Rapport : installation et configuration d’apache Hadoop et exécution d’un programme mapreduce à nœud unique et à nœuds multiples.

Filière : DATA INE1

Binôme : Mouna Ali & Soumane Khaoula

---

configuration de Hadoop dans un cluster à nœud unique.

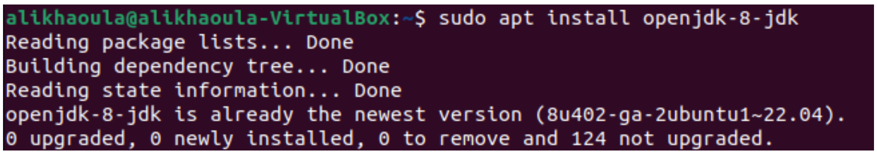
Installation d’Ubuntu dans une machine virtuelle :

1. télécharger et installer un logiciel de virtualisation.
2. télécharger le fichier iso d’Ubuntu depuis le site officiel.
3. créer une nouvelle machine virtuelle.
4. démarrer la machine virtuelle avec le fichier iso d’Ubuntu.
5. suivre l’assistant d’installation.
6. configurer les préférences, partitionner le disque, créer un utilisateur.
7. terminer l’installation et redémarrer.
8. installer les mises à jour et les pilotes.
9. Ubuntu est prêt à être utilisé!

Configuration de Hadoop :

1) installer java jdk 8.

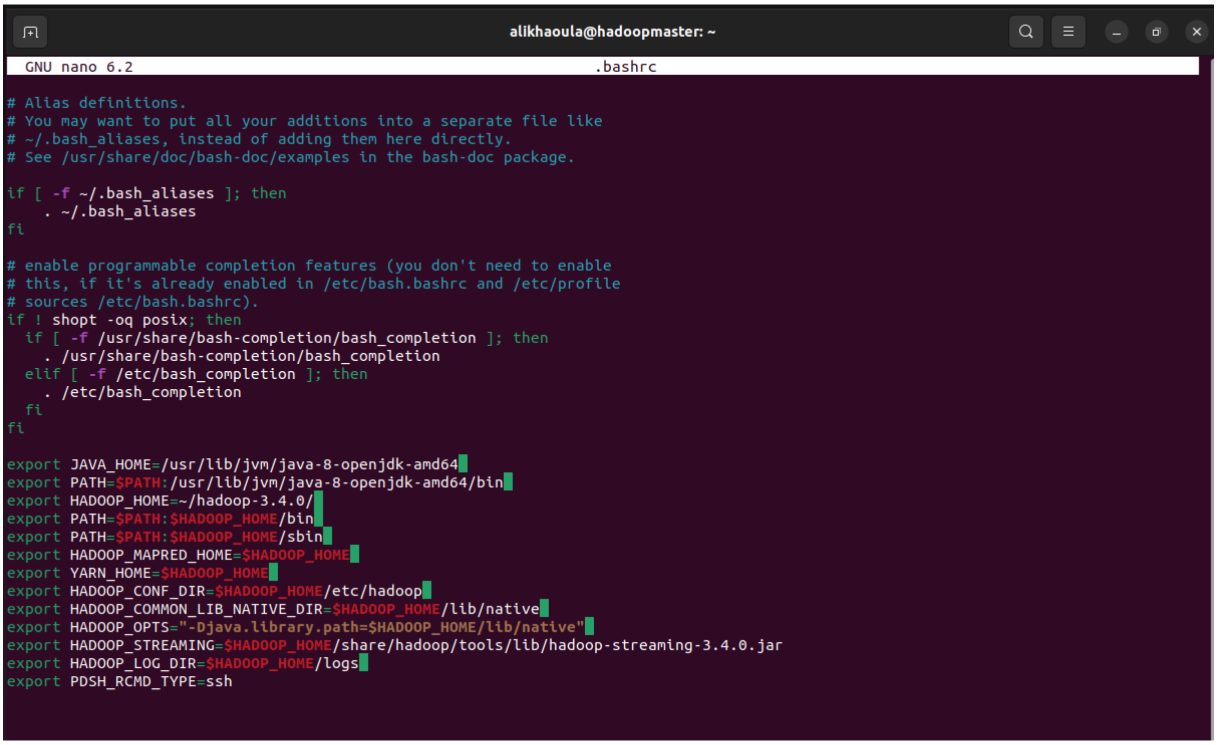
```
sudo apt install openjdk-8-jdk
```



```
alikhaoula@alikhaoula-VirtualBox:~$ sudo apt install openjdk-8-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openjdk-8-jdk is already the newest version (8u402-ga-2ubuntu1~22.04).
0 upgraded, 0 newly installed, 0 to remove and 124 not upgraded.
```

2)ajouter les commandes suivante au fichier . bashrc :

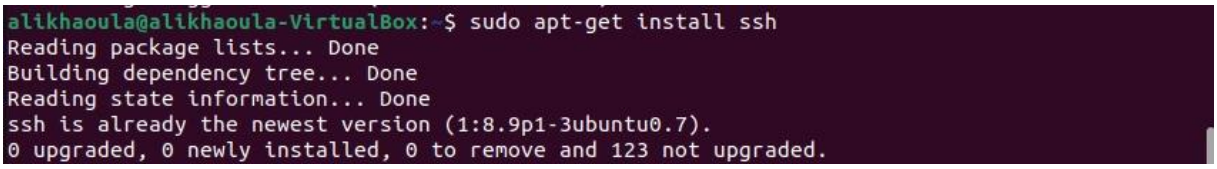
sudo nano .bashrc



```
  GNU nano 6.2                                        .bashrc

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

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

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

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin
export HADOOP_HOME=~/hadoop-3.4.0/
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$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_STREAMING=$HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming-3.4.0.jar
export HADOOP_LOG_DIR=$HADOOP_HOME/logs
export PDSH_RCMD_TYPE=ssh
```

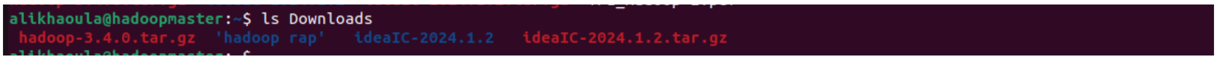
3)installer ssh:

sudo apt - get install ssh



```
alikhaoula@alikhaoula-VirtualBox:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
ssh is already the newest version (1:8.9p1-3ubuntu0.7).
0 upgraded, 0 newly installed, 0 to remove and 123 not upgraded.
```

4) télécharger le fichier tar de Hadoop sur le site : apache.hadoop.org



```
alikhaoula@hadoopmaster:~$ ls Downloads
 hadoop-3.4.0.tar.gz  'hadoop rap'   ideaIC-2024.1.2   ideaIC-2024.1.2.tar.gz
alikhaoula@hadoopmaster:~$
```

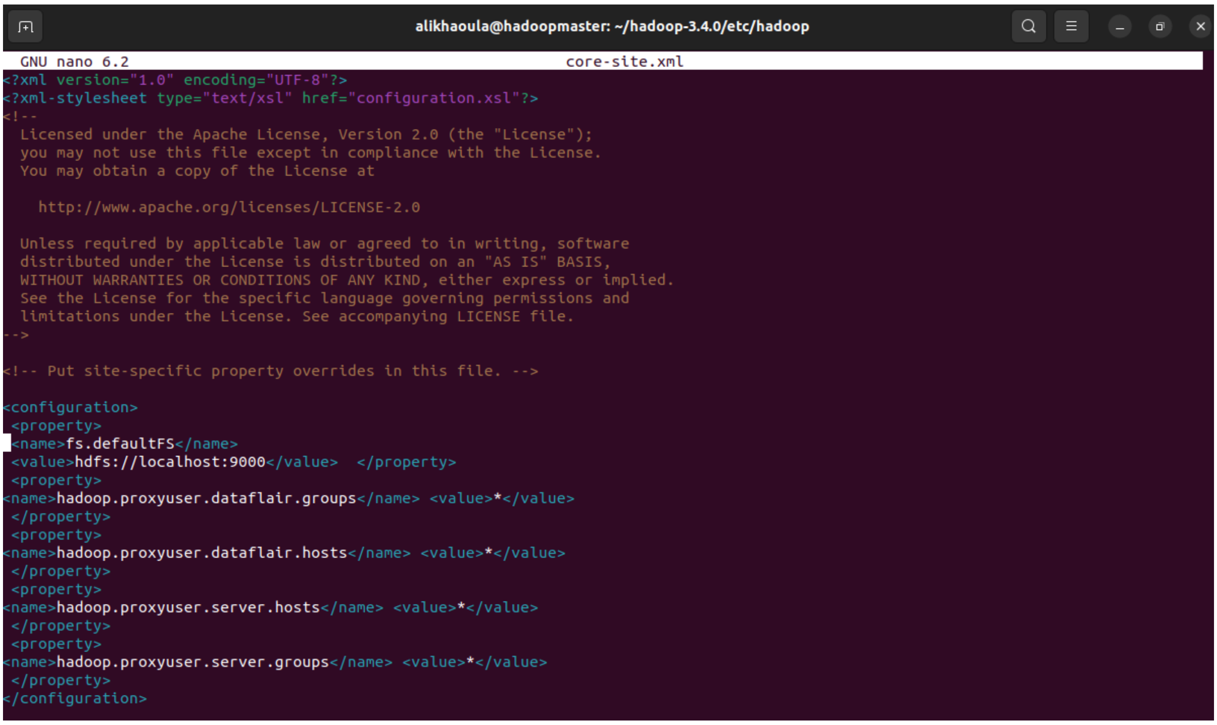
5) extraire le fichier tar

6) ouvrir le fichier hadoop-env.sh aui se trouve dans hadoop-3.4.0/etc/hadoop et définir le chemin suivant pour java_home :

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

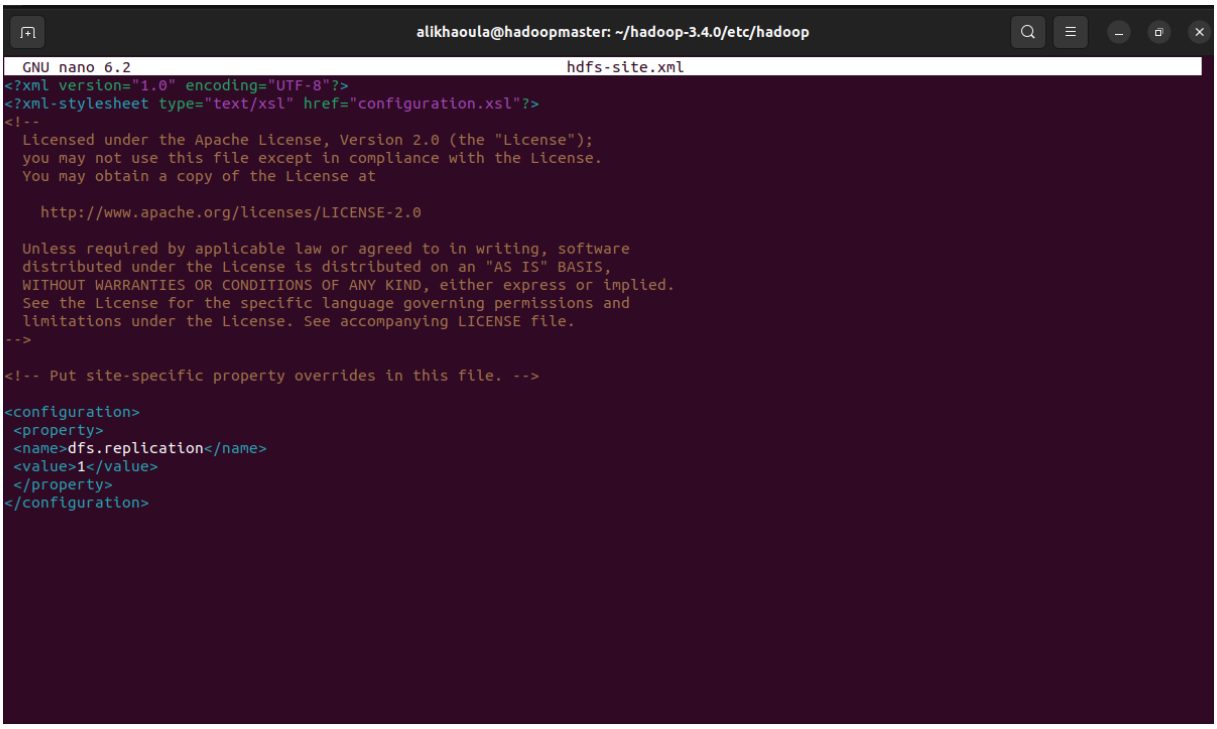
7) configure les fichier xml:

sudo nano core-site.xml



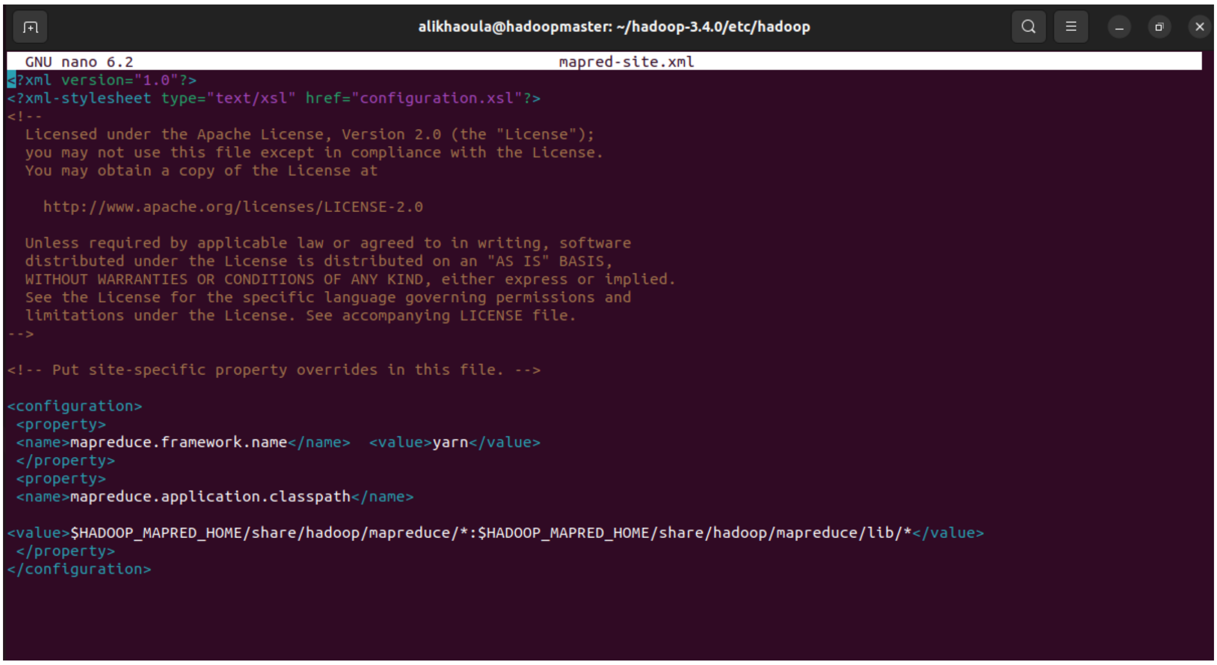
```
  GNU nano 6.2                                         core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

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

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

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

<configuration>
 <property>
 <name>fs.defaultFS</name>
 <value>hdfs://localhost:9000</value>  </property>
 <property>
<name>hadoop.proxyuser.dataflair.groups</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.dataflair.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.hosts</name> <value>*</value>
 </property>
 <property>
<name>hadoop.proxyuser.server.groups</name> <value>*</value>
 </property>
</configuration>
```

sudo nano hdfs-site.xml



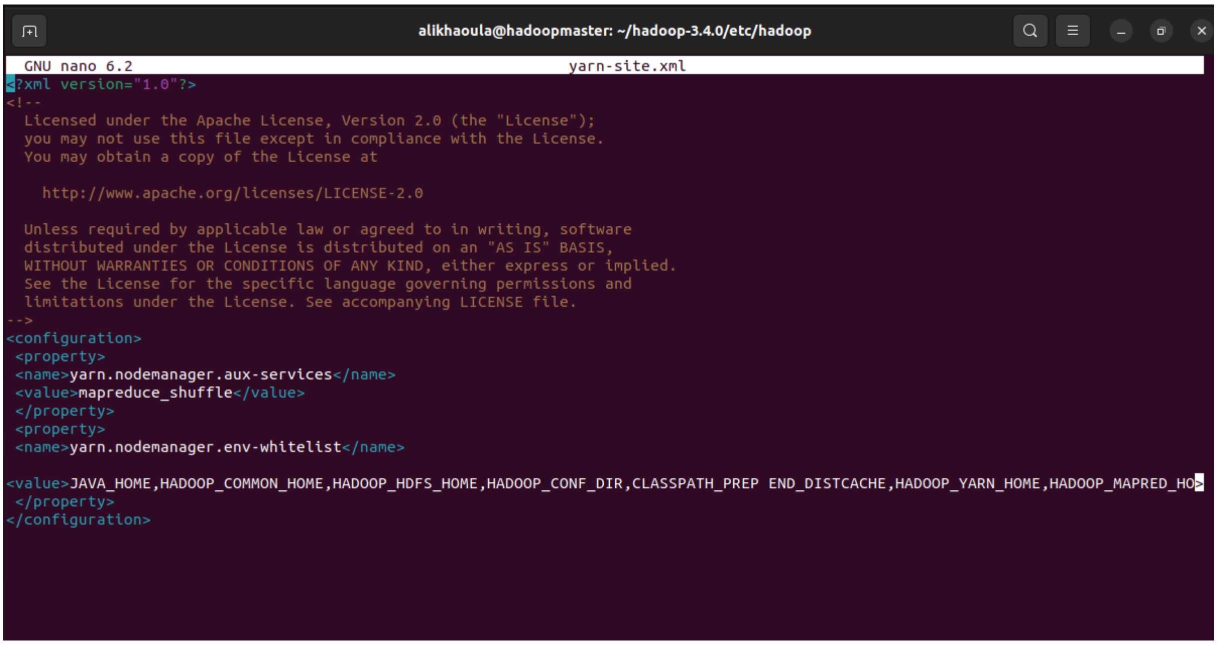
```
  GNU nano 6.2                                    hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  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>
</configuration>
```

sudo nano mapred-site.xml



```
  GNU nano 6.2                                          mapred-site.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>
 <property>
 <name>mapreduce.application.classpath</name>

<value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>
 </property>
</configuration>
```

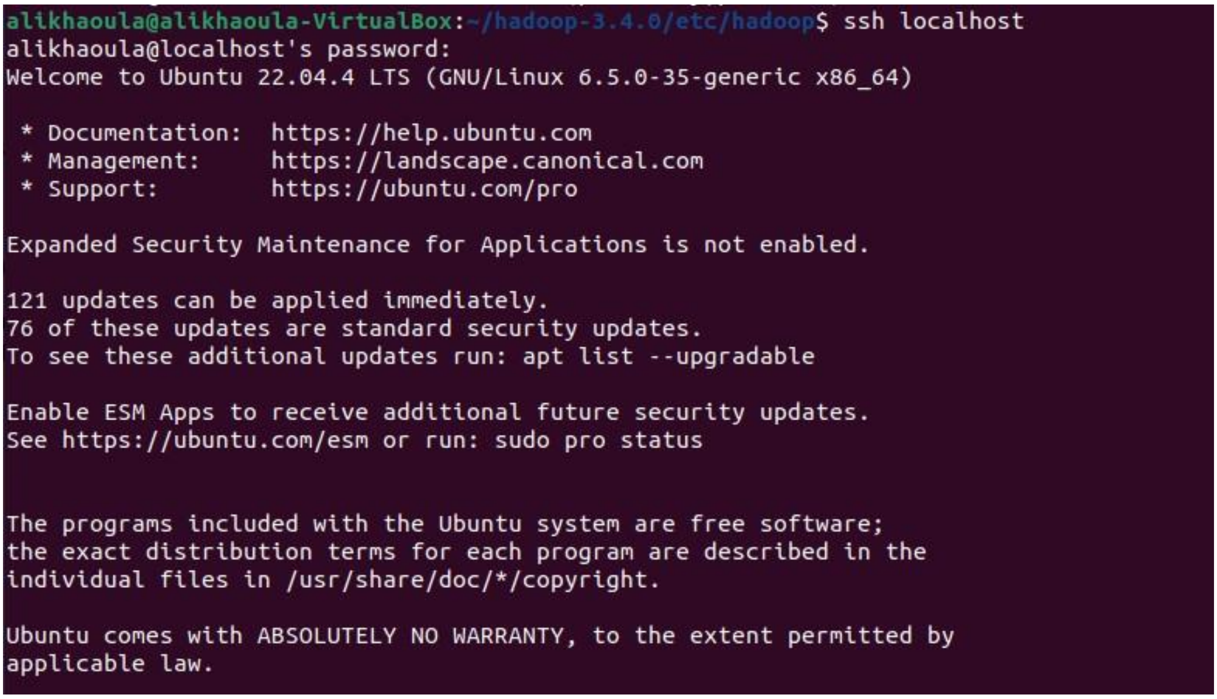
sudo nano yarn-site.xml



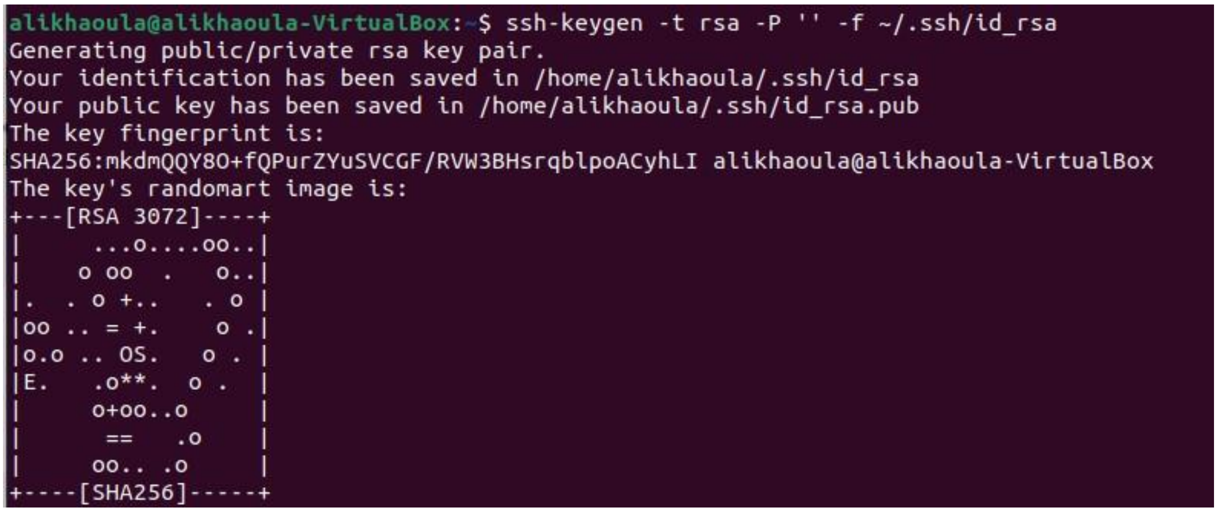
```
  GNU nano 6.2                                           yarn-site.xml
<?xml version="1.0"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

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

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

8) ssh:

ssh localhost
ssh-keygen -t rsa -p '' -f ~/.ssh/id_rsa
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
chmod 0600 ~/.ssh/authorized_keys
hadoop-3.4.0/bin/hdfs namenode -format



```
alikhaoula@alikhaoula-VirtualBox:~/hadoop-3.4.0/etc/hadoop$ ssh localhost
alikhaoula@localhost's password:
Welcome to Ubuntu 22.04.4 LTS (GNU/Linux 6.5.0-35-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.

