Lab 01: A Gentle Introduction to Hadoop

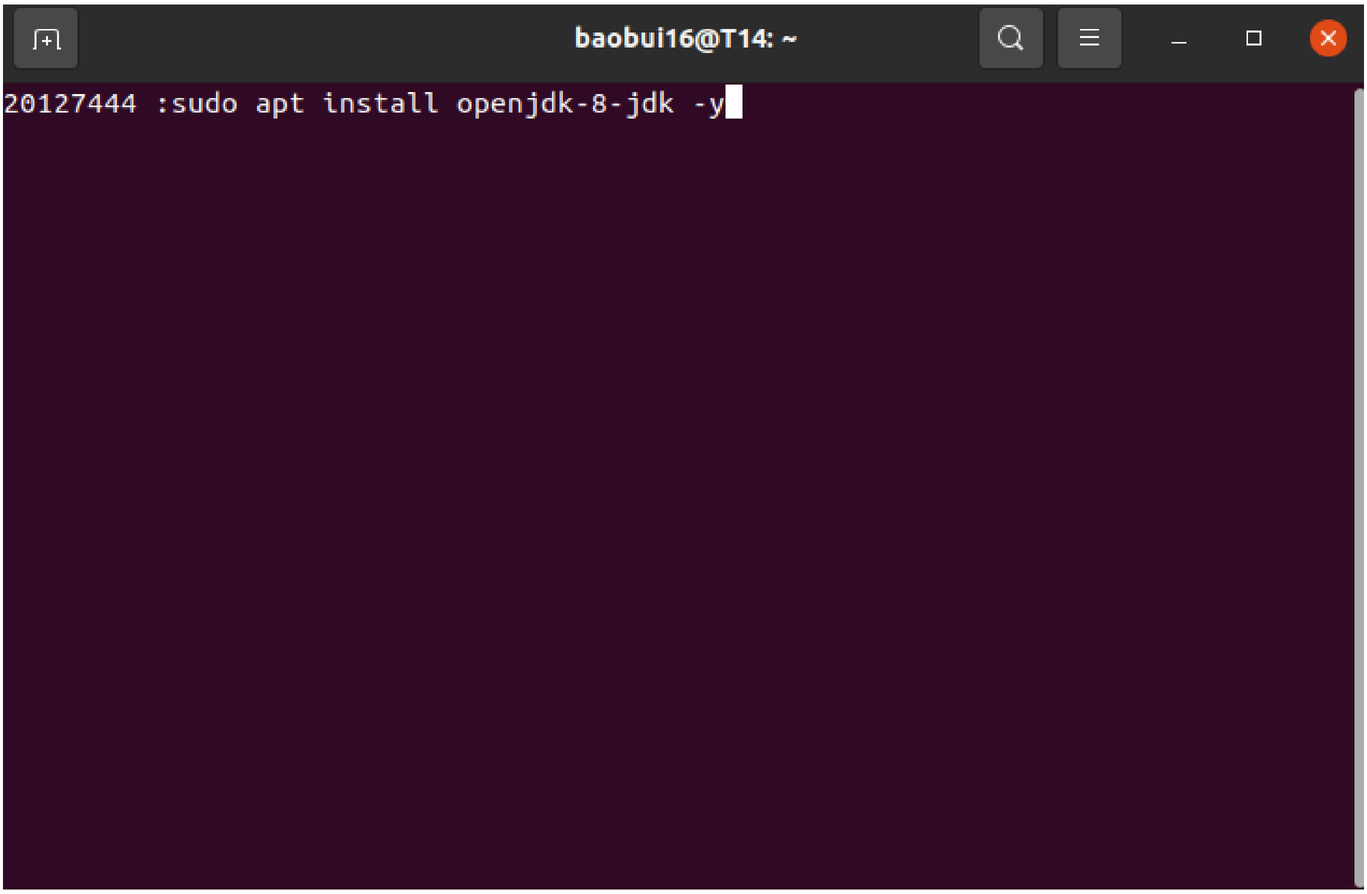
How much work, in percent (%),have you finished in each section?

My team works in section

```
    |  Section 1 |  Section 2 | Section 3  |  Section 4 |
    |------------|------------|------------|------------|
    |     100%   |     100%   |    100%    |     25%    |


    (In section 4 we just done section 4.1 and we can't do section 4.2)
```

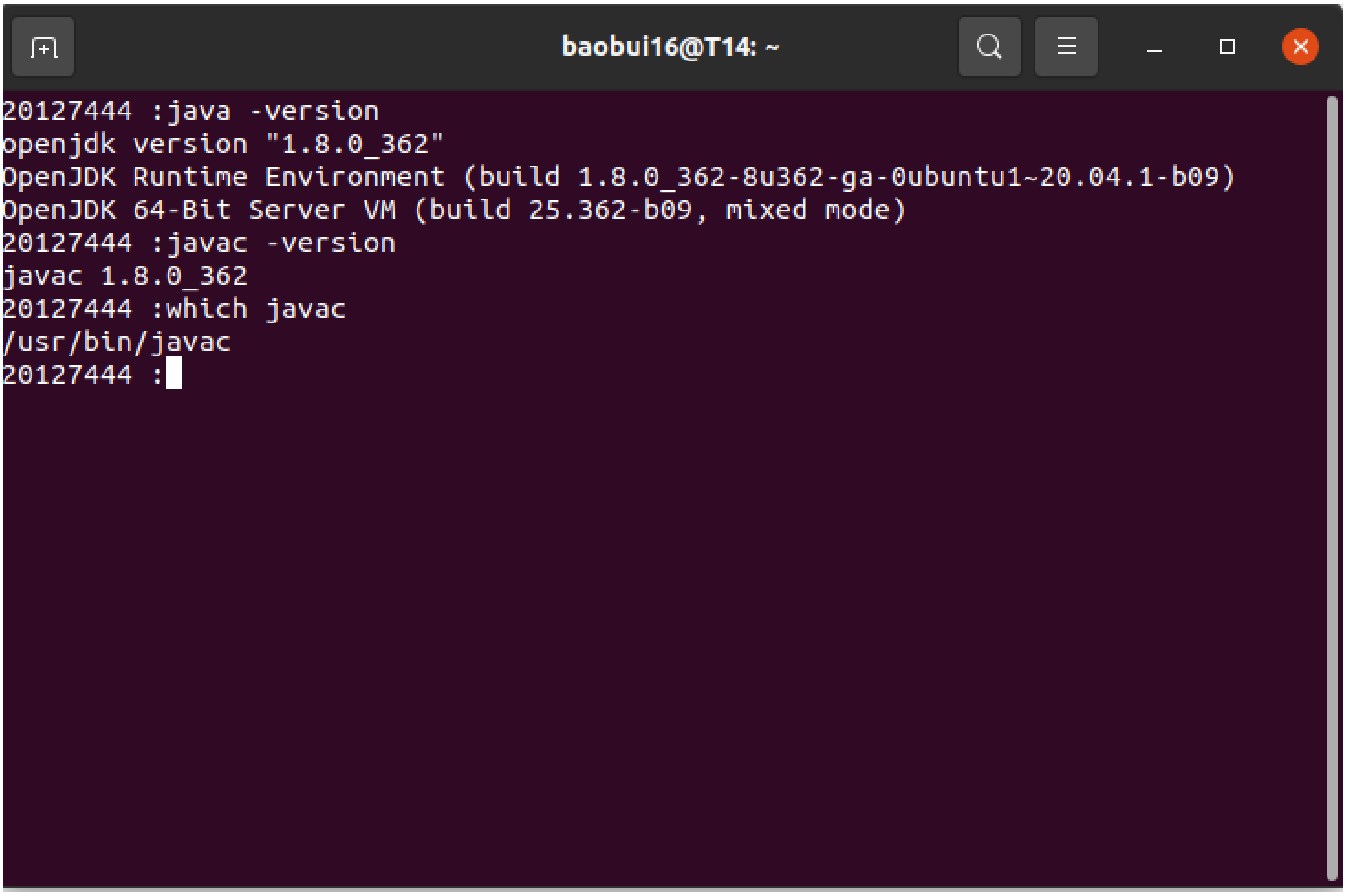
Setting up Single-node Hadoop Cluster

- Install java



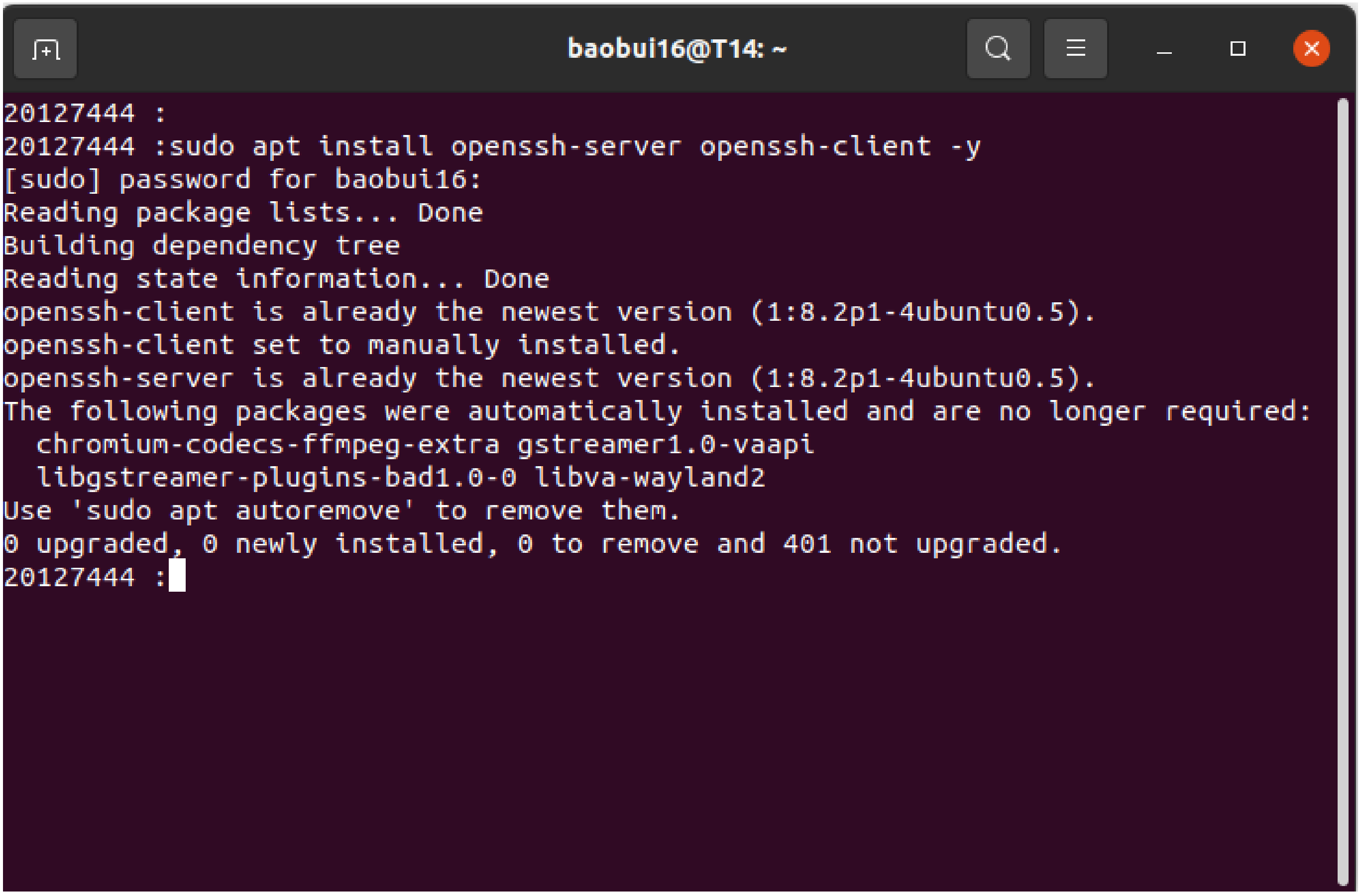
```
20127444 :sudo apt install openjdk-8-jdk -y
```

- Check java settings and path



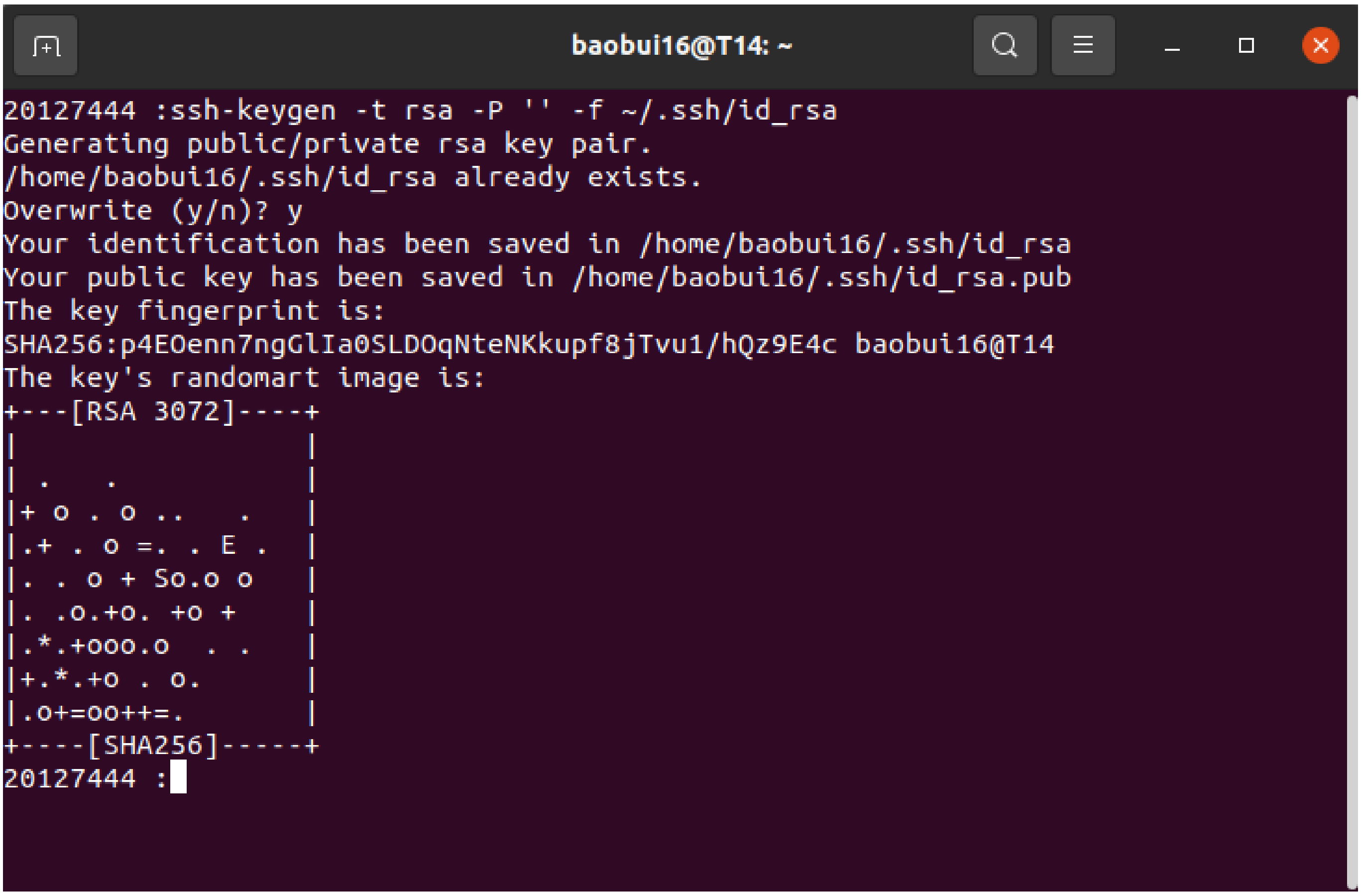
```
20127444 :java -version
openjdk version "1.8.0_362"
OpenJDK Runtime Environment (build 1.8.0_362-8u362-ga-0ubuntu1~20.04.1-b09)
OpenJDK 64-Bit Server VM (build 25.362-b09, mixed mode)
20127444 :javac -version
javac 1.8.0_362
20127444 :which javac
/usr/bin/javac
20127444 :
```

