Lab 01: A Gentle Introduction to Hadoop

Setting up Single-node Hadoop Cluster

Introduction to MapReduce

Running a warm-up problem: Word Count

Bonus

TEAM HugeData:

MSSV	FULLNAME	TASKS
20120560	Cao Đinh Quí	1, 2.1.3, write report
20120089	Lê Xuân Hoàng	1, 2.1.1, 4
20120130	Đinh Thị Hoàng Linh	1, 2.1.4, 4
20120397	Bùi Quang Tùng	1, 2.1.2, 3

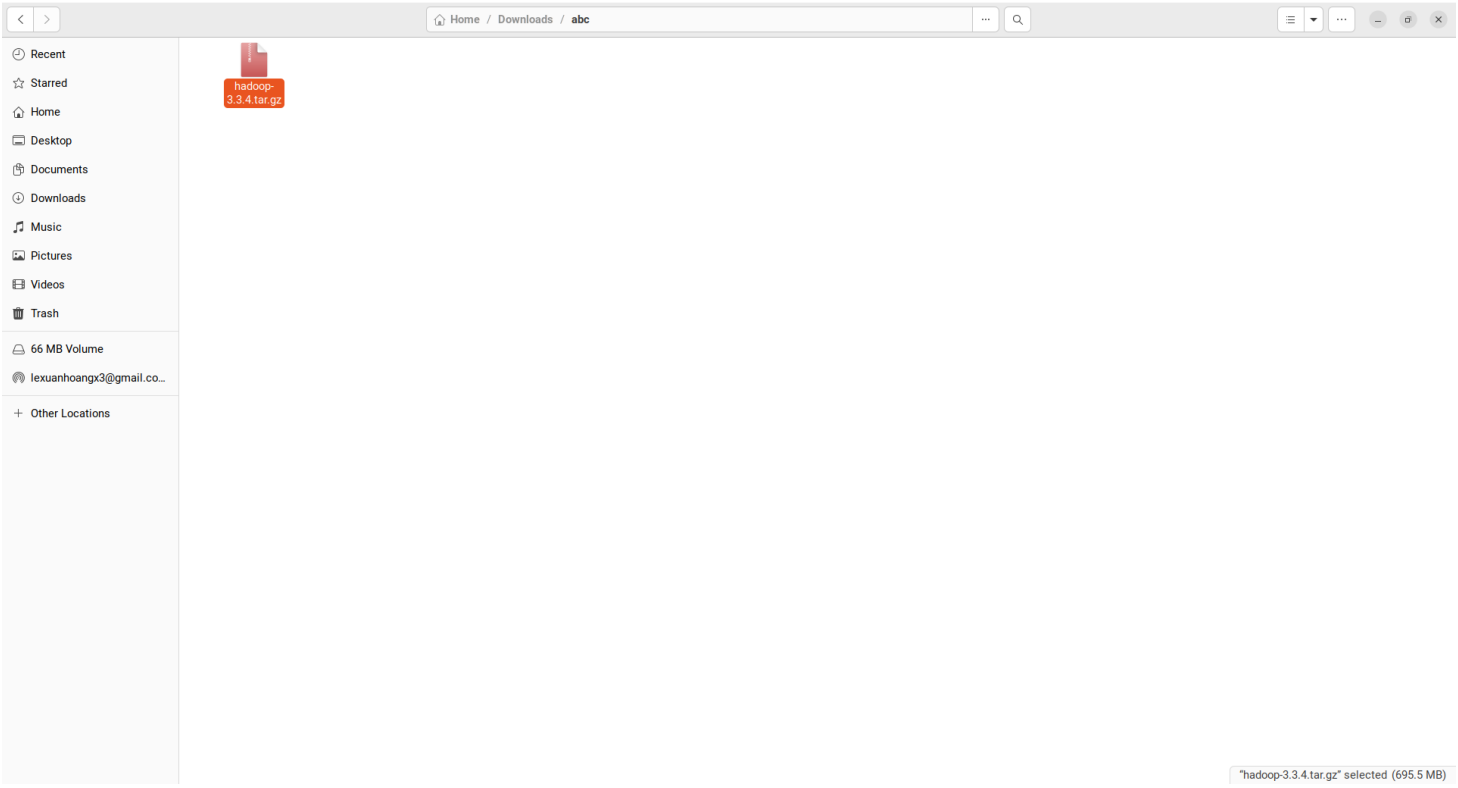
COMPLETION RATE:

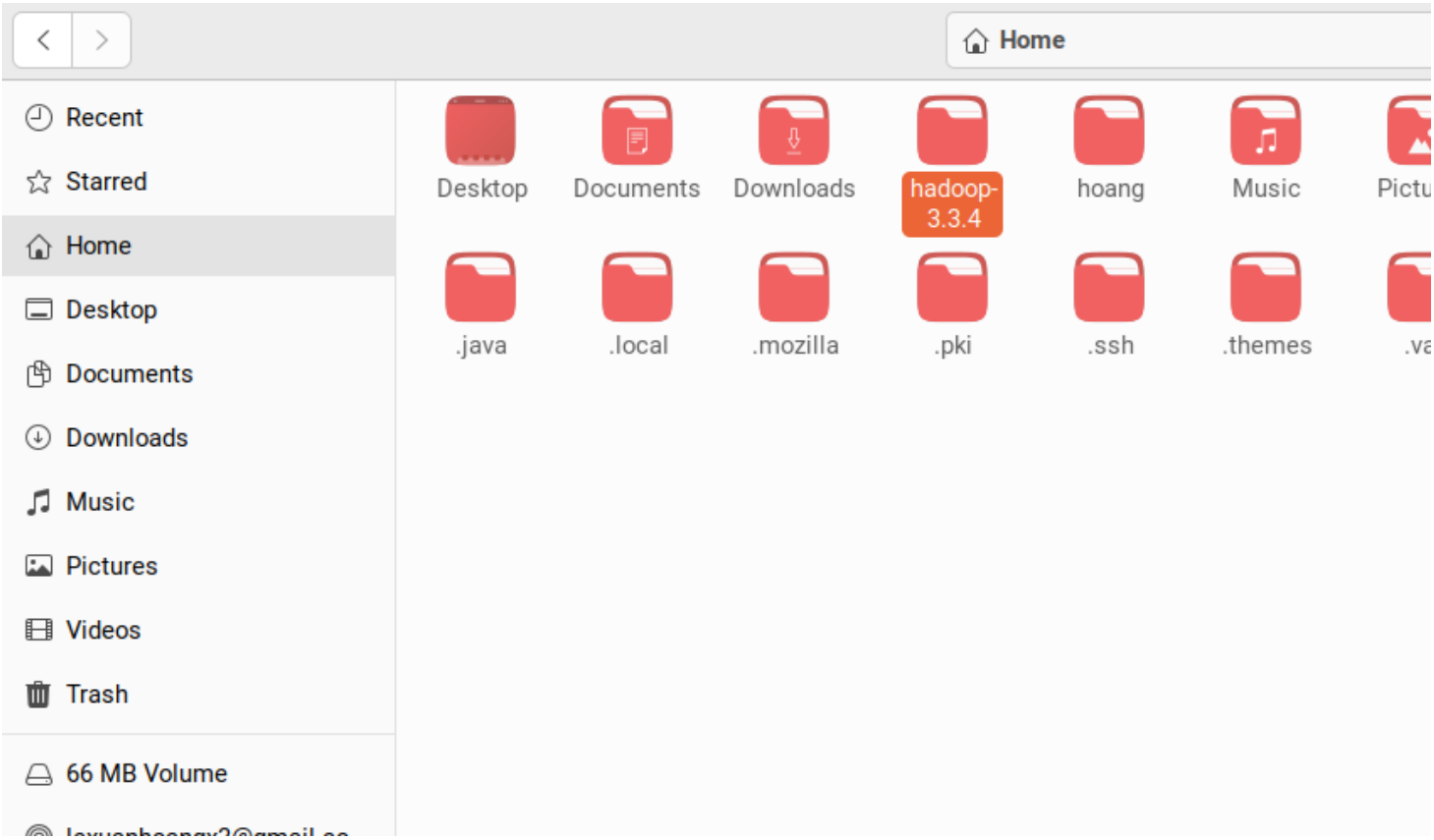
SESSION	RATE
1	100%
2	100%
3	100%
4	?

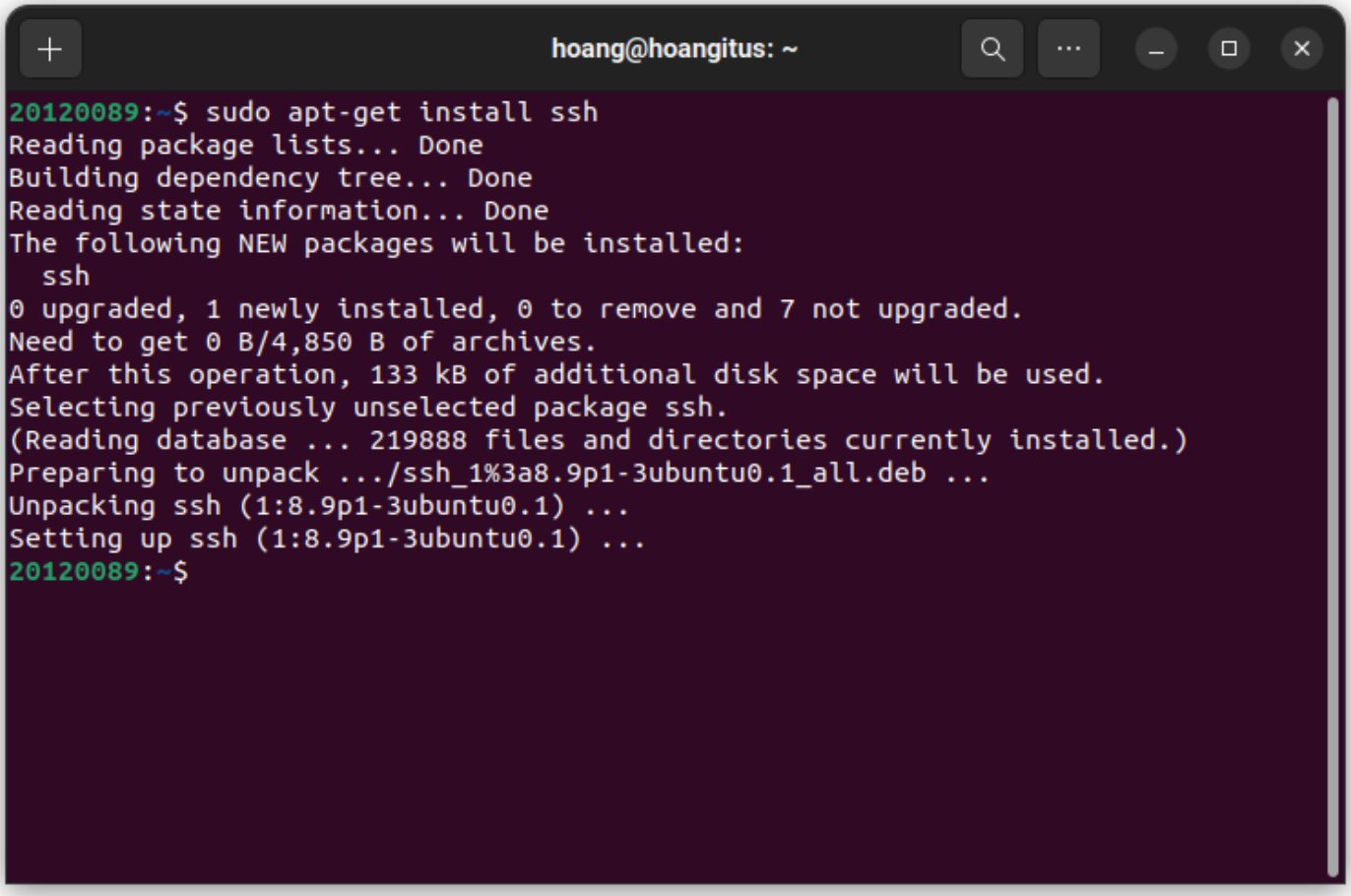
1. Setting up Single-node Hadoop Cluster

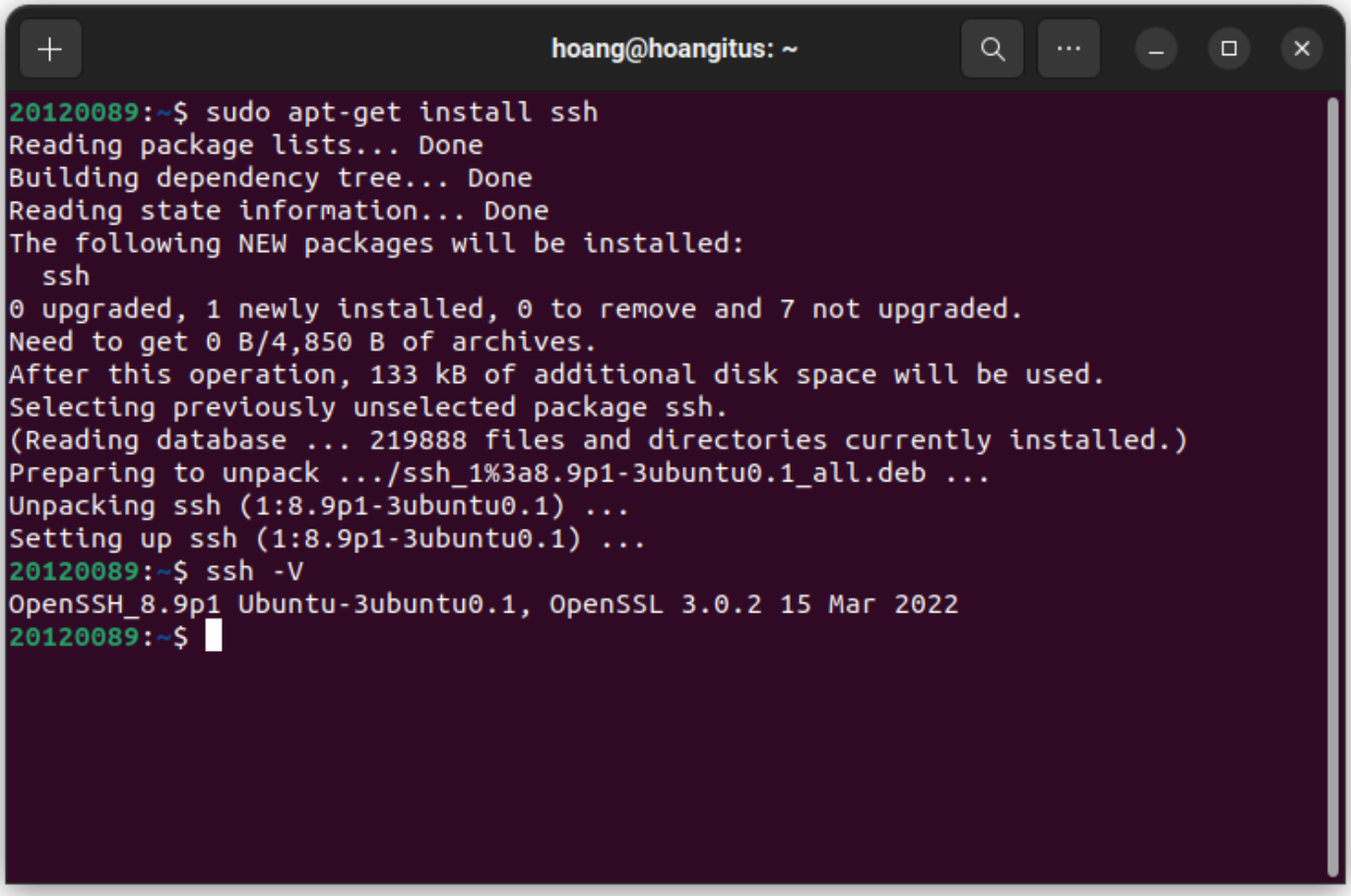
All members finished successfully.

