


Matière : Big Data et architectures associées Enseignante responsable : Leila Baccour Enseignantes de TP : Hana MALLEK, Maysam Chaari, Samar Akremi	Section : D-LSI ADBD	AU : 2023-2024 
--	--------------------------------	---

TP2 : Installation et mise en marche Hadoop

I. Objectifs

L'objectif principal de ce TP est de suivre les étapes d'installation du Framework Apache Hadoop en modifiant les configurations des différents fichiers nécessaires pour le fonctionnement de ses composants.

II. Mise en place

1. Ouvrir le dossier Téléchargements, vous allez trouver le fichier `hadoop-2.10.2.tar.gz`
2. Extraire le dossier zippé avec le bouton droit puis extraire ici
3. Créer un dossier `hadoop` dans `/usr/local` à travers la commande suivante :
`sudo mkdir /usr/local/hadoop`
4. Lancer la commande suivante pour déplacer le dossier `hadoop-2.10.2` dans `/usr/local/hadoop`
`sudo mv Téléchargements/hadoop-2.10.2 /usr/local/hadoop`
5. Vérifier l'emplacement Hadoop à travers : **`ls /usr/local/hadoop`**
6. Définir les alias suivants dans le fichier caché `.bashrc` qui se trouve dans le dossier personnel

Notez: Le chemin de notre JDK est `/usr/lib/jvm/java-1.8.0-openjdk-amd64`

- Lancer la commande **`sudo gedit .bashrc`** ou **`sudo nano .bashrc`** et ajouter les lignes suivantes à la fin de ce fichier :

```
#HADOOP VARIABLES
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop/hadoop-2.10.2
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#END
```
- Ctrl+ X pour enregistrer et quitter, valider par O puis Entrée
- Redémarrer le terminal ou lancer la commande **`source .bashrc`** afin que les alias

ajoutés soient pris en compte.

III. Configurer hadoop sur un cluster à un simple nœud

Pour la configuration de Hadoop, il faut modifier 5 fichiers qui sont dans le dossier :

/usr/local/hadoop/hadoop-2.10.2/etc/hadoop/

1	2	3	4	5
hadoop-env.sh	core-site.xml	hdfs-site.xml	mapred-site.xml	yarn-site.xml

1. Modifier le fichier **hadoop-env.sh**

Le fichier **hadoop-env.sh** est essentiel au processus de démarrage des démons de Hadoop, qui sont les composants responsables du bon fonctionnement de la plateforme. Ces démons comprennent :

- **Namenode**, (Gérer le nommage des données)
- **Secondary Namenode**, (effectuer des opérations de sauvegarde et de maintenance pour le Namenode principal)
- **Datanode**, (Stocker les données ans HDFS)
- **JobTracker**, (Coordonner et superviser l'exécution du MapReduce.)
- **TaskTraker**. (Exécuter les tâches MapReduce)

Étant donné que Hadoop est développé en Java, il est nécessaire de définir le chemin vers le JDK (Kit de développement Java) afin de permettre l'activation de ses démons.

1.1. Ouvrir le fichier hadoop-env.sh à travers la commande suivante :

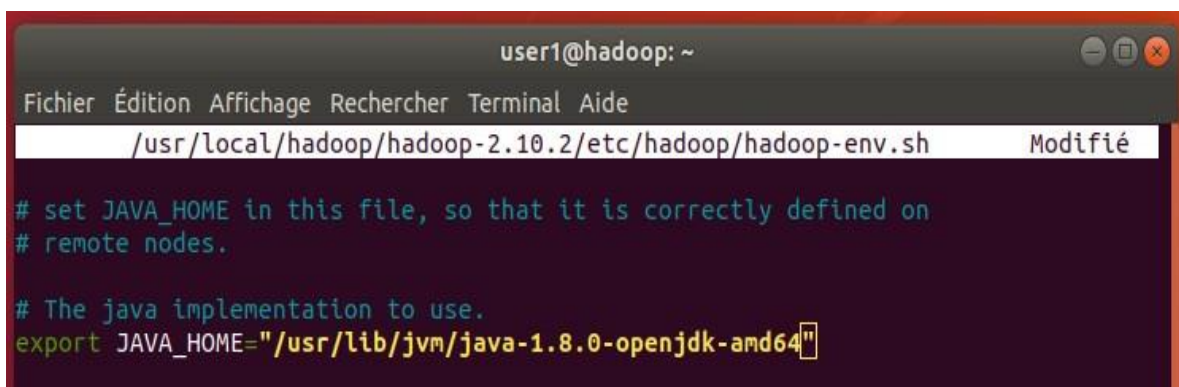
sudo nano \$HADOOP_INSTALL/etc/hadoop/hadoop-env.sh

ou bien

sudo nano /usr/local/hadoop/hadoop-2.10.2/etc/hadoop/hadoop-env.sh

1.2 Modifier la ligne où il y a « export JAVA_HOME={...} » et modifiez le chemin par **/usr/lib/jvm/java-1.8.0-openjdk-amd64**