- Install openSSH

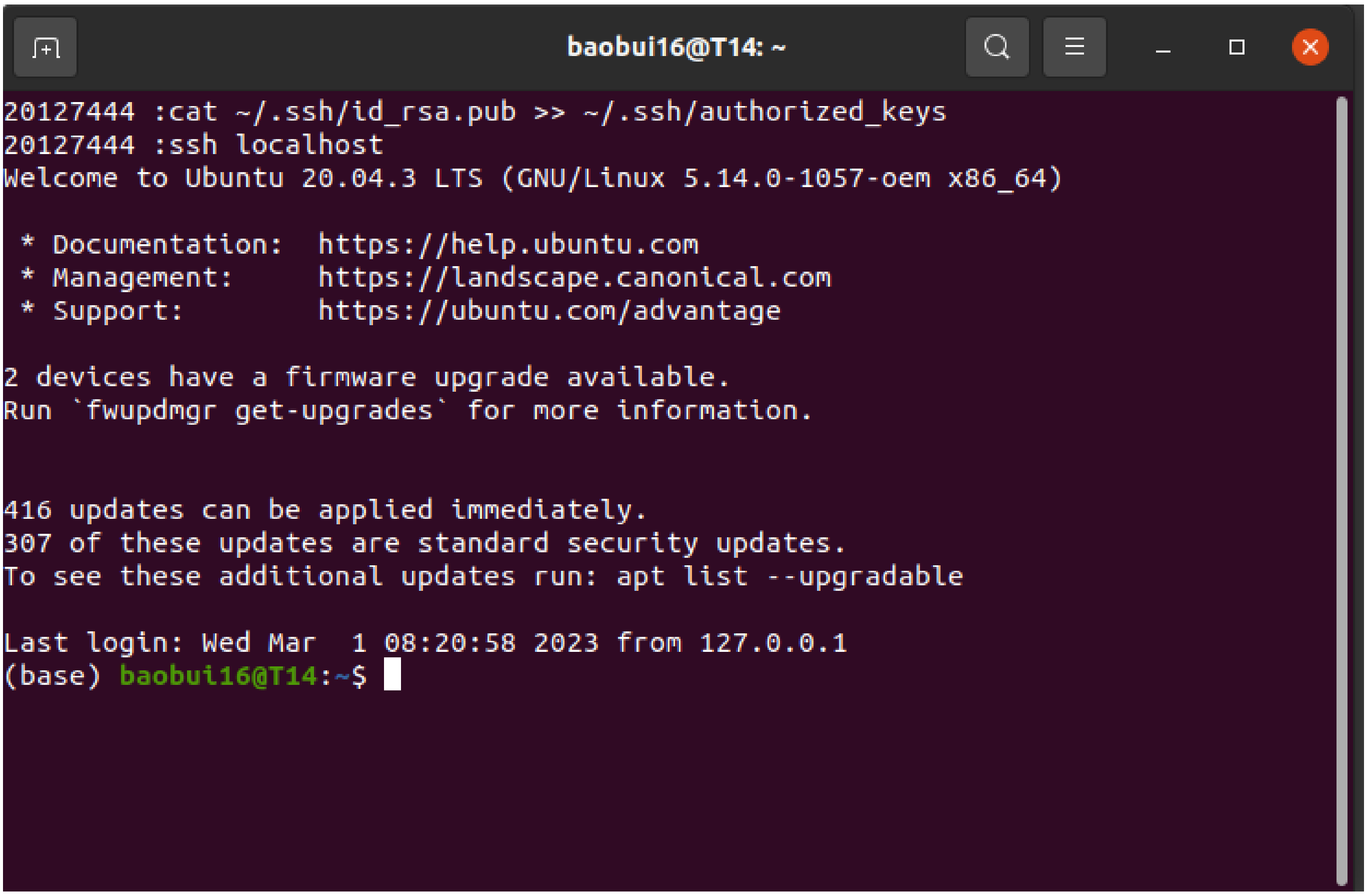


```
20127444 :
20127444 :sudo apt install openssh-server openssh-client -y
[sudo] password for baobui16:
Reading package lists... Done
Building dependency tree
Reading state information... Done
openssh-client is already the newest version (1:8.2p1-4ubuntu0.5).
openssh-client set to manually installed.
openssh-server is already the newest version (1:8.2p1-4ubuntu0.5).
The following packages were automatically installed and are no longer required:
  chromium-codecs-ffmpeg-extra gstreamer1.0-vaapi
  libgstreamer-plugins-bad1.0-0 libva-wayland2
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 401 not upgraded.
20127444 :
```

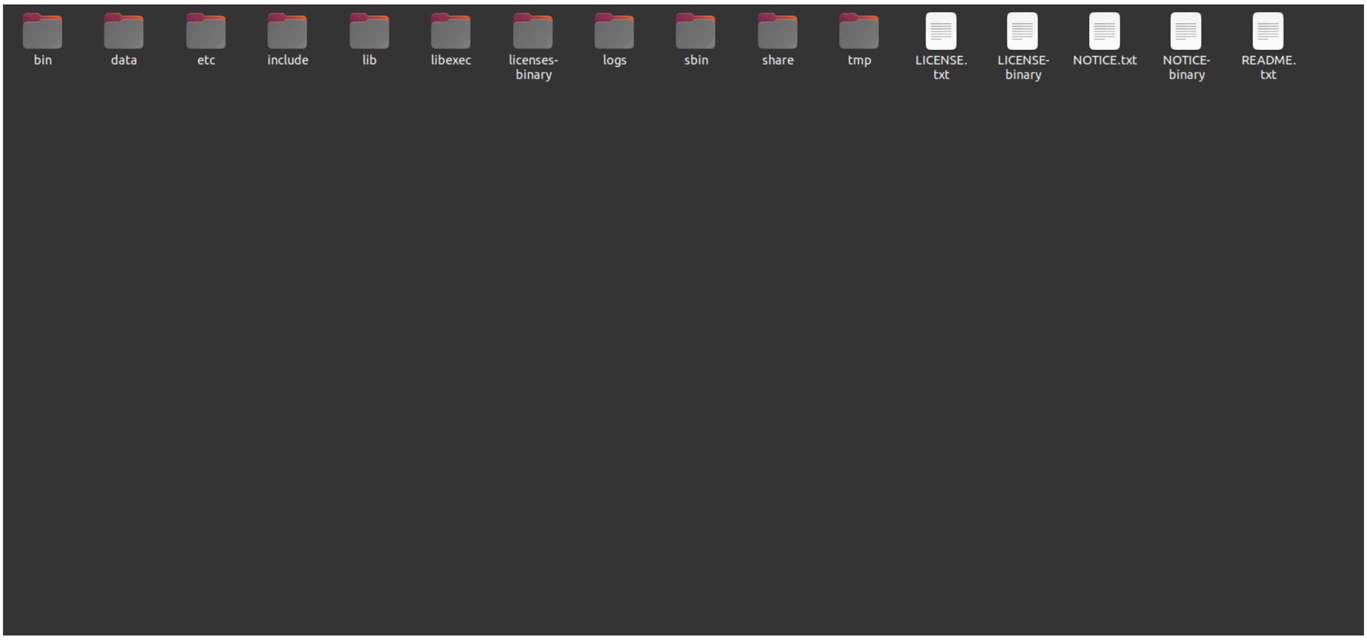
- Create and Install SSH Certificates

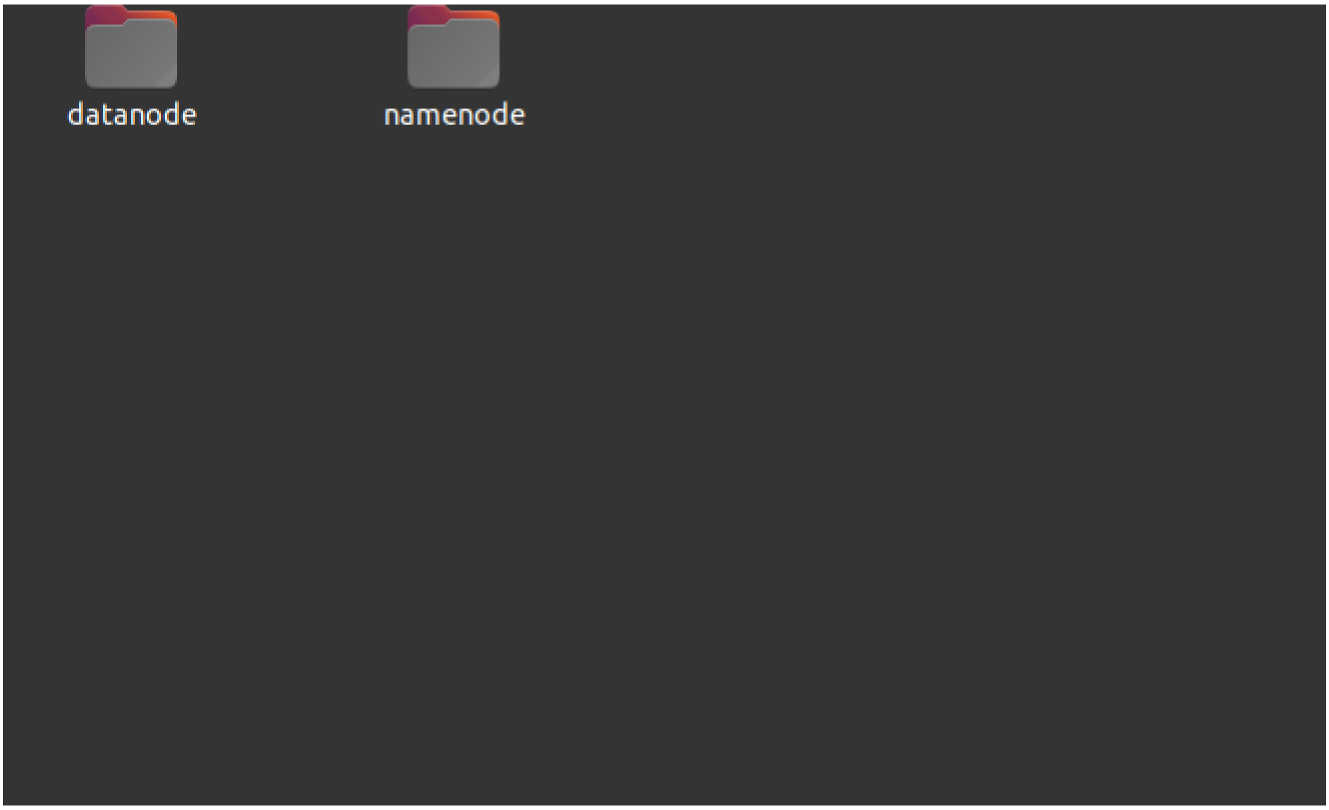


```
20127444 :ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
/home/baobui16/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /home/baobui16/.ssh/id_rsa
Your public key has been saved in /home/baobui16/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:p4EOenn7ngGlIa0SLDOqNteNKkupf8jTvu1/hQz9E4c baobui16@T14
The key's randomart image is:
+---[RSA 3072]----+
|                 |
| .   .           |
|+ o . o ..   .   |
|.+ . o =. . E .  |
|. . o + So.o o   |
|. .o.+o. +o +    |
|.*.+ooo.o  . .   |
|+.*.+o . o.      |
|.o+=oo++=.       |
+----[SHA256]-----+
20127444 :
```

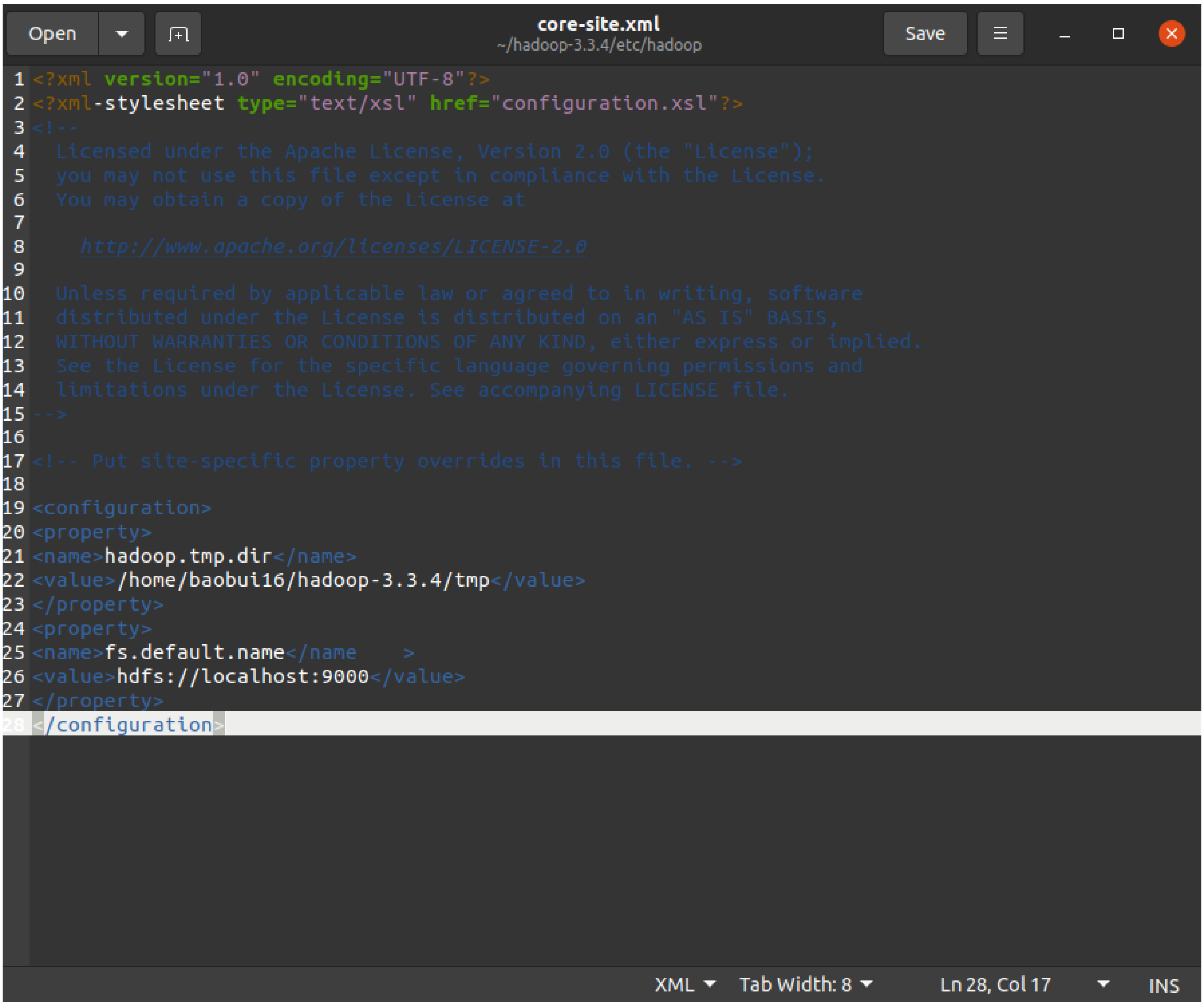


```
20127444 :cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
20127444 :ssh localhost
Welcome to Ubuntu 20.04.3 LTS (GNU/Linux 5.14.0-1057-oem x86_64)

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

2 devices have a firmware upgrade available.
Run `fwupdmgr get-upgrades` for more information.


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

Last login: Wed Mar  1 08:20:58 2023 from 127.0.0.1
(base) baobui16@T14:~$ 
```