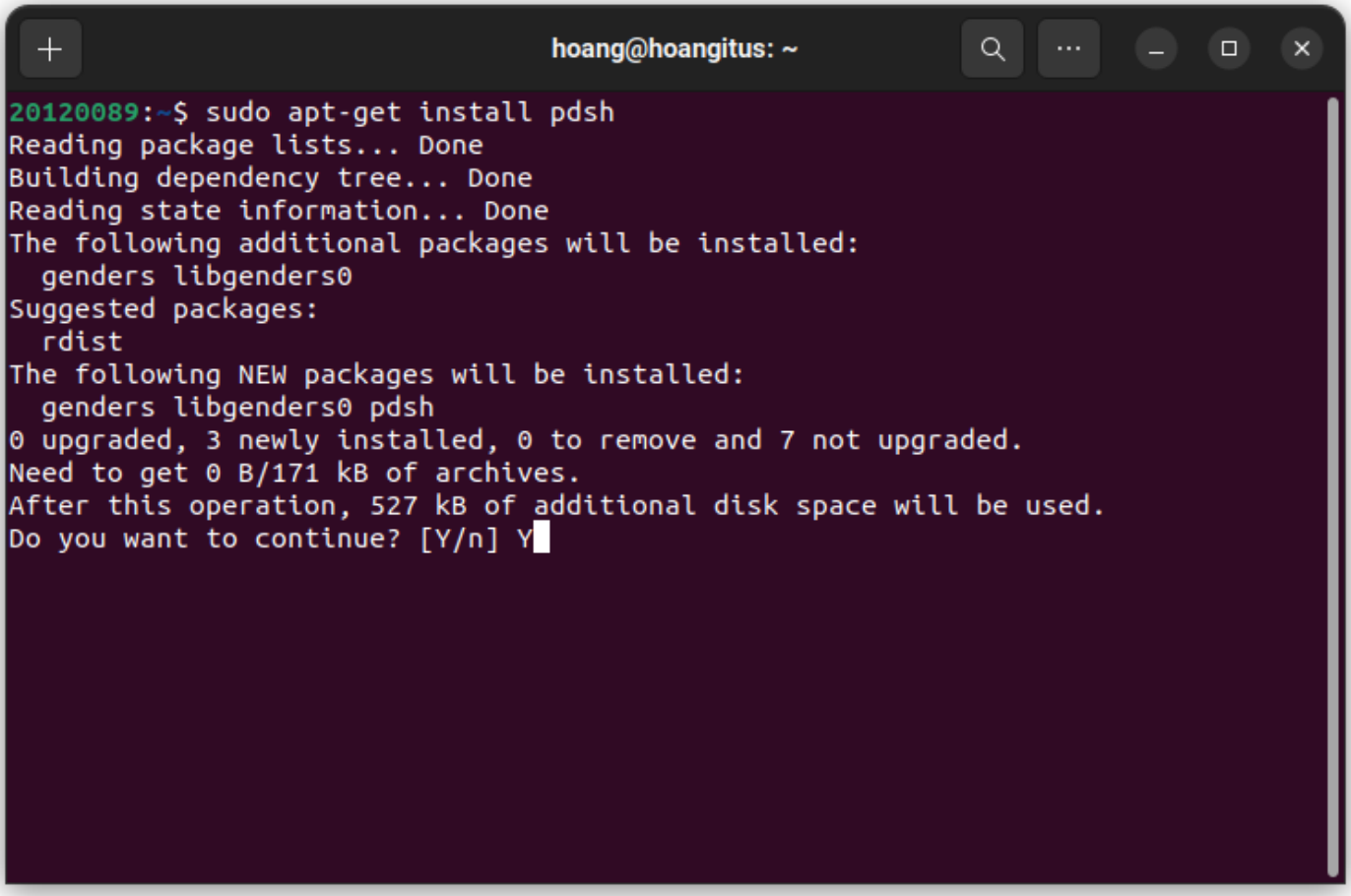
The process:

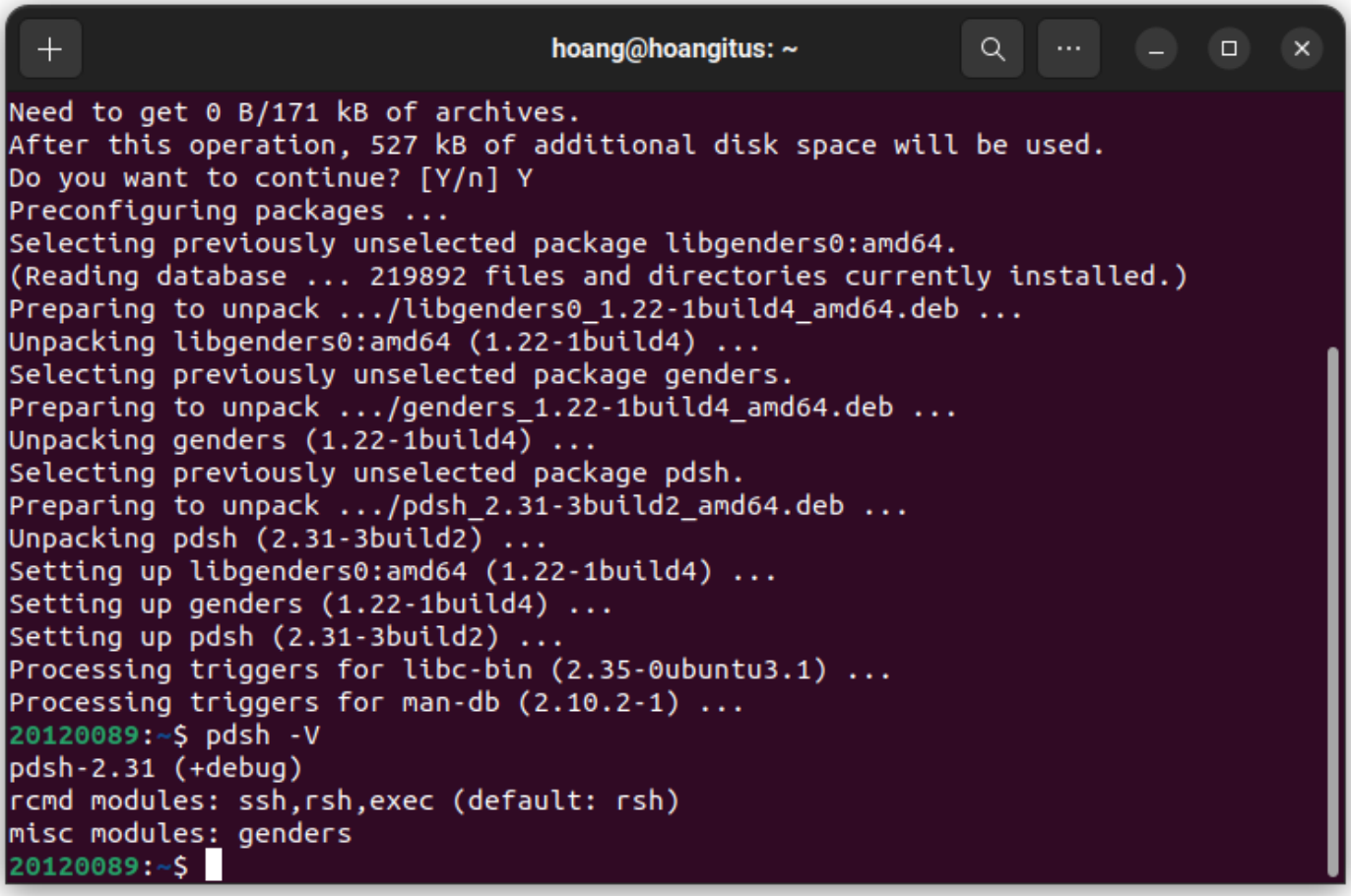


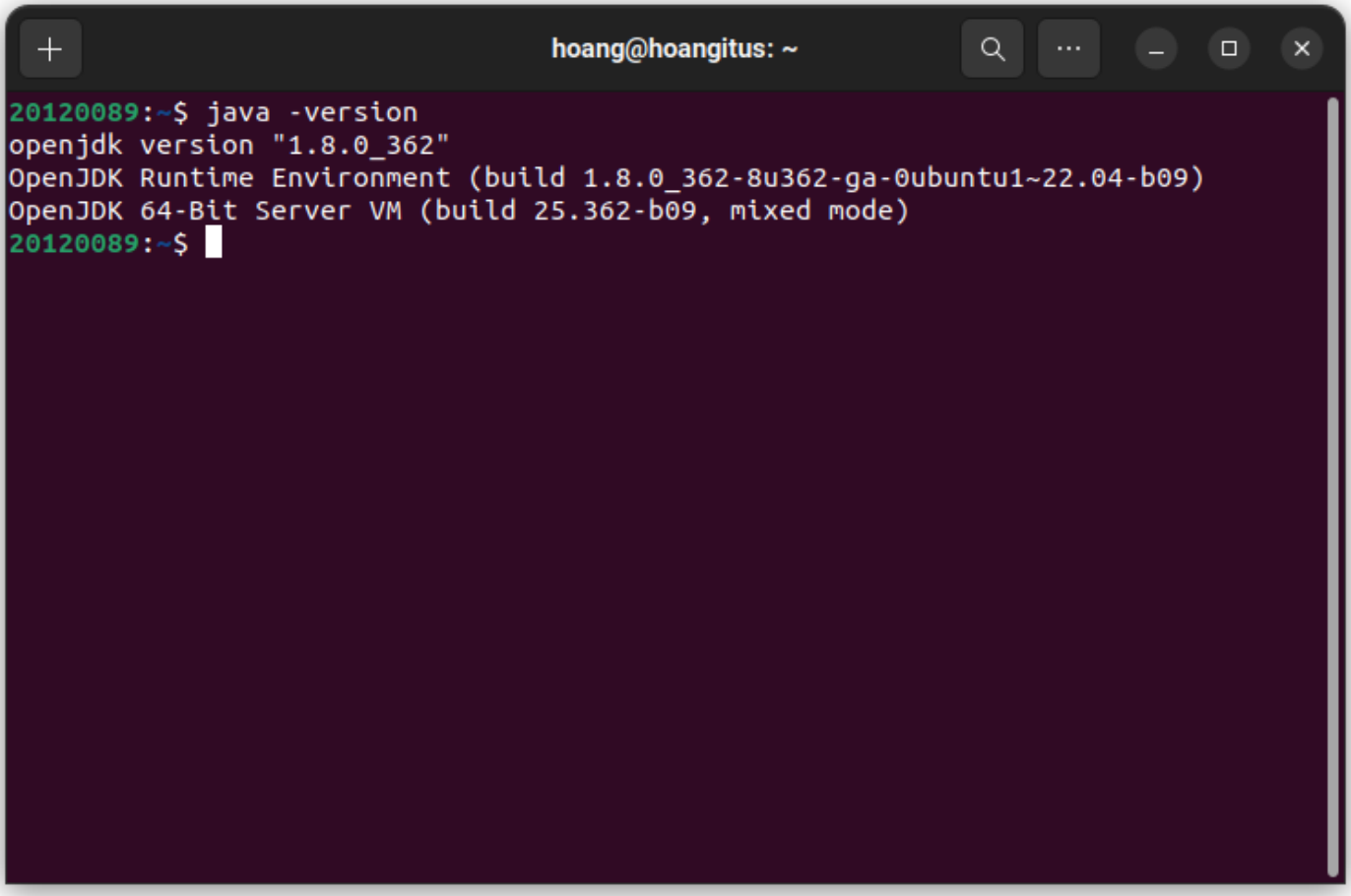

Home
Recent
Starred
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
66 MB Volume
Desktop
Documents
Downloads
hadoop-3.3.4
hoang
Music
.java
.local
.mozilla
.pki
.ssh
.themes

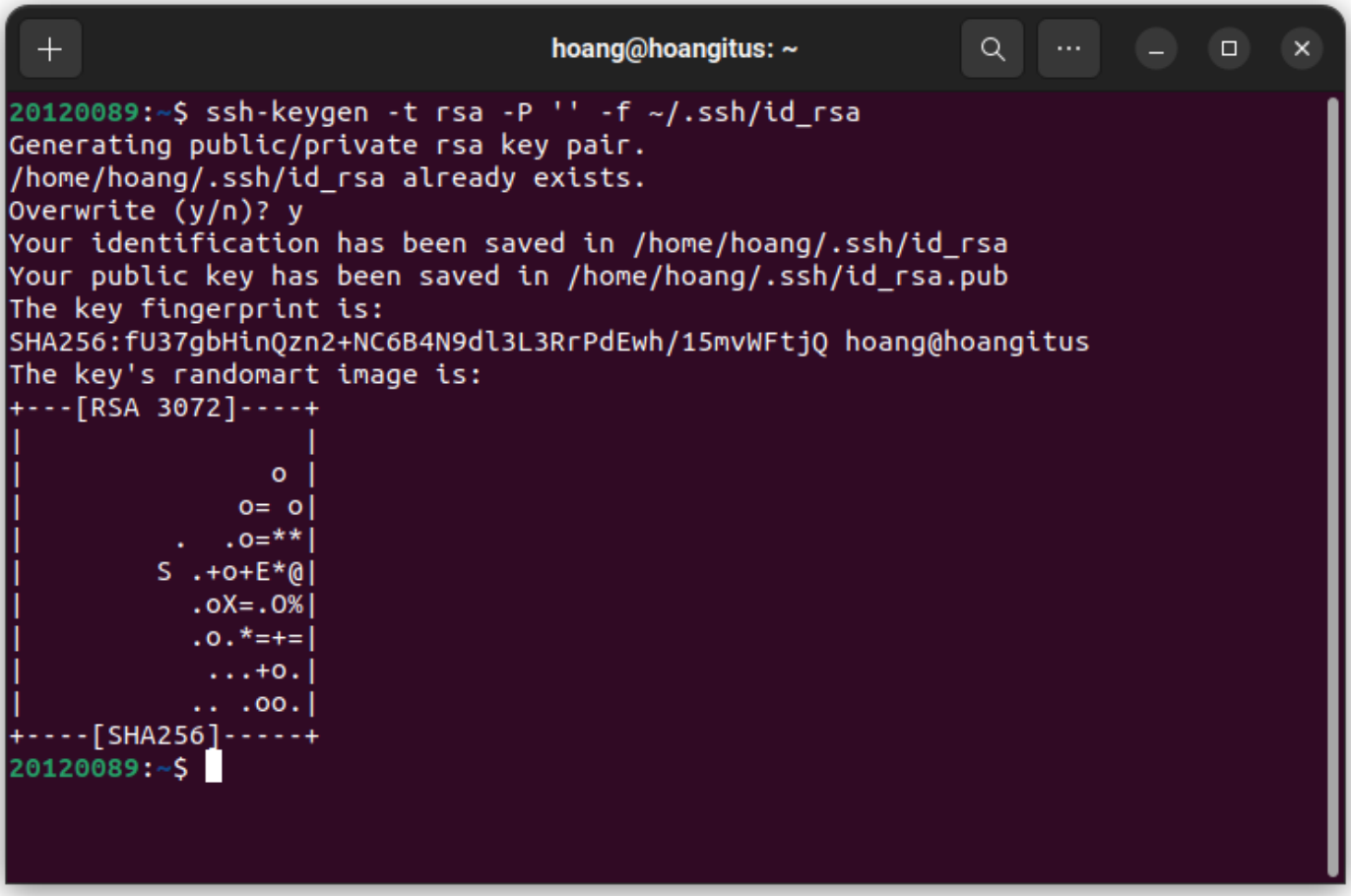

hoang@hoangitus: ~
20120089:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
ssh
0 upgraded, 1 newly installed, 0 to remove and 7 not upgraded.
Need to get 0 B/4,850 B of archives.
After this operation, 133 kB of additional disk space will be used.
Selecting previously unselected package ssh.
(Reading database ... 219888 files and directories currently installed.)
Preparing to unpack .../ssh_1%3a8.9p1-3ubuntu0.1_all.deb ...
Unpacking ssh (1:8.9p1-3ubuntu0.1) ...
Setting up ssh (1:8.9p1-3ubuntu0.1) ...
20120089:~$

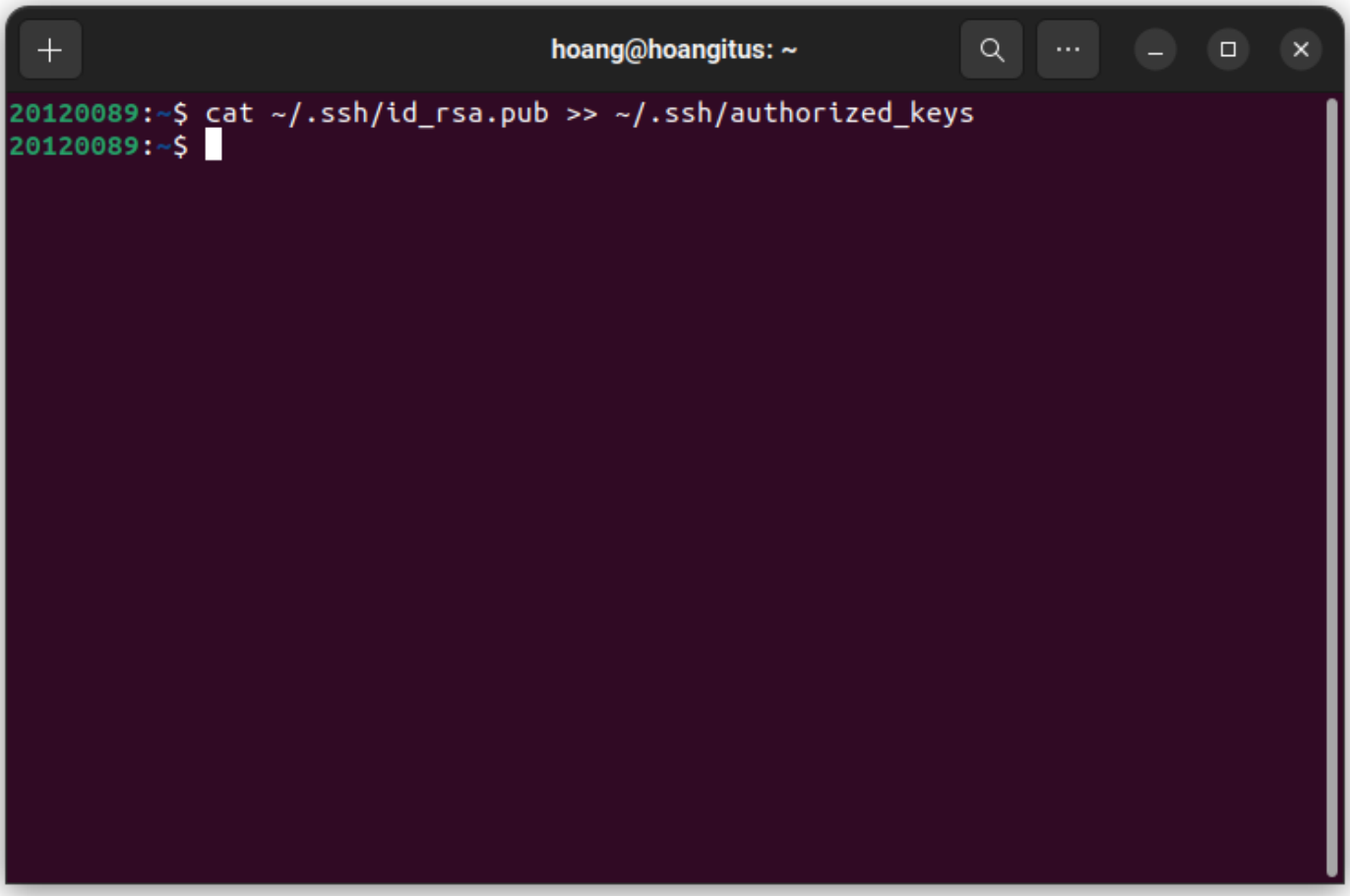

hoang@hoangitus: ~
20120089:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following NEW packages will be installed:
ssh
0 upgraded, 1 newly installed, 0 to remove and 7 not upgraded.
Need to get 0 B/4,850 B of archives.
After this operation, 133 kB of additional disk space will be used.
Selecting previously unselected package ssh.
(Reading database ... 219888 files and directories currently installed.)
Preparing to unpack .../ssh_1%3a8.9p1-3ubuntu0.1_all.deb ...
Unpacking ssh (1:8.9p1-3ubuntu0.1) ...
Setting up ssh (1:8.9p1-3ubuntu0.1) ...
20120089:~$ ssh -V
OpenSSH_8.9p1 Ubuntu-3ubuntu0.1, OpenSSL 3.0.2 15 Mar 2022
20120089:~$

