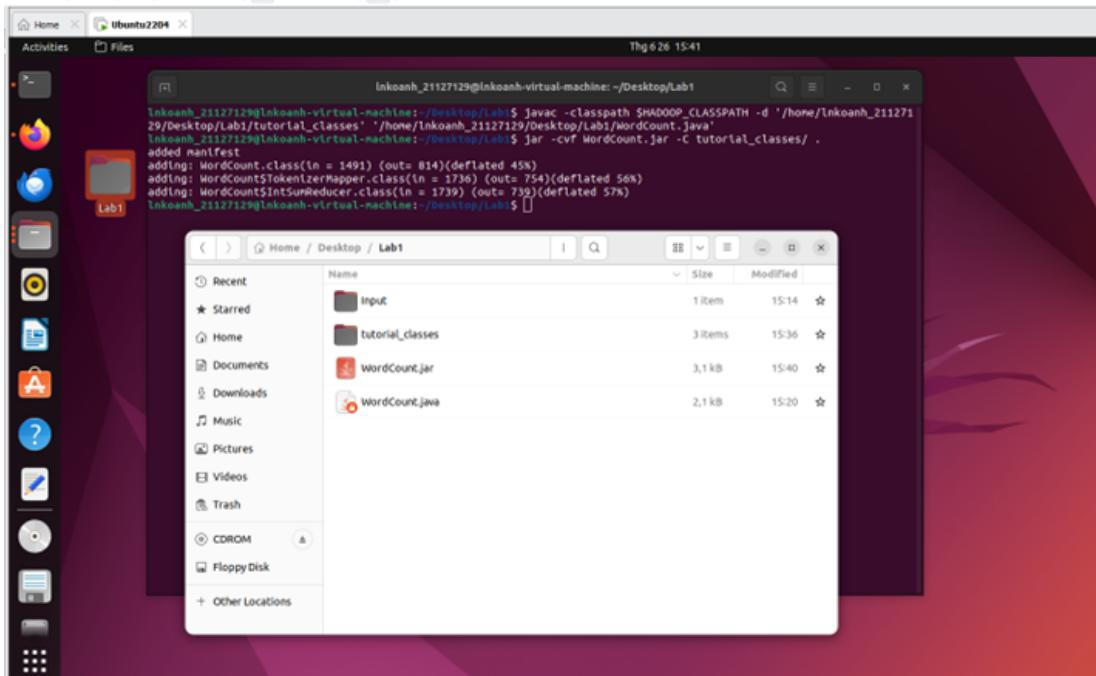
- Now, let see the directories and files we placed in the file system



