Fode Mangane

Installer et configurer un cluster Hadoop à 3 nœuds

Nous commencerons par installer Hadoop sur les trois nœuds, à savoir node1, node2 et nodemaster, en suivant simplement la documentation officielle de Hadoop(le lien ici).

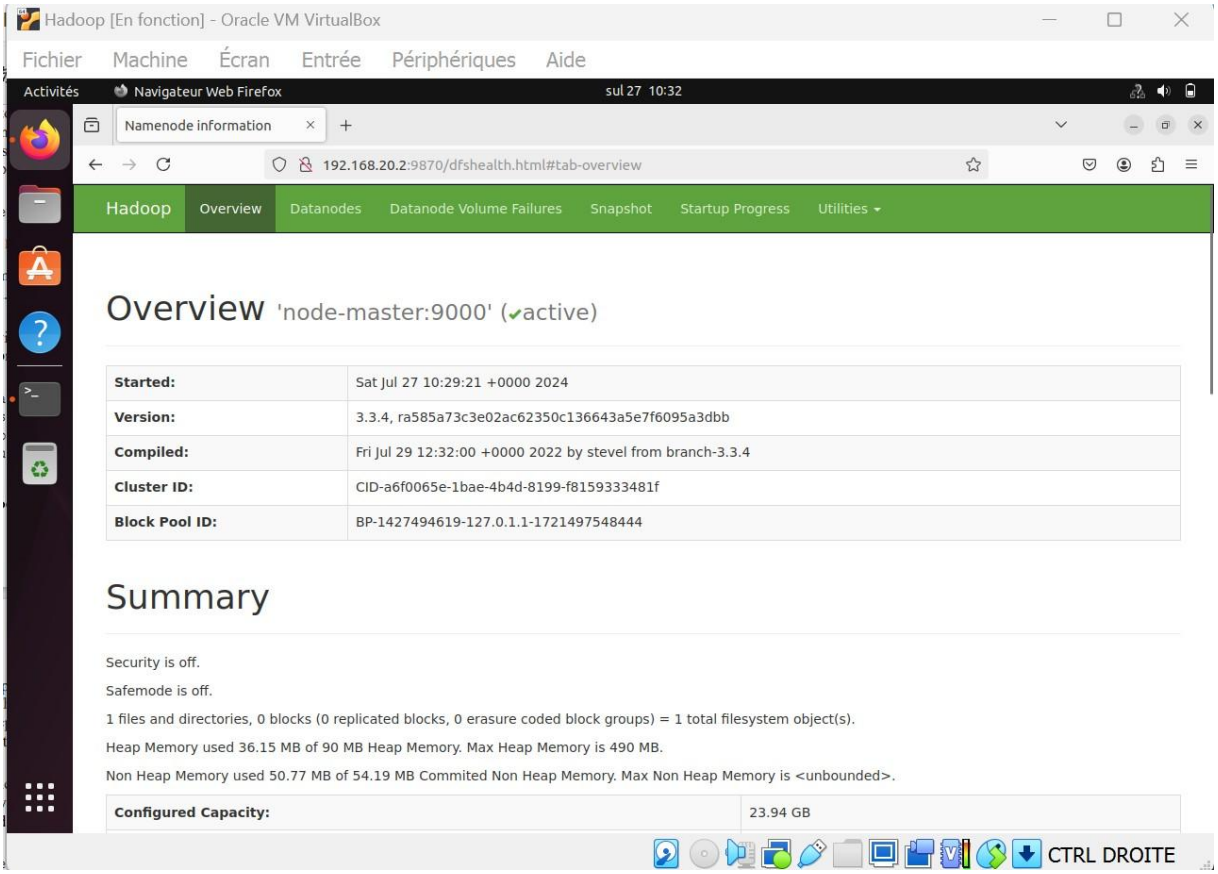
Notes :