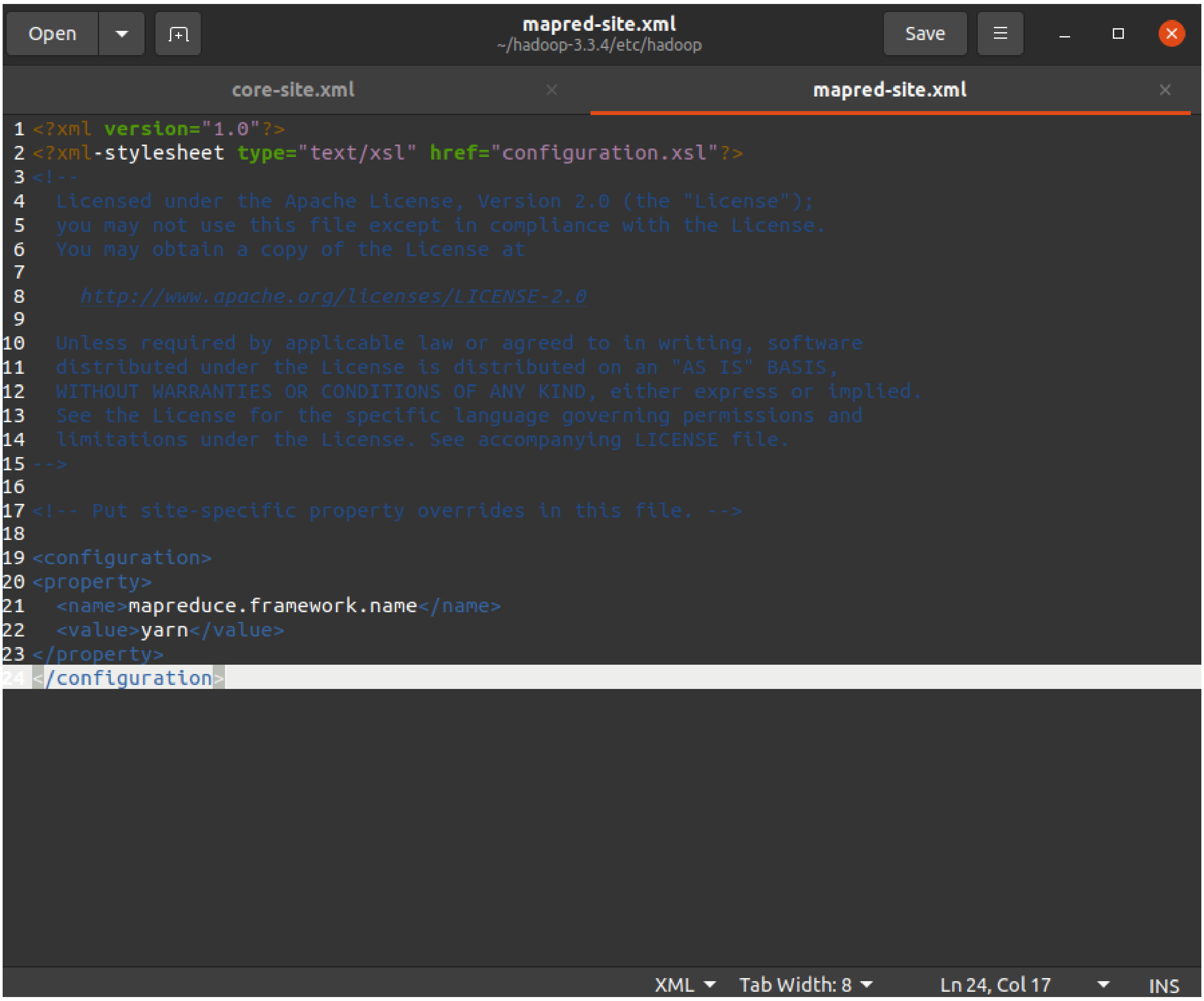


- Configure file core-site.xml



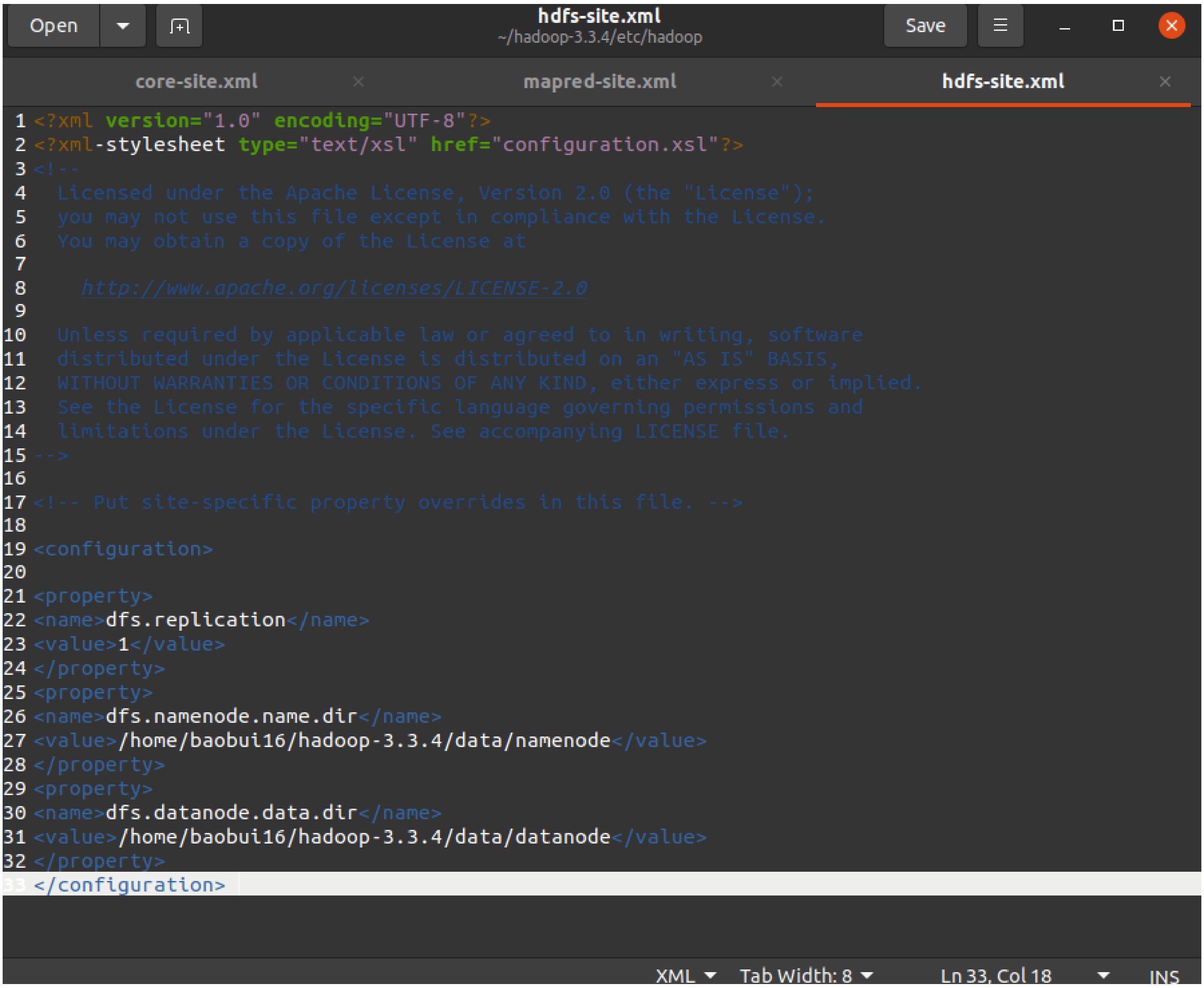
```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

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

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

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

<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/home/baobui16/hadoop-3.3.4/tmp</value>
</property>
<property>
<name>fs.default.name</name    >
<value>hdfs://localhost:9000</value>
</property>
</configuration>
```

- Configure the file mapred-site.xml



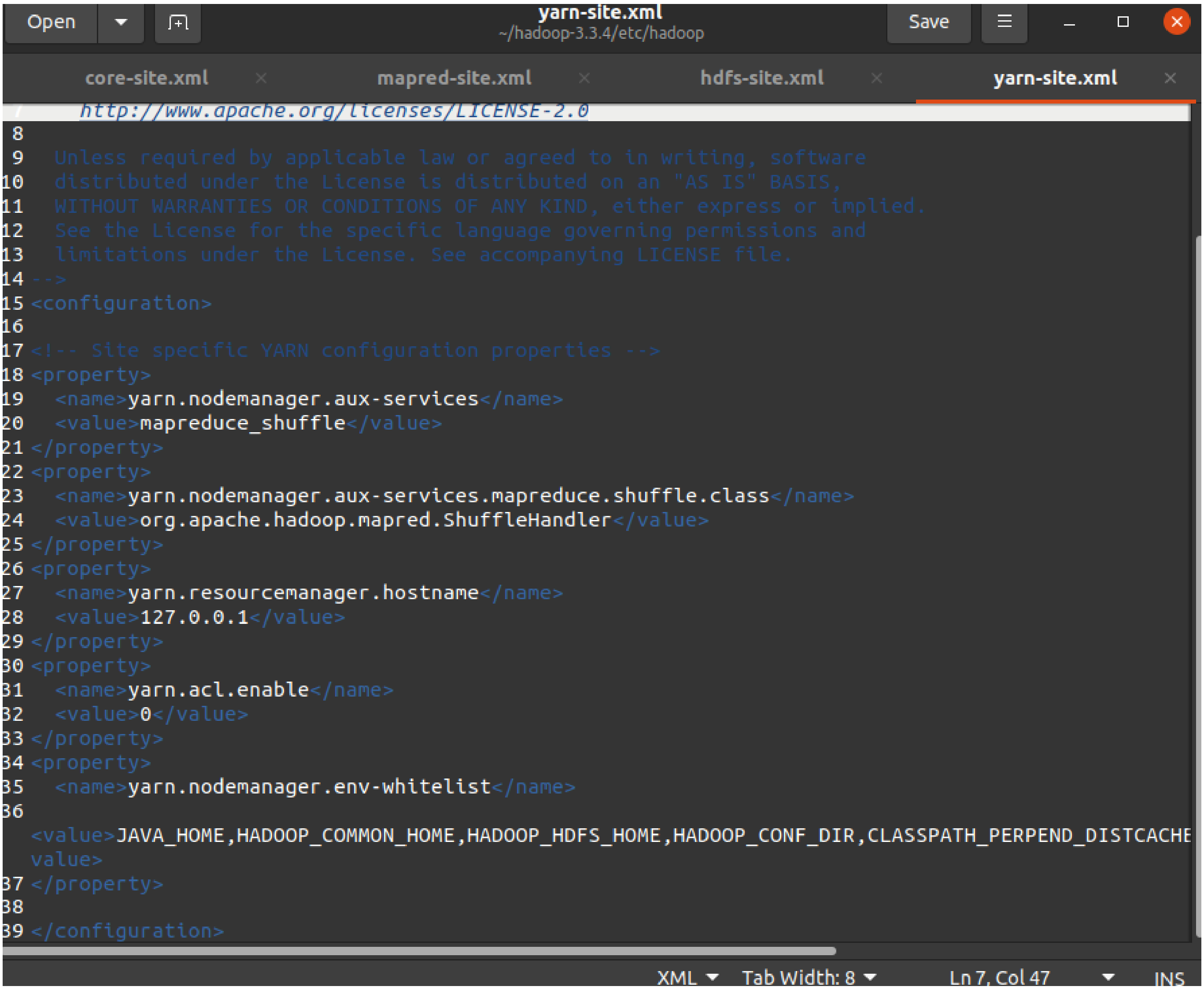
```xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

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

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

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

<configuration>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>
```

- Configure file hdfs-site.xml



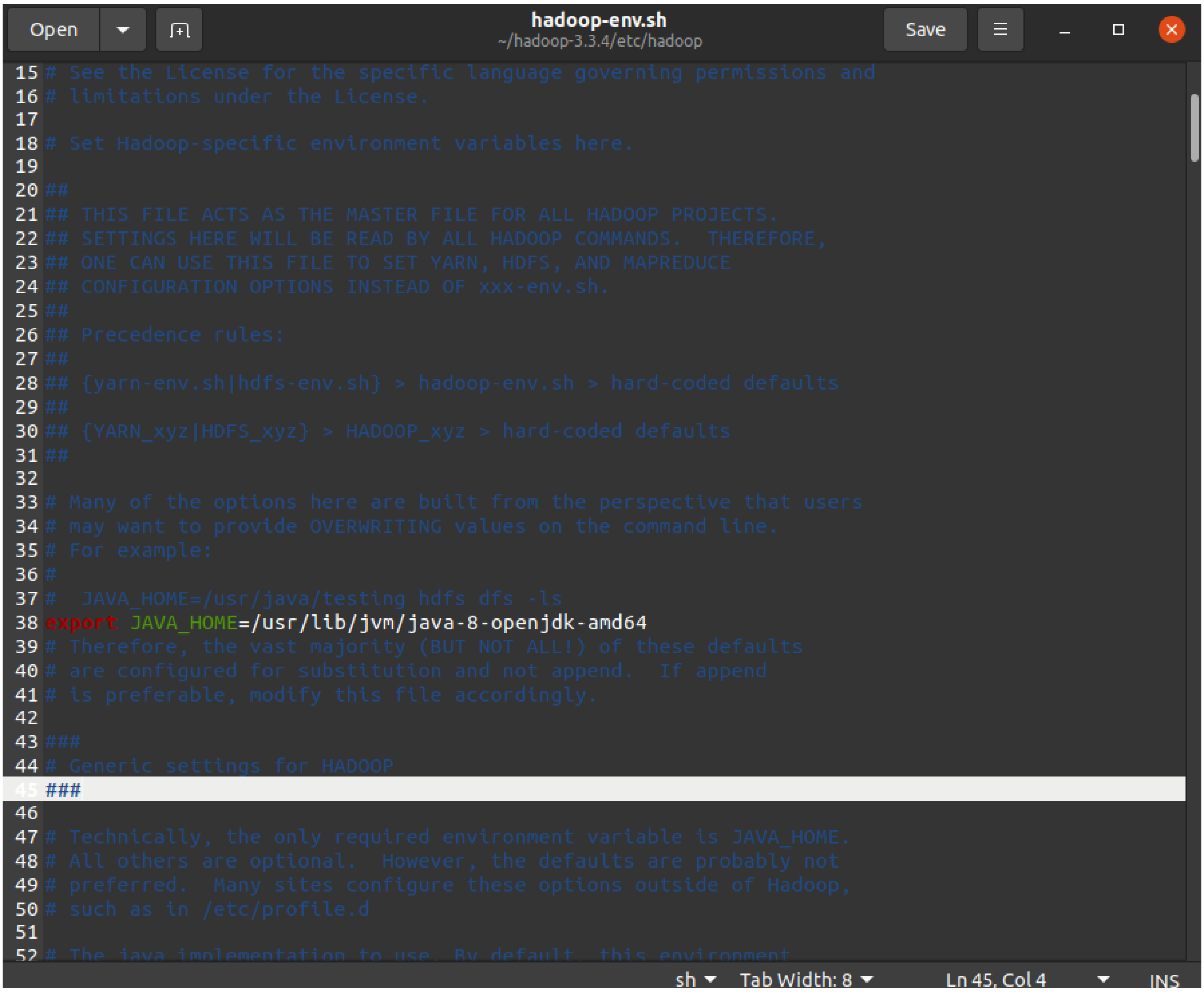
```xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    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/baobui16/hadoop-3.3.4/data/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/baobui16/hadoop-3.3.4/data/datanode</value>
</property>
</configuration>
```

- Configure file yarn-site.xml



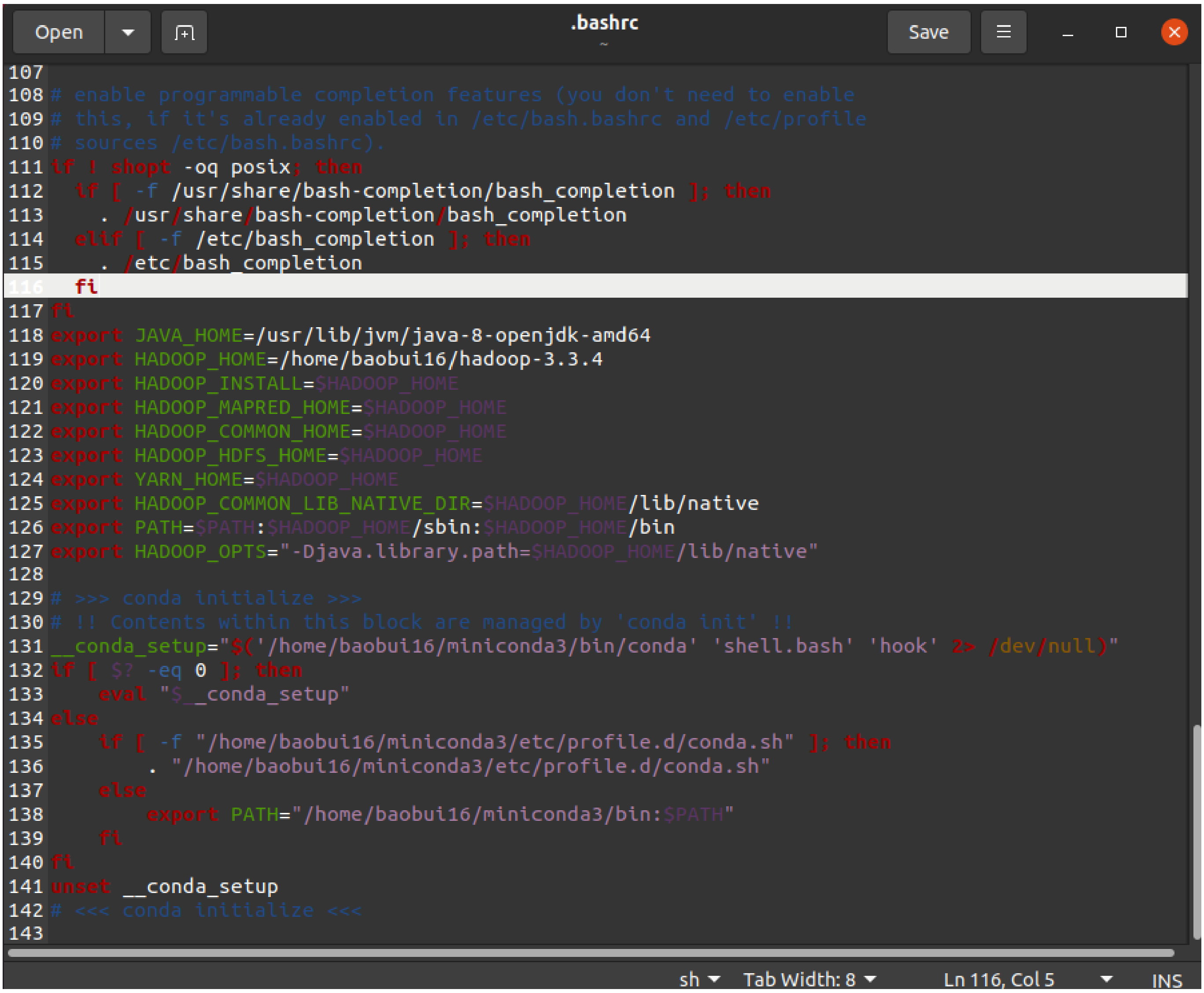
```
    http://www.apache.org/licenses/LICENSE-2.0

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

<!-- Site specific YARN configuration properties -->
<property>
  <name>yarn.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
value>
</property>

</configuration>
```