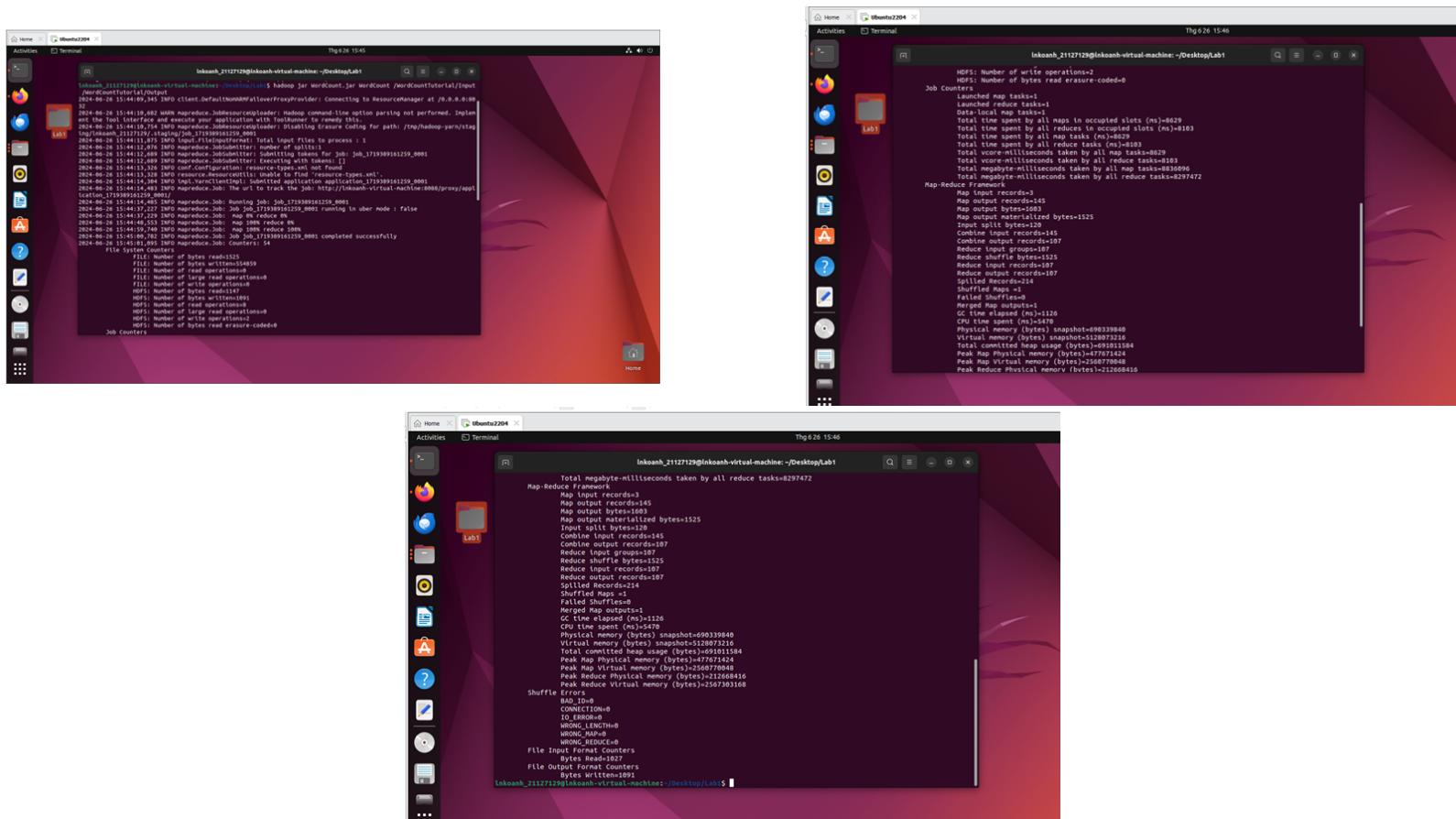
- Compile the "WordCount.java" file



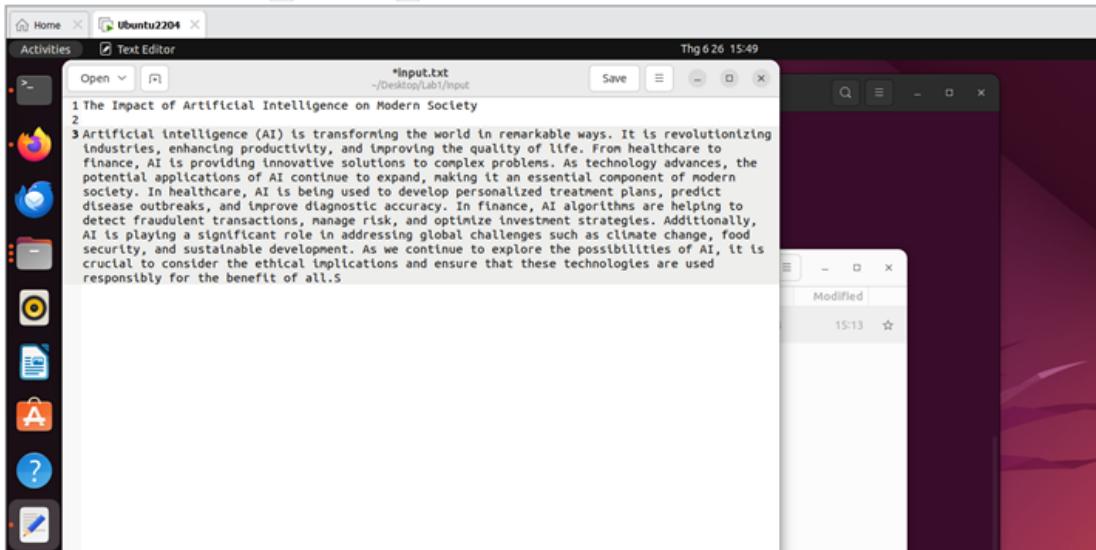
- Putting the output files in one jar file