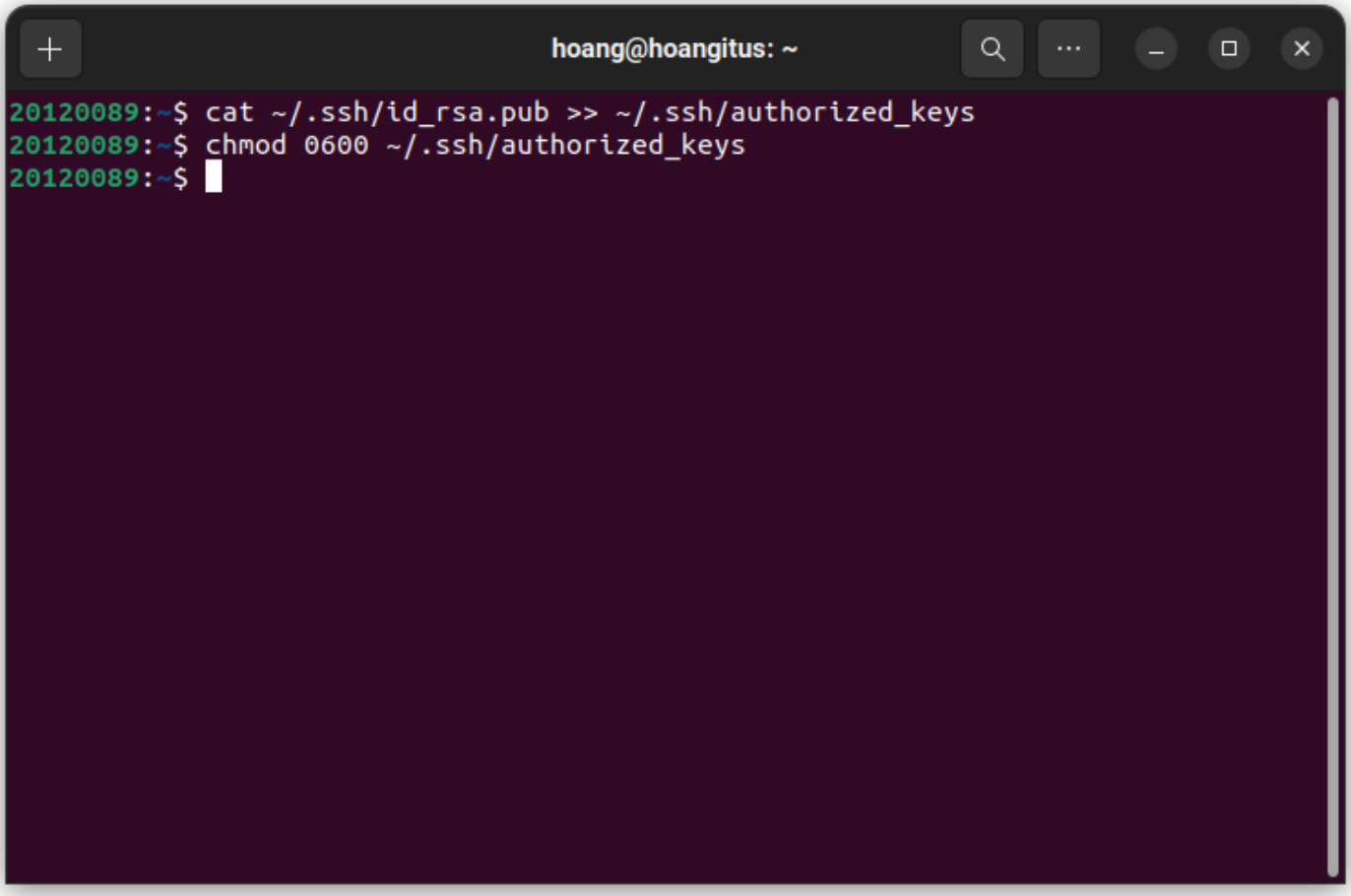

hoang@hoangitus: ~
20120089:~$ sudo apt-get install pdsh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
genders libgenders0
Suggested packages:
rdist
The following NEW packages will be installed:
genders libgenders0 pdsh
0 upgraded, 3 newly installed, 0 to remove and 7 not upgraded.
Need to get 0 B/171 kB of archives.
After this operation, 527 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y

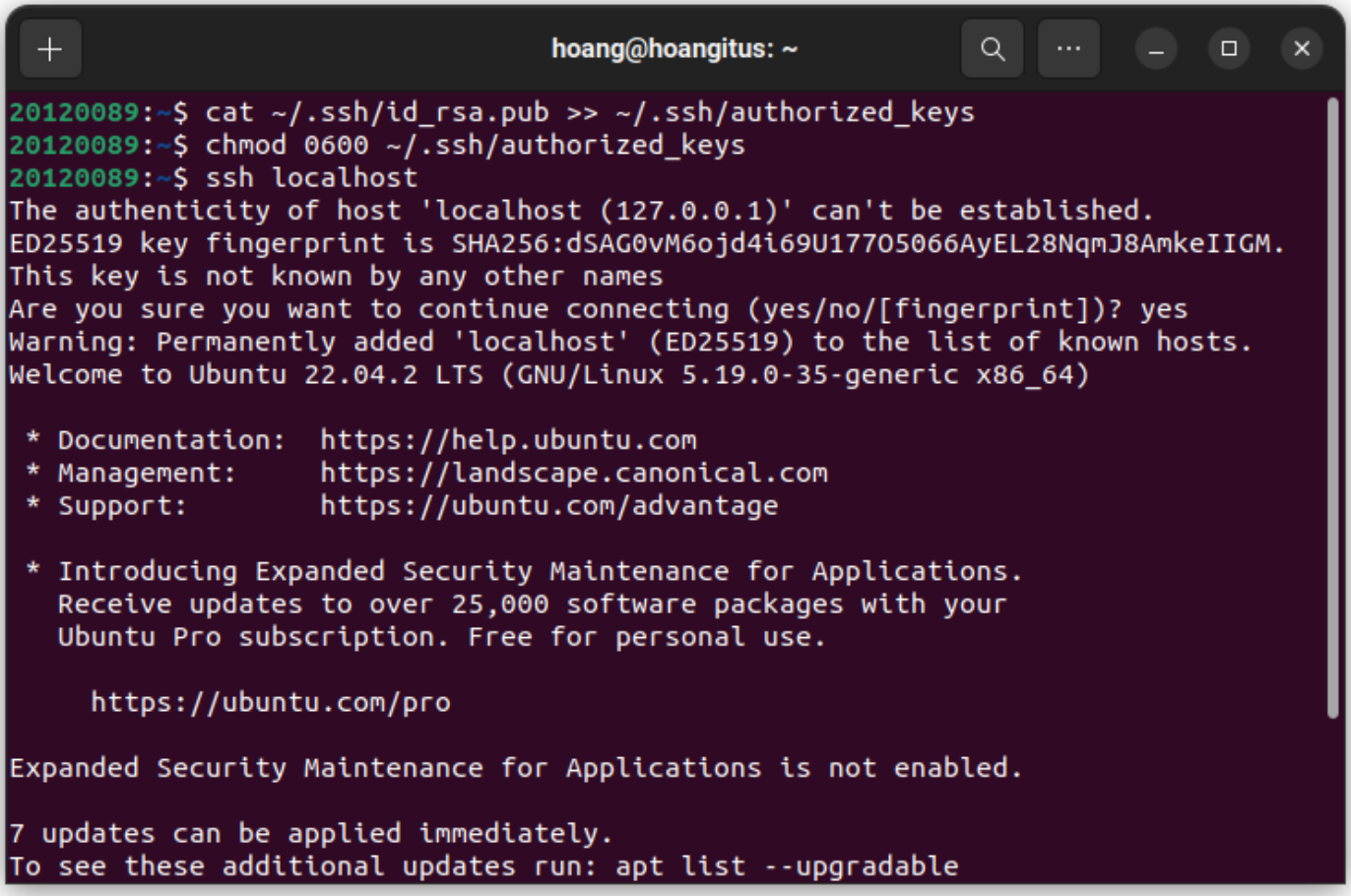

hoang@hoangitus: ~
Need to get 0 B/171 kB of archives.
After this operation, 527 kB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Preconfiguring packages ...
Selecting previously unselected package libgenders0:amd64.
(Reading database ... 219892 files and directories currently installed.)
Preparing to unpack .../libgenders0_1.22-1build4_amd64.deb ...
Unpacking libgenders0:amd64 (1.22-1build4) ...
Selecting previously unselected package genders.
Preparing to unpack .../genders_1.22-1build4_amd64.deb ...
Unpacking genders (1.22-1build4) ...
Selecting previously unselected package pdsh.
Preparing to unpack .../pdsh_2.31-3build2_amd64.deb ...
Unpacking pdsh (2.31-3build2) ...
Setting up libgenders0:amd64 (1.22-1build4) ...
Setting up genders (1.22-1build4) ...
Setting up pdsh (2.31-3build2) ...
Processing triggers for libc-bin (2.35-0ubuntu3.1) ...
Processing triggers for man-db (2.10.2-1) ...
20120089:~$ pdsh -V
pdsh-2.31 (+debug)
rcmd modules: ssh,rsh,exec (default: rsh)
misc modules: genders
20120089:~$

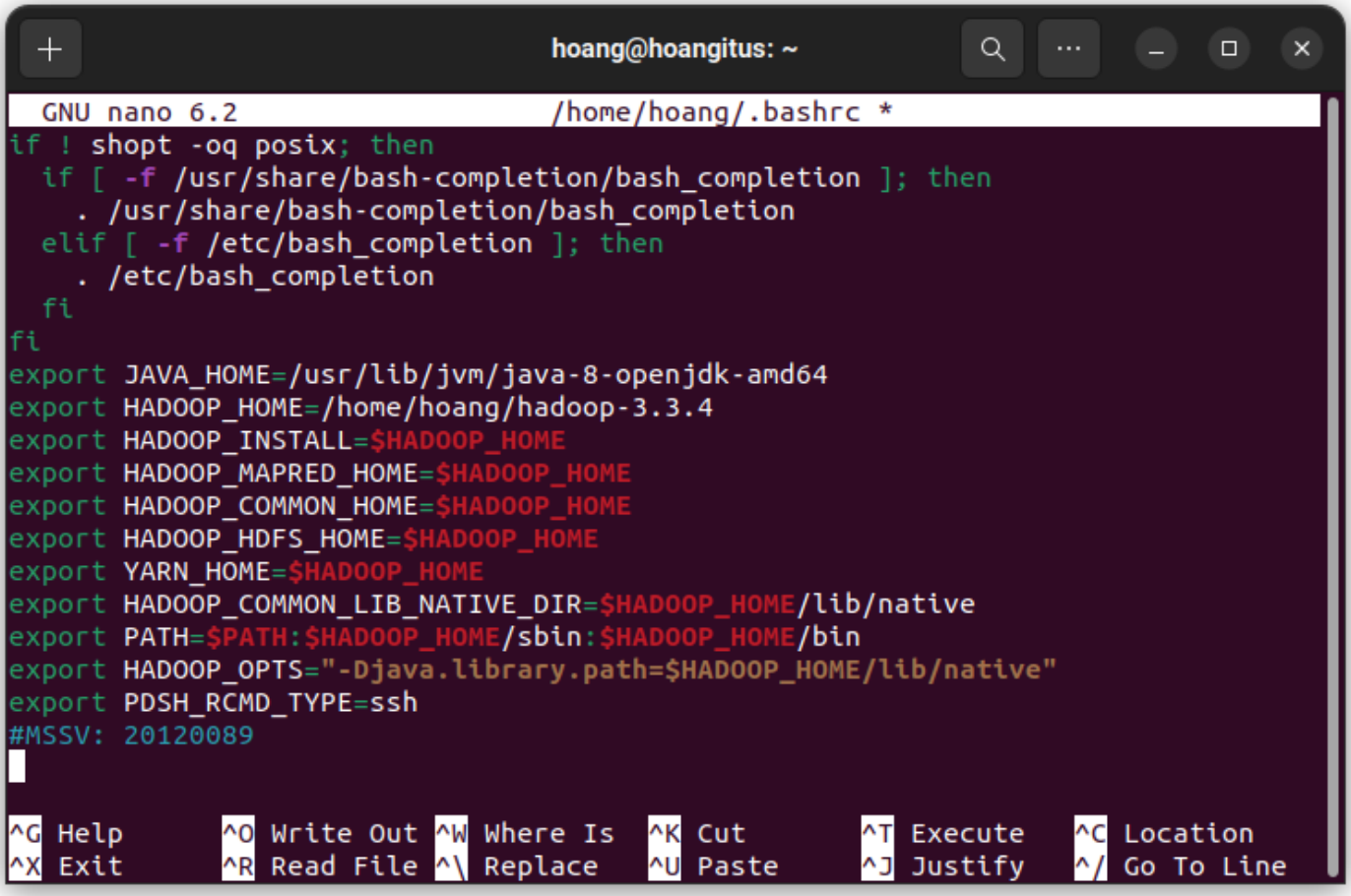

hoang@hoangitus: ~
20120089:~$ java -version
openjdk version "1.8.0_362"
OpenJDK Runtime Environment (build 1.8.0_362-8u362-ga-0ubuntu1~22.04-b09)
OpenJDK 64-Bit Server VM (build 25.362-b09, mixed mode)
20120089:~$

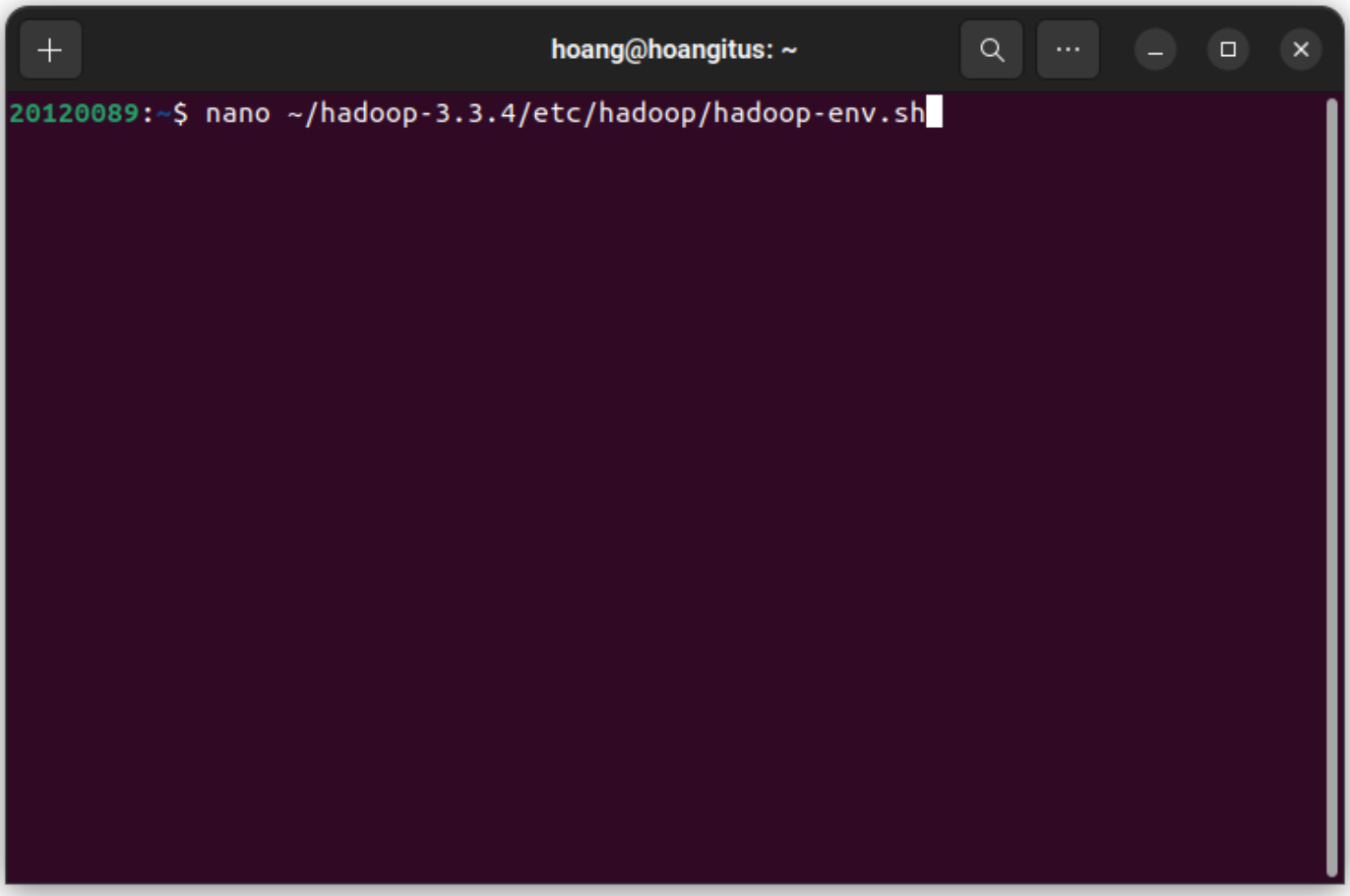

hoang@hoangitus: ~
20120089:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/home/hoang/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/hoang/.ssh/id_rsa
Your public key has been saved in /home/hoang/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:fU37gbHinQzn2+NC6B4N9dl3L3RrPdEwh/15mvWFtjQ hoang@hoangitus
The key's randomart image is:
+---[RSA 3072]----+
| |
| o |
| o= o|
| . .o=**|
| S .+o+E*@|
| .oX=.O%|
| .o.*=+=|
| ...+o.|
| .. .oo.|
+----[SHA256]-----+
20120089:~$