Node1= 192.168.1.4

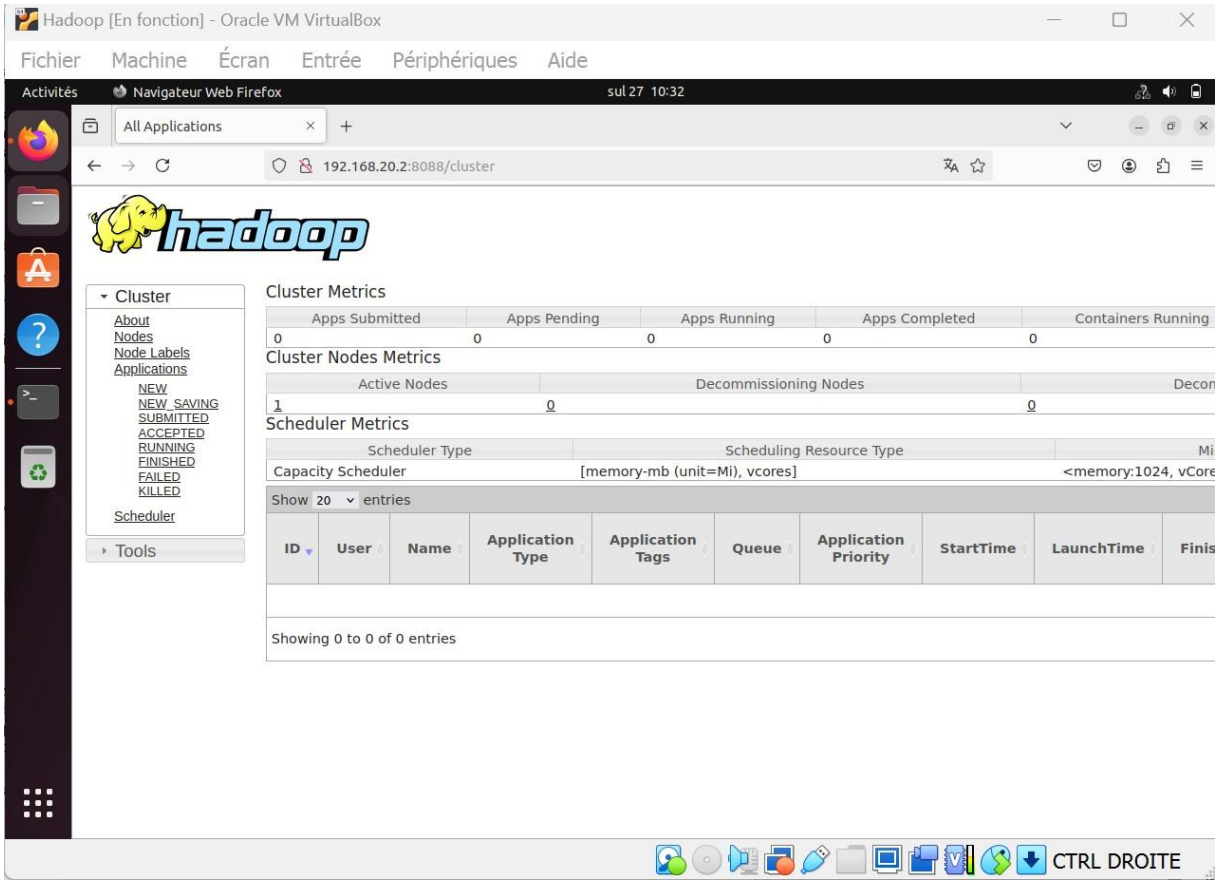
Node2= 192.168.1.6

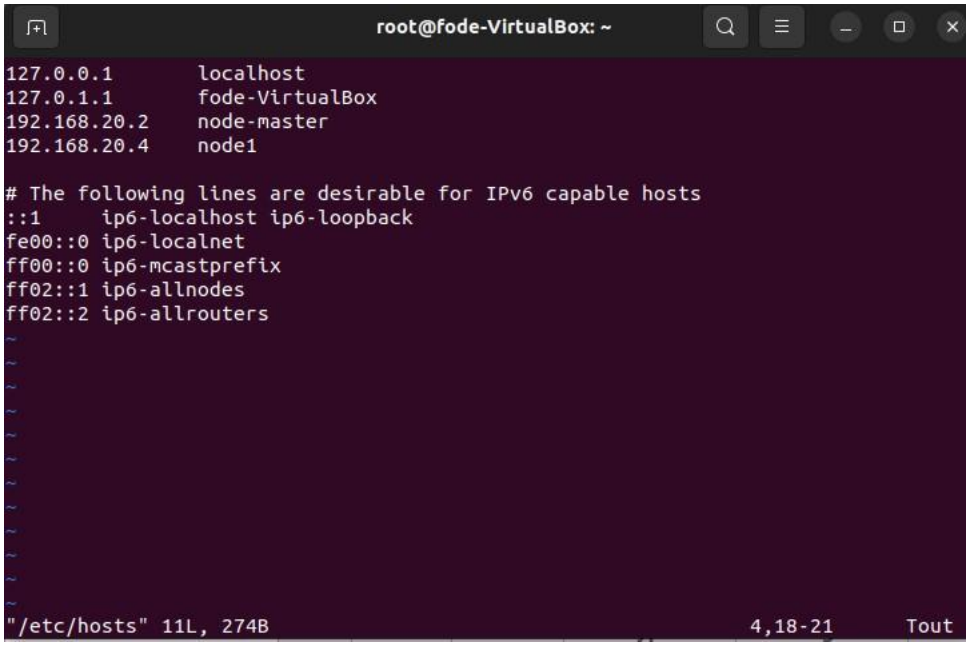
Node-master= 192.168.1.2

```
hadoop@fode-VirtualBox:/root$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@fode-VirtualBox:/root$
```

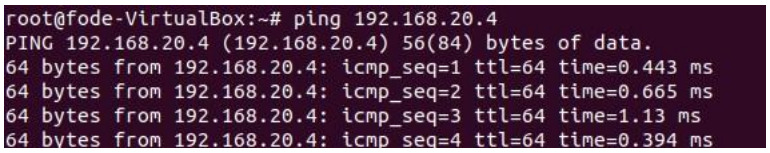


Fodemangane@gmail.com





Sur node-master le ping passe



Sur node1

```
root@fode-VirtualBox:~# ping 192.168.20.2
PING 192.168.20.2 (192.168.20.2) 56(84) bytes of data.
64 bytes from 192.168.20.2: icmp_seq=1 ttl=64 time=1.54 ms
64 bytes from 192.168.20.2: icmp_seq=2 ttl=64 time=0.415 ms
64 bytes from 192.168.20.2: icmp_seq=3 ttl=64 time=0.604 ms
64 bytes from 192.168.20.2: icmp_seq=4 ttl=64 time=0.359 ms
```

```
The key fingerprint is:
SHA256:E0I/j5kXOyOkBDNmW6Xh5J5qJ6LlpJ6/uUcAxsCpvYc root@fode-VirtualBox
The key's randomart image is:
+---[RSA 4096]----+
|= . * =..        |
| * o @ +         |
|o.. . * = .      |
|. .. o = B o     |
|   o. + S *      |
|  E .o   + o     |
|  +.= .          |
| *.o.+           |
|+oo=+            |
+----[SHA256]-----+
root@fode-VirtualBox:~# ssh-copy-id -i $HOME/.ssh/id_rsa.pub hadoop@node-master
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/root/.ssh/id_rsa
.pub"
The authenticity of host 'node-master (192.168.20.2)' can't be established.
ED25519 key fingerprint is SHA256:2MM5tYXLJlvhuBV5MLKkaQ65ZrDx0GNxBAY3/DUOS1o.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])?
```

Vim ~/hadoop/etc/hadoop/workers

```
root@fode-VirtualBox: ~
node1
localhost
~
```

Vim /home/hadoop/hadoop/etc/hadoop/yarn-site.xml

```
root@fode-VirtualBox: ~
        <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CL
SSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_HOME,PATH,LANG,TZ,HADOOP_MAPRE
_HOME</value>
</property>

<property>
<name>yarn.nodemanager.resource.memory-mb</name>
<value> 1536</value>
</property>
<property>
<name>yarn.scheduler.maximum-allocation-mb</name>
<value> 1536</value>
</property>

<property>
<name>yarn.scheduler.minimum-allocation-mb</name>
<value> 128</value>
</property>
<property>
<name>yarn.nodemanager.vmem-check-enabled</name>
<value>false</value>
</property>

-- INSERTION --                                        34,11          95%
```

Vim /home/hadoop/hadoop/etc/hadoop/mapred-site.xml

```
root@fode-VirtualBox: ~

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

<property>
<name>yarn.app.mapreduce.am.resource.mb</name>
<value>512</value>
</property>

<property>
<name>mapreduce.map.memory.mb</name>
<value>256</value>

</property>
<property>
<name>mapreduce.reduce.memory.mb</name>
<value>256</value>
</property>

</configuration>
~
~
~
-- INSERTION --                                          44,17         Bas
```

```
hadoop@node-master:~$ scp hadoop-3.3.4.tar.gz  node1:/home/hadoop
hadoop-3.3.4.tar.gz                           100%  663MB 117.1MB/s   00:05
hadoop@node-master:~$
```

```
hadoop@node-master:~$ ssh node1
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 6.5.0-17-generic x86_64)

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

La maintenance de sécurité étendue pour Applications n'est pas activée.

219 mises à jour peuvent être appliquées immédiatement.
149 de ces mises à jour sont des mises à jour de sécurité.
Pour afficher ces mises à jour supplémentaires, exécuter : apt list --upgradable

Activez ESM Apps pour recevoir des futures mises à jour de sécurité supplémentai
res.
Visitez https://ubuntu.com/esm ou executez : sudo pro status

Failed to connect to https://changelogs.ubuntu.com/meta-release-lts. Check your
Internet connection or proxy settings

Last login: Thu Aug  1 19:20:48 2024 from 192.168.20.4
```

```
hadoop@node-master:~$ tar -xzf hadoop-3.3.4.tar.gz
hadoop@node-master:~$
```

```
hadoop@node-master:~$ cd /home/hadoop
hadoop@node-master:~$ tar -xzf hadoop-3.3.4.tar.gz
hadoop@node-master:~$ mv hadoop-3.3.4 hadoop
hadoop@node-master:~$ exit
déconnexion
Connection to node1 closed.
hadoop@node-master:~$
```

```
hadoop@node-master:~$ ssh node1
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 6.5.0-17-generic x86_64)

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

La maintenance de sécurité étendue pour Applications n'est pas activée.

219 mises à jour peuvent être appliquées immédiatement.
149 de ces mises à jour sont des mises à jour de sécurité.
Pour afficher ces mises à jour supplémentaires, exécuter : apt list --upgra

Activez ESM Apps pour recevoir des futures mises à jour de sécurité supplém
res.
Visitez https://ubuntu.com/esm ou executez : sudo pro status

Failed to connect to https://changelogs.ubuntu.com/meta-release-lts. Check
Internet connection or proxy settings

Last login: Thu Aug  1 19:22:39 2024 from 192.168.20.4
hadoop@node-master:~$ mkdir -p /home/hadoop/hadoop
hadoop/              hadoop-3.3.4.tar.gz
```

```
*hadoop@node2:~$ cd /home/hadoop/
hadoop@node2:~$ scp hadoop-3.3.4.tar.gz  node2:/home/hadoop
hadoop-3.3.4.tar.gz                           61%  408MB 141.7MB/s   00:01 ETA
```

```
hadoop@node2:~$ sudo tar -xzf hadoop-3.3.4.tar.gz
[sudo] Mot de passe de hadoop :
hadoop@node2:~$
```

```
hadoop@node2:~$ sudo tar -xzf hadoop-3.3.4.tar.gz
[sudo] Mot de passe de hadoop :
hadoop@node2:~$ mv hadoop-3.3.4 hadoop
mv: impossible de déplacer 'hadoop-3.3.4' vers 'hadoop': Permission non accordée
hadoop@node2:~$ sudo mv hadoop-3.3.4 hadoop
hadoop@node2:~$
```

```
hadoop@node-master:/root$ for node in node1 node2; do
scp ~/hadoop/etc/hadoop/* $node:/home/hadoop/hadoop/etc/hadoop/;
done
capacity-scheduler.xml                      100% 9213     3.2MB/s   00:00
configuration.xsl                           100% 1335     2.5MB/s   00:00
container-executor.cfg                      100% 2567     1.6MB/s   00:00
core-site.xml                               100%  774     1.7MB/s   00:00
hadoop-env.cmd                              100% 3999   804.0KB/s   00:00
hadoop-env.sh                               100%   16KB   4.8MB/s   00:00
hadoop-metrics2.properties                  100% 3321     2.6MB/s   00:00
hadoop-policy.xml                           100%   11KB   8.8MB/s   00:00
hadoop-user-functions.sh.example            100% 3414     3.3MB/s   00:00
hdfs-rbf-site.xml                           100%  683   874.3KB/s   00:00
hdfs-site.xml                               100%  775   488.8KB/s   00:00
httpfs-env.sh                               100% 1484     1.4MB/s   00:00
```

```
hadoop@node-master:/root$ hdfs namenode -format
2024-08-01 20:34:32,907 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = node-master/192.168.20.2
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.3.4
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/
hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop/share/hadoop/common/
```

```
2024-08-01 20:36:17,646 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at node-master/192.168.20.2
************************************************************/
hadoop@node-master:/root$
```

```
hadoop@node-master:/root$ start-dfs.sh
Starting namenodes on [node-master]
Starting datanodes
Starting secondary namenodes [node-master]
hadoop@node-master:/root$
```

```
hadoop@node-master:/root$ jps
6651 SecondaryNameNode
6220 NameNode
6781 Jps
hadoop@node-master:/root$
```

```
hadoop@node-master:/root$ stop-dfs.sh
Stopping namenodes on [node-master]
Stopping datanodes
Stopping secondary namenodes [node-master]
hadoop@node-master:/root$
```

```
hadoop@node-master:/root$ hdfs dfsadmin -report
Configured Capacity: 0 (0 B)
Present Capacity: 0 (0 B)
DFS Remaining: 0 (0 B)
DFS Used: 0 (0 B)
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

-------------------------------------------------
hadoop@node-master:/root$
```

```
hadoop@node-master:/root$ hdfs dfsadmin -help
hdfs dfsadmin performs DFS administrative commands.
Note: Administrative commands can only be run with superuser permission.
The full syntax is:

hdfs dfsadmin
        [-report [-live] [-dead] [-decommissioning] [-enteringmaintenance] [-inm
aintenance]]
        [-safemode <enter | leave | get | wait | forceExit>]
        [-saveNamespace [-beforeShutdown]]
        [-rollEdits]
```