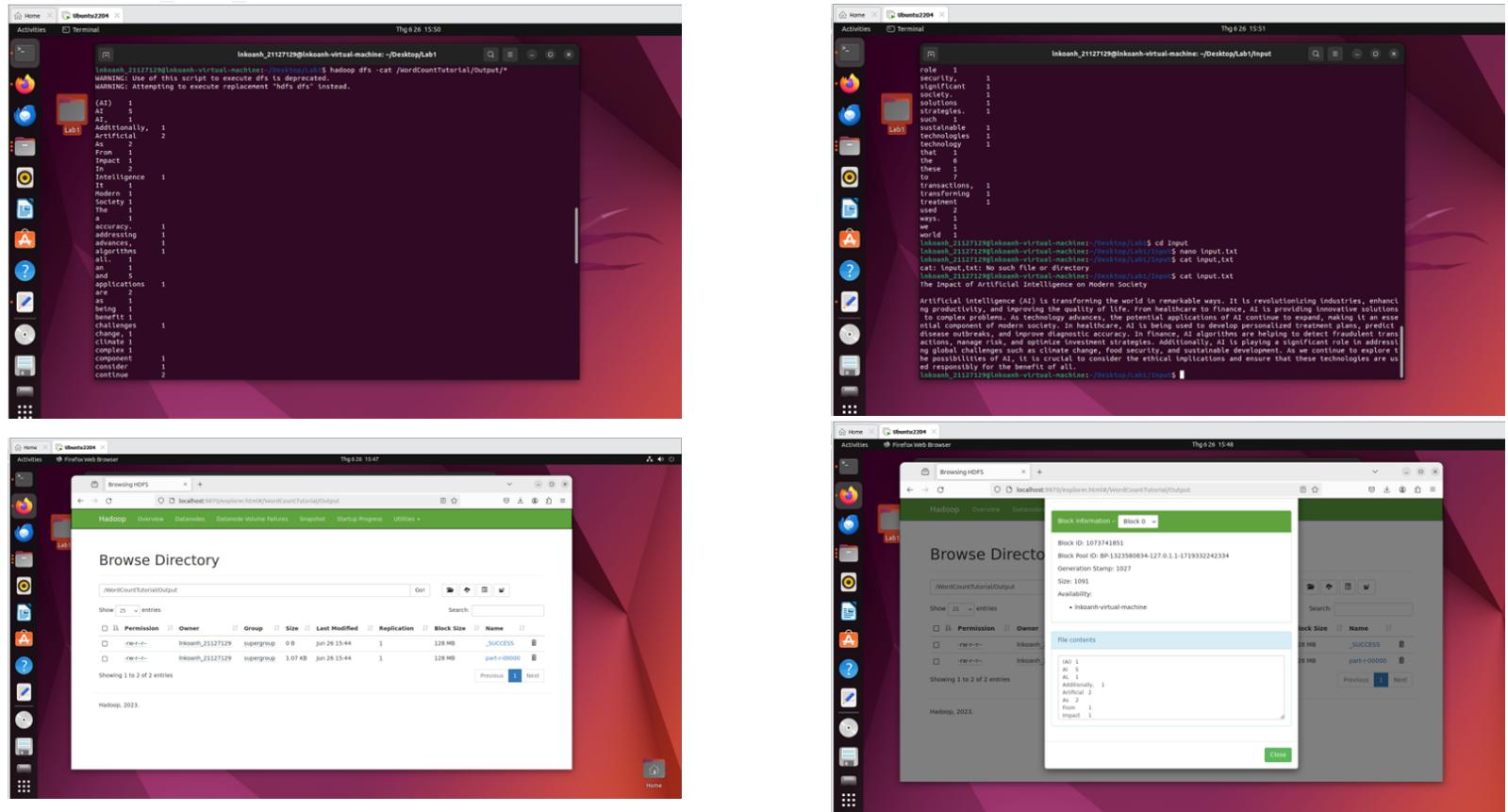
- Running the jar file on Hadoop



- The input file

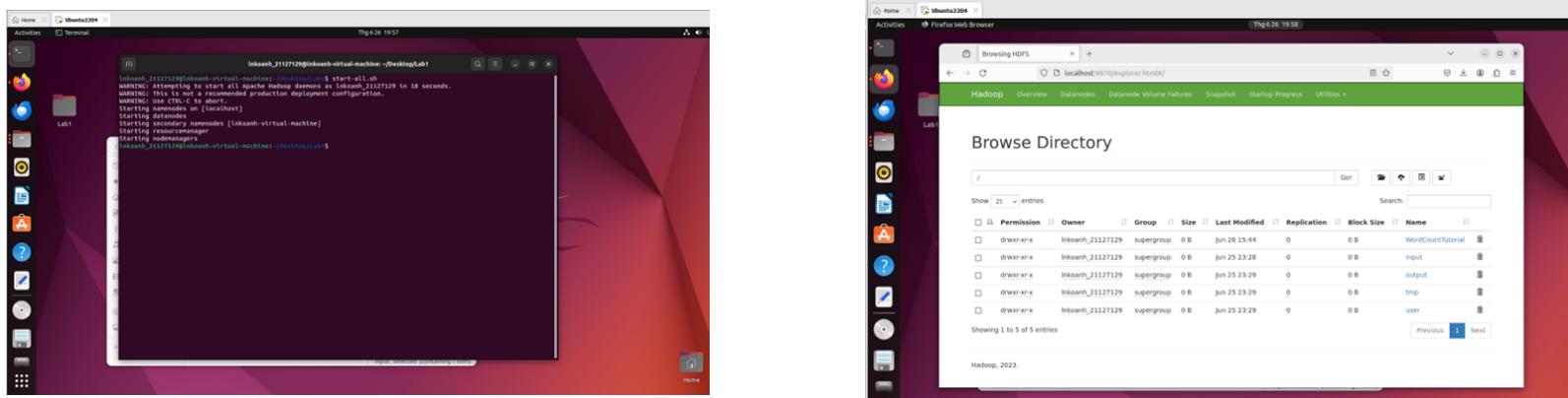