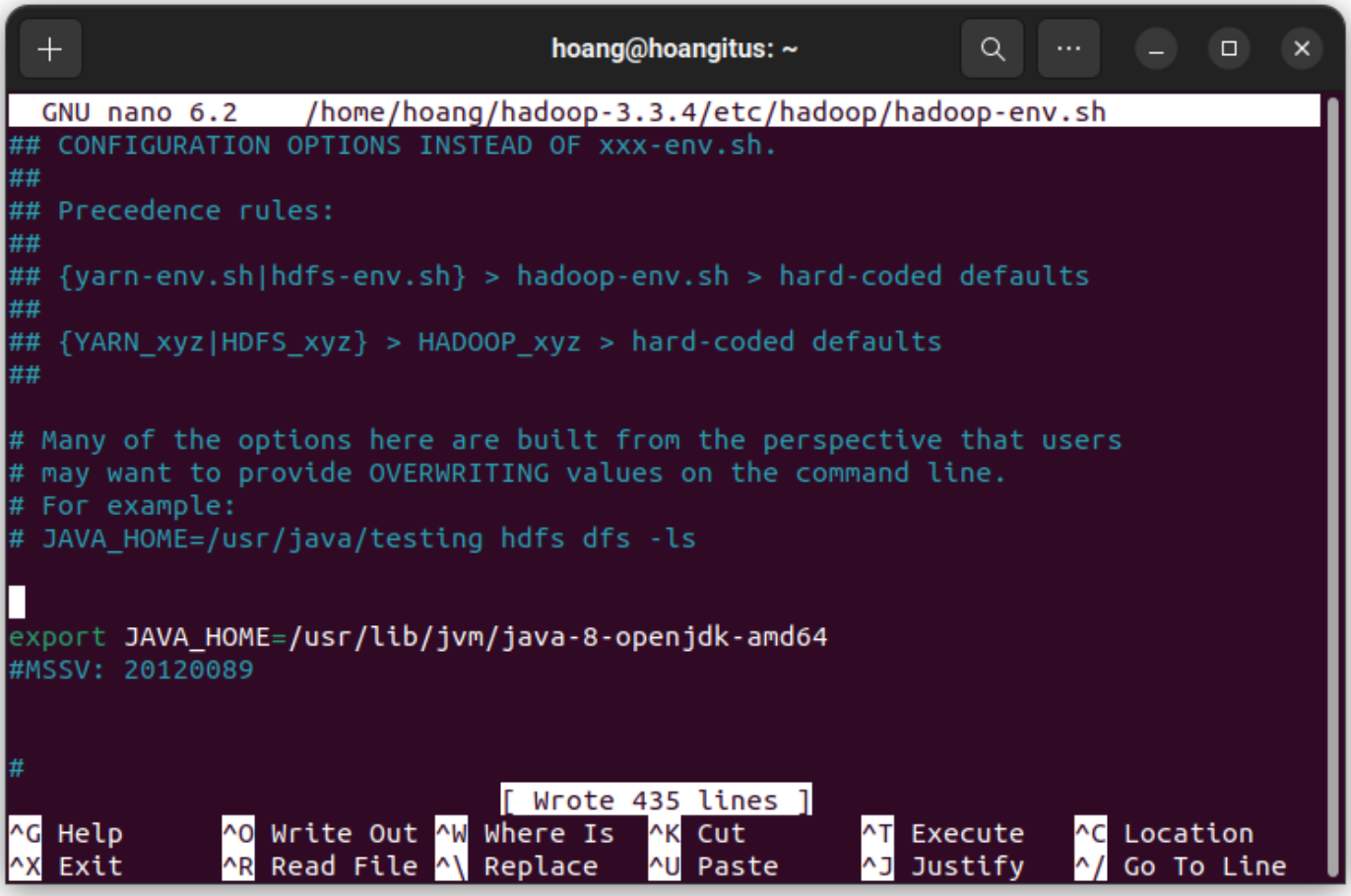

hoang@hoangitus: ~
20120089:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
20120089:~$

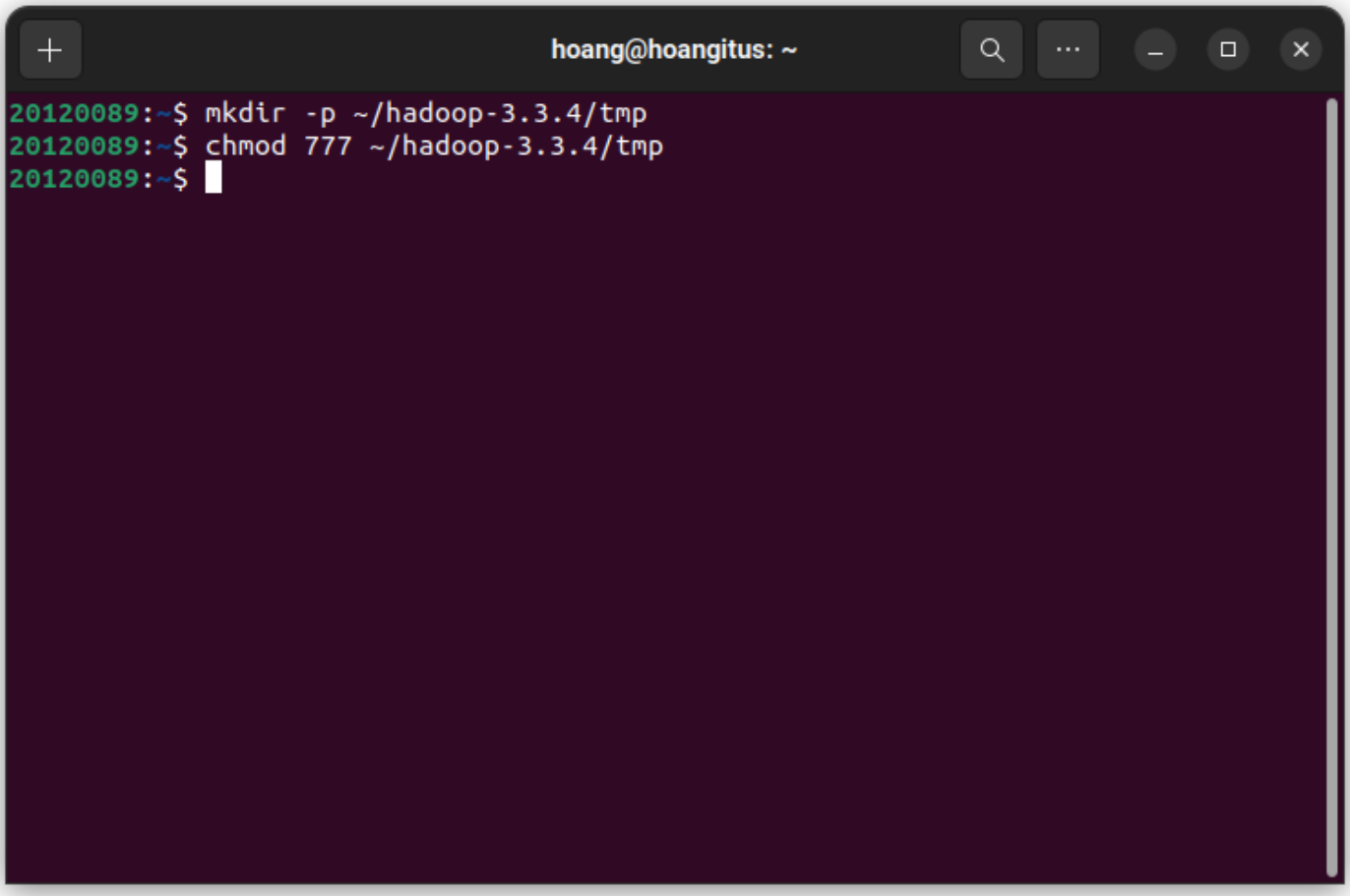

hoang@hoangitus: ~
20120089:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
20120089:~$ chmod 0600 ~/.ssh/authorized_keys
20120089:~$

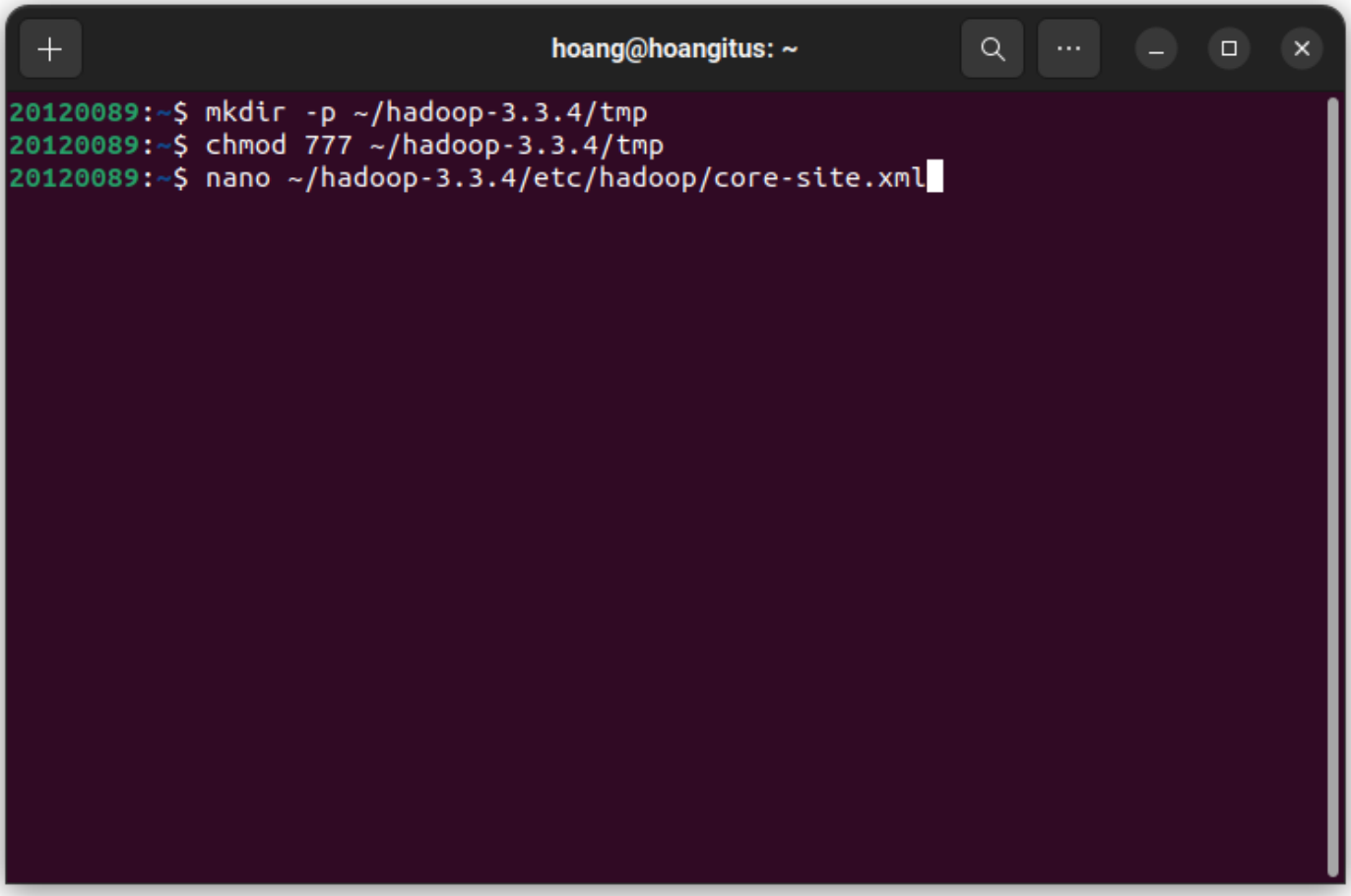

hoang@hoangitus: ~
20120089:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
20120089:~$ chmod 0600 ~/.ssh/authorized_keys
20120089:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:dSAG0vM6ojd4i69U177O5066AyEL28NqmJ8AmkeIIGM.
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 22.04.2 LTS (GNU/Linux 5.19.0-35-generic x86_64)
* Documentation: https://help.ubuntu.com
* Management: https://landscape.canonical.com
* Support: https://ubuntu.com/advantage
* Introducing Expanded Security Maintenance for Applications.
Receive updates to over 25,000 software packages with your
Ubuntu Pro subscription. Free for personal use.
https://ubuntu.com/pro
Expanded Security Maintenance for Applications is not enabled.
7 updates can be applied immediately.
To see these additional updates run: apt list --upgradable

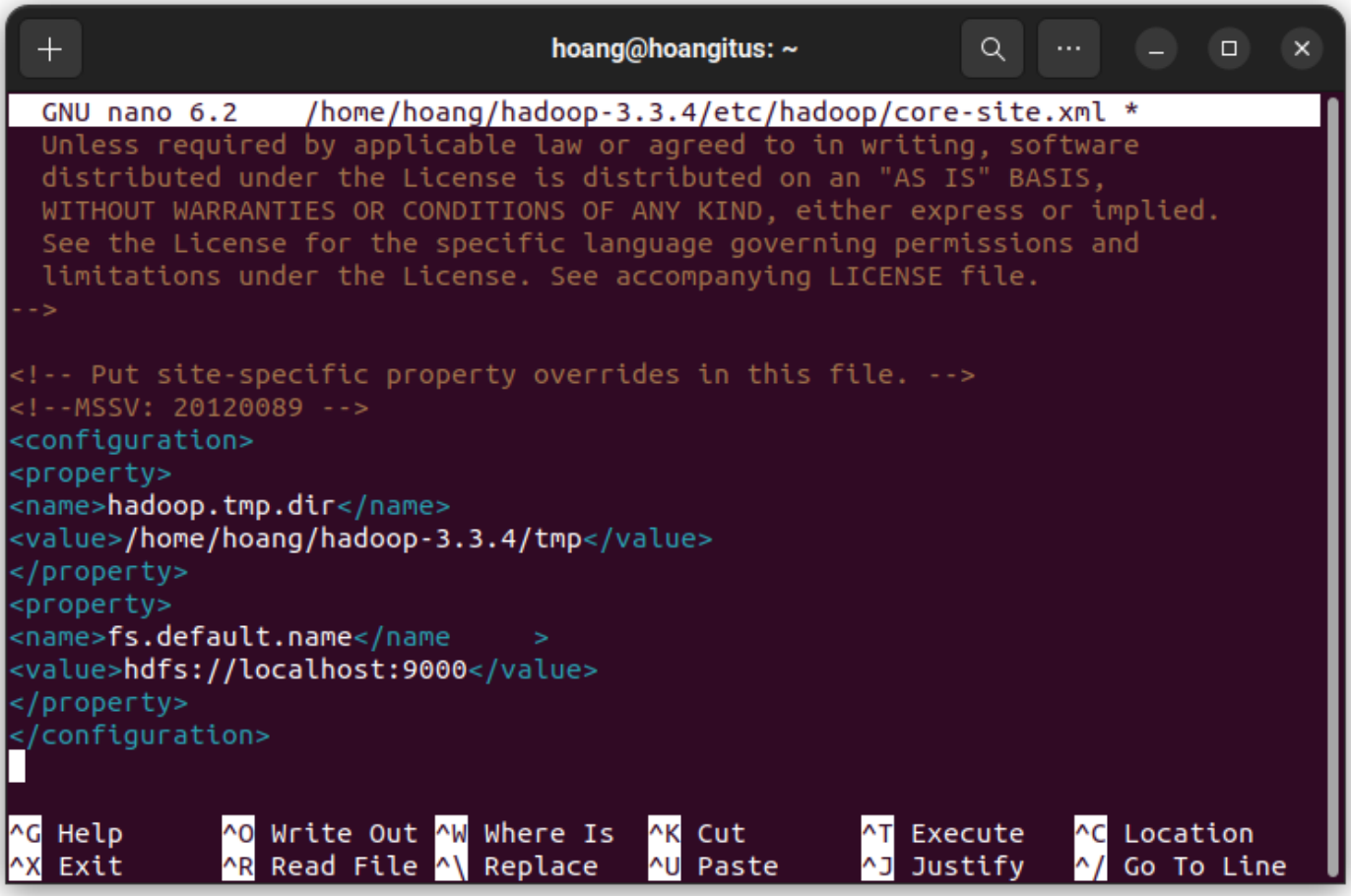

hoang@hoangitus: ~
GNU nano 6.2 /home/hoang/.bashrc *
if ! shopt -oq posix; then
if [-f /usr/share/bash-completion/bash_completion]; then
. /usr/share/bash-completion/bash_completion
elif [-f /etc/bash_completion]; then
. /etc/bash_completion
fi
fi
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/home/hoang/hadoop-3.3.4
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export PDSH_RCMD_TYPE=ssh
#MSSV: 20120089
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

