

**Rapport de :**

**TP 1 : “Installation et configuration d'Apache Hadoop et  
exécution d'un programme MapReduce dans un cluster  
hadoop à nœud unique et à nœuds multiples.”**

**Réalisé par :**

→ Riali Mouad

→ Addi Kamal

**encadré par :**

→ Pr. D.Zaidouni

## Table de matières :

- I. Installation et configuration d'un nœud unique d'Apache Hadoop 3.2.1 .**
  - 1. Création d'un utilisateur hduser**
  - 2. Mise en place de la clé ssh**
  - 3. Installation de JAVA 8**
  - 4. Installation d'Apache Hadoop 3.2.1**
  - 5. Configuration d'Apache Hadoop 3.2.1**
- II. Exécution d'un programme Map/Reduce dans un cluster à nœud unique .**
- III. Configuration d'un cluster multi-nœuds d'Apache Hadoop .**
  - 1. Attribution statique d'adresse IP à la machine hadoopmaster**
  - 2. Modification des fichiers de configuration de hadoop**
  - 3. Clonage de la machine hadoopmaster**
  - 4. Modification à faire dans les machines slave1 et slave2**
  - 5. Connexion entre les machines du cluster**
- IV. Exécution d'un programme Map/Reduce dans un cluster multi-nœuds .**



ORACLE  
**VirtualBox**



Pré-requis techniques :

x *Oracle VM VirtualBox-6.0 :*

Oracle VM VirtualBox (anciennement VirtualBox) est un logiciel libre de virtualisation publié par Oracle.

Lien de Téléchargement :

ORACLE  
**VirtualBox**



[https://download.virtualbox.org/virtualbox/6.0.12/virtualbox-6.0\\_6.0.12-133076~Ubuntu~bionic\\_amd64.deb](https://download.virtualbox.org/virtualbox/6.0.12/virtualbox-6.0_6.0.12-133076~Ubuntu~bionic_amd64.deb)

x *Ubuntu 18.04.3* :

Ubuntu est un système d'exploitation GNU/Linux basé sur la distribution Linux Debian. Il est développé, commercialisé et maintenu pour les ordinateurs individuels (desktop), les serveurs (Server) et les objets connectés (Core) par la société Canonical.



Lien de Téléchargement de la version Ubuntu 20.04 :

<https://ubuntu.com/download/desktop/thank-you?version=20.04.1&architecture=amd64>

x *Apache Hadoop version=3.2.1* :

est un framework libre et open source écrit en Java destiné à faciliter la création d'applications distribuées et échelonnables permettant aux applications de travailler avec des milliers de nœuds et des pétaoctets de données. Ainsi chaque nœud est constitué de machines standard regroupées en grappe.



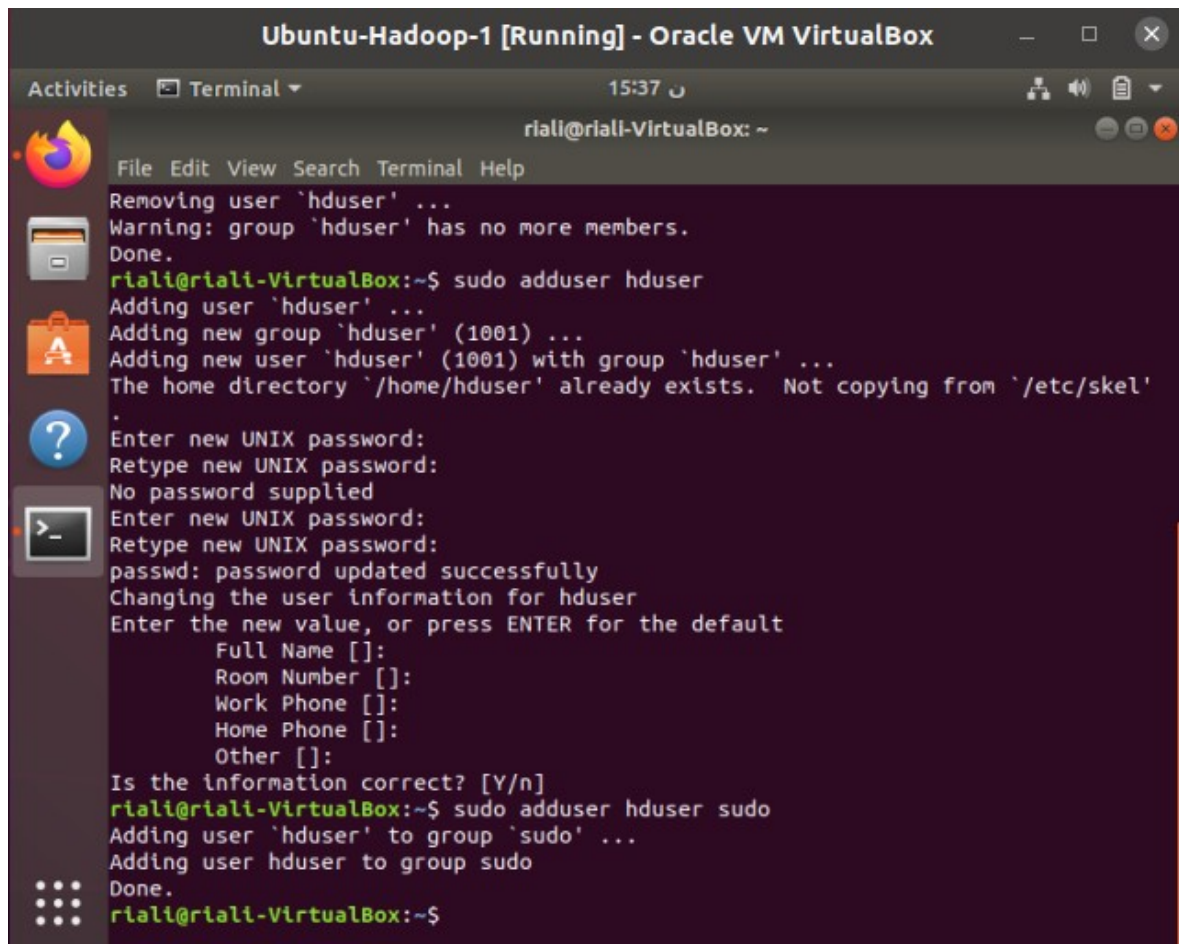
Lien de Téléchargement :

<https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz>

pour **Java 8** : <https://github.com/sanyoushi/java-buildpack.git>

# I. Installation et configuration d'un nœud unique d'Apache Hadoop 3.2.1

## Création d'un utilisateur hduser:



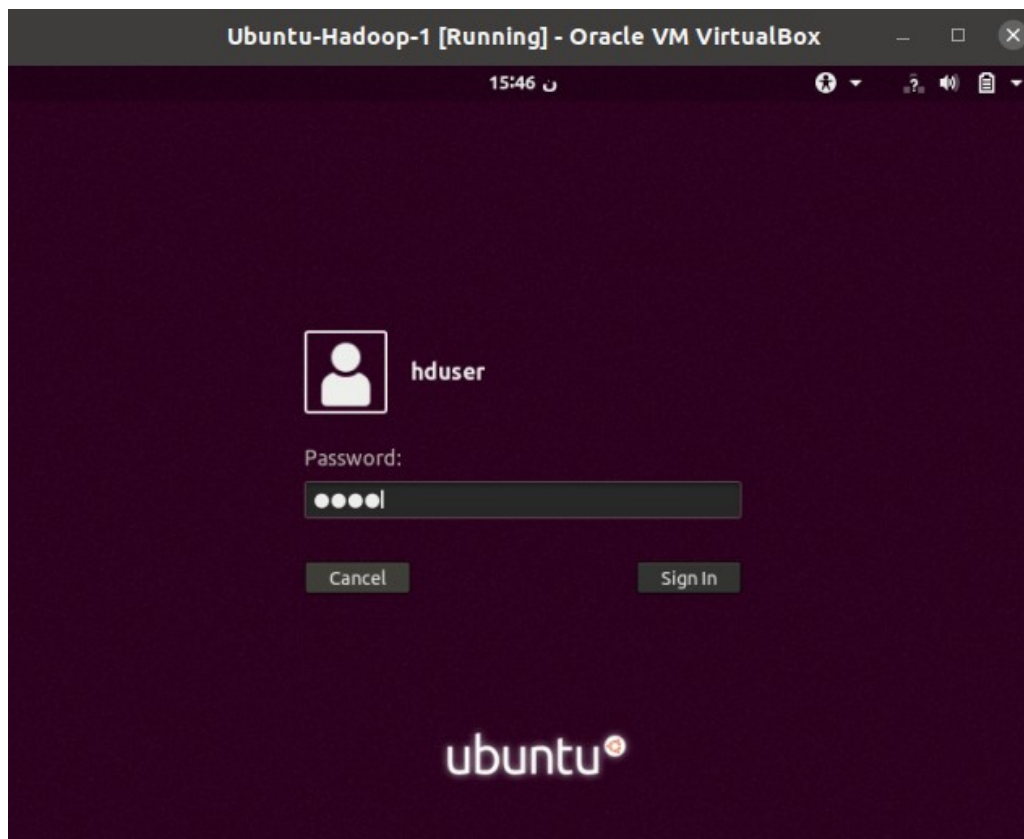
```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
Activities Terminal 15:37
riali@riali-VirtualBox: ~
File Edit View Search Terminal Help
Removing user `hduser' ...
Warning: group `hduser' has no more members.
Done.
riali@riali-VirtualBox:~$ sudo adduser hduser
Adding user `hduser' ...
Adding new group `hduser' (1001) ...
Adding new user `hduser' (1001) with group `hduser' ...
The home directory `/home/hduser' already exists. Not copying from `/etc/skel'
.
Enter new UNIX password:
Retype new UNIX password:
No password supplied
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hduser
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]
riali@riali-VirtualBox:~$ sudo adduser hduser sudo
Adding user `hduser' to group `sudo' ...
Adding user hduser to group sudo
Done.
riali@riali-VirtualBox:~$
```

- Pour ajouter un nouveau “user” qu’on va l’appeler “hduser” , il faut juste appeler la commande suivante : `sudo adduser hduser`
- Ensuite on va appeler la commande : `sudo adduser hduser sudo`

pour inclure le nouveau user dans le groupe sudo pour faire de lui un sudoer dont l’objectif est de lui permettre à exécuter des commandes en tant que superutilisateur .

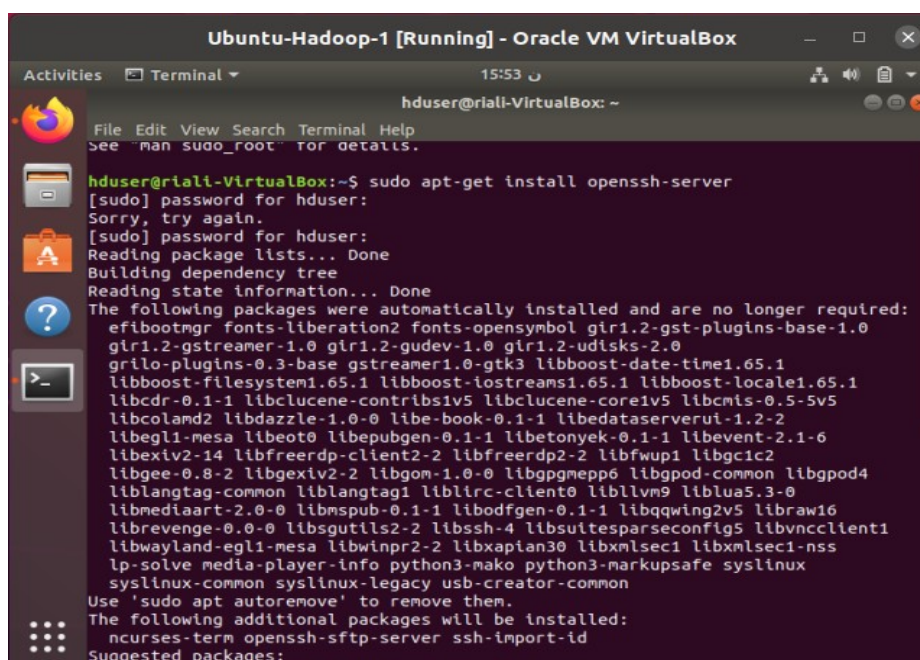
## Mise en place de la clé ssh :

