Installation Hadoop

Sous WSL 2

http://hadoop.apache.org/docs/current/

Installation de la distribution Ubuntu

L’installation est effectuée à partir de Microsoft Store

Ubuntu 20.04 LTS

Canonical Group Limited • Outils de développement > Utilitaires

Partager

Ubuntu 20.04 LTS on Windows allows you to use Ubuntu Terminal and run Ubuntu command line utilities including bash, ssh, git, apt and many

Lors du premier lancement de la distribution, il vous sera demandé de créer le compte administrateur

```
Installing, this may take a few minutes...
Please create a default UNIX user account. The username does not need to match your Windows username.
For more information visit: https://aka.ms/wslusers
Enter new UNIX username: hadoop
```

La distribution est souvent livrée avec un catalogue de packages minimal ou pas à jour, alors il est recommandé mettre à jour à jour une distribution avant son utilisation:

```
hadoop@DESKTOP-M2J7LR1:~$ sudo apt update && sudo apt upgrade
[sudo] password for hadoop:
Get:1 http://security.ubuntu.com/ubuntu focal-security InRelease [109 kB]
Hit:2 http://archive.ubuntu.com/ubuntu focal InRelease
```

Installation Java

- sudo apt install default-jre
- Sudo apt install default-jdk

```
hadoop@DESKTOP-M2J7LR1:/mnt/c/Users/nadir$ java -version
openjdk version "11.0.10" 2021-01-19
OpenJDK Runtime Environment (build 11.0.10+9-Ubuntu-0ubuntu1.20.04)
OpenJDK 64-Bit Server VM (build 11.0.10+9-Ubuntu-0ubuntu1.20.04, mixed mode, sharing)
hadoop@DESKTOP-M2J7LR1:/mnt/c/Users/nadir$ javac -version
javac 11.0.10
```

← Pour vérifier les versions installé

Obtenir le chemin d'installation de java (Nous allons définir la variable d'environnement JAVA_HOME ultérieurement)

```
hadoop@DESKTOP-M2J7LR1:/mnt/c/Users/nadir$ sudo update-alternatives --config java
There is only one alternative in link group java (providing /usr/bin/java): /usr/lib/jvm/java-11-openjdk-amd64/bin/java
Nothing to configure.
```

Installation et configuration ssh et pdsh

- sudo apt install ssh
- sudo apt install pdsh
- Génération de la paire des clés RSA

```
hadoop@DESKTOP-M2J7LR1:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:NM3AV0jpxQ0F00975BotnLcPHZLxGLBHhkt8ewmuaKw hadoop@DESKTOP-M2J7LR1
The key's randomart image is:
+---[RSA 3072]----+
|       ...o*BO.  |
|        .+++++o o|
|        oo+o+oOBo|
|       . ....*=**|
|        S . . o*+|
|         + .  o..|
|        o      ..|
|       E        .|
|                 |
+----[SHA256]-----+
```

Ajout de la clé publique à la liste des clés autorisées :
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

```
hadoop@DESKTOP-M2J7LR1:~$ ls -l -a .ssh
total 20
drwx------ 2 hadoop hadoop 4096 Mar 13 22:21 .
drwxr-xr-x 4 hadoop hadoop 4096 Mar 13 22:10 ..
-rw-r--r-- 1 hadoop hadoop  576 Mar 13 22:21 authorized_keys
-rw------- 1 hadoop hadoop 2610 Mar 13 22:10 id_rsa
-rw-r--r-- 1 hadoop hadoop  576 Mar 13 22:10 id_rsa.pub
```

Redémarrer ssh

```
hadoop@DESKTOP-M2J7LR1:~$ sudo service ssh restart
```

```
hadoop@DESKTOP-M2J7LR1:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:+3y6rEl2R3LdBU7kAo2Q8pIfpeA40l0h2RMrZh6/PsI.
Are you sure you want to continue connecting (yes/no/[fingerprint])? y
Please type 'yes', 'no' or the fingerprint: yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 20.04.2 LTS (GNU/Linux 5.4.72-microsoft-standard-WSL2 x86_64)

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