121 updates can be applied immediately.
76 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.
```

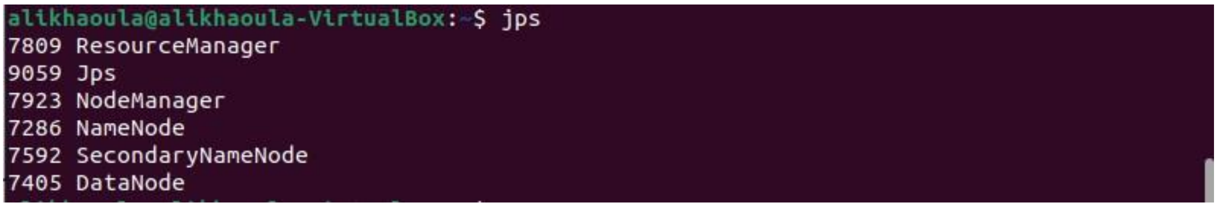


```
alikhaoula@alikhaoula-VirtualBox:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Your identification has been saved in /home/alikhaoula/.ssh/id_rsa
Your public key has been saved in /home/alikhaoula/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:mkdmQQY8O+fQPurZYuSVCGF/RVW3BHsrqblpoACyhLI alikhaoula@alikhaoula-VirtualBox
The key's randomart image is:
+---[RSA 3072]----+
|     ...o....oo..|
|    o oo  .   o..|
|.  . o +..   . o |
|oo .. = +.    o .|
|o.o .. OS.   o . |
|E.   .o**.  o .  |
|     o+oo..o     |
|      ==   .o    |
|     oo.. .o     |
+----[SHA256]-----+
```

9) formater le fichier système:

export pdsh_rcmd_type=ssh

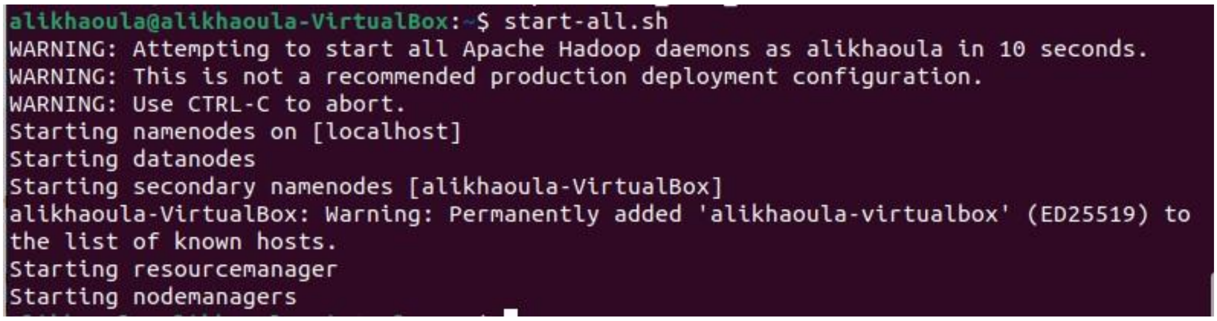
10) verifier si la configuration est bien faite :



```
alikhaoula@alikhaoula-VirtualBox:~$ jps
7809 ResourceManager
9059 Jps
7923 NodeManager
7286 NameNode
7592 SecondaryNameNode
7405 DataNode
```

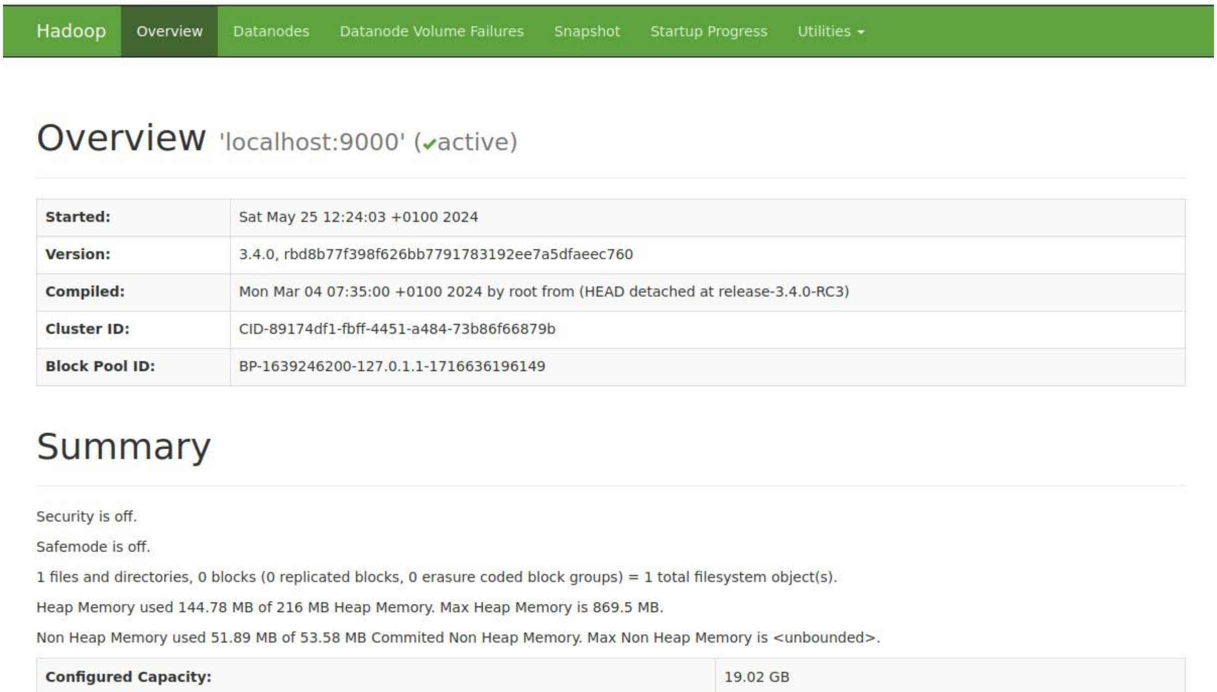
11) démarrer Hadoop:

start-all.sh



```
alikhaoula@alikhaoula-VirtualBox:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as alikhaoula 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 [alikhaoula-VirtualBox]
alikhaoula-VirtualBox: Warning: Permanently added 'alikhaoula-virtualbox' (ED25519) to
the list of known hosts.
Starting resourcemanager
Starting nodemanagers
```

localhost :9870/ :



Overview 'localhost:9000' (✓active)

Started:	Sat May 25 12:24:03 +0100 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled:	Mon Mar 04 07:35:00 +0100 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-89174df1-fbff-4451-a484-73b86f66879b
Block Pool ID:	BP-1639246200-127.0.1.1-1716636196149

Summary

Security is off.

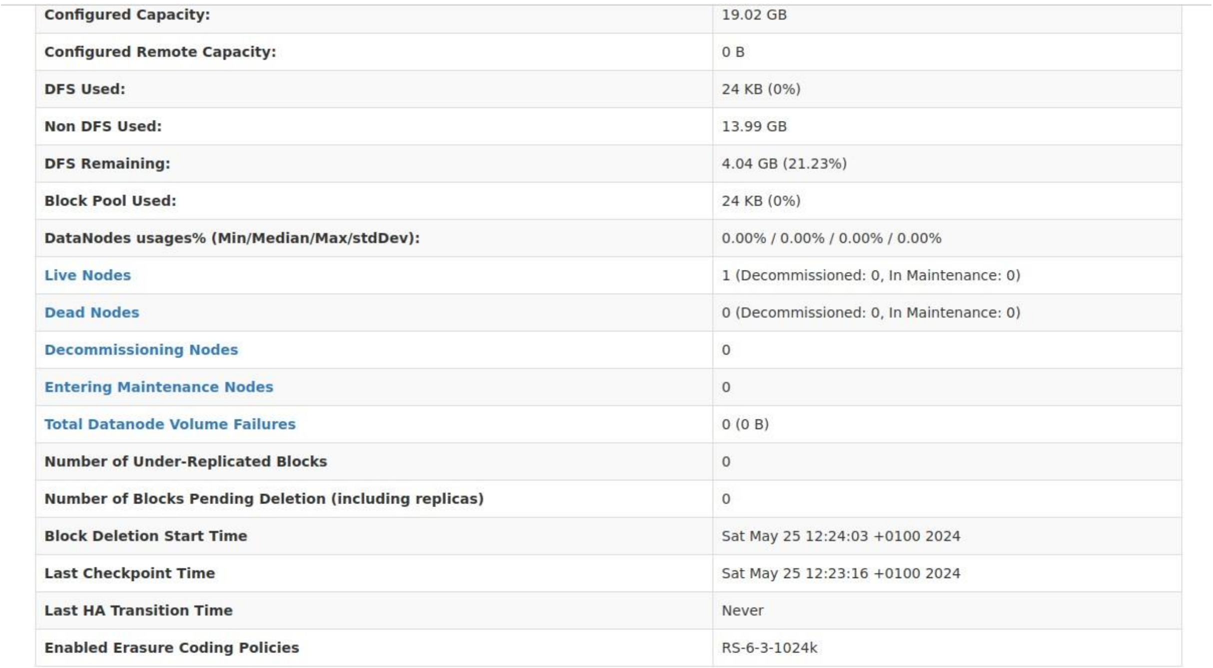
Safemode is off.

1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

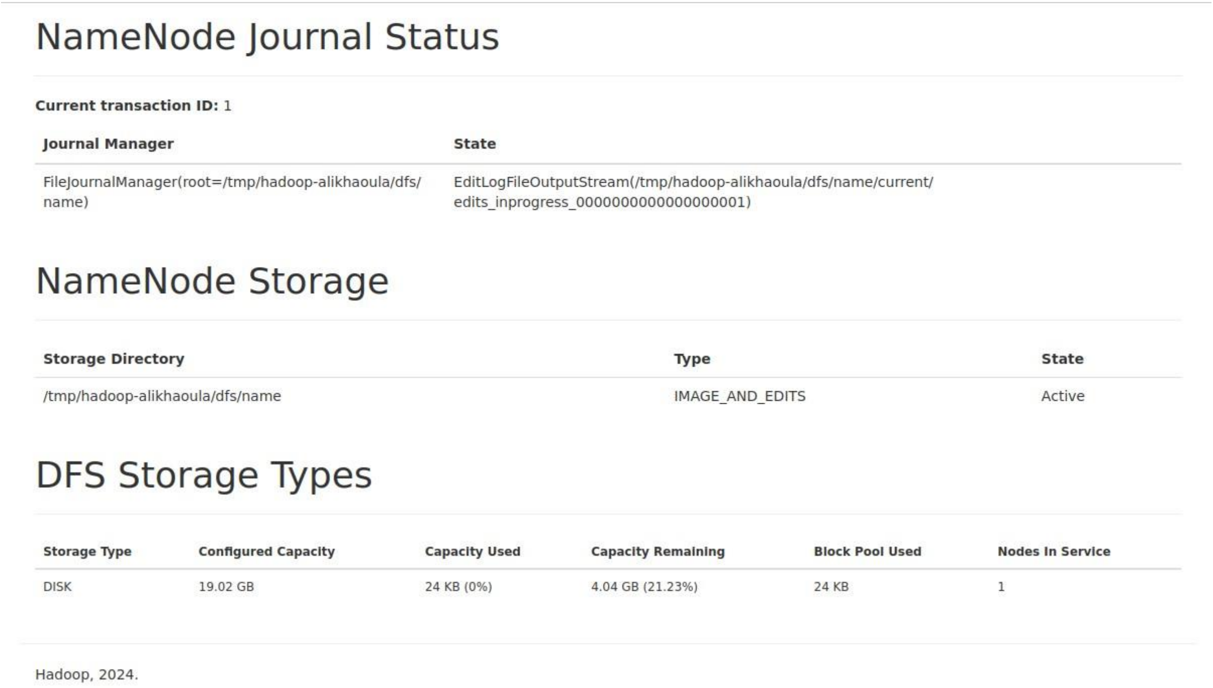
Heap Memory used 144.78 MB of 216 MB Heap Memory. Max Heap Memory is 869.5 MB.

Non Heap Memory used 51.89 MB of 53.58 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	19.02 GB



Configured Capacity:	19.02 GB
Configured Remote Capacity:	0 B
DFS Used:	24 KB (0%)
Non DFS Used:	13.99 GB
DFS Remaining:	4.04 GB (21.23%)
Block Pool Used:	24 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Sat May 25 12:24:03 +0100 2024
Last Checkpoint Time	Sat May 25 12:23:16 +0100 2024
Last HA Transition Time	Never
Enabled Erasure Coding Policies	RS-6-3-1024k



NameNode Journal Status

Current transaction ID: 1

Journal Manager	State
FileJournalManager(root=/tmp/hadoop-alikhaoula/dfs/name)	EditLogFileOutputStream(/tmp/hadoop-alikhaoula/dfs/name/current/edits_inprogress_0000000000000000001)

NameNode Storage

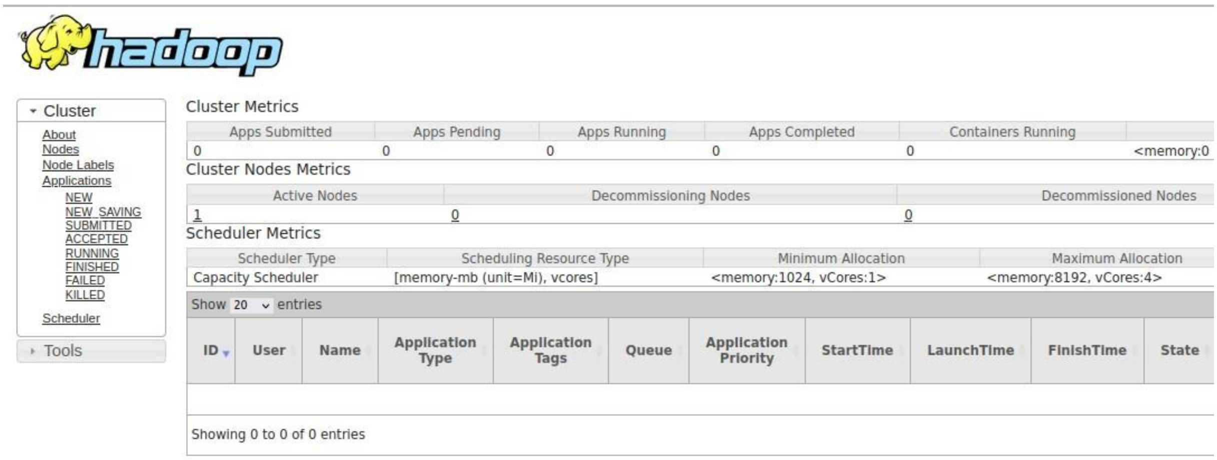
Storage Directory	Type	State
/tmp/hadoop-alikhaoula/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	19.02 GB	24 KB (0%)	4.04 GB (21.23%)	24 KB	1

Hadoop, 2024.

localhost:8088/



Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	
0	0	0	0	0	<memory:0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

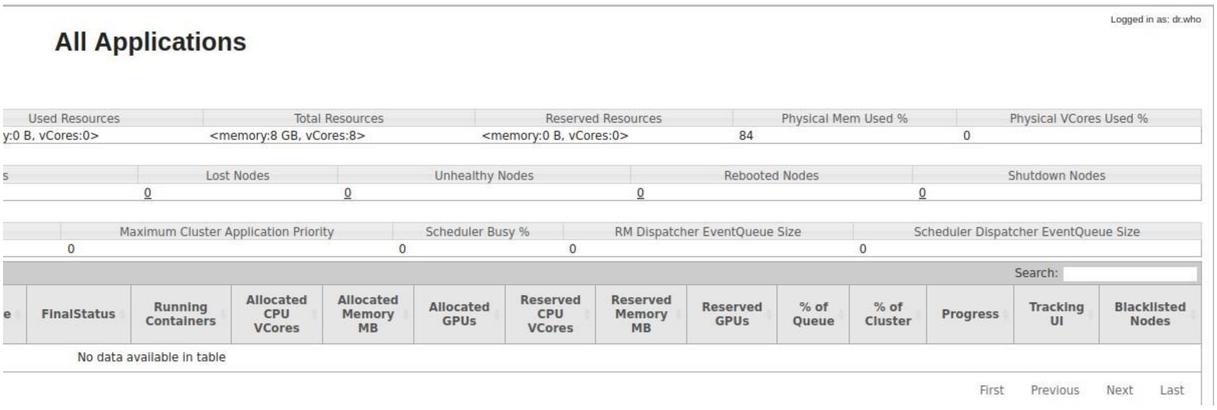
Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>

Show 20 entries

ID	User	Name	Application Type	Application Tags	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State

Showing 0 to 0 of 0 entries



Logged in as: dr.who

All Applications

Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
y:0 B, vCores:0>	<memory:8 GB, vCores:8>	<memory:0 B, vCores:0>	84	0

s	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
	0	0	0	0

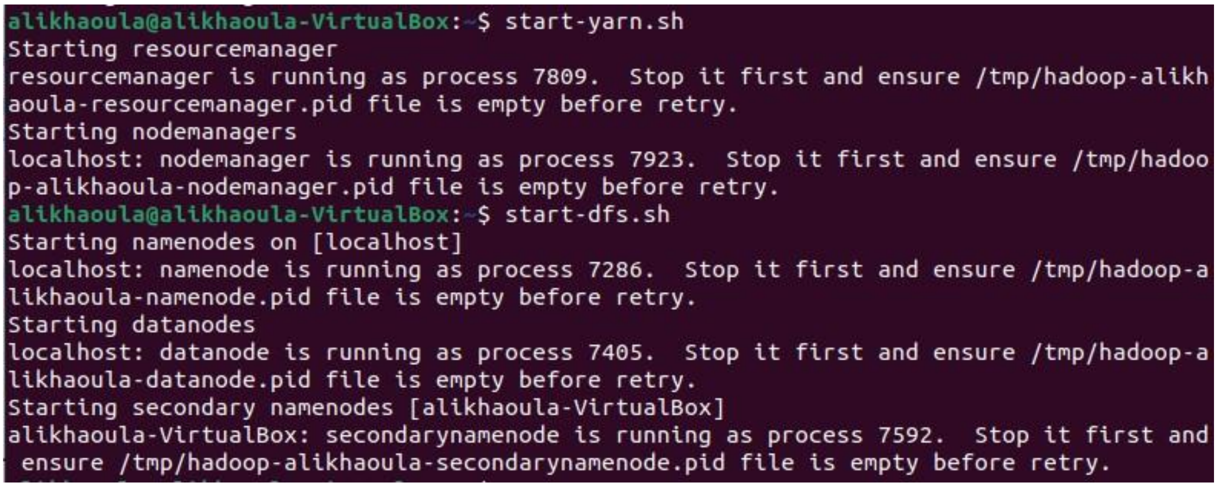
	Maximum Cluster Application Priority	Scheduler Busy %	RM Dispatcher EventQueue Size	Scheduler Dispatcher EventQueue Size
	0	0	0	0

Search:

e	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Allocated GPUs	Reserved CPU VCores	Reserved Memory MB	Reserved GPUs	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
	No data available in table												

start-yarn.sh

start-hdfs.sh

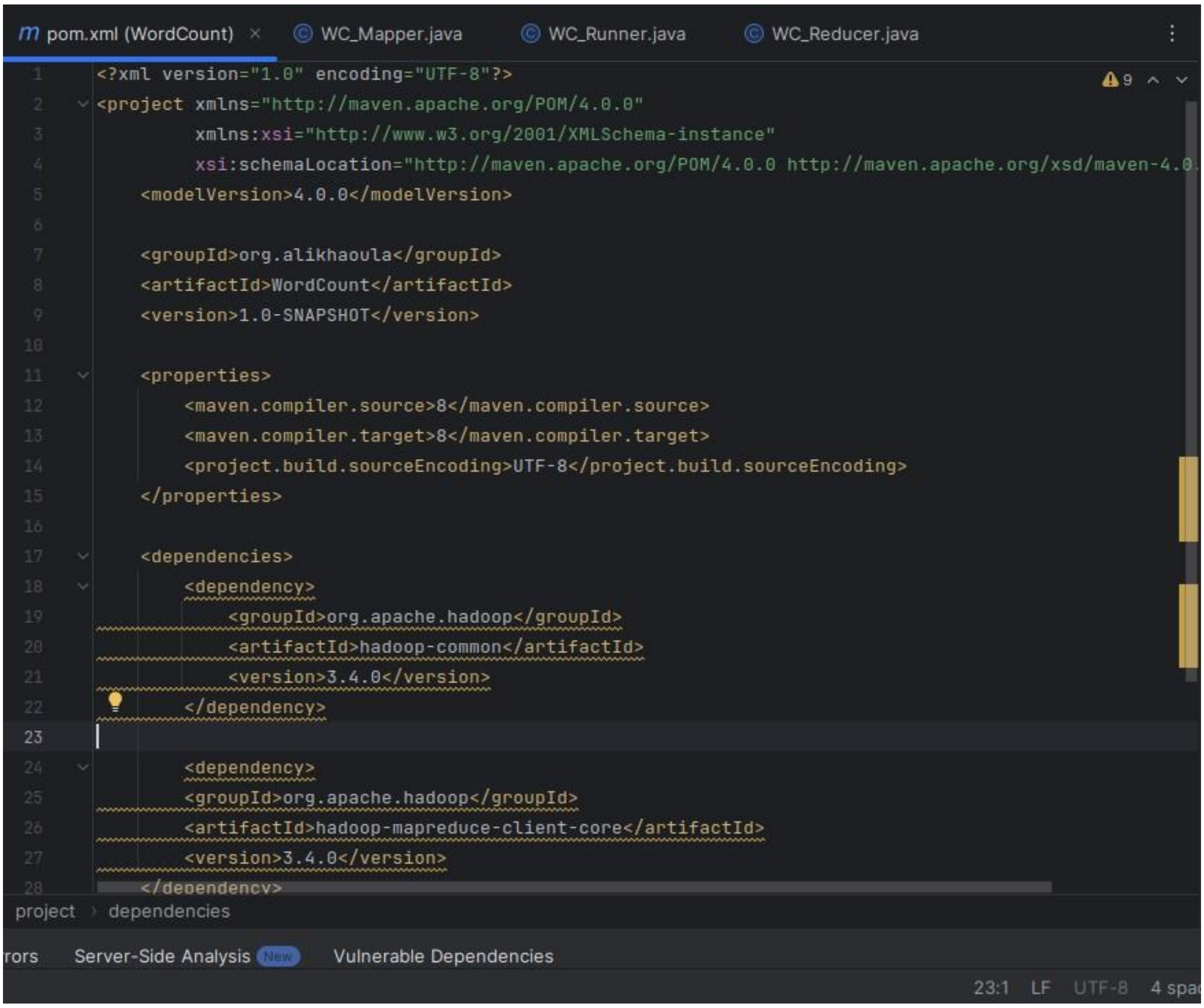


```
alikhaoula@alikhaoula-VirtualBox:~$ start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 7809.  Stop it first and ensure /tmp/hadoop-alikh
aoula-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 7923.  Stop it first and ensure /tmp/hadoo
p-alikhaoula-nodemanager.pid file is empty before retry.
alikhaoula@alikhaoula-VirtualBox:~$ start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 7286.  Stop it first and ensure /tmp/hadoop-a
likhaoula-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 7405.  Stop it first and ensure /tmp/hadoop-a
likhaoula-datanode.pid file is empty before retry.
Starting secondary namenodes [alikhaoula-VirtualBox]
alikhaoula-VirtualBox: secondarynamenode is running as process 7592.  Stop it first and
 ensure /tmp/hadoop-alikhaoula-secondarynamenode.pid file is empty before retry.
```

Programme MapReduce: WordCount.

1) Télécharger IntelliJ idea.
2) Créer un nouveau projet Maven.

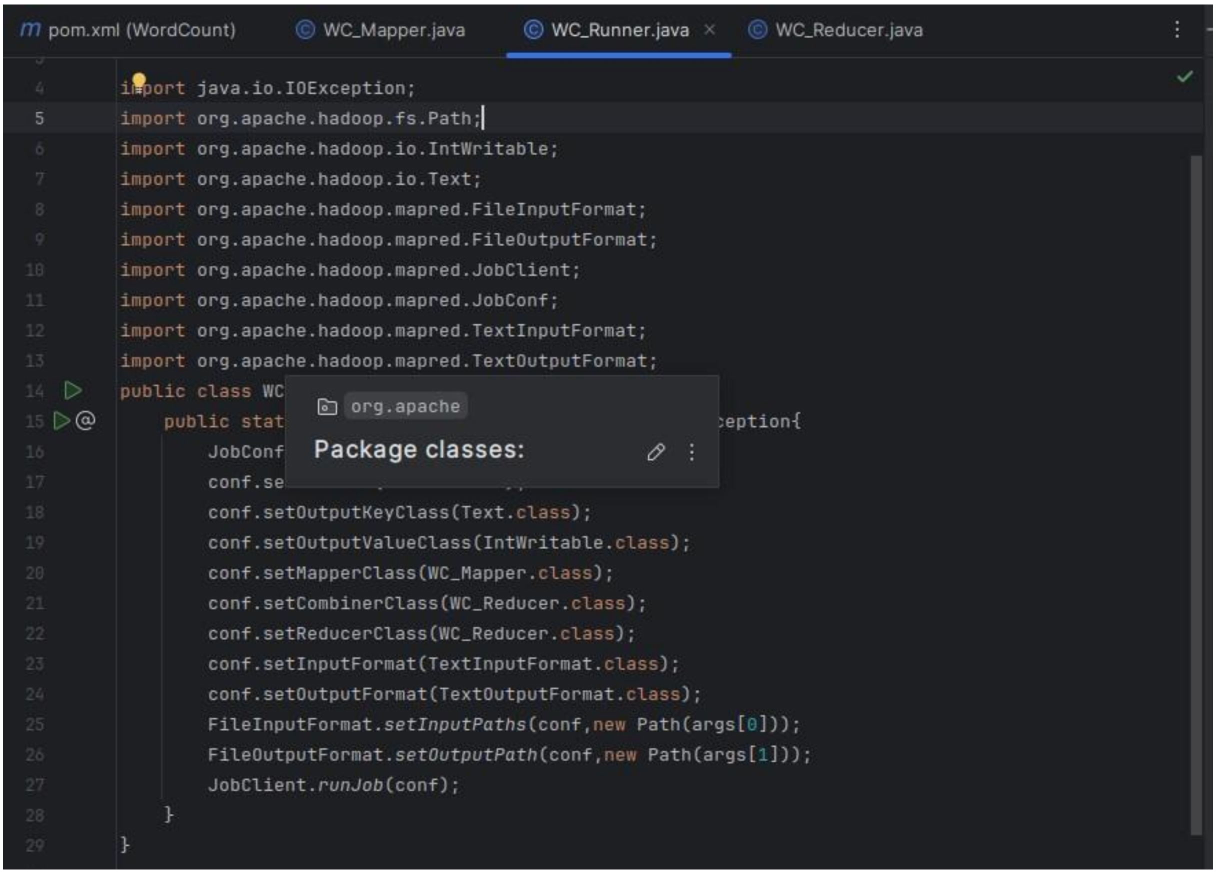
3) Ajouter les dependencies



```
<?xml version="1.0" encoding="UTF-8"?>
<project xmlns="http://maven.apache.org/POM/4.0.0"
         xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
         xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xsd/maven-4.0.
    <modelVersion>4.0.0</modelVersion>

    <groupId>org.alikhaoula</groupId>
    <artifactId>WordCount</artifactId>
    <version>1.0-SNAPSHOT</version>

    <properties>
        <maven.compiler.source>8</maven.compiler.source>
        <maven.compiler.target>8</maven.compiler.target>
        <project.build.sourceEncoding>UTF-8</project.build.sourceEncoding>
    </properties>

    <dependencies>
        <dependency>
            <groupId>org.apache.hadoop</groupId>
            <artifactId>hadoop-common</artifactId>
            <version>3.4.0</version>
        </dependency>