En premier lieu , il faut qu'on se connecte par l'utilisateur : hduser



## Installer le serveur openssh :

Sur ubuntu 18.04 ,  
pour installer le  
serveur openssh , il  
ne faut qu'appeler la  
commande suivante :  
`sudo apt-get install  
openssh-server`

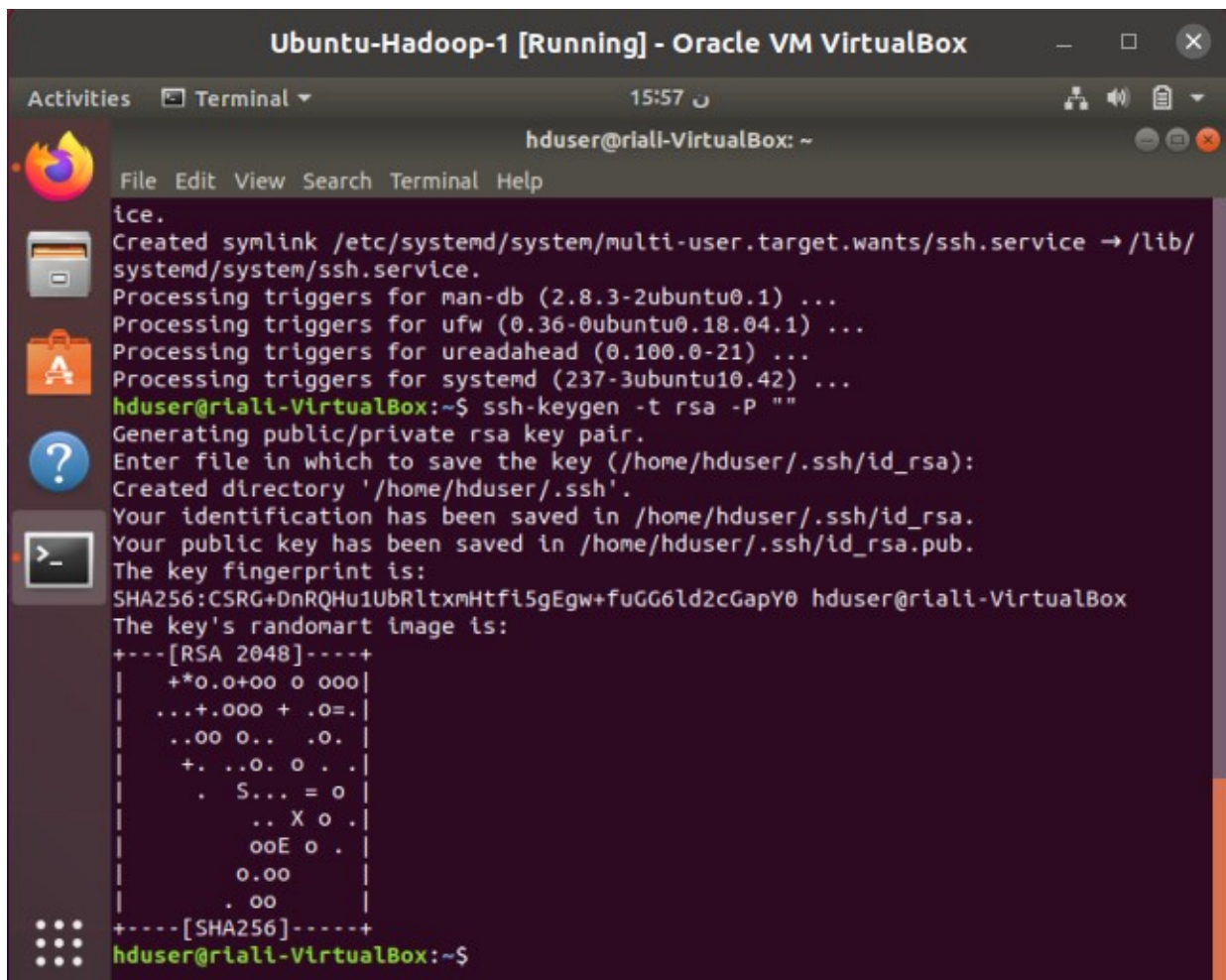




## Mettre en place la clé ssh pour son propre compte :

Dans cette étape, on va comprendre pourquoi on doit se connecter par hduser, tout simplement , car il faut mettre en place une clé ssh pour celui-ci, pour qu'il puisse dans ce qui suit, se connecter aux autres machines de cluster qui vont être clonées à partir de cette machine-là, alors il faut taper ces commandes-là :

- `ssh-keygen -t rsa -P ""`
- `cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys`
- `chmod 0600 ~/.ssh/authorized_keys`



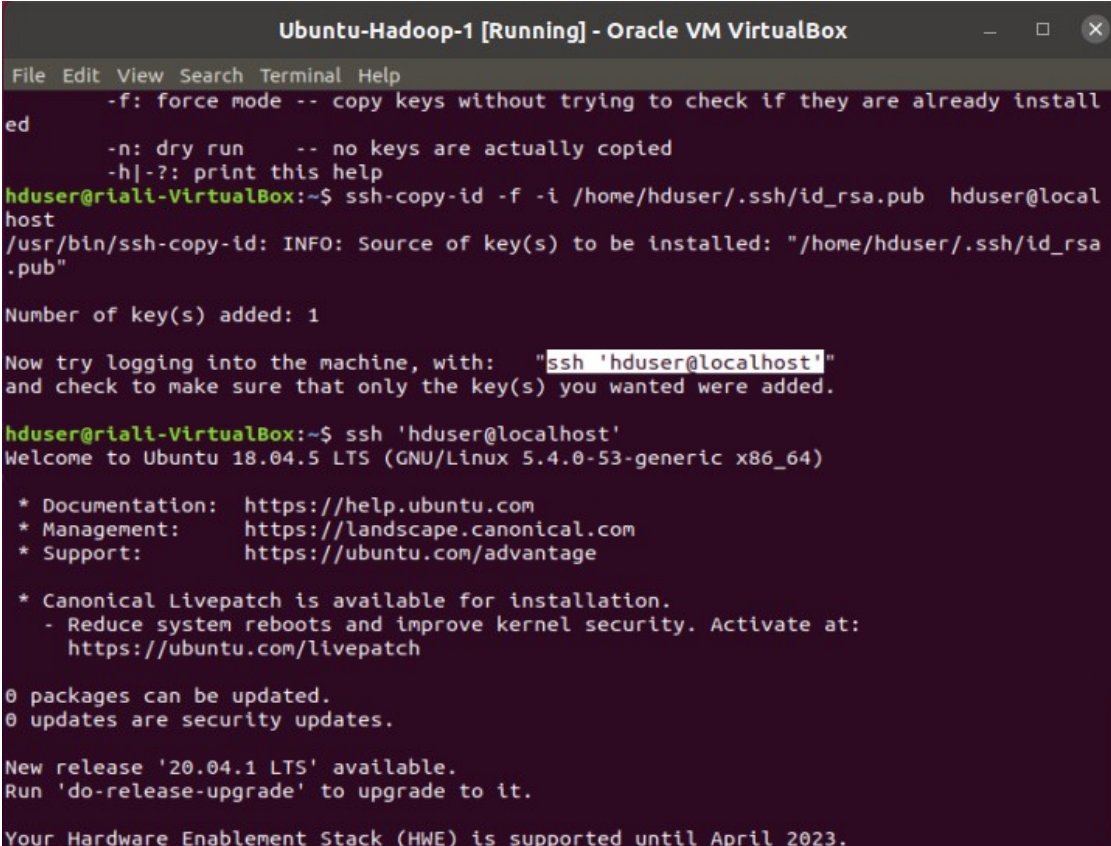
```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
Activities Terminal 15:57
hduser@riali-VirtualBox: ~
File Edit View Search Terminal Help
ice.
Created symlink /etc/systemd/system/multi-user.target.wants/ssh.service → /lib/
systemd/system/ssh.service.
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
Processing triggers for ufw (0.36-0ubuntu0.18.04.1) ...
Processing triggers for ureadahead (0.100.0-21) ...
Processing triggers for systemd (237-3ubuntu10.42) ...
hduser@riali-VirtualBox:~$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hduser/.ssh/id_rsa):
Created directory '/home/hduser/.ssh'.
Your identification has been saved in /home/hduser/.ssh/id_rsa.
Your public key has been saved in /home/hduser/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:CSRG+DnRQHw1UbRltxmHtfi5gEgw+fuGG6ld2cGapY0 hduser@riali-VirtualBox
The key's randomart image is:
+---[RSA 2048]---+
|  +*o.o+oo o ooo|
|  ...+.ooo + .o=.|
|  ..oo o.. .o. |
|  +. ..o. o . . |
|  . S... = o |
|  .. X o . |
|  ooE o . |
|  o.o |
|  . oo |
+-----[SHA256]-----+
hduser@riali-VirtualBox:~$
```

## Copier la clé public sur le serveur localhost :

Après la mise e place de la cle ssh pou le compte hduser, c'est le moment pour copier la clé public sur le serveur localhost :

- `ssh-copy-id -i /home/hduser/.ssh/id_rsa.pub`  
[hduser@localhost](#)

Enfin on va tester la connexion a localhost par la commande suivante : `ssh 'hduser@localhost'`



```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
File Edit View Search Terminal Help
-f: force mode -- copy keys without trying to check if they are already install
ed
-n: dry run -- no keys are actually copied
-h|-?: print this help
hduser@riali-VirtualBox:~$ ssh-copy-id -f -i /home/hduser/.ssh/id_rsa.pub hduser@local
host
/usr/bin/ssh-copy-id: INFO: Source of key(s) to be installed: "/home/hduser/.ssh/id_rsa
.pub"

Number of key(s) added: 1

Now try logging into the machine, with: "ssh 'hduser@localhost'"
and check to make sure that only the key(s) you wanted were added.

hduser@riali-VirtualBox:~$ ssh 'hduser@localhost'
Welcome to Ubuntu 18.04.5 LTS (GNU/Linux 5.4.0-53-generic x86_64)

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

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

0 packages can be updated.
0 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Your Hardware Enablement Stack (HWE) is supported until April 2023.
```



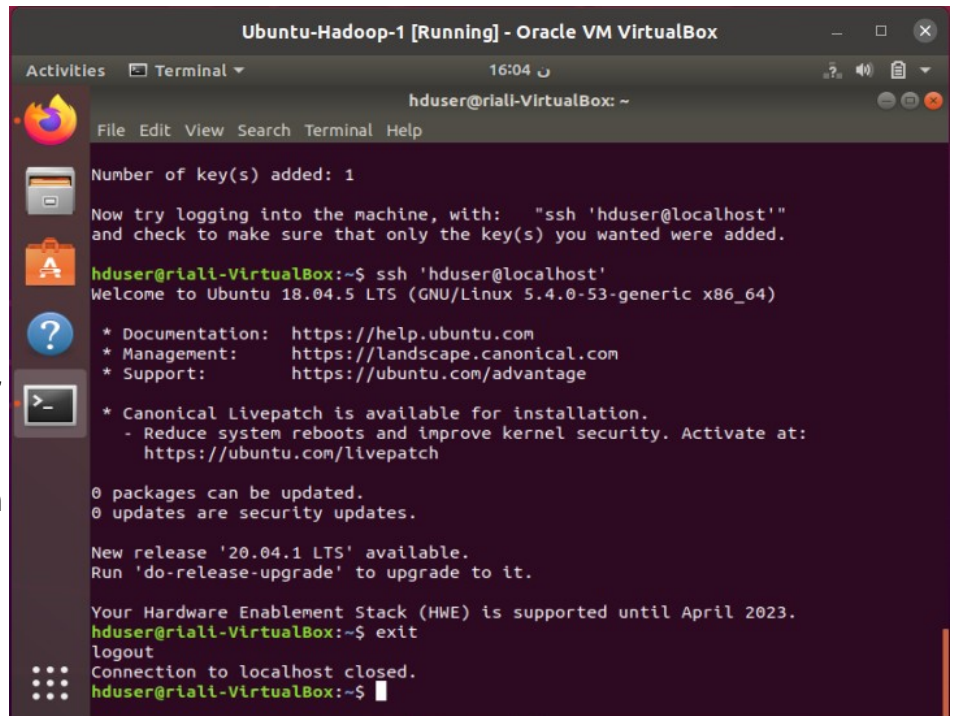
## Installation de JAVA 8 :

