

Rapport de TP - Configuration de Hadoop en Mode Multi-nœuds

Réalisé par : LAABID ABDESSAMAD

Github : https://github.com/aplusInDev/hadoop_tps/tree/main/tp3

1. Introduction

Ce travail pratique vise à configurer un cluster Hadoop en mode multi-nœuds, comprenant un nœud maître et deux nœuds esclaves. Cette configuration permettra d'exploiter pleinement les capacités de traitement distribué de Hadoop pour des applications Big Data.

Le cluster mis en place inclut les composants essentiels de l'écosystème Hadoop :

- HDFS (Hadoop Distributed File System) pour le stockage distribué
- YARN (Yet Another Resource Negotiator) pour la gestion des ressources
- MapReduce pour le traitement distribué des données

2. Objectifs

- Configurer un cluster Hadoop fonctionnel avec une architecture maître-esclave
- Mettre en place HDFS en mode distribué
- Configurer YARN pour la gestion des ressources du cluster
- Exécuter des jobs MapReduce sur le cluster
- Comprendre les mécanismes de réplication et de distribution des données

3. Environnement de travail

Architecture du cluster :

- 1 nœud maître : **master**
- 2 nœuds esclaves : **slave1, slave2**

Configuration matérielle :

- Machines virtuelles avec système d'exploitation Linux
- 4 Go de RAM par nœud
- 2 core pour chaque nœud
- Disque dur avec espace suffisant pour le stockage HDFS

Version des logiciels :

- Hadoop : 3.4.0
- Java : JDK 11

4. Méthodologie

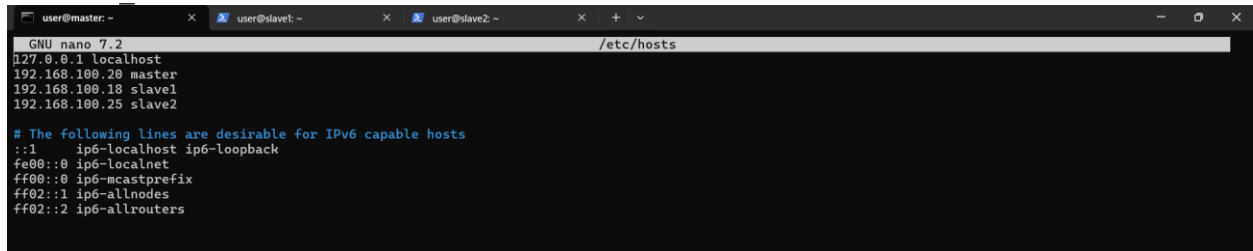
4.1 Préparation de l'environnement

Configuration des noms d'hôtes

La première étape consiste à configurer les noms d'hôtes sur tous les nœuds pour faciliter la communication au sein du cluster.

Fichier `/etc/hosts` (à ajouter sur tous les nœuds) :

```
<MASTER_IP> master
<SLAVE1_IP> slave1
<SLAVE2_IP> slave2
```



```
GNU nano 7.2 /etc/hosts
127.0.0.1 localhost
192.168.100.20 master
192.168.100.18 slave1
192.168.100.25 slave2

# 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
```

Configuration SSH sans mot de passe

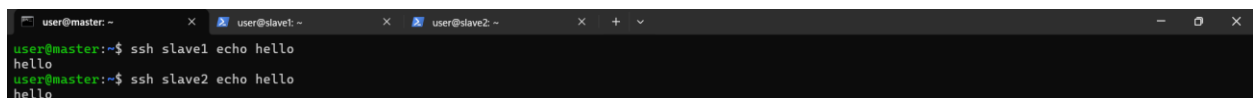
Pour permettre au nœud maître de communiquer avec les nœuds esclaves sans intervention manuelle, nous avons configuré SSH pour une authentification par clé :

Sur le nœud maître :

```
# Génération de la clé SSH
ssh-keygen -t rsa -P "" -f ~/.ssh/id_rsa
```

```
# Copie de la clé vers tous les nœuds
ssh-copy-id user@master
ssh-copy-id user@slave1
ssh-copy-id user@slave2
```

Vérification de la connexion SSH sans mot de passe :



```
user@master:~$ ssh slave1 echo hello
hello
user@master:~$ ssh slave2 echo hello
hello
```

4.2 Installation de Hadoop

L'installation de Hadoop a été réalisée sur tous les nœuds du cluster :

```
# Téléchargement et extraction de l'archive Hadoop
wget https://downloads.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0.tar.gz
tar -xzf hadoop-3.4.0.tar.gz -C /home/user/
```