```
hadoop@DESKTOP-M2J7LR1:~$ ls -l -a .ssh
total 24
drwx------ 2 hadoop hadoop 4096 Mar 13 22:26 .
drwxr-xr-x 5 hadoop hadoop 4096 Mar 13 22:26 ..
-rw-r--r-- 1 hadoop hadoop  576 Mar 13 22:21 authorized_keys
-rw------- 1 hadoop hadoop 2610 Mar 13 22:10 id_rsa
-rw-r--r-- 1 hadoop hadoop  576 Mar 13 22:10 id_rsa.pub
-rw-r--r-- 1 hadoop hadoop  222 Mar 13 22:26 known_hosts
```

fichiers dans ↓ ssh

```
hadoop@DESKTOP-M2J7LR1:~$ ssh localhost
Welcome to Ubuntu 20.04.2 LTS (GNU/Linux 5.4.72-microsoft-standard-WSL2 x86_64)

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

  System information as of Sat Mar 13 22:31:24 +01 2021

  System load:  0.0               Processes:             15
  Usage of /:   0.7% of 250.98GB  Users logged in:       0
  Memory usage: 1%                IPv4 address for eth0: 172.26.15.230
  Swap usage:   0%


0 updates can be installed immediately.
0 of these updates are security updates.
```

↑ connexion réussie.

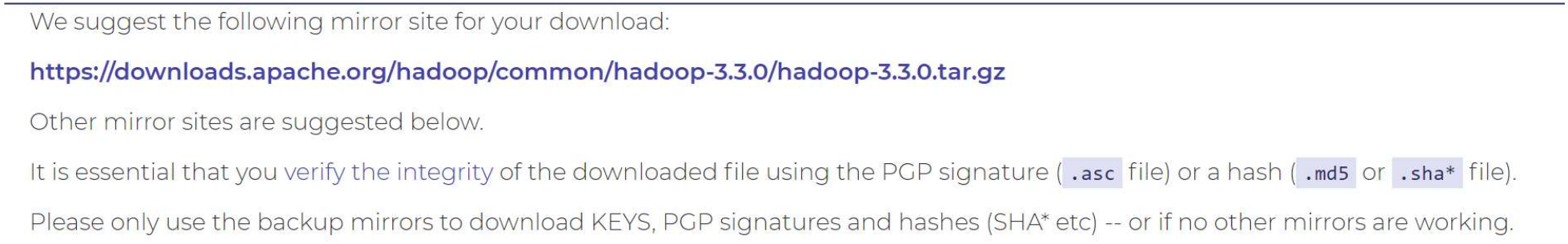
Téléchargement

```
hadoop@DESKTOP-M2J7LR1:~$ sudo mkdir bigdata
[sudo] password for hadoop:
hadoop@DESKTOP-M2J7LR1:~$ cd bigdata
```

Créer un dossier pour installer hadoop.

Télécharger hadoop:

sudo wget https://downloads.apache.org/hadoop/common/hadoop-3.3.0/hadoop-3.3.0.tar.gz



Décompresser l'archive:

sudo tar -xvzf hadoop-3.3.0.tar.gz

```
hadoop@DESKTOP-M2J7LR1:~/bigdata$ ls
hadoop-3.3.0  hadoop-3.3.0.tar.gz
```

Si la decompression a été effectuée sans erreurs, l'archive peut être supprimée:

sudo rm hadoop-3.3.0.tar.gz

Configurer les variables d'environnement

Ajouter les variables d'environnement suivantes dans le fichier ~/.bashrc et après executer la commande: source ~/.bashrc

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=~/bigdata/hadoop-3.3.0
export PATH=$JAVA_HOME/bin:$HADOOP_HOME/bin:$HADOOP_HOME/sbin:$PATH
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/hadoop
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
```

Configuration Hadoop

Les fichiers de configuration Hadoop sont localisés dans le dossier HADOOP_HOME/etc/hadoop

Nous allons modifier les fichiers de configuration suivants:

- hadoop-env.sh
- core-site.xml
- hdfs-site.xml
- mapred-site.xml
- yarn-site.xml

Ajouter les variables suivantes dans le fichier hadoop-env.sh et après exécuter la commande :

source $HADOOP_HOME/etc/hadoop/hadoop-env.sh

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HDFS_NAMENODE_USER="hadoop"
export HDFS_DATANODE_USER="hadoop"
export HDFS_SECONDARYNAMENODE_USER="hadoop"
export YARN_RESOURCEMANAGER_USER="hadoop"
export YARN_NODEMANAGER_USER="hadoop"
```

Configuration hadoop core-site.xml