Comme on a déjà mentionné au début, Hadoop est programmé par java, alors il est impérativement d'avoir un JDK pour une bonne configuration et un bon fonctionnement de Hadoop , alors dans ce TP on a choisi d'installer

le JAVA 8, D'abord on va télécharger [jdk-8u71-linux-x64.tar.gz](http://www.oracle.com/technetwork/java/javase-downloads-1344955.html) et après on va suite les étapes suivantes :

- `tar -zxvf jdk-8u71-linux-x64.tar.gz`
- `mv jdk1.8.0_71/ /opt/java/`

ps : on a déjà créé le rep /opt/java



```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
16:04
hduser@riali-VirtualBox: ~
File Edit View Search Terminal Help

Number of key(s) added: 1
Now try logging into the machine, with: "ssh 'hduser@localhost'"
and check to make sure that only the key(s) you wanted were added.

hduser@riali-VirtualBox:~$ ssh 'hduser@localhost'
Welcome to Ubuntu 18.04.5 LTS (GNU/Linux 5.4.0-53-generic x86_64)

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

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

0 packages can be updated.
0 updates are security updates.

New release '20.04.1 LTS' available.
Run 'do-release-upgrade' to upgrade to it.

Your Hardware Enablement Stack (HWE) is supported until April 2023.
hduser@riali-VirtualBox:~$ exit
logout
Connection to localhost closed.
hduser@riali-VirtualBox:~$
```

```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
Activities Terminal 18:19
hduser@riali-VirtualBox:/opt/java/jdk1.8.0_71
File Edit View Search Terminal Help
tar: jdk1.8.0_71: Cannot mkdir: Permission denied
tar: jdk1.8.0_71/jre/lib/ext/dnsns.jar: Cannot open: No such file or directory
jdk1.8.0_71/jre/lib/ext/jfxrt.jar
tar: jdk1.8.0_71: Cannot mkdir: Permission denied
^C
riali@riali-VirtualBox:/home/hduser/Documents$ sudo tar -zxvf jdk-8u71-linux-i586.tar.gz
jdk1.8.0_71/
jdk1.8.0_71/THIRDPARTYLICENSEREADME.txt
jdk1.8.0_71/src.zip
jdk1.8.0_71/man/
jdk1.8.0_71/man/ja_JP.UTF-8/
jdk1.8.0_71/man/ja_JP.UTF-8/man1/
jdk1.8.0_71/man/ja_JP.UTF-8/man1/jjs.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/jstatd.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/javadoc.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/javaws.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/jrunscript.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/jvisualvm.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/extcheck.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/pack200.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/rmiregistry.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/unpack200.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/schemagen.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/appletviewer.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/jstat.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/wsgen.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/jmc.1
jdk1.8.0_71/man/ja_JP.UTF-8/man1/jdb.1
```

Après le déplacement du répertoire jdk1.8.0\_71 vers opt/java, on va mettre à jour les liens par défaut de jDkjdkj, pour cela on va appeler les commandes suivantes dans le nvx répertoire de java :

- update-alternatives --install /usr/bin/java java /opt/java/jdk1.8.0\_71/bin/java 100

```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
Activities Terminal 18:20
hduser@riali-VirtualBox:/opt/java/jdk1.8.0_71
File Edit View Search Terminal Help
jdk1.8.0_71 jdk-8u71-linux-i586.tar.gz
riali@riali-VirtualBox:/home/hduser/Documents$ sudo mv jdk1.8.0_71/ /opt/java/
riali@riali-VirtualBox:/home/hduser/Documents$ cd /opt/java/jdk1.8.0_71/
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ sudo update-alternatives --install /usr/bin/java java /opt/java/jdk1.8.0_71/bin/java 100
update-alternatives: unknown option '--install/usr/bin/java'

Use 'update-alternatives --help' for program usage information.
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ sudo update-alternatives --install /usr/bin/java java /opt/java/jdk1.8.0_71/bin/java 100
update-alternatives: unknown option '--install/usr/bin/java'

Use 'update-alternatives --help' for program usage information.
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ sudo update-alternatives --install /usr/bin/java java /opt/java/jdk1.8.0_71/bin/java 100
update-alternatives: --install needs <link> <name> <path> <priority>

Use 'update-alternatives --help' for program usage information.
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ sudo update-alternatives --install /usr/bin/java java /opt/java/jdk1.8.0_71/bin/java 100
update-alternatives: using /opt/java/jdk1.8.0_71/bin/java to provide /usr/bin/java (java) in auto mode
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ update-alternatives --config java

There is only one alternative in link group java (providing /usr/bin/java): /opt/java/jdk1.8.0_71/bin/java
Nothing to configure.
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ sudo update-alternatives --install /usr/bin/javac javac /opt/java/jdk1.8.0_71/bin/javac 100
```

- `update-alternatives --config java`

et on va refaire la même chose pour javac :

- `update-alternatives --install /usr/bin/javac javac /opt/java/jdk1.8.0_71/bin/javac 100`
- `update-alternatives -config javac`
- `sudo nano` OU `sudo vim` OU `sudo gedit /etc/profile` , pour pouvoir modifier le fichier et ajouter ces trois lignes vers sa fin :

`export JAVA_HOME=/opt/java/jdk1.8.0_71/`

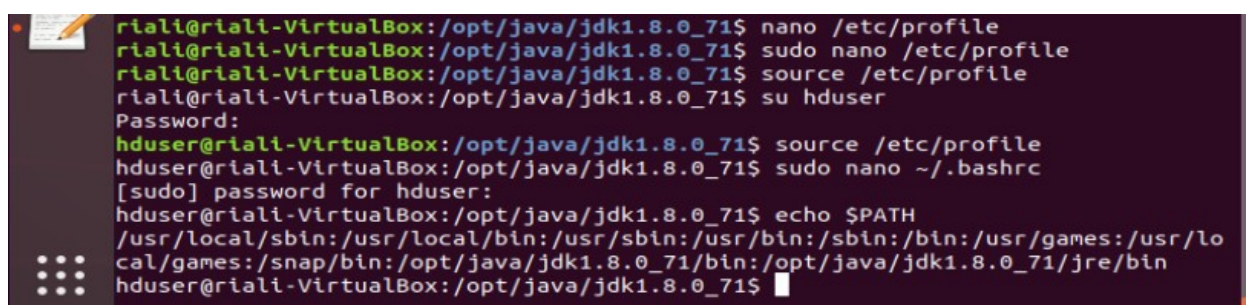
`export JRE_HOME=/opt/java/jdk1.8.0._71/jre`

`export`

`PATH=$PATH:/opt/java/jdk1.8.0_71/bin:/opt/java/jdk1.8.0_71/jre/bin`

- refaire les mêmes opérations pour le fichier : `~/.bashrc`
- Exécuter les 2 commandes suivantes : `source /etc/profile` ET `source ~/.bashrc`

ET VOILA →



```

riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ nano /etc/profile
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ sudo nano /etc/profile
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ source /etc/profile
riali@riali-VirtualBox:/opt/java/jdk1.8.0_71$ su hduser
Password:
hduser@riali-VirtualBox:/opt/java/jdk1.8.0_71$ source /etc/profile
hduser@riali-VirtualBox:/opt/java/jdk1.8.0_71$ sudo nano ~/.bashrc
[sudo] password for hduser:
hduser@riali-VirtualBox:/opt/java/jdk1.8.0_71$ echo $PATH
/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/games:/usr/local/games:/snap/bin:/opt/java/jdk1.8.0_71/bin:/opt/java/jdk1.8.0_71/jre/bin
hduser@riali-VirtualBox:/opt/java/jdk1.8.0_71$

```

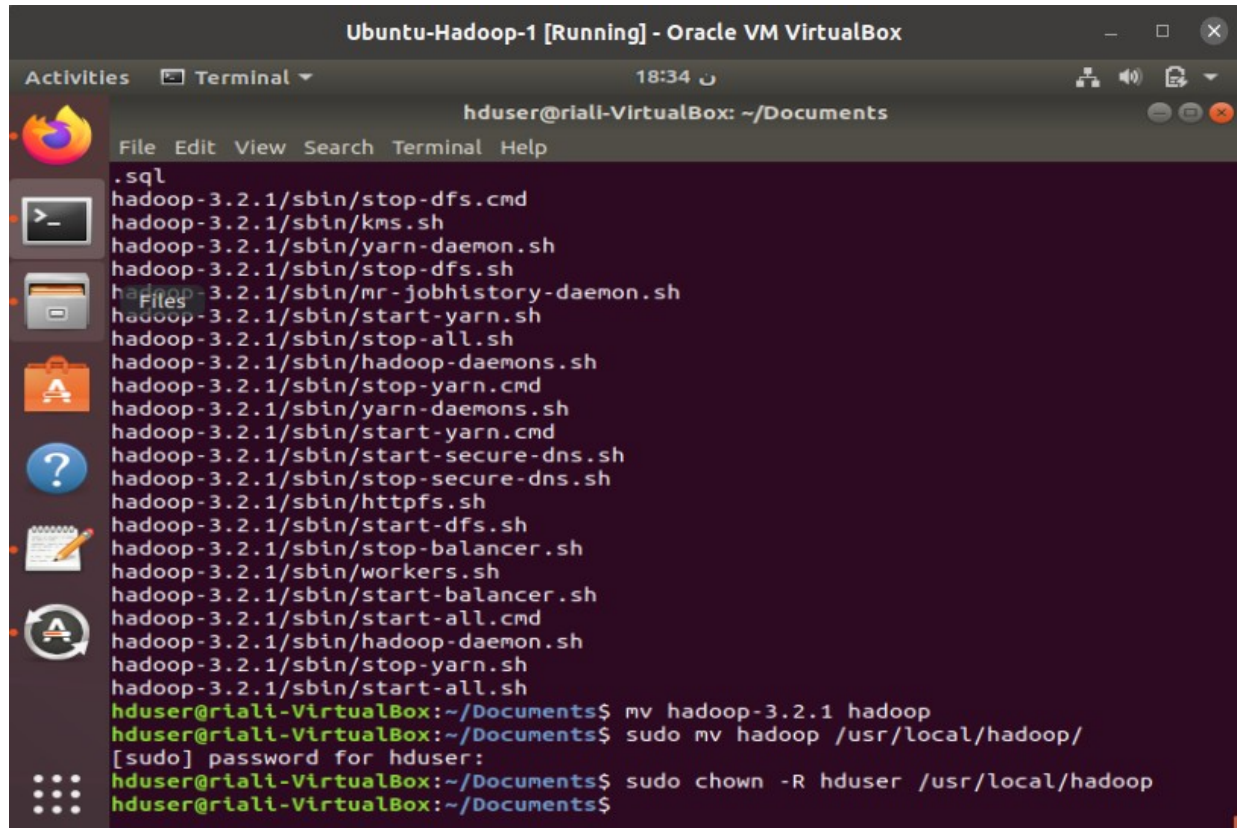
ici on observe que le variable \$PATH nous donne le correcte chemin vers JAVA 8



