Table des matières

Installation d'un cluster single-node

/!\ testé sur Centos 6.6 /!\

- Mise à jour des paquets et installations

```
yum -y update
yum -y install wget openssh openssh-clients nmap java-1.7.0-openjdk-devel.x86_64 java-1.7.0-openjdk.x86_64
```

- Configuration du fichier **/etc/hosts** :

On utilise des FQDN. Dans le fichier ci-dessous on trouve les autres noeuds qu'on rajoutera par la suite.

```
127.0.0.1   localhost localhost.localdomain localhost4 localhost4.localdomain4
::1         localhost localhost.localdomain localhost6 localhost6.localdomain6
192.168.2.5    hadoop-namenode.localdomain hadoop-namenode
192.168.2.11    hadoop-datanode1.localdomain hadoop-datanode1
192.168.2.12    hadoop-datanode2.localdomain hadoop-datanode2
```

- Désactivation de SELinux

```
sed -i "%enforcing%disabled%g" /etc/selinux/config
reboot
```

- Création du user/group Hadoop

```
groupadd hadoop
useradd -g hadoop hadoop
passwd hadoop

cat >> /home/hadoop/.bashrc << "EOF"

export JAVA_HOME=/usr/lib/jvm/jre-1.7.0-openjdk.x86_64
PATH=$PATH:$JAVA_HOME/bin

export HADOOP_INSTALL=/opt/hadoop
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export MAPRED_HOME=$YARN_INSTALL

export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin

EOF
```

- Téléchargement et décompression d'Hadoop

```
cd /opt/
wget http://wwwftp.ciril.fr/pub/apache/hadoop/common/hadoop-2.5.2/hadoop-2.5.2.tar.gz
tar vzxf hadoop-2.5.2.tar.gz
ln -s hadoop-2.5.2 hadoop
```

- Configuration de Hadoop

```
mv /opt/hadoop/etc/hadoop/core-site.xml /opt/hadoop/etc/hadoop/core-site.xml.bak

cat > /opt/hadoop/etc/hadoop/core-site.xml << "EOF"
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/opt/HDFS/tmp</value>
</property>
```

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

```
mv /opt/hadoop/etc/hadoop/mapred-site.xml /opt/hadoop/etc/hadoop/mapred-site.xml.bak

cat > /opt/hadoop/etc/hadoop/mapred-site.xml << "EOF"
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>mapreduce.jobtracker.address</name>
<value>local</value>
</property>
</configuration>
EOF
mv /opt/hadoop/etc/hadoop/yarn-site.xml /opt/hadoop/etc/hadoop/yarn-site.xml.bak
cat > /opt/hadoop/etc/hadoop/yarn-site.xml << "EOF"
<?xml version="1.0"?>
<configuration>
</configuration>
EOF
```

```
mv /opt/hadoop/etc/hadoop/hdfs-site.xml /opt/hadoop/etc/hadoop/hdfs-site.xml.bak

cat > /opt/hadoop/etc/hadoop/hdfs-site.xml << "EOF"
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/opt/HDFS/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/opt/HDFS/datanode</value>
</property>
</configuration>
EOF
```

- Création des répertoires + droits

```
mkdir /opt/HDFS
mkdir -p /opt/HDFS/namenode
mkdir -p /opt/HDFS/datanode
chown -R hadoop:hadoop /opt/HDFS
chown -R hadoop:hadoop /opt/hadoop
chown -R hadoop:hadoop /opt/hadoop-2.5.2
```

- Création d'un script de démarrage

```
cat > /etc/init.d/starthadoop << "EOF"

. /etc/init.d/functions


RETVAL=$?

case "$1" in
start)
echo $"Starting Hadoop server"
/bin/su - hadoop -c start-dfs.sh
/bin/su - hadoop -c start-yarn.sh
;;

stop)
echo $"Stopping Hadoop server"
/bin/su - hadoop -c stop-dfs.sh
/bin/su - hadoop -c stop-yarn.sh
```

```
;;

*)
echo $"Usage: $0 {start|stop}"
exit 1
;;
esac

exit $RETVAL
EOF

chmod u+x /etc/init.d/starthadoop
```

Ajout d'un datanode à un cluster existant

Pré-requis

- Copies de clés ssh des users **hadoop** sur chaque noeud
- Install de java et d'hadoop sur chaque noeud
- Copies des fichiers /etc/hosts sur chaque noeud

 Toutes les modifs des fichiers ci-dessous doivent être reportées sur chaque noeud