Configuration des variables d'environnement dans le fichier `.bashrc` sur tous les nœuds :

```
export HADOOP_HOME=/home/user/hadoop-3.4.0
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

4.3 Configuration du cluster Hadoop

Fichier workers (sur le nœud maître)

Création du fichier `$HADOOP_HOME/etc/hadoop/workers` pour spécifier les nœuds esclaves :

```
slave1
slave2
```

Configuration de core-site.xml (tous les nœuds)

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/home/user/tmpdata</value>
  </property>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://master:9000</value>
  </property>
</configuration>
```

Configuration de hdfs-site.xml (nœud maître)

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>/home/user/dfsdata/namenode</value>
  </property>
  <property>
```

```

    <name>dfs.datanode.data.dir</name>
    <value>/home/user/dfsdata/datanode</value>
  </property>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.http-address</name>
    <value>master:9870</value>
  </property>
  <property>
    <name>dfs.permissions</name>
    <value>false</value>
  </property>
</configuration>

```

Configuration de hdfs-site.xml (nœud esclave)

```

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>/home/user/dfsdata/datanode</value>
  </property>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.permissions</name>
    <value>false</value>
  </property>
</configuration>

```

Configuration de mapred-site.xml (tous les nœuds)

```

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>mapreduce.framework.name</name>
    <value>yarn</value>
  </property>
  <property>
    <name>mapreduce.jobhistory.address</name>
    <value>master:10020</value>
  </property>
  <property>
    <name>mapreduce.jobhistory.webapp.address</name>
    <value>master:19888</value>
  </property>
  <property>
    <name>yarn.app.mapreduce.am.resource.mb</name>

```

```

    <value>1536</value>
  </property>
</property>
  <name>mapreduce.map.memory.mb</name>
  <value>512</value>
</property>
</property>
  <name>mapreduce.reduce.memory.mb</name>
  <value>1024</value>
</property>
</configuration>

```

Configuration de yarn-site.xml (nœud maître)

```

<?xml version="1.0"?>
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>master</value>
  </property>
  <property>
    <name>yarn.resourcemanager.webapp.address</name>
    <value>master:8088</value>
  </property>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>
    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
  </property>
  <property>
    <name>yarn.resourcemanager.resource-tracker.address</name>
    <value>master:8025</value>
  </property>
  <property>
    <name>yarn.resourcemanager.scheduler.address</name>
    <value>master:8030</value>
  </property>
  <property>
    <name>yarn.resourcemanager.address</name>
    <value>master:8050</value>
  </property>
</configuration>

```

Configuration de yarn-site.xml (nœuds esclaves)

```
<?xml version="1.0"?>
<configuration>
  <property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
  </property>
  <property>
    <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
    <value>org.apache.hadoop.mapred.ShuffleHandler</value>
  </property>
  <property>
    <name>yarn.resourcemanager.hostname</name>
    <value>hadoop-master</value>
  </property>
  <property>
    <name>yarn.acl.enable</name>
    <value>0</value>
  </property>
  <property>
    <name>yarn.nodemanager.env-whitelist</name>
    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
  </property>
  <property>
    <name>yarn.nodemanager.resource.memory-mb</name>
    <value>4096</value>
  </property>
  <property>
    <name>yarn.scheduler.maximum-allocation-mb</name>
    <value>4096</value>
  </property>
  <property>
    <name>yarn.scheduler.minimum-allocation-mb</name>
    <value>512</value>
  </property>
</configuration>
```

4.4 Démarrage du cluster

Une fois la configuration terminée, nous avons formaté le NameNode :

```
# Formatage du NameNode (uniquement sur le nœud maître)
hdfs namenode -format
```

Ensuite, nous avons démarré les services HDFS et YARN :

```
# Démarrage de HDFS
start-dfs.sh

# Démarrage de YARN
start-yarn.sh
```

Vérification des processus en cours d'exécution :

Sur le nœud maître

```
user@master:~$ jps
2822 SecondaryNameNode
8359 Jps
6045 ResourceManager
2598 NameNode
4991 NodeManager
```

Sur les nœuds esclaves

```
user@slave1:~$ jps
3462 NodeManager
3306 DataNode
3597 Jps
```

5. Tests et validation

5.1 Vérification du système de fichiers HDFS

Nous avons effectué des tests pour vérifier le bon fonctionnement du système de fichiers HDFS :