- Finally, we can check the result by using Terminal or Hadoop Web Interface

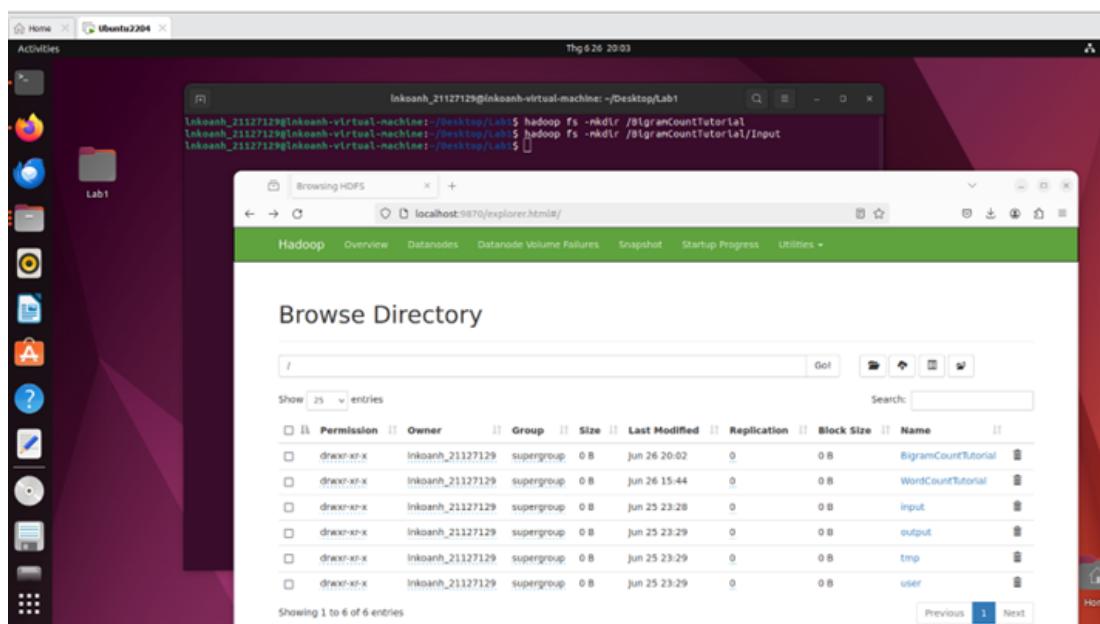