```
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
```

Configuration HDFS

créer un répertoire pour stocker toutes les données du nœud maître (namenode) et un autre pour stocker les données (datanode). Dans exemple:

```
hadoop@DESKTOP-M2J7LR1:~/bigdata/atelier0$ sudo mkdir $HADOOP_HOME/data
hadoop@DESKTOP-M2J7LR1:~/bigdata/atelier0$ sudo mkdir $HADOOP_HOME/data/namenode
hadoop@DESKTOP-M2J7LR1:~/bigdata/atelier0$ sudo mkdir $HADOOP_HOME/data/datanode
```

hdfs-site.xml

```
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>~/bigdata/hadoop-3.3.0/data/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>~/bigdata/hadoop-3.3.0/data/datanode</value>
  </property>
</configuration>
```

Par défaut replication = 3

Configuration MapReduce: mapred-site.xml

```
<configuration>
  <property>
      <name>mapreduce.framework.name</name>
      <value>yarn</value>
  </property>
  <property>
      <name>mapreduce.application.classpath</name>
      <value>~/bigdata/Hadoop-
3.3.0/share/hadoop/mapreduce/*: ~/bigdata/Hadoop-3.3.0
/share/hadoop/mapreduce/lib/*</value>
  </property>
</configuration>
```

Yarn: yarn-site.xml

```
<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_DI
R,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
  </property>
</configuration>
```

Formatage HDFS

Exécuter la commande: hdfs namenode -format

```
2021-03-01 08:03:55,046 INFO snapshot.SnapshotManager: SkipList is disabled
2021-03-01 08:03:55,061 INFO util.GSet: Computing capacity for map cachedBlocks
2021-03-01 08:03:55,061 INFO util.GSet: VM type       = 64-bit
2021-03-01 08:03:55,061 INFO util.GSet: 0.25% max memory 966.7 MB = 2.4 MB
2021-03-01 08:03:55,061 INFO util.GSet: capacity      = 2^18 = 262144 entries
2021-03-01 08:03:55,092 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.window.num.buckets = 10
2021-03-01 08:03:55,092 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.num.users = 10
2021-03-01 08:03:55,092 INFO metrics.TopMetrics: NNTop conf: dfs.namenode.top.windows.minutes = 1,5,25
2021-03-01 08:03:55,108 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2021-03-01 08:03:55,139 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry
 time is 600000 millis
2021-03-01 08:03:55,139 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2021-03-01 08:03:55,139 INFO util.GSet: VM type       = 64-bit
2021-03-01 08:03:55,139 INFO util.GSet: 0.029999999329447746% max memory 966.7 MB = 297.0 KB
2021-03-01 08:03:55,139 INFO util.GSet: capacity      = 2^15 = 32768 entries
2021-03-01 08:03:55,233 INFO namenode.FSImage: Allocated new BlockPoolId: BP-891857813-172.18.99.250-1614614635217
2021-03-01 08:03:55,358 INFO common.Storage: Storage directory C:\Tools\hadoop-3.3.0\data\dfs\namenode1 has been success
fully formatted.
2021-03-01 08:03:55,452 INFO namenode.FSImageFormatProtobuf: Saving image file C:\Tools\hadoop-3.3.0\data\dfs\namenode1\
current\fsimage.ckpt_0000000000000000000 using no compression
2021-03-01 08:03:55,671 INFO namenode.FSImageFormatProtobuf: Image file C:\Tools\hadoop-3.3.0\data\dfs\namenode1\current
\fsimage.ckpt_0000000000000000000 of size 396 bytes saved in 0 seconds .
2021-03-01 08:03:55,702 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >= 0
2021-03-01 08:03:55,749 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2021-03-01 08:03:55,749 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at WinDev2101Eval/172.18.99.250
************************************************************/