```
user@master:~$ # Création d'un répertoire dans HDFS
user@master:~$ hdfs dfs -mkdir -p /tp3/test
user@master:~$ echo "hello world!" > test_hdfs.txt
user@master:~$ hdfs dfs -put test_hdfs.txt /tp3/test
user@master:~$ # Vérification du fichier et de sa réplcation
user@master:~$ hdfs dfs -ls /tp3/test
Found 1 items
-rw-r--r--  2 user supergroup        13 2025-05-06 15:55 /tp3/test/test_hdfs.txt
user@master:~$ hdfs fsck /tp3/test/test_hdfs.txt -files -blocks -locations
Connecting to namenode via http://master:9870/fsck?ugi=user&files=1&blocks=1&locations=1&path=%2Ftp3%2Ftest%2Ftest_hdfs.txt
FSCK started by user (auth:SIMPLE) from /192.168.100.20 for path /tp3/test/test_hdfs.txt at Tue May 06 15:56:46 UTC 2025

/tp3/test/test_hdfs.txt 13 bytes, replicated: replication=2, 1 block(s): OK
0. BP-226074085-192.168.100.20-1746539216608:blk_1073741875_1051 len=13 Live_repl=2 [DatanodeInfoWithStorage[192.168.100.18:9866,DS-99e8d161-b1b1-44bc-9561-f70449b89492,DISK], DatanodeInfoWithStorage[192.168.100.25:9866,DS-48d2cbcd-72c6-49a6-b139-11e7517bd6f9,DISK]]

Status: HEALTHY
Number of data-nodes:  2
Number of racks:       1
Total dirs:            0
Total symlinks:        0

Replicated Blocks:
Total size:           13 B
Total files:          1
Total blocks (validated): 1 (avg. block size 13 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 2
Average block replication: 2.0
Missing blocks:        0
Corrupt blocks:         0
Missing replicas:       0 (0.0 %)
Blocks queued for replication: 0

Erasure Coded Block Groups:
Total size:           0 B
Total files:          0
Total block groups (validated): 0
Minimally erasure-coded block groups: 0
Over-erasure-coded block groups: 0
Under-erasure-coded block groups: 0
Unsatisfactory placement block groups: 0
Average block group size: 0.0
Missing block groups:  0
Corrupt block groups:  0
Missing internal blocks: 0
Blocks queued for replication: 0
FSCK ended at Tue May 06 15:56:46 UTC 2025 in 15 milliseconds
```

5.2 Test d'un job MapReduce

Exécution des jobs de TP précédents

Nous avons également exécuté les jobs MapReduce développés lors des TP précédents pour analyser les données météorologiques :

Job 1 : Température maximale par année

```
user@master: ~$ bash hadoop-tps/tp2/apply_1.sh
packageJobJar: [/tmp/hadoop-unjar13916758689632666012/] [] /tmp/streamjob17538157509715134994.jar tmpDir=null
2025-05-06 16:10:28,415 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/192.168.100.20:8050
2025-05-06 16:10:28,616 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/192.168.100.20:8050
2025-05-06 16:10:29,466 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/user/.staging/job_1746546722942_0001
2025-05-06 16:10:31,223 INFO mapred.FileInputFormat: Total input files to process : 1
2025-05-06 16:10:31,461 INFO mapreduce.JobSubmitter: number of splits:2
2025-05-06 16:10:31,991 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746546722942_0001
2025-05-06 16:10:31,991 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 16:10:32,325 INFO conf.Configuration: resource-types.xml not found
2025-05-06 16:10:32,326 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-05-06 16:10:33,039 INFO impl.YarnClientImpl: Submitted application application_1746546722942_0001
2025-05-06 16:10:33,126 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1746546722942_0001/
2025-05-06 16:10:33,129 INFO mapreduce.Job: Running job: job_1746546722942_0001
2025-05-06 16:10:43,783 INFO mapreduce.Job: Job job_1746546722942_0001 running in uber mode : false
2025-05-06 16:10:43,787 INFO mapreduce.Job: map 0% reduce 0%
2025-05-06 16:10:56,911 INFO mapreduce.Job: map 50% reduce 0%
2025-05-06 16:10:58,650 INFO mapreduce.Job: map 100% reduce 0%
2025-05-06 16:11:05,985 INFO mapreduce.Job: map 100% reduce 100%
2025-05-06 16:11:06,006 INFO mapreduce.Job: Job job_1746546722942_0001 completed successfully
2025-05-06 16:11:06,157 INFO mapreduce.Job: Counters: 54

File System Counters
  FILE: Number of bytes read=1337
  FILE: Number of bytes written=938640
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=4720
  HDFS: Number of bytes written=634
  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
  Rack-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=19849
  Total time spent by all reduces in occupied slots (ms)=5355
  Total time spent by all map tasks (ms)=19849
  Total time spent by all reduce tasks (ms)=5355
  Total vcore-milliseconds taken by all map tasks=19849
  Total vcore-milliseconds taken by all reduce tasks=5355
  Total megabyte-milliseconds taken by all map tasks=20325376
  Total megabyte-milliseconds taken by all reduce tasks=5483520