Configuration

- Fichiers **/opt/hadoop/etc/hadoop/masters** et **/opt/hadoop/etc/slaves** :

```
[hadoop@hadoop-namenode hadoop]$ cat masters
hadoop-namenode.localdomain

[hadoop@hadoop-namenode hadoop]$ cat slaves
hadoop-namenode.localdomain
hadoop-datanode1.localdomain
```

- Fichier **/opt/hadoop/etc/hadoop/core-site.xml**

```
[hadoop@hadoop-namenode hadoop]$ cat core-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>hadoop.tmp.dir</name>
    <value>/opt/HDFS/tmp</value>
  </property>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://hadoop-namenode.localdomain:9000</value>
  </property>
</configuration>
```

- Fichier **/opt/hadoop/etc/hadoop/hdfs-site.xml**

```
[hadoop@hadoop-namenode hadoop]$ cat hdfs-site.xml

<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>dfs.replication</name>
    <value>2</value>
  </property>
  <property>
    <name>dfs.namenode.name.dir</name>
    <value>file:/opt/HDFS/namenode</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:/opt/HDFS/datanode</value>
  </property>
```

```
</configuration>
```

- Fichier **/opt/hadoop/etc/hadoop/mapred-site.xml**

```
[hadoop@hadoop-namenode hadoop]$ cat mapred-site.xml

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<configuration>
  <property>
    <name>mapreduce.jobtracker.address</name>
    <value>hadoop-namenode:54311</value>
  </property>
</configuration>
```

Opérations de base

Démarrage et vérifications

```
[hadoop@hadoop-namenode /]$ start-dfs.sh
14/11/27 13:51:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [hadoop-namenode.localdomain]
hadoop-namenode.localdomain: starting namenode, logging to /opt/hadoop-2.5.2/logs/hadoop-hadoop-namenode-hadoop-namenode.localdomain.out
hadoop-namenode.localdomain: starting datanode, logging to /opt/hadoop-2.5.2/logs/hadoop-hadoop-datanode-hadoop-namenode.localdomain.out
hadoop-datanode1.localdomain: starting datanode, logging to /opt/hadoop-2.5.2/logs/hadoop-hadoop-datanode-hadoop-datanode1.localdomain.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /opt/hadoop-2.5.2/logs/hadoop-hadoop-secondarynamenode-hadoop-namenode.localdomain.out
14/11/27 13:51:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

note : le message *Unable to load native-hadoop library for your platform... using builtin-java classes where applicable* n'est pas bloquant. Il indique juste qu'Hadoop ne dispose pas des librairies natives (car machine 64 bits VS librairies 32 bits).

Check des process

```
[hadoop@hadoop-namenode /]$ jps
7720 DataNode
7898 SecondaryNameNode
7626 NameNode
6521 NodeManager
8006 Jps
6428 ResourceManager
```

```
[hadoop@hadoop-namenode /]$ ssh hadoop-datanode1 jps
11511 DataNode
11580 Jps
```

Statut de HDFS

```
[hadoop@hadoop-namenode /]$ hdfs dfsadmin -report
Configured Capacity: 6438158336 (6.00 GB)
Present Capacity: 3252350976 (3.03 GB)
DFS Remaining: 3252289536 (3.03 GB)
DFS Used: 61440 (60 KB)
DFS Used%: 0.00%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0

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

Name: 192.168.2.11:50010 (hadoop-datanode1.localdomain)
Hostname: hadoop-datanode1.localdomain
Decommission Status : Normal
Configured Capacity: 3219079168 (3.00 GB)
DFS Used: 28672 (28 KB)
Non DFS Used: 1588588544 (1.48 GB)
DFS Remaining: 1630461952 (1.52 GB)
DFS Used%: 0.00%
DFS Remaining%: 50.65%
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: 1
Last contact: Thu Nov 27 13:54:47 CET 2014


Name: 192.168.2.5:50010 (hadoop-namenode.localdomain)
Hostname: hadoop-namenode.localdomain
Decommission Status : Normal
Configured Capacity: 3219079168 (3.00 GB)
DFS Used: 32768 (32 KB)
Non DFS Used: 1597218816 (1.49 GB)
DFS Remaining: 1621827584 (1.51 GB)
DFS Used%: 0.00%
DFS Remaining%: 50.38%
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: 1
Last contact: Thu Nov 27 13:54:47 CET 2014
```

Manipulations de fichiers/répertoires

```
hadoop fs -ls /
hadoop fs -mkdir /user
hadoop fs -mkdir /user/hduser
hadoop fs -chown hduser /user/hduser
hadoop fs -chown :hadoop /user/hduser
hadoop fs -ls /user
```

Arrêter un node particulier

- Arrêt du datanode **hadoop-datanode1**