### Installation d'Apache Hadoop 3.2.1

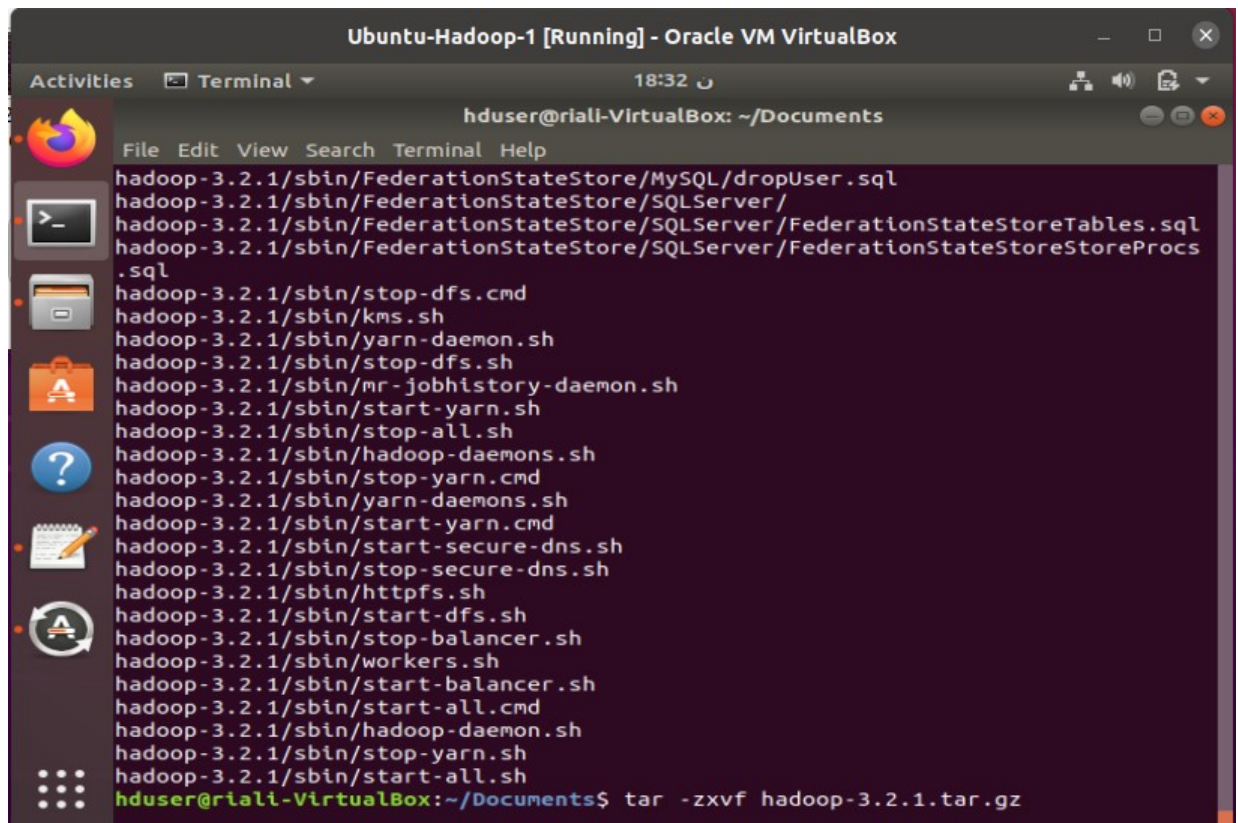
On va émettre les étapes faites pour Java 8 , sauf qu'on va déplacer le répertoire vers : `"/usr/local/"`

On aura alors :



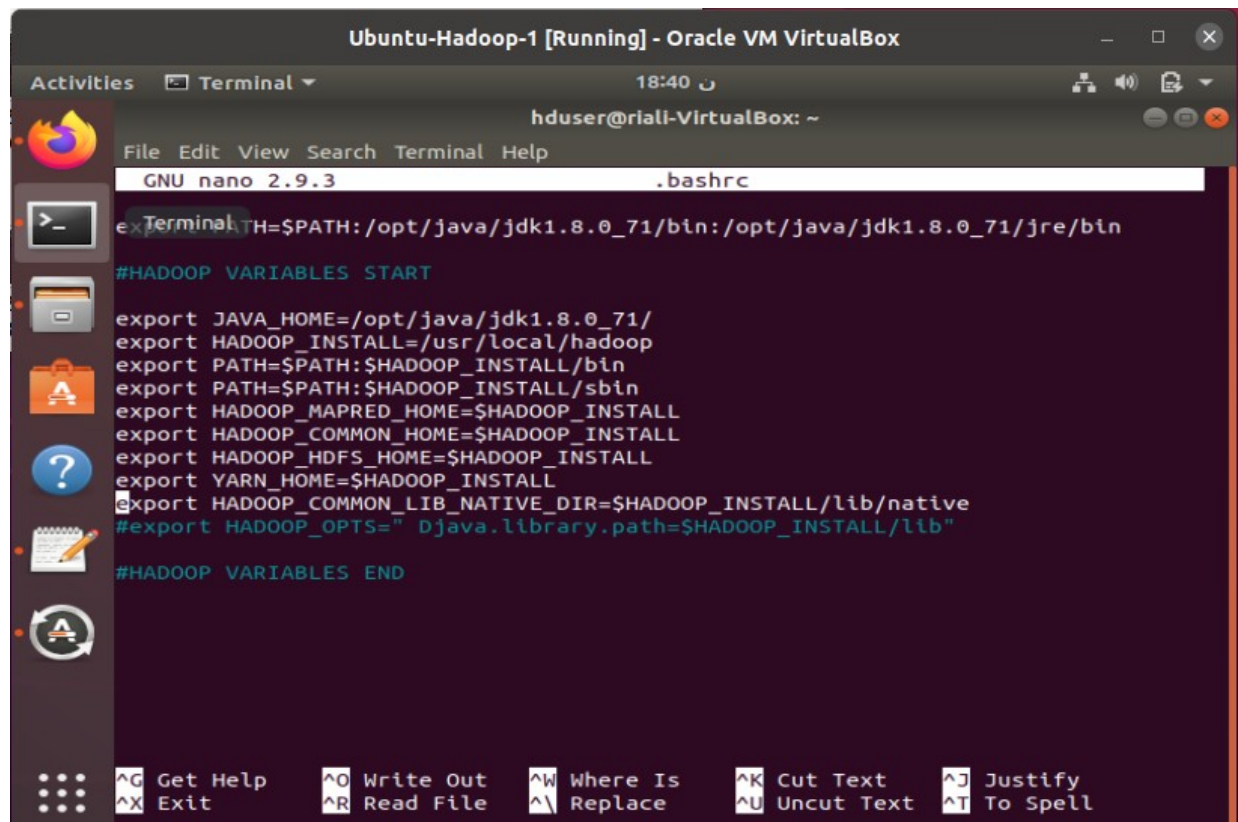
```
hduser@riali-VirtualBox: ~/Documents
hadoop-3.2.1/sbin/stop-dfs.cmd
hadoop-3.2.1/sbin/kms.sh
hadoop-3.2.1/sbin/yarn-daemon.sh
hadoop-3.2.1/sbin/stop-dfs.sh
hadoop-3.2.1/sbin/mr-jobhistory-daemon.sh
hadoop-3.2.1/sbin/start-yarn.sh
hadoop-3.2.1/sbin/stop-all.sh
hadoop-3.2.1/sbin/hadoop-daemons.sh
hadoop-3.2.1/sbin/stop-yarn.cmd
hadoop-3.2.1/sbin/yarn-daemons.sh
hadoop-3.2.1/sbin/start-yarn.cmd
hadoop-3.2.1/sbin/start-secure-dns.sh
hadoop-3.2.1/sbin/stop-secure-dns.sh
hadoop-3.2.1/sbin/httpfs.sh
hadoop-3.2.1/sbin/start-dfs.sh
hadoop-3.2.1/sbin/stop-balancer.sh
hadoop-3.2.1/sbin/workers.sh
hadoop-3.2.1/sbin/start-balancer.sh
hadoop-3.2.1/sbin/start-all.cmd
hadoop-3.2.1/sbin/hadoop-daemon.sh
hadoop-3.2.1/sbin/stop-yarn.sh
hadoop-3.2.1/sbin/start-all.sh
hduser@riali-VirtualBox:~/Documents$ mv hadoop-3.2.1 hadoop
hduser@riali-VirtualBox:~/Documents$ sudo mv hadoop /usr/local/hadoop/
[sudo] password for hduser:
hduser@riali-VirtualBox:~/Documents$ sudo chown -R hduser /usr/local/hadoop
hduser@riali-VirtualBox:~/Documents$
```

la ligne : `" sudo chown -R hduser /usr/local/hadoop"` donne le droit à hduser d'opérer sur les fichiers et répertoires inclus dans `/usr/local/hadoop` de manière réursive.



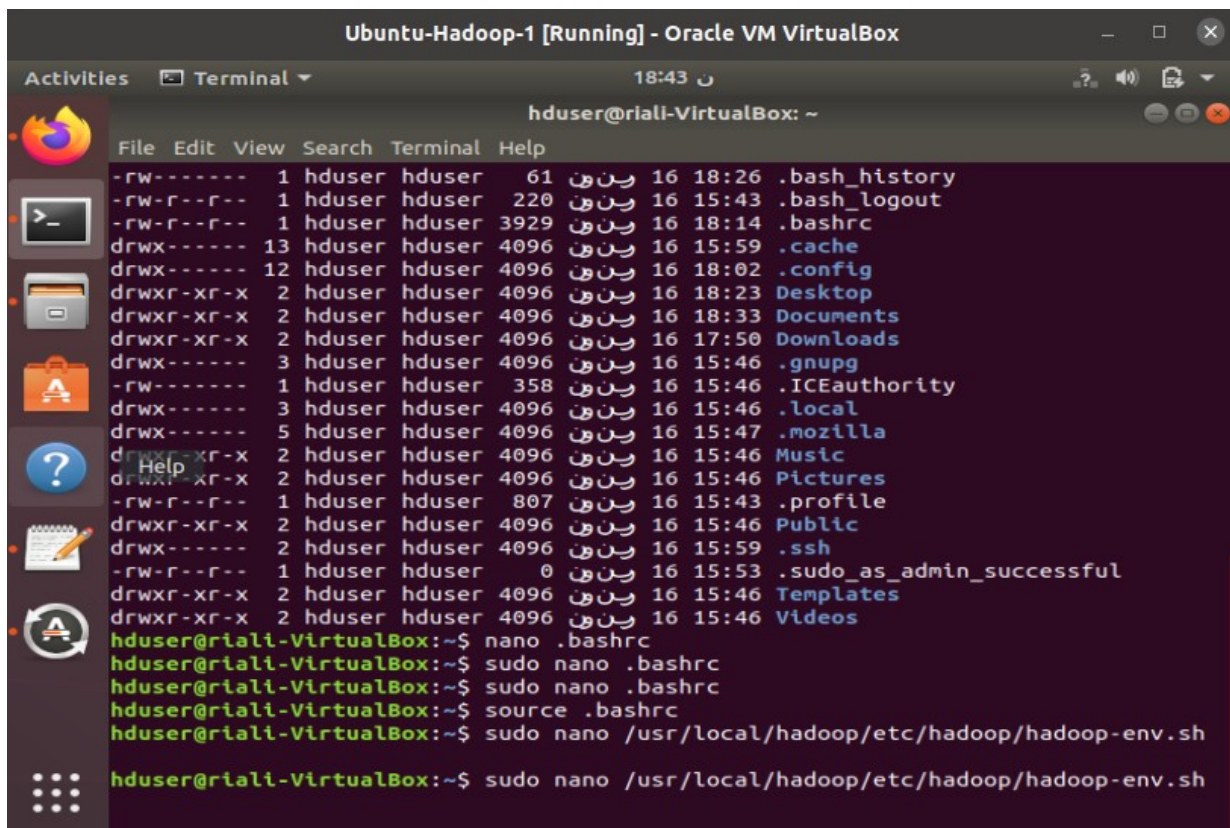
```
hduser@riall-VirtualBox: ~/Documents
hadoop-3.2.1/sbin/FederationStateStore/MySQL/dropUser.sql
hadoop-3.2.1/sbin/FederationStateStore/SQLServer/
hadoop-3.2.1/sbin/FederationStateStore/SQLServer/FederationStateStoreTables.sql
hadoop-3.2.1/sbin/FederationStateStore/SQLServer/FederationStateStoreStoreProcs
.sql
hadoop-3.2.1/sbin/stop-dfs.cmd
hadoop-3.2.1/sbin/kfs.sh
hadoop-3.2.1/sbin/yarn-daemon.sh
hadoop-3.2.1/sbin/stop-dfs.sh
hadoop-3.2.1/sbin/mr-jobhistory-daemon.sh
hadoop-3.2.1/sbin/start-yarn.sh
hadoop-3.2.1/sbin/stop-all.sh
hadoop-3.2.1/sbin/hadoop-daemons.sh
hadoop-3.2.1/sbin/stop-yarn.cmd
hadoop-3.2.1/sbin/yarn-daemons.sh
hadoop-3.2.1/sbin/start-yarn.cmd
hadoop-3.2.1/sbin/start-secure-dns.sh
hadoop-3.2.1/sbin/stop-secure-dns.sh
hadoop-3.2.1/sbin/httpfs.sh
hadoop-3.2.1/sbin/start-dfs.sh
hadoop-3.2.1/sbin/stop-balancer.sh
hadoop-3.2.1/sbin/workers.sh
hadoop-3.2.1/sbin/start-balancer.sh
hadoop-3.2.1/sbin/start-all.cmd
hadoop-3.2.1/sbin/hadoop-daemon.sh
hadoop-3.2.1/sbin/stop-yarn.sh
hadoop-3.2.1/sbin/start-all.sh
hduser@riall-VirtualBox:~/Documents$ tar -zxvf hadoop-3.2.1.tar.gz
```