Map-Reduce Framework
  Map input records=100
  Map output records=100
  Map output bytes=1311
  Map output materialized bytes=1343
  Input split bytes=206
  Combine input records=0
  Combine output records=0
  Reduce input groups=99
  Reduce shuffle bytes=1343
  Reduce input records=100
  Reduce output records=62
  Spilled Records=200
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=254
  CPU time spent (ms)=3880
  Physical memory (bytes) snapshot=801404800
  Virtual memory (bytes) snapshot=7262711808
  Total committed heap usage (bytes)=618659040
  Peak Map Physical memory (bytes)=298082304
  Peak Map Virtual memory (bytes)=2271117312
  Peak Reduce Physical memory (bytes)=237084672
  Peak Reduce Virtual memory (bytes)=2722119680

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=4514
File Output Format Counters
  Bytes Written=634

2025-05-06 16:11:06,158 INFO streaming.StreamJob: Output directory: /output_1
Results: -----
1900 -2.0
1903 -41.0
1907 -42.0
1910 -33.0
1914 -44.0
```

Job 2 : Nombre de mois avec température > seuil (0)


```
user@master: ~$ bash hadoop_tps/tp2/apply_2.sh 0
hadoop_tps/tp2/apply_2.sh: line 1: [: command not found
packageJobJar: [/tmp/hadoop-unjar4888876739400836898/] [] /tmp/streamjob964613826103748894.jar tmpDir=null
2025-05-06 16:47:51,945 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/192.168.100.20:8050
2025-05-06 16:47:52,235 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at master/192.168.100.20:8050
2025-05-06 16:47:52,773 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/user/.staging/job_1746546722942_0003
2025-05-06 16:47:54,195 INFO mapred.FileInputFormat: Total input files to process : 1
2025-05-06 16:47:54,637 INFO mapreduce.JobSubmitter: number of splits:2
2025-05-06 16:47:55,187 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1746546722942_0003
2025-05-06 16:47:55,188 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 16:47:55,525 INFO conf.Configuration: resource-types.xml not found