1.3 Taper Ctrl X puis O puis Entrée pour enregistrer les modifications apportées.



```
user1@hadoop: ~  
Fichier Édition Affichage Rechercher Terminal Aide  
/usr/local/hadoop/hadoop-2.10.2/etc/hadoop/hadoop-env.sh Modifié  
# set JAVA_HOME in this file, so that it is correctly defined on  
# remote nodes.  
# The java implementation to use.  
export JAVA_HOME="/usr/lib/jvm/java-1.8.0-openjdk-amd64"
```

2. Modifier le fichier **core-site.xml**

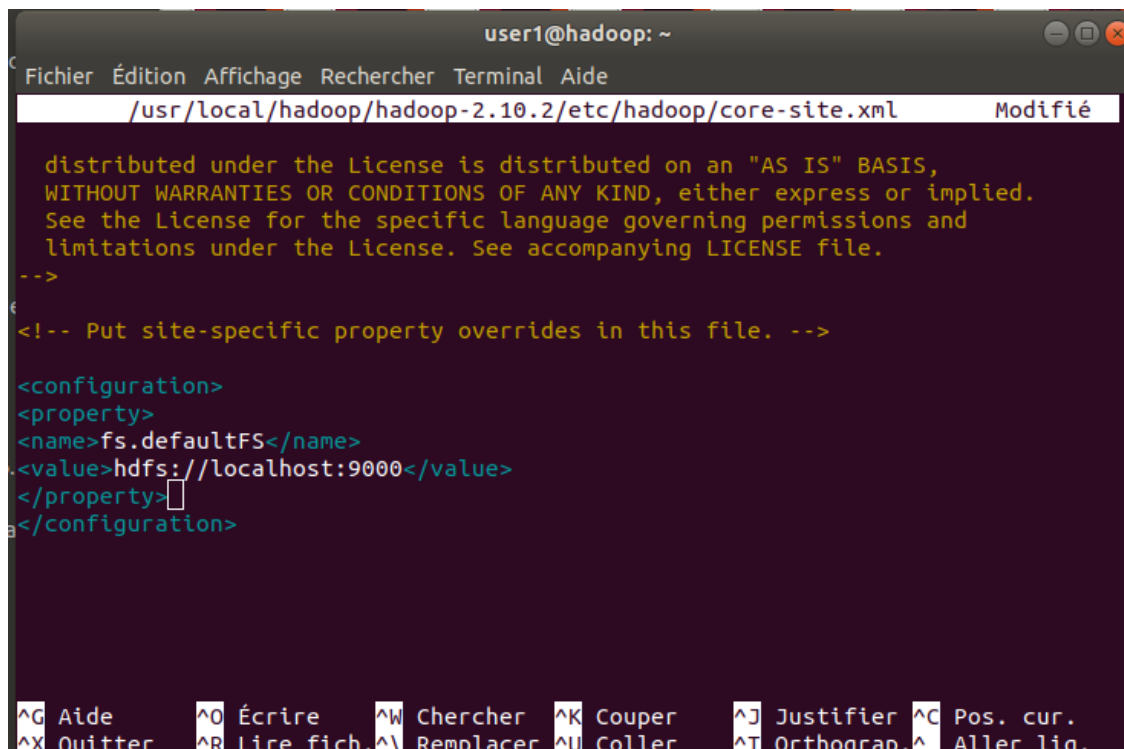
Le fichier core-site.xml a pour objectif d'informer les démons de Hadoop que le namenode s'exécute localement (localhost) sur le port 9000, ce port est associé au système de fichiers HDFS.