Ainsi, il faut modifier les fichier “.bashrc” et “/etc/profile” et ajouter les chemins vers \$HADOOP\_HOME ...etc,comme il est clair dans l’image au dessous :



```
GNU nano 2.9.3 .bashrc
export PATH=$PATH:/opt/java/jdk1.8.0_71/bin:/opt/java/jdk1.8.0_71/jre/bin
#HADOOP VARIABLES START
export JAVA_HOME=/opt/java/jdk1.8.0_71/
export HADOOP_INSTALL=/usr/local/hadoop
export PATH=$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"
#HADOOP VARIABLES END
```

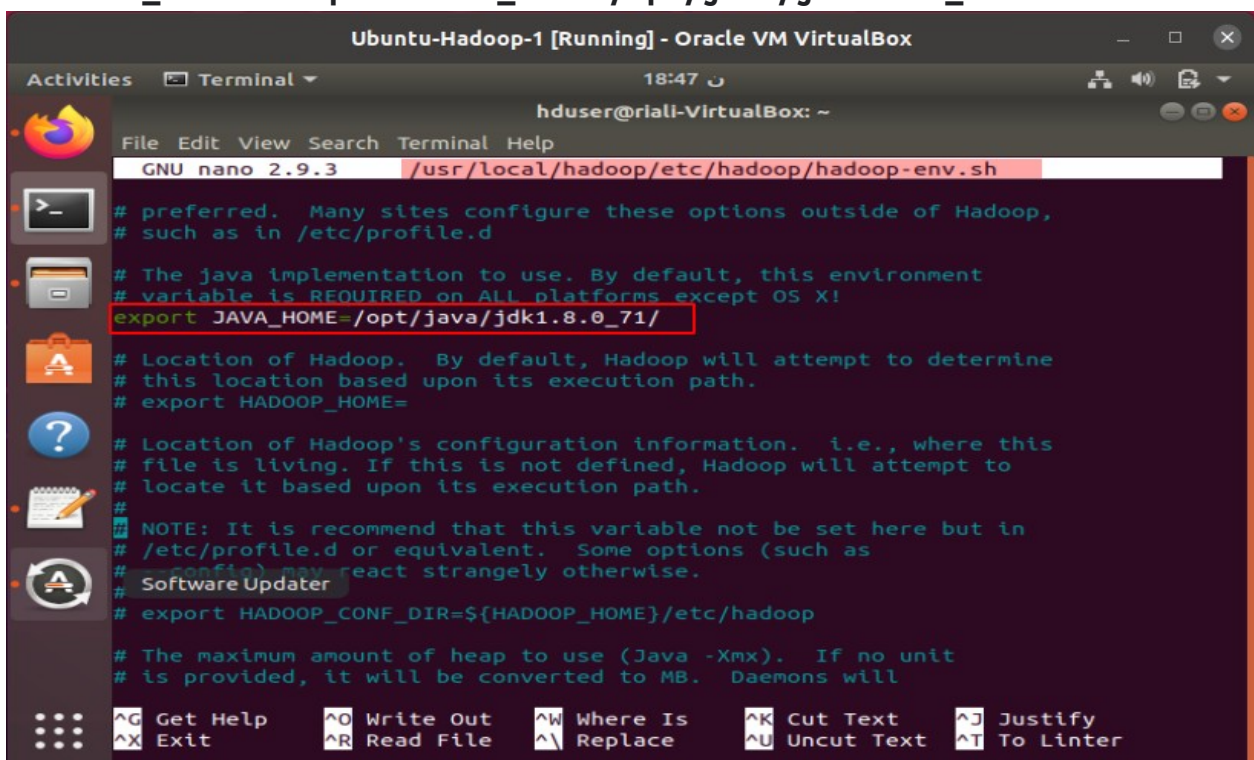




```
hduser@riali-VirtualBox: ~  
File Edit View Search Terminal Help  
-rw----- 1 hduser hduser 61 16 18:26 .bash_history  
-rw-r--r-- 1 hduser hduser 220 16 15:43 .bash_logout  
-rw-r--r-- 1 hduser hduser 3929 16 18:14 .bashrc  
drwx----- 13 hduser hduser 4096 16 15:59 .cache  
drwx----- 12 hduser hduser 4096 16 18:02 .config  
drwxr-xr-x 2 hduser hduser 4096 16 18:23 Desktop  
drwxr-xr-x 2 hduser hduser 4096 16 18:33 Documents  
drwxr-xr-x 2 hduser hduser 4096 16 17:50 Downloads  
drwx----- 3 hduser hduser 4096 16 15:46 .gnupg  
-rw----- 1 hduser hduser 358 16 15:46 .ICEauthority  
drwx----- 3 hduser hduser 4096 16 15:46 .local  
drwx----- 5 hduser hduser 4096 16 15:47 .mozilla  
drwxr-xr-x 2 hduser hduser 4096 16 15:46 Music  
drwxr-xr-x 2 hduser hduser 4096 16 15:46 Pictures  
-rw-r--r-- 1 hduser hduser 807 16 15:43 .profile  
drwxr-xr-x 2 hduser hduser 4096 16 15:46 Public  
drwx----- 2 hduser hduser 4096 16 15:59 .ssh  
-rw-r--r-- 1 hduser hduser 0 16 15:53 .sudo_as_admin_successful  
drwxr-xr-x 2 hduser hduser 4096 16 15:46 Templates  
drwxr-xr-x 2 hduser hduser 4096 16 15:46 Videos  
hduser@riali-VirtualBox:~$ nano .bashrc  
hduser@riali-VirtualBox:~$ sudo nano .bashrc  
hduser@riali-VirtualBox:~$ sudo nano .bashrc  
hduser@riali-VirtualBox:~$ source .bashrc  
hduser@riali-VirtualBox:~$ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh  
hduser@riali-VirtualBox:~$ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh
```

Maintenant, on ouvre le fichier  
/usr/local/hadoop/etc/hadoop/hadoop-env.sh et on modifie la  
variable d'environnement

**JAVA\_HOME : export JAVA\_HOME=/opt/java/jdk1.8.0\_71**



```
GNU nano 2.9.3 /usr/local/hadoop/etc/hadoop/hadoop-env.sh  
# preferred. Many sites configure these options outside of Hadoop,  
# such as in /etc/profile.d  
  
# The java implementation to use. By default, this environment  
# variable is REQUIRED on ALL platforms except OS X!  
export JAVA_HOME=/opt/java/jdk1.8.0_71/  
  
# Location of Hadoop. By default, Hadoop will attempt to determine  
# this location based upon its execution path.  
# export HADOOP_HOME=  
  
# Location of Hadoop's configuration information. i.e., where this  
# file is living. If this is not defined, Hadoop will attempt to  
# locate it based upon its execution path.  
#  
# NOTE: It is recommend that this variable not be set here but in  
# /etc/profile.d or equivalent. Some options (such as  
# --config) may react strangely otherwise.  
#  
# export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop  
  
# The maximum amount of heap to use (Java -Xmx). If no unit  
# is provided, it will be converted to MB. Daemons will  
  
^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify  
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Linter
```

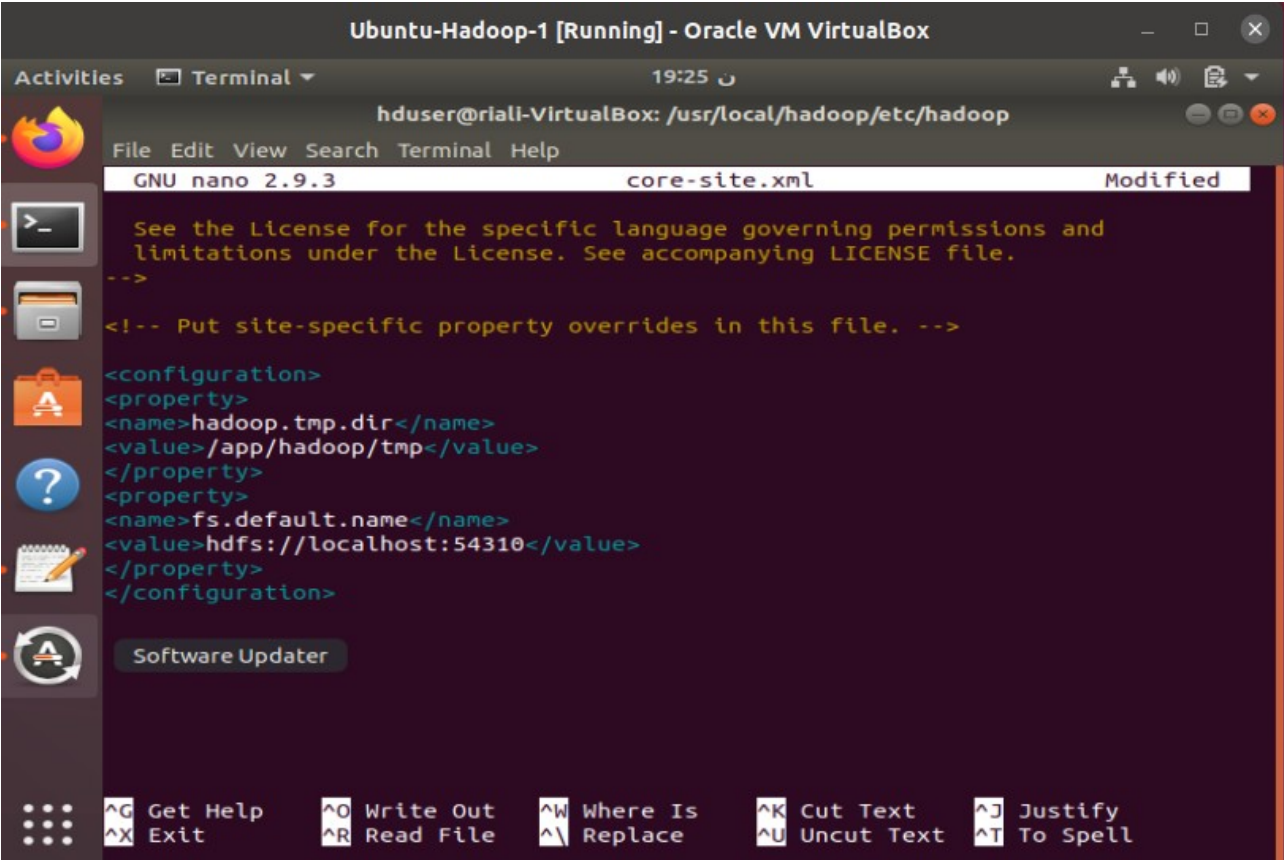


On crée le répertoire des fichiers temporaires de hadoop :

- `sudo mkdir -p /app/hadoop/tmp`
- `sudo chown hduser /app/hadoop/tmp`

### Modification des fichiers de configuration de Hadoop :

Nous allons maintenant passer au répertoire «/usr/local/hadoop» Commençons avec “core-site.xml” :