2025-05-06 16:47:55,530 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2025-05-06 16:47:55,703 INFO impl.YarnClientImpl: Submitted application application_1746546722942_0003
2025-05-06 16:47:55,763 INFO mapreduce.Job: The url to track the job: http://master:8088/proxy/application_1746546722942_0003/
2025-05-06 16:47:55,767 INFO mapreduce.Job: Running job: job_1746546722942_0003
2025-05-06 16:48:05,615 INFO mapreduce.Job: Job job_1746546722942_0003 running in uber mode : false
2025-05-06 16:48:05,623 INFO mapreduce.Job: map 0% reduce 0%
2025-05-06 16:48:17,688 INFO mapreduce.Job: map 50% reduce 0%
2025-05-06 16:48:18,721 INFO mapreduce.Job: map 100% reduce 0%
2025-05-06 16:48:26,176 INFO mapreduce.Job: map 100% reduce 100%
2025-05-06 16:48:27,294 INFO mapreduce.Job: Job job_1746546722942_0003 completed successfully
2025-05-06 16:48:27,471 INFO mapreduce.Job: Counters: 54
  File System Counters
    FILE: Number of bytes read=482
    FILE: Number of bytes written=936930
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=4720
    HDFS: Number of bytes written=55
    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
    Rack-local map tasks=2
    Total time spent by all maps in occupied slots (ms)=17281
    Total time spent by all reduces in occupied slots (ms)=5497
    Total time spent by all map tasks (ms)=17281
    Total time spent by all reduce tasks (ms)=5497
    Total vcore-milliseonds taken by all map tasks=17281
    Total vcore-milliseonds taken by all reduce tasks=5497
    Total megabyte-milliseonds taken by all map tasks=17695744
    Total megabyte-milliseonds taken by all reduce tasks=5628928
  Map-Reduce Framework
    Map input records=100
    Map output records=48
    Map output bytes=380
    Map output materialized bytes=488
    Input split bytes=206
    Combine input records=0
    Combine output records=0
    Reduce input groups=46
    Reduce shuffle bytes=488
    Reduce input records=48
    Reduce output records=1
    Spilled Records=96
    Shuffled Maps =2
    Failed Shuffles=0
    Merged Map outputs=2
    GC time elapsed (ms)=276