9870 et 9864

192.168.20.2:9870/dfshealth.html#tab-overview

Hadoop | Overview | Datanodes | Datanode Volume Failures | Snapshot | Startup Progress | Utilities

Overview 'node-master:9000' (active)

Started:	Thu Aug 01 21:07:27 +0000 2024
Version:	3.3.4, ra585a73c3e02ac62350c136643a5e7f6095a3dbb
Compiled:	Fri Jul 29 12:32:00 +0000 2022 by stevel from branch-3.3.4
Cluster ID:	CID-4aaa06d5-8535-43a8-bb45-94a68feb2732
Block Pool ID:	BP-1237826746-192.168.20.2-1722544577381

Summary

Security is off.

Safemode is off.

19 files and directories, 3 blocks (3 replicated blocks, 0 erasure coded block groups) = 22 total filesystem object(s).

Heap Memory used 45.33 MB of 77 MB Heap Memory. Max Heap Memory is 490 MB.

Non Heap Memory used 55.63 MB of 58.56 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	23.94 GB
Configured Remote Capacity:	0 B
DFS Used:	315.17 KB (0%)
Non DFS Used:	17.42 GB
DFS Remaining:	5.28 GB (22.06%)
Block Pool Used:	315.17 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

```
hadoop@node-master:/root$ hdfs dfs -mkdir -p /user/hadoop
hadoop@node-master:/root$ hdfs dfs -mkdir books
hadoop@node-master:/root$ cd /home/hadoop
hadoop@node-master:~$ wget -O alice.txt https://www.gutenberg.org/fles/11/11-0.t
xt
```

```
hadoop@node-master:~$ wget -O holmes.txt https://www.gutenberg.org/ebooks/1661.t
xt.utf-8
--2024-08-01 21:26:32--  https://www.gutenberg.org/ebooks/1661.txt.utf-8
Résolution de www.gutenberg.org (www.gutenberg.org)… 152.19.134.47, 2610:28:3090
:3000:0:bad:cafe:47
Connexion à www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:443… connecté.
requête HTTP transmise, en attente de la réponse… 302 Found
Emplacement : http://www.gutenberg.org/cache/epub/1661/pg1661.txt [suivant]
--2024-08-01 21:26:33--  http://www.gutenberg.org/cache/epub/1661/pg1661.txt
Connexion à www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:80… connecté.
requête HTTP transmise, en attente de la réponse… 302 Found
Emplacement : https://www.gutenberg.org/cache/epub/1661/pg1661.txt [suivant]
--2024-08-01 21:26:34--  https://www.gutenberg.org/cache/epub/1661/pg1661.txt
Connexion à www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:443… connecté.
requête HTTP transmise, en attente de la réponse… 200 OK
Taille : 607648 (593K) [text/plain]
Enregistre : ‘holmes.txt’

holmes.txt          100%[===================>] 593,41K   561KB/s    ds 1,1s

2024-08-01 21:26:36 (561 KB/s) - ‘holmes.txt’ enregistré [607648/607648]

hadoop@node-master:~$
```

```
hadoop@node-master:~$ wget -O frankenstein.txt https://www.gutenberg.org/ebooks/
84.txt.utf-8
--2024-08-01 21:27:04--  https://www.gutenberg.org/ebooks/84.txt.utf-8
Résolution de www.gutenberg.org (www.gutenberg.org)… 152.19.134.47, 2610:28:3090
:3000:0:bad:cafe:47
```

```
hadoop@node-master:~$ hdfs dfs -put alice.txt holmes.txt frankenstein.txt books
2024-08-01 21:28:31,378 WARN hdfs.DataStreamer: DataStreamer Exception
org.apache.hadoop.ipc.RemoteException(java.io.IOException): File /user/hadoop/bo
oks/holmes.txt._COPYING_ could only be written to 0 of the 1 minReplication node
s. There are 0 datanode(s) running and 0 node(s) are excluded in this operation.
```