- Configure file hadoop-env.sh



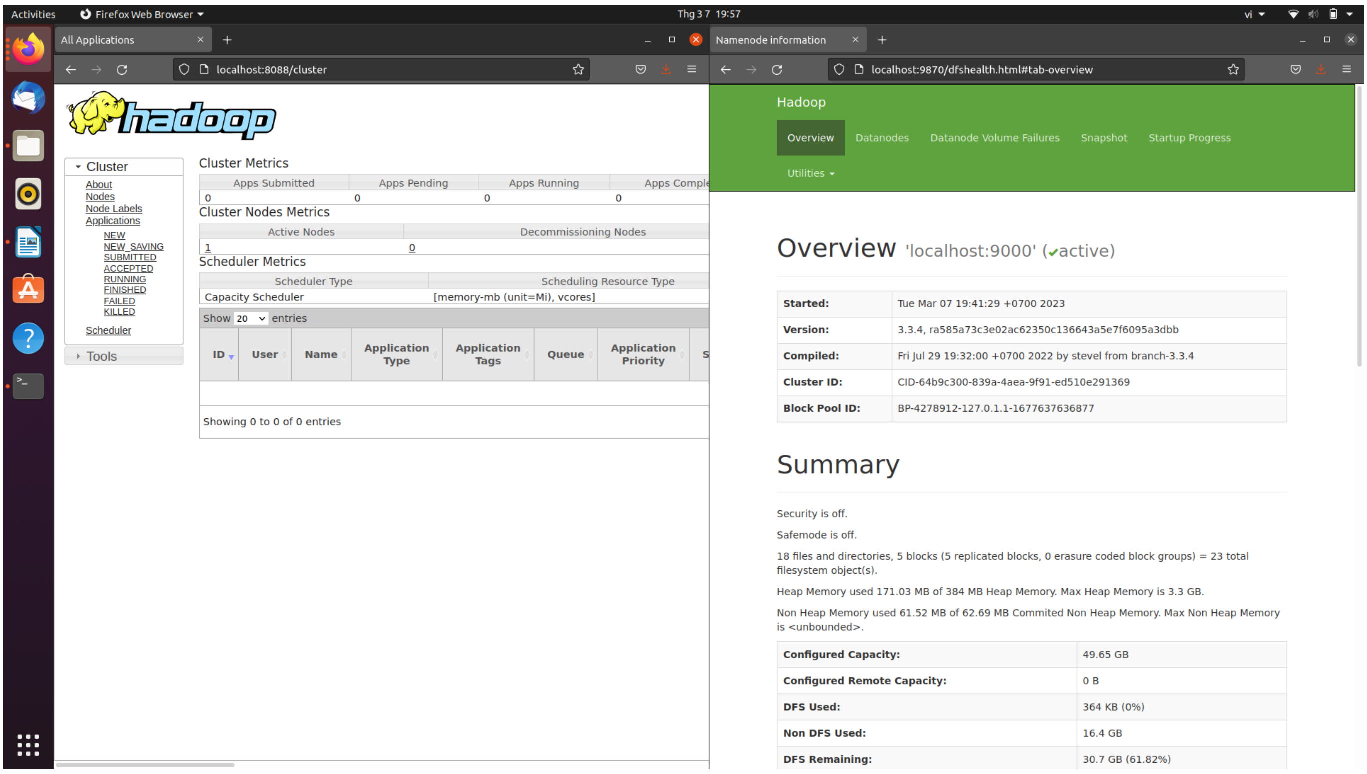
```
# See the License for the specific language governing permissions and
# limitations under the License.

# Set Hadoop-specific environment variables here.

##
## THIS FILE ACTS AS THE MASTER FILE FOR ALL HADOOP PROJECTS.
## SETTINGS HERE WILL BE READ BY ALL HADOOP COMMANDS.  THEREFORE,
## ONE CAN USE THIS FILE TO SET YARN, HDFS, AND MAPREDUCE
## CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.
##
## Precedence rules:
##
## {yarn-env.sh|hdfs-env.sh} > hadoop-env.sh > hard-coded defaults
##
## {YARN_xyz|HDFS_xyz} > HADOOP_xyz > hard-coded defaults
##

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

###
# Generic settings for HADOOP
###

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

# The java implementation to use. By default, this environment
```

- Configure file bash



```
# 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 JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/home/baobui16/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"

# >>> conda initialize >>>
# !! Contents within this block are managed by 'conda init' !!
__conda_setup="$('/home/baobui16/miniconda3/bin/conda' 'shell.bash' 'hook' 2> /dev/null)"
if [ $? -eq 0 ]; then
    eval "$__conda_setup"
else
    if [ -f "/home/baobui16/miniconda3/etc/profile.d/conda.sh" ]; then
        . "/home/baobui16/miniconda3/etc/profile.d/conda.sh"
    else
        export PATH="/home/baobui16/miniconda3/bin:$PATH"
    fi
fi
unset __conda_setup
# <<< conda initialize <<<
```

- Successful installation



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Comple
0	0	0	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes
1	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type
Capacity Scheduler	[memory-mb (unit=Mi), vcores]

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	S

Showing 0 to 0 of 0 entries

Overview 'localhost:9000' (✓active)

Started:	Tue Mar 07 19:41:29 +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-64b9c300-839a-4aea-9f91-ed510e291369
Block Pool ID:	BP-4278912-127.0.1.1-1677637636877

Summary

Security is off.

Safemode is off.

18 files and directories, 5 blocks (5 replicated blocks, 0 erasure coded block groups) = 23 total filesystem object(s).

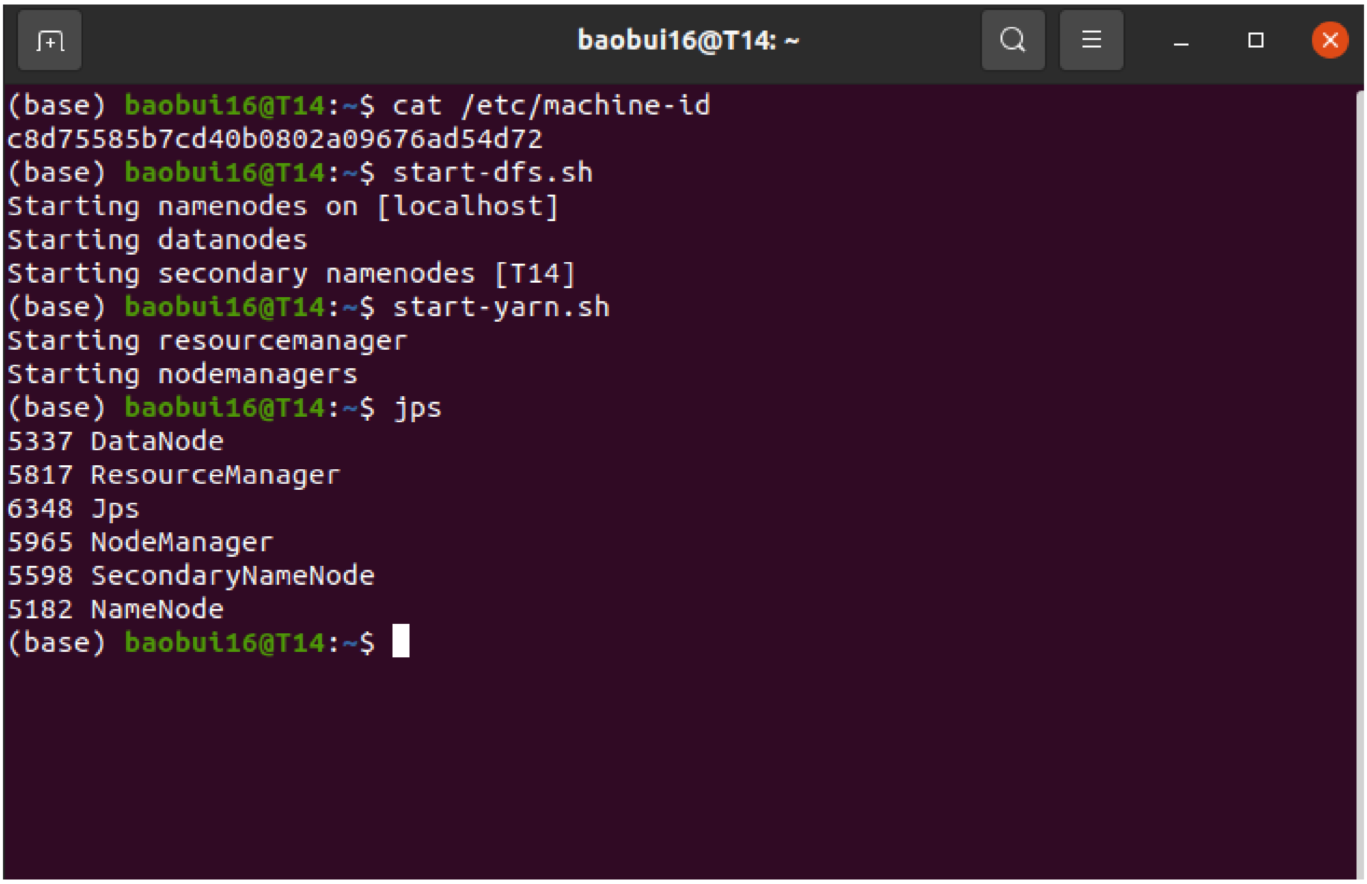
Heap Memory used 171.03 MB of 384 MB Heap Memory. Max Heap Memory is 3.3 GB.

Non Heap Memory used 61.52 MB of 62.69 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	49.65 GB
Configured Remote Capacity:	0 B
DFS Used:	364 KB (0%)
Non DFS Used:	16.4 GB
DFS Remaining:	30.7 GB (61.82%)

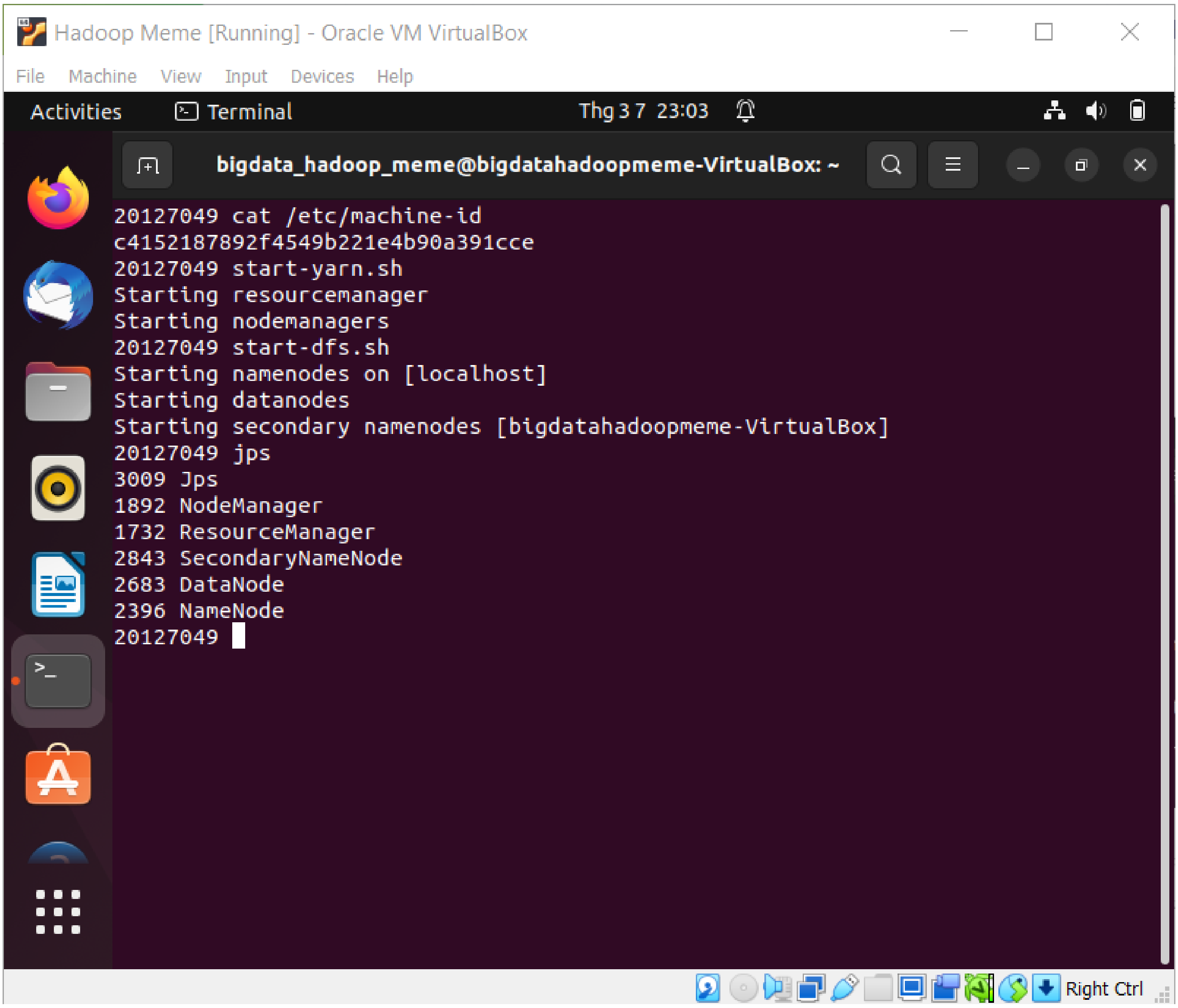
SET SUCCESS OF TEAM MEMBERS:

- 20127444 - Bùi Duy Bảo



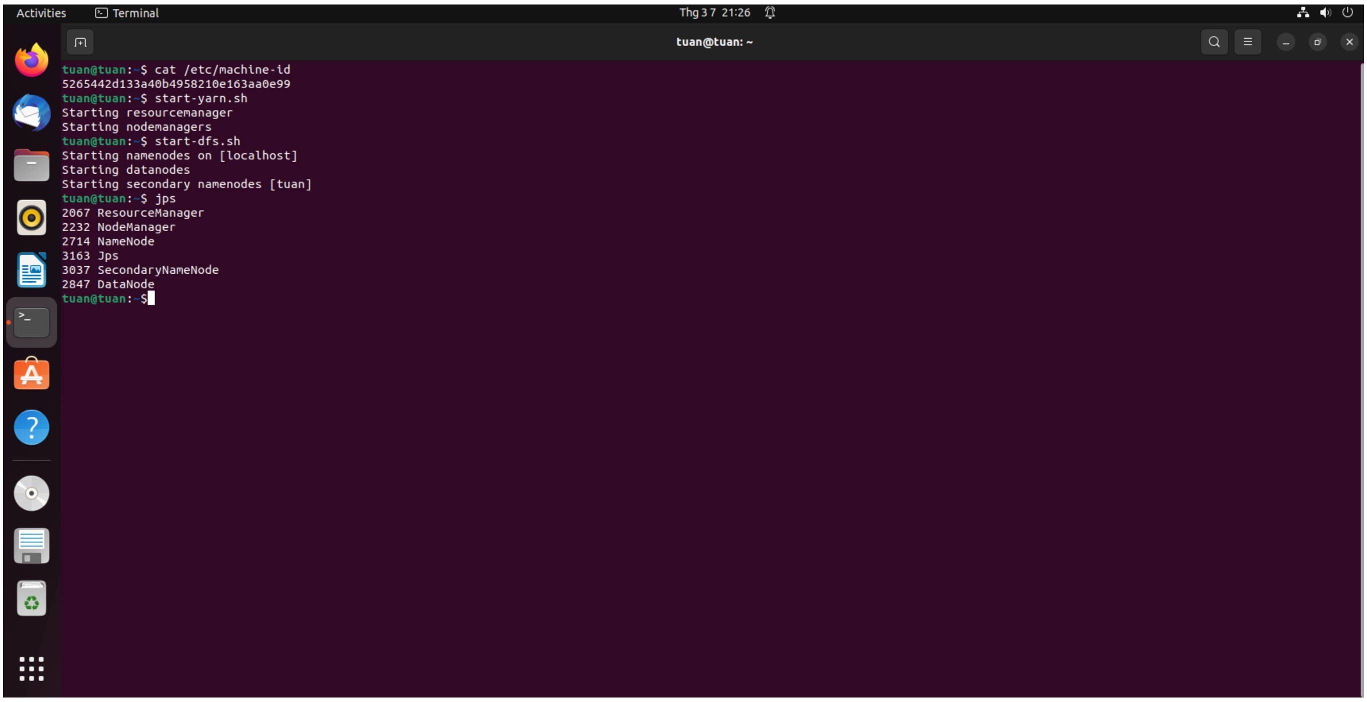
```
(base) baobui16@T14:~$ cat /etc/machine-id
c8d75585b7cd40b0802a09676ad54d72
(base) baobui16@T14:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [T14]
(base) baobui16@T14:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
(base) baobui16@T14:~$ jps
5337 DataNode
5817 ResourceManager
6348 Jps
5965 NodeManager
5598 SecondaryNameNode
5182 NameNode
(base) baobui16@T14:~$
```

- 20127049 - Nguyễn Đức Minh


Hadoop Meme [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Thg 3 7 23:03
bigdata_hadoop_meme@bigdatahadoopmeme-VirtualBox: ~
Right Ctrl