    CPU time spent (ms)=3800
    Physical memory (bytes) snapshot=766980096
    Virtual memory (bytes) snapshot=7273168896
    Total committed heap usage (bytes)=620756992
    Peak Map Physical memory (bytes)=274190336
    Peak Map Virtual memory (bytes)=2276450304
    Peak Reduce Physical memory (bytes)=224935936
    Peak Reduce Virtual memory (bytes)=2721779712
  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=4514
  File Output Format Counters
    Bytes Written=55
2025-05-06 16:48:27,472 INFO streaming.StreamJob: Output directory: /output_2
Results for Months with Temperature > 0: -----
Number of months with temperature above threshold: 12
Deleted /output_2
Deleted /data/tp2
```

6. Problèmes rencontrés et solutions

6.1 Configuration des services YARN

Problème : Les services NodeManager ne démarraient pas correctement sur les nœuds esclaves, ce qui empêchait l'exécution des jobs MapReduce. Cette erreur était visible dans l'interface web de YARN où les applications restaient bloquées dans l'état "ACCEPTED".

Solution :

1. Vérification des processus en cours d'exécution avec la commande `jps`
2. Redémarrage manuel des services NodeManager sur les nœuds esclaves avec la commande :

7. Interfaces Web

Le cluster Hadoop dispose de plusieurs interfaces web pour la supervision et l'administration :

- **HDFS NameNode** : `http://<master-ip_address>:9870`
 - Vue d'ensemble du système de fichiers
 - Statut des DataNodes
 - Utilisation du stockage

Overview 'master:9000' (✔active)

Started:	Tue May 06 14:47:15 +0100 2025
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaec760
Compiled:	Mon Mar 04 07:35:00 +0100 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-3f19a66e-7a13-45d4-9c87-696efbdfd9e5
Block Pool ID:	BP-226074005-192.168.100.20-1746539216608

Summary

Security is off.
Safemode is off.
37 files and directories, 20 blocks (20 replicated blocks, 0 erasure coded block groups) = 57 total filesystem object(s).
Heap Memory used 56.86 MB of 187 MB Heap Memory. Max Heap Memory is 980 MB.