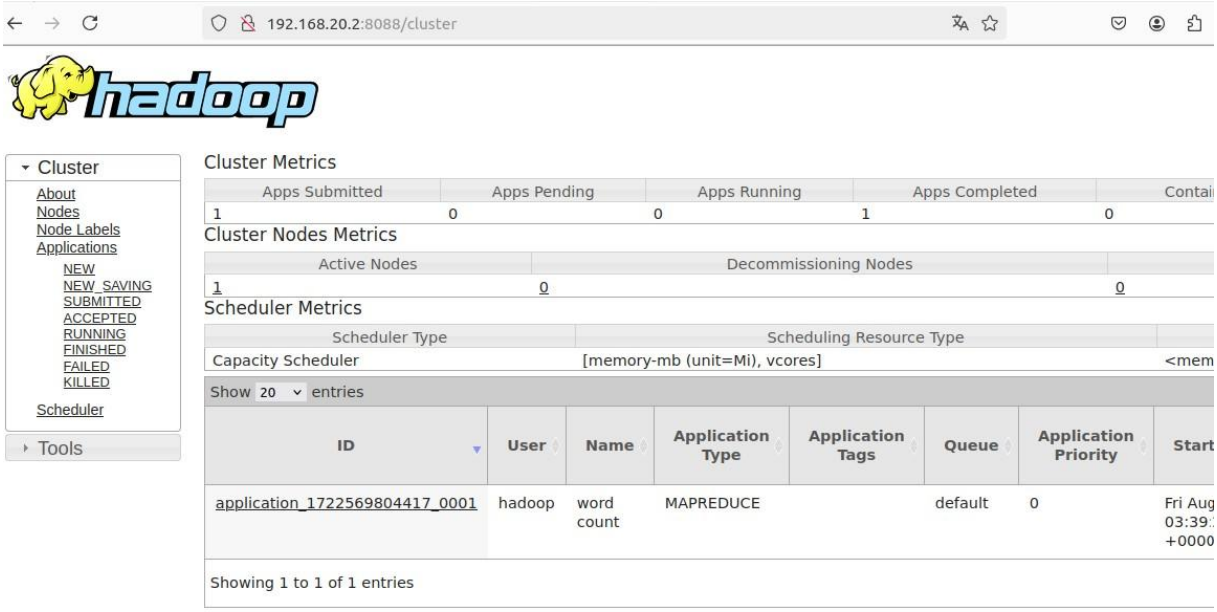
```
hadoop@node-master:~$ hdfs dfs -ls books
Found 1 items
-rw-r--r--   1 hadoop supergroup          0 2024-08-01 21:28 books/alice.txt
hadoop@node-master:~$
```

```
hadoop@node-master:~$ hdfs dfs -get books/alice.txt
get: `alice.txt': File exists
hadoop@node-master:~$
```

```
hadoop@node-master:~$ hdfs dfs -cat books/alice.txt
hadoop@node-master:~$
```

```
hadoop@node-master:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@node-master:~$ stop-yarn.sh
Stopping nodemanagers
Stopping resourcemanager
hadoop@node-master:~$
```

```
hadoop@node-master:~$ yarn application -list
2024-08-01 21:36:33,115 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /0.0.0.0:8032
2024-08-01 21:36:34,316 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0
.0.0:8032. Already tried 0 time(s); retry policy is RetryUpToMaximumCountWithFix
edSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-08-01 21:36:35,319 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0
.0.0:8032. Already tried 1 time(s); retry policy is RetryUpToMaximumCountWithFix
edSleep(maxRetries=10, sleepTime=1000 MILLISECONDS)
2024-08-01 21:36:36,321 INFO ipc.Client: Retrying connect to server: 0.0.0.0/0.0
```




```
hadoop@node-master:~$ yarn jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-exa
mples-3.3.4.jar wordcount "books/*" output
2024-08-02 03:39:17,198 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting
to ResourceManager at /0.0.0.0:8032
2024-08-02 03:39:18,239 INFO mapreduce.JobResourceUploader: Disabling Erasure Codin
g for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1722569804417_0001
2024-08-02 03:39:20,670 INFO input.FileInputFormat: Total input files to process :
1
2024-08-02 03:39:20,750 INFO mapreduce.JobSubmitter: number of splits:1
2024-08-02 03:39:21,004 INFO mapreduce.JobSubmitter: Submitting tokens for job: job
_1722569804417_0001
2024-08-02 03:39:21,005 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-08-02 03:39:21,200 INFO conf.Configuration: resource-types.xml not found
2024-08-02 03:39:21,201 INFO resource.ResourceUtils: Unable to find 'resource-types
.xml'.
2024-08-02 03:39:21,581 INFO impl.YarnClientImpl: Submitted application application
_1722569804417_0001
```

On ajoute

<property>

<name>dfs.datanode.data.dir</name>

<value>/home/hadoop/data/dataNode</value>

</property>

À

```
nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
```

```
hadoop@node-master:~$ jps
11683 NameNode
12148 SecondaryNameNode
11909 DataNode
12267 Jps
```

```
hadoop@node-master:~$ hdfs dfs -ls output
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2024-08-02 03:39 output/_SUCCESS
-rw-r--r--   1 hadoop supergroup          0 2024-08-02 03:39 output/part-r-00000
hadoop@node-master:~$ hdfs dfs -cat output/part-r-00000
hadoop@node-master:~$
```

```
hadoop@node-master:~$ wget https://archive.apache.org/dist/spark/spark-2.2.0/spark-
2.2.0-bin-hadoop2.7.tgz
--2024-08-02 03:50:48--  https://archive.apache.org/dist/spark/spark-2.2.0/spark-2.
2.0-bin-hadoop2.7.tgz
Résolution de archive.apache.org (archive.apache.org)… 65.108.204.189, 2a01:4f9:1a:
a084::2
Connexion à archive.apache.org (archive.apache.org)|65.108.204.189|:443… connecté.
requête HTTP transmise, en attente de la réponse… 200 OK
Taille : 203728858 (194M) [application/x-gzip]
Enregistre : 'spark-2.2.0-bin-hadoop2.7.tgz'

.2.0-bin-hadoop2.7.t   0%[                    ] 546,76K   129KB/s    tps 25m 38s
```

```
hadoop@node-master:~$ tar -xvf spark-2.2.0-bin-hadoop2.7.tgz
spark-2.2.0-bin-hadoop2.7/
spark-2.2.0-bin-hadoop2.7/NOTICE
spark-2.2.0-bin-hadoop2.7/jars/
spark-2.2.0-bin-hadoop2.7/jars/parquet-common-1.8.2.jar
spark-2.2.0-bin-hadoop2.7/jars/bonecp-0.8.0.RELEASE.jar
spark-2.2.0-bin-hadoop2.7/jars/commons-net-2.2.jar
```

```
hadoop@node-master:~$ mv spark-2.2.0-bin-hadoop2.7 spark
hadoop@node-master:~$
```

```
hadoop@node-master:~$ mv spark-2.2.0-bin-hadoop2.7 spark
hadoop@node-master:~$ vim /home/hadoop/.profile
hadoop@node-master:~$ vim /home/hadoop/.profile
hadoop@node-master:~$ source /home/hadoop/.profile
hadoop@node-master:~$ echo $HADOOP_CONF_DIR
/home/hadoop/hadoop/etc/hadoop
hadoop@node-master:~$ echo $SPARK_HOME
/home/hadoop/spark
hadoop@node-master:~$ echo $LD_LIBRARY_PATH
/home/hadoop/hadoop/lib/native:
hadoop@node-master:~$ mv $SPARK_HOME/conf/spark-defaults.conf.template $SPARK_HOME/
conf/spark-defaults.conf
hadoop@node-master:~$
```

```
PATH=/home/hadoop/spark/bin:$PATH
export HADOOP_CONF_DIR=/home/hadoop/hadoop/etc/hadoop
export SPARK_HOME=/home/hadoop/spark
export LD_LIBRARY_PATH=/home/hadoop/hadoop/lib/native:$LD_LIBRARY_PATH
"~/.profile" 32L, 1004B                                          32,1
```

```
spark.master yarn
"~/spark/conf/spark-defaults.conf" 29L, 1311B
```