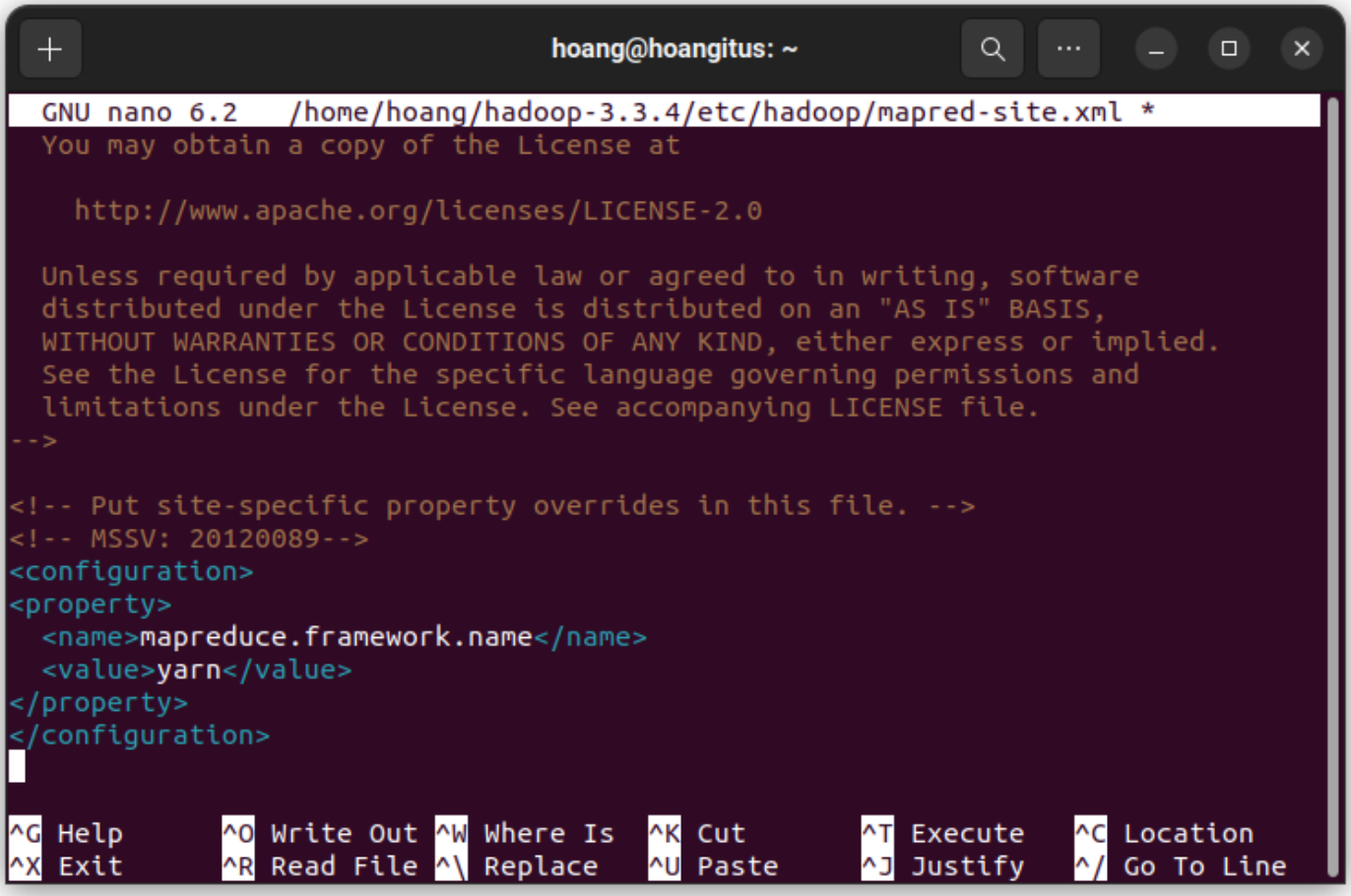

hoang@hoangitus: ~
20120089:~$ nano ~/hadoop-3.3.4/etc/hadoop/hadoop-env.sh

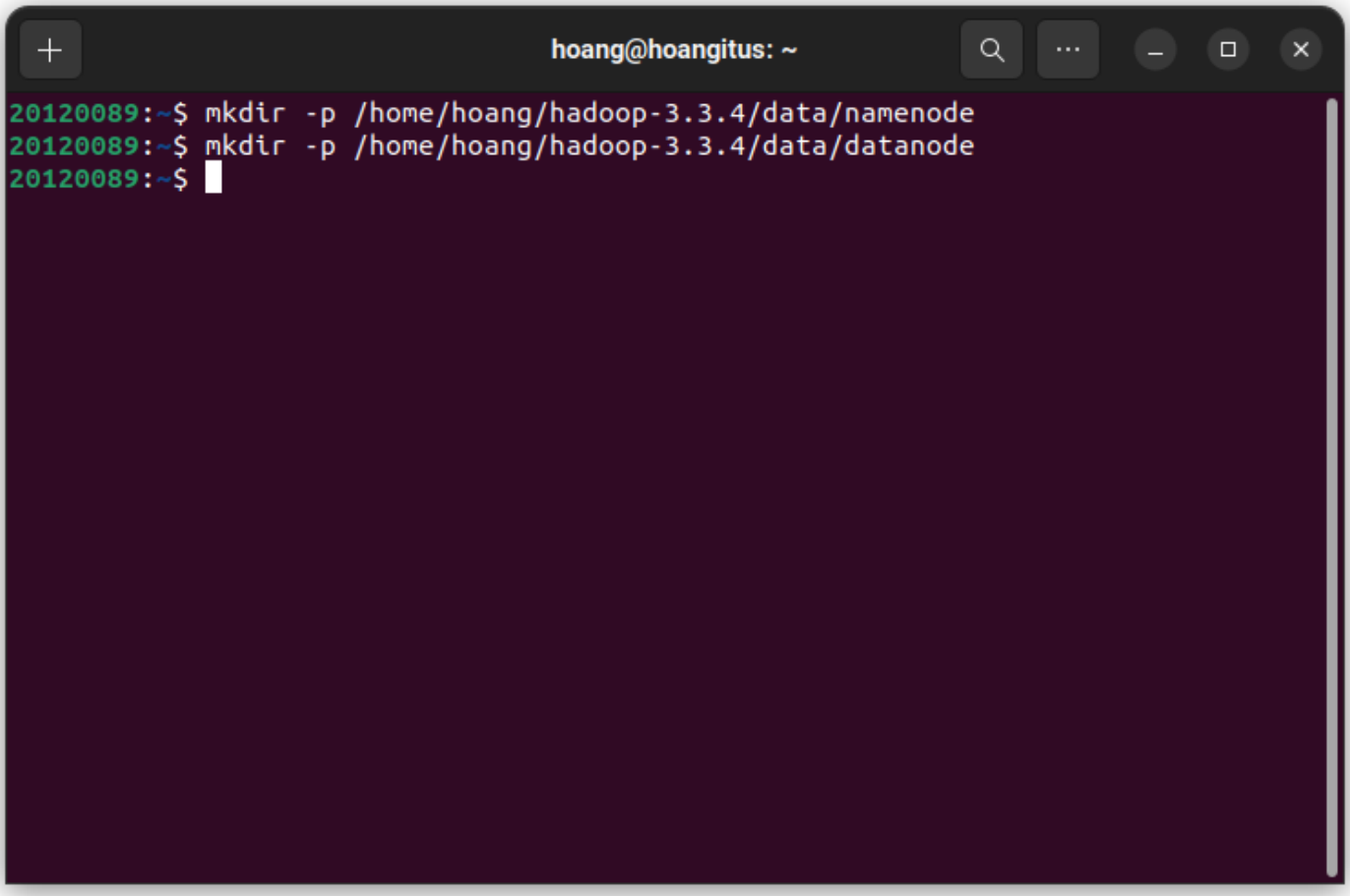

hoang@hoangitus: ~
GNU nano 6.2 /home/hoang/hadoop-3.3.4/etc/hadoop/hadoop-env.sh
CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.
##
Precedence rules:
##
{yarn-env.sh|hdfs-env.sh} > hadoop-env.sh > hard-coded defaults
##
{YARN_xyz|HDFS_xyz} > HADOOP_xyz > hard-coded defaults
##
Many of the options here 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
#MSSV: 20120089
#
[Wrote 435 lines]
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

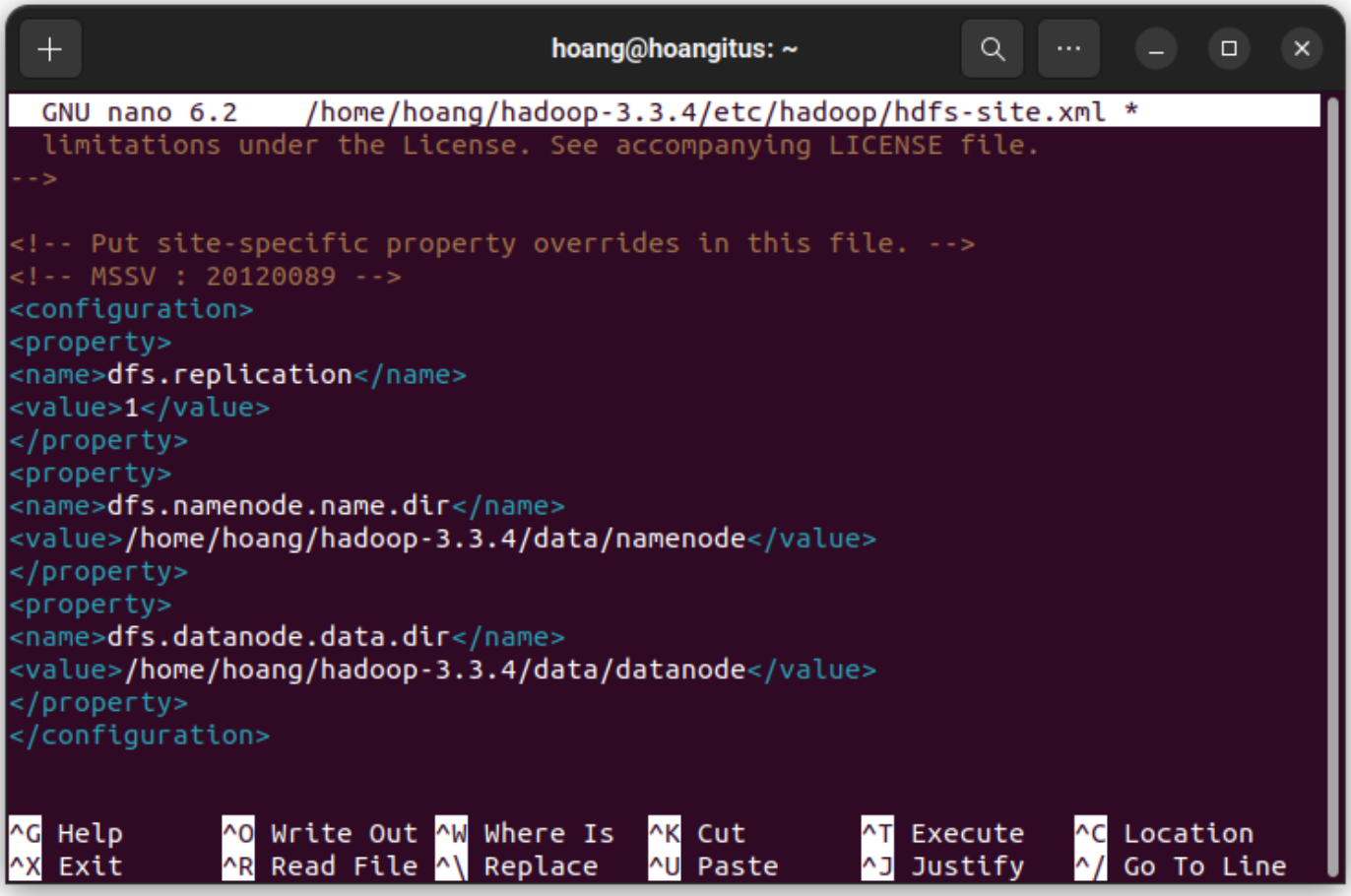

hoang@hoangitus: ~
20120089:~$ mkdir -p ~/hadoop-3.3.4/tmp
20120089:~$ chmod 777 ~/hadoop-3.3.4/tmp
20120089:~$

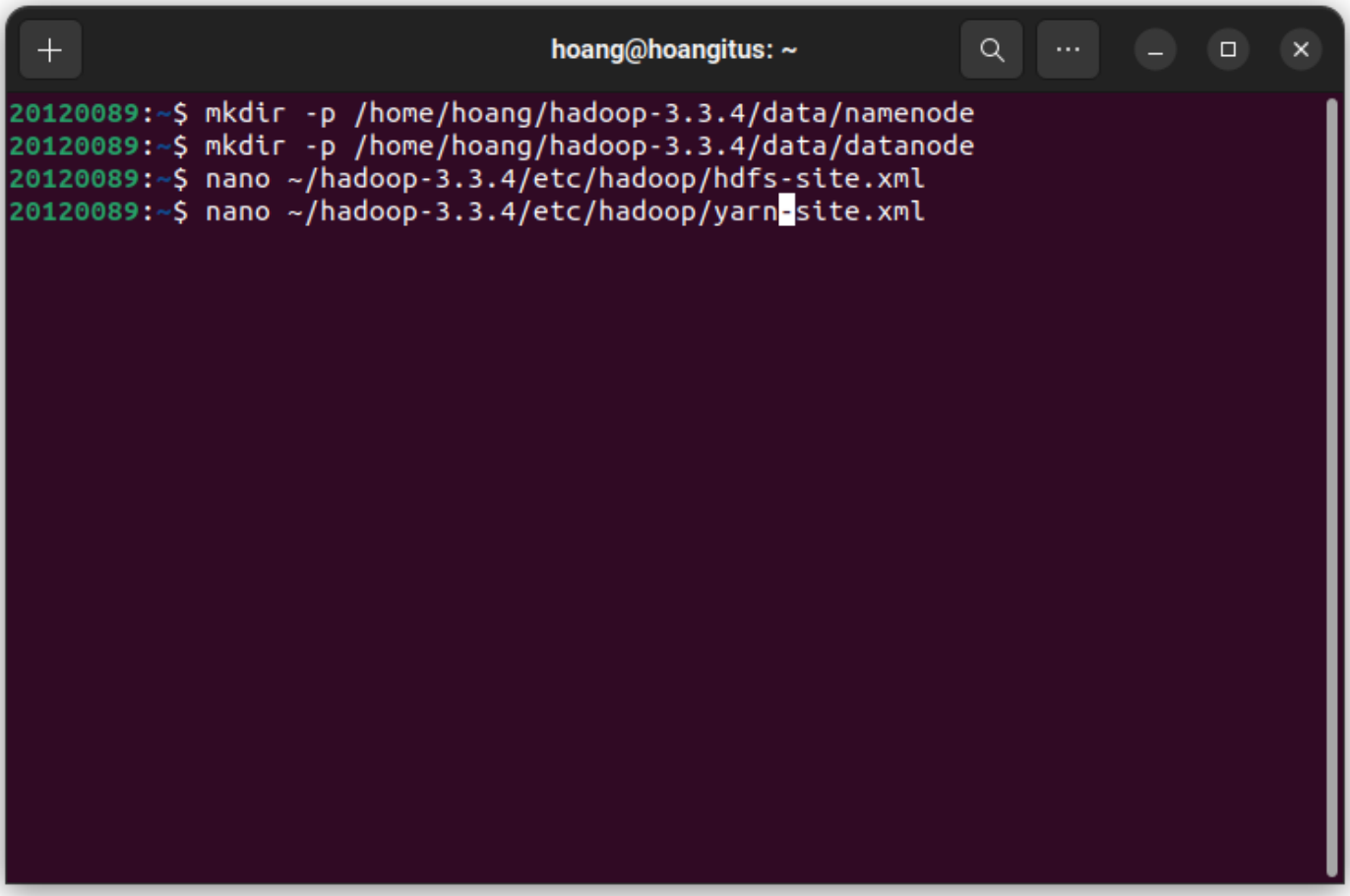

hoang@hoangitus: ~
20120089:~$ mkdir -p ~/hadoop-3.3.4/tmp
20120089:~$ chmod 777 ~/hadoop-3.3.4/tmp
20120089:~$ nano ~/hadoop-3.3.4/etc/hadoop/core-site.xml

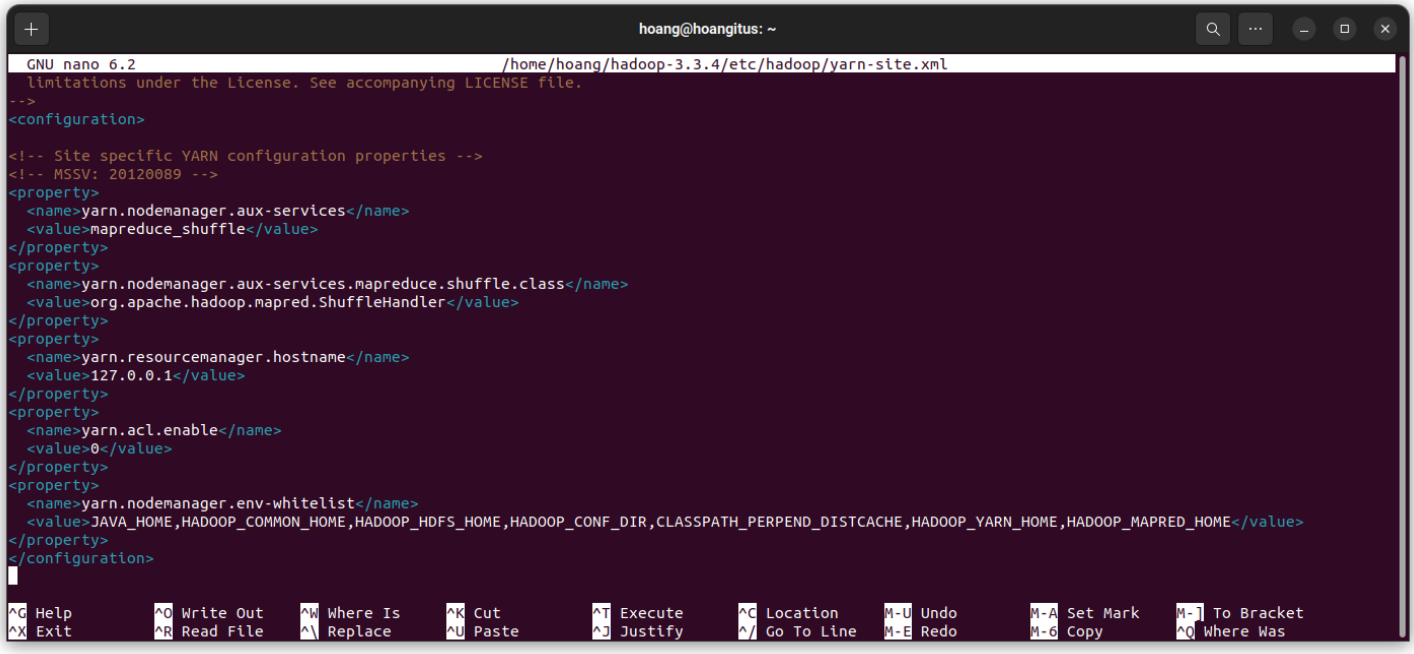

hoang@hoangitus: ~
GNU nano 6.2 /home/hoang/hadoop-3.3.4/etc/hadoop/core-site.xml *
Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<!--MSSV: 20120089 -->
<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/hoang/hadoop-3.3.4/tmp</value>
</property>
<property>
<name>fs.default.name</name >
<value>hdfs://localhost:9000</value>
</property>
</configuration>
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