        <dependency>
        <groupId>org.apache.hadoop</groupId>
        <artifactId>hadoop-mapreduce-client-core</artifactId>
        <version>3.4.0</version>
    </dependency>
```

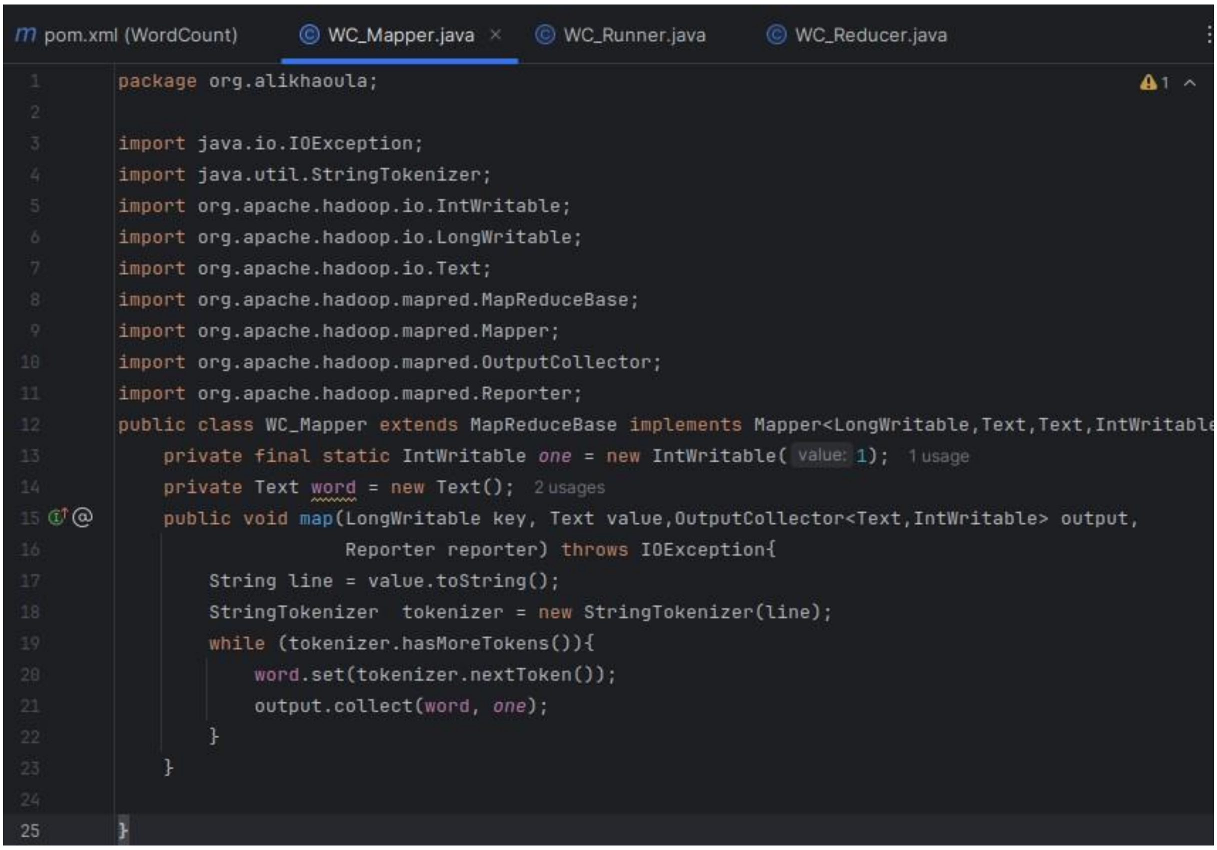
4) Créer les classes nécessaires :

- Le Runner :



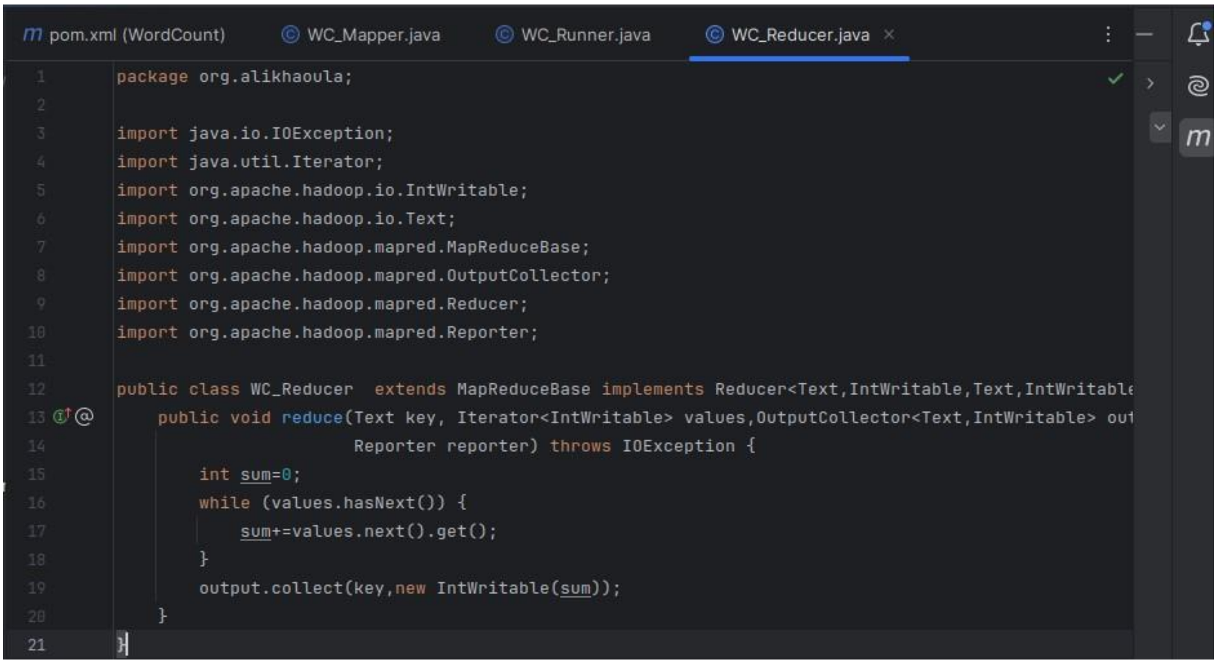
```
import java.io.IOException;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.TextOutputFormat;
public class WC
    public stat                                          ception{
        JobConf
        conf.se
        conf.setOutputKeyClass(Text.class);
        conf.setOutputValueClass(IntWritable.class);
        conf.setMapperClass(WC_Mapper.class);
        conf.setCombinerClass(WC_Reducer.class);
        conf.setReducerClass(WC_Reducer.class);
        conf.setInputFormat(TextInputFormat.class);
        conf.setOutputFormat(TextOutputFormat.class);
        FileInputFormat.setInputPaths(conf,new Path(args[0]));
        FileOutputFormat.setOutputPath(conf,new Path(args[1]));
        JobClient.runJob(conf);
    }
}
```

- Le Mapper :



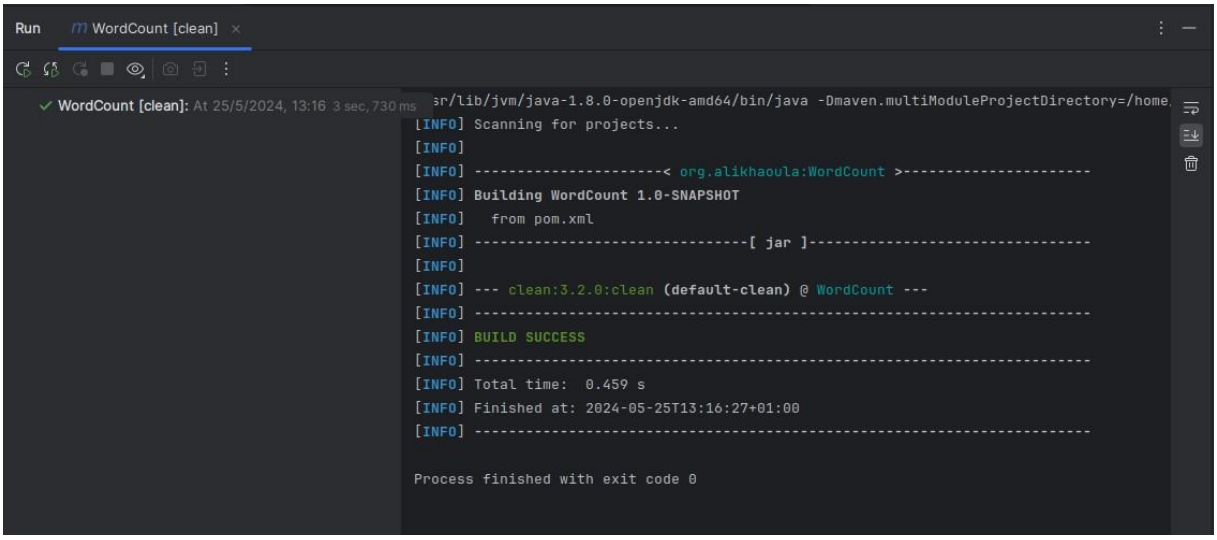
```java
package org.alikhaoula;

import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reporter;
public class WC_Mapper extends MapReduceBase implements Mapper<LongWritable,Text,Text,IntWritable
    private final static IntWritable one = new IntWritable(1);
    private Text word = new Text();
    public void map(LongWritable key, Text value,OutputCollector<Text,IntWritable> output,
                    Reporter reporter) throws IOException{
        String line = value.toString();
        StringTokenizer  tokenizer = new StringTokenizer(line);
        while (tokenizer.hasMoreTokens()){
            word.set(tokenizer.nextToken());
            output.collect(word, one);
        }
    }

}
```

- Le Reducer :



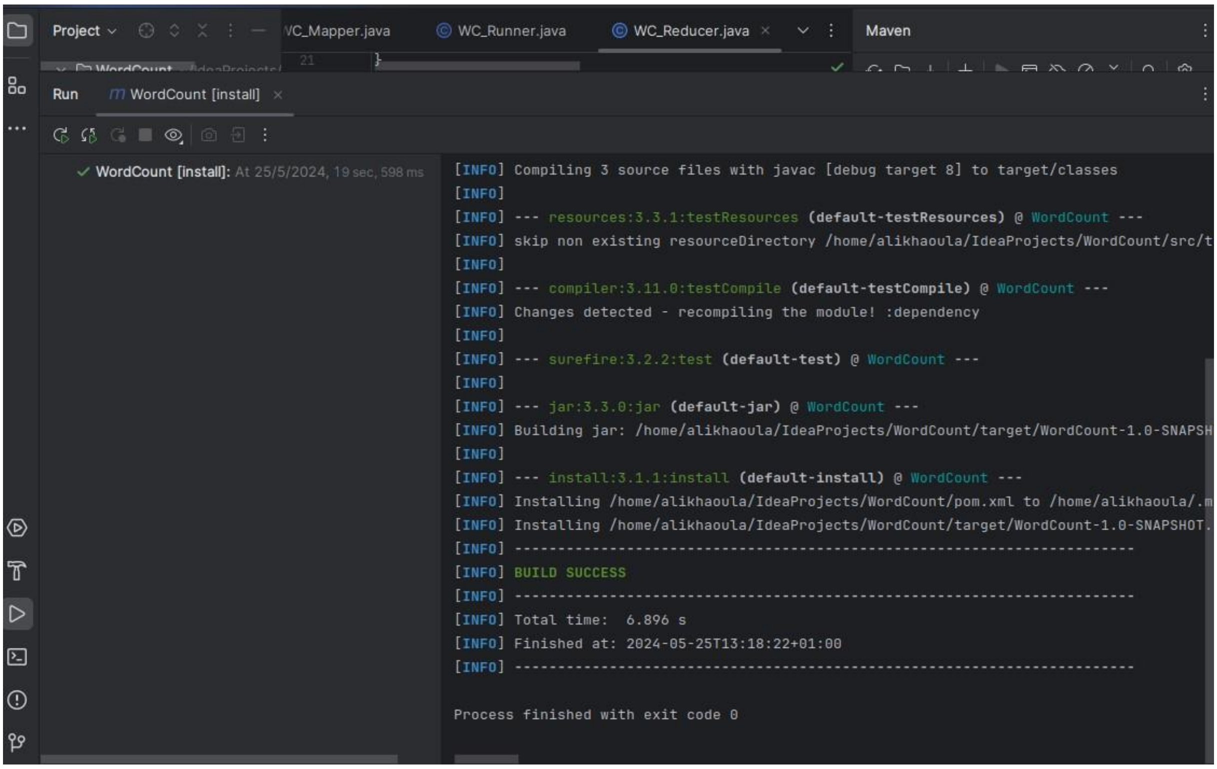
```java
package org.alikhaoula;

import java.io.IOException;
import java.util.Iterator;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;

public class WC_Reducer  extends MapReduceBase implements Reducer<Text,IntWritable,Text,IntWritable
    public void reduce(Text key, Iterator<IntWritable> values,OutputCollector<Text,IntWritable> out
                       Reporter reporter) throws IOException {
        int sum=0;
        while (values.hasNext()) {
            sum+=values.next().get();
        }
        output.collect(key,new IntWritable(sum));
    }
}
```

5) Maven clean :



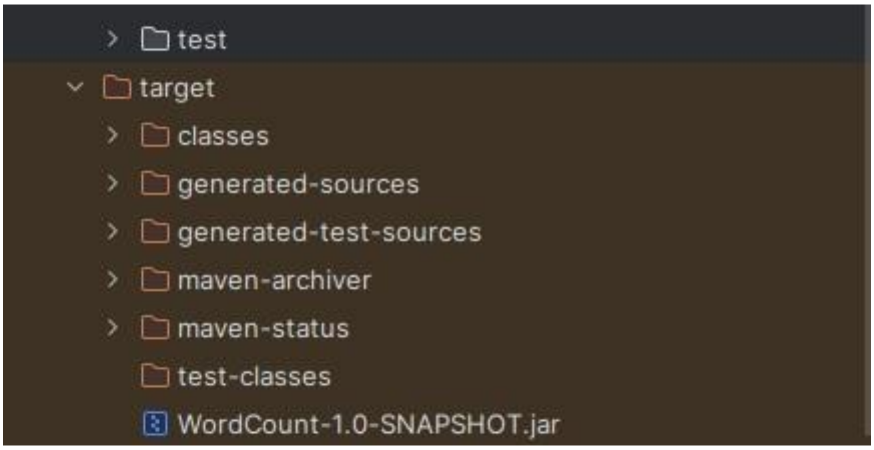
```
sr/lib/jvm/java-1.8.0-openjdk-amd64/bin/java -Dmaven.multiModuleProjectDirectory=/home
[INFO] Scanning for projects...
[INFO]
[INFO] ---------------------< org.alikhaoula:WordCount >---------------------
[INFO] Building WordCount 1.0-SNAPSHOT
[INFO]   from pom.xml
[INFO] --------------------------------[ jar ]---------------------------------
[INFO]
[INFO] --- clean:3.2.0:clean (default-clean) @ WordCount ---
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  0.459 s
[INFO] Finished at: 2024-05-25T13:16:27+01:00
[INFO] ------------------------------------------------------------------------

Process finished with exit code 0
```

6) Maeven install :



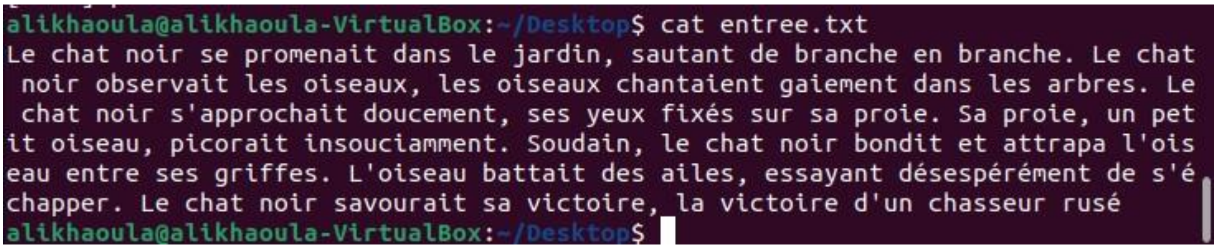
```
[INFO] Compiling 3 source files with javac [debug target 8] to target/classes
[INFO]
[INFO] --- resources:3.3.1:testResources (default-testResources) @ WordCount ---
[INFO] skip non existing resourceDirectory /home/alikhaoula/IdeaProjects/WordCount/src/t
[INFO]
[INFO] --- compiler:3.11.0:testCompile (default-testCompile) @ WordCount ---
[INFO] Changes detected - recompiling the module! :dependency
[INFO]
[INFO] --- surefire:3.2.2:test (default-test) @ WordCount ---
[INFO]
[INFO] --- jar:3.3.0:jar (default-jar) @ WordCount ---
[INFO] Building jar: /home/alikhaoula/IdeaProjects/WordCount/target/WordCount-1.0-SNAPSH
[INFO]
[INFO] --- install:3.1.1:install (default-install) @ WordCount ---
[INFO] Installing /home/alikhaoula/IdeaProjects/WordCount/pom.xml to /home/alikhaoula/.m
[INFO] Installing /home/alikhaoula/IdeaProjects/WordCount/target/WordCount-1.0-SNAPSHOT.
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  6.896 s
[INFO] Finished at: 2024-05-25T13:18:22+01:00
[INFO] ------------------------------------------------------------------------

Process finished with exit code 0
```

D’où la creation du fichier target :

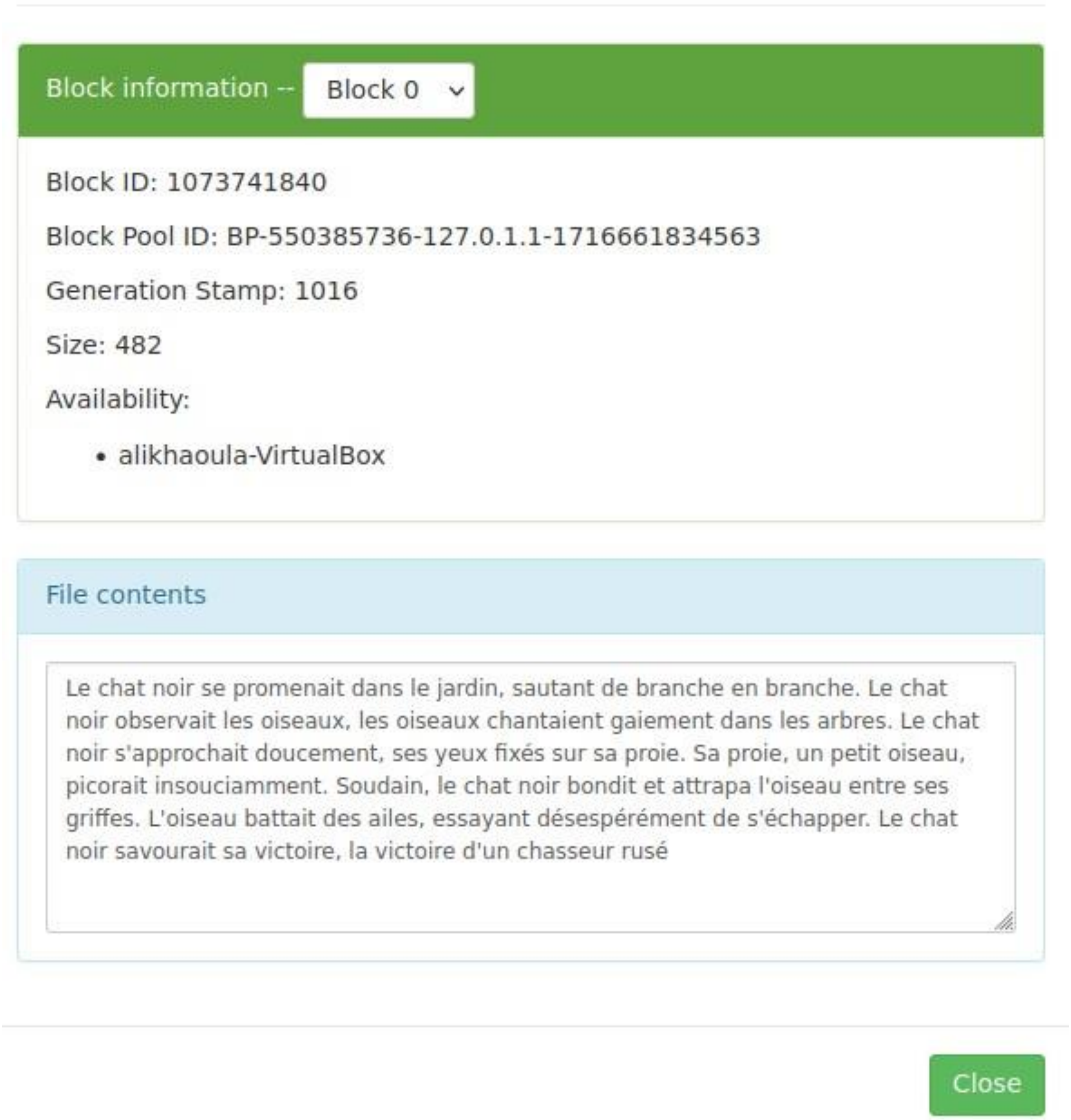


```
    > test
  ˅ target
    > classes
    > generated-sources
    > generated-test-sources
    > maven-archiver
    > maven-status
      test-classes
      WordCount-1.0-SNAPSHOT.jar
```

7) Créer un fichier texte pour tester le programme :



```
alikhaoula@alikhaoula-VirtualBox:~/Desktop$ cat entree.txt
Le chat noir se promenait dans le jardin, sautant de branche en branche. Le chat
 noir observait les oiseaux, les oiseaux chantaient gaiement dans les arbres. Le
 chat noir s'approchait doucement, ses yeux fixés sur sa proie. Sa proie, un pet
it oiseau, picorait insouciamment. Soudain, le chat noir bondit et attrapa l'ois
eau entre ses griffes. L'oiseau battait des ailes, essayant désespérément de s'é
chapper. Le chat noir savourait sa victoire, la victoire d'un chasseur rusé
alikhaoula@alikhaoula-VirtualBox:~/Desktop$
```



Block ID: 1073741840

Block Pool ID: BP-550385736-127.0.1.1-1716661834563

Generation Stamp: 1016

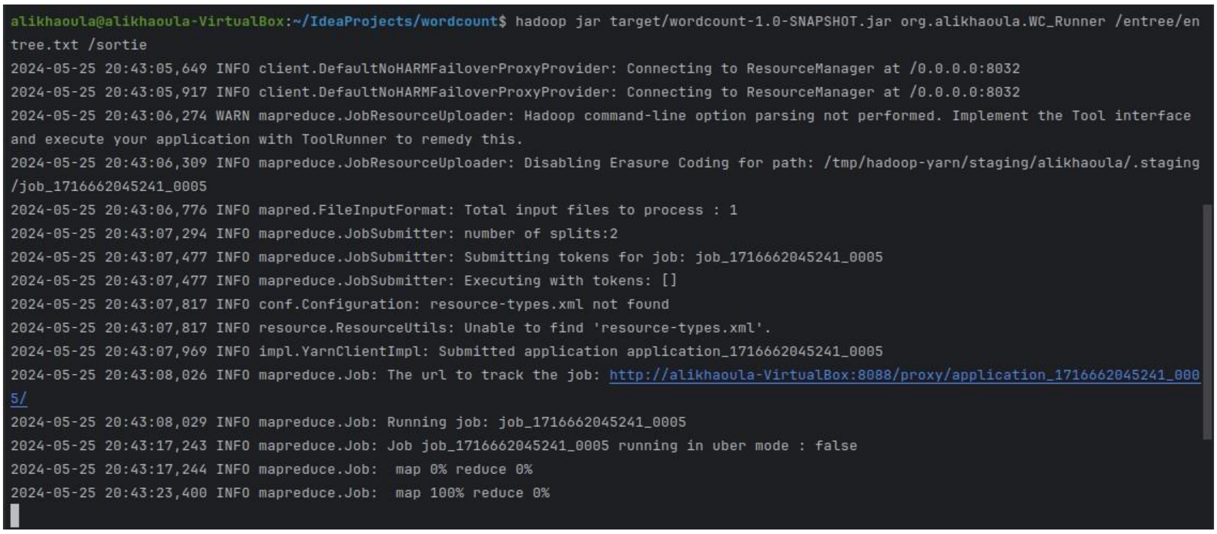
Size: 482

Availability:

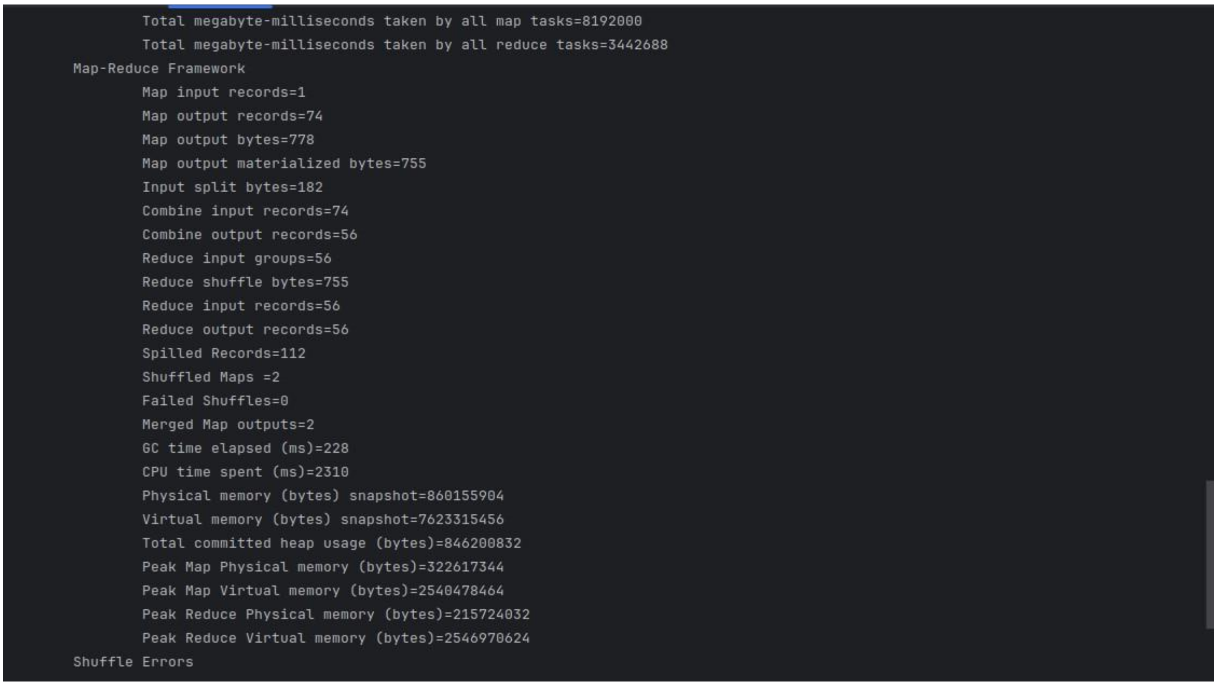
- alikhaoula-VirtualBox