The screenshot shows a terminal window titled "Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox". The terminal is running the nano text editor, editing the file `core-site.xml` located at `/usr/local/hadoop/etc/hadoop/`. The editor's status bar indicates "GNU nano 2.9.3" and "core-site.xml Modified". The content of the file is as follows:

```
See the License for the specific language governing permissions and
limitations under the License. See accompanying LICENSE file.
-->

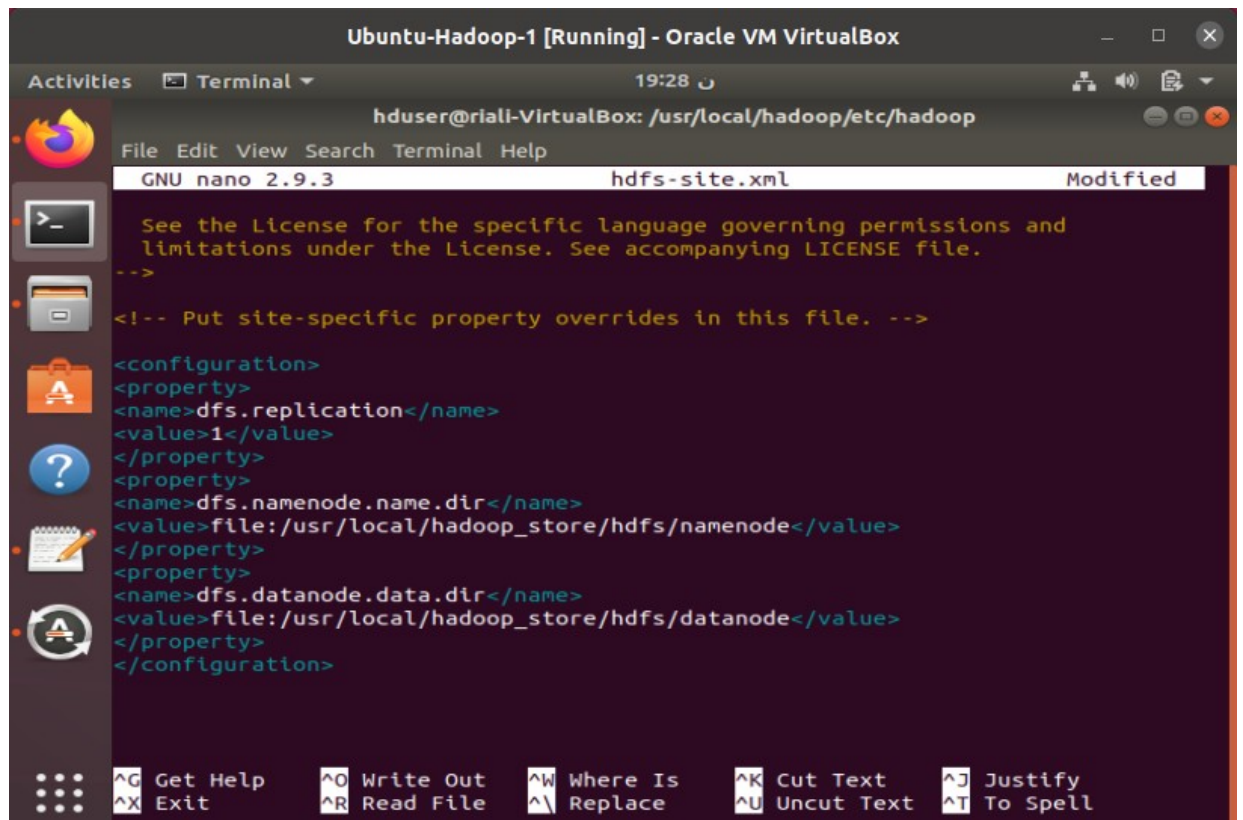
<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
<name>hadoop.tmp.dir</name>
<value>/app/hadoop/tmp</value>
</property>
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:54310</value>
</property>
</configuration>
```

At the bottom of the terminal, a "Software Updater" button is visible. The terminal also shows a menu bar with options like "File", "Edit", "View", "Search", "Terminal", and "Help", and a bottom status bar with keyboard shortcuts.

- **hadoop.tmp.dir** : C’est une base locale pour des répertoires temporaires.
- **fs.default.name** : Le nom du système de fichiers par défaut. Un URI dont le schéma et l'autorité déterminent l'implémentation FileSystem

Puis on va passer au “**hdfs-site.xml**” :



```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
19:28
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop
GNU nano 2.9.3 hdfs-site.xml Modified

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. -->

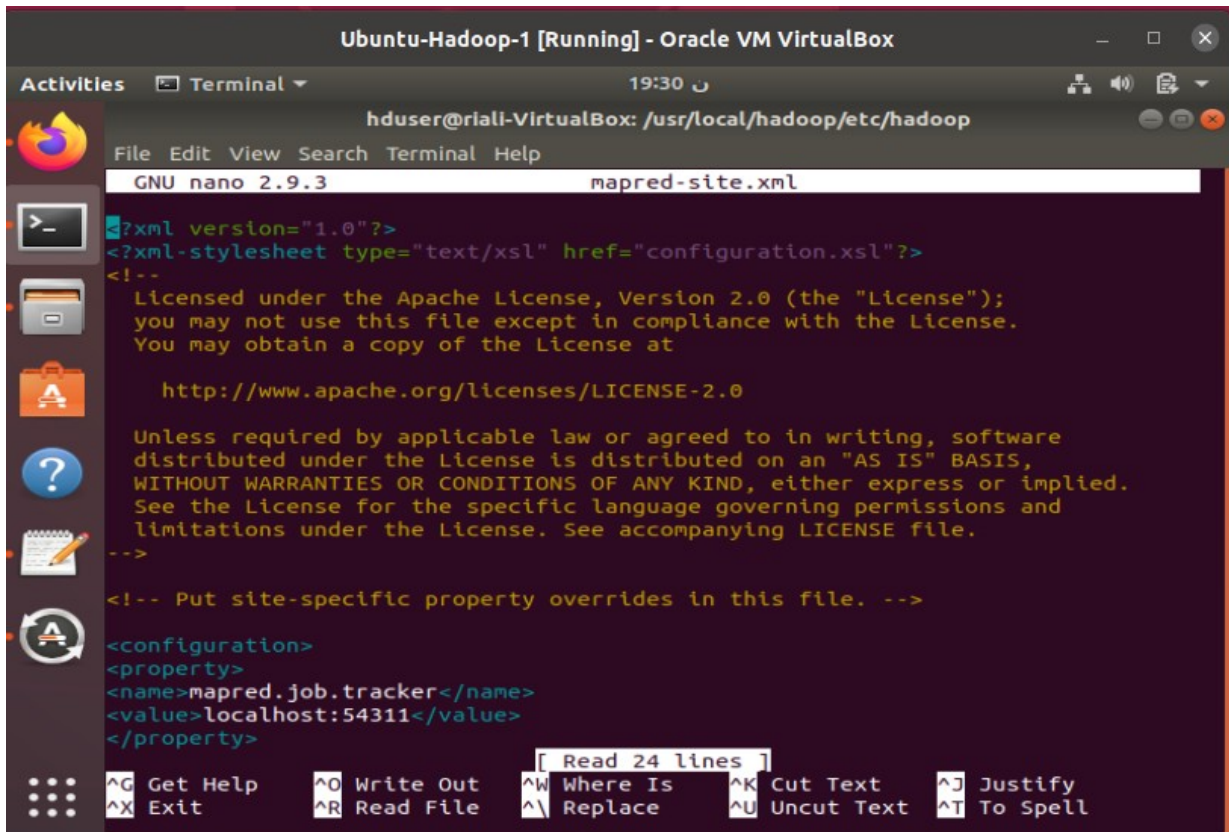
<configuration>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:/usr/local/hadoop_store/hdfs/namenode</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:/usr/local/hadoop_store/hdfs/datanode</value>
</property>
</configuration>

^G Get Help      ^O Write Out    ^W Where Is     ^K Cut Text     ^J Justify
^X Exit          ^R Read File    ^\ Replace      ^U Uncut Text   ^T To Spell
```

- **dfs.replication** : Lorsque nous stockons les fichiers dans HDFS, la structure hadoop divise le fichier en un ensemble de blocs (64 Mo ou 128 Mo), puis ces blocs seront répliqués sur les nœuds du cluster. La configuration dfs. la réplication consiste à spécifier le nombre de réplifications requises.
- **dfs.namenode.name.dir** : Détermine où sur le système de fichiers local le nœud de nom DFS doit stocker la table de noms (fsimage). S'il s'agit d'une liste de répertoires séparés par des virgules, la table de noms est répliquée dans tous les répertoires, pour la redondance.
- **Dfs.datanode.data.dir** : Détermine où sur le système de fichiers local un nœud de données DFS doit stocker ses

blocs. S'il s'agit d'une liste de répertoires séparés par des virgules, les données seront stockées dans tous les répertoires nommés, généralement sur différents appareils. Les répertoires qui n'existent pas sont ignorés.

→ **mapred-site.xml** :



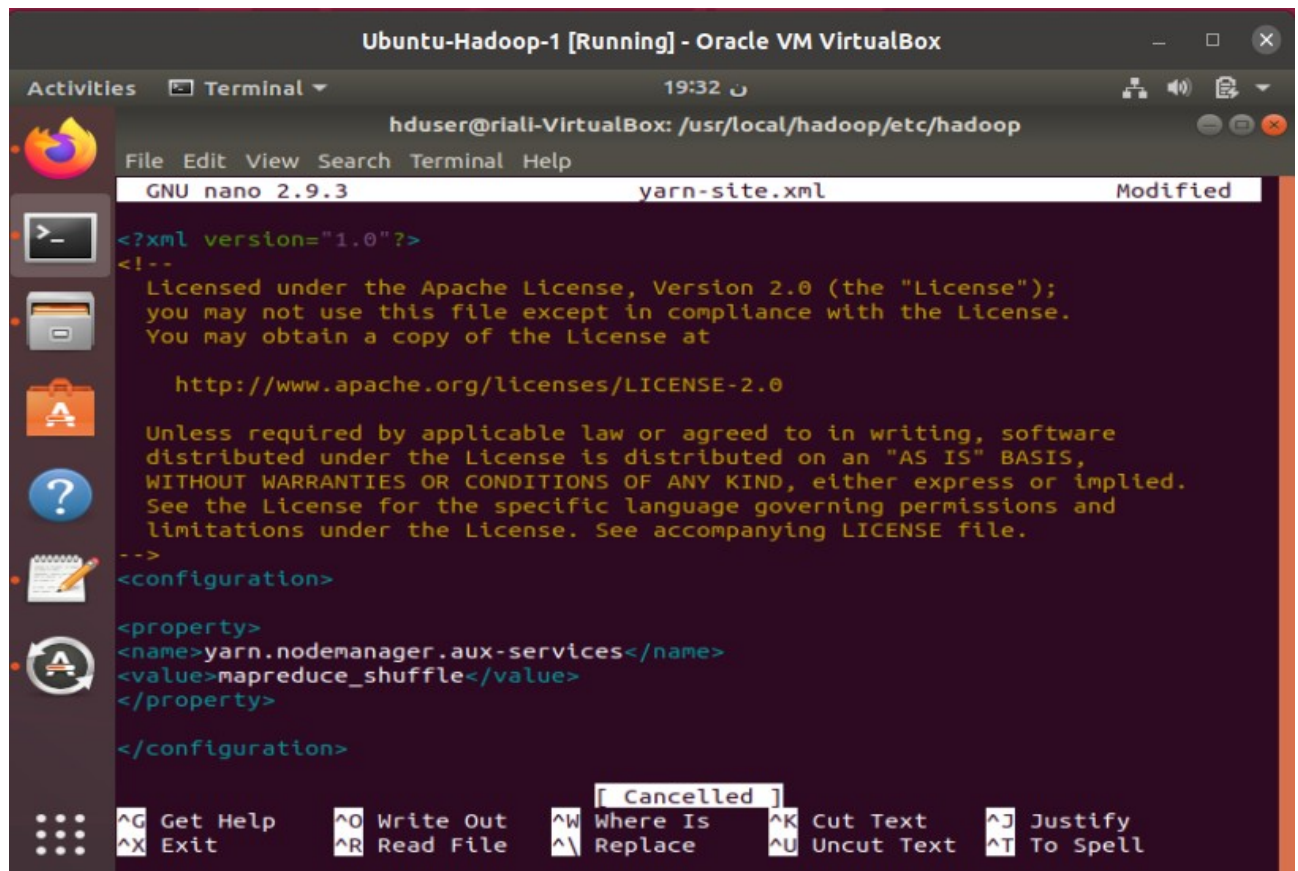
```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
hduser@riall-VirtualBox: /usr/local/hadoop/etc/hadoop
GNU nano 2.9.3 mapred-site.xml
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
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>mapred.job.tracker</name>
<value>localhost:54311</value>
</property>
[ Read 24 lines ]
^G Get Help      ^O Write Out    ^W Where Is     ^K Cut Text     ^J Justify
^X Exit          ^R Read File    ^\ Replace      ^U Uncut Text  ^T To Spell
```

- **mapred.job.tracker** : L'hôte et le port sur lesquels s'exécute le suivi des travaux MapReduce. Si "local", les travaux sont exécutés en cours de processus comme une seule carte et réduisent la tâche.