3.2 Bigrams Count

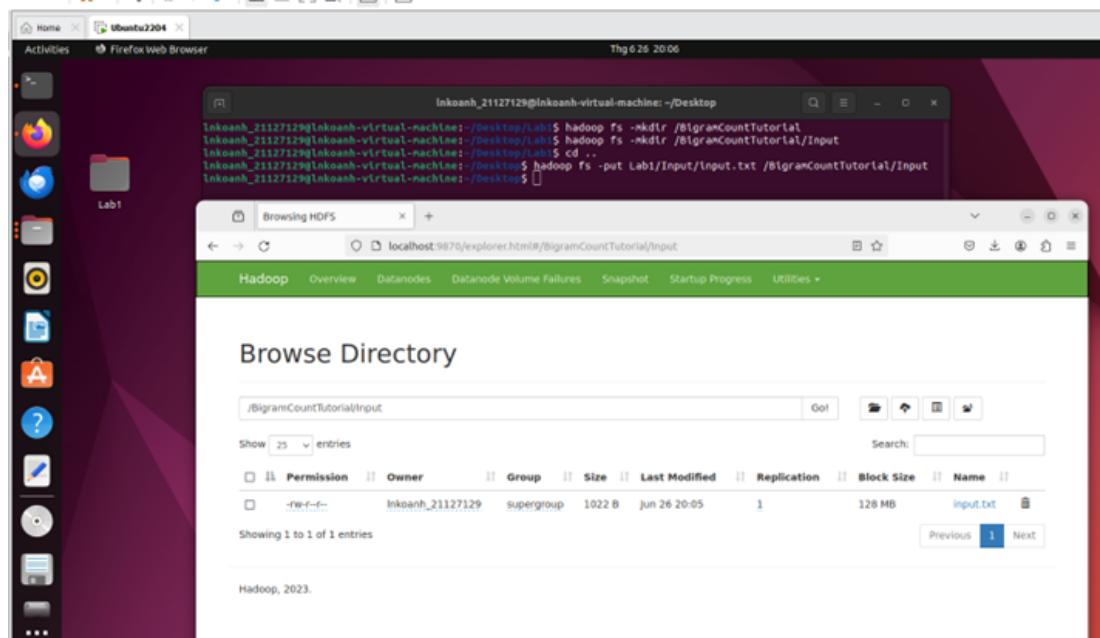
- Firstly, we need to start all the Apache Hadoop Daemons.