C:\bigdata>
```

Démarrage du système HDFS

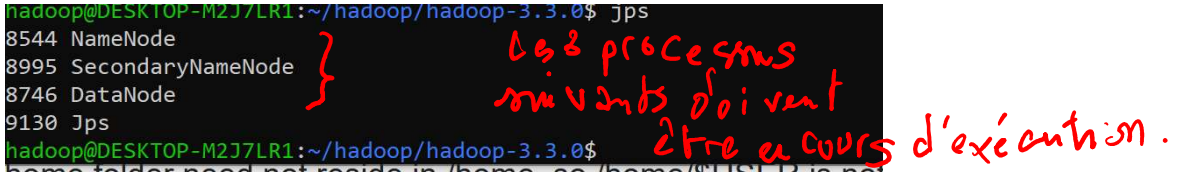
Démarrer le système de fichiers HDFS, en exécutant le script: $HADOOP_HOME/sbin/start-dfs.sh

Si vous rencontrez l'erreur .

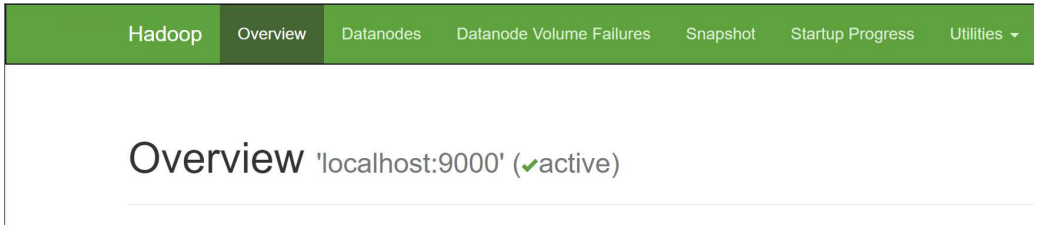
Starting namenodes on [localhost]

pdsh@DESKTOP-M2J7LR1: localhost: rcmd: socket:
Permission denied

Ajoutez dans .bashrc : `export PDSH_RCMD_TYPE=ssh` et exécuter `source ~/.bashrc`



Interface web par défaut du namenode:
http://localhost:9870/



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 40.41 MB of 62.45 MB Heap Memory. Max Heap Memory is 966.69 MB.

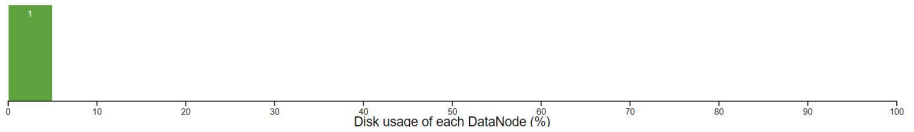
Non Heap Memory used 60.4 MB of 61.55 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	126.46 GB
Configured Remote Capacity:	0 B
DFS Used:	321 B (0%)
Non DFS Used:	62.96 GB
DFS Remaining:	63.5 GB (50.21%)
Block Pool Used:	321 B (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)

Datanode Information

In service · Down · Decommissioning · Decommissioned · Decommissioned & dead
Entering Maintenance · In Maintenance · In Maintenance & dead

Datanode usage histogram



Commandes HDFS

- Démarrer le terminal
- Commandes HDFS: hadoop fs
  - Pour copier un fichier dans HDFS: -copyFromLocal fichier.txt
  - Pour afficher le contenu: –ls
  - Créer une copie d’un fichier: -cp f1 f2
  - Copier un fichier à partir de hdfs dans le système local: -copyToLocal f
  - Supprimer un fichier: -rm f