```
Le chat noir se promenait dans le jardin, sautant de branche en branche. Le chat
noir observait les oiseaux, les oiseaux chantaient gaiement dans les arbres. Le chat
noir s'approchait doucement, ses yeux fixés sur sa proie. Sa proie, un petit oiseau,
picorait insouciamment. Soudain, le chat noir bondit et attrapa l'oiseau entre ses
griffes. L'oiseau battait des ailes, essayant désespérément de s'échapper. Le chat
noir savourait sa victoire, la victoire d'un chasseur rusé
```

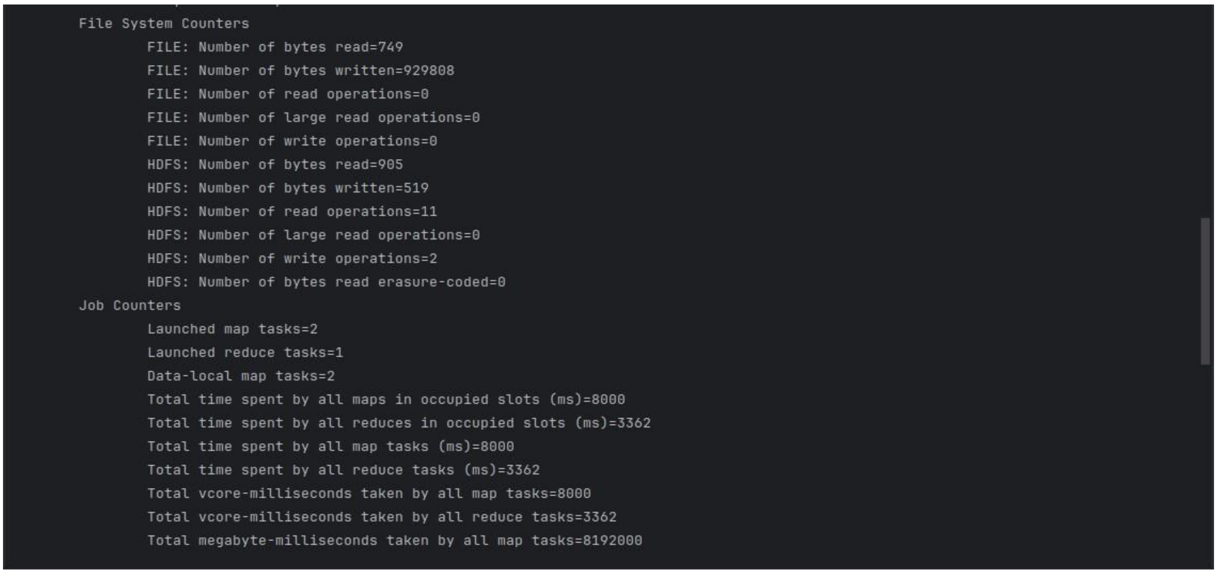
8) Tester le programme :



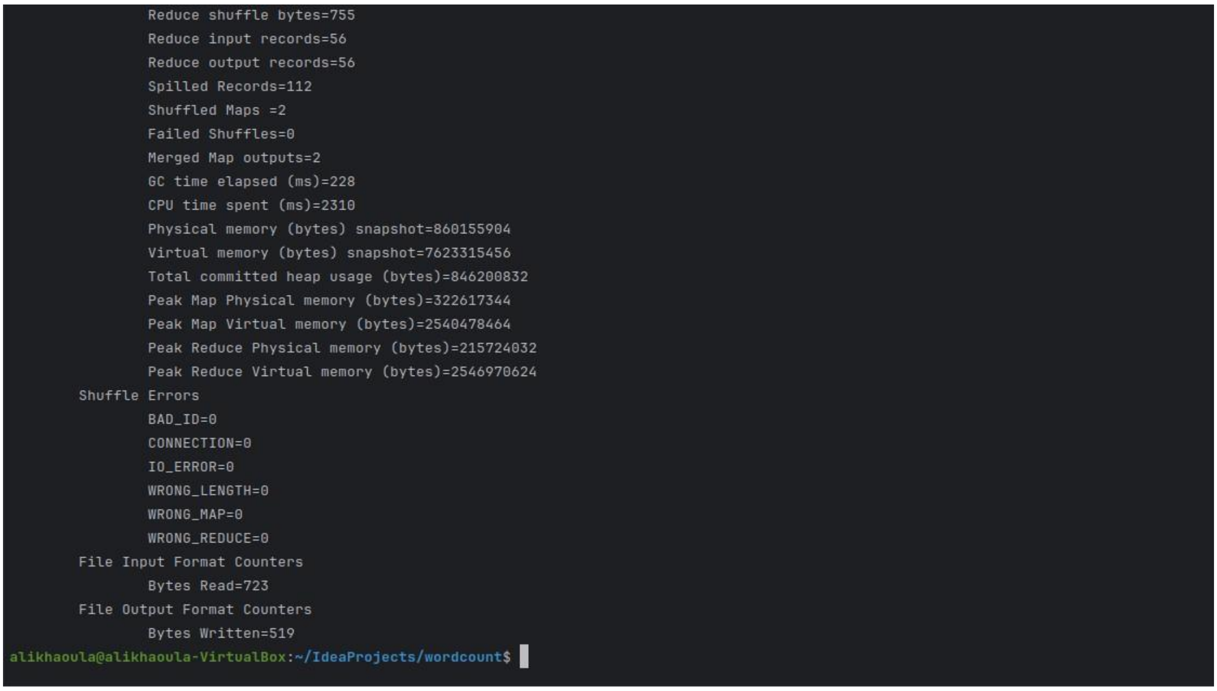
```
alikhaoula@alikhaoula-VirtualBox:~/IdeaProjects/wordcount$ hadoop jar target/wordcount-1.0-SNAPSHOT.jar org.alikhaoula.WC_Runner /entree/en
tree.txt /sortie
2024-05-25 20:43:05,649 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-05-25 20:43:05,917 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /0.0.0.0:8032
2024-05-25 20:43:06,274 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface
and execute your application with ToolRunner to remedy this.
2024-05-25 20:43:06,309 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/alikhaoula/.staging
/job_1716662045241_0005
2024-05-25 20:43:06,776 INFO mapred.FileInputFormat: Total input files to process : 1
2024-05-25 20:43:07,294 INFO mapreduce.JobSubmitter: number of splits:2
2024-05-25 20:43:07,477 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1716662045241_0005
2024-05-25 20:43:07,477 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-05-25 20:43:07,817 INFO conf.Configuration: resource-types.xml not found
2024-05-25 20:43:07,817 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-05-25 20:43:07,969 INFO impl.YarnClientImpl: Submitted application application_1716662045241_0005
2024-05-25 20:43:08,026 INFO mapreduce.Job: The url to track the job: http://alikhaoula-VirtualBox:8088/proxy/application_1716662045241_000
5/
2024-05-25 20:43:08,029 INFO mapreduce.Job: Running job: job_1716662045241_0005
2024-05-25 20:43:17,243 INFO mapreduce.Job: Job job_1716662045241_0005 running in uber mode : false
2024-05-25 20:43:17,244 INFO mapreduce.Job:  map 0% reduce 0%
2024-05-25 20:43:23,400 INFO mapreduce.Job:  map 100% reduce 0%
```



```
                Total megabyte-milliseconds taken by all map tasks=8192000
                Total megabyte-milliseconds taken by all reduce tasks=3442688
        Map-Reduce Framework
                Map input records=1
                Map output records=74
                Map output bytes=778
                Map output materialized bytes=755
                Input split bytes=182
                Combine input records=74
                Combine output records=56
                Reduce input groups=56
                Reduce shuffle bytes=755
                Reduce input records=56
                Reduce output records=56
                Spilled Records=112
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=228
                CPU time spent (ms)=2310
                Physical memory (bytes) snapshot=860155904
                Virtual memory (bytes) snapshot=7623315456
                Total committed heap usage (bytes)=846200832
                Peak Map Physical memory (bytes)=322617344
                Peak Map Virtual memory (bytes)=2540478464
                Peak Reduce Physical memory (bytes)=215724032
                Peak Reduce Virtual memory (bytes)=2546970624
        Shuffle Errors
```

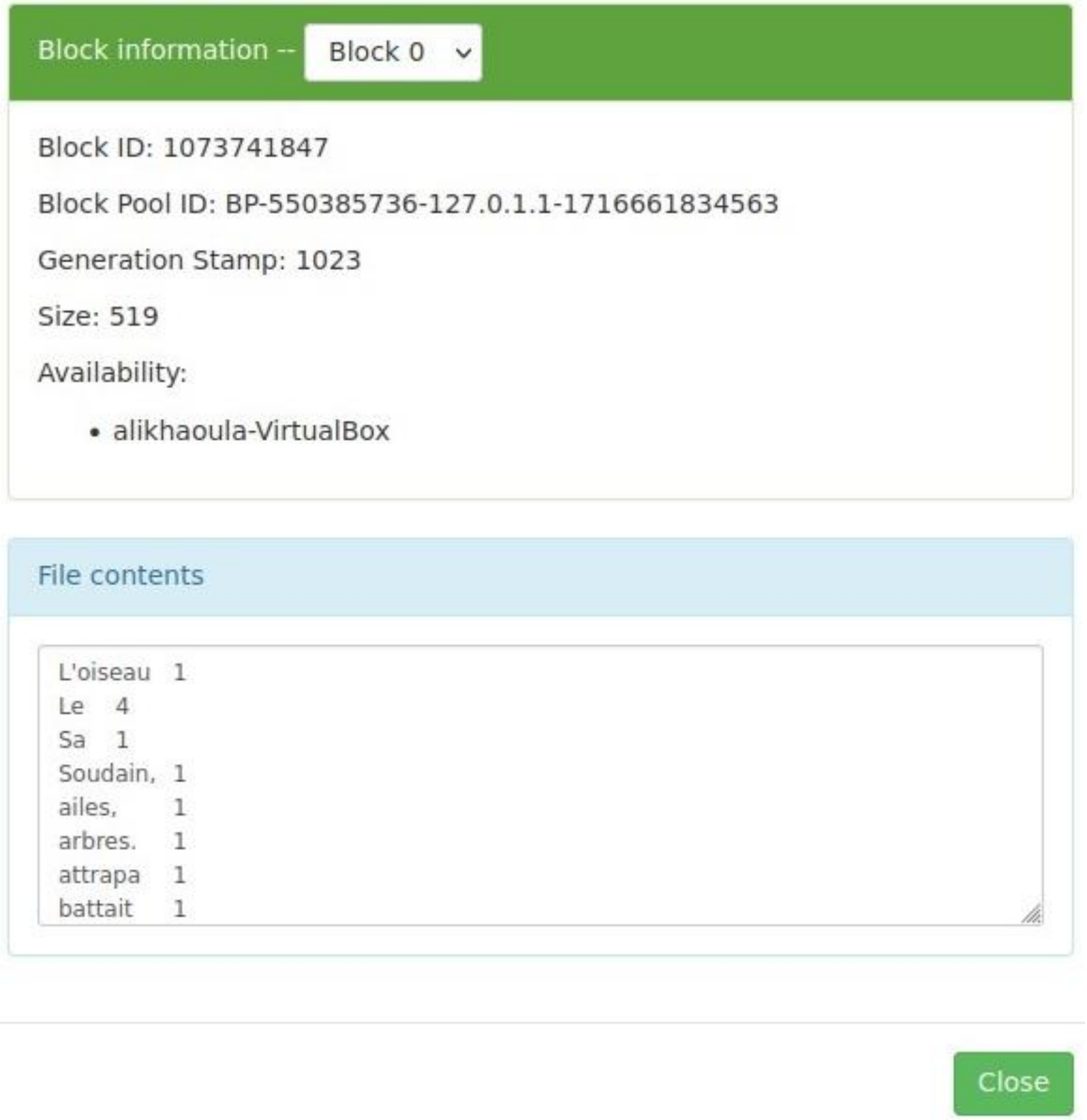


```
        File System Counters
                FILE: Number of bytes read=749
                FILE: Number of bytes written=929808
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=905
                HDFS: Number of bytes written=519
                HDFS: Number of read operations=11
                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=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=8000
                Total time spent by all reduces in occupied slots (ms)=3362
                Total time spent by all map tasks (ms)=8000
                Total time spent by all reduce tasks (ms)=3362
                Total vcore-milliseconds taken by all map tasks=8000
                Total vcore-milliseconds taken by all reduce tasks=3362
                Total megabyte-milliseconds taken by all map tasks=8192000
```



```
                Reduce shuffle bytes=755
                Reduce input records=56
                Reduce output records=56
                Spilled Records=112
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=228
                CPU time spent (ms)=2310
                Physical memory (bytes) snapshot=860155904
                Virtual memory (bytes) snapshot=7623315456
                Total committed heap usage (bytes)=846200832
                Peak Map Physical memory (bytes)=322617344
                Peak Map Virtual memory (bytes)=2540478464
                Peak Reduce Physical memory (bytes)=215724032
                Peak Reduce Virtual memory (bytes)=2546970624
        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=723
        File Output Format Counters
                Bytes Written=519
alikhaoula@alikhaoula-VirtualBox:~/IdeaProjects/wordcount$
```

9) Résultat :



Block ID: 1073741847

Block Pool ID: BP-550385736-127.0.1.1-1716661834563

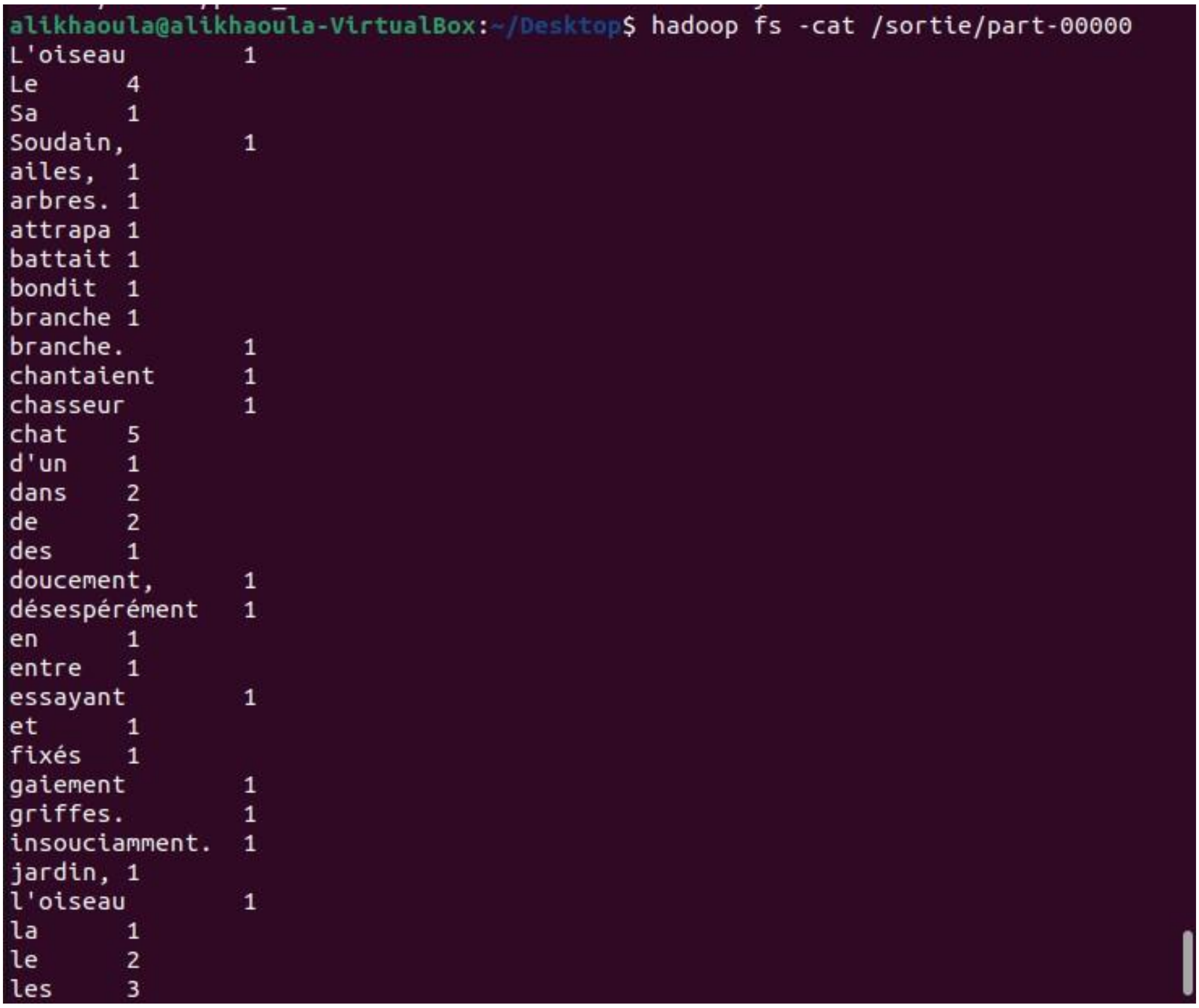
Generation Stamp: 1023

Size: 519

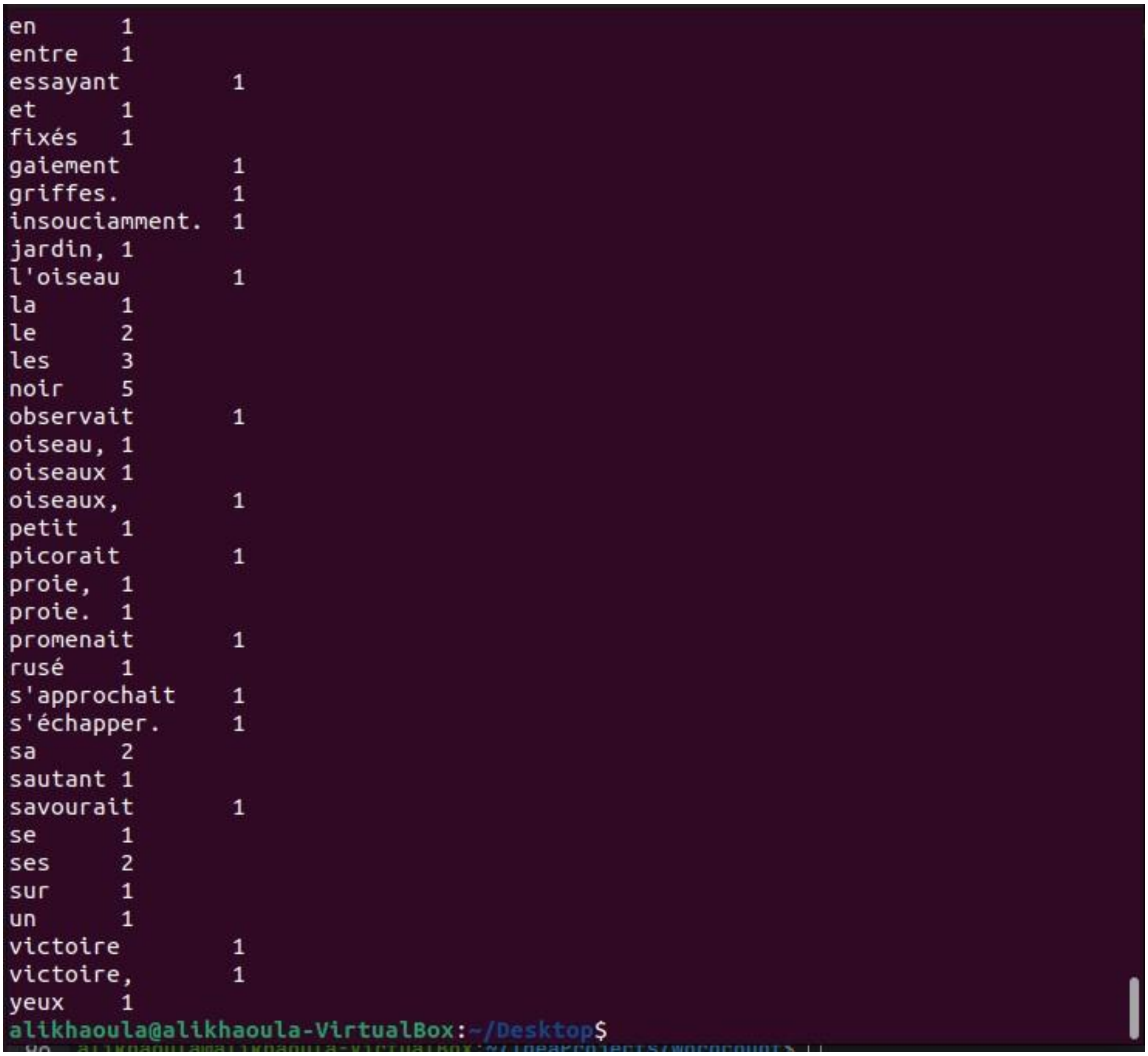
Availability:

- alikhaoula-VirtualBox

```
L'oiseau	1
Le	4
Sa	1
Soudain,	1
ailes,	1
arbres.	1
attrapa	1
battait	1
```



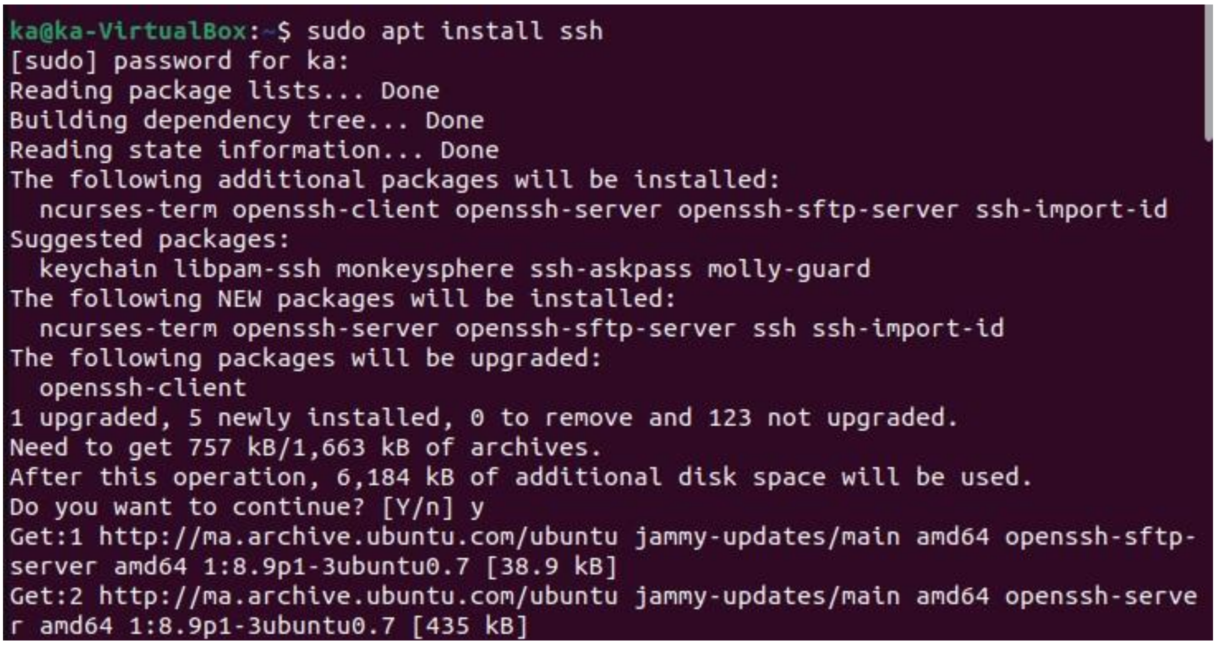
```
alikhaoula@alikhaoula-VirtualBox:~/Desktop$ hadoop fs -cat /sortie/part-00000
L'oiseau        1
Le      4
Sa      1
Soudain,        1
ailes,  1
arbres. 1
attrapa 1
battait 1
bondit  1
branche 1
branche.        1
chantaient      1
chasseur        1
chat    5
d'un    1
dans    2
de      2
des     1
doucement,      1
désespérément   1
en      1
entre   1
essayant        1
et      1
fixés   1
gaiement        1
griffes.        1
insouciamment.  1
jardin, 1
l'oiseau        1
la      1
le      2
les     3
```