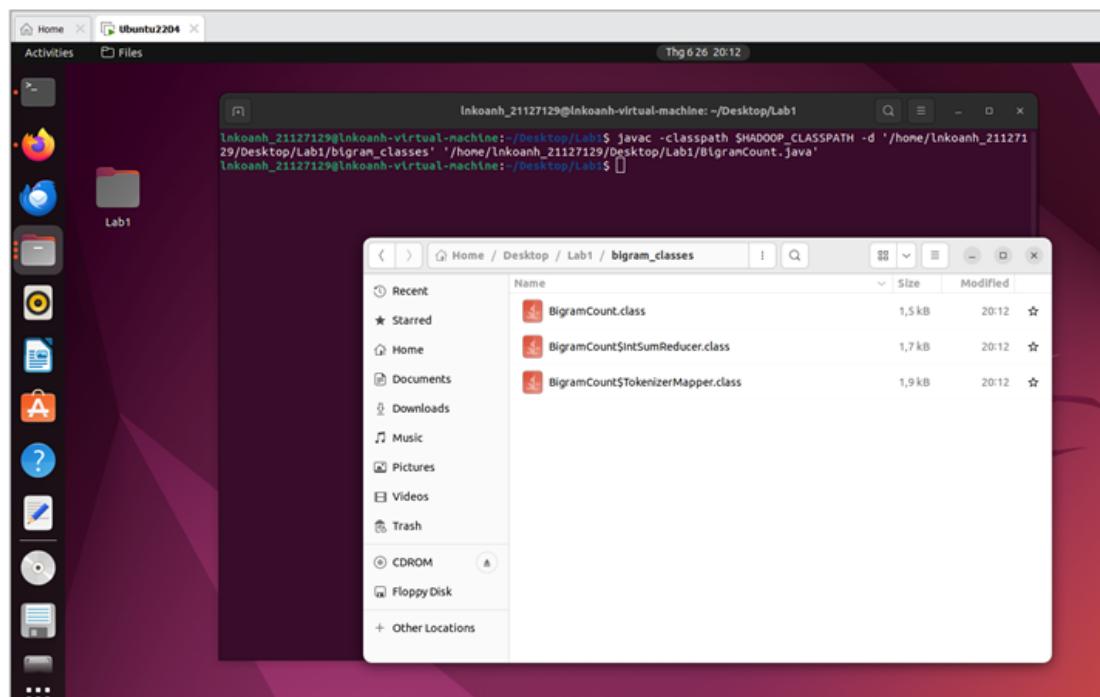
- Create directoys on HDFS



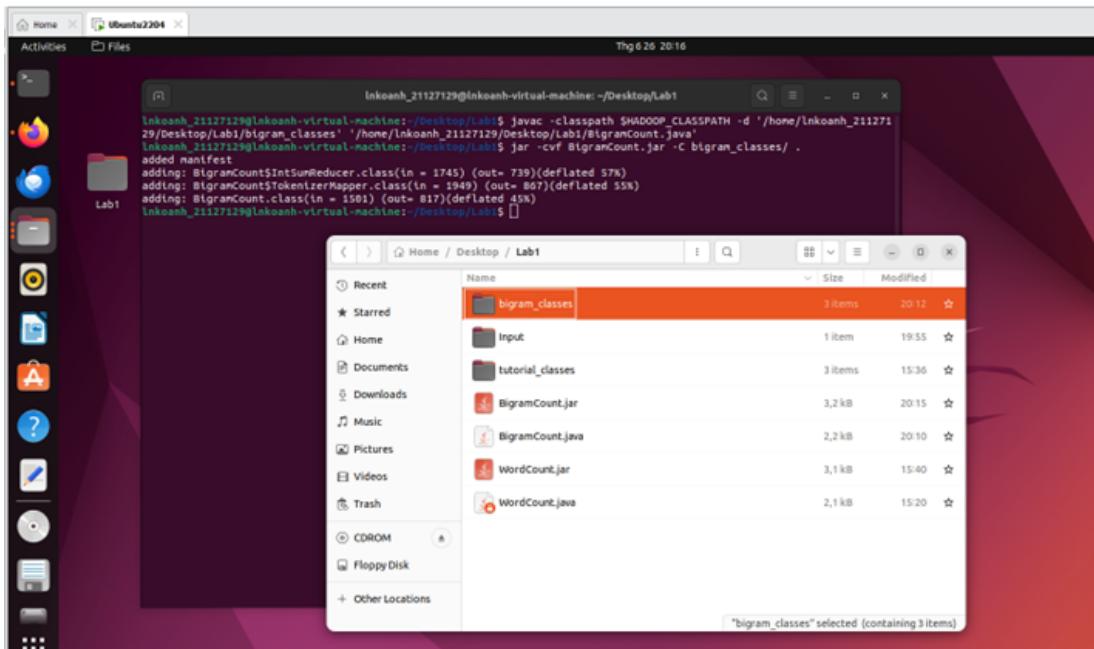