→ yarn-site.xml :



```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
19:32
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop
GNU nano 2.9.3 yarn-site.xml Modified

<?xml version="1.0"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

Unless required by applicable law or agreed to in writing, software
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.
-->
<configuration>

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

</configuration>

[Cancelled]
^G Get Help      ^O Write Out    ^W Where Is     ^K Cut Text     ^J Justify
^X Exit          ^R Read File    ^\ Replace      ^U Uncut Text  ^T To Spell
```

- **yarn.nodemanager.aux-services** : Le nom du service auxiliaire

## Formatage du Namenode :

Pour pouvoir lancer le service **Hadoop Distributed File System** , il faut d'abord formater le **Namenode** par la commande suivante : **"hdfs namenode -format"** comme il est clair au dessous :



```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
19:45
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop

hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.xml
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ sudo nano mapred-site.xml
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ sudo nano yarn-site.xml
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ hdfs namenode -format
WARNING: /usr/local/hadoop/logs does not exist. Creating.
/usr/local/hadoop/libexec/hadoop-functions.sh: line 1808: /opt/java/jdk1.8.0_71
//bin/java: No such file or directory
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ hdfs namenode -format
/usr/local/hadoop/libexec/hadoop-functions.sh: line 1808: /opt/java/jdk1.8.0_71
//bin/java: No such file or directory
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ sudo nano /usr/local/hado
op/libexec/hadoop-functions.sh
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ ls
capacity-scheduler.xml          kms-log4j.properties
configuration.xsl               kms-site.xml
container-executor.cfg          log4j.properties
core-site.xml                   mapred-env.cmd
hadoop-env.cmd                  mapred-env.sh
hadoop-env.sh                   mapred-queues.xml.template
hadoop-metrics2.properties      mapred-site.xml
hadoop-policy.xml               shellprofile.d
hadoop-user-functions.sh.example ssl-client.xml.example
hdfs-site.xml                   ssl-server.xml.example
https-env.sh                    user_ec_policies.xml.template
https-log4j.properties          workers
https-signature.secret          yarn-env.cmd
https-site.xml                  yarn-env.sh
kms-acls.xml                     yarnservice-log4j.properties
kms-env.sh                       yarn-site.xml
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop$ sudo nano hadoop-env.sh
```

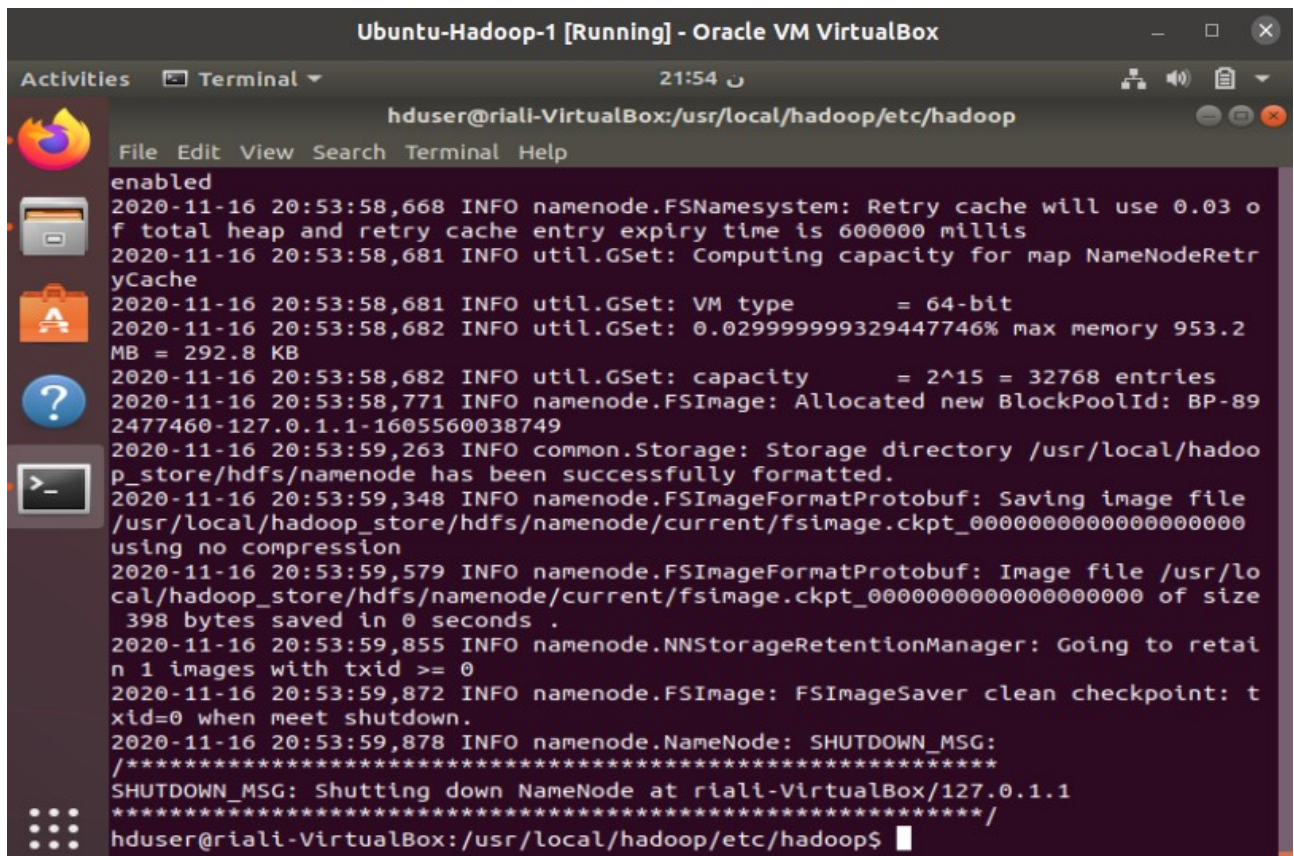
Malheureusement une erreur est survenue et voir le message d'erreur on peut bien estimer qu'il faut modifier le variable **JAVA\_HOME** dans le fichier **hadoop-env.sh** :

```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
19:46
hduser@riali-VirtualBox: /usr/local/hadoop/etc/hadoop
GNU nano 2.9.3 hadoop-env.sh Modified

#
# Therefore, the vast majority (BUT NOT ALL!) of these defaults
# are configured for substitution and not append.  If append
# is preferable, modify this file accordingly.
###
# Generic settings for HADOOP
###
# Technically, the only required environment variable is JAVA_HOME.
# All others are optional.  However, the defaults are probably not
# preferred.  Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d
# The java implementation to use.  By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/opt/java/jdk1.8.0_71/
# Location of Hadoop.  By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=
# Location of Hadoop's configuration information.  i.e., where this
# file is living.  If this is not defined, Hadoop will attempt to
# locate it based upon its execution path.
Cancelled
^G Get Help ^O Write Out ^W Where Is ^K Cut Text ^J Justify
^X Exit ^R Read File ^\ Replace ^U Uncut Text ^T To Linter
```

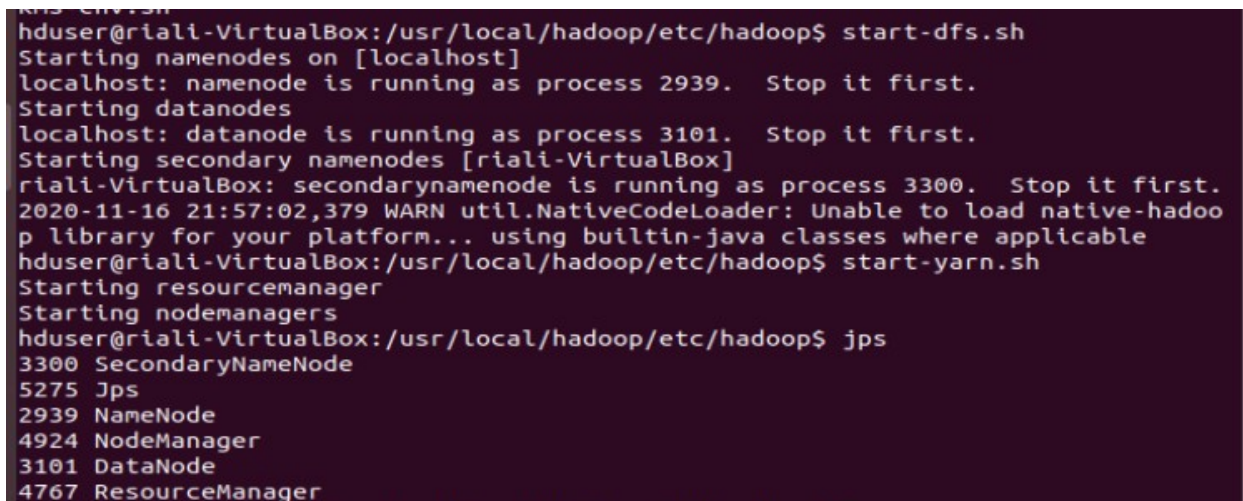


Après avoir changer le variable **JAVA\_HOME** on peut d'abord relancer notre commande du formatage du **Namenode** :



```
Ubuntu-Hadoop-1 [Running] - Oracle VM VirtualBox
Activities Terminal 21:54
hduser@riali-VirtualBox:/usr/local/hadoop/etc/hadoop
File Edit View Search Terminal Help
enabled
2020-11-16 20:53:58,668 INFO namenode.FSNamesystem: Retry cache will use 0.03 o
f total heap and retry cache entry expiry time is 600000 millis
2020-11-16 20:53:58,681 INFO util.GSet: Computing capacity for map NameNodeRetr
yCache
2020-11-16 20:53:58,681 INFO util.GSet: VM type          = 64-bit
2020-11-16 20:53:58,682 INFO util.GSet: 0.0299999999329447746% max memory 953.2
MB = 292.8 KB
2020-11-16 20:53:58,682 INFO util.GSet: capacity          = 2^15 = 32768 entries
2020-11-16 20:53:58,771 INFO namenode.FSImage: Allocated new BlockPoolId: BP-89
2477460-127.0.1.1-1605560038749
2020-11-16 20:53:59,263 INFO common.Storage: Storage directory /usr/local/hadoo
p_store/hdfs/namenode has been successfully formatted.
2020-11-16 20:53:59,348 INFO namenode.FSImageFormatProtobuf: Saving image file
/usr/local/hadoop_store/hdfs/namenode/current/fsimage.ckpt_000000000000000000
using no compression
2020-11-16 20:53:59,579 INFO namenode.FSImageFormatProtobuf: Image file /usr/lo
cal/hadoop_store/hdfs/namenode/current/fsimage.ckpt_000000000000000000 of size
398 bytes saved in 0 seconds .
2020-11-16 20:53:59,855 INFO namenode.NNStorageRetentionManager: Going to retai
n 1 images with txid >= 0
2020-11-16 20:53:59,872 INFO namenode.FSImage: FSImageSaver clean checkpoint: t
xid=0 when meet shutdown.
2020-11-16 20:53:59,878 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at riali-VirtualBox/127.0.1.1
*****/
hduser@riali-VirtualBox:/usr/local/hadoop/etc/hadoop$
```

- Ensuite, on va lancer les deux commandes suivantes :
- **start-dfs.sh** : Démarre les démons Hadoop DFS, le namenode et les datanodes
- **start-yarn.sh** : démarre le serveur MapReduce
- **jps** : Pour s'assurer que tout fonctionne, utiliser l'outil jps pour lister les processus Java en cours d'exécution



```
hduser@riali-VirtualBox:/usr/local/hadoop/etc/hadoop$ start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 2939. Stop it first.
Starting datanodes
localhost: datanode is running as process 3101. Stop it first.
Starting secondary namenodes [riali-VirtualBox]
riali-VirtualBox: secondarynamenode is running as process 3300. Stop it first.
2020-11-16 21:57:02,379 WARN util.NativeCodeLoader: Unable to load native-hadoo
p library for your platform... using builtin-java classes where applicable
hduser@riali-VirtualBox:/usr/local/hadoop/etc/hadoop$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hduser@riali-VirtualBox:/usr/local/hadoop/etc/hadoop$ jps
3300 SecondaryNameNode
5275 Jps
2939 NameNode
4924 NodeManager
3101 DataNode
4767 ResourceManager
```