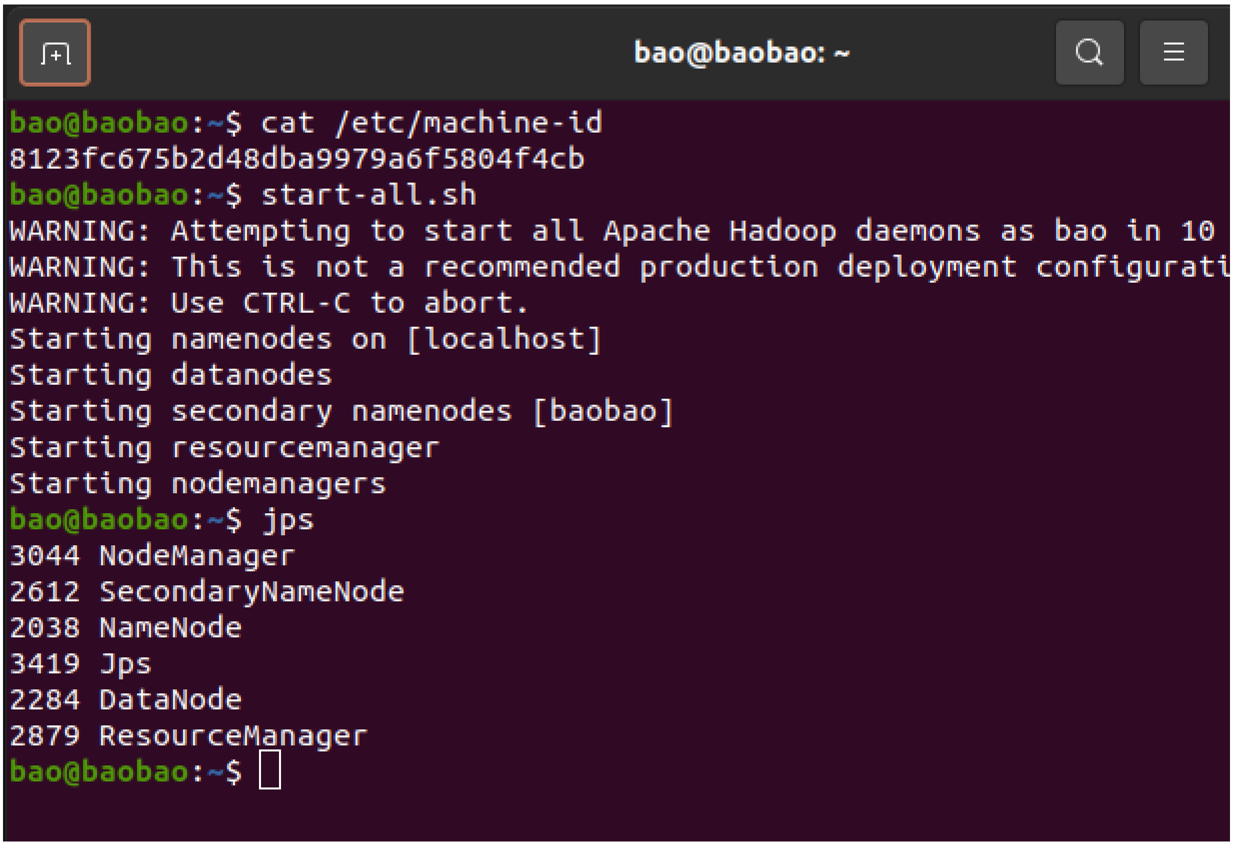
```
20127049 cat /etc/machine-id
c4152187892f4549b221e4b90a391cce
20127049 start-yarn.sh
Starting resourcemanager
Starting nodemanagers
20127049 start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bigdatahadoopmeme-VirtualBox]
20127049 jps
3009 Jps
1892 NodeManager
1732 ResourceManager
2843 SecondaryNameNode
2683 DataNode
2396 NameNode
20127049
```

- 20127092 - Nguyễn Minh Tuấn



```
tuan@tuan:~$ cat /etc/machine-id
5265442d133a40b4958210e163aa0e99
tuan@tuan:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
tuan@tuan:~$ start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [tuan]
tuan@tuan:~$ jps
2067 ResourceManager
2232 NodeManager
2714 NameNode
3163 Jps
3037 SecondaryNameNode
2847 DataNode
tuan@tuan:~$
```

- 20127448 - Nguyễn Thái Bảo



```
bao@baobao:~$ cat /etc/machine-id
8123fc675b2d48dba9979a6f5804f4cb
bao@baobao:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as bao in 10
WARNING: This is not a recommended production deployment configurati
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [baobao]
Starting resourcemanager
Starting nodemanagers
bao@baobao:~$ jps
3044 NodeManager
2612 SecondaryNameNode
2038 NameNode
3419 Jps
2284 DataNode
2879 ResourceManager
bao@baobao:~$ 
```

Introduction to MapReduce

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

The input keys-values represent the input data that is read from HDFS or other data sources

The intermediate keys-values are generated by the map tasks during the processing of the input data, and are in a different format or schema than the input data. These intermediate pairs are then sorted and shuffled across the cluster, and sent to the reduce tasks for further processing.

The output keys-values are the final key-value pairs generated by the reduce tasks, and are the result of aggregating, summarizing, or transforming the intermediate pairs

Overall, the input keys-values, intermediate keys-values, and output keys-values are all important components of the MapReduce data processing model, and they are all related to each other in the sense that they represent different stages in the processing of data on a distributed cluster.

2. **How does MapReduce deal with node failures?**

MapReduce deals with node failure by being designed to be fault-tolerant and by using some techniques, for example data replication and job restart even node failure appears.

About by being designed to be fault-tolerant: it means that MapReduce can handle node failures and continue processing the job. If a node fails, the tasks running on that node are automatically reassigned to other available nodes. The framework also periodically pings the nodes to check if they are still alive. If a node does not respond, it is assumed to have failed, and its tasks are reassigned to other nodes.

About by using data replication: MapReduce uses this technique to ensure that data is not lost when node failure happens, each block of data is replicated across multiple nodes in the cluster. If one node fails, the data is still available on the other nodes, and the job can continue processing.

About by using job restart: in this situation, the entire job may need to be restarted. This is because if a node fails while processing a task, the output of that task may be lost. If the output of a task is lost, any subsequent tasks that depend on that output will need to be rerun.

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

The locality is input data is stored on the local disks of the machines that make up cluster. It's manager by GFS which divides each file into 64 MB blocks, and stores several copies of each block (typically 3 copies) on different machines,if it had any node fail ,GFS would be specify another node that contains the copy in order to comply with the minimum requirement, 3 copies must be obtained

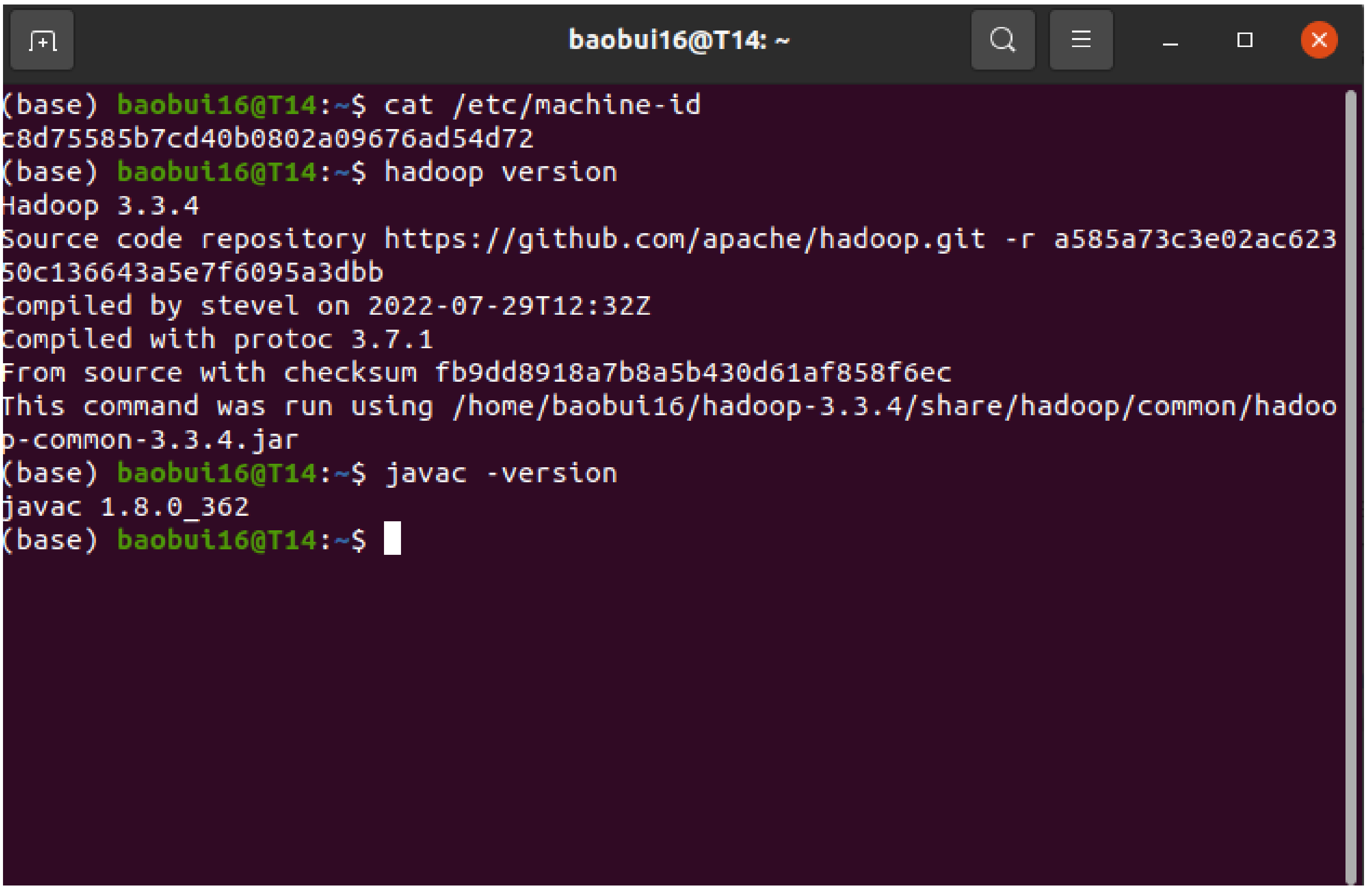
It's used when running large MapReduce operations on a significant fraction of the workers in a cluster, most input data is read locally and consumes so it not need to spend network bandwidth

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

Combiner Function is similar to small Reduce phases of each Mapper on local disk which helps the process decrease number of pairs (key-value) before Reduce phases. As we see, reducing pairs of key-value on local disk decreases the workload that enhance bandwidth quality to run the process faster.

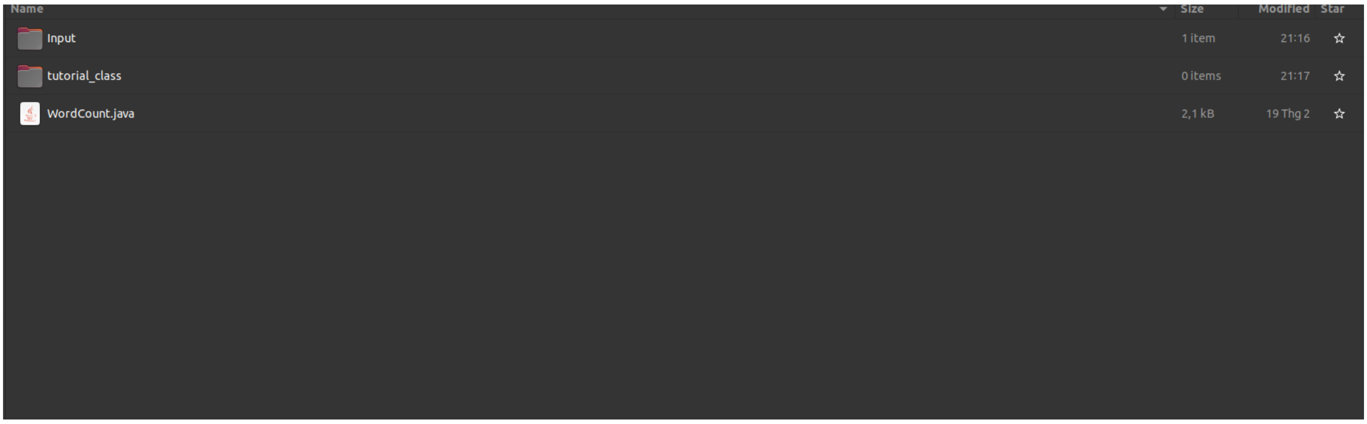
Running a warm-up problem: Word Count

- Check Hadoop version

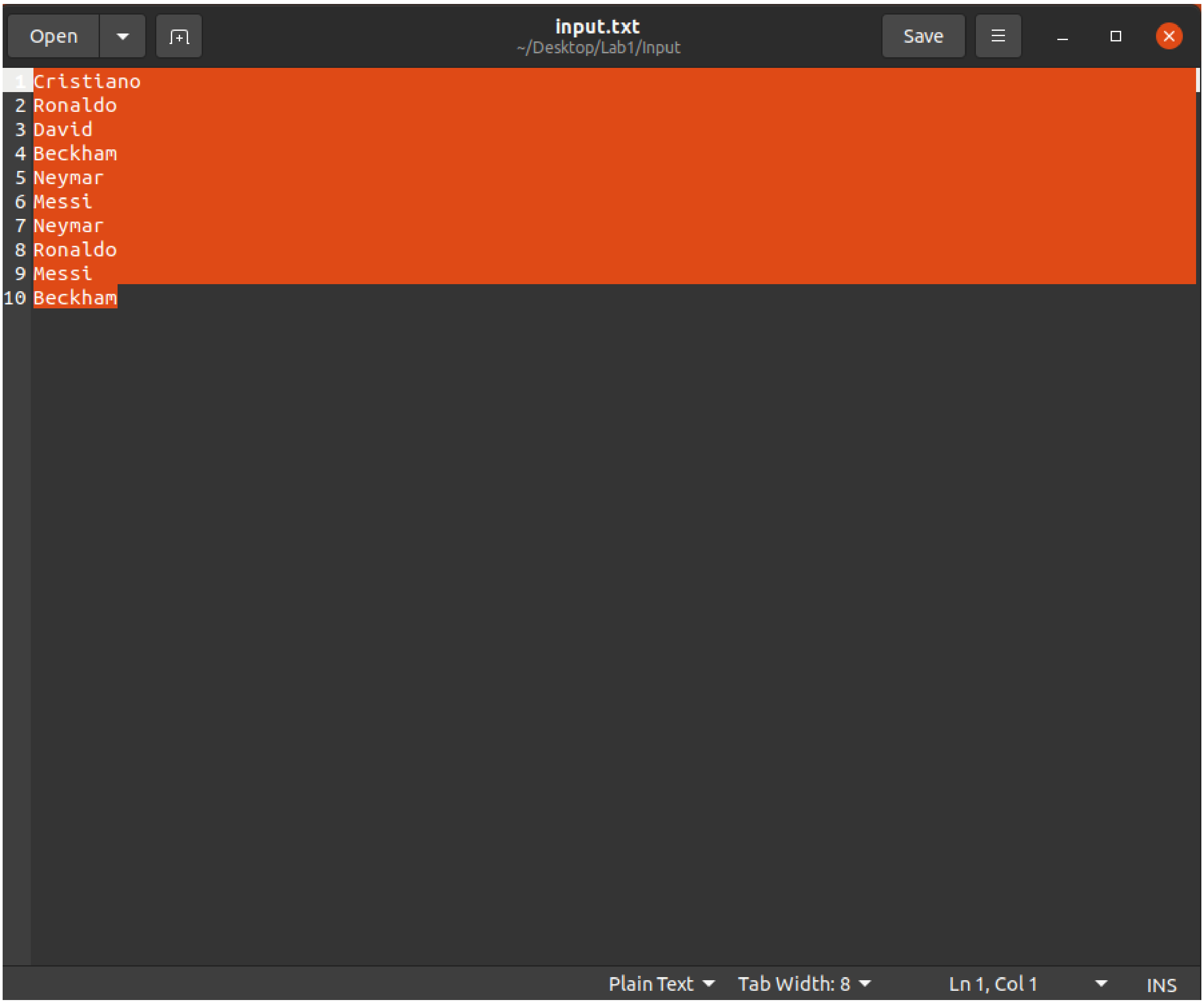


```
(base) baobui16@T14:~$ cat /etc/machine-id
c8d75585b7cd40b0802a09676ad54d72
(base) baobui16@T14:~$ hadoop version
Hadoop 3.3.4
Source code repository https://github.com/apache/hadoop.git -r a585a73c3e02ac623
50c136643a5e7f6095a3dbb
Compiled by stevel on 2022-07-29T12:32Z
Compiled with protoc 3.7.1
From source with checksum fb9dd8918a7b8a5b430d61af858f6ec
This command was run using /home/baobui16/hadoop-3.3.4/share/hadoop/common/hadoo
p-common-3.3.4.jar
(base) baobui16@T14:~$ javac -version
javac 1.8.0_362
(base) baobui16@T14:~$
```