2.1. Ouvrir le fichier core-site.xml à travers la commande suivante :

```
sudo nano $HADOOP_INSTALL/etc/hadoop/core-site.xml
```

2.2. Ajouter à la fin du fichier les lignes suivantes dans la balise configuration la balise property suivante:

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



```
user1@hadoop: ~
Fichier Édition Affichage Rechercher Terminal Aide
/usr/local/hadoop/hadoop-2.10.2/etc/hadoop/core-site.xml Modifié

distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

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

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

2.3. Taper Ctrl X puis O puis Entrée pour enregistrer les modifications apportées.

3. Modifier le fichier **hdfs-site.xml**

Le fichier hdfs-site.xml contient des informations cruciales pour Hadoop et son système HDFS, notamment le **nombre de répliquions** (propriété 1), qui est défini à 1

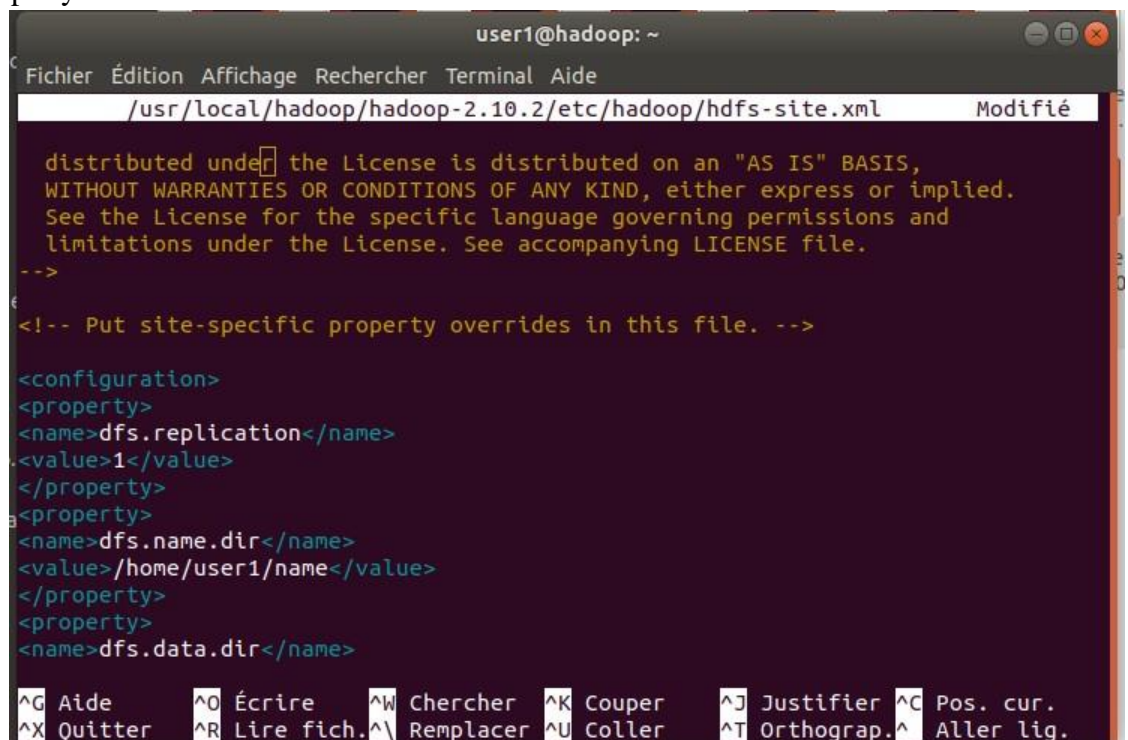
seule réplication dans notre puisque nous utilisons une seule machine. L'adresse de l'historique des transactions du NameNode (propriété 2) ainsi que l'adresse du stockage des blocs par les DataNode (propriété 3).

- 3.1. Ouvrir le fichier core-site.xml à travers la commande suivante :

sudo nano \$HADOOP_INSTALL/etc/hadoop/hdfs-site.xml

- 3.2. Ajouter à la fin du fichier les lignes suivantes dans la balise configuration les balises property suivantes:

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.name.dir</name>
<value>/home/user1/name</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/home/user1/data</value>
</property>
```



```
user1@hadoop: ~
Fichier Édition Affichage Rechercher Terminal Aide
/usr/local/hadoop/hadoop-2.10.2/etc/hadoop/hdfs-site.xml Modifié

distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->
<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.name.dir</name>
<value>/home/user1/name</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/home/user1/data</value>
</property>
</configuration>

^G Aide    ^O Écrire  ^W Chercher ^K Couper   ^J Justifier ^C Pos. cur.
^X Quitter ^R Lire fich. ^\ Remplacer ^U Coller   ^T Orthograp. ^_ Aller lig.
```

- 3.3. Taper Ctrl X puis O puis Entrée pour enregistrer les modifications apportées.