Non Heap Memory used 69.68 MB of 72.31 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	39.03 GB
Configured Remote Capacity:	0 B
DFS Used:	3.36 MB (0.01%)
Non DFS Used:	19.03 GB
DFS Remaining:	17.97 GB (46.04%)
Block Pool Used:	3.36 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.01% / 0.01% / 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	12
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Tue May 06 14:47:15 +0100 2025
Last Checkpoint Time	Tue May 06 14:46:56 +0100 2025
Last HA Transition Time	Never
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 282	
Journal Manager	State
FileJournalManager(root=/home/user/dfsdata/namenode)	EditLogFileOutputStream(/home/user/dfsdata/namenode/current/edits_inprogress_000000000000000282)

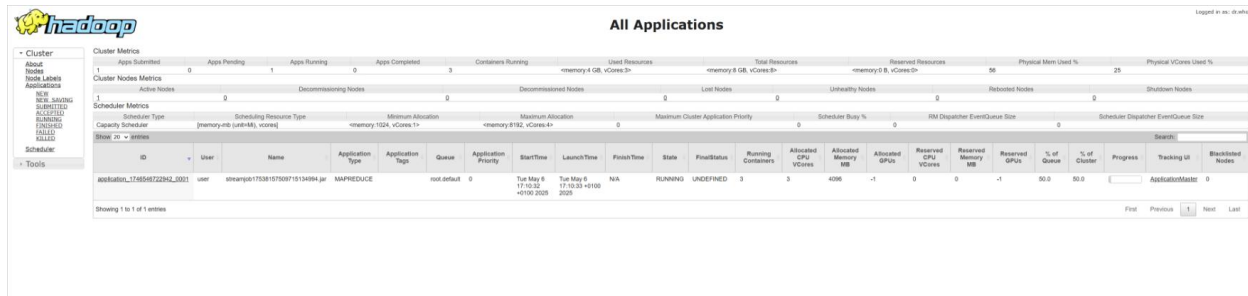
NameNode Storage

Storage Directory	Type	State
/home/user/dfsdata/namenode	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	39.03 GB	3.36 MB (0.01%)	17.97 GB (46.04%)	3.36 MB	2

- **YARN ResourceManager** : http://<master-ip_address>:8088
 - Suivi des applications
 - Utilisation des ressources du cluster
 - Statut des NodeManagers



The screenshot shows the Hadoop YARN ResourceManager web interface. The top section displays 'All Applications' with a summary of cluster metrics. Below this, there are several tabs for different views: Cluster, Applications, Nodes, and Scheduler. The 'Applications' tab is selected, showing a table of running applications. The table has columns for ID, User, Name, Application Type, Application Tags, Queue, Application Priority, Start Time, Launch Time, Final Time, State, Final Status, Running Containers, Allocated CPU V-Cores, Allocated Memory MB, Allocated GPUs, Reserved CPU V-Cores, Reserved Memory MB, Reserved GPUs, % of Queue, % of Cluster, Progress, Tracking UI, and Blacklisted Nodes. A single application is listed with ID 'application_1718548722642_0001', User 'user', Name 'streamjob1718548722642_0001', Application Type 'MAPREDUCE', Application Tags 'root:default', Queue 