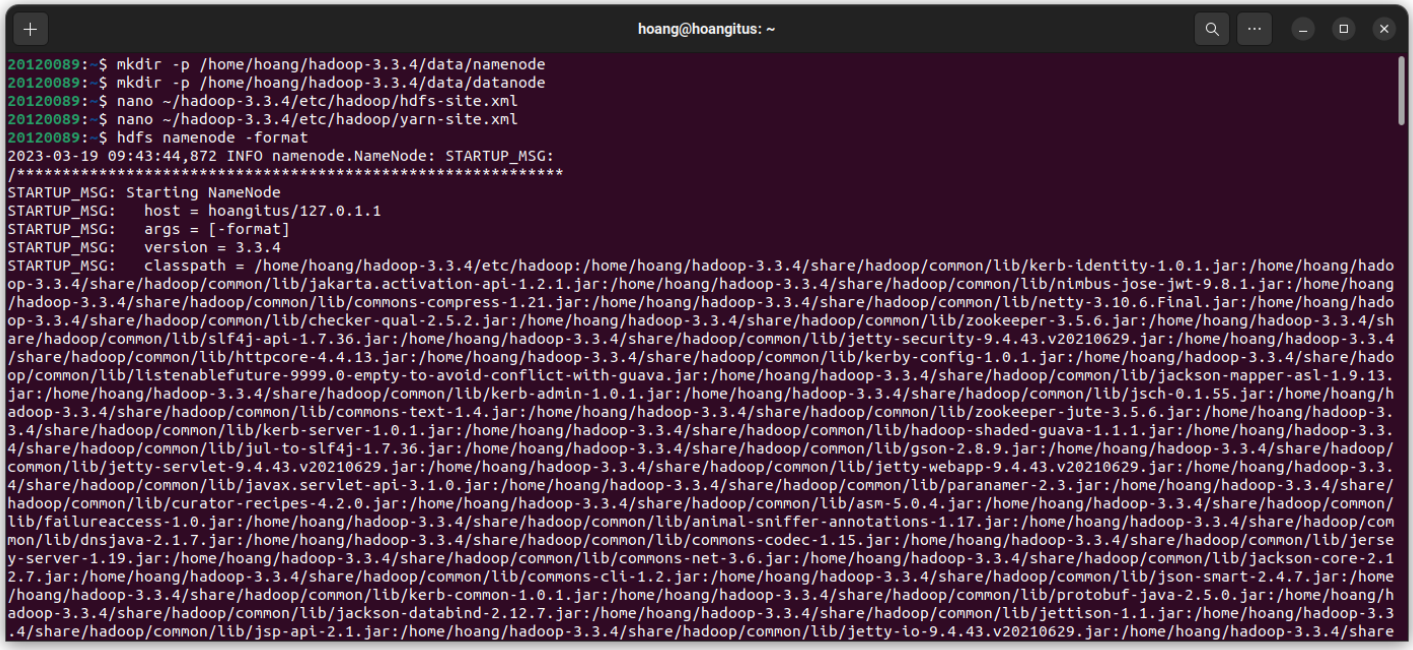

hoang@hoangitus: ~
GNU nano 6.2 /home/hoang/hadoop-3.3.4/etc/hadoop/mapred-site.xml *
You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0
Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
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. -->
<!-- MSSV: 20120089-->
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

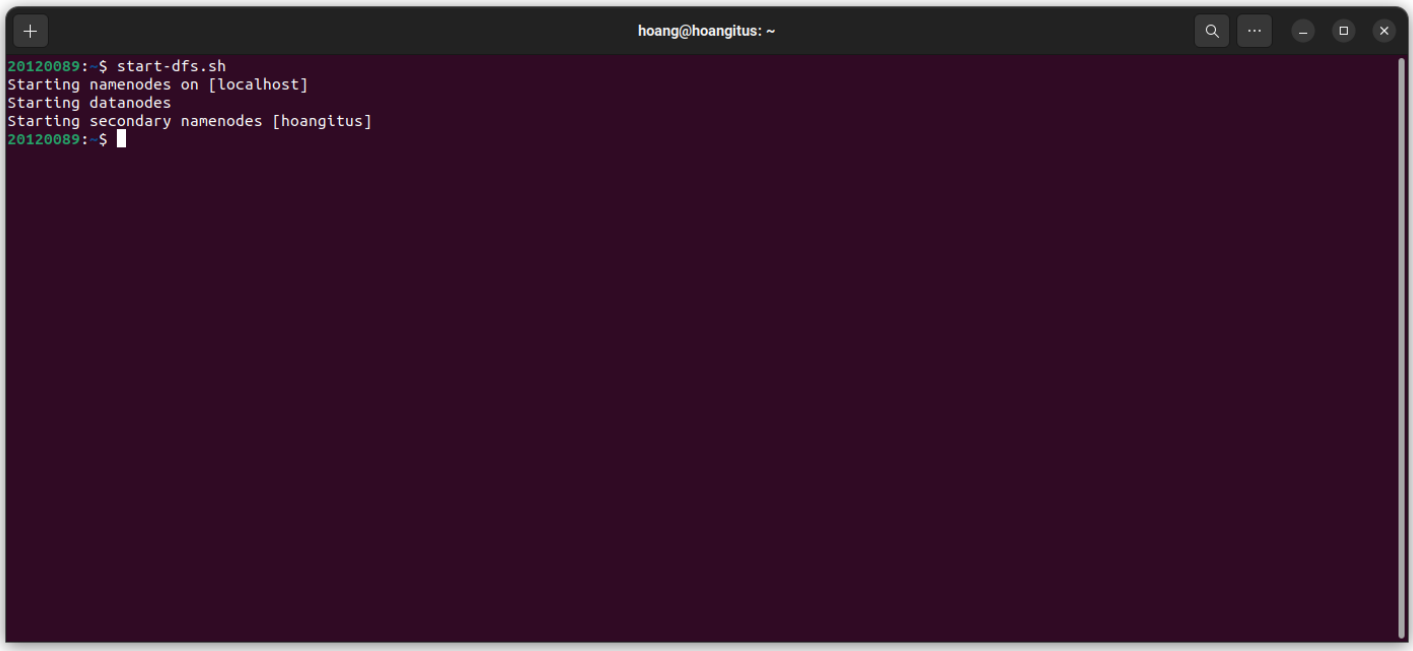

hoang@hoangitus: ~
20120089:~$ mkdir -p /home/hoang/hadoop-3.3.4/data/namenode
20120089:~$ mkdir -p /home/hoang/hadoop-3.3.4/data/datanode
20120089:~$

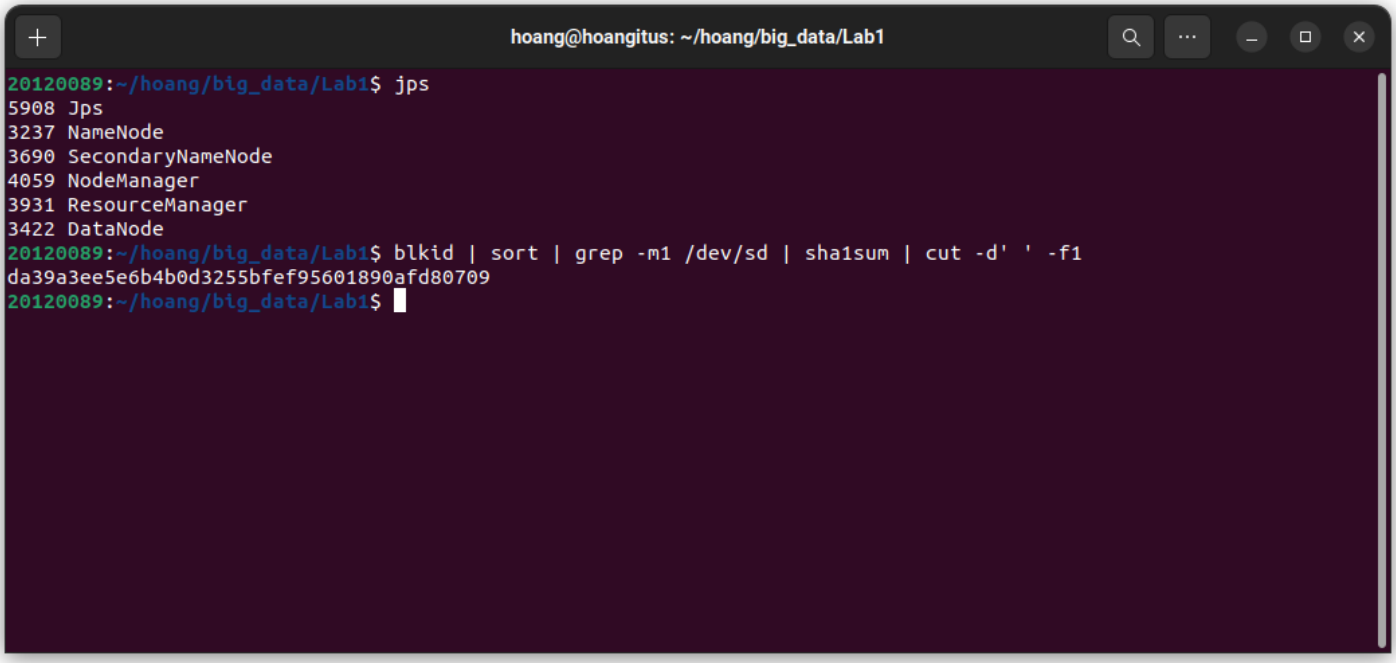

hoang@hoangitus: ~
GNU nano 6.2 /home/hoang/hadoop-3.3.4/etc/hadoop/hdfs-site.xml *
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->
<!-- MSSV : 20120089 -->
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>/home/hoang/hadoop-3.3.4/data/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/hoang/hadoop-3.3.4/data/datanode</value>
</property>
</configuration>
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

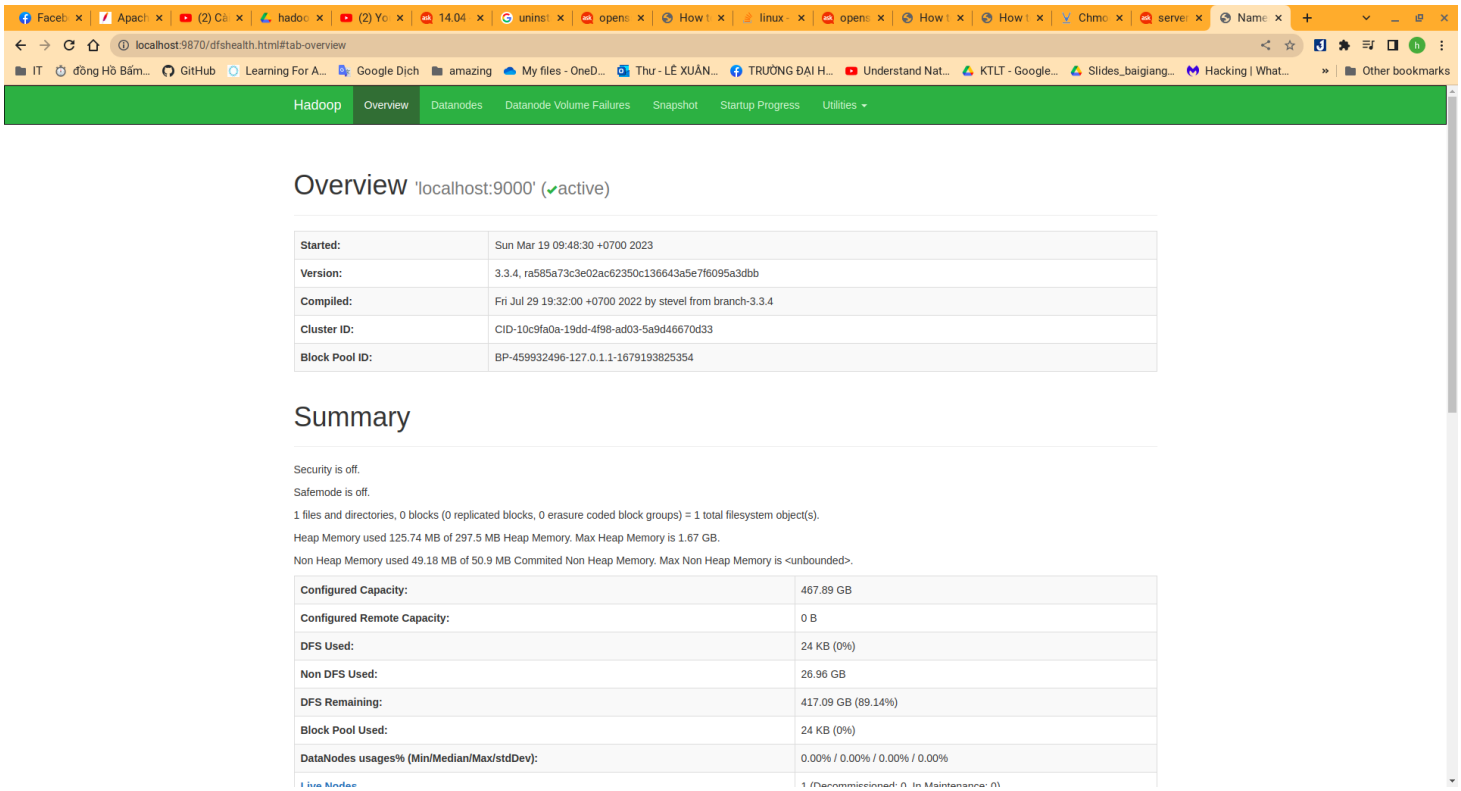

hoang@hoangitus: ~
20120089:~$ mkdir -p /home/hoang/hadoop-3.3.4/data/namenode
20120089:~$ mkdir -p /home/hoang/hadoop-3.3.4/data/datanode
20120089:~$ nano ~/hadoop-3.3.4/etc/hadoop/hdfs-site.xml
20120089:~$ nano ~/hadoop-3.3.4/etc/hadoop/yarn-site.xml

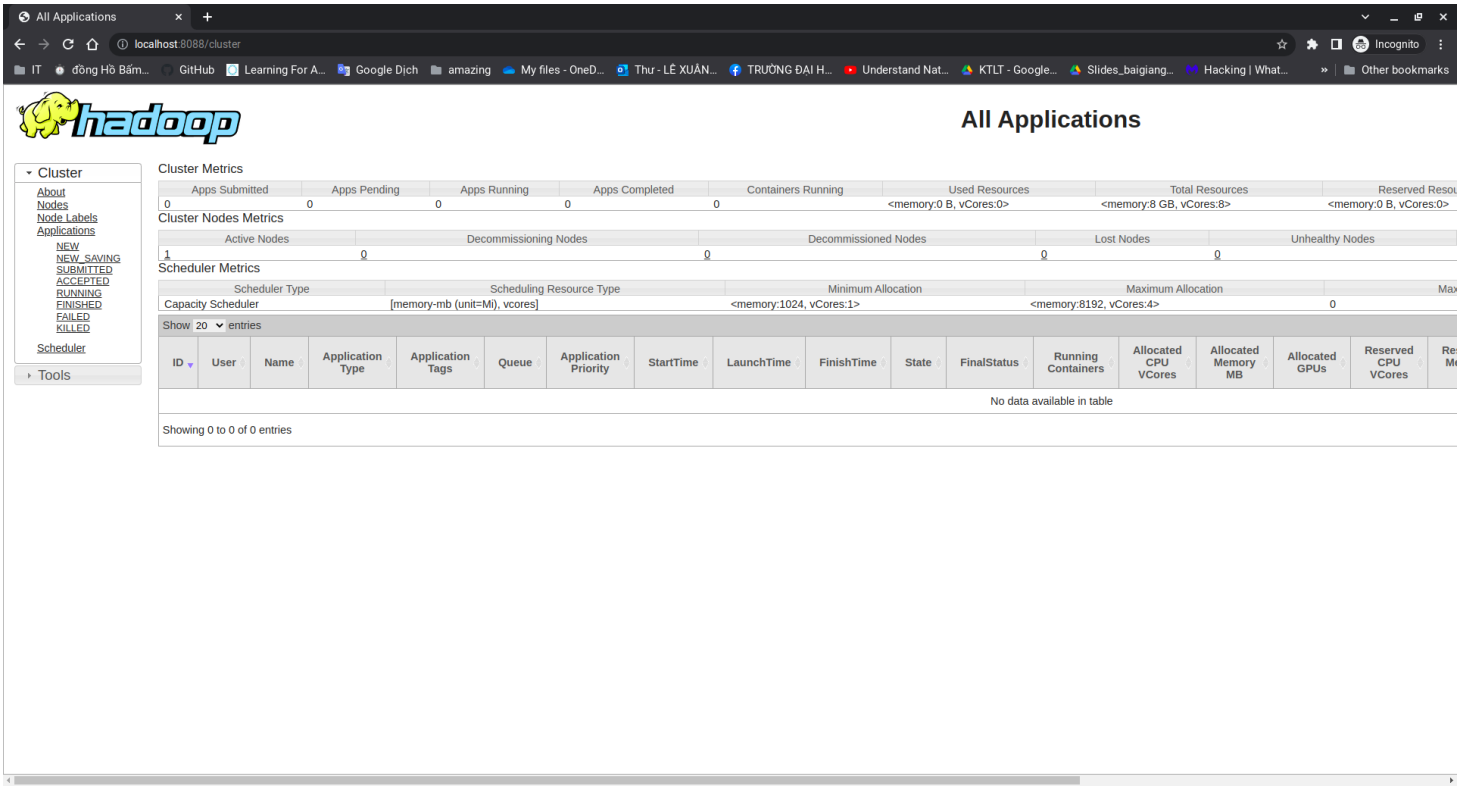

hoang@hoangitus: ~
GNU nano 6.2 /home/hoang/hadoop-3.3.4/etc/hadoop/yarn-site.xml
limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<!-- Site specific YARN configuration properties -->
<!-- MSSV: 20120089 -->
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</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_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>
^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location M-U Undo M-A Set Mark M-] To Bracket
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line M-E Redo M-6 Copy ^Q Where Was