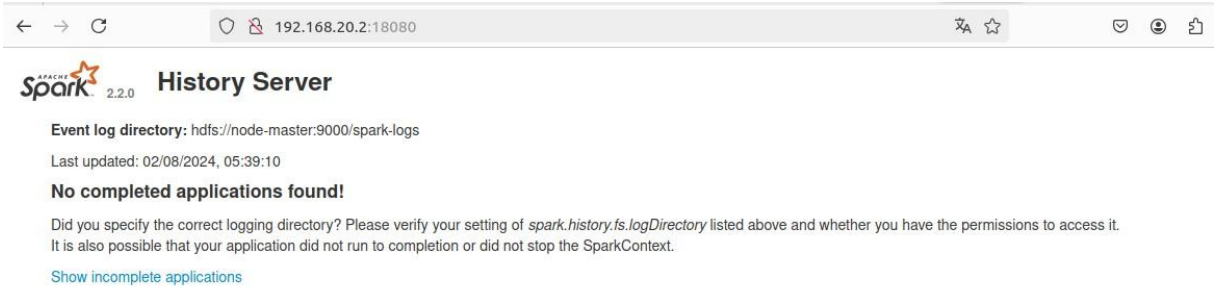
```
<configuration>
        <property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>1536</value>
</property>

</configuration>
~
~
"~/hadoop/etc/hadoop/yarn-site.xml" 21L, 738B
```

```
spark.master yarn
spark.driver.memory 512m
spark.yarn.am.memory 512m
spark.executor.memory 512m
spark.eventLog.enabled true
spark.eventLog.dir hdfs://node-master:9000/spark-logs
"~/spark/conf/spark-defaults.conf" 34L, 1471B
```

```
hadoop@node-master:~$ spark-submit --deploy-mode client --class org.apache.spark
.examples.SparkPi \$SPARK_HOME/examples/jars/spark-examples_2.11-2.2.0.jar 10
```

```
hadoop@node-master:/root$ hdfs dfs -mkdir /spark-logs
hadoop@node-master:/root$
```

```
hadoop@node-master:~$ $SPARK_HOME/sbin/start-history-server.sh
starting org.apache.spark.deploy.history.HistoryServer, logging to /home/hadoop
spark/logs/spark-hadoop-org.apache.spark.deploy.history.HistoryServer-1-node-ma
ter.out
hadoop@node-master:~$
```




```
ter.out
hadoop@node-master:~$ wget -O alice.txt https://www.gutenberg.org/fles/11/11-0.t
xt
--2024-08-02 05:41:08--  https://www.gutenberg.org/fles/11/11-0.txt
Résolution de www.gutenberg.org (www.gutenberg.org)… 152.19.134.47, 2610:28:3090
:3000:0:bad:cafe:47
Connexion à www.gutenberg.org (www.gutenberg.org)|152.19.134.47|:443… connecté.
```

```
hadoop@node-master:~$ hdfs dfs -mkdir inputs
hadoop@node-master:~$ hdfs dfs -put alice.txt inputs
hadoop@node-master:~$ spark-shell
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).

Failed to initialize compiler: object java.lang.Object in compiler mirror not fo
```

Part 2

Fode Mangane:

Application de Traitement de Données Distribuées : Implémentation de HDFS, MapReduce et Apache Spark sur Hadoop

Prérequis :

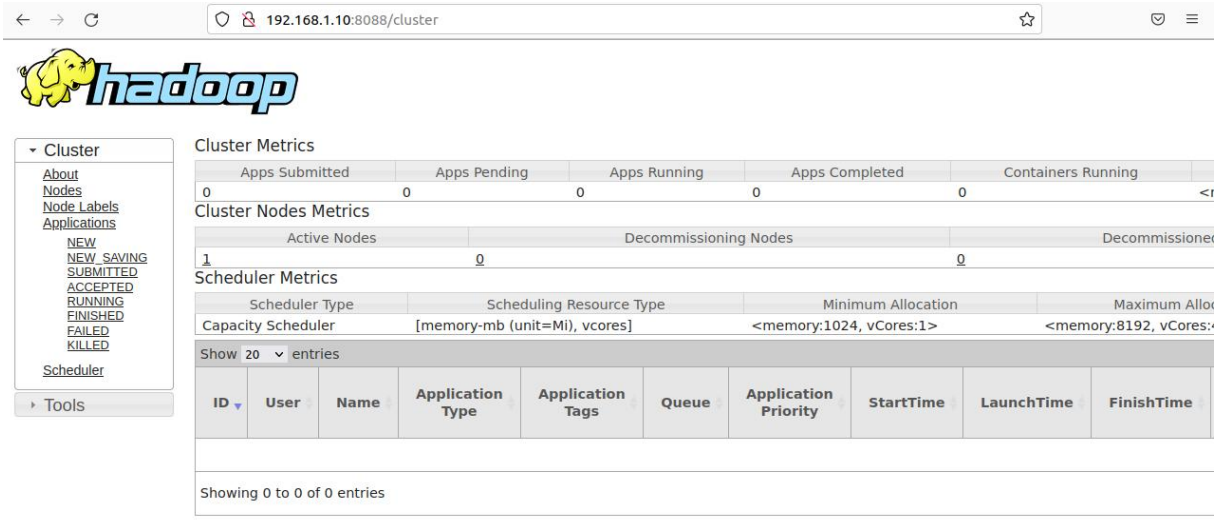
Pour commencer l'installation et la configuration, les composants suivants doivent être installés et fonctionnels :

- YARN
- DFS (Distributed File System)
- Spark

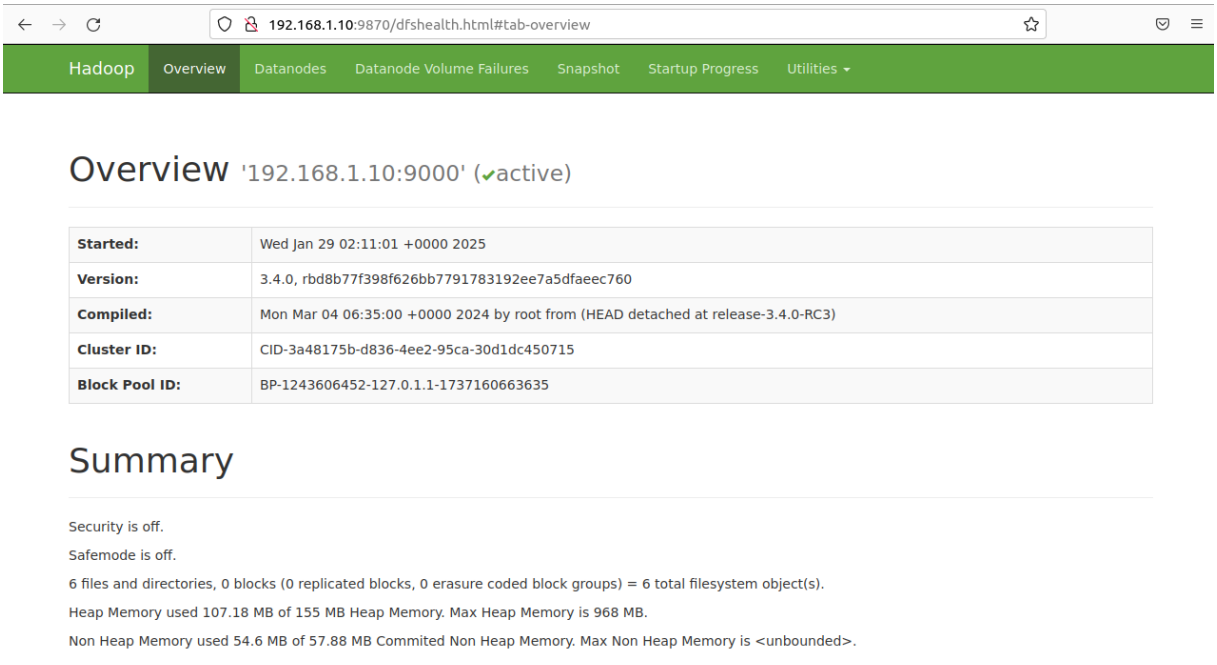
On démarre les services :

```
hadoop@Ubuntu:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@Ubuntu:~$ start-dfs.sh
Starting namenodes on [192.168.1.10]
Starting datanodes
Starting secondary namenodes [Ubuntu]
hadoop@Ubuntu:~$
```