```
en      1
entre   1
essayant        1
et      1
fixés   1
gaiement        1
griffes.        1
insouciamment.  1
jardin, 1
l'oiseau        1
la      1
le      2
les     3
noir    5
observait       1
oiseau, 1
oiseaux 1
oiseaux,        1
petit   1
picorait        1
proie,  1
proie.  1
promenait       1
rusé    1
s'approchait    1
s'échapper.     1
sa      2
sautant 1
savourait       1
se      1
ses     2
sur     1
un      1
victoire        1
victoire,       1
yeux    1
alikhaoula@alikhaoula-VirtualBox:~/Desktop$
```

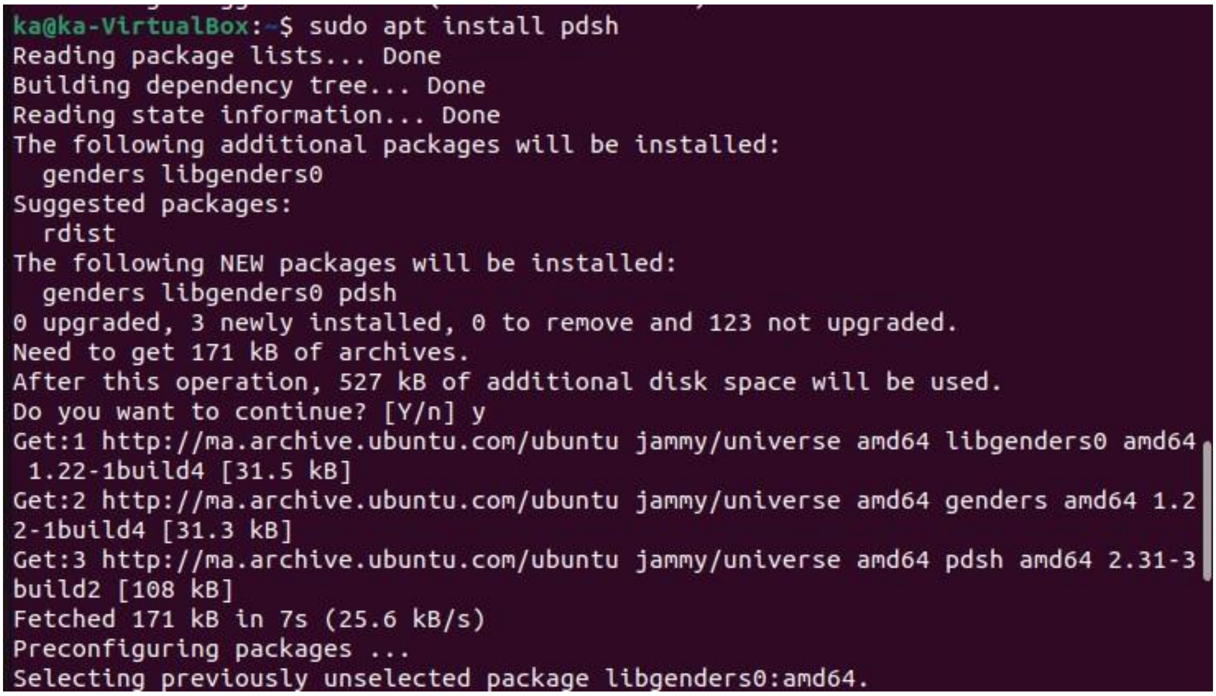
Configuration de Hadoop dans un cluster à nœuds multiples.

1) Dans les paramètres VB, assurez-vous que votre carte réseau est définie sur accès par ponts.
2) Installer ssh :



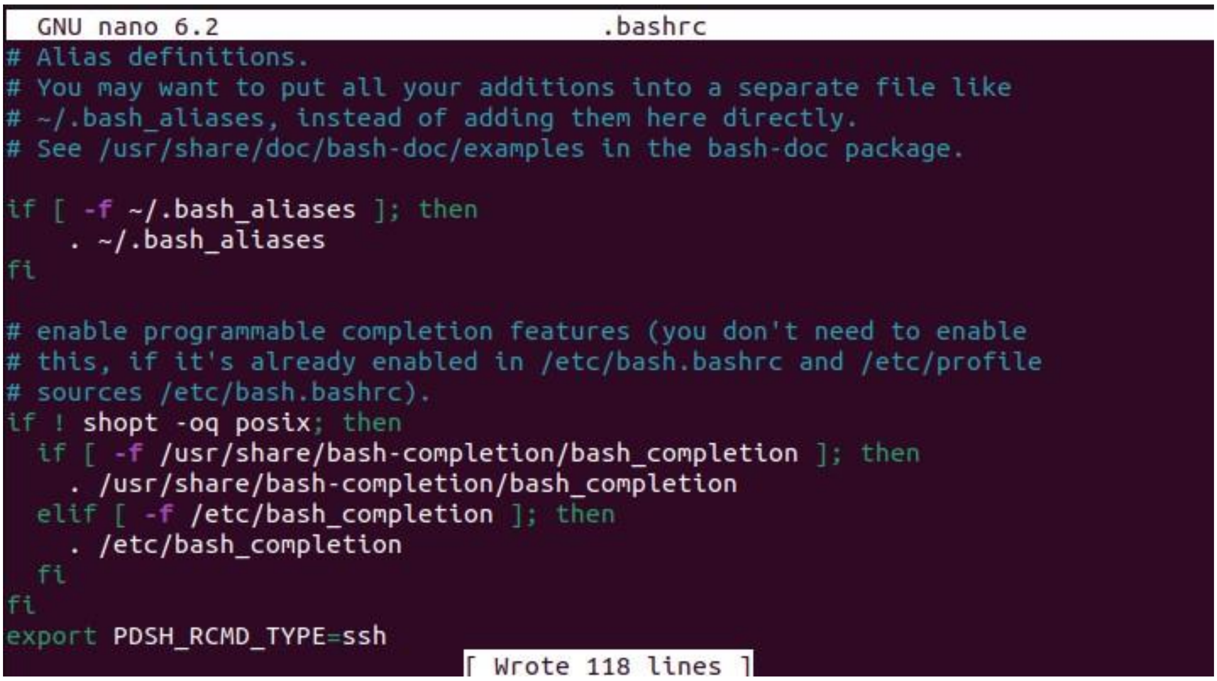
```
ka@ka-VirtualBox:~$ sudo apt install ssh
[sudo] password for ka:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ncurses-term openssh-client openssh-server openssh-sftp-server ssh-import-id
Suggested packages:
  keychain libpam-ssh monkeysphere ssh-askpass molly-guard
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
The following packages will be upgraded:
  openssh-client
1 upgraded, 5 newly installed, 0 to remove and 123 not upgraded.
Need to get 757 kB/1,663 kB of archives.
After this operation, 6,184 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://ma.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-sftp-
server amd64 1:8.9p1-3ubuntu0.7 [38.9 kB]
Get:2 http://ma.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-serve
r amd64 1:8.9p1-3ubuntu0.7 [435 kB]
```

3) Installer pdsh :



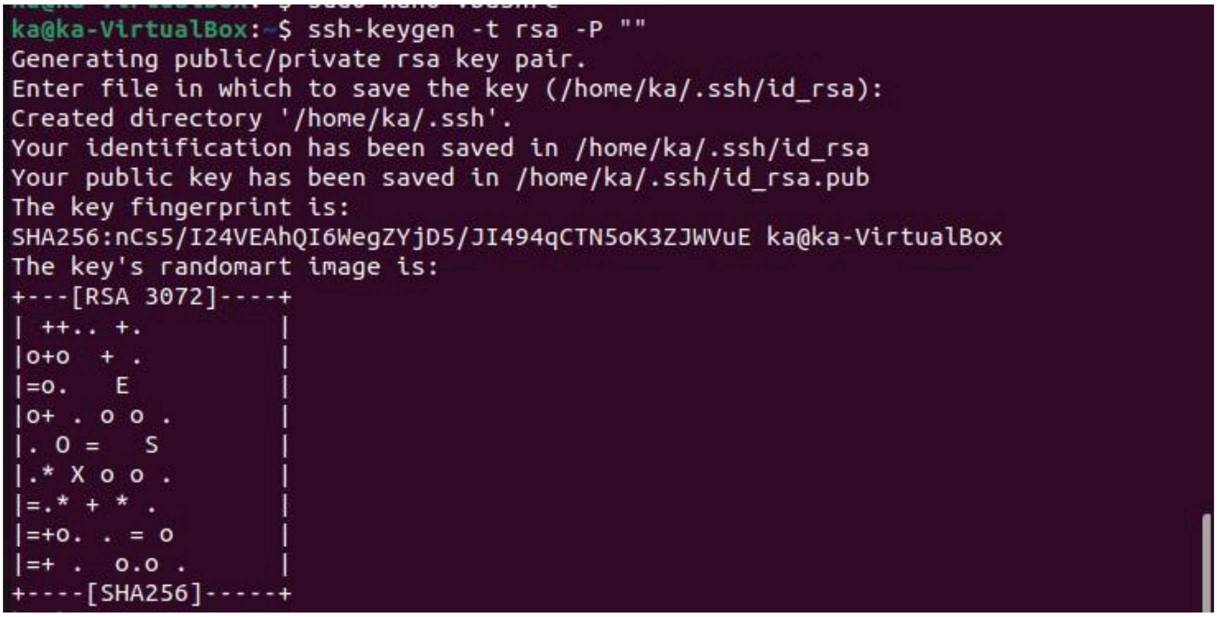
```
ka@ka-VirtualBox:~$ sudo apt install pdsh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  genders libgenders0
Suggested packages:
  rdist
The following NEW packages will be installed:
  genders libgenders0 pdsh
0 upgraded, 3 newly installed, 0 to remove and 123 not upgraded.
Need to get 171 kB of archives.
After this operation, 527 kB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://ma.archive.ubuntu.com/ubuntu jammy/universe amd64 libgenders0 amd64
 1.22-1build4 [31.5 kB]
Get:2 http://ma.archive.ubuntu.com/ubuntu jammy/universe amd64 genders amd64 1.2
2-1build4 [31.3 kB]
Get:3 http://ma.archive.ubuntu.com/ubuntu jammy/universe amd64 pdsh amd64 2.31-3
build2 [108 kB]
Fetched 171 kB in 7s (25.6 kB/s)
Preconfiguring packages ...
Selecting previously unselected package libgenders0:amd64.
```

Ouvrir le fichier .bashrc et ajouter `export PDSH_RCMD_TYPE=ssh`



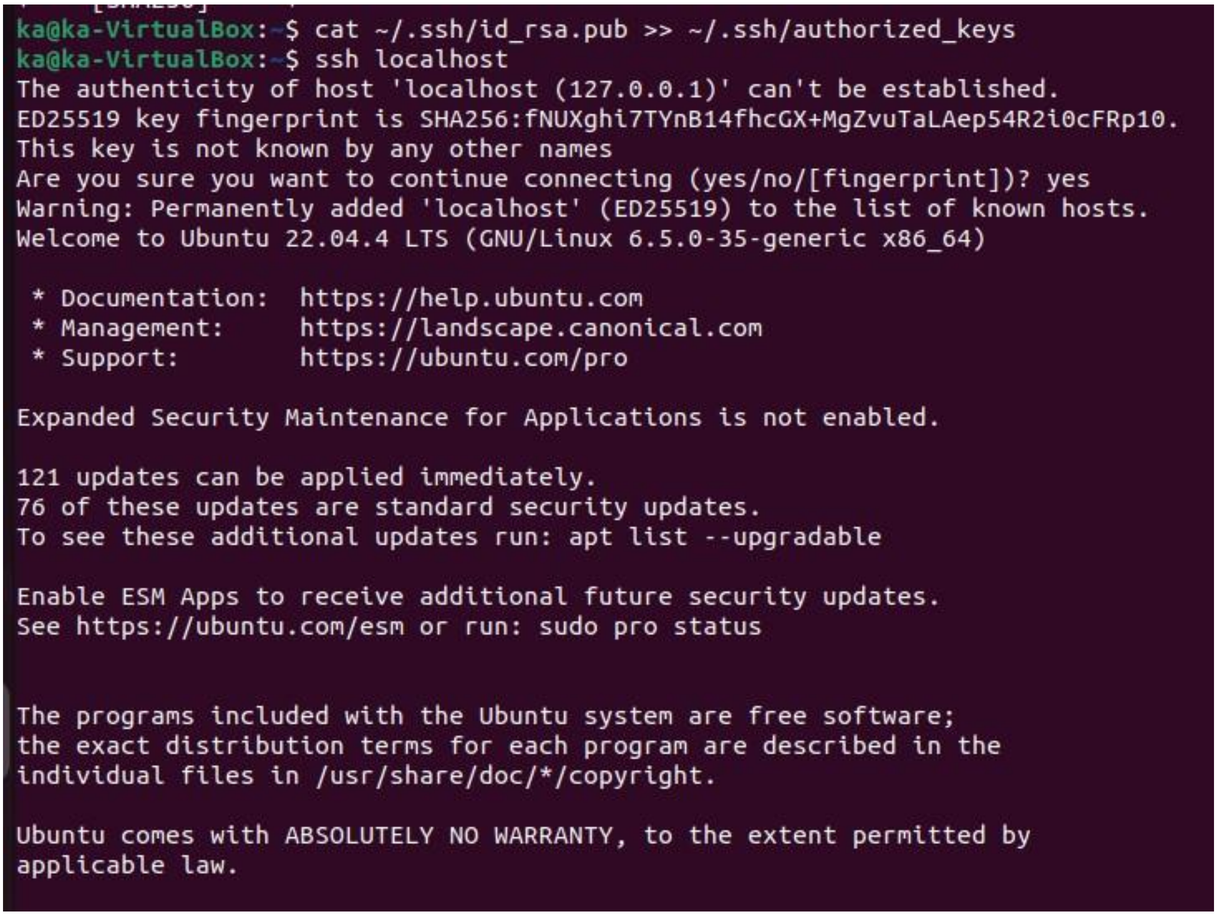
```
  GNU nano 6.2                        .bashrc
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

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

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
export PDSH_RCMD_TYPE=ssh
                              [ Wrote 118 lines ]
```

4) Générer une clé ssh :



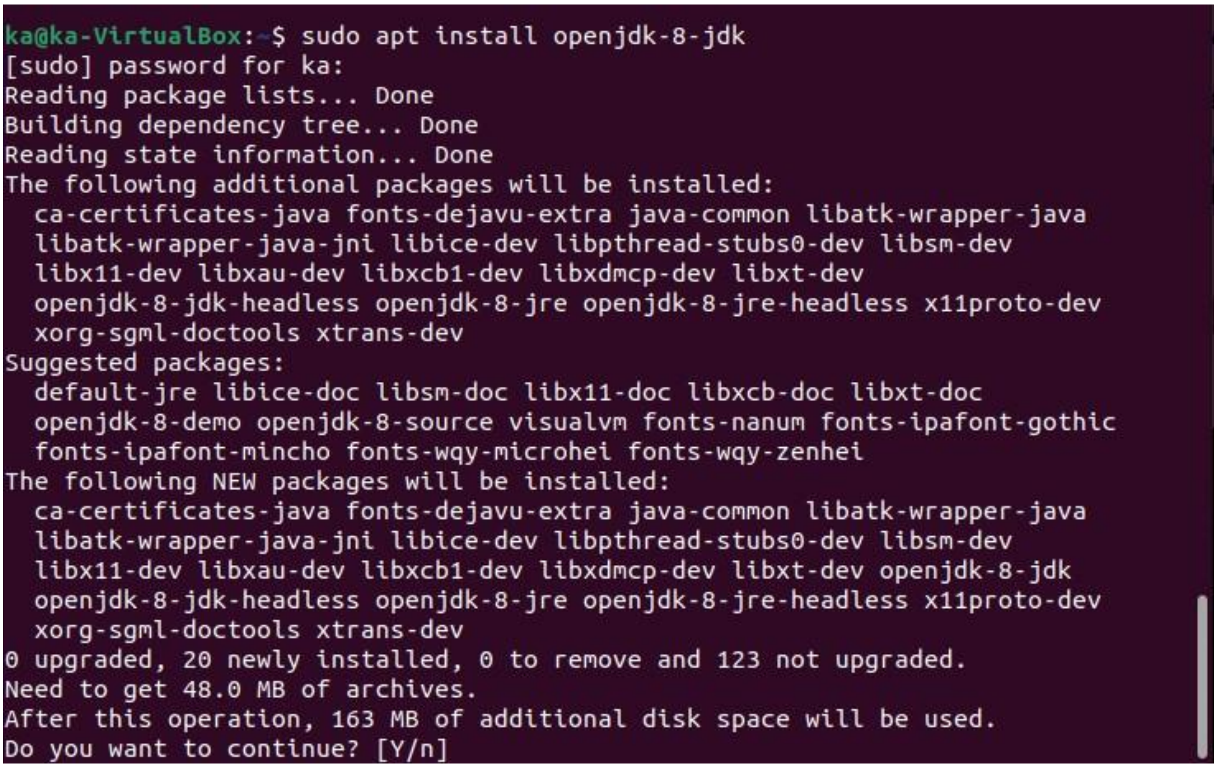
```
ka@ka-VirtualBox:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/ka/.ssh/id_rsa):
Created directory '/home/ka/.ssh'.
Your identification has been saved in /home/ka/.ssh/id_rsa
Your public key has been saved in /home/ka/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:nCs5/I24VEAhQI6WegZYjD5/JI494qCTN5oK3ZJWVuE ka@ka-VirtualBox
The key's randomart image is:
+---[RSA 3072]----+
| ++.. +.         |
|o+o  + .         |
|=o.   E          |
|o+ . o o .       |
|. O =   S        |
|.* X o o .       |
|=.* + * .        |
|=+o. . = o       |
|=+ .  o.o .      |
+----[SHA256]-----+
```

5) Copier les clés autorisées pour donner les permissions nécessaires.
Tester si tout marche bien.

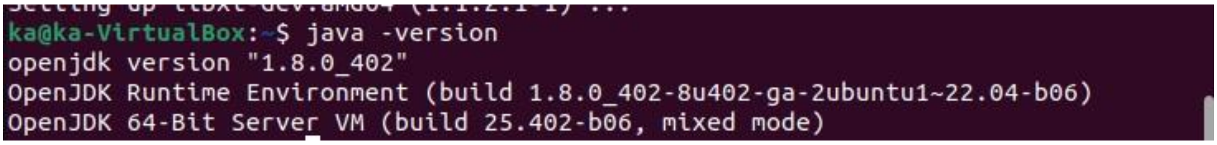


```
ka@ka-VirtualBox:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
ka@ka-VirtualBox:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:fNUXghi7TYnB14fhcGX+MgZvuTaLAep54R2i0cFRp10.
This key is not known by any other names
Are 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 6.5.0-35-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.

121 updates can be applied immediately.
76 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.
```

6) Installer java 8.

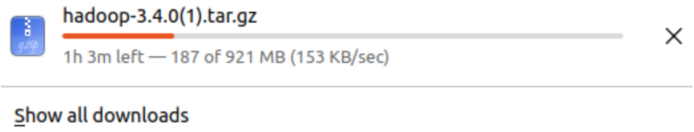


```
ka@ka-VirtualBox:~$ sudo apt install openjdk-8-jdk
[sudo] password for ka:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni 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 visualvm 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 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 123 not upgraded.
Need to get 48.0 MB of archives.
After this operation, 163 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```



```
ka@ka-VirtualBox:~$ java -version
openjdk version "1.8.0_402"
OpenJDK Runtime Environment (build 1.8.0_402-8u402-ga-2ubuntu1~22.04-b06)
OpenJDK 64-Bit Server VM (build 25.402-b06, mixed mode)
```

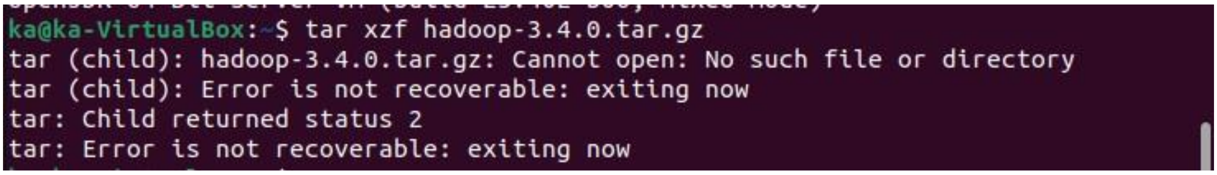
7) Télécharger et installer Hadoop depuis : apache.hadoop.org.



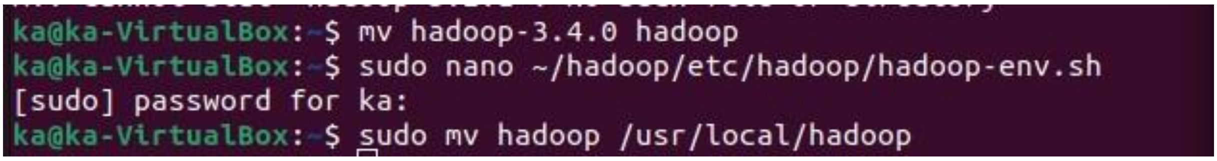
hadoop-3.4.0(1).tar.gz

1h 3m left — 187 of 921 MB (153 KB/sec)

8) Extraire le fichier zip de Hadoop et renommer hadoop-3.4.0 et déplacer le fichier :

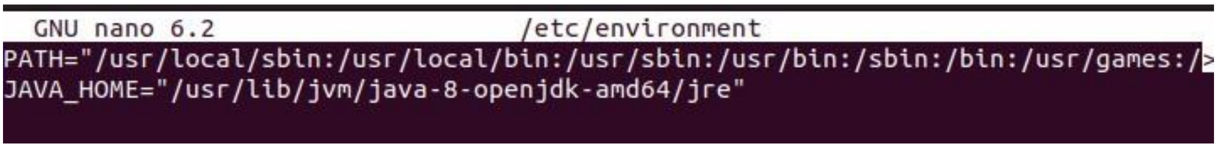


```
ka@ka-VirtualBox:~$ tar xzf hadoop-3.4.0.tar.gz
tar (child): hadoop-3.4.0.tar.gz: Cannot open: No such file or directory
tar (child): Error is not recoverable: exiting now
tar: Child returned status 2
tar: Error is not recoverable: exiting now
```



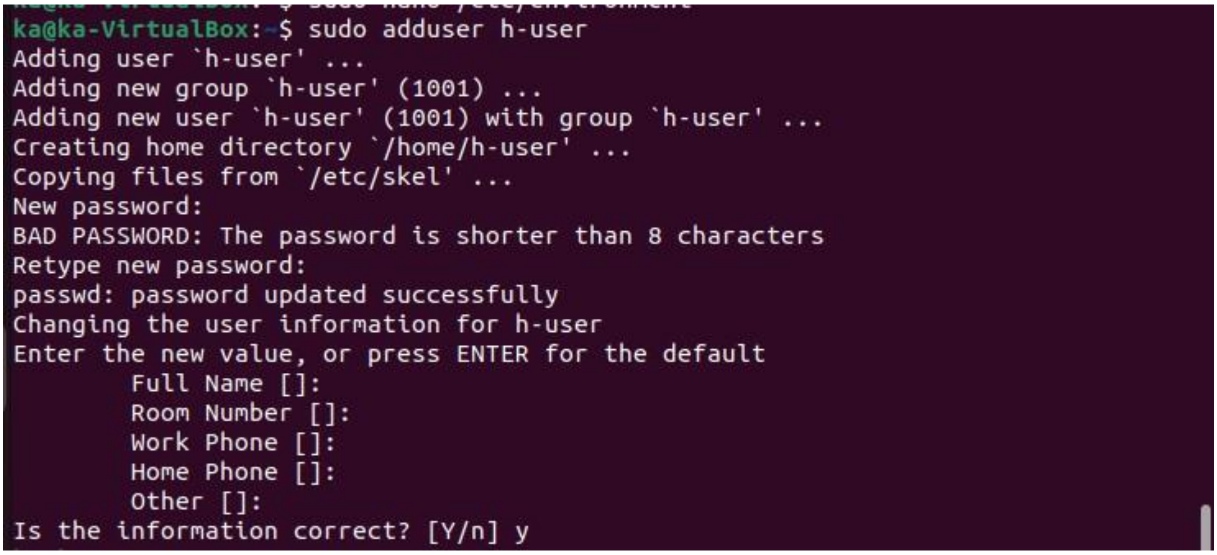
```
ka@ka-VirtualBox:~$ mv hadoop-3.4.0 hadoop
ka@ka-VirtualBox:~$ sudo nano ~/hadoop/etc/hadoop/hadoop-env.sh
[sudo] password for ka:
ka@ka-VirtualBox:~$ sudo mv hadoop /usr/local/hadoop
```

9) Configurer hadoop path :



```
  GNU nano 6.2                        /etc/environment
PATH="/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/>
JAVA_HOME="/usr/lib/jvm/java-8-openjdk-amd64/jre"
```

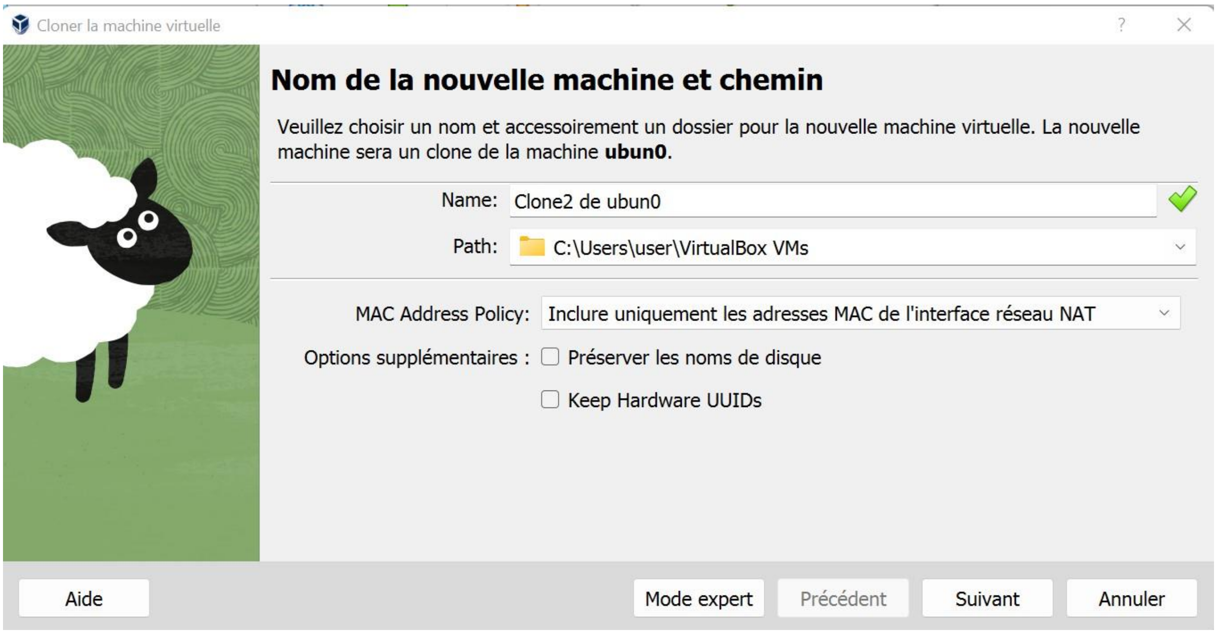
10) Créer un utilisateur spécifique pour Hadoop, et donner lui les permissions nécessaire pour travailler à l'intérieur du dossier hadoop :



```
ka@ka-VirtualBox:~$ sudo adduser h-user
Adding user `h-user' ...
Adding new group `h-user' (1001) ...
Adding new user `h-user' (1001) with group `h-user' ...
Creating home directory `/home/h-user' ...
Copying files from `/etc/skel' ...
New password:
BAD PASSWORD: The password is shorter than 8 characters
Retype new password:
passwd: password updated successfully
Changing the user information for h-user
Enter the new value, or press ENTER for the default
        Full Name []:
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] y
```

```
sudo usermod -aG hadoopuser h-user
sudo chown h-user:root -R /usr/local/hadoop/
sudo chmod g+rwx -R /usr/local/hadoop/
sudo adduser h-user sudo
```

11) Créer 2 clones de la machine virtuelle actuelle :



Nom de la nouvelle machine et chemin

Veuillez choisir un nom et accessoirement un dossier pour la nouvelle machine virtuelle. La nouvelle machine sera un clone de la machine **ubun0**.

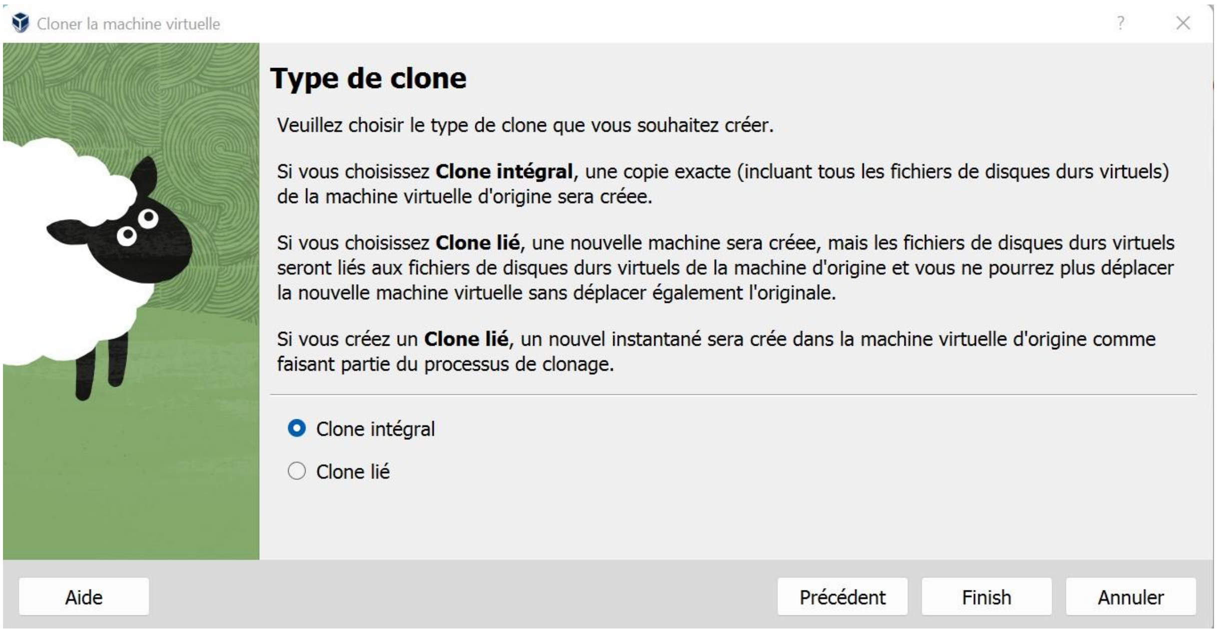
Name: Clone2 de ubun0 ✓

Path: C:\Users\user\VirtualBox VMs

MAC Address Policy: Inclure uniquement les adresses MAC de l'interface réseau NAT

Options supplémentaires : □ Préserver les noms de disque

□ Keep Hardware UUIDs



Type de clone

Veuillez choisir le type de clone que vous souhaitez créer.

Si vous choisissez **Clone intégral**, une copie exacte (incluant tous les fichiers de disques durs virtuels) de la machine virtuelle d'origine sera créee.

Si vous choisissez **Clone lié**, une nouvelle machine sera créee, mais les fichiers de disques durs virtuels seront liés aux fichiers de disques durs virtuels de la machine d'origine et vous ne pourrez plus déplacer la nouvelle machine virtuelle sans déplacer également l'originale.

Si vous créez un **Clone lié**, un nouvel instantané sera crée dans la machine virtuelle d'origine comme faisant partie du processus de clonage.