- The input directory contains the file input.txt

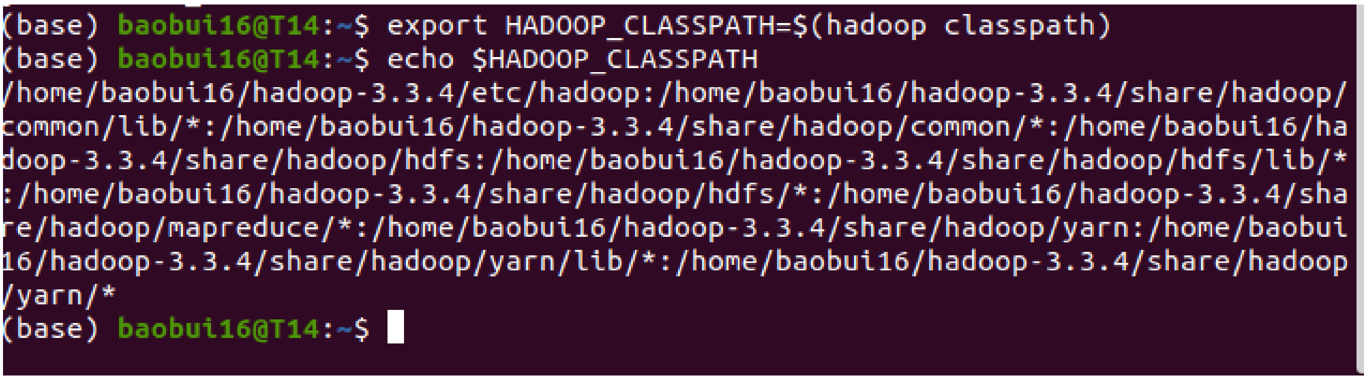


Name	Size	Modified	Star
Input	1 item	21:16	☆
tutorial_class	0 items	21:17	☆
WordCount.java	2,1 kB	19 Thg 2	☆



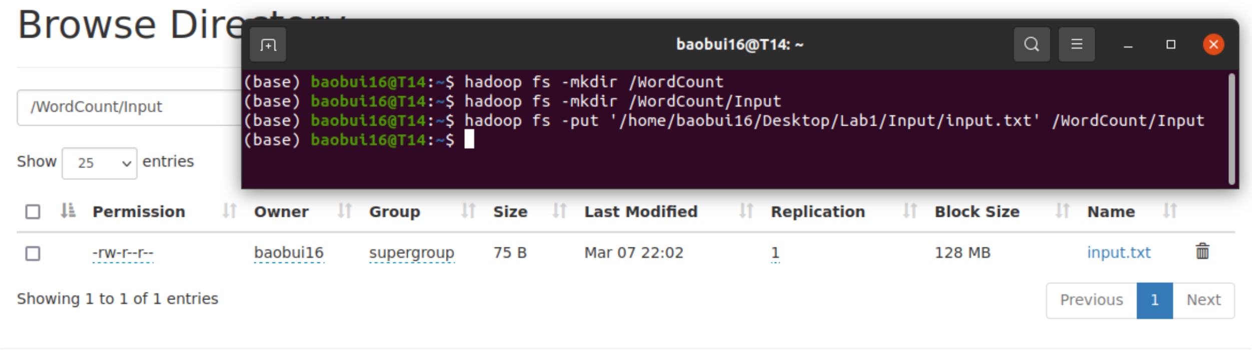
```
Cristiano
Ronaldo
David
Beckham
Neymar
Messi
Neymar
Ronaldo
Messi
Beckham
```

- Enter the following path to export the Hadoop classpath to bash
- Make sure it's now exported



```
(base) baobui16@T14:~$ export HADOOP_CLASSPATH=$(hadoop classpath)
(base) baobui16@T14:~$ echo $HADOOP_CLASSPATH
/home/baobui16/hadoop-3.3.4/etc/hadoop:/home/baobui16/hadoop-3.3.4/share/hadoop/
common/lib/*:/home/baobui16/hadoop-3.3.4/share/hadoop/common/*:/home/baobui16/ha
doop-3.3.4/share/hadoop/hdfs:/home/baobui16/hadoop-3.3.4/share/hadoop/hdfs/lib/*
:/home/baobui16/hadoop-3.3.4/share/hadoop/hdfs/*:/home/baobui16/hadoop-3.3.4/sha
re/hadoop/mapreduce/*:/home/baobui16/hadoop-3.3.4/share/hadoop/yarn:/home/baobui
16/hadoop-3.3.4/share/hadoop/yarn/lib/*:/home/baobui16/hadoop-3.3.4/share/hadoop
/yarn/*
(base) baobui16@T14:~$
```

- Create this directory on HDFS and put input.txt . file



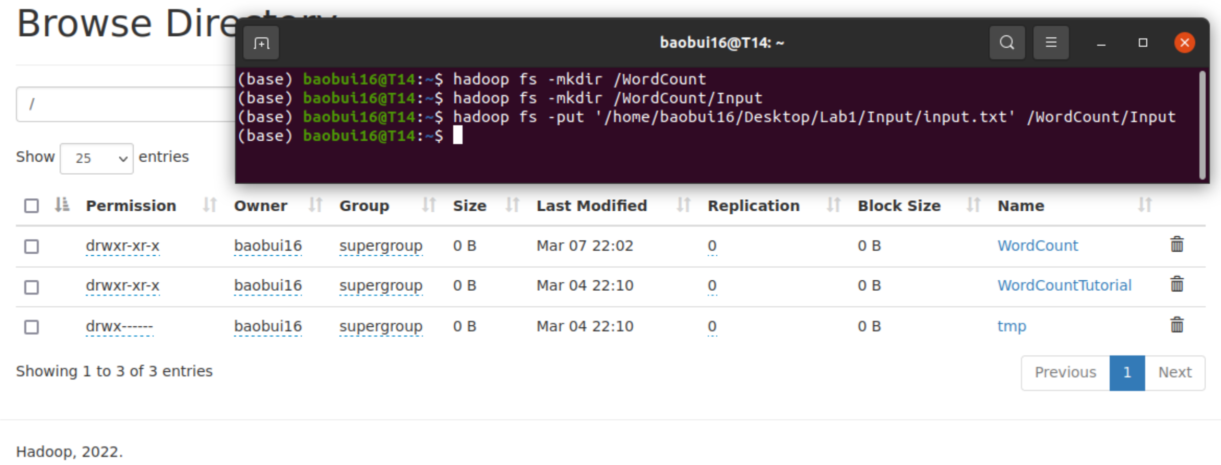
Browse Dire

/WordCount/Input

```
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount/Input
(base) baobui16@T14:~$ hadoop fs -put '/home/baobui16/Desktop/Lab1/Input/input.txt' /WordCount/Input
(base) baobui16@T14:~$
```

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	baobui16	supergroup	75 B	Mar 07 22:02	1	128 MB	input.txt	🗑

Showing 1 to 1 of 1 entries



Browse Dire

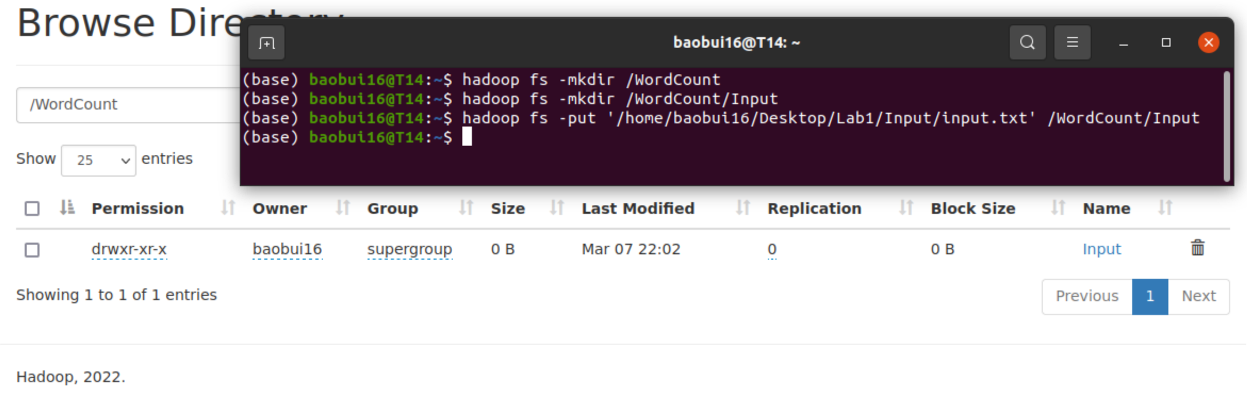
/

```
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount/Input
(base) baobui16@T14:~$ hadoop fs -put '/home/baobui16/Desktop/Lab1/Input/input.txt' /WordCount/Input
(base) baobui16@T14:~$
```

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxr-xr-x	baobui16	supergroup	0 B	Mar 07 22:02	0	0 B	WordCount	🗑
□	drwxr-xr-x	baobui16	supergroup	0 B	Mar 04 22:10	0	0 B	WordCountTutorial	🗑
□	drwx------	baobui16	supergroup	0 B	Mar 04 22:10	0	0 B	tmp	🗑

Showing 1 to 3 of 3 entries

Hadoop, 2022.



Browse Dire

/WordCount

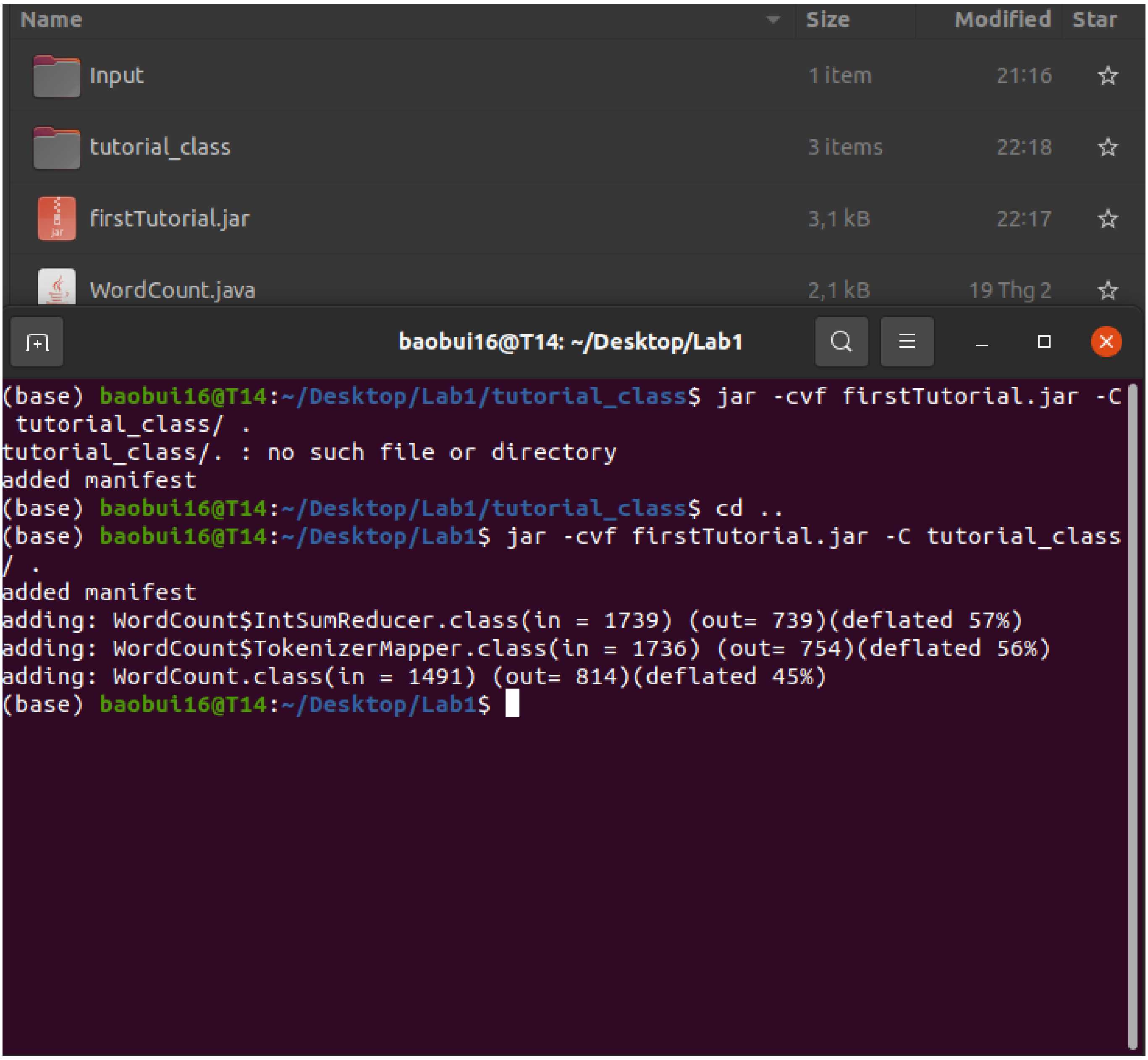
```
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount/Input
(base) baobui16@T14:~$ hadoop fs -put '/home/baobui16/Desktop/Lab1/Input/input.txt' /WordCount/Input
(base) baobui16@T14:~$
```

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	drwxr-xr-x	baobui16	supergroup	0 B	Mar 07 22:02	0	0 B	Input	🗑

Showing 1 to 1 of 1 entries

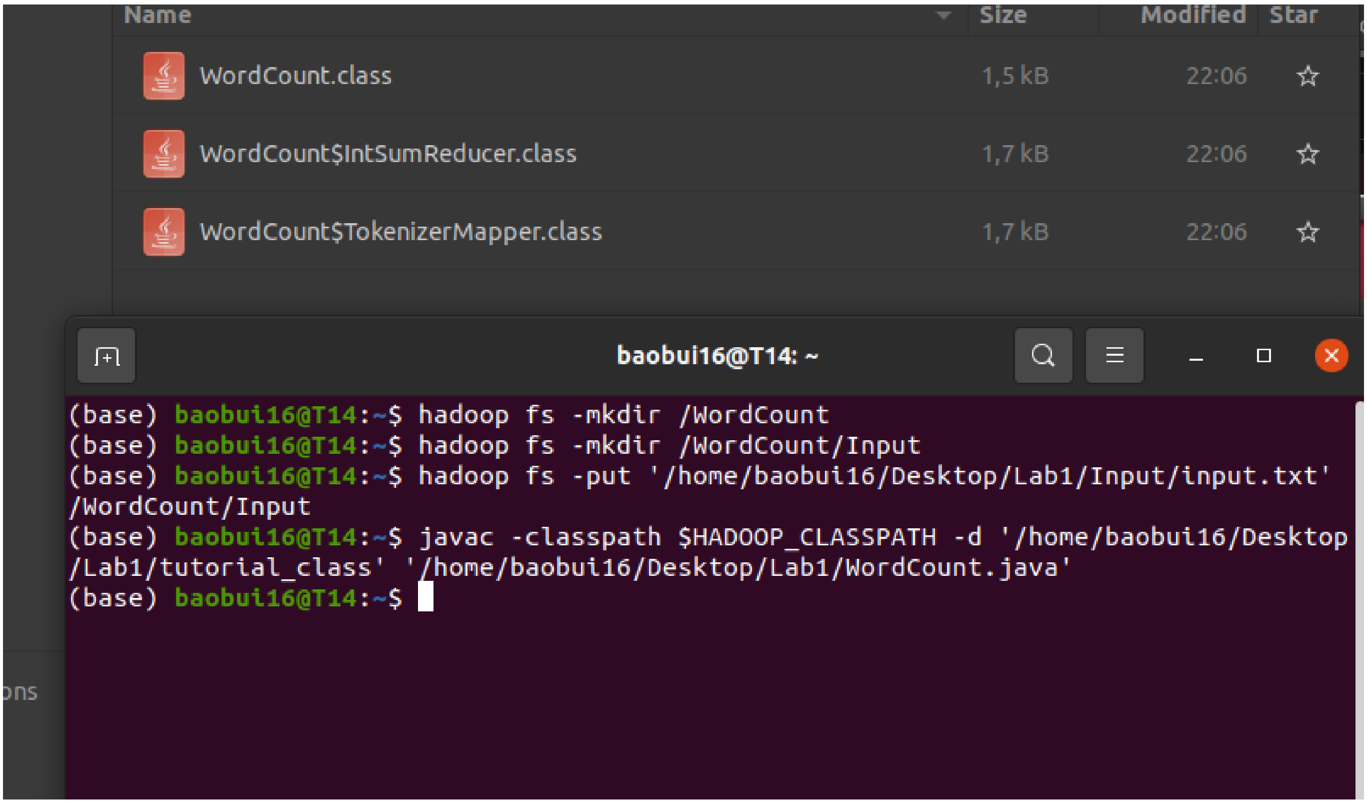
Hadoop, 2022.