'root:default', Application Priority '0', Start Time 'Tue May 6 17:16:32 +0100 2020', Launch Time 'Tue May 6 17:16:33 +0100 2020', Final Time 'NA', State 'RUNNING', Final Status 'UNDEFINED', Running Containers '3', Allocated CPU V-Cores '3', Allocated Memory MB '4096', Allocated GPUs '-1', Reserved CPU V-Cores '0', Reserved Memory MB '0', Reserved GPUs '-1', % of Queue '50.0', % of Cluster '50.0', Progress '0', Tracking UI 'ApplicationMaster', and Blacklisted Nodes '0'.

8. Conclusion

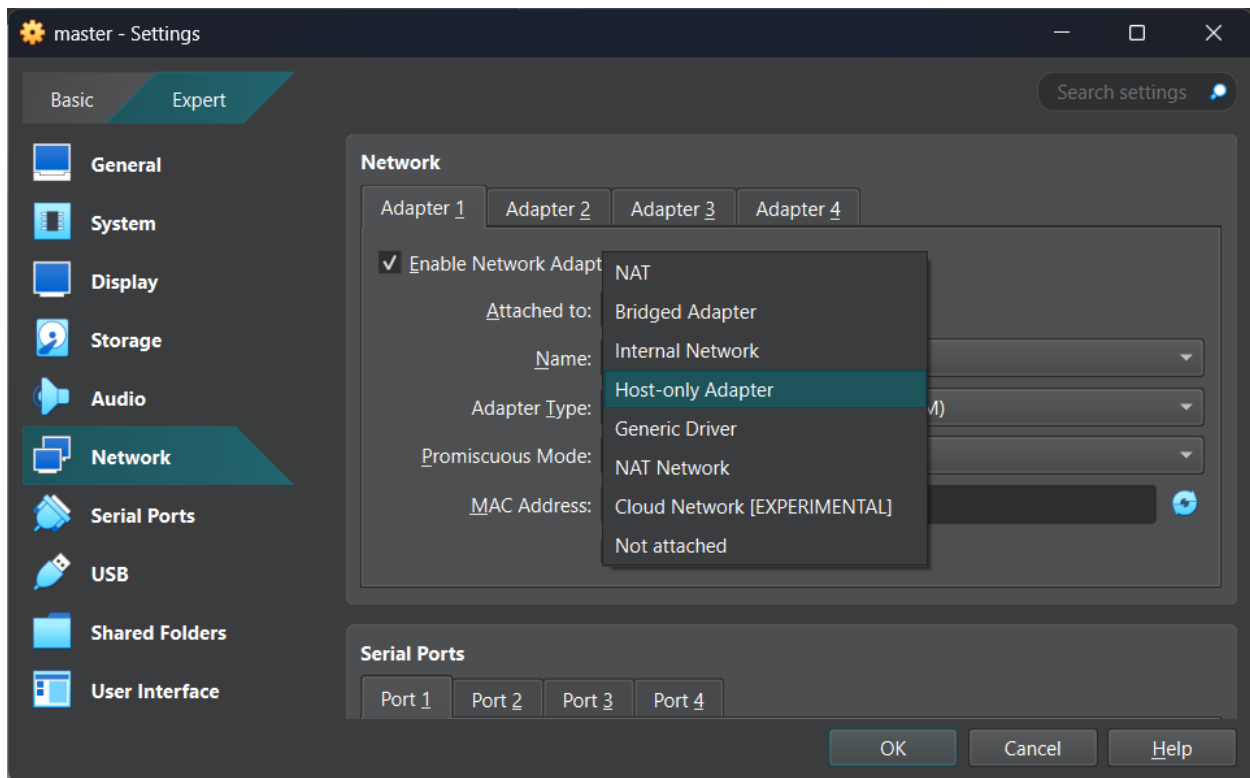
La mise en place d'un cluster Hadoop en mode multi-nœuds a permis de comprendre les aspects fondamentaux de la configuration d'un environnement de traitement distribué. Cette configuration offre plusieurs avantages par rapport à un déploiement en mode standalone :

- **Scalabilité** : possibilité d'ajouter des nœuds pour augmenter la capacité de traitement
- **Haute disponibilité** : réplication des données pour assurer la tolérance aux pannes
- **Performance** : distribution des tâches de calcul sur plusieurs nœuds

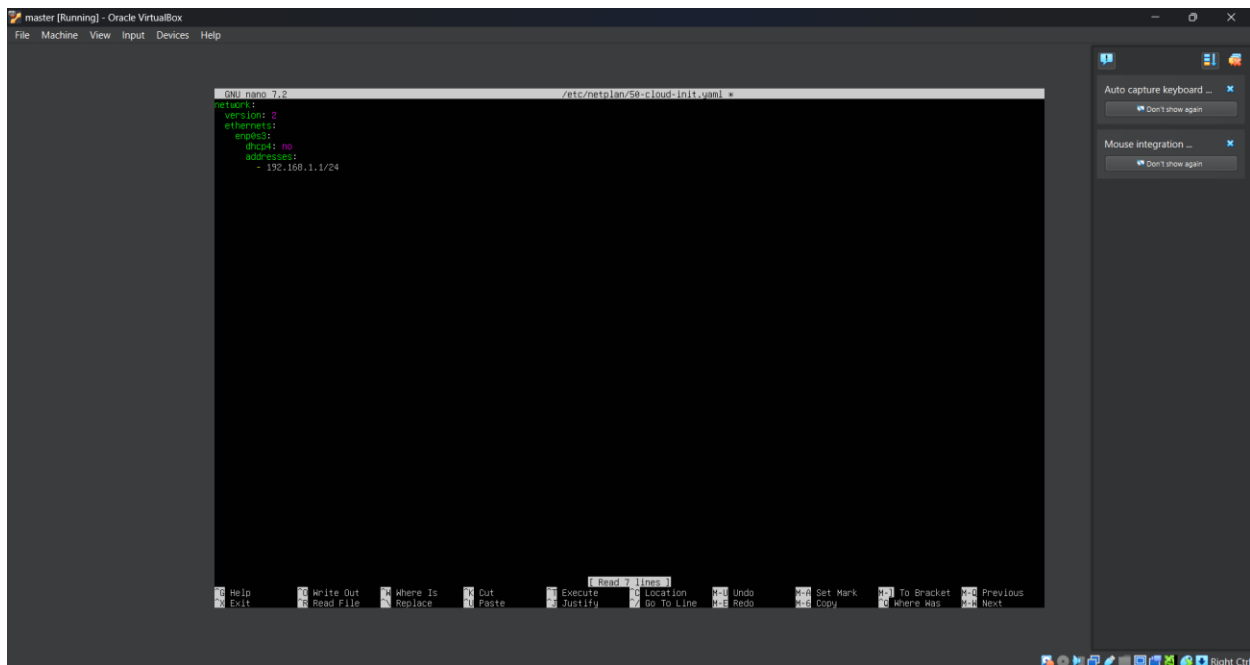
Les tests réalisés ont démontré le bon fonctionnement du cluster pour le stockage distribué avec HDFS et pour le traitement distribué avec YARN et MapReduce. Les jobs MapReduce développés lors des TP's précédents ont été exécutés avec succès sur notre infrastructure distribuée.

9. Annexe

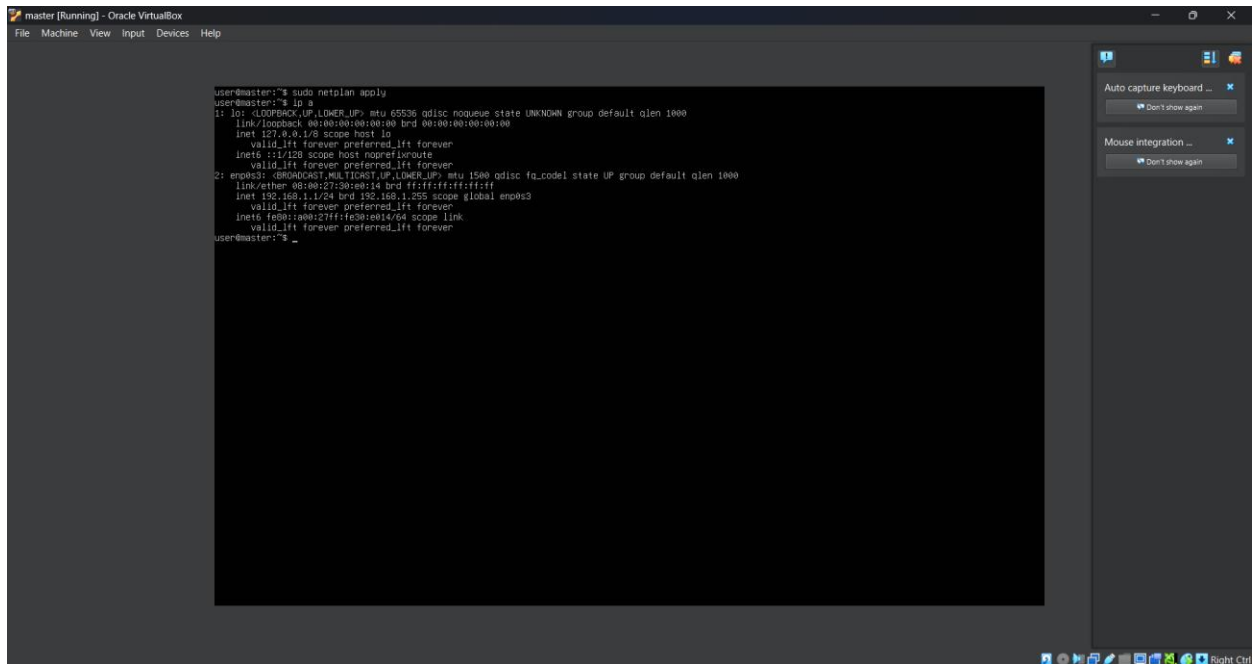
Remplaçons Nat adapter par Host-Only adapter



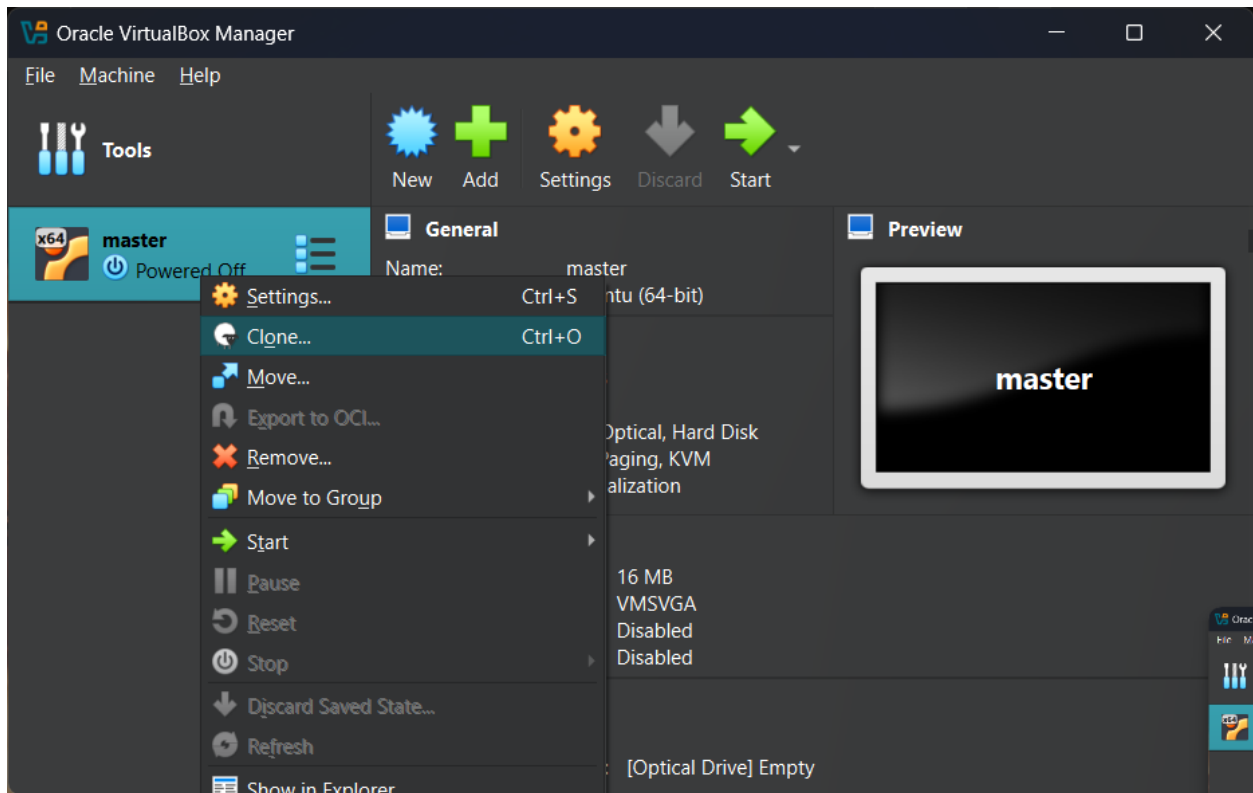
Modification de configurations de netplan







Application de modification




Clone de nœud master pour la création des nœuds slaves :



 Clone Virtual Machine





New Machine Name and Path

Name:

slave1

✓

Path:

C:\Users\aplu\VirtualBox VMs

✓

Clone Type

☒ Full Clone

☐ Linked Clone

Snapshots

☒ Current Machine State

☐ Everything

Additional Options

MAC Address Policy:

Generate new MAC addresses for all network adapters

Additional Options:

☐ Keep Disk Names

☐ Keep Hardware UUIDs

Help

Back

Finish

Cancel