```
hadoop@DESKTOP-M2J7LR1:~$ start-yarn.sh
```

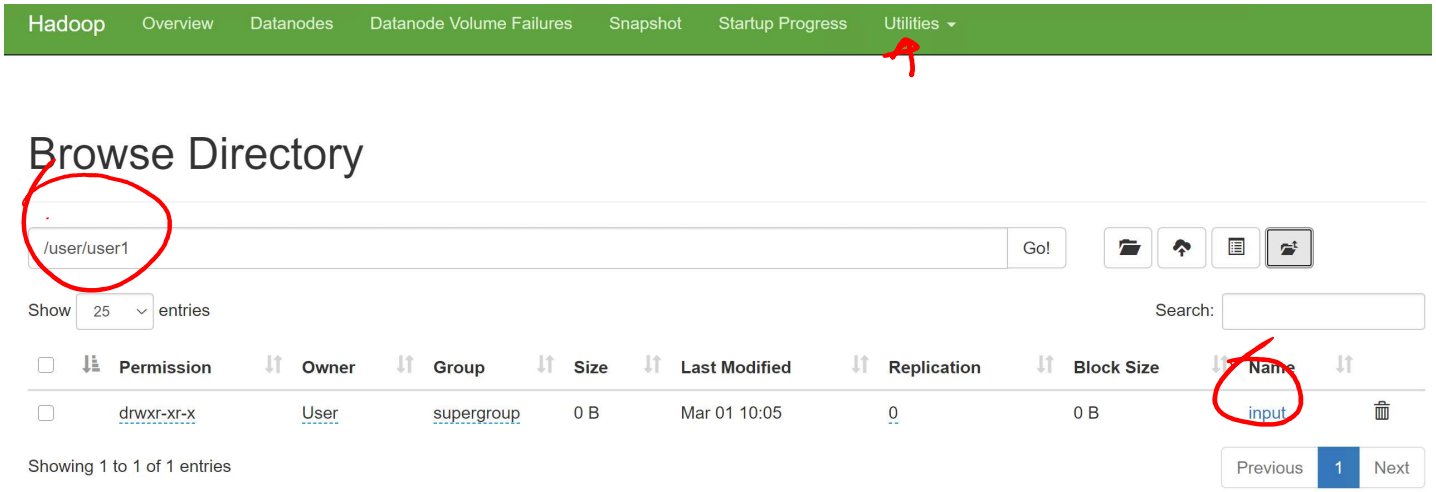
```
hadoop@DESKTOP-M2J7LR1:~$ jps
12690 Jps
12038 NodeManager
11431 DataNode
11880 ResourceManager
11658 SecondaryNameNode
11246 NameNode
```

Démarrer yarn: Resource Manager

Atelier

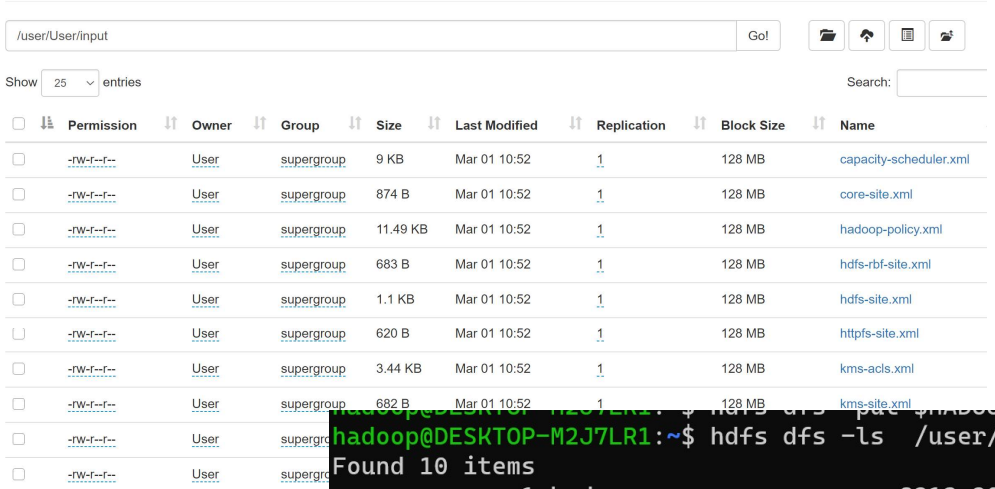
- Créer l'arborescence des dossiers suivante: "user/hadoop/input" dans hdfs:

```
C:\bigdata>hdfs dfs -mkdir /user

>hdfs dfs -mkdir /user/User

>hdfs dfs -mkdir /user/User/input
```



Copier les fichiers xml du dossier etc/hadoop dans le dossier hdfs input