- Run file jar on Hadoop



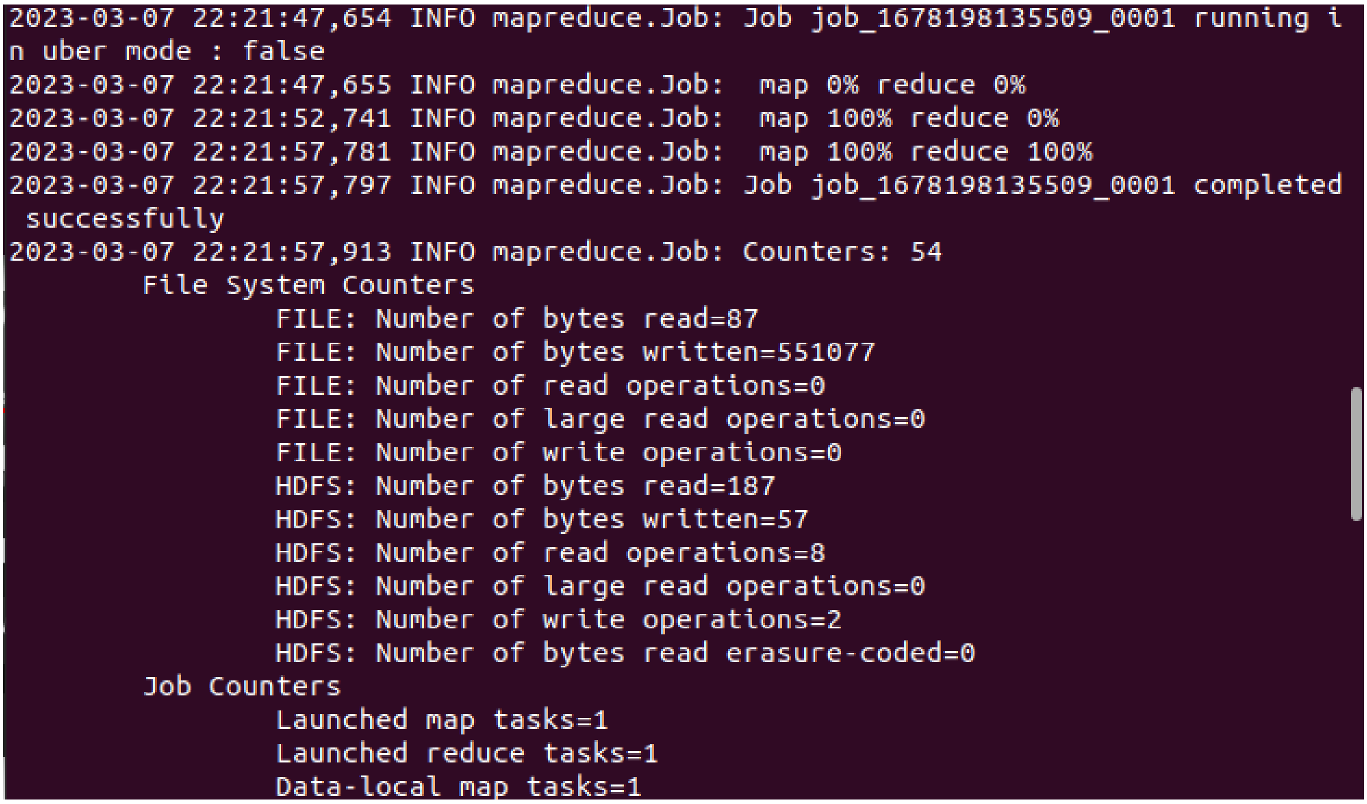
Name	Size	Modified	Star
Input	1 item	21:16	☆
tutorial_class	3 items	22:18	☆
firstTutorial.jar	3,1 kB	22:17	☆
WordCount.java	2,1 kB	19 Thg 2	☆

```
(base) baobui16@T14:~/Desktop/Lab1/tutorial_class$ jar -cvf firstTutorial.jar -C
 tutorial_class/ .
tutorial_class/. : no such file or directory
added manifest
(base) baobui16@T14:~/Desktop/Lab1/tutorial_class$ cd ..
(base) baobui16@T14:~/Desktop/Lab1$ jar -cvf firstTutorial.jar -C tutorial_class
/ .
added manifest
adding: WordCount$IntSumReducer.class(in = 1739) (out= 739)(deflated 57%)
adding: WordCount$TokenizerMapper.class(in = 1736) (out= 754)(deflated 56%)
adding: WordCount.class(in = 1491) (out= 814)(deflated 45%)
(base) baobui16@T14:~/Desktop/Lab1$ 
```



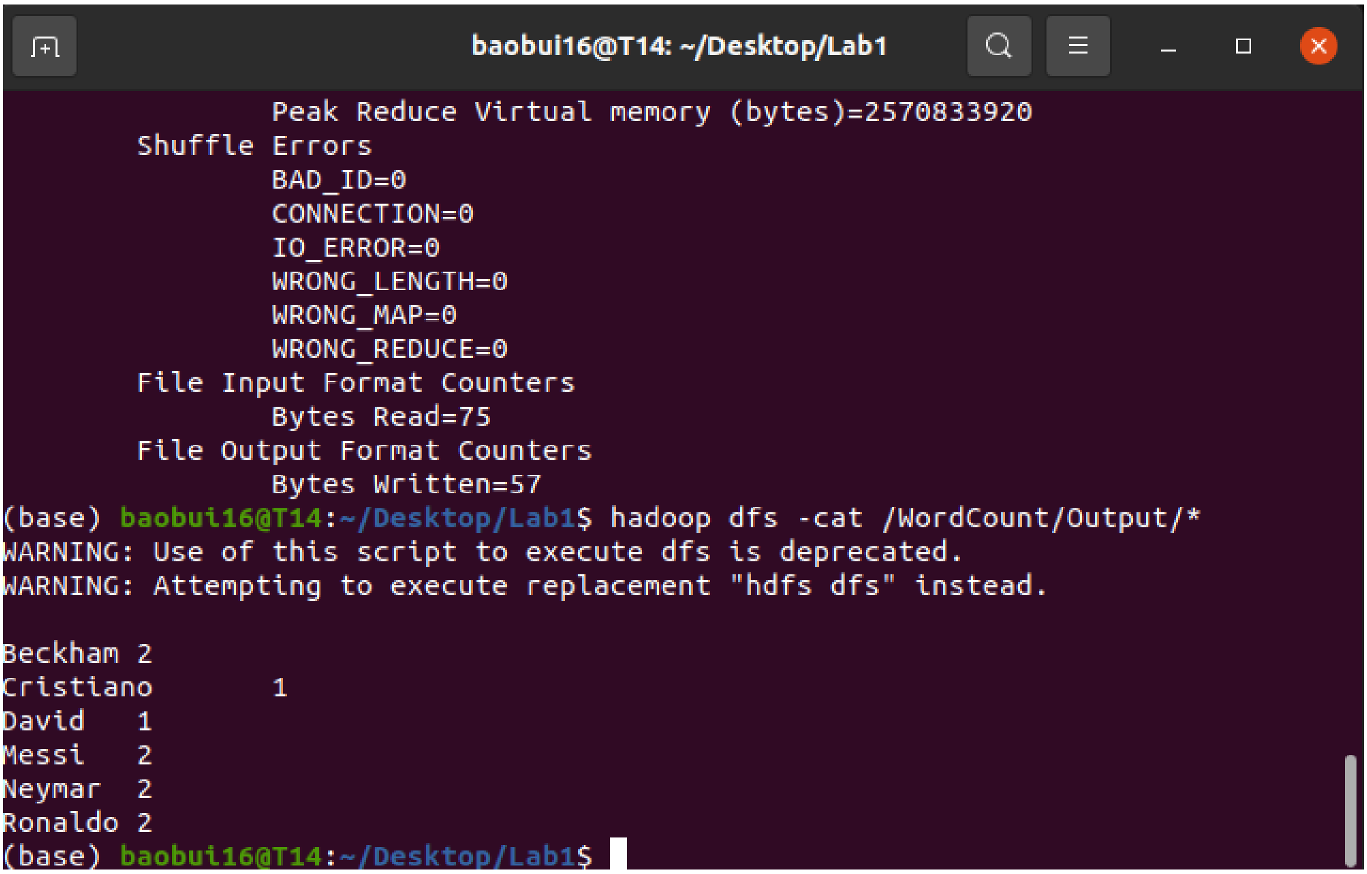
Name	Size	Modified	Star
WordCount.class	1,5 kB	22:06	☆
WordCount$IntSumReducer.class	1,7 kB	22:06	☆
WordCount$TokenizerMapper.class	1,7 kB	22:06	☆

```
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount
(base) baobui16@T14:~$ hadoop fs -mkdir /WordCount/Input
(base) baobui16@T14:~$ hadoop fs -put '/home/baobui16/Desktop/Lab1/Input/input.txt'
/WordCount/Input
(base) baobui16@T14:~$ javac -classpath $HADOOP_CLASSPATH -d '/home/baobui16/Desktop
/Lab1/tutorial_class' '/home/baobui16/Desktop/Lab1/WordCount.java'
(base) baobui16@T14:~$ 
```

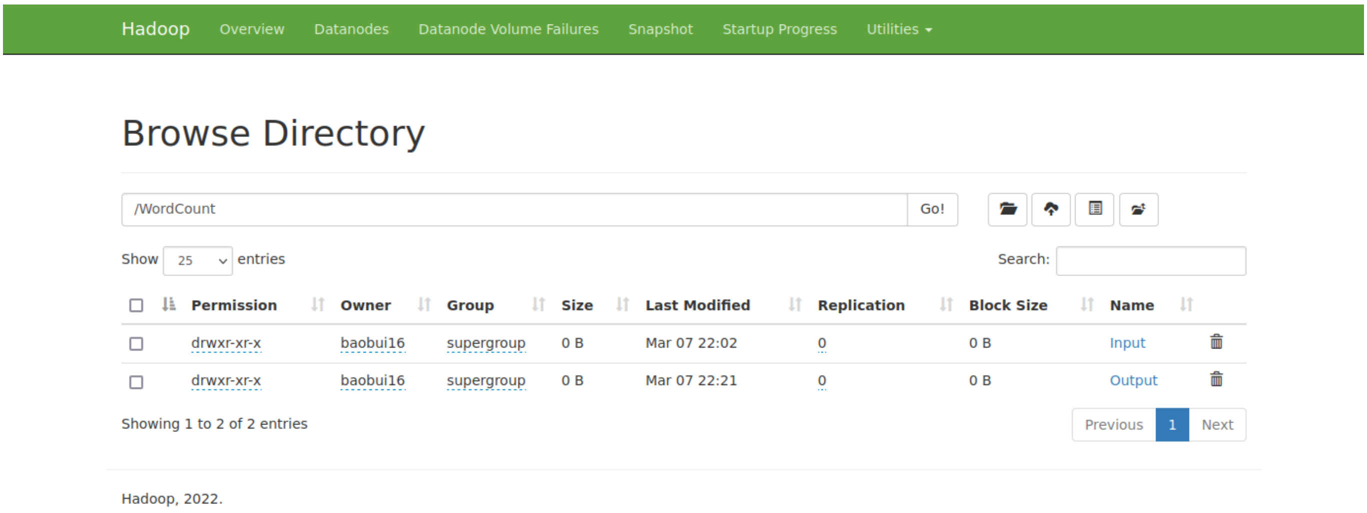


```
2023-03-07 22:21:47,654 INFO mapreduce.Job: Job job_1678198135509_0001 running i
n uber mode : false
2023-03-07 22:21:47,655 INFO mapreduce.Job:  map 0% reduce 0%
2023-03-07 22:21:52,741 INFO mapreduce.Job:  map 100% reduce 0%
2023-03-07 22:21:57,781 INFO mapreduce.Job:  map 100% reduce 100%
2023-03-07 22:21:57,797 INFO mapreduce.Job: Job job_1678198135509_0001 completed
 successfully
2023-03-07 22:21:57,913 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=87
                FILE: Number of bytes written=551077
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=187
                HDFS: Number of bytes written=57
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
```

- Output



```
                Peak Reduce Virtual memory (bytes)=2570833920
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=75
        File Output Format Counters
                Bytes Written=57
(base) baobui16@T14:~/Desktop/Lab1$ hadoop dfs -cat /WordCount/Output/*
WARNING: Use of this script to execute dfs is deprecated.
WARNING: Attempting to execute replacement "hdfs dfs" instead.

Beckham 2
Cristiano       1
David   1
Messi   2
Neymar  2
Ronaldo 2
(base) baobui16@T14:~/Desktop/Lab1$
```



Browse Directory

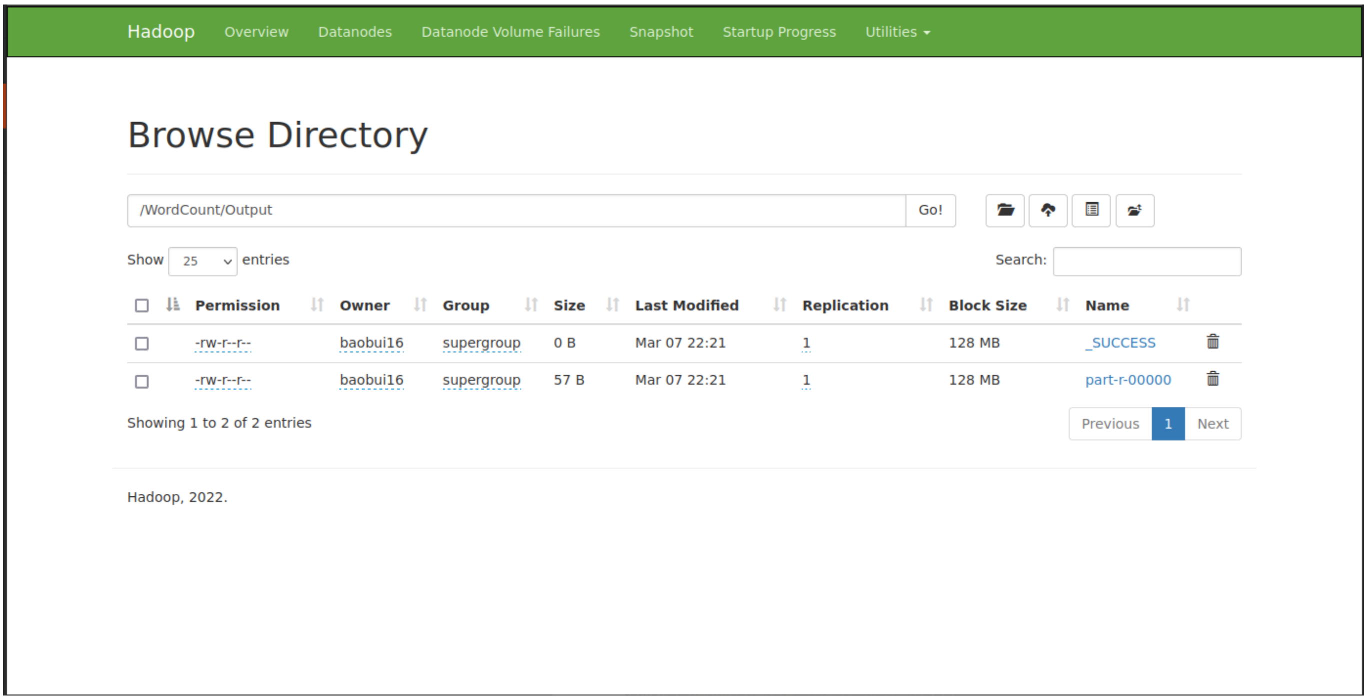
/WordCount Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
□	drwxr-xr-x	baobui16	supergroup	0 B	Mar 07 22:02	0	0 B	Input
□	drwxr-xr-x	baobui16	supergroup	0 B	Mar 07 22:21	0	0 B	Output

Showing 1 to 2 of 2 entries

Hadoop, 2022.



Browse Directory

/WordCount/Output Go!

Show 25 entries Search:

☐	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
☐	-rw-r--r--	baobui16	supergroup	0 B	Mar 07 22:21	1	128 MB	_SUCCESS
☐	-rw-r--r--	baobui16	supergroup	57 B	Mar 07 22:21	1	128 MB	part-r-00000

Showing 1 to 2 of 2 entries

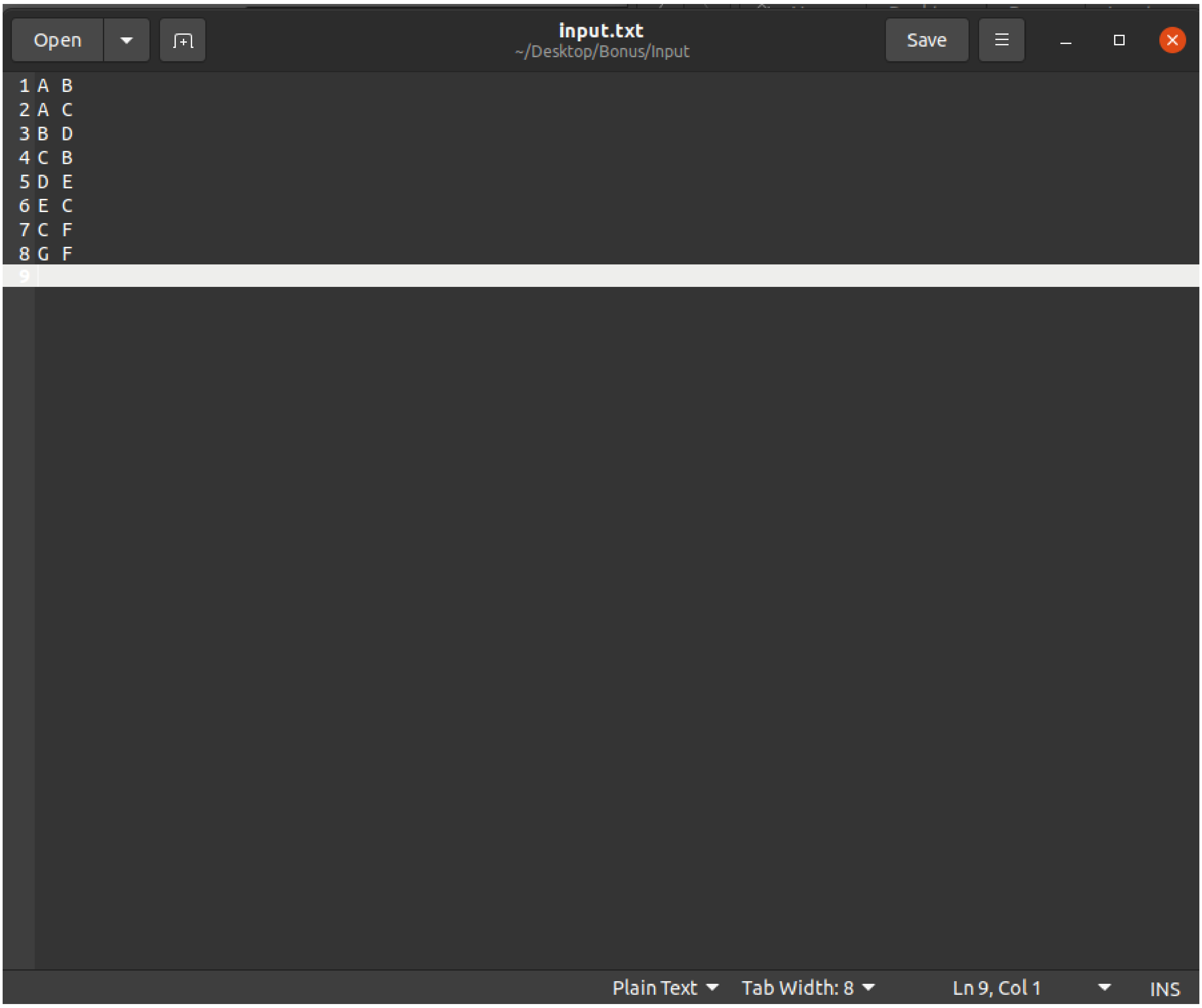
Previous 1 Next

Hadoop, 2022.