- Upload file "input.txt" on HDFS



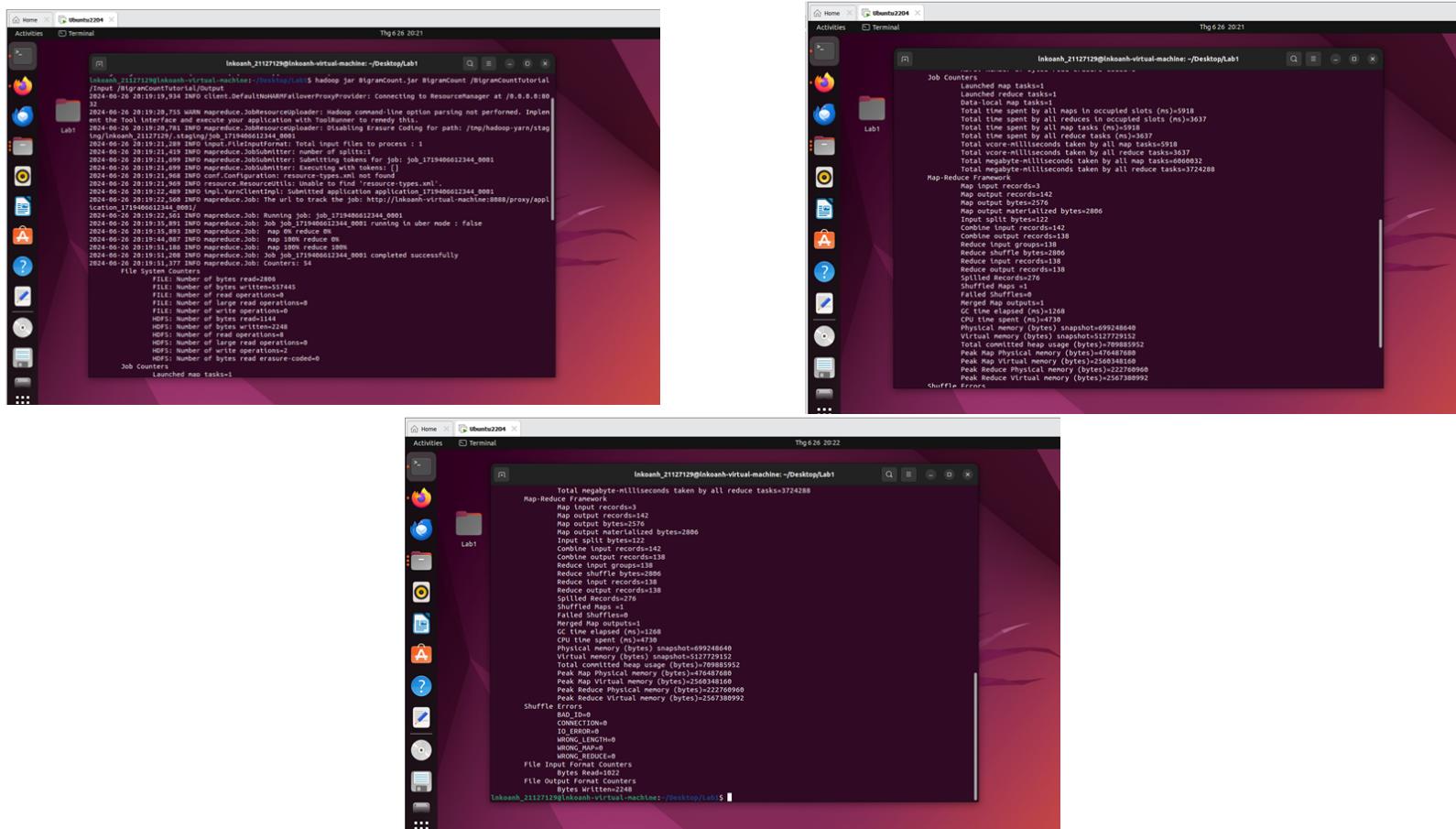
- Compile file "BigramCount.java"