```
~$ hdfs dfs -put $HADOOP_HOME/etc/hadoop/*.xml /user/hadoop/input
```



/user/User/input Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	User	supergroup	9 KB	Mar 01 10:52	1	128 MB	capacity-scheduler.xml
-rw-r--r--	User	supergroup	874 B	Mar 01 10:52	1	128 MB	core-site.xml
-rw-r--r--	User	supergroup	11.49 KB	Mar 01 10:52	1	128 MB	hadoop-policy.xml
-rw-r--r--	User	supergroup	683 B	Mar 01 10:52	1	128 MB	hdfs-rbf-site.xml
-rw-r--r--	User	supergroup	1.1 KB	Mar 01 10:52	1	128 MB	hdfs-site.xml
-rw-r--r--	User	supergroup	620 B	Mar 01 10:52	1	128 MB	httpfs-site.xml
-rw-r--r--	User	supergroup	3.44 KB	Mar 01 10:52	1	128 MB	kms-acls.xml
-rw-r--r--	User	supergroup	682 B	Mar 01 10:52	1	128 MB	kms-site.xml
-rw-r--r--	User	supergroup					
-rw-r--r--	User	supergroup					

```
hadoop@DESKTOP-M2J7LR1:~$ hdfs dfs -ls  /user/hadoop/input
Found 10 items
-rw-r--r--   1 hadoop supergroup       9213 2021-03-14 01:48 /user/hadoop/input/capacity-scheduler.xml
-rw-r--r--   1 hadoop supergroup        884 2021-03-14 01:48 /user/hadoop/input/core-site.xml
-rw-r--r--   1 hadoop supergroup      11765 2021-03-14 01:48 /user/hadoop/input/hadoop-policy.xml
-rw-r--r--   1 hadoop supergroup        683 2021-03-14 01:48 /user/hadoop/input/hdfs-rbf-site.xml
-rw-r--r--   1 hadoop supergroup       1116 2021-03-14 01:48 /user/hadoop/input/hdfs-site.xml
-rw-r--r--   1 hadoop supergroup        620 2021-03-14 01:48 /user/hadoop/input/httpfs-site.xml
-rw-r--r--   1 hadoop supergroup       3518 2021-03-14 01:48 /user/hadoop/input/kms-acls.xml
-rw-r--r--   1 hadoop supergroup        682 2021-03-14 01:48 /user/hadoop/input/kms-site.xml
-rw-r--r--   1 hadoop supergroup        758 2021-03-14 01:48 /user/hadoop/input/mapred-site.xml
-rw-r--r--   1 hadoop supergroup        989 2021-03-14 01:48 /user/hadoop/input/yarn-site.xml
```

Exécuter un script mapreduce

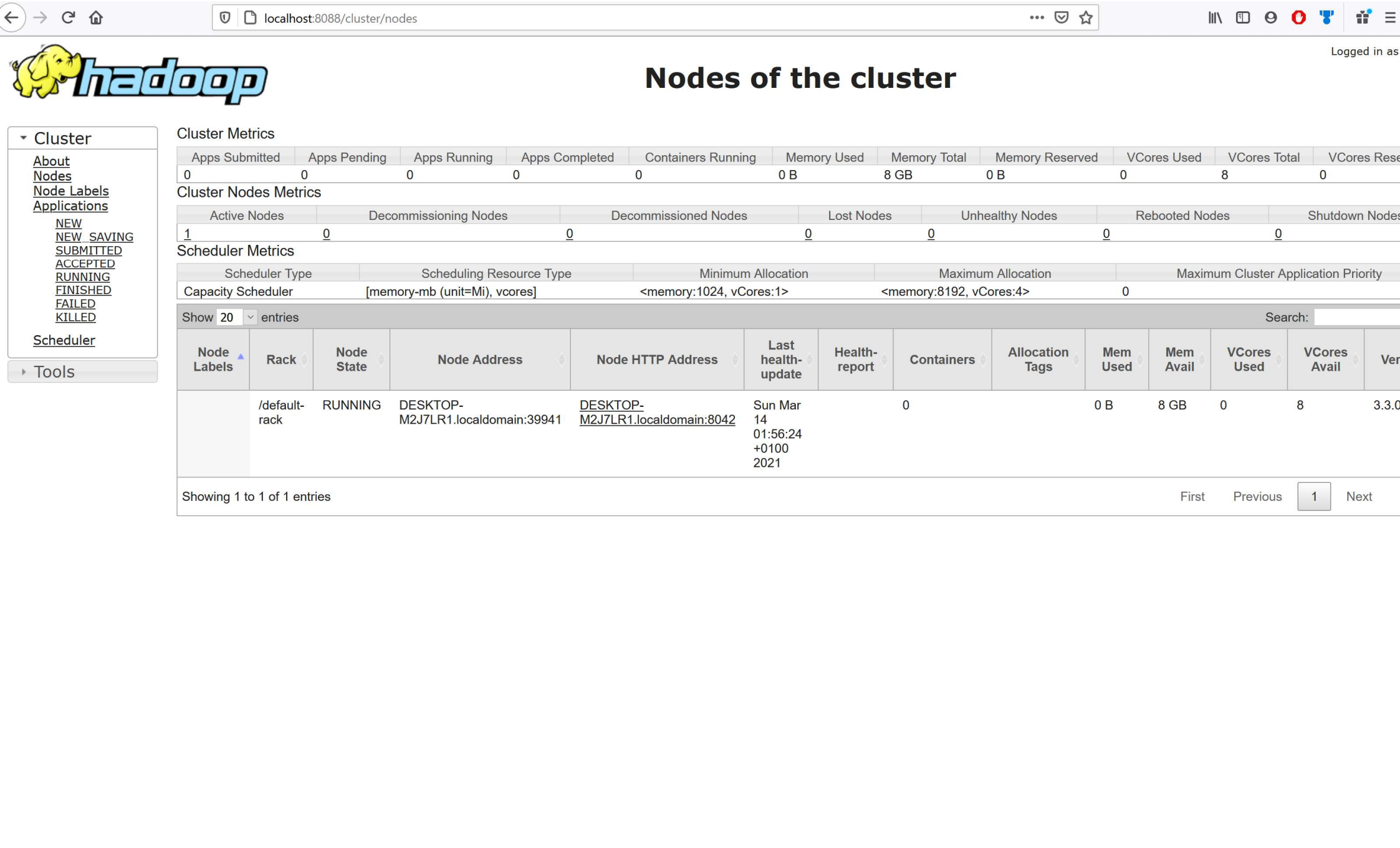
```
$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar grep input output 'dfs[a-z.]+'
```