4. Modifier le fichier **mapred-site.xml**

Le fichier mapred-site.xml indique au package MapReduce qu'il s'exécutera en tant qu'application YARN, ce qui implique une séparation entre la gestion des ressources et la gestion des traitements.

- 4.1. Faire une copie du fichier `mapred-site.xml.template` sous le nom `mapred-site.xml` avec la commande :

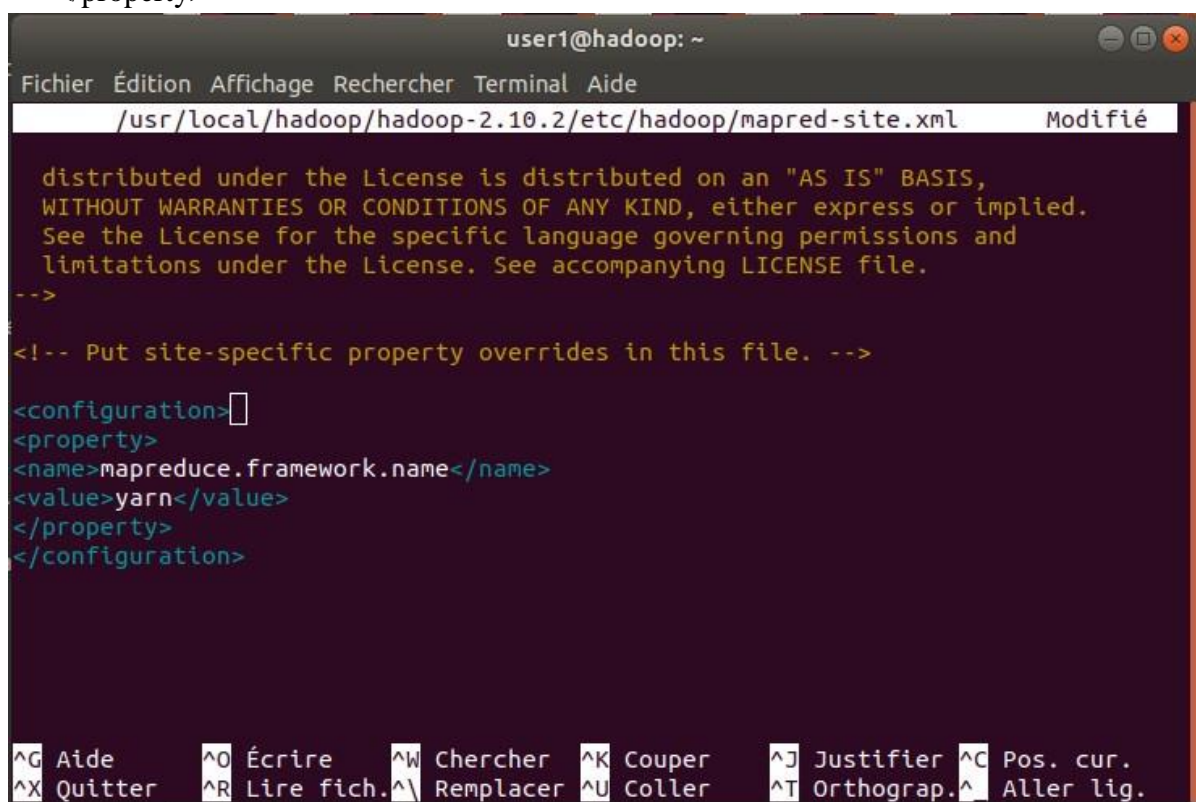
```
sudo cp $HADOOP_INSTALL/etc/hadoop/mapred-site.xml.template  
$HADOOP_INSTALL/etc/hadoop/mapred-site.xml
```

- 4.2. Ouvrir le fichier `mapred-site.xml` à travers la commande suivante :

```
sudo nano $HADOOP_INSTALL/etc/hadoop/mapred-site.xml
```

- 4.3. Dans la balise configuration, ajoutez les lignes suivantes :

```
<property>  
<name>mapreduce.framework.name</name>  
<value>yarn</value>  
</property>
```



```
user1@hadoop: ~  
Fichier Édition Affichage Rechercher Terminal Aide  
/usr/local/hadoop/hadoop-2.10.2/etc/hadoop/mapred-site.xml Modifié  
  
distributed under the License is distributed on an "AS IS" BASIS,  
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.  
See the License for the specific language governing permissions and  
limitations under the License. See accompanying LICENSE file.  
-->  
  
<!-- Put site-specific property overrides in this file. -->  
  
<configuration>  
<property>  
<name>mapreduce.framework.name</name>  
<value>yarn</value>  
</property>  
</configuration>
```

- 4.4. Taper Ctrl X puis O puis Entrée pour enregistrer les modifications apportées.