Resource manager :



Namenode :



Manipulation du contenu HDFS

- **Avant de commencer, nous préférons que notre utilisateur du système (fodehadoop) appartienne à un groupe spécifique. Nous choisissons dans ce cas de créer un groupe « hadoop » pour la manipulation de HDFS.**

```
hadoop@Ubuntu:~$ sudo adduser fodehadoop
Adding user `fodehadoop' ...
Adding new group `fodehadoop' (1003) ...
Adding new user `fodehadoop' (1002) with group `fodehadoop' ...
Creating home directory `/home/fodehadoop' ...
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 fodehadoop
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]
hadoop@Ubuntu:~$
```

Appliquer immédiatement les modifications pour un utilisateur

```
hadoop@Ubuntu:~$ newgrp hadoopgroup
hadoop@Ubuntu:~$
```

- **Nous ajoutons notre utilisateur du système à ce groupe.**

```
hadoop@Ubuntu:~$ sudo adduser hadoop hadoopgroup
Adding user `hadoop' to group `hadoopgroup' ...
Adding user hadoop to group hadoopgroup
Done.
hadoop@Ubuntu:~$
```

- **Nous devons changer par la suite le propriétaire du dossier hadoop. L'utilisation de « chown » avec l'option « -R » (R pour récursive) permet de changer le propriétaire du dossier et aussi des dossiers et fichiers contenus à l'intérieur de ce dossier.**

```
hadoop@Ubuntu:~$ sudo chown -R hadoop:hadoopgroup /usr/local/hadoop
hadoop@Ubuntu:~$
```

```
hadoop@Ubuntu:~$ sudo addgroup hadoopgroup
[sudo] password for hadoop:
Adding group `hadoopgroup' (GID 1002) ...
Done.
hadoop@Ubuntu:~$
```

- **Concernant l'usage du chemin relatif dans le système de fichier HDFS, nous avons constaté que sur la version proposée par la fondation Apache, il était nécessaire d'initialiser nous-même le répertoire /user/[USER].**

- **Nous allons donc nous intéresser à créer un répertoire via -mkdir et donner un droit d'accès via -chown à un utilisateur du système. On va donc créer le répertoire /user/[USER] et lui donner les droits pour l'utilisateur nom_user et pour le groupe hadoop.**

Créer des répertoires et attribuer des droits :

```
hadoop@Ubuntu:~$ newgrp hadoopgroup
hadoop@Ubuntu:~$ hadoop fs -mkdir /user
hadoop@Ubuntu:~$ hadoop fs -mkdir /user/hadoop
hadoop@Ubuntu:~$ hadoop fs -chown hadoop:hadoopgroup /user/hadoop
hadoop@Ubuntu:~$
```

- **Copie de fichiers**

```
hadoop@Ubuntu:~/Documents$ touch test.txt
hadoop@Ubuntu:~/Documents$ ls
test.txt
hadoop@Ubuntu:~/Documents$ cd
hadoop@Ubuntu:~$
hadoop@Ubuntu:~$
hadoop@Ubuntu:~$ hadoop fs -put ~/Documents/test.txt /user/hadoop/
hadoop@Ubuntu:~$
hadoop@Ubuntu:~$ hadoop fs -ls /user/hadoop/
Found 1 items
-rw-r--r--   1 hadoop hadoopgroup          0 2025-01-18 13:20 /user/hadoop/test.
txt
hadoop@Ubuntu:~$
```

- **Suppression de fichiers**

```
hadoop@Ubuntu:~$ hadoop fs -mkdir /user/hadoop/mon_repertoire
hadoop@Ubuntu:~$ hadoop fs -mv /user/hadoop/test.txt /user/hadoop/mon_repertoire
/
hadoop@Ubuntu:~$
```

```
hadoop@Ubuntu:~$ hadoop fs -rm /user/hadoop/mon_repertoire/test.txt
Deleted /user/hadoop/mon_repertoire/test.txt
hadoop@Ubuntu:~$ hadoop fs -rm -r /user/hadoop/mon_repertoire
Deleted /user/hadoop/mon_repertoire
hadoop@Ubuntu:~$
```

Dans le fichier test.txt, j'ai écrit : "Welcome to Fode test"

```
hadoop@Ubuntu:~$ hadoop fs -mkdir /user/hadoop/input
hadoop@Ubuntu:~$ hadoop fs -put ~/Documents/test.txt /user/hadoop/input/
```

```
hadoop@Ubuntu:~$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce
-examples-*.jar wordcount /user/hadoop/input /user/hadoop/output
2025-01-18 13:41:06,356 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /0.0.0.0:8032
2025-01-18 13:41:06,727 INFO mapreduce.JobResourceUploader: Disabling Erasure Co
ding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1737197843567_0001
2025-01-18 13:41:07,424 INFO input.FileInputFormat: Total input files to process
 : 1
2025-01-18 13:41:08,316 INFO mapreduce.JobSubmitter: number of splits:1
2025-01-18 13:41:08,938 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1737197843567_0001
2025-01-18 13:41:08,940 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-01-18 13:41:09,086 INFO conf.Configuration: resource-types.xml not found
2025-01-18 13:41:09,086 INFO resource.ResourceUtils: Unable to find 'resource-ty
pes.xml'.
2025-01-18 13:41:09,547 INFO impl.YarnClientImpl: Submitted application applicat
```

Et voilà les résultats du test

```
hadoop@Ubuntu:~$ hadoop fs -cat /user/hadoop/output/part-r-00000
Welcome 1
fode    1
test    1
to      1
hadoop@Ubuntu:~$
```

Programmation Hadoop – Implémentation des classes Driver, MAP et REDUCE

J'ai créé les fichiers WCount.java, WCountMap.java, WCountReduce.java, map.py et reduce.py, contenant le code nécessaire au fonctionnement de mon application.

```
hadoop@Ubuntu:~$ vim /home/hadoop/WCount.java
hadoop@Ubuntu:~$ vim /home/hadoop/WCountMap.java
hadoop@Ubuntu:~$ vim /home/hadoop/WCountReduce.java
hadoop@Ubuntu:~$ vim /home/hadoop/map.py
hadoop@Ubuntu:~$ vim /home/hadoop/reduce.py
hadoop@Ubuntu:~$ javac -classpath `hadoop classpath` -d /home/hadoop/classes WCo
unt.java WCountMap.java WCountReduce.java
hadoop@Ubuntu:~$
```

```
hadoop@Ubuntu:~$ ls
classes                 Music                          src
Desktop                 output                         Templates
Documents               Pictures                       Videos
Downloads               Public                         WCount.jar
hadoop-3.4.0.tar.gz     reduce.py                      WCount.java
hdfs                    snap                           WCountMap.java
input.txt               spark-3.2.1-bin-hadoop3.2.tgz  WCountReduce.java
map.py                  spark-wordcount
hadoop@Ubuntu:~$
```

Compilation

```
hadoop@Ubuntu:~$ javac -classpath `hadoop classpath` -d /home/hadoop/classes WCo
unt.java WCountMap.java WCountReduce.java
hadoop@Ubuntu:~$
```

J'ai généré un fichier JAR pour mon application

```
hadoop@Ubuntu:~$ jar -cvf /home/hadoop/WCount.jar -C /home/hadoop/classes/ .
added manifest
adding: hadoop/(in = 0) (out= 0)(stored 0%)
adding: hadoop/wordcount/(in = 0) (out= 0)(stored 0%)
adding: hadoop/wordcount/WCount.class(in = 1770) (out= 945)(deflated 46%)
adding: hadoop/wordcount/WCountMap.class(in = 1693) (out= 740)(deflated 56%)
adding: hadoop/wordcount/WCountReduce.class(in = 1608) (out= 673)(deflated 58%)
hadoop@Ubuntu:~$
```