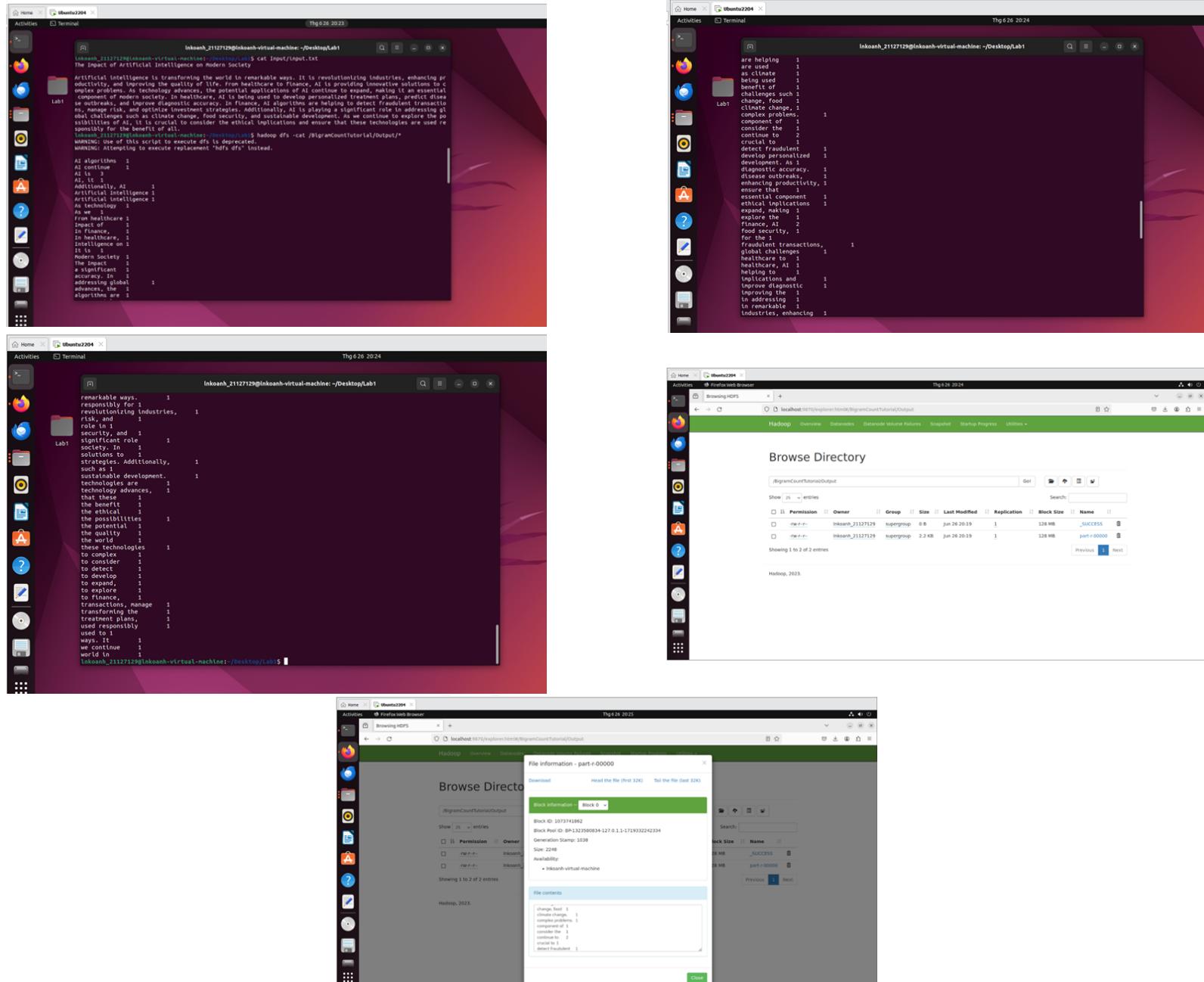
- Putting files in one jar file



- Running the jar file on Hadoop



- Finally, we can check the result by using Terminal or Hadoop Web Interface



4 Reflection

4.1 Challenges and Bugs Encountered

- The stage of installing the virtual machine and the Ubuntu operating system encountered difficulties in selecting the version and configuration.
- It is not possible to create a folder on HDFS due to the default root directory starting at /user/username.
- In section 4: Setting up Fully Distributed Mode, our team is not located near each other and has not yet found an appropriate solution, so we have not been able to complete the requirement.

4.2 How We Overcame It

- During our setup of the virtual machine and Ubuntu OS, we researched different versions, selected the most suitable for our project needs, and followed a detailed YouTube tutorial for installation guidance. We verified each step against the tutorial, adjusted configurations as necessary, and documented the process for future reference and troubleshooting. This method ensured a smooth setup and optimized our Ubuntu environment effectively.
- It is not possible to create a folder on HDFS due to the default root directory starting at /user/username. To resolve this issue, the folder needs to be created in that directory first, or / should be added at the beginning of the path in any command.

4.3 Lessons Learned

Setting up our Hadoop cluster and running MapReduce programs provided valuable insights into big data processing and distributed computing. We delved deep into distributed systems, mastering key components like NameNode and DataNode, enhancing our cluster architecture knowledge.

The process significantly improved our Linux command skills as we handled tasks such as OS installation, user permissions, network configurations, and software setups. Troubleshooting various issues strengthened our system administration capabilities, particularly in diagnosing configuration errors and optimizing cluster performance.

In summary, our Hadoop experience was educational, boosting our understanding of distributed systems, refining Linux proficiency, and enhancing problem-solving skills crucial for managing

large-scale data processing.

Bibliography

- [1] Apache Hadoop 3.3.6 – Hadoop: Setting up a Single Node Cluster <https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html>
- [2] Apache Hadoop 3.3.6 – MapReduce Tutorial https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example%3A_WordCount_v1.0
- [3] Code With Arjun. (2022, October 2). Install Hadoop on Ubuntu (22.04 / 20.04 LTS) | HDFS | Namenode | DataNode | Big Data Analytics [Video]. YouTube <https://www.youtube.com/watch?v=S1bi-uzPtnw>
- [4] Code With Arjun. (2022b, October 12). MapReduce Word Count Example using Hadoop and Java [Video]. YouTube. <https://www.youtube.com/watch?v=qgBu8Go1SyM>
- [5] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters (research.google) <https://research.google/pubs/mapreduce-simplified-data-processing-on-large-clusters/>