● Clone intégral

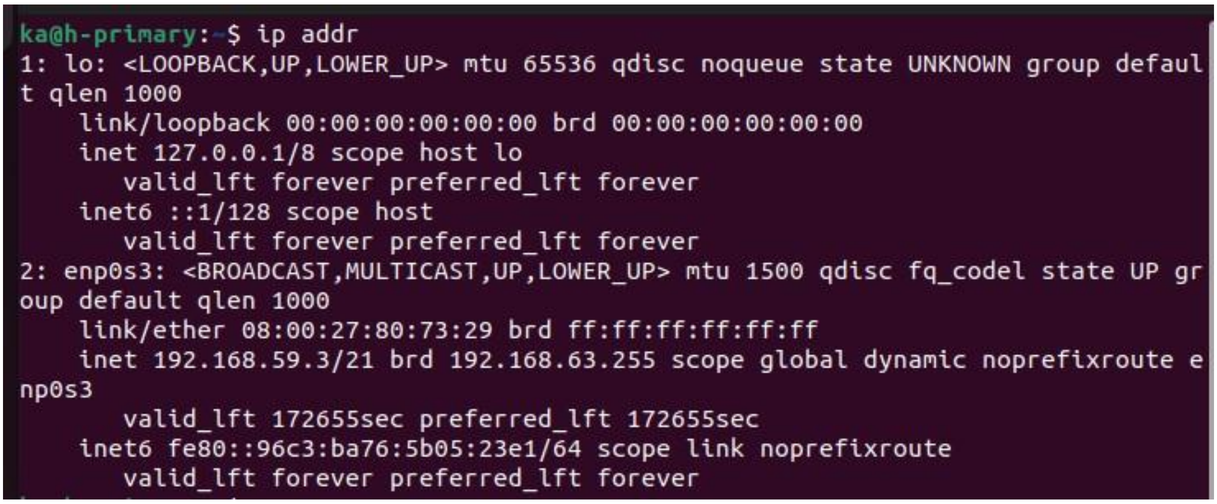
○ Clone lié

12) Changer les hostnames dans chaque machine :

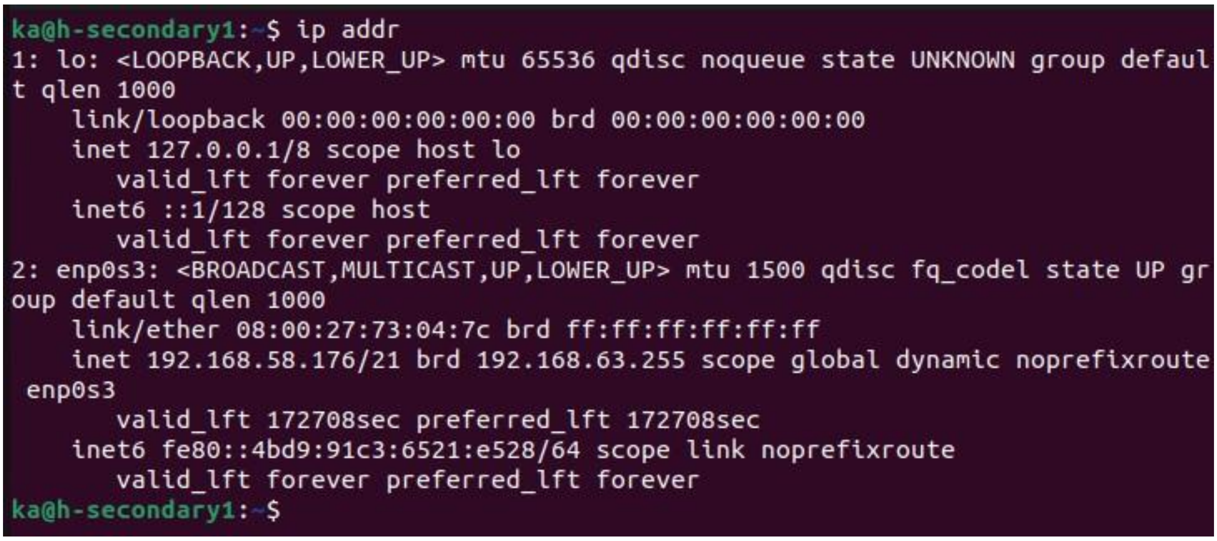
h-primary , h-secondary1et h-secondary2 dans /etc/hostname

puis redemarrer les machines.

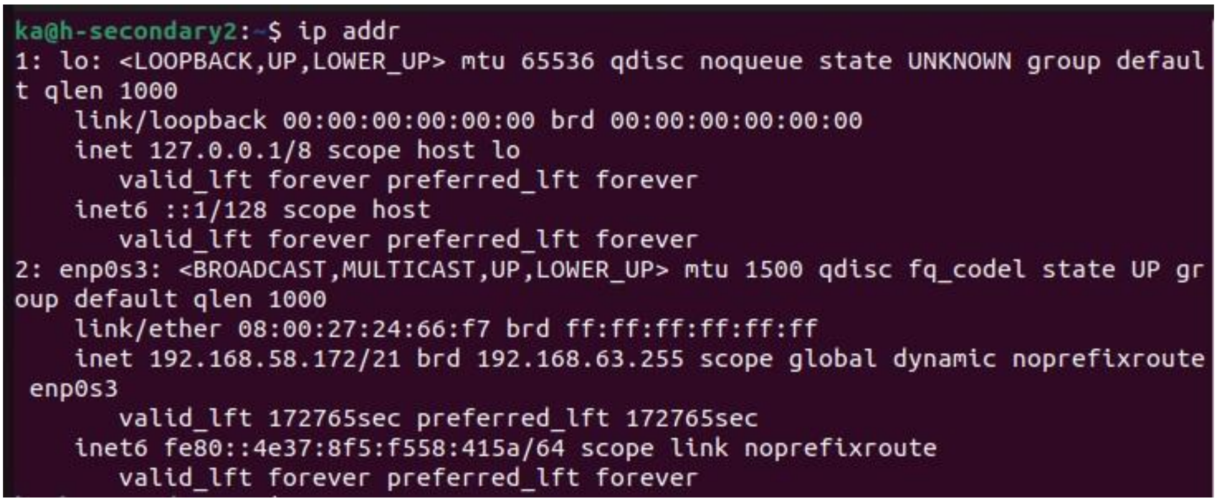
13) Chercher les addresses ip des machines :



```
ka@h-primary:~$ ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group defaul
t qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP gr
oup default qlen 1000
    link/ether 08:00:27:80:73:29 brd ff:ff:ff:ff:ff:ff
    inet 192.168.59.3/21 brd 192.168.63.255 scope global dynamic noprefixroute e
np0s3
       valid_lft 172655sec preferred_lft 172655sec
    inet6 fe80::96c3:ba76:5b05:23e1/64 scope link noprefixroute
       valid_lft forever preferred_lft forever
```

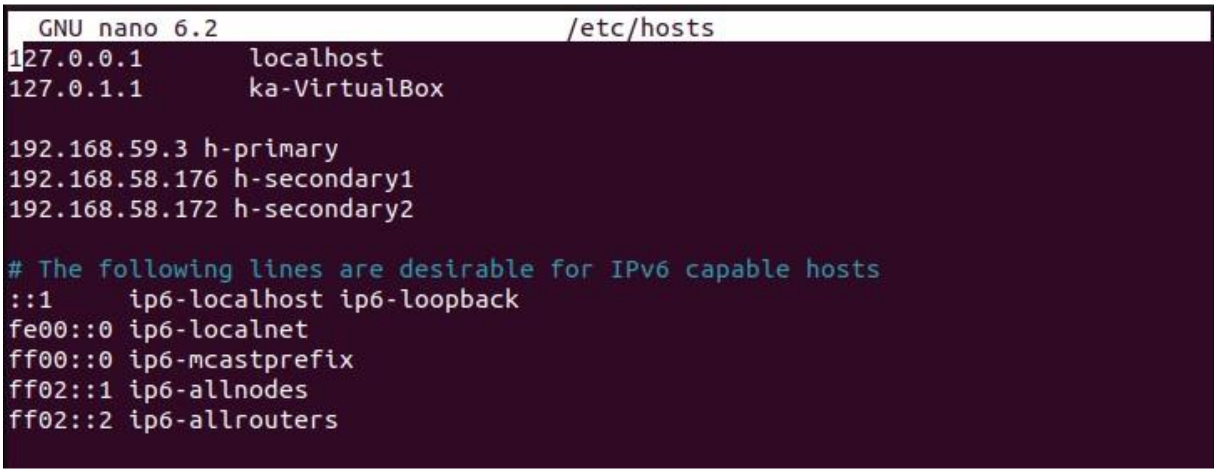


```
ka@h-secondary1:~$ ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group defaul
t qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP gr
oup default qlen 1000
    link/ether 08:00:27:73:04:7c brd ff:ff:ff:ff:ff:ff
    inet 192.168.58.176/21 brd 192.168.63.255 scope global dynamic noprefixroute
 enp0s3
       valid_lft 172708sec preferred_lft 172708sec
    inet6 fe80::4bd9:91c3:6521:e528/64 scope link noprefixroute
       valid_lft forever preferred_lft forever
ka@h-secondary1:~$
```



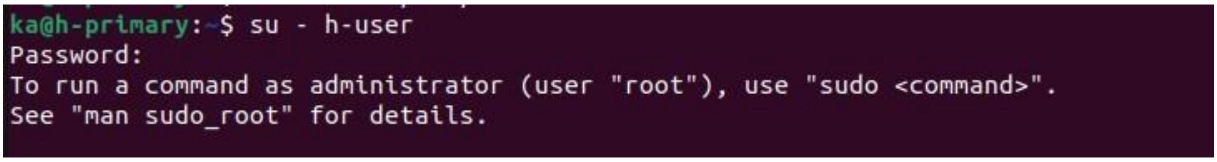
```
ka@h-secondary2:~$ ip addr
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group defaul
t qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP gr
oup default qlen 1000
    link/ether 08:00:27:24:66:f7 brd ff:ff:ff:ff:ff:ff
    inet 192.168.58.172/21 brd 192.168.63.255 scope global dynamic noprefixroute
 enp0s3
       valid_lft 172765sec preferred_lft 172765sec
    inet6 fe80::4e37:8f5:f558:415a/64 scope link noprefixroute
       valid_lft forever preferred_lft forever
```

14) Modifier le fichier hosts de chaque machine :



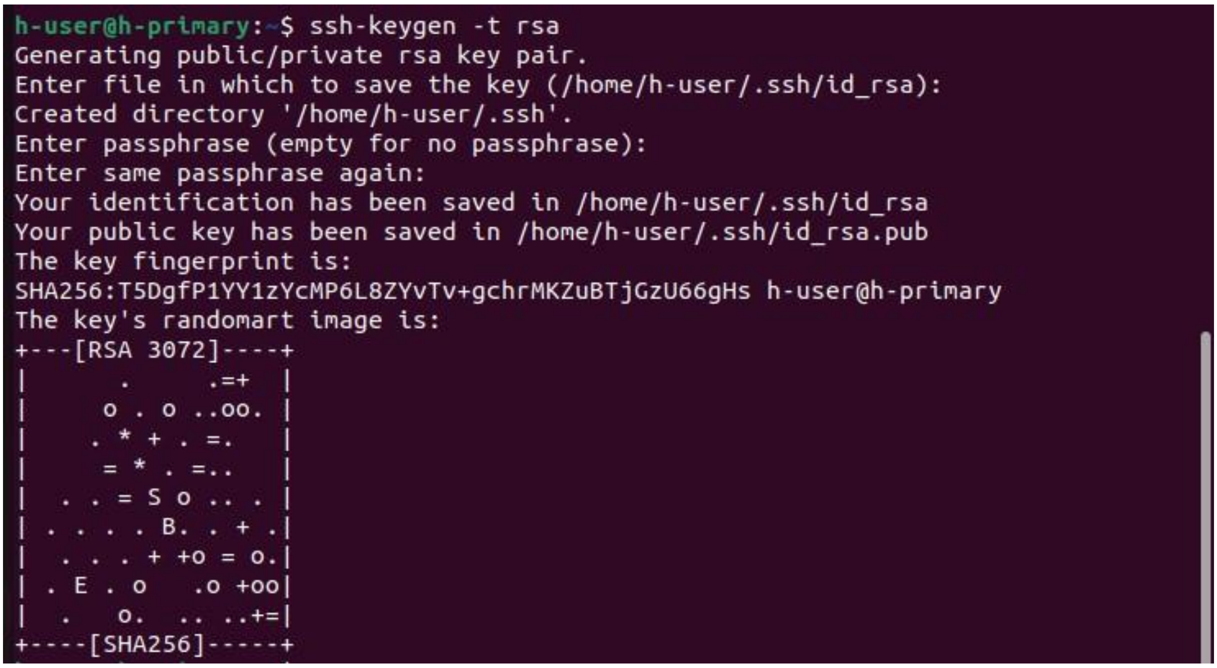
```
  GNU nano 6.2                    /etc/hosts
127.0.0.1       localhost
127.0.1.1       ka-VirtualBox

192.168.59.3 h-primary
192.168.58.176 h-secondary1
192.168.58.172 h-secondary2

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

15) Configurer le ssh sur le primaire avec le user qu'on a créé :



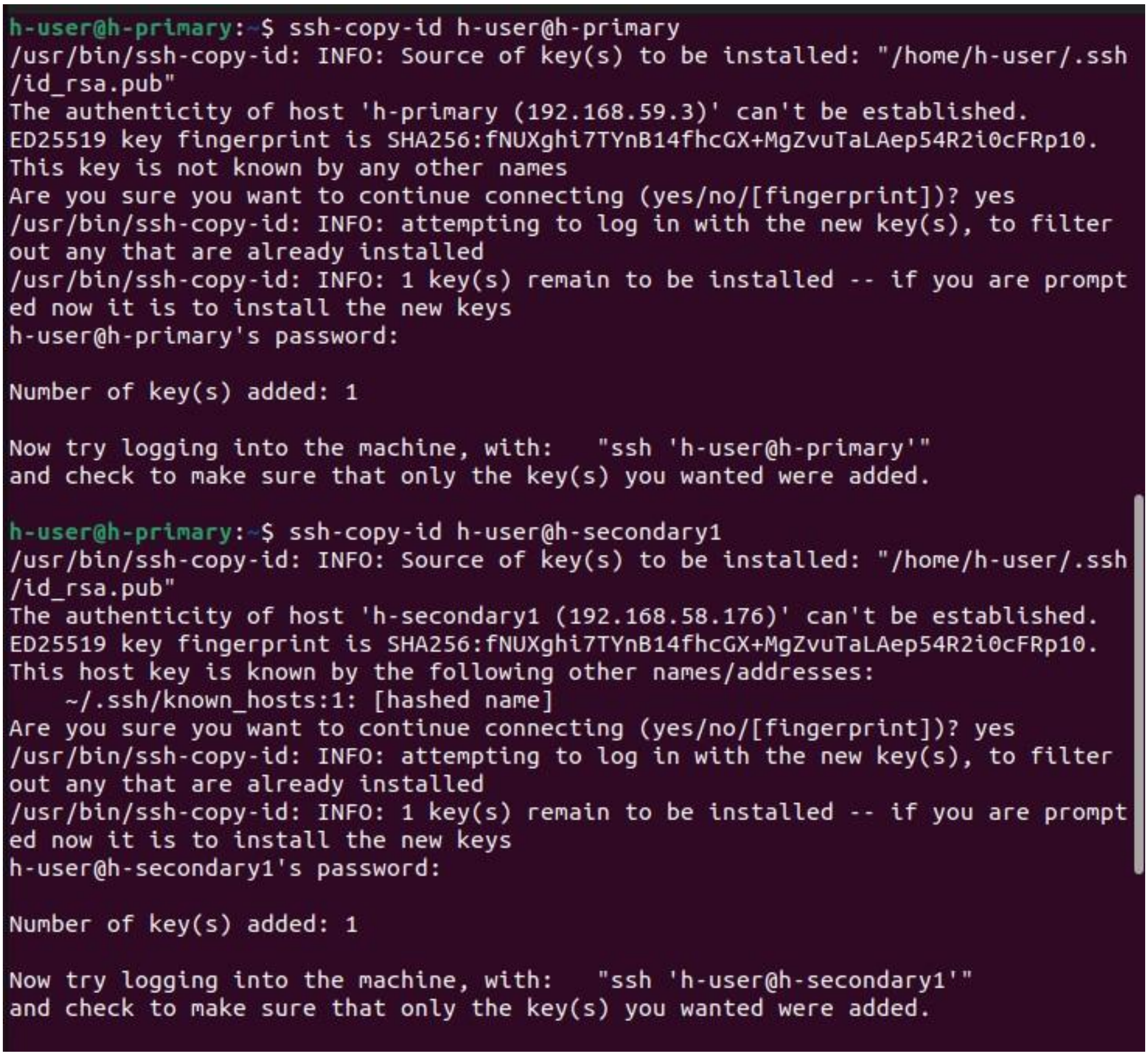
```
ka@h-primary:~$ su - h-user
Password:
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.
```

16) Générer une clé pour ce user :

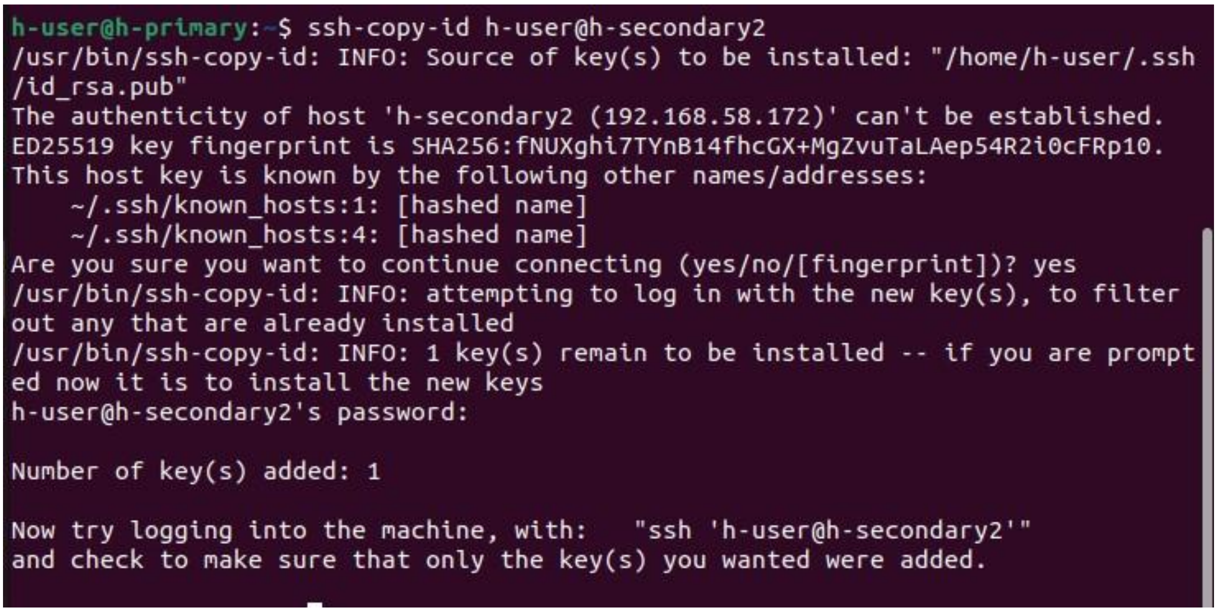


```
h-user@h-primary:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/h-user/.ssh/id_rsa):
Created directory '/home/h-user/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/h-user/.ssh/id_rsa
Your public key has been saved in /home/h-user/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:T5DgfP1YY1zYcMP6L8ZYvTv+gchrMKZuBTjGzU66gHs h-user@h-primary
The key's randomart image is:
+---[RSA 3072]----+
|      .     .=+  |
|     o . o ..oo. |
|    . * + . =.   |
|     = * . =..   |
|  . . = S o .. . |
| . . . . B. . + .|
|  . . . + +o = o.|
| . E . o   .o +oo|
|  .   o.  .. ..+=|
+----[SHA256]-----+
```

17) Copier les clés ssh dans les machines secondaires 'esclaves'



```
h-user@h-primary:~$ ssh-copy-id h-user@h-primary
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/h-user/.ssh
/id_rsa.pub"
The authenticity of host 'h-primary (192.168.59.3)' can't be established.
ED25519 key fingerprint is SHA256:fNUXghi7TYnB14fhcGX+MgZvuTaLAep54R2i0cFRp10.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter
out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompt
ed now it is to install the new keys
h-user@h-primary's password:

Number of key(s) added: 1

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

h-user@h-primary:~$ ssh-copy-id h-user@h-secondary1
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/h-user/.ssh
/id_rsa.pub"
The authenticity of host 'h-secondary1 (192.168.58.176)' can't be established.
ED25519 key fingerprint is SHA256:fNUXghi7TYnB14fhcGX+MgZvuTaLAep54R2i0cFRp10.
This host key is known by the following other names/addresses:
    ~/.ssh/known_hosts:1: [hashed name]
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter
out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompt
ed now it is to install the new keys
h-user@h-secondary1's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'h-user@h-secondary1'"
and check to make sure that only the key(s) you wanted were added.
```

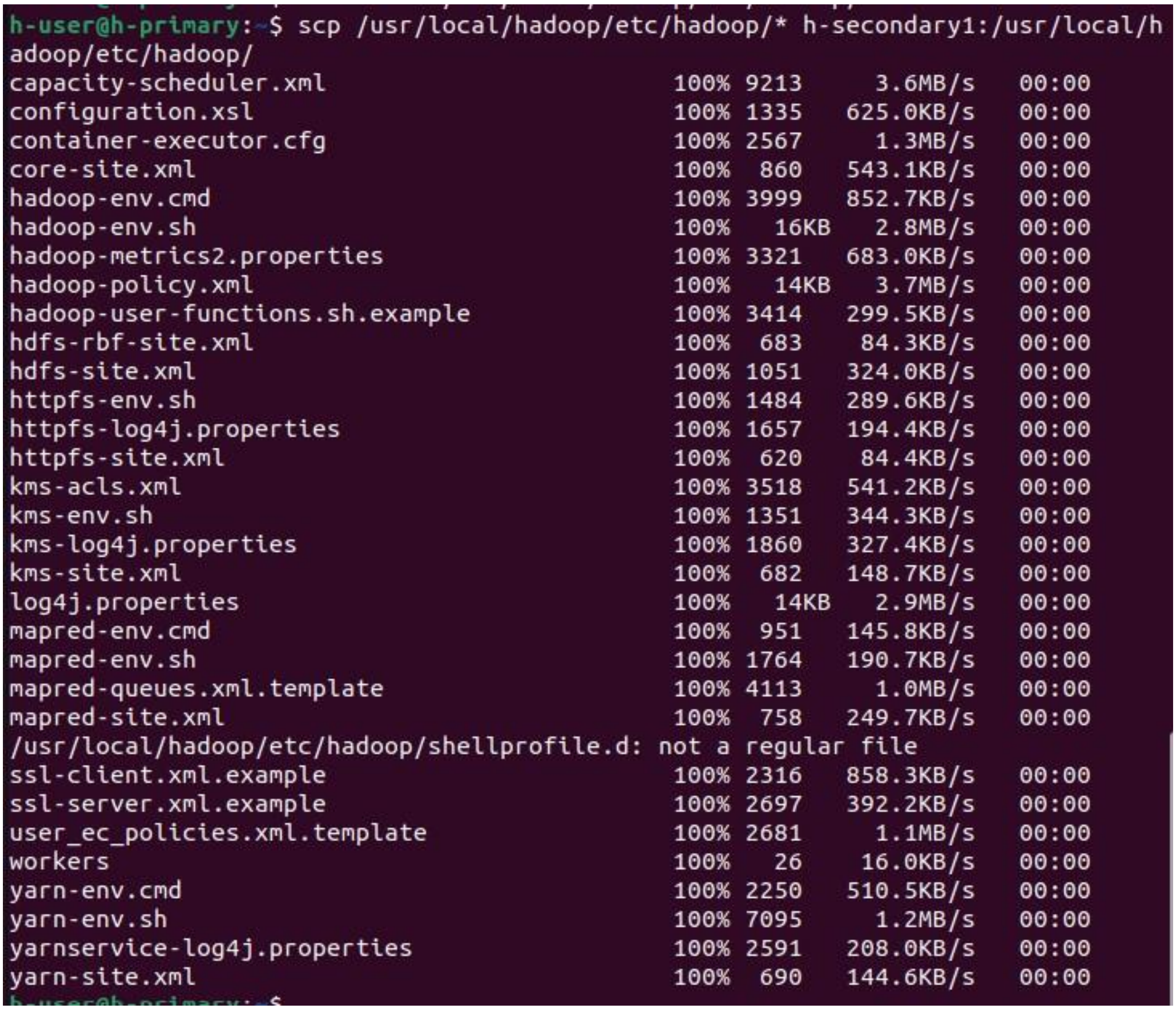


```
h-user@h-primary:~$ ssh-copy-id h-user@h-secondary2
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/h-user/.ssh
/id_rsa.pub"
The authenticity of host 'h-secondary2 (192.168.58.172)' can't be established.
ED25519 key fingerprint is SHA256:fNUXghi7TYnB14fhcGX+MgZvuTaLAep54R2i0cFRp10.
This host key is known by the following other names/addresses:
    ~/.ssh/known_hosts:1: [hashed name]
    ~/.ssh/known_hosts:4: [hashed name]
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
/usr/bin/ssh-copy-id: INFO: attempting to log in with the new key(s), to filter
out any that are already installed
/usr/bin/ssh-copy-id: INFO: 1 key(s) remain to be installed -- if you are prompt
ed now it is to install the new keys
h-user@h-secondary2's password:

Number of key(s) added: 1

Now try logging into the machine, with:   "ssh 'h-user@h-secondary2'"
and check to make sure that only the key(s) you wanted were added.
```

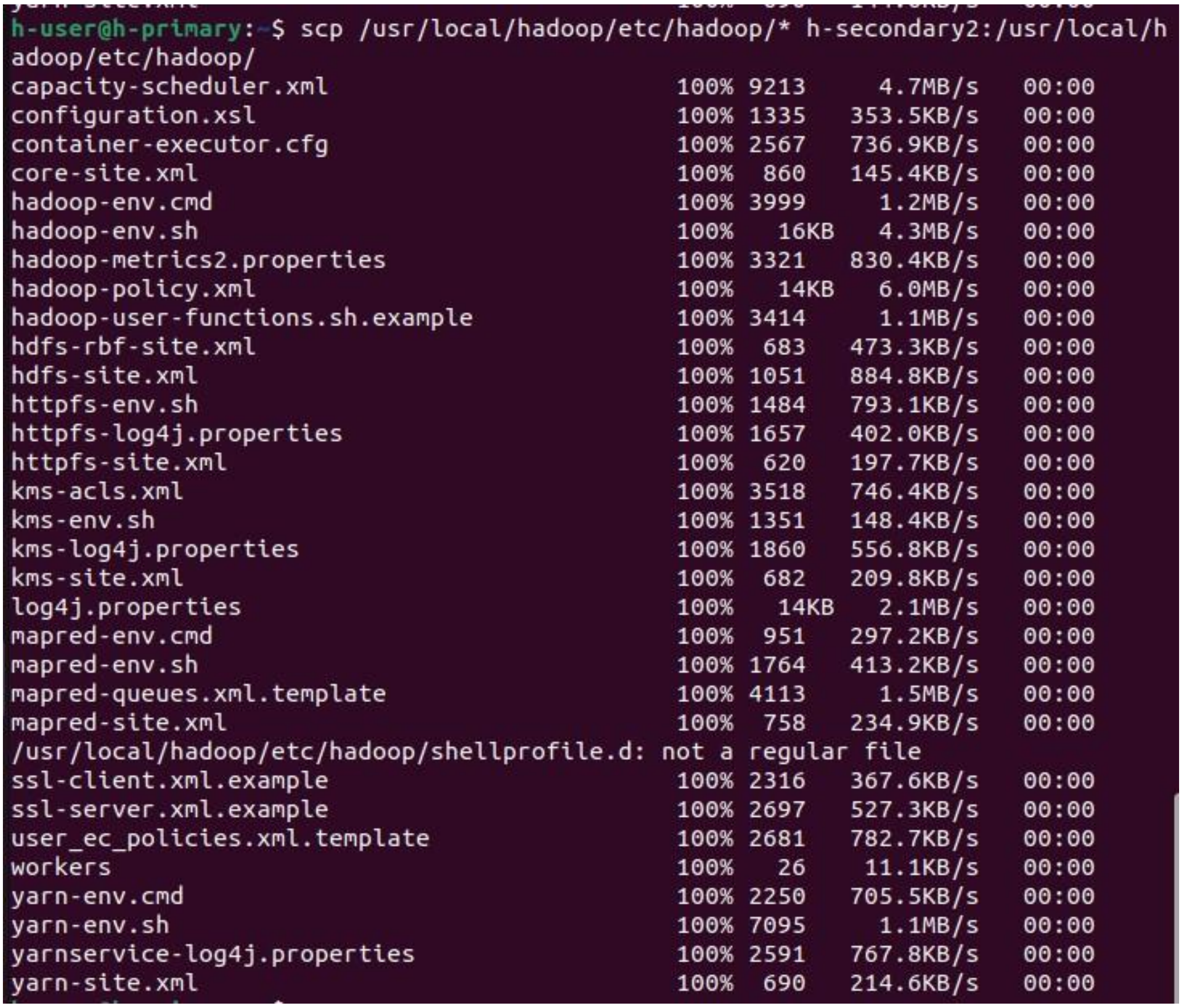
18) Configurer le service port de Hadoop ainsi que hdfs :

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

<property>
<name>fs.defaultFS</name>
<value>hdfs://h-primary:9000</value>
</property>

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

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

19) Copier ces configurations dans les autres machines (slaves) :

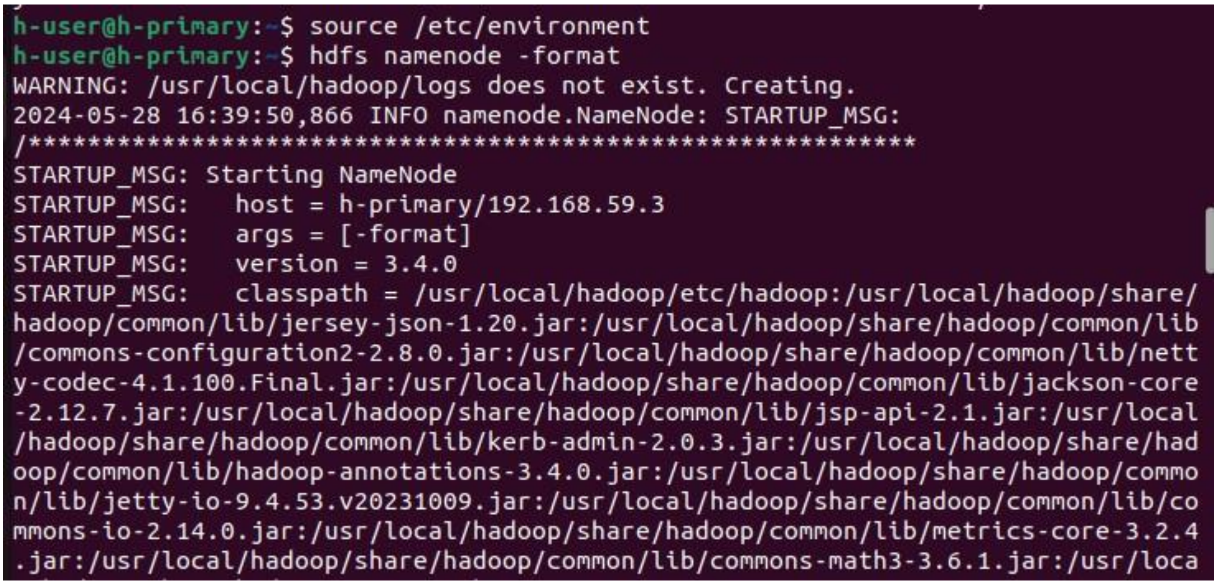


```
h-user@h-primary:~$ scp /usr/local/hadoop/etc/hadoop/* h-secondary1:/usr/local/h
adoop/etc/hadoop/
capacity-scheduler.xml                        100% 9213     3.6MB/s   00:00
configuration.xsl                             100% 1335   625.0KB/s   00:00
container-executor.cfg                        100% 2567     1.3MB/s   00:00
core-site.xml                                 100%  860   543.1KB/s   00:00
hadoop-env.cmd                                100% 3999   852.7KB/s   00:00
hadoop-env.sh                                 100%   16KB   2.8MB/s   00:00
hadoop-metrics2.properties                    100% 3321   683.0KB/s   00:00
hadoop-policy.xml                             100%   14KB   3.7MB/s   00:00
hadoop-user-functions.sh.example              100% 3414   299.5KB/s   00:00
hdfs-rbf-site.xml                             100%  683    84.3KB/s   00:00
hdfs-site.xml                                 100% 1051   324.0KB/s   00:00
httpfs-env.sh                                 100% 1484   289.6KB/s   00:00
httpfs-log4j.properties                       100% 1657   194.4KB/s   00:00
httpfs-site.xml                               100%  620    84.4KB/s   00:00
kms-acls.xml                                  100% 3518   541.2KB/s   00:00
kms-env.sh                                    100% 1351   344.3KB/s   00:00
kms-log4j.properties                          100% 1860   327.4KB/s   00:00
kms-site.xml                                  100%  682   148.7KB/s   00:00
log4j.properties                              100%   14KB   2.9MB/s   00:00
mapred-env.cmd                                100%  951   145.8KB/s   00:00
mapred-env.sh                                 100% 1764   190.7KB/s   00:00
mapred-queues.xml.template                    100% 4113     1.0MB/s   00:00
mapred-site.xml                               100%  758   249.7KB/s   00:00
/usr/local/hadoop/etc/hadoop/shellprofile.d: not a regular file
ssl-client.xml.example                        100% 2316   858.3KB/s   00:00
ssl-server.xml.example                        100% 2697   392.2KB/s   00:00
user_ec_policies.xml.template                 100% 2681     1.1MB/s   00:00
workers                                       100%   26    16.0KB/s   00:00
yarn-env.cmd                                  100% 2250   510.5KB/s   00:00
yarn-env.sh                                   100% 7095     1.2MB/s   00:00
yarnservice-log4j.properties                  100% 2591   208.0KB/s   00:00
yarn-site.xml                                 100%  690   144.6KB/s   00:00
h-user@h-primary:~$
```



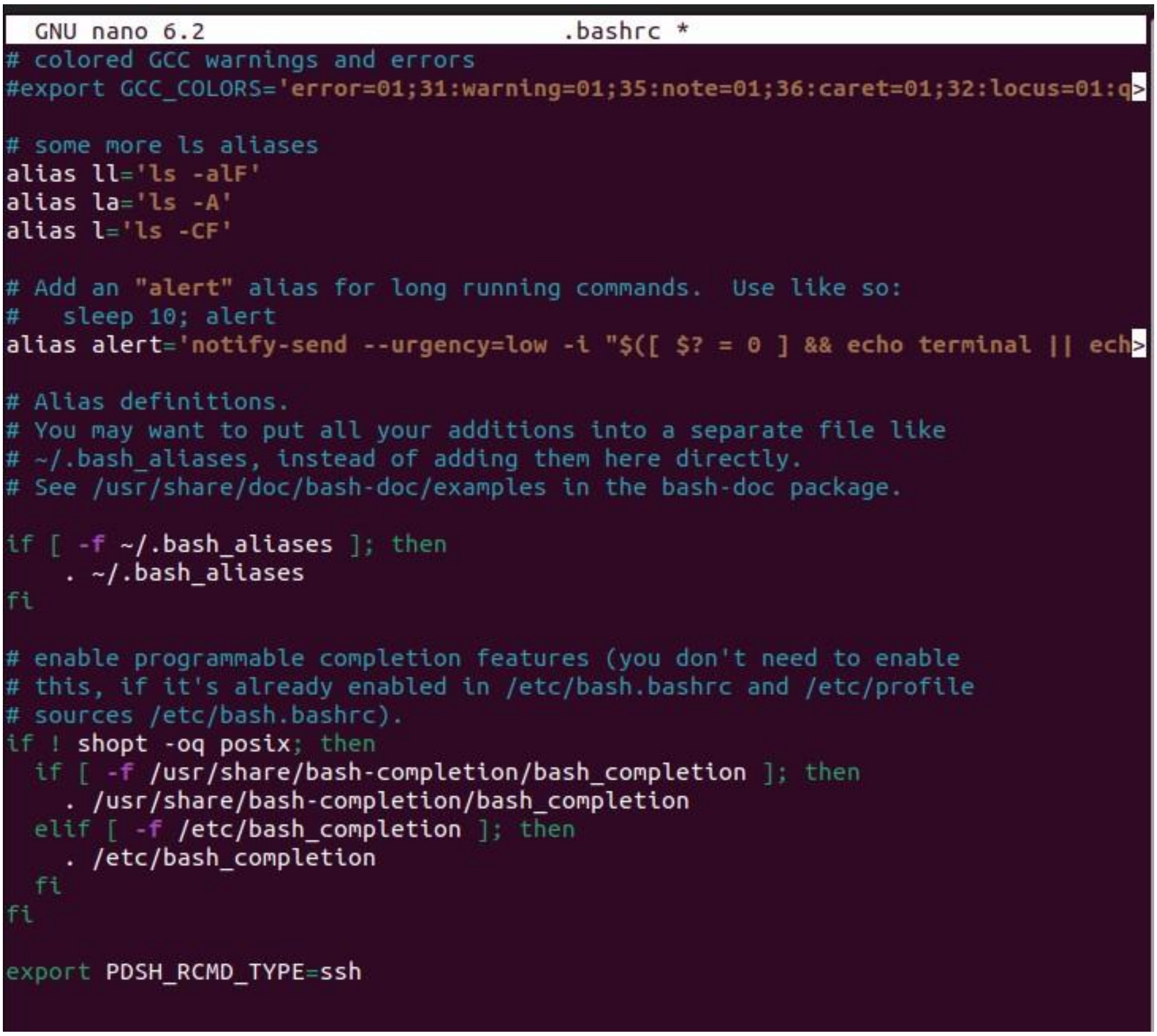
```
h-user@h-primary:~$ scp /usr/local/hadoop/etc/hadoop/* h-secondary2:/usr/local/h
adoop/etc/hadoop/
capacity-scheduler.xml                         100% 9213     4.7MB/s   00:00
configuration.xsl                              100% 1335   353.5KB/s   00:00
container-executor.cfg                         100% 2567   736.9KB/s   00:00
core-site.xml                                  100%  860   145.4KB/s   00:00
hadoop-env.cmd                                 100% 3999     1.2MB/s   00:00
hadoop-env.sh                                  100%   16KB   4.3MB/s   00:00
hadoop-metrics2.properties                     100% 3321   830.4KB/s   00:00
hadoop-policy.xml                              100%   14KB   6.0MB/s   00:00
hadoop-user-functions.sh.example               100% 3414     1.1MB/s   00:00
hdfs-rbf-site.xml                              100%  683   473.3KB/s   00:00
hdfs-site.xml                                  100% 1051   884.8KB/s   00:00
httpfs-env.sh                                  100% 1484   793.1KB/s   00:00
httpfs-log4j.properties                        100% 1657   402.0KB/s   00:00
httpfs-site.xml                                100%  620   197.7KB/s   00:00
kms-acls.xml                                   100% 3518   746.4KB/s   00:00
kms-env.sh                                     100% 1351   148.4KB/s   00:00
kms-log4j.properties                           100% 1860   556.8KB/s   00:00
kms-site.xml                                   100%  682   209.8KB/s   00:00
log4j.properties                               100%   14KB   2.1MB/s   00:00
mapred-env.cmd                                 100%  951   297.2KB/s   00:00
mapred-env.sh                                  100% 1764   413.2KB/s   00:00
mapred-queues.xml.template                     100% 4113     1.5MB/s   00:00
mapred-site.xml                                100%  758   234.9KB/s   00:00
/usr/local/hadoop/etc/hadoop/shellprofile.d: not a regular file
ssl-client.xml.example                         100% 2316   367.6KB/s   00:00
ssl-server.xml.example                         100% 2697   527.3KB/s   00:00
user_ec_policies.xml.template                  100% 2681   782.7KB/s   00:00
workers                                        100%   26    11.1KB/s   00:00
yarn-env.cmd                                   100% 2250   705.5KB/s   00:00
yarn-env.sh                                    100% 7095     1.1MB/s   00:00
yarnservice-log4j.properties                   100% 2591   767.8KB/s   00:00
yarn-site.xml                                  100%  690   214.6KB/s   00:00
```

20) Formater et démarrer le système HDFS

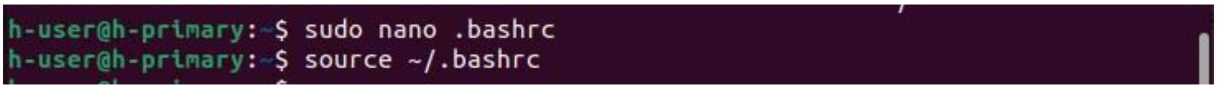


```
h-user@h-primary:~$ source /etc/environment
h-user@h-primary:~$ hdfs namenode -format
WARNING: /usr/local/hadoop/logs does not exist. Creating.
2024-05-28 16:39:50,866 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = h-primary/192.168.59.3
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.4.0
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/
hadoop/common/lib/jersey-json-1.20.jar:/usr/local/hadoop/share/hadoop/common/lib
/commons-configuration2-2.8.0.jar:/usr/local/hadoop/share/hadoop/common/lib/nett
y-codec-4.1.100.Final.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-core
-2.12.7.jar:/usr/local/hadoop/share/hadoop/common/lib/jsp-api-2.1.jar:/usr/local
/hadoop/share/hadoop/common/lib/kerb-admin-2.0.3.jar:/usr/local/hadoop/share/had
oop/common/lib/hadoop-annotations-3.4.0.jar:/usr/local/hadoop/share/hadoop/commo
n/lib/jetty-io-9.4.53.v20231009.jar:/usr/local/hadoop/share/hadoop/common/lib/co
mmons-io-2.14.0.jar:/usr/local/hadoop/share/hadoop/common/lib/metrics-core-3.2.4
.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-math3-3.6.1.jar:/usr/loca
```

21) Vérifier si .bahsrc est configuré :

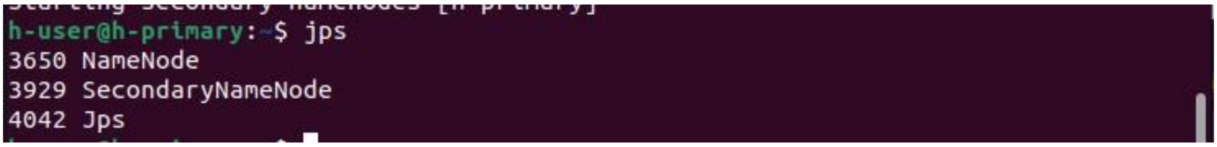


```
  GNU nano 6.2                              .bashrc *
# colored GCC warnings and errors
#export GCC_COLORS='error=01;31:warning=01;35:note=01;36:caret=01;32:locus=01:q>

# some more ls aliases
alias ll='ls -alF'
alias la='ls -A'
alias l='ls -CF'

# Add an "alert" alias for long running commands.  Use like so:
#   sleep 10; alert
alias alert='notify-send --urgency=low -i "$([ $? = 0 ] && echo terminal || ech>

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

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

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

export PDSH_RCMD_TYPE=ssh
```

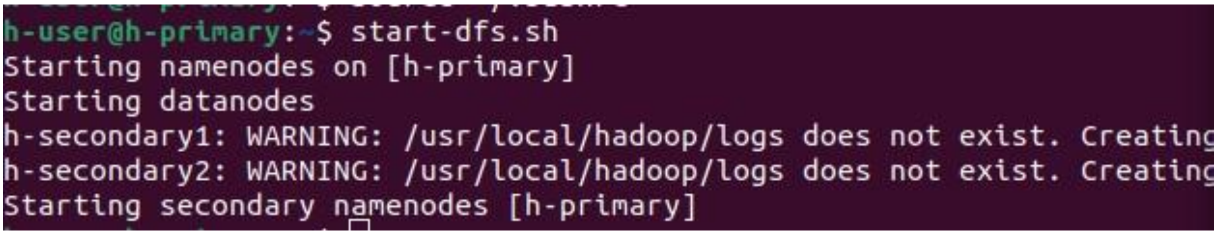


```
h-user@h-primary:~$ sudo nano .bashrc
h-user@h-primary:~$ source ~/.bashrc
```

22) Démarrer le service et Vérifier si les machines fonctionnent correctement :

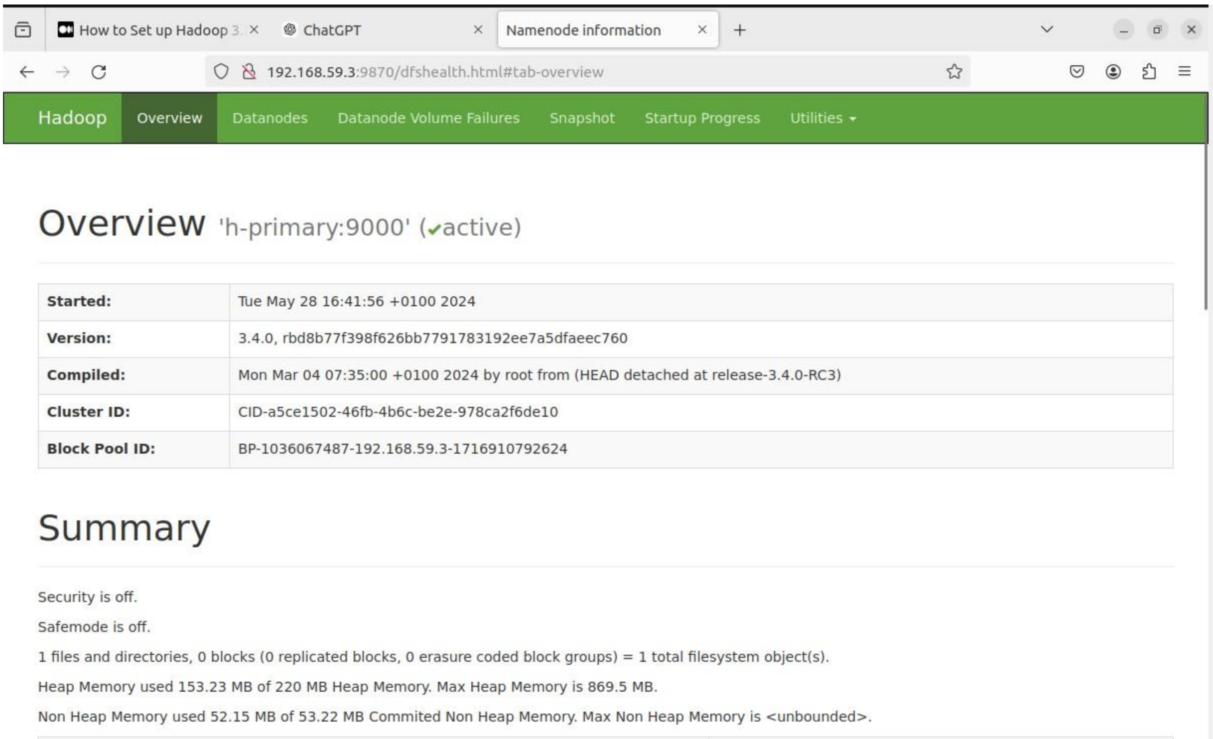


```
h-user@h-primary:~$ jps
3650 NameNode
3929 SecondaryNameNode
4042 Jps
```



```
h-user@h-primary:~$ start-dfs.sh
Starting namenodes on [h-primary]
Starting datanodes
h-secondary1: WARNING: /usr/local/hadoop/logs does not exist. Creating
h-secondary2: WARNING: /usr/local/hadoop/logs does not exist. Creating
Starting secondary namenodes [h-primary]
```

23) Résultat :



Overview 'h-primary:9000' (✓active)

Started:	Tue May 28 16:41:56 +0100 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled:	Mon Mar 04 07:35:00 +0100 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-a5ce1502-46fb-4b6c-be2e-978ca2f6de10
Block Pool ID:	BP-1036067487-192.168.59.3-1716910792624

Summary

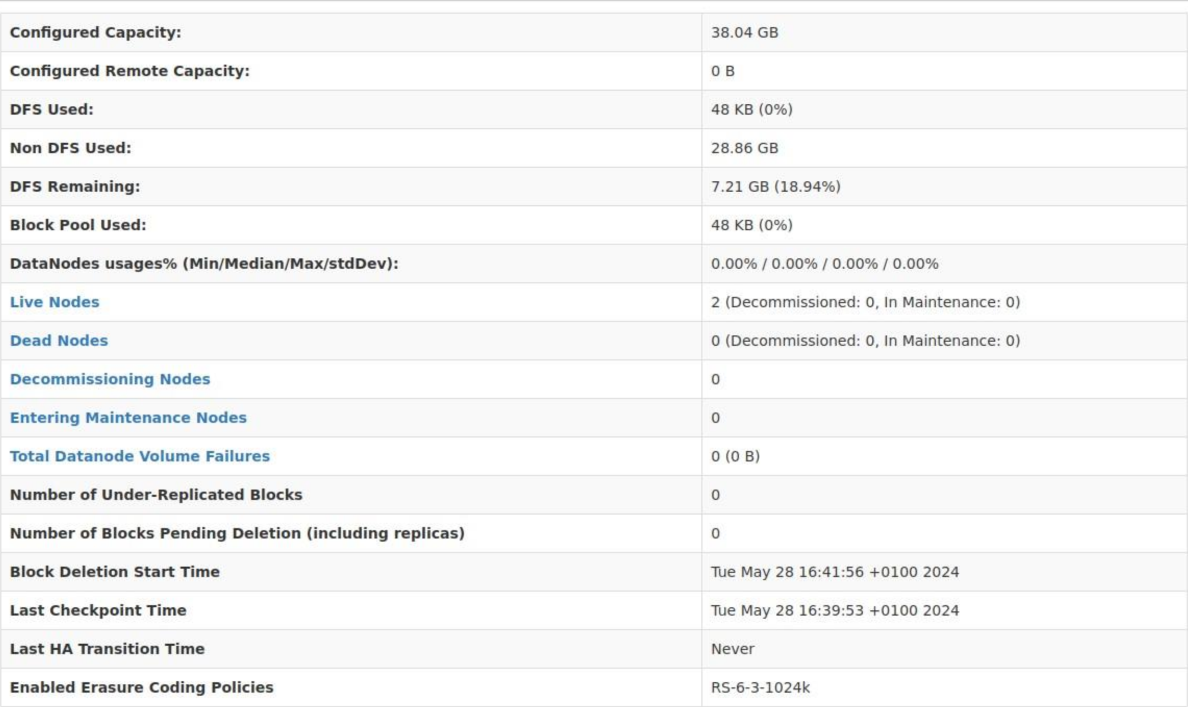
Security is off.

Safemode is off.

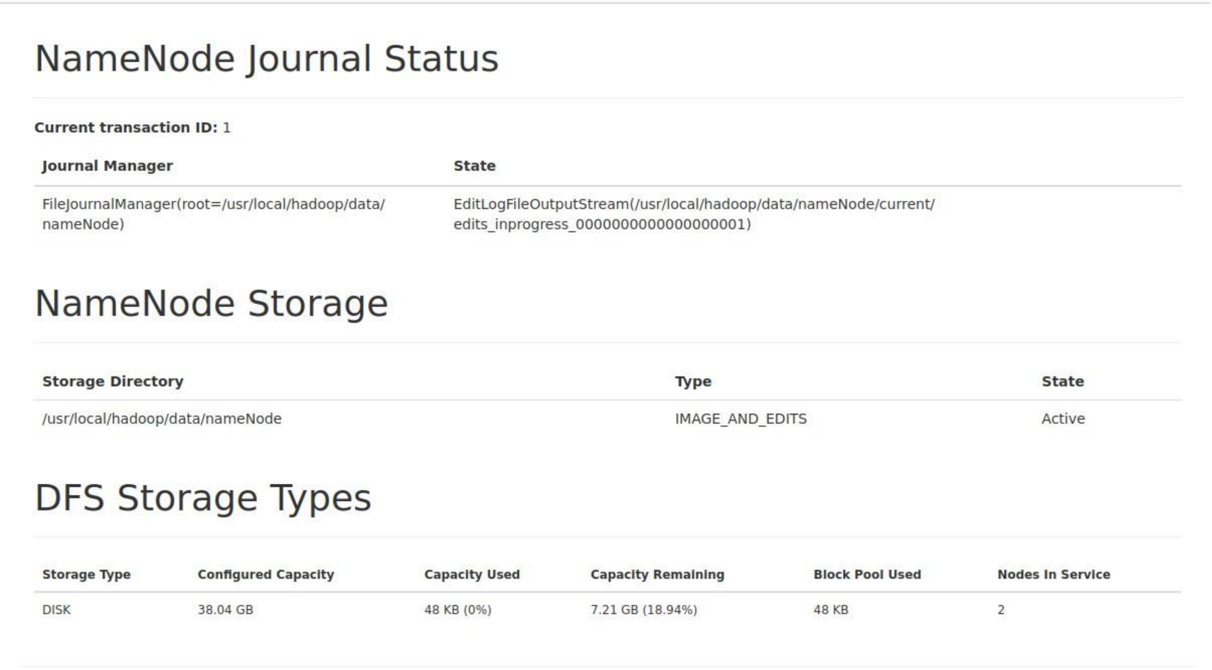
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).

Heap Memory used 153.23 MB of 220 MB Heap Memory. Max Heap Memory is 869.5 MB.

Non Heap Memory used 52.15 MB of 53.22 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.



Configured Capacity:	38.04 GB
Configured Remote Capacity:	0 B
DFS Used:	48 KB (0%)
Non DFS Used:	28.86 GB
DFS Remaining:	7.21 GB (18.94%)
Block Pool Used:	48 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	2 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Tue May 28 16:41:56 +0100 2024
Last Checkpoint Time	Tue May 28 16:39:53 +0100 2024
Last HA Transition Time	Never
Enabled Erasure Coding Policies	RS-6-3-1024k



NameNode Journal Status

Current transaction ID: 1

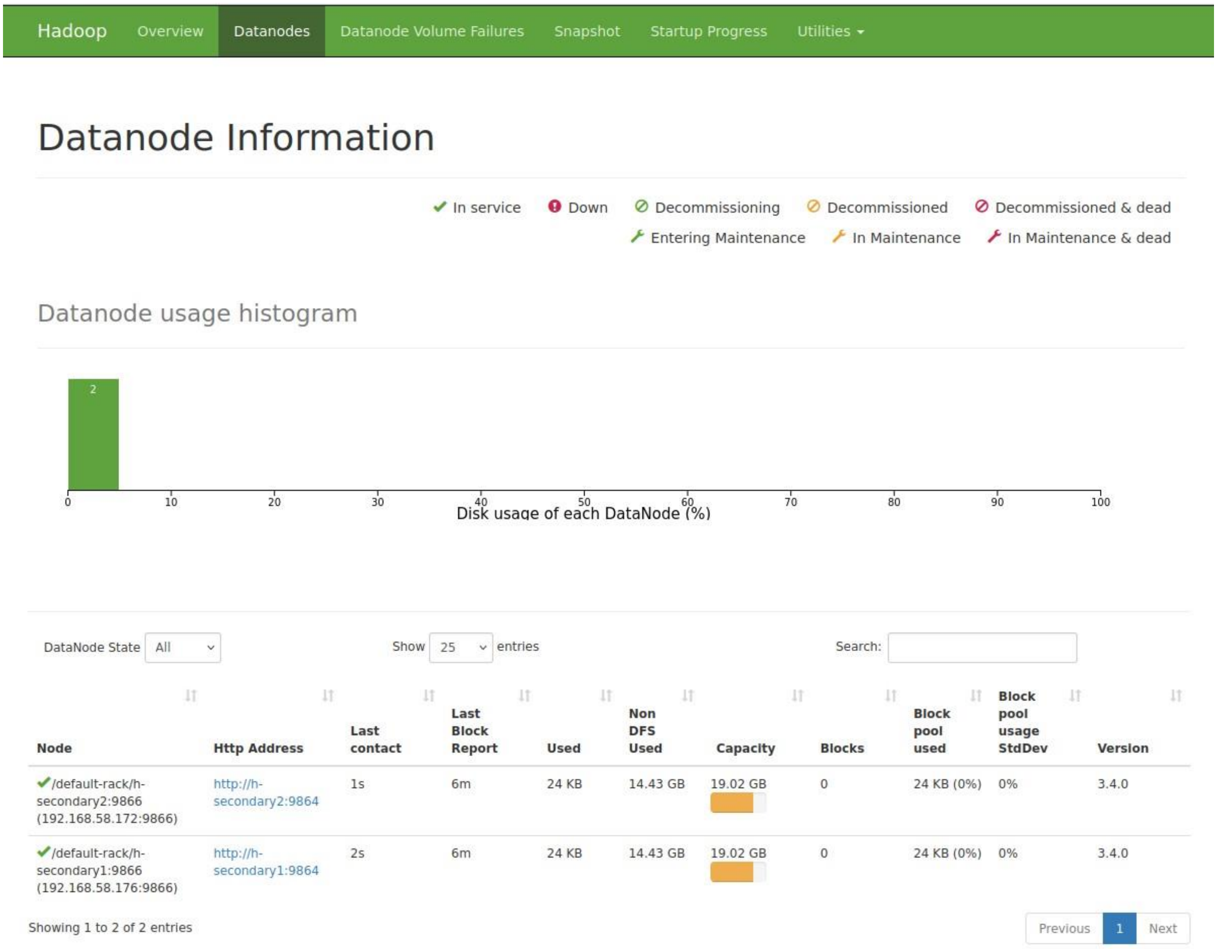
Journal Manager	State
FileJournalManager(root=/usr/local/hadoop/data/nameNode)	EditLogFileOutputStream(/usr/local/hadoop/data/nameNode/current/edits_inprogress_0000000000000000001)

NameNode Storage

Storage Directory	Type	State
/usr/local/hadoop/data/nameNode	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	38.04 GB	48 KB (0%)	7.21 GB (18.94%)	48 KB	2



Datanode Information

✓ In service ● Down ⊘ Decommissioning ⊘ Decommissioned ⊘ Decommissioned & dead
Entering Maintenance In Maintenance In Maintenance & dead

Datanode usage histogram

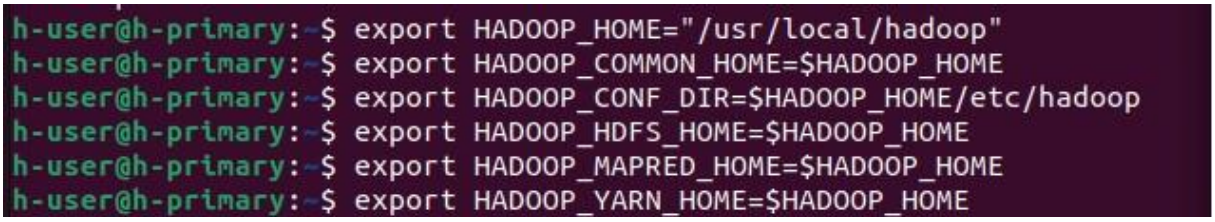
DataNode State All Show 25 entries Search:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Block pool usage StdDev	Version
✓/default-rack/h-secondary2:9866 (192.168.58.172:9866)	http://h-secondary2:9864	1s	6m	24 KB	14.43 GB	19.02 GB	0	24 KB (0%)	0%	3.4.0
✓/default-rack/h-secondary1:9866 (192.168.58.176:9866)	http://h-secondary1:9864	2s	6m	24 KB	14.43 GB	19.02 GB	0	24 KB (0%)	0%	3.4.0

Showing 1 to 2 of 2 entries

Previous 1 Next

24) Configurer yarn :

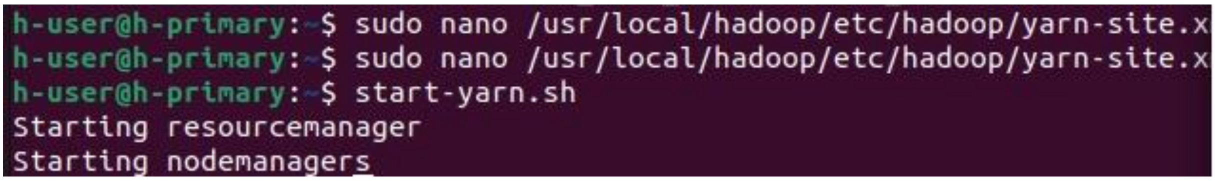