Lancement et exécution du programme Hadoop

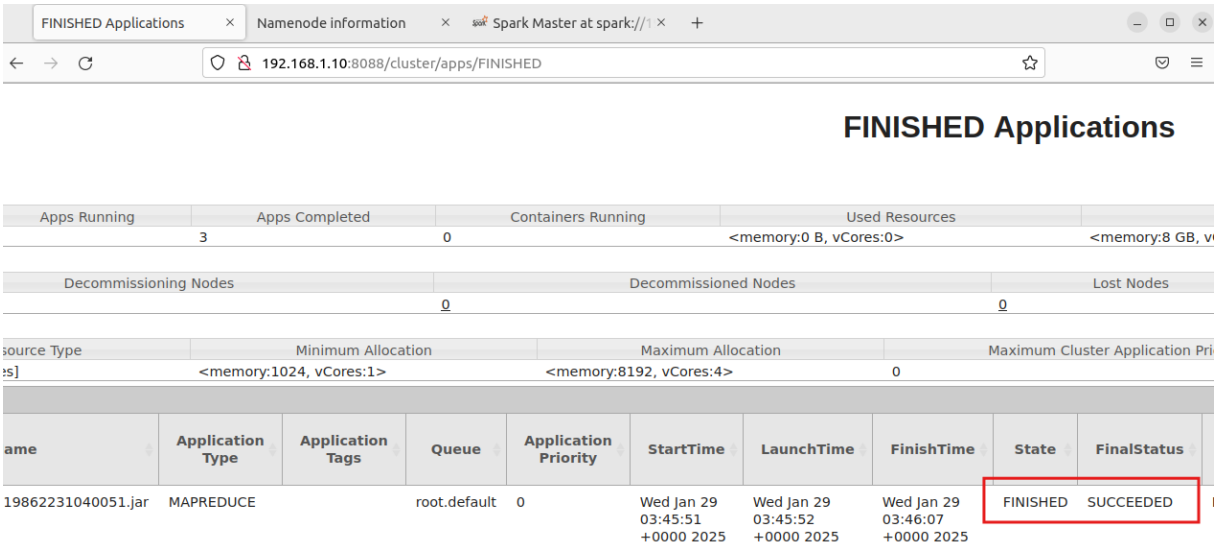
```
hadoop@Ubuntu:~$ hadoop jar /home/hadoop/WCount.jar hadoop.wordcount.WCount /use
r/hadoop/input /user/hadoop/output
2025-01-18 14:12:39,996 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /0.0.0.0:8032
Exception in thread "main" org.apache.hadoop.mapred.FileAlreadyExistsException:
Output directory hdfs://192.168.1.10:9000/user/hadoop/output already exists
        at org.apache.hadoop.mapreduce.lib.output.FileOutputFormat.checkOutputSp
ecs(FileOutputFormat.java:164)
```

Exécution du programme avec Hadoop Streaming

```
hadoop@Ubuntu:~$ hadoop jar $HADOOP_HOME/share/hadoop/tools/lib/hadoop-streaming
-*.jar \
    -input /user/hadoop/input \
    -output /user/hadoop/output \
    -mapper /home/hadoop/map.py \
    -reducer /home/hadoop/reduce.py
packageJobJar: [/tmp/hadoop-unjar4075629422349118048/] [] /tmp/streamjob12318802
968651101781.jar tmpDir=null
2025-01-18 14:36:36,023 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /0.0.0.0:8032
2025-01-18 14:36:36,125 INFO client.DefaultNoHARMFailoverProxyProvider: Connecti
ng to ResourceManager at /0.0.0.0:8032
2025-01-18 14:36:36,312 INFO mapreduce.JobResourceUploader: Disabling Erasure Co
ding for path: /tmp/hadoop-yarn/staging/hadoop/.staging/job_1737197843567_0003
2025-01-18 14:36:36,533 INFO mapred.FileInputFormat: Total input files to proces
s : 1
2025-01-18 14:36:37,003 INFO mapreduce.JobSubmitter: number of splits:2
2025-01-18 14:36:37,143 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1737197843567_0003
```

Verfication :

L'application fonctionne correctement et est accessible via l'interface graphique à l'adresse 192.168.1.10:8088.

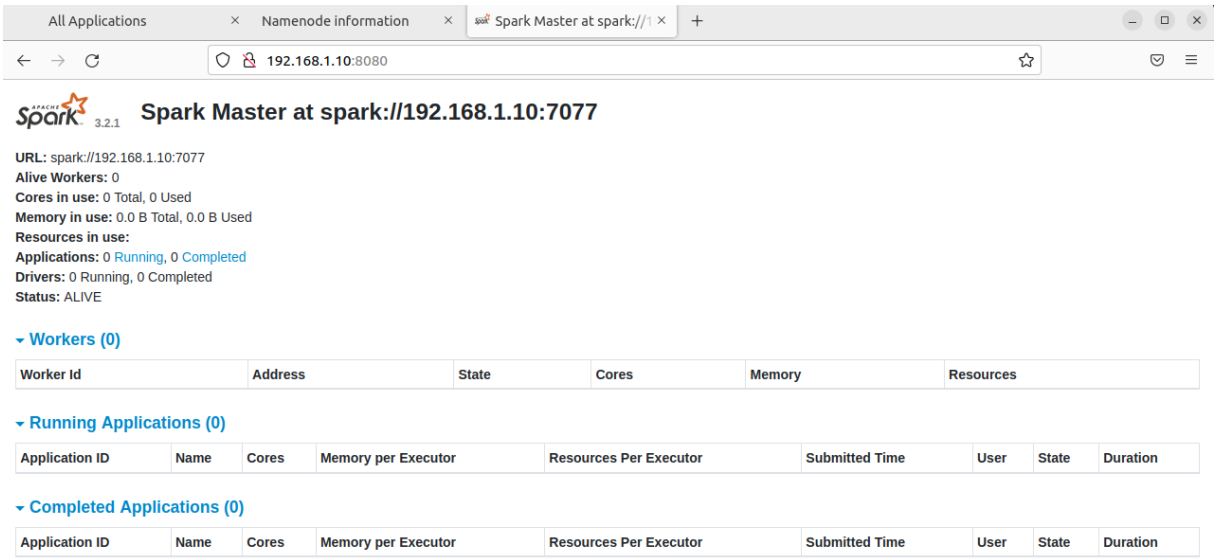


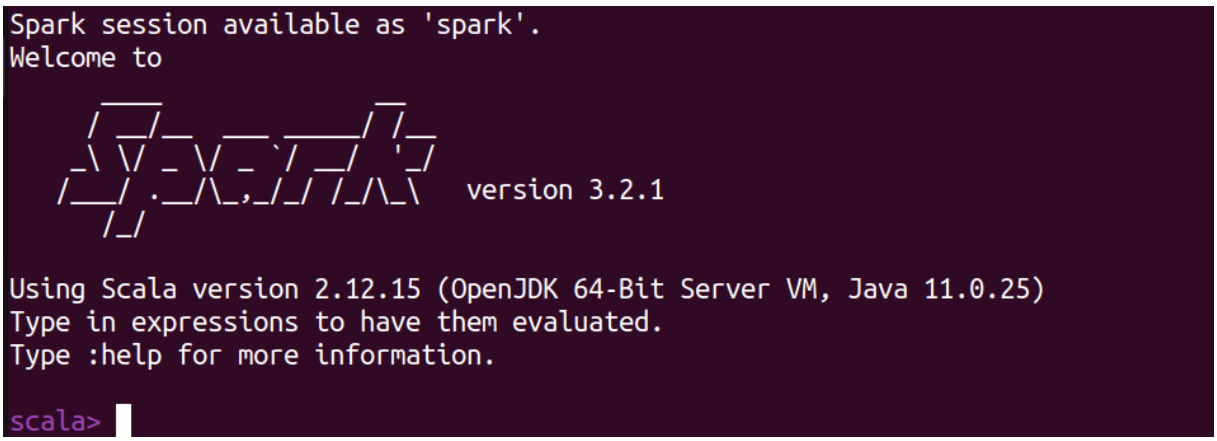
SPARK

Démarrage de spark :

```
hadoop@Ubuntu:~$ $SPARK_HOME/sbin/start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark/logs/spark
-hadoop-org.apache.spark.deploy.master.Master-1-Ubuntu.out
hadoop@Ubuntu:~$
```