```
ssh hadoop-datanode1 /opt/hadoop/sbin/hadoop-daemon.sh --config /opt/hadoop/etc/hadoop/ stop datanode
```

Définir un namenode sur un noeud quelconque

- Dans **/opt/hadoop/etc/hadoop/hdfs-site.xml**, rajouter :

```
<property>
<name>dfs.namenode.secondary.http-address</name>
<value>hadoop-datanode2.localdomain:50090</value>
</property>
```

Simuler un crash d'un datanode

```
ssh hadoop-datanode1 /opt/hadoop/sbin/hadoop-daemon.sh --config /opt/hadoop/etc/hadoop/ stop datanode
ssh hadoop-datanode1 rm -rf /opt/HDFS
ssh hadoop-datanode1 /opt/hadoop/sbin/hadoop-daemon.sh --config /opt/hadoop/etc/hadoop/ start datanode
```

Lancement de jobs MapReduce

Calcul de pi

```
[hadoop@hadoop-namenode ~]$ hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.2.jar pi 10 100
Number of Maps  = 10
Samples per Map = 100
Wrote input for Map #0
Wrote input for Map #1
Wrote input for Map #2
Wrote input for Map #3
Wrote input for Map #4
Wrote input for Map #5
```

```
Wrote input for Map #6
Wrote input for Map #7
Wrote input for Map #8
Wrote input for Map #9
Starting Job
14/11/27 16:23:42 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
14/11/27 16:23:42 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
14/11/27 16:23:42 INFO input.FileInputFormat: Total input paths to process : 10
14/11/27 16:23:42 INFO mapreduce.JobSubmitter: number of splits:10
14/11/27 16:23:42 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local930526865_0001
...
...
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=1180
        File Output Format Counters
                Bytes Written=97
Job Finished in 2.929 seconds
Estimated value of Pi is 3.14800000000000000000
```

Calcul occurences de mots dans des livres

- Récupérer des livres au format .txt sur http://www.gutenberg.org

```
[hadoop@hadoop-namenode ~]$ hadoop fs -mkdir /user/hadoop/ebooks
[hadoop@hadoop-namenode ~]$ hadoop fs -put *.txt /user/hadoop/books

[hadoop@hadoop-namenode ~]$ hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.2.jar wordcount books output
14/11/27 16:27:16 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
14/11/27 16:27:16 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
14/11/27 16:27:16 INFO input.FileInputFormat: Total input paths to process : 3
14/11/27 16:27:16 INFO mapreduce.JobSubmitter: number of splits:3
14/11/27 16:27:16 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1050799219_0001
...
...
        File Input Format Counters
                Bytes Read=2856705
        File Output Format Counters
                Bytes Written=525222
```

- Voir le résultat

```
[hadoop@hadoop-namenode ~]$  hadoop fs -cat /user/hadoop/output/part-r-00000 |egrep  -w "^the|^house"
house   134
house!  1
house,  73
house,' 2
house--which    1
house-door      2
house-door,     3
house-door.     1
house-door.'    1
house-door;     1
house-keeping,  1
house-keeping.  1
house-top,      1
house.  38
house." 1
house.' 3
house:  1
house;  11
house?" 2
house?' 1
the     34498
the)    1
the---- 1
the.    2
the...  2
the.... 1
the.....        2
the.............        1
the:    1
the]    5
```

Sudoku

- Soit le sudoku ci-dessous :

```
[hadoop@hadoop-namenode ~]$ cat puzzle1.dta
8 5 ? 3 9 ? ? ? ?
? ? 2 ? ? ? ? ? ?
? ? 6 ? 1 ? ? ? 2
? ? 4 ? ? 3 ? 5 9
? ? 8 9 ? 1 4 ? ?
3 2 ? 4 ? ? 8 ? ?
9 ? ? ? 8 ? 5 ? ?
? ? ? ? ? ? 2 ? ?
? ? ? ? 4 5 ? 7 8
```

```
[hadoop@hadoop-namenode ~]$ hadoop jar /opt/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.5.2.jar sudoku puzzle1.dta
Solving puzzle1.dta
8 5 1 3 9 2 6 4 7
4 3 2 6 7 8 1 9 5
7 9 6 5 1 4 3 8 2
6 1 4 8 2 3 7 5 9
5 7 8 9 6 1 4 2 3
3 2 9 4 5 7 8 1 6
9 4 7 2 8 6 5 3 1
1 8 5 7 3 9 2 6 4
2 6 3 1 4 5 9 7 8

Found 1 solutions
```

Troubleshooting

ClusterID mismatch for namenode and datanodes

```
cat /tmp/hadoop-hdfs/dfs/name/current/VERSION
```

⇒ Recopier l'ID dans *data/VERSION* ⇒ Redémarrer datanode et namenode