Bonus

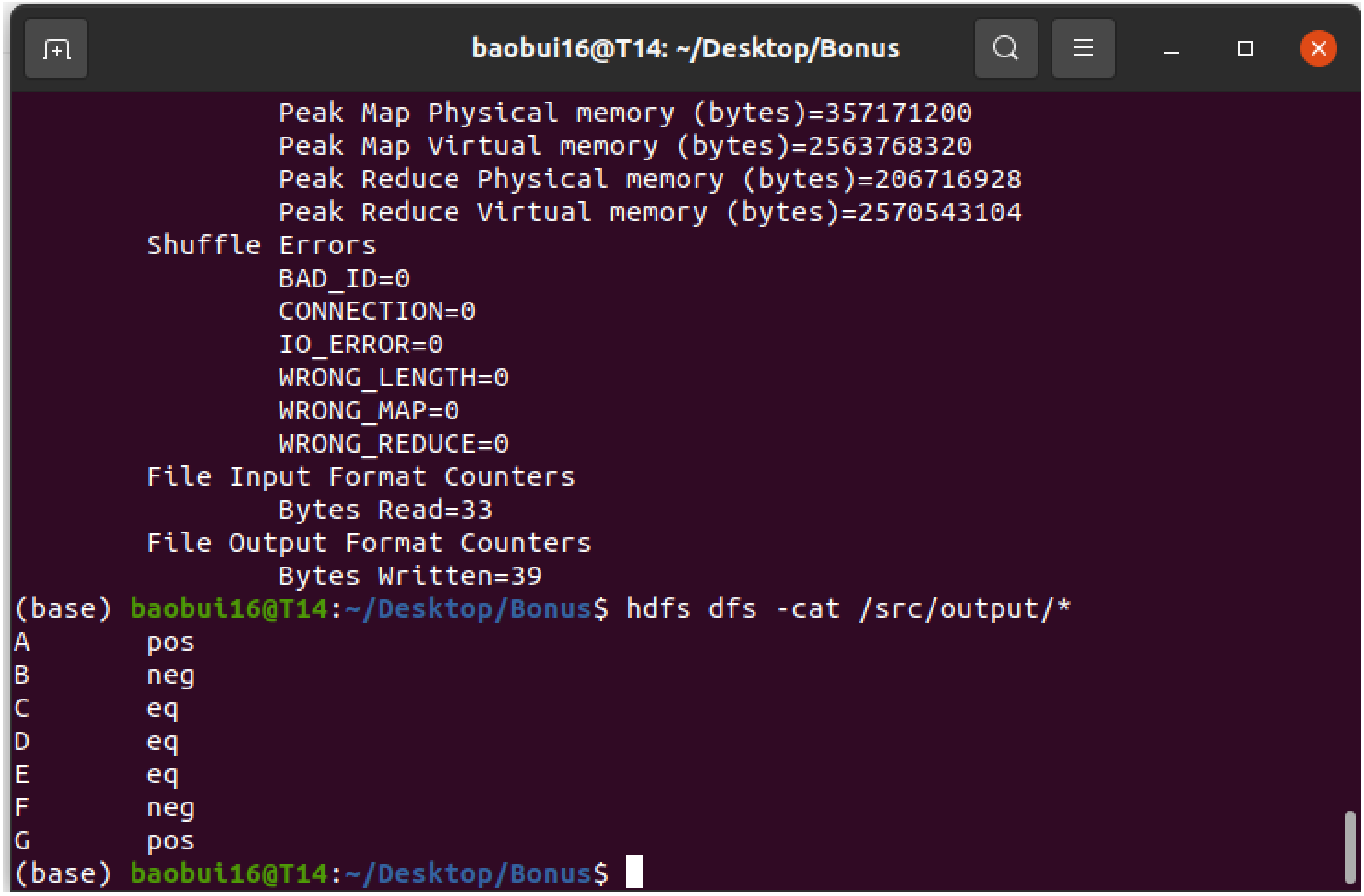
4.1 Extended Word Count: Unhealthy relationships

- Input



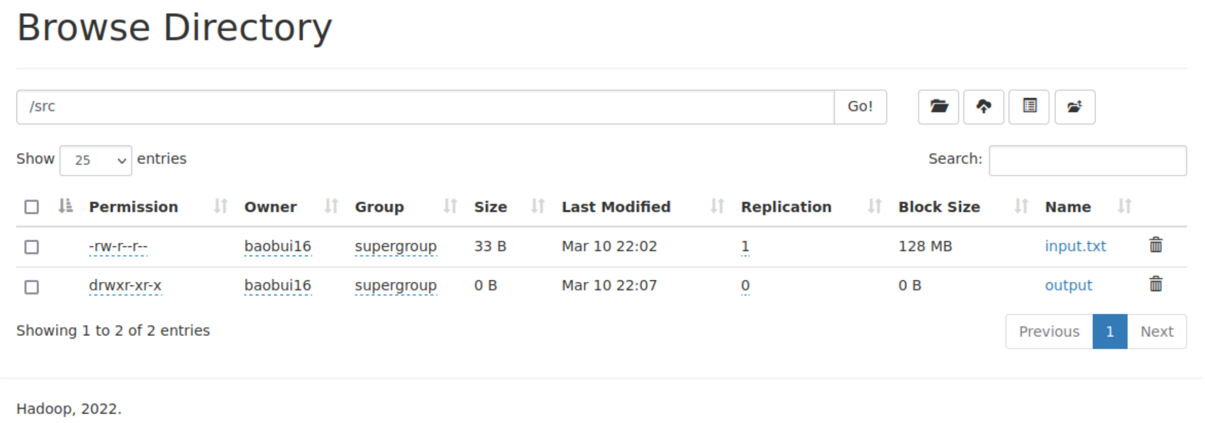
```
A  B
A  C
B  D
C  B
D  E
E  C
C  F
G  F

```

- Output



```
                Peak Map Physical memory (bytes)=357171200
                Peak Map Virtual memory (bytes)=2563768320
                Peak Reduce Physical memory (bytes)=206716928
                Peak Reduce Virtual memory (bytes)=2570543104
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=33
        File Output Format Counters
                Bytes Written=39
(base) baobui16@T14:~/Desktop/Bonus$ hdfs dfs -cat /src/output/*
A       pos
B       neg
C       eq
D       eq
E       eq
F       neg
G       pos
(base) baobui16@T14:~/Desktop/Bonus$
```

- Directory containing output on hadoop



Browse Directory

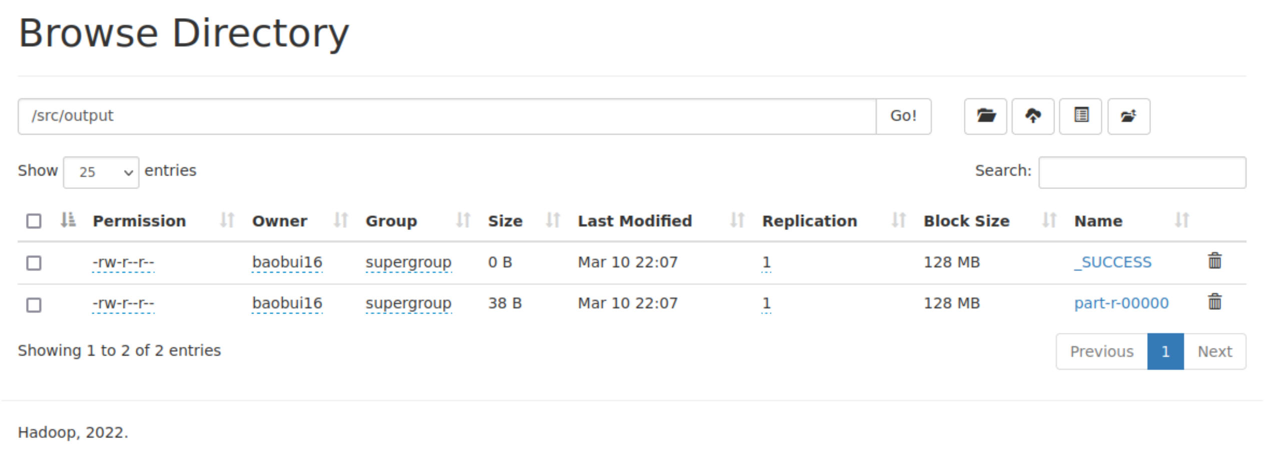
/src Go!

Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	baobui16	supergroup	33 B	Mar 10 22:02	1	128 MB	input.txt	🗑
□	drwxr-xr-x	baobui16	supergroup	0 B	Mar 10 22:07	0	0 B	output	🗑

Showing 1 to 2 of 2 entries

Hadoop, 2022.



Browse Directory

/src/output Go!

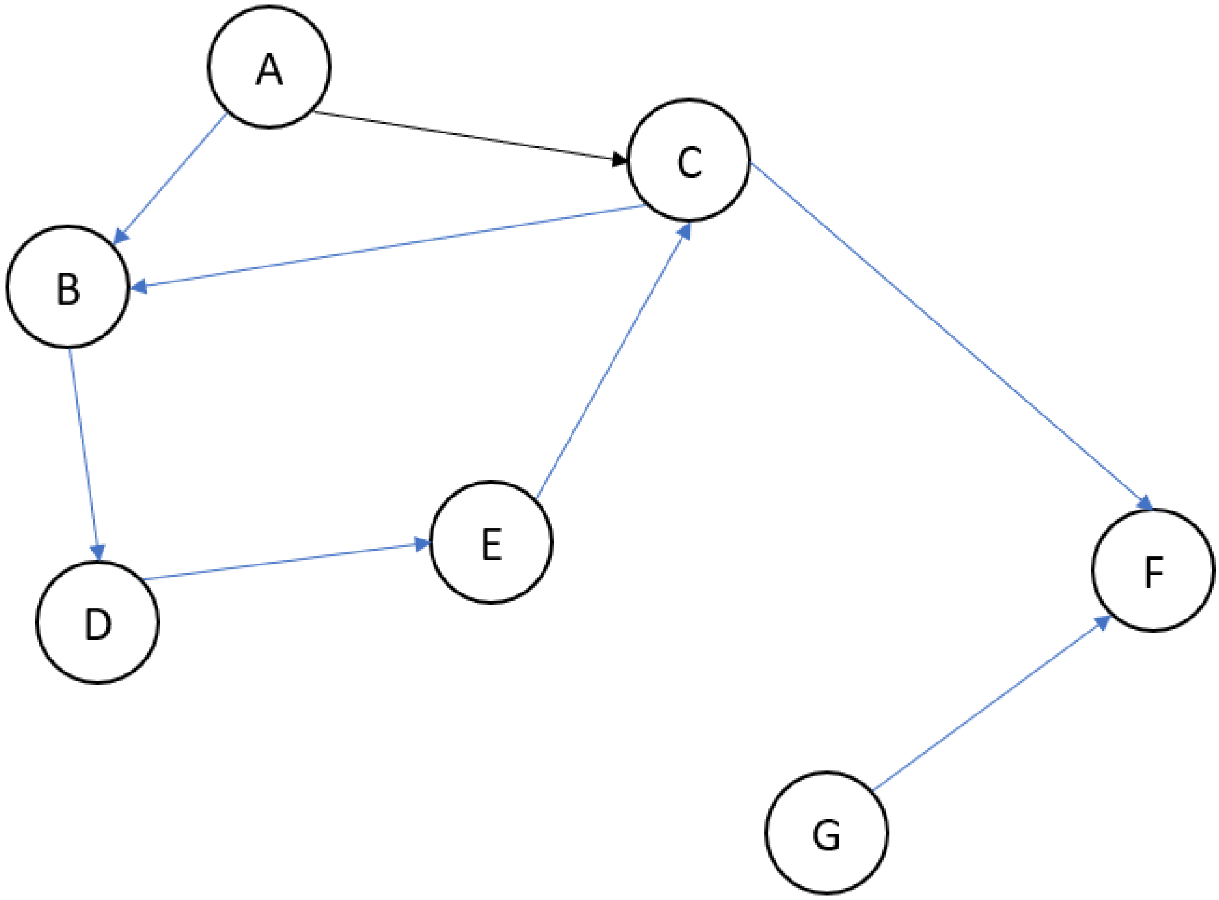
Show 25 entries Search:

□	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	
□	-rw-r--r--	baobui16	supergroup	0 B	Mar 10 22:07	1	128 MB	_SUCCESS	🗑
□	-rw-r--r--	baobui16	supergroup	38 B	Mar 10 22:07	1	128 MB	part-r-00000	🗑

Showing 1 to 2 of 2 entries

Hadoop, 2022.

- Visualize example



Reflection of your team. (Does your journey to the deadline have any bugs? How have you overcome it? What have you learned after this process? If you cannot overcome the bugs, describe where the bottlenecks are in your work.)

- In process , we have some problem about setting up Hadoop in make enviremnt so we need to uninstall ubuntu and resetup after that
- when we run Wordcount and expand wordcount problems , we had error in output of map and input of reduce not correct so we search in google and see solusion by some people
- After this lab , we can operate MapReduce and custom MapReduce in high level.
- We try to install full hadoop distribution mode but can't.

References

- Two Cloudera version of WordCount problem:
    - https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount2.html
    - https://docs.cloudera.com/documentation/other/tutorial/CDH5/topics/ht_wordcount3.html

- Apache Hadoop
    - [1] Apache Hadoop 3.3.4 – Hadoop Cluster Setup. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/ClusterSetup.html.

    - [2] Apache Hadoop 3.3.4 – Hadoop in Secure Mode. https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/SecureMode.html.

    - [3] Apache Hadoop 3.3.4 – Hadoop: Setting up a Single Node Cluster. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html.

    - [4] Apache Hadoop 3.3.4 – MapReduce Tutorial https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html#Example:_WordCount_v1.0.

- Book:
    - MapReduce Design Patterns [Donald Miner, Adam Shook, 2012]
    - [5] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In OSDI'04: Sixth Symposium on Operating System Design and Implementation, pages 137–150, San Francisco, CA, 2004.

- All of StackOverflow link related.
    - https://stackoverflow.com/questions/56153007/java-lang-exception-java-io-ioexception-wrong-value-class-while-setting-hadoop?fbclid=IwAR1UNjMqPVvXy9nMberqIHAYFNZ02q0we1BAoqZOGUhmEFArVmbnBfyoCiQ

    - https://stackoverflow.com/questions/58272650/what-exactly-does-data-locality-mean-in-hadoop?fbclid=IwAR0jSqFAXazrxBUA7uO8sZsGWbhkry6SYJAf4PtBEbNfilKoZcl6HTBbdLk

- Another link

- https://community.cloudera.com/t5/Support-Questions/class-org-apache-hadoop-io-MapWritable-is-not-class-org/m-p/136602?fbclid=IwAR3VazsXyHo688u4UyNvsBdj_a_2hOwuJ-KpGZD6fg4aMwW8trQlpz-KArA
- https://www.thoughtworks.com/insights/decoder/d/data-locality?fbclid=IwAR1BwoHh0Q8-hylq51ylms9NIWQI300GMKdZm4yydbQkEDBWhrbcqQoM554
- https://www.tutorialspoint.com/map_reduce/map_reduce_quick_guide.html
- https://www.youtube.com/watch?v=6sK3LDY7Pp4&t=483s
- https://www.youtube.com/watch?v=MZ_FUEnbrR4