```
hadoop@DESKTOP-M2J7LR1:~$ hdfs dfs -ls /user/hadoop
Found 2 items
drwxr-xr-x   - hadoop supergroup          0 2021-03-14 01:48 /user/hadoop/input
drwxr-xr-x   - hadoop supergroup          0 2021-03-14 01:54 /user/hadoop/output
hadoop@DESKTOP-M2J7LR1:~$ hdfs dfs -ls /user/hadoop/output
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2021-03-14 01:54 /user/hadoop/output/_SUCCESS
-rw-r--r--   1 hadoop supergroup         77 2021-03-14 01:54 /user/hadoop/output/part-r-00000
hadoop@DESKTOP-M2J7LR1:~$
```

L'exécution a réussi

Résultat

```
hadoop@DESKTOP-M2J7LR1:~$ hdfs dfs -text /user/hadoop/output/part-r-00000
1       dfsadmin
1       dfs.replication
1       dfs.namenode.name.dir
1       dfs.datanode.data.dir
```


localhost:8088/cluster/nodes
Logged in as
hadoop
Nodes of the cluster
Cluster
About
Nodes
Node Labels
Applications
NEW
NEW_SAVING
SUBMITTED
ACCEPTED
RUNNING
FINISHED
FAILED
KILLED
Scheduler
Tools
Cluster Metrics
Apps Submitted
Apps Pending
Apps Running
Apps Completed
Containers Running
Memory Used
Memory Total
Memory Reserved
VCores Used
VCores Total
VCores Rese
0
0
0
0
0
0 B
8 GB
0 B
0
8
0
Cluster Nodes Metrics
Active Nodes
Decommissioning Nodes
Decommissioned Nodes
Lost Nodes
Unhealthy Nodes
Rebooted Nodes
Shutdown Nodes
1
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:
Node Labels
Rack
Node State
Node Address
Node HTTP Address
Last health-update
Health-report
Containers
Allocation Tags
Mem Used
Mem Avail
VCores Used
VCores Avail
Ver
/default-rack
RUNNING
DESKTOP-M2J7LR1.localdomain:39941
DESKTOP-M2J7LR1.localdomain:8042
Sun Mar 14 01:56:24 +0100 2021
0
0 B
8 GB
0
8
3.3.0
Showing 1 to 1 of 1 entries
First
Previous
1
Next