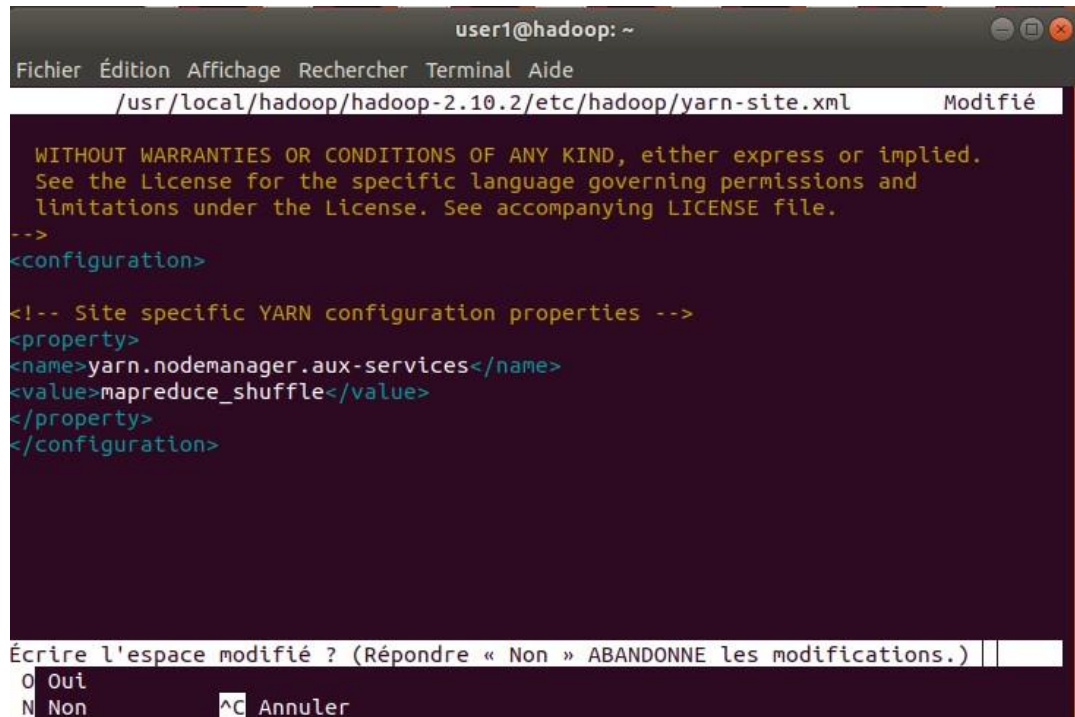
5. Modifier le fichier **yarn-site.xml**

Le fichier `yarn-site.xml` configure le Node Manager pour qu'il mette en place un service auxiliaire spécifique qui guide le MapReduce sur la manière de gérer efficacement le transfert des données (shuffling) au sein de l'infrastructure YARN.

```
sudo nano $HADOOP_INSTALL/etc/hadoop/yarn-site.xml
```

5.1. Dans la balise configuration, ajoutez les lignes suivantes :

```
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
```



5.2. Taper Ctrl X puis O puis Entrée pour enregistrer les modifications apportées.

IV. Vérifier l'installation

Après avoir achevé la configuration de Hadoop, il est essentiel d'effectuer une vérification de son installation. Cette étape nécessite de formater le NameNode à chaque démarrage des services Hadoop, ce qui permet de réinitialiser les fichiers temporaires tout en préservant l'architecture des fichiers existants.

hdfs namenode -format

Remarque: Ce formatage ne supprime pas les fichiers de données essentiels stockés dans HDFS. Au lieu de cela, il supprime les métadonnées et les fichiers temporaires qui ont été générés pendant les précédentes opérations Hadoop.