hoang@hoangitus: ~
20120089:~$ mkdir -p /home/hoang/hadoop-3.3.4/data/namenode
20120089:~$ mkdir -p /home/hoang/hadoop-3.3.4/data/datanode
20120089:~$ nano ~/hadoop-3.3.4/etc/hadoop/hdfs-site.xml
20120089:~$ nano ~/hadoop-3.3.4/etc/hadoop/yarn-site.xml
20120089:~$ hdfs namenode -format
2023-03-19 09:43:44,872 INFO namenode.NameNode: STARTUP_MSG:
/**
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = hoangitus/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.3.4
STARTUP_MSG: classpath = /home/hoang/hadoop-3.3.4/etc/hadoop:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/kerb-identity-1.0.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jakarta.activation-api-1.2.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/nimbus-jose-jwt-9.8.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/commons-compress-1.21.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/netty-3.10.6.Final.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/checker-qual-2.5.2.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/zookeeper-3.5.6.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/slf4j-api-1.7.36.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jetty-security-9.4.43.v20210629.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/httpcore-4.4.13.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/kerby-config-1.0.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/listenablefuture-9999.0-empty-to-avoid-conflict-with-guava.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/kerb-admin-1.0.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jsch-0.1.55.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/commons-text-1.4.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/zookeeper-jute-3.5.6.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/kerb-server-1.0.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/hadoop-shaded-guava-1.1.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jul-to-slf4j-1.7.36.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/gson-2.8.9.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jetty-servlet-9.4.43.v20210629.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jetty-webapp-9.4.43.v20210629.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/javax.servlet-api-3.1.0.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/paranamer-2.3.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/curator-recipes-4.2.0.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/asm-5.0.4.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/failureaccess-1.0.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/animal-sniffer-annotations-1.17.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/dnsjava-2.1.7.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/commons-codec-1.15.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jersey-server-1.19.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/commons-net-3.6.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jackson-core-2.12.7.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/commons-cli-1.2.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/json-smart-2.4.7.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/kerb-common-1.0.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jackson-databind-2.12.7.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jettison-1.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jsp-api-2.1.jar:/home/hoang/hadoop-3.3.4/share/hadoop/common/lib/jetty-io-9.4.43.v20210629.jar:/home/hoang/hadoop-3.3.4/share

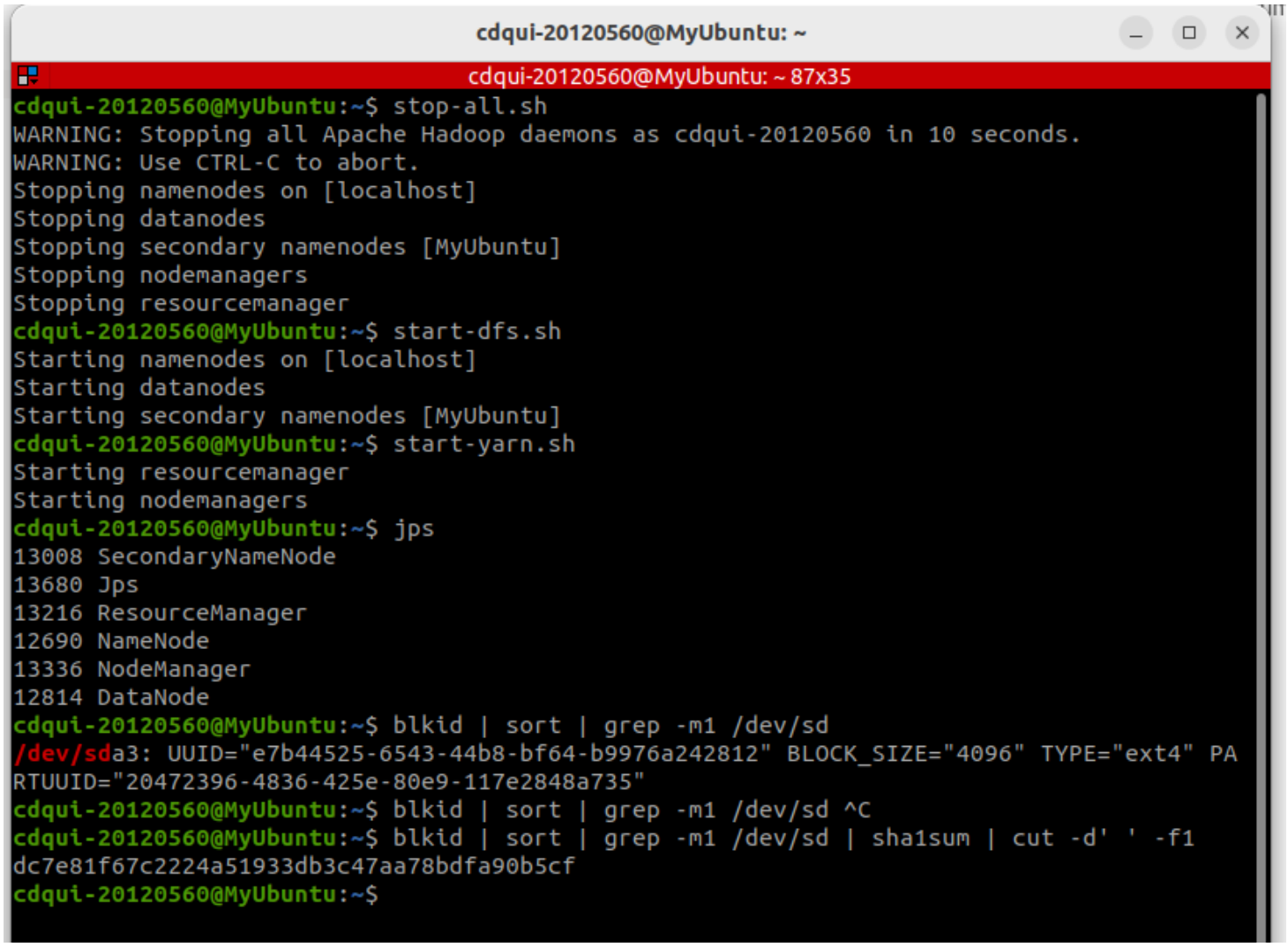

hoang@hoangitus: ~
20120089:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hoangitus]
20120089:~$

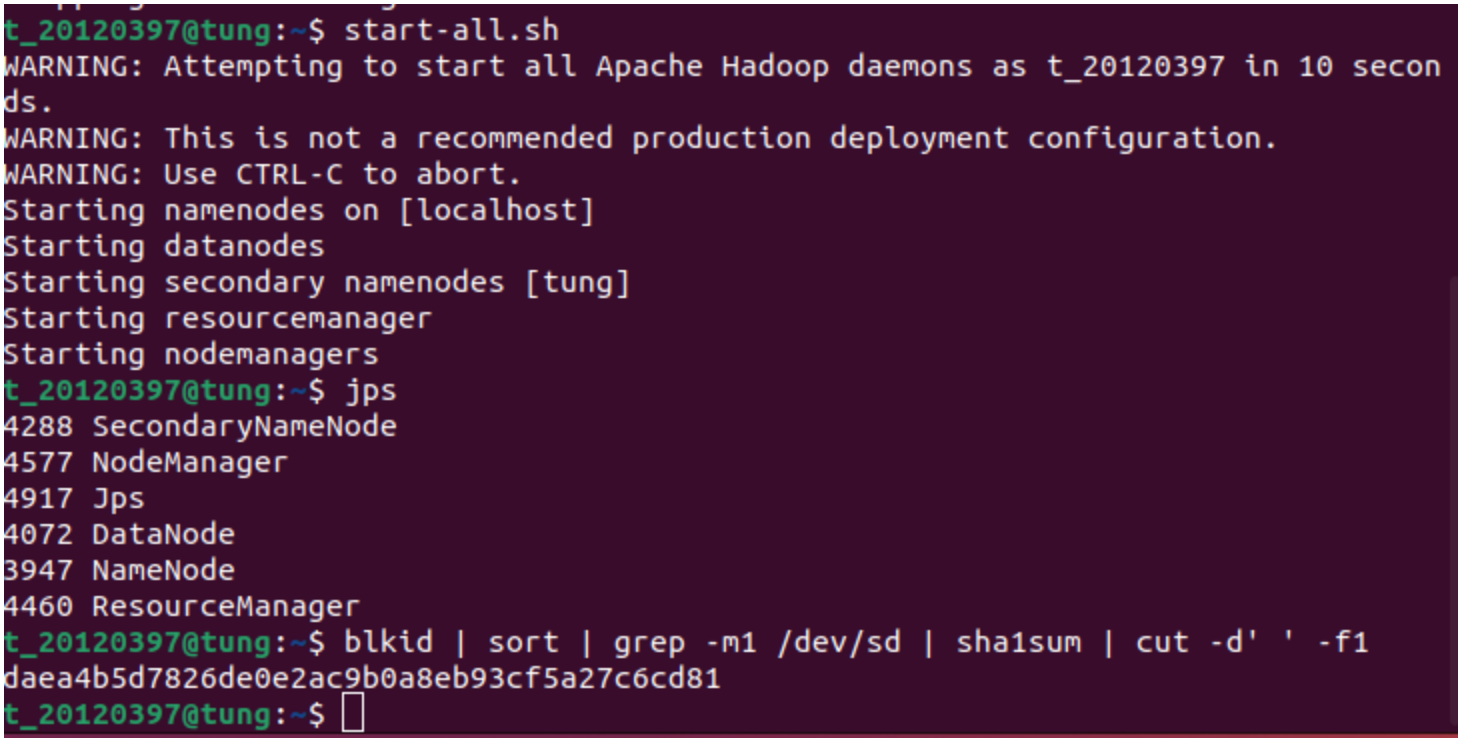

hoang@hoangitus: ~/hoang/big_data/Lab1
20120089:~/hoang/big_data/Lab1$ jps
5908 Jps
3237 NameNode
3690 SecondaryNameNode
4059 NodeManager
3931 ResourceManager
3422 DataNode
20120089:~/hoang/big_data/Lab1$ blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
20120089:~/hoang/big_data/Lab1$

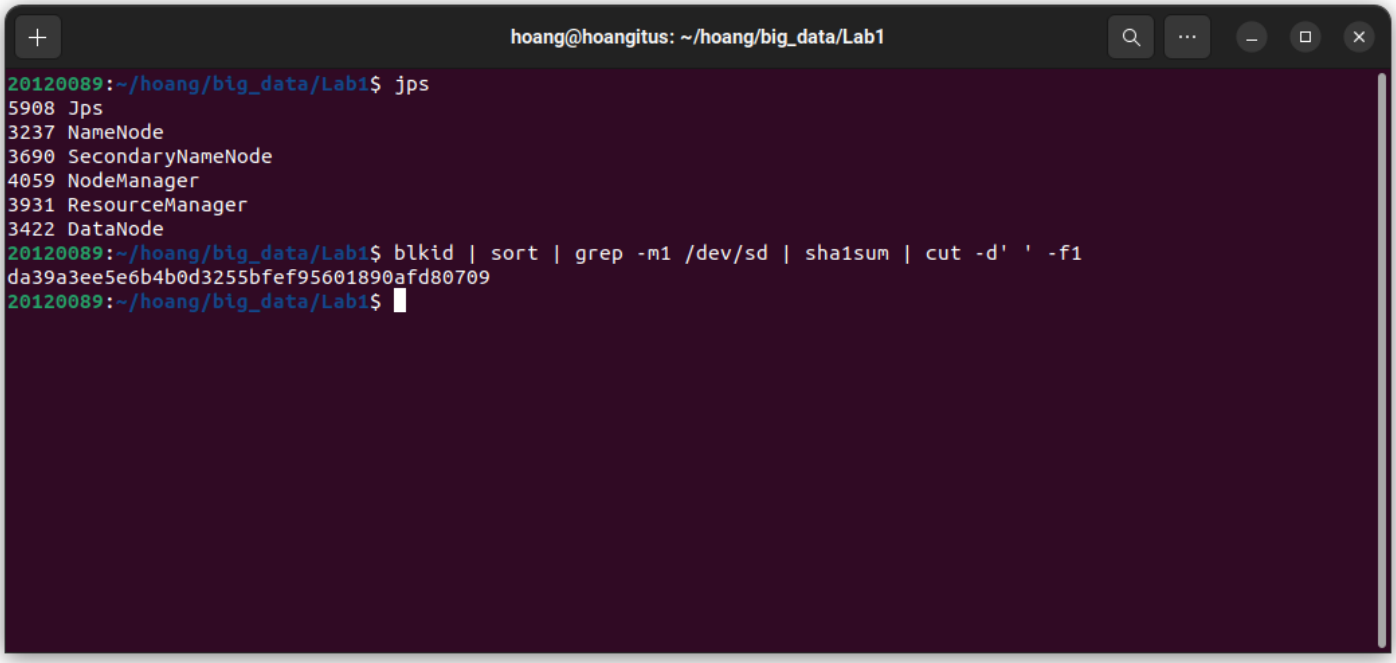

localhost:9870/dfshealth.html#tab-overview
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Overview 'localhost:9000' (active)
Started: Sun Mar 19 09:48:30 +0700 2023
Version: 3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled: Fri Jul 29 19:32:00 +0700 2022 by stevel from branch-3.3.4
Cluster ID: CID-10c9fa0a-19dd-4f98-ad03-5a9d46670d33
Block Pool ID: BP-459932496-127.0.1.1-1679193825354
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 125.74 MB of 297.5 MB Heap Memory. Max Heap Memory is 1.67 GB.
Non Heap Memory used 49.18 MB of 50.9 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.
Configured Capacity: 467.89 GB
Configured Remote Capacity: 0 B
DFS Used: 24 KB (0%)
Non DFS Used: 26.96 GB
DFS Remaining: 417.09 GB (89.14%)
Block Pool Used: 24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev): 0.00% / 0.00% / 0.00% / 0.00%

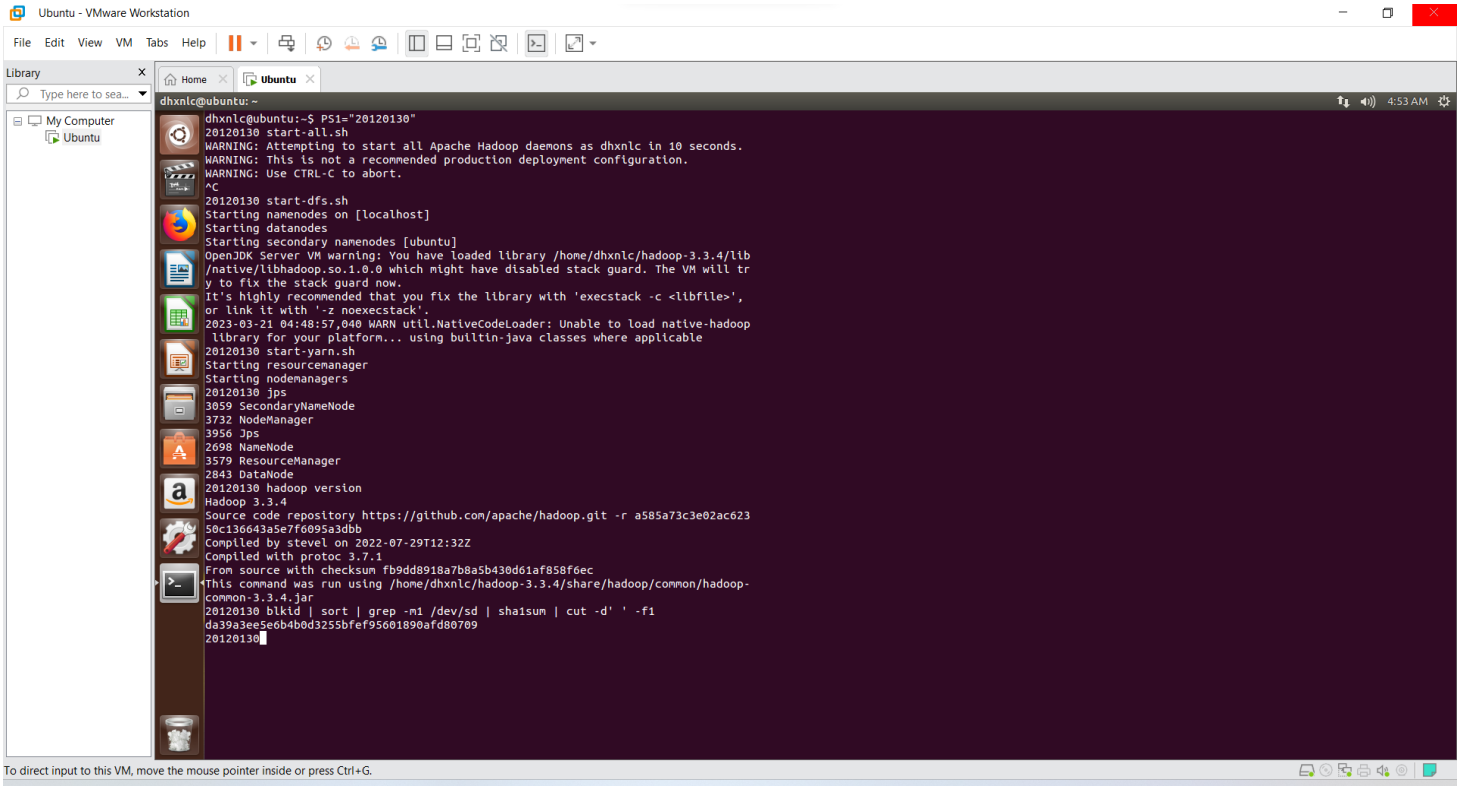

All Applications
localhost:8088/cluster
Incognito
IT
đồng Hồ Bấm...
GitHub
Learning For A...
Google Dịch
amazing
My files - OneD...
Thư - LÊ XUÂN...
TRƯỜNG ĐẠI H...
Understand Nat...
KTLT - Google...
Slides_baigiang...
Hacking | What...
Other bookmarks
hadoop
All Applications
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted
Apps Pending
Apps Running
Apps Completed
Containers Running
Used Resources
Total Resources
Reserved Resou
0
0
0
0
0
<memory:0 B, vCores:0>
<memory:8 GB, vCores:8>
<memory:0 B, vCores:0>
Cluster Nodes Metrics
Active Nodes
Decommissioning Nodes
Decommissioned Nodes
Lost Nodes
Unhealthy Nodes
1
0
0
0
0
Scheduler Metrics
Scheduler Type
Scheduling Resource Type
Minimum Allocation
Maximum Allocation
Max
Capacity Scheduler
[memory-mb (unit=Mi), vcores]
<memory:1024, vCores:1>
<memory:8192, vCores:4>
0
Show 20 entries
ID
User
Name
Application Type
Application Tags
Queue
Application Priority
StartTime
LaunchTime
FinishTime
State
FinalStatus
Running Containers
Allocated CPU VCores
Allocated Memory MB
Allocated GPUs
Reserved CPU VCores
No data available in table
Showing 0 to 0 of 0 entries