```
h-user@h-primary:~$ export HADOOP_HOME="/usr/local/hadoop"
h-user@h-primary:~$ export HADOOP_COMMON_HOME=$HADOOP_HOME
h-user@h-primary:~$ export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
h-user@h-primary:~$ export HADOOP_HDFS_HOME=$HADOOP_HOME
h-user@h-primary:~$ export HADOOP_MAPRED_HOME=$HADOOP_HOME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

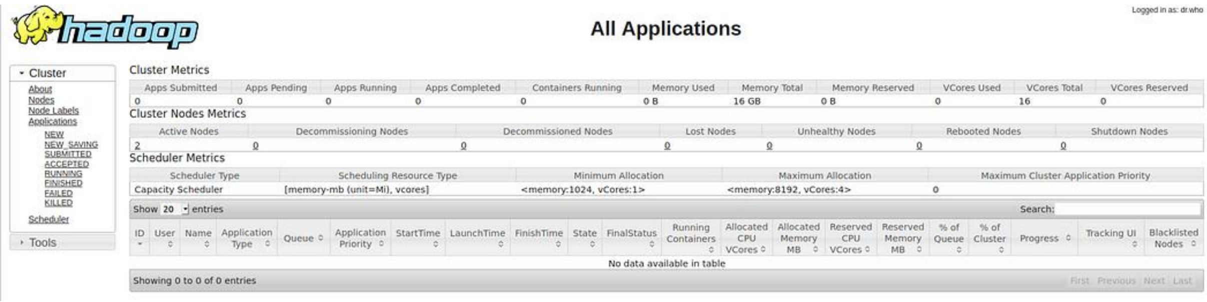
25) Changer la configuration de yarn dans les deux esclaves (h-secondary1 et h-secondary2)

```
sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.xml
<property>
<name>yarn.resourcemanager.hostname</name>
<value>h-primary</value>
</property>
```

26) Démarrer yarn:



```
h-user@h-primary:~$ sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.x
h-user@h-primary:~$ sudo nano /usr/local/hadoop/etc/hadoop/yarn-site.x
h-user@h-primary:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```



Logged in as: dr.who

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	16 GB	0 B	0	16	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
2	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0

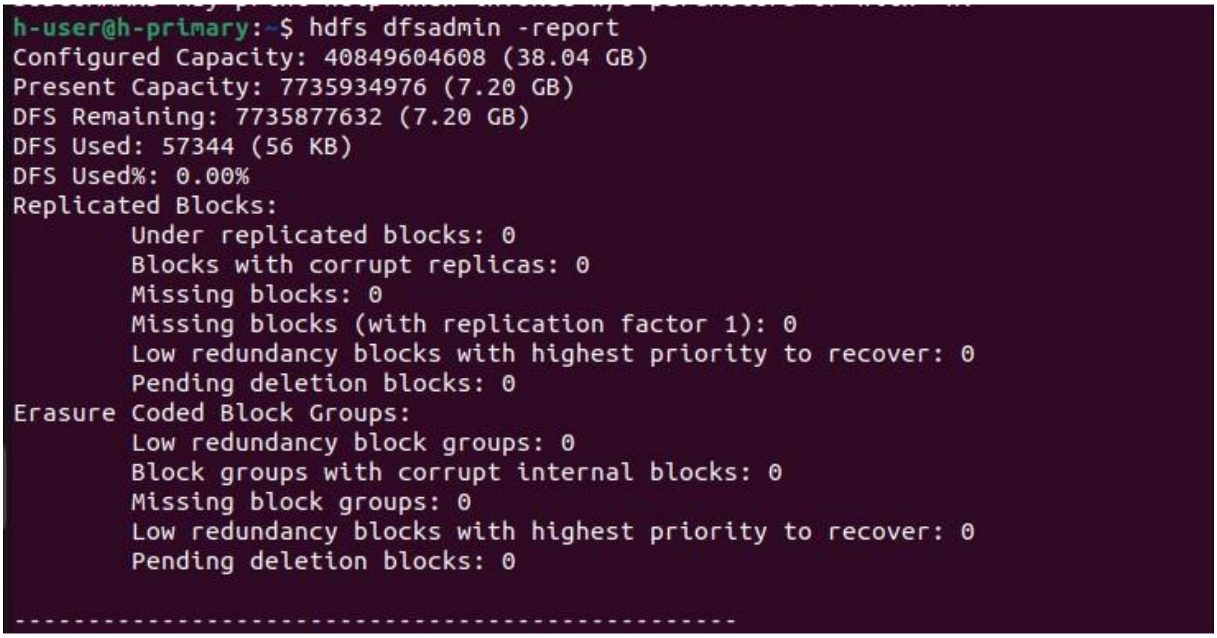
Show 20 entries Search:

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU VCores	Allocated Memory MB	Reserved CPU VCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes

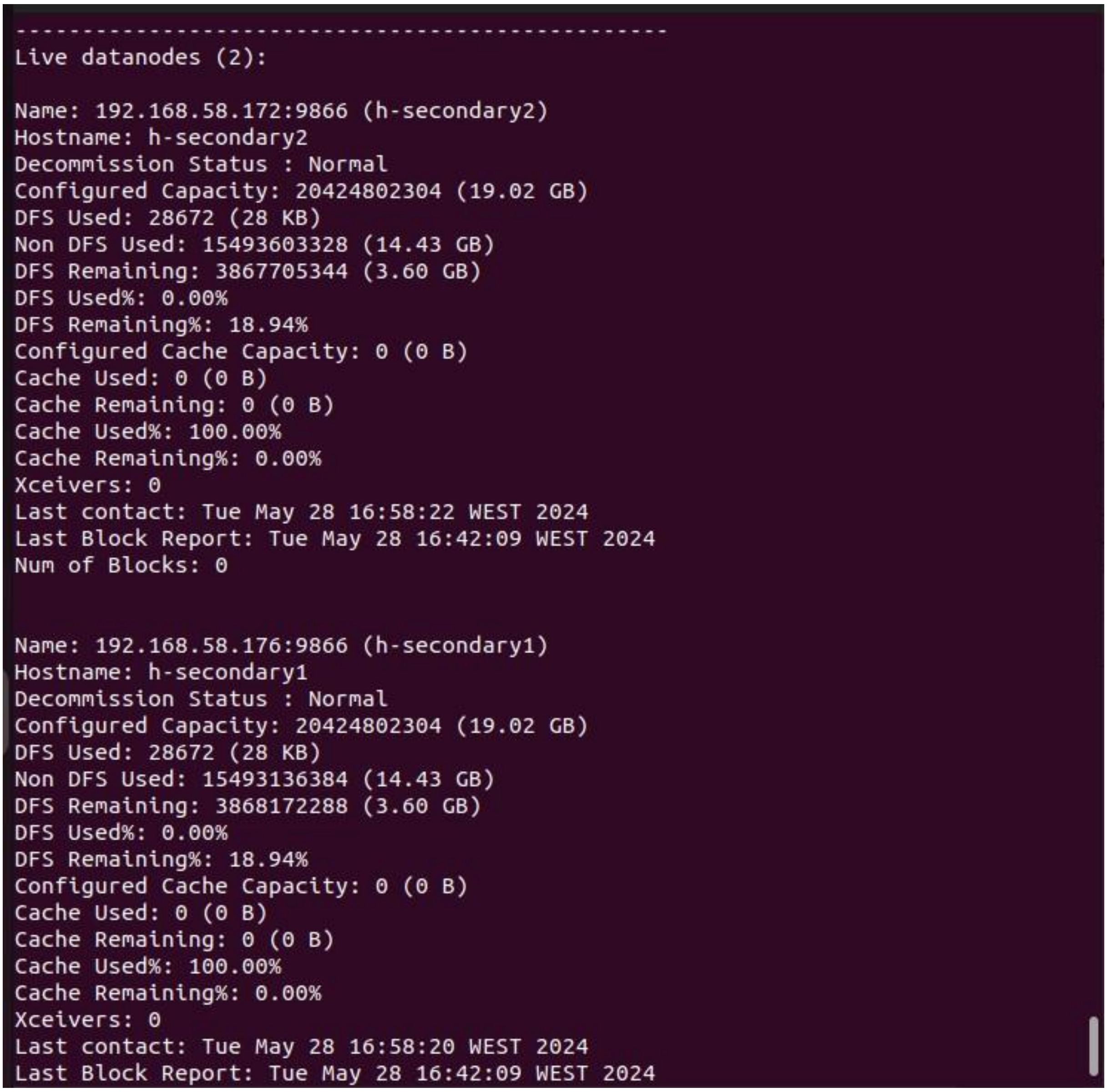
No data available in table

Showing 0 to 0 of 0 entries First Previous Next Last

27) Vérifier si tout est bien configuré en utilisant hdfs dfsadmin -report :

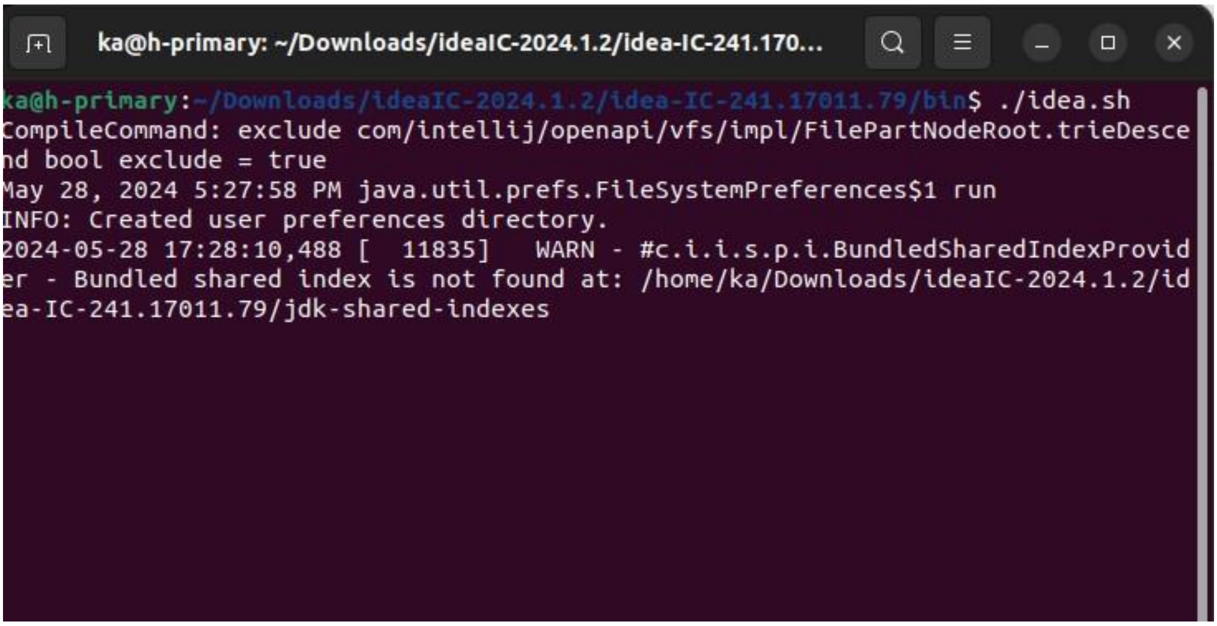


```
h-user@h-primary:~$ hdfs dfsadmin -report
Configured Capacity: 40849604608 (38.04 GB)
Present Capacity: 7735934976 (7.20 GB)
DFS Remaining: 7735877632 (7.20 GB)
DFS Used: 57344 (56 KB)
DFS Used%: 0.00%
Replicated Blocks:
        Under replicated blocks: 0
        Blocks with corrupt replicas: 0
        Missing blocks: 0
        Missing blocks (with replication factor 1): 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0
Erasure Coded Block Groups:
        Low redundancy block groups: 0
        Block groups with corrupt internal blocks: 0
        Missing block groups: 0
        Low redundancy blocks with highest priority to recover: 0
        Pending deletion blocks: 0

-------------------------------------------------
```

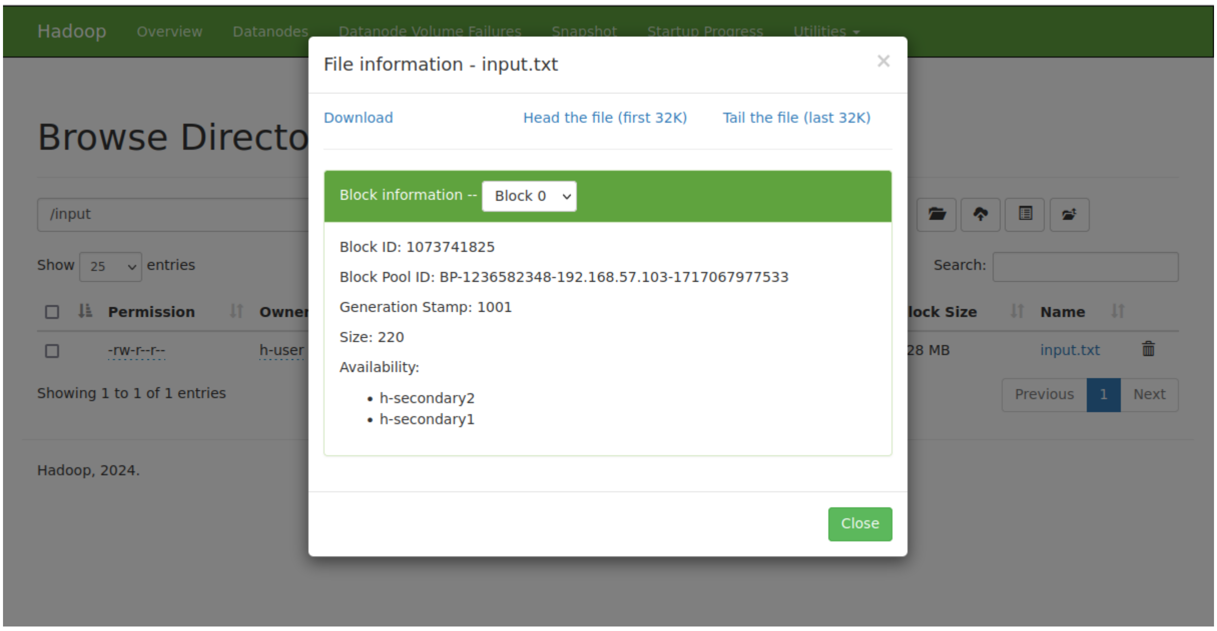


```
-------------------------------------------------
Live datanodes (2):

Name: 192.168.58.172:9866 (h-secondary2)
Hostname: h-secondary2
Decommission Status : Normal
Configured Capacity: 20424802304 (19.02 GB)
DFS Used: 28672 (28 KB)
Non DFS Used: 15493603328 (14.43 GB)
DFS Remaining: 3867705344 (3.60 GB)
DFS Used%: 0.00%
DFS Remaining%: 18.94%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Tue May 28 16:58:22 WEST 2024
Last Block Report: Tue May 28 16:42:09 WEST 2024
Num of Blocks: 0


Name: 192.168.58.176:9866 (h-secondary1)
Hostname: h-secondary1
Decommission Status : Normal
Configured Capacity: 20424802304 (19.02 GB)
DFS Used: 28672 (28 KB)
Non DFS Used: 15493136384 (14.43 GB)
DFS Remaining: 3868172288 (3.60 GB)
DFS Used%: 0.00%
DFS Remaining%: 18.94%
Configured Cache Capacity: 0 (0 B)
Cache Used: 0 (0 B)
Cache Remaining: 0 (0 B)
Cache Used%: 100.00%
Cache Remaining%: 0.00%
Xceivers: 0
Last contact: Tue May 28 16:58:20 WEST 2024
Last Block Report: Tue May 28 16:42:09 WEST 2024
```

28) installer intellij idea pour appliquer le programme MapReduce :


ka@h-primary: ~/Downloads/idealC-2024.1.2/idea-IC-241.170...

```
ka@h-primary:~/Downloads/ideaIC-2024.1.2/idea-IC-241.17011.79/bin$ ./idea.sh
CompileCommand: exclude com/intellij/openapi/vfs/impl/FilePartNodeRoot.trieDesce
nd bool exclude = true
May 28, 2024 5:27:58 PM java.util.prefs.FileSystemPreferences$1 run
INFO: Created user preferences directory.
2024-05-28 17:28:10,488 [  11835]   WARN - #c.i.i.s.p.i.BundledSharedIndexProvid
er - Bundled shared index is not found at: /home/ka/Downloads/ideaIC-2024.1.2/id
ea-IC-241.17011.79/jdk-shared-indexes
```



File information - input.txt

Block ID: 1073741825

Block Pool ID: BP-1236582348-192.168.57.103-1717067977533

Generation Stamp: 1001

Size: 220

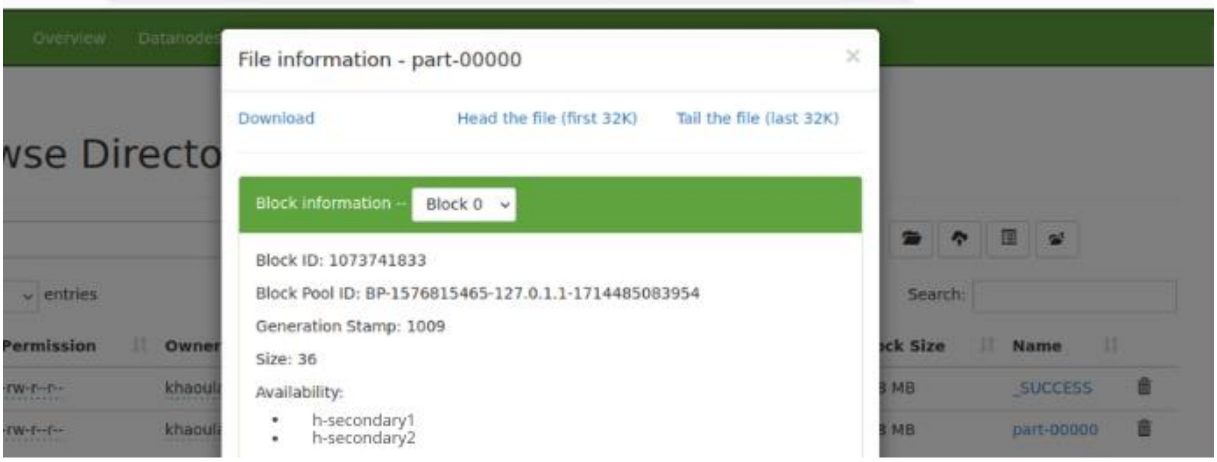
Availability:

- h-secondary2
- h-secondary1

Programme MapReduce: WordCount.

On reprend les mêmes étapes qui ont été faites dans la première partie relative au cluster à nœud unique.

Voici les résultats obtenus :



File information - part-00000

Block ID: 1073741833

Block Pool ID: BP-1576815465-127.0.1.1-1714485083954

Generation Stamp: 1009

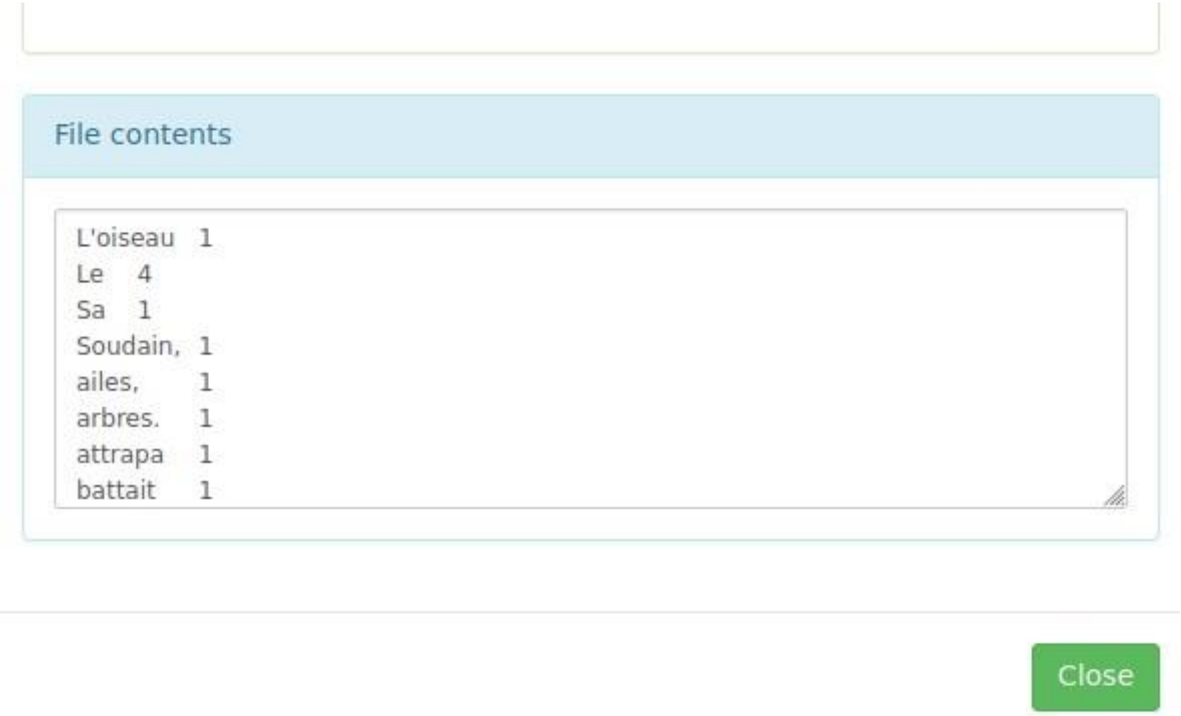
Size: 36

Availability:

- h-secondary1
- h-secondary2

Permission	Owner	ck Size	Name
-rw-r--r--	khaoul	3 MB	_SUCCESS
-rw-r--r--	khaoul	3 MB	part-00000

On obtient le même résultat.



```
L'oiseau  1
Le    4
Sa    1
Soudain,  1
ailes,    1
arbres.   1
attrapa   1
battait   1
```

Fin.