1. Vérifier les services actifs

1.1. Démarrer le système hadoop

- a) Avant de démarrer hadoop, vérifiez les services actifs avec la commande **jps**

```
user1@hadoop:~$ jps
3581 Jps
user1@hadoop:~$
```

- b) Démarrer le système Hadoop par la commande :

start-all.sh

ou bien

start-dfs.sh start-yarn.sh

```
user1@hadoop: ~
Fichier Édition Affichage Rechercher Terminal Aide
3581 Jps
user1@hadoop:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
23/09/30 11:44:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
localhost: starting namenode, logging to /usr/local/hadoop/hadoop-2.10.2/logs/ha
dooop-user1-namenode-hadoop.out
localhost: starting datanode, logging to /usr/local/hadoop/hadoop-2.10.2/logs/ha
dooop-user1-datanode-hadoop.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ECDSA key fingerprint is SHA256:97oD8vHkRZDMRceK6xMdrG80vXOX/xvdcn9os33hT/E.
Are you sure you want to continue connecting (yes/no)? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ECDSA) to the list of known hosts
.
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/hadoop-2.10.2/
logs/hadoop-user1-secondarynamenode-hadoop.out
23/09/30 11:46:13 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/hadoop-2.10.2/logs/yarn-u
ser1-resourcemanager-hadoop.out
```

1.2. Vérifier les services actifs

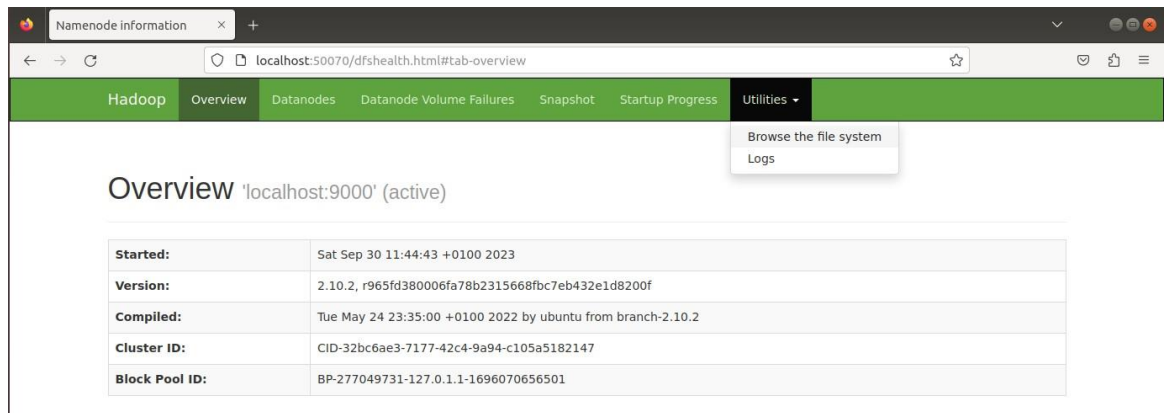
Après le démarrage du Hadoop, vérifier les services actifs par la commande : **jps**

Il est impératif que les six démons soient en activité pour déterminer si l'installation a été effectuée avec succès. Le bon fonctionnement est un indicateur crucial de la réussite de l'installation du Hadoop.

```
yarn-user1-nodemanager-hadoop.out
user1@hadoop:~$ jps
4355 ResourceManager
3940 DataNode
4196 SecondaryNameNode
4487 NodeManager
3790 NameNode
4543 Jps
user1@hadoop:~$
```

1.3. Visualiser l'interface web du NameNode

Avec votre navigateur web de la machine virtuelle, vous pouvez accéder à l'interface web NameNode via <http://localhost:50070/>



The screenshot shows a web browser window with the title 'Namenode information'. The address bar displays 'localhost:50070/dfshealth.html#tab-overview'. The navigation bar includes links for Hadoop, Overview (selected), Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. A dropdown menu for Utilities shows 'Browse the file system' and 'Logs'. The main content area is titled 'Overview 'localhost:9000' (active)' and contains a table with the following information:

Started:	Sat Sep 30 11:44:43 +0100 2023
Version:	2.10.2, r965fd380006fa78b2315668fbc7eb432e1d8200f
Compiled:	Tue May 24 23:35:00 +0100 2022 by ubuntu from branch-2.10.2
Cluster ID:	CID-32bc6ae3-7177-42c4-9a94-c105a5182147
Block Pool ID:	BP-277049731-127.0.1.1-1696070656501