Interface Spark :





La commande **mvn clean package** permet de nettoyer le projet en supprimant les fichiers de compilations précédentes et de générer un fichier JAR contenant l'application compilée.

```
hadoop@Ubuntu:~/spark-wordcount$ mvn clean package
[INFO] Scanning for projects...
[INFO]
[INFO] -------------------< com.example:spark-wordcount >--------------------
[INFO] Building spark-wordcount 1.0-SNAPSHOT
[INFO] --------------------------------[ jar ]---------------------------------
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/
plugins/maven-clean-plugin/2.5/maven-clean-plugin-2.5.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/apache/maven/p
lugins/maven-clean-plugin/2.5/maven-clean-plugin-2.5.pom (3.9 kB at 3.0 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/
plugins/maven-plugins/22/maven-plugins-22.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/apache/maven/p
lugins/maven-plugins/22/maven-plugins-22.pom (13 kB at 67 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/
maven-parent/21/maven-parent-21.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/apache/maven/m
aven-parent/21/maven-parent-21.pom (26 kB at 38 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/apache
/10/apache-10.pom
```

```
s/plexus-archiver/2.1/plexus-archiver-2.1.jar (184 kB at 244 kB/s)
Downloaded from central: https://repo.maven.apache.org/maven2/commons-lang/commo
ns-lang/2.1/commons-lang-2.1.jar (208 kB at 252 kB/s)
[INFO] Building jar: /home/hadoop/spark-wordcount/target/spark-wordcount-1.0-SNA
PSHOT.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  03:10 min
[INFO] Finished at: 2025-01-20T00:43:42Z
[INFO] ------------------------------------------------------------------------
hadoop@Ubuntu:~/spark-wordcount$
```

TEST:

```
hadoop@Ubuntu:~/spark-wordcount$ echo "Bonjour Spark Bonjour Hadoop" > input.txt

hadoop@Ubuntu:~/spark-wordcount$ spark-submit --class com.example.WordCount \
    --master local \
    target/spark-wordcount-1.0-SNAPSHOT.jar \
    /path/to/input.txt /path/to/output
25/01/20 00:46:51 WARN Utils: Your hostname, Ubuntu resolves to a loopback addre
ss: 127.0.1.1; using 192.168.148.143 instead (on interface ens33)
25/01/20 00:46:51 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another
address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/us
r/local/spark/jars/spark-unsafe_2.12-3.2.1.jar) to constructor java.nio.DirectBy
teBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.u
nsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflect
```

VERIFICATION avec SPARK:

```
hadoop@Ubuntu:~/spark-wordcount$ ls ~/output
part-00000  _SUCCESS
hadoop@Ubuntu:~/spark-wordcount$ cat ~/output/part-00000
(Spark,1)
(Bonjour,2)
(Hadoop,1)
hadoop@Ubuntu:~/spark-wordcount$
```

Mise en place d'un projet Maven pour le développement de l'application Spark

```
hadoop@Ubuntu:~$ cd spark-wordcount/
hadoop@Ubuntu:~/spark-wordcount$ ls
pom.xml  src  target
hadoop@Ubuntu:~/spark-wordcount$ vim pom.xml
```

```
hadoop@Ubuntu:~/spark-wordcount$ cd src/main/java/com/example/
hadoop@Ubuntu:~/spark-wordcount/src/main/java/com/example$ ls
WordCount.java
hadoop@Ubuntu:~/spark-wordcount/src/main/java/com/example$ vim WordCount.java
```

```
hadoop@Ubuntu:~/spark-wordcount$ mkdir -p src/main/java/com/example
hadoop@Ubuntu:~/spark-wordcount$ cd src/main/java/com/example
hadoop@Ubuntu:~/spark-wordcount/src/main/java/com/example$ ls
hadoop@Ubuntu:~/spark-wordcount/src/main/java/com/example$ vim WordCount.java
hadoop@Ubuntu:~/spark-wordcount/src/main/java/com/example$
```

Nous plaçons notre code de configuration dans ces deux fichiers, puis nous procédons à la compilation.

```
hadoop@Ubuntu:~/spark-wordcount$ mvn clean package
[INFO] Scanning for projects...
[INFO]
[INFO] -------------------< com.example:spark-wordcount >--------------------
[INFO] Building spark-wordcount 1.0-SNAPSHOT
[INFO] --------------------------------[ jar ]---------------------------------
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/
plugins/maven-clean-plugin/2.5/maven-clean-plugin-2.5.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/apache/maven/p
lugins/maven-clean-plugin/2.5/maven-clean-plugin-2.5.pom (3.9 kB at 5.5 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/
plugins/maven-plugins/22/maven-plugins-22.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/apache/maven/p
lugins/maven-plugins/22/maven-plugins-22.pom (13 kB at 88 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/maven/
maven-parent/21/maven-parent-21.pom
Downloaded from central: https://repo.maven.apache.org/maven2/org/apache/maven/m
aven-parent/21/maven-parent-21.pom (26 kB at 160 kB/s)
Downloading from central: https://repo.maven.apache.org/maven2/org/apache/apache
```

Fin

```
Downloaded from central: https://repo.maven.apache.org/maven2/commons-lang/commo
ns-lang/2.1/commons-lang-2.1.jar (208 kB at 384 kB/s)
[INFO] Building jar: /home/hadoop/spark-wordcount/target/spark-wordcount-1.0-SNA
PSHOT.jar
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  02:00 min
[INFO] Finished at: 2025-01-29T04:02:39Z
[INFO] ------------------------------------------------------------------------
hadoop@Ubuntu:~/spark-wordcount$
```

Exécuter l'application Spark :

Préparation d'un fichier d'entrée :

Crée un fichier d'entrée pour tester l'application. Par exemple :

```
hadoop@Ubuntu:~/spark-wordcount$ echo "Bonjour Spark Bonjour Hadoop" > input.txt

hadoop@Ubuntu:~/spark-wordcount$
```

<u>Exécuter</u> l'application avec Spark : Utilise spark-submit pour lancer le programme :

```
hadoop@Ubuntu:~/spark-wordcount$ spark-submit --class com.example.WordCount \
    --master local \
    target/spark-wordcount-1.0-SNAPSHOT.jar \
    /home/hadoop/spark-wordcount/input.txt /home/hadoop/spark-wordcount/output
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/op
t/spark/jars/spark-unsafe_2.12-3.2.1.jar) to constructor java.nio.DirectByteBuff
er(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.u
nsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflect
ive access operations
WARNING: All illegal access operations will be denied in a future release
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
25/01/29 04:20:58 INFO SparkContext: Running Spark version 3.2.1
25/01/29 04:20:58 WARN NativeCodeLoader: Unable to load native-hadoop library fo
r your platform... using builtin-java classes where applicable
25/01/29 04:20:58 INFO ResourceUtils: ==============================================
====================
25/01/29 04:20:58 INFO ResourceUtils: No custom resources configured for spark.d
```

VERIFICATION :

```
hadoop@Ubuntu:~/spark-wordcount$ cat /home/hadoop/spark-wordcount/output/part-00
000
(Spark,1)
(Bonjour,2)
(Hadoop,1)
hadoop@Ubuntu:~/spark-wordcount$
```