Result:


cdqui-20120560@MyUbuntu: ~
cdqui-20120560@MyUbuntu: ~ 87x35
cdqui-20120560@MyUbuntu:~$ stop-all.sh
WARNING: Stopping all Apache Hadoop daemons as cdqui-20120560 in 10 seconds.
WARNING: Use CTRL-C to abort.
Stopping namenodes on [localhost]
Stopping datanodes
Stopping secondary namenodes [MyUbuntu]
Stopping nodemanagers
Stopping resourcemanager
cdqui-20120560@MyUbuntu:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [MyUbuntu]
cdqui-20120560@MyUbuntu:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
cdqui-20120560@MyUbuntu:~$ jps
13008 SecondaryNameNode
13680 Jps
13216 ResourceManager
12690 NameNode
13336 NodeManager
12814 DataNode
cdqui-20120560@MyUbuntu:~$ blkid | sort | grep -m1 /dev/sd
/dev/sda3: UUID="e7b44525-6543-44b8-bf64-b9976a242812" BLOCK_SIZE="4096" TYPE="ext4" PA
RTUUID="20472396-4836-425e-80e9-117e2848a735"
cdqui-20120560@MyUbuntu:~$ blkid | sort | grep -m1 /dev/sd ^C
cdqui-20120560@MyUbuntu:~$ blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
dc7e81f67c2224a51933db3c47aa78bdfa90b5cf
cdqui-20120560@MyUbuntu:~$


t_20120397@tung:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as t_20120397 in 10 secon
ds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [tung]
Starting resourcemanager
Starting nodemanagers
t_20120397@tung:~$ jps
4288 SecondaryNameNode
4577 NodeManager
4917 Jps
4072 DataNode
3947 NameNode
4460 ResourceManager
t_20120397@tung:~$ blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
daea4b5d7826de0e2ac9b0a8eb93cf5a27c6cd81
t_20120397@tung:~$


hoang@hoangitus: ~/hoang/big_data/Lab1
20120089:~/hoang/big_data/Lab1$ jps
5908 Jps
3237 NameNode
3690 SecondaryNameNode
4059 NodeManager
3931 ResourceManager
3422 DataNode
20120089:~/hoang/big_data/Lab1$ blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
20120089:~/hoang/big_data/Lab1$


Ubuntu - VMware Workstation
File Edit View VM Tabs Help
Library
Type here to sea...
My Computer
Ubuntu
Home
Ubuntu
dhxnlc@ubuntu: ~
4:53 AM
dhxnlc@ubuntu:~$ PS1="20120130"
20120130 start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as dhxnlc in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
^C
20120130 start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ubuntu]
OpenJDK Server VM warning: You have loaded library /home/dhxnlc/hadoop-3.3.4/lib
/native/libhadoop.so.1.0.0 which might have disabled stack guard. The VM will tr
y to fix the stack guard now.
It's highly recommended that you fix the library with 'execstack -c <libfile>',
or link it with '-z noexecstack'.
2023-03-21 04:48:57,040 WARN util.NativeCodeLoader: Unable to load native-hadoop
library for your platform... using builtin-java classes where applicable
20120130 start-yarn.sh
Starting resourcemanager
Starting nodemanagers
20120130 jps
3059 SecondaryNameNode
3732 NodeManager
3956 Jps
2698 NameNode
3579 ResourceManager
2843 DataNode
20120130 hadoop version
Hadoop 3.3.4
Source code repository https://github.com/apache/hadoop.git -r a585a73c3e02ac623
50c136643a5e7f6095a3dbb
Compiled by stevel on 2022-07-29T12:32Z
Compiled with protoc 3.7.1
From source with checksum fb9dd8918a7b8a5b430d61af858f6ec
This command was run using /home/dhxnlc/hadoop-3.3.4/share/hadoop/common/hadoop-
common-3.3.4.jar
20120130 blkid | sort | grep -m1 /dev/sd | sha1sum | cut -d' ' -f1
da39a3ee5e6b4b0d3255bfef95601890afd80709
20120130
To direct input to this VM, move the mouse pointer inside or press Ctrl+G.

2. Introduction to MapReduce

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

- Input keys-values: The initial data is divided into multiple input keys-values to be fed into MapReduce for processing.
- Intermediate keys-values: Generated from the input keys-values by the Map function. Its key is the result of the Map function's processing, and its value is information to be used in the Reduce function.
- Output keys-values: Generated from intermediate keys-values. The intermediate keys-values are sorted by key and partitioned across reducers. The reducers perform the Reduce function on the groups of intermediate keys-values and generate output keys-values.
- It can be said that input keys-values, intermediate keys-values, and output keys-values are interrelated as the output of one function serves as the input for the next function in the MapReduce process.

2.2 How does MapReduce deal with node failures?

- Redundant storage: MapReduce replicates data across multiple nodes in the cluster to ensure that if one node fails, the data can still be accessed and processed.
- Task tracking: MapReduce tracks completed tasks and tasks that are currently running. If a node fails while running a task, the task can be automatically restarted on another node.
- Job checkpointing: MapReduce periodically stores the intermediate output of a job on disk. If a node fails, the job can be restarted from the last checkpoint instead of starting from scratch, reducing processing time.
- Node monitoring: MapReduce continuously monitors the health of nodes in the cluster through heartbeat. If a node becomes unresponsive or fails, MapReduce can automatically remove it from the cluster and redistribute its tasks to other nodes.

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

- The meaning:
  - In MapReduce Hadoop, locality refers to processing data at or near its physical storage location. Locality is an important aspect of the Hadoop MapReduce framework, and it relates to the principle of processing data at or near its physical storage location. By prioritizing locality, Hadoop can reduce network load and improve system performance.
- Use case:

- Locality aims to reduce the amount of data that needs to be transferred over the network, reduce network load, and improve system performance. Hadoop achieves locality by attempting to schedule tasks on nodes where their input data is stored, which is known as data locality. This is made possible by Hadoop storing data in a distributed manner across a cluster of standard hardware nodes, with each node responsible for processing a portion of the data.
- When a task is scheduled on a node, the MapReduce framework attempts to read data from the local disk of that node first. Only when the data is not available locally, it is accessed from a remote node. By prioritizing data locality, Hadoop can significantly reduce the amount of data transmitted over the network, which is important for efficiently processing large data sets.

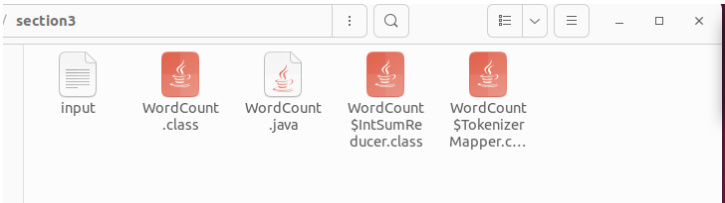
2.4. . Which problem is addressed by introducing a combiner function to the MapReduce model?

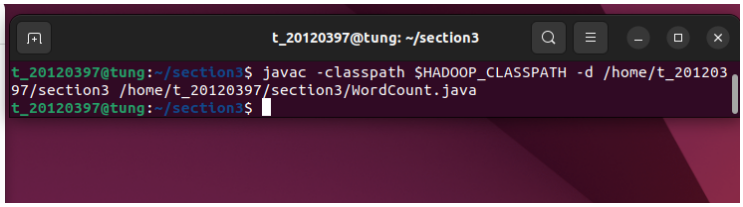
- In some cases, the Map tasks return multiple instances of the same <key,value> pair. The Combiner function summarizes these instances into a single <key,value> pair then transfers the result to the Reduce tasks, thus reducing the workload on them and speeding up the whole operation.

3. Running a warm-up problem: Word Count

The process and result:

```
t_20120397@tung:~/section3$ export HADOOP_CLASSPATH=${HADOOP_HOME}/share/hadoop/
common/hadoop-common-3.3.4.jar:${HADOOP_HOME}/share/hadoop/mapreduce/hadoop-mapr
educe-client-core-3.3.4.jar
t_20120397@tung:~/section3$ echo $HADOOP_CLASSPATH
/home/t_20120397/hadoop/share/hadoop/common/hadoop-common-3.3.4.jar:/home/t_2012
0397/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-client-core-3.3.4.jar
```





```
t_20120397@tung:~/section3$ jps
15587 Jps
12536 ResourceManager
12121 DataNode
11995 NameNode
12315 SecondaryNameNode
12655 NodeManager
```

```
t_20120397@tung:~/section3$ hadoop fs -ls /output_l1_s3
Found 2 items
-rw-r--r--   1 t_20120397 supergroup          0 2023-03-20 12:00 /output_l1_s3/_SUCCESS
-rw-r--r--   1 t_20120397 supergroup       1516 2023-03-20 12:00 /output_l1_s3/part-r-00000
```

Browse Directory

/input_l1_s3 Go!

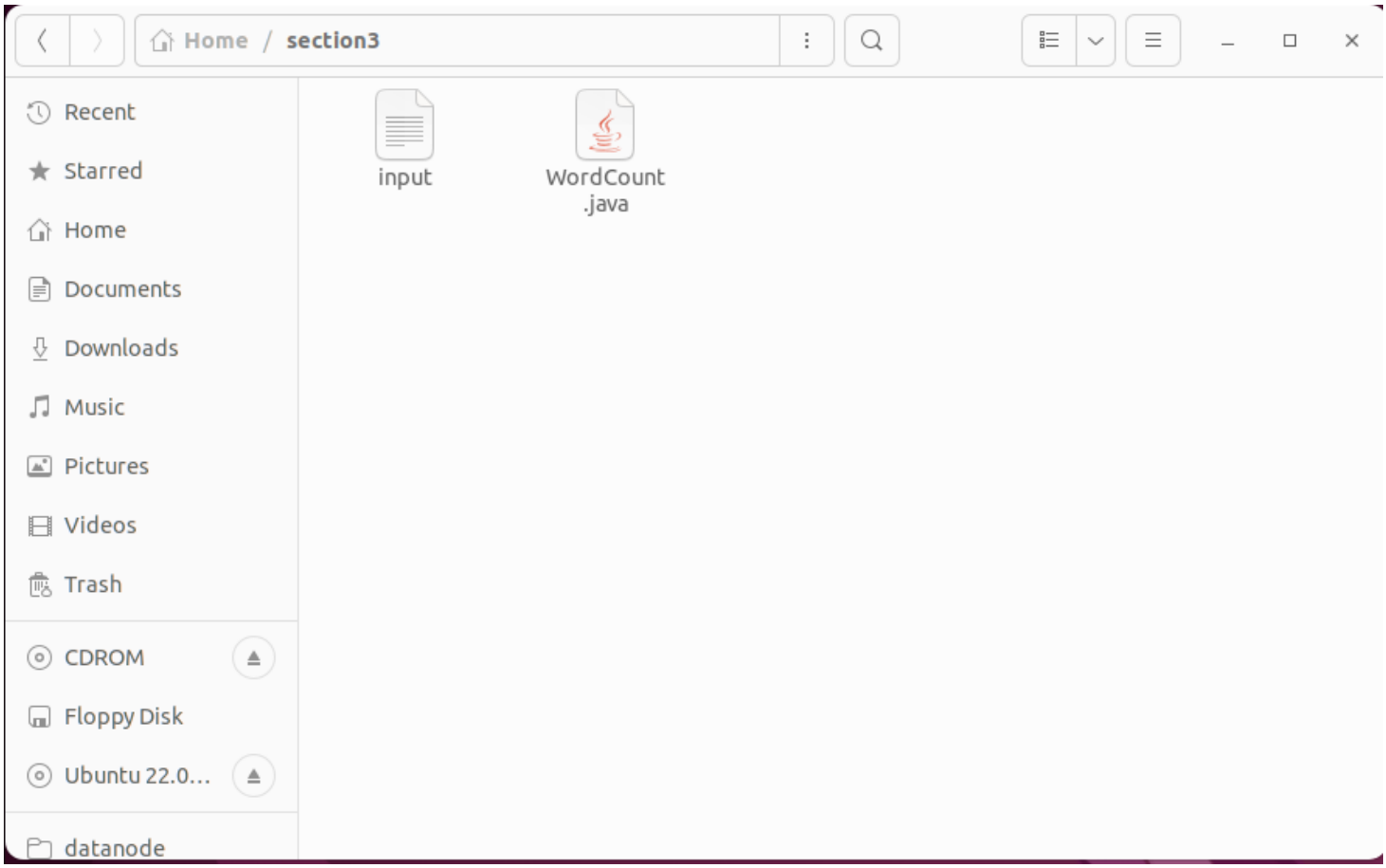
Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	t_20120397	supergroup	1.71 KB	Mar 20 11:57	1	128 MB	input.txt

Showing 1 to 1 of 1 entries Previous 1 Next

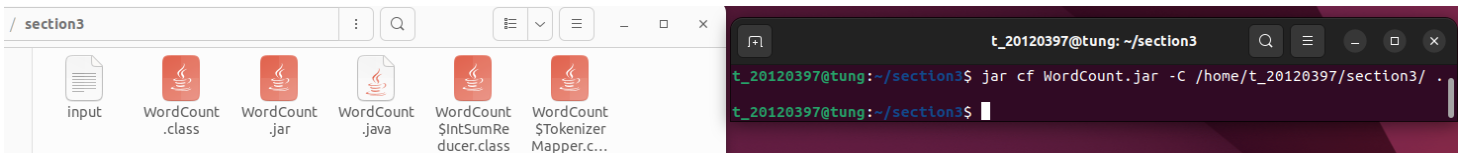
Hadoop, 2022.

```
t_20120397@tung: ~/section3
copyFromLocal: `/home/t_20120397/section3/input.txt': No such file or directory
t_20120397@tung:~/section3$ hadoop fs -copyFromLocal /home/t_20120397/section3/i
nput.txt /input_l1_s3/input.txt
t_20120397@tung:~/section3$
```


Home / section3
Recent
Starred
Home
Documents
Downloads
Music
Pictures
Videos
Trash
CDROM
Floppy Disk
Ubuntu 22.0…
datanode
input
WordCount
.java

```
t_20120397@tung:~/section3$ hadoop fs -cat /output_l1_s3/part-r-00000
(jar/executable 1
(multi-terabyte 1
(see    2
(thousands      1
A       1
Architecture    2
Distributed     1
File    1
Guide)  1
Guide). 1
HDFS    1
Hadoop  3
MRAppMaster     1
MapReduce       4
Minimally,      1
NodeManager     1
ResourceManager 1
ResourceManager,        1
System  1
The     4
These,  1
This    1
Typically       2
YARN    1
a       5
abstract-classes.       1
across  1
aggregate       1
allows  1
already 1
amounts 1
and     11
and/or  1
application     1
applications    2
appropriate     1
are     5
assumes 1
bandwidth       1
both    1
by      1
```

```
t_20120397@tung:~/section3$ hadoop jar WordCount.jar WordCount /input_l1_s3/input.txt /output_l1_s3
2023-03-20 11:59:32,277 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /127.0.0.1:
8032
2023-03-20 11:59:33,301 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Imple
ent the Tool interface and execute your application with ToolRunner to remedy this.
2023-03-20 11:59:33,337 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/stag
ing/t_20120397/.staging/job_1679283655462_0004
2023-03-20 11:59:33,854 INFO input.FileInputFormat: Total input files to process : 1
2023-03-20 11:59:34,097 INFO mapreduce.JobSubmitter: number of splits:1
2023-03-20 11:59:34,572 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1679283655462_0004
```



4. Bonus

4.1. Extended Word Count: Unhealthy relationships

- Input:
  A D
  A B
  B C
  D B
  B E
  E C
- Output:
  A pos
  B eq
  C neg
  D eq
  E eq

4.2. Setting up Fully Distributed Mode

References