## II. Exécution d'un programme Map/Reduce dans un cluster à nœud unique :

Accéder aux services de Hadoop via le navigateur :

localhost:8088/cluster 67%

### hadoop All Applications

**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
0	0	0	0	0	0 B

**Cluster Nodes Metrics**

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes
1	0	0	0

**Scheduler Metrics**

Scheduler Type	Scheduling Resource Type	Minimum Allocation
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	LaunchTime	FinishTime	State	FinalStatus	Running Containers
No data available in table											

Showing 0 to 0 of 0 entries

localhost:9870/dfshealth.html 80%

## Overview 'localhost:54310' (active)

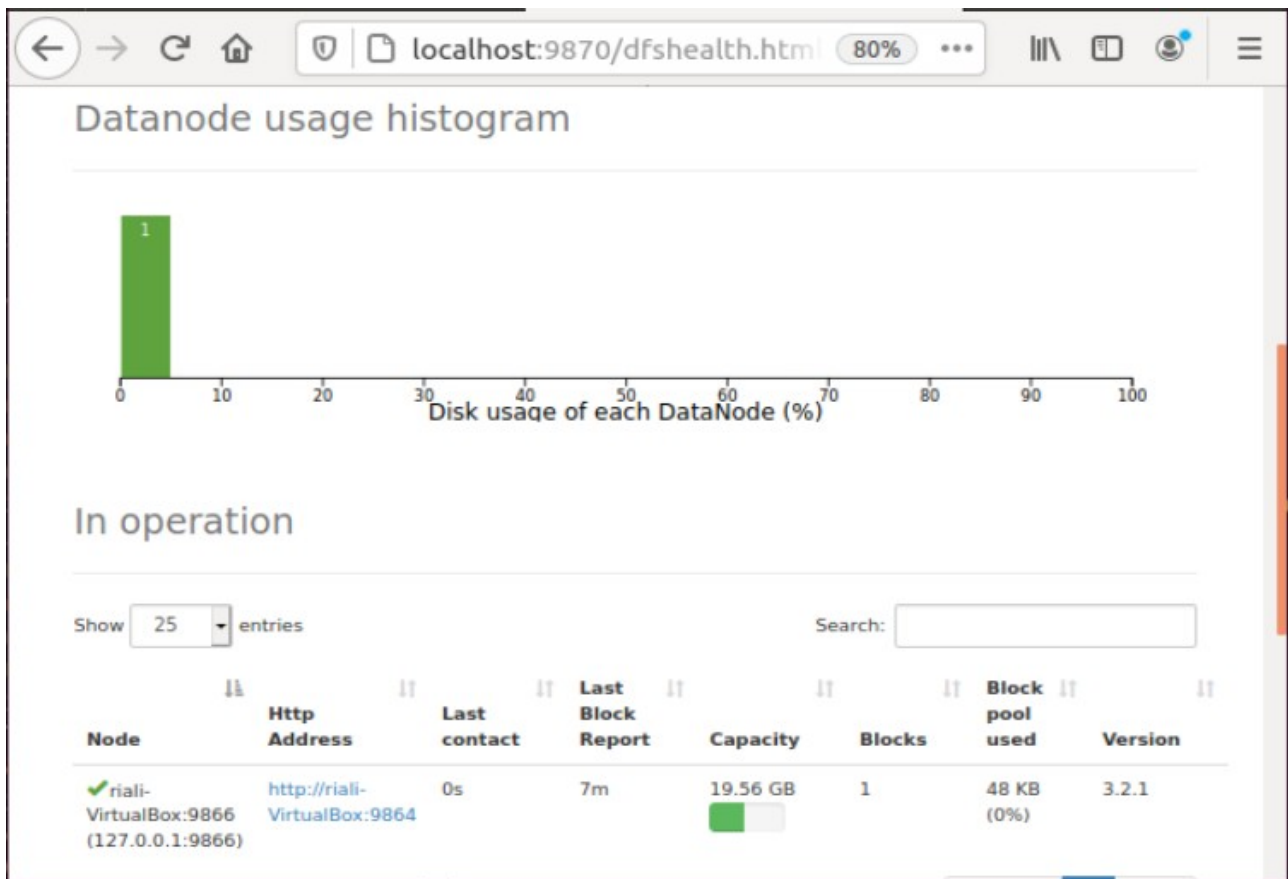
<b>Started:</b>	Tue Nov 17 12:00:05 +0100 2020
<b>Version:</b>	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
<b>Compiled:</b>	Tue Sep 10 16:56:00 +0100 2019 by rohithsharmaks from branch-3.2.1
<b>Cluster ID:</b>	CID-e24b372d-e17d-4461-939f-6495884d0aba
<b>Block Pool ID:</b>	BP-892477460-127.0.1.1-1605560038749

## Summary

Security is off.

Safemode is off.

2 files and directories, 1 blocks (1 replicated blocks, 0 erasure coded block groups) = 3 total filesystem object(s).



### Exécution d'un programme Map/Reduce :

Pour executer le programme **Map/Reduce**, on va executer les commandes suivantes :

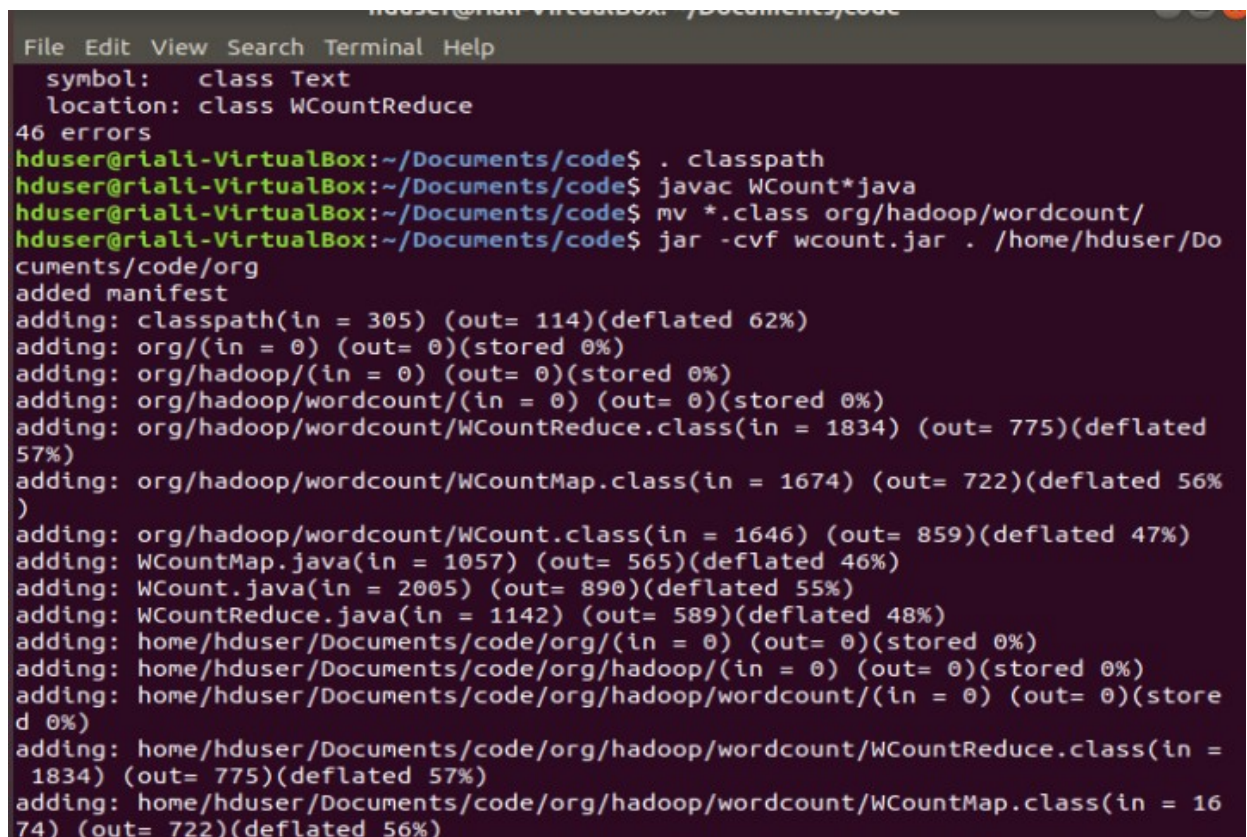
- **hdfs dfsadmin -report** : “pour tester le bon fonctionnement du service **hdfs**”

```
-----
Live datanodes (1):

Name: 127.0.0.1:9866 (localhost)
Hostname: riali-VirtualBox
Decommission Status : Normal
Configured Capacity: 21001486336 (19.56 GB)
DFS Used: 28672 (28 KB)
Non DFS Used: 9721094144 (9.05 GB)
DFS Remaining: 10189950976 (9.49 GB)
DFS Used%: 0.00%
DFS Remaining%: 48.52%
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: Mon Nov 16 23:04:40 WET 2020
Last Block Report: Mon Nov 16 23:02:13 WET 2020
Num of Blocks: 0

hduser@riali-VirtualBox:/usr/local/hadoop/etc/hadoop$
```

- `cd /home/hduser/Documents/code/` : accéder au repertoire contenant les fichier `wordcount` et le fichier `poeme.txt`
- `mkdir -p org/hadoop/wordcount/` : creer les repertoires `org` ET `hadoop` ET `wordcount`
- `. classpath` : executer le fichier “classpath”
- `javac WCount*.java` : executer tous les fichier {Wcount\*.java}
- `mv *.class org/hadoop/wordcount/`
- `jar -cvf wcount.jar . /home/hduser/Documents/code/org` : Generer le .jar



```

File Edit View Search Terminal Help
symbol: class Text
location: class WCountReduce
46 errors
hduser@riali-VirtualBox:~/Documents/code$ . classpath
hduser@riali-VirtualBox:~/Documents/code$ javac WCount*.java
hduser@riali-VirtualBox:~/Documents/code$ mv *.class org/hadoop/wordcount/
hduser@riali-VirtualBox:~/Documents/code$ jar -cvf wcount.jar . /home/hduser/Do
cuments/code/org
added manifest
adding: classpath(in = 305) (out= 114)(deflated 62%)
adding: org/(in = 0) (out= 0)(stored 0%)
adding: org/hadoop/(in = 0) (out= 0)(stored 0%)
adding: org/hadoop/wordcount/(in = 0) (out= 0)(stored 0%)
adding: org/hadoop/wordcount/WCountReduce.class(in = 1834) (out= 775)(deflated
57%)
adding: org/hadoop/wordcount/WCountMap.class(in = 1674) (out= 722)(deflated 56%
)
adding: org/hadoop/wordcount/WCount.class(in = 1646) (out= 859)(deflated 47%)
adding: WCountMap.java(in = 1057) (out= 565)(deflated 46%)
adding: WCount.java(in = 2005) (out= 890)(deflated 55%)
adding: WCountReduce.java(in = 1142) (out= 589)(deflated 48%)
adding: home/hduser/Documents/code/org/(in = 0) (out= 0)(stored 0%)
adding: home/hduser/Documents/code/org/hadoop/(in = 0) (out= 0)(stored 0%)
adding: home/hduser/Documents/code/org/hadoop/wordcount/(in = 0) (out= 0)(store
d 0%)
adding: home/hduser/Documents/code/org/hadoop/wordcount/WCountReduce.class(in =
1834) (out= 775)(deflated 57%)
adding: home/hduser/Documents/code/org/hadoop/wordcount/WCountMap.class(in = 16
74) (out= 722)(deflated 56%)

```

- `cd /usr/local/hadoop/` : accéder au repertoire de hadoop
- `bin/hdfs dfs -put /home/hduser/Documents/code/poeme.txt /` : Copier “poeme.txt” de fichiers local vers le système de fichiers de destination. Lit également l'entrée de stdin et écrit dans le système de fichiers de destination.



- `bin/hdfs dfs -ls /` : Pour un répertoire comme notre cas, il renvoie la liste de ses enfants directs comme dans Unix
- `cd /home/hduser/Documents/code/`
- `hadoop jar wcount.jar org.hadoop.wordcount.WCount /poeme.txt /results`

```
hduser@riali-VirtualBox:~/Documents/code$ cd /usr/local/hadoop/
hduser@riali-VirtualBox:~/Documents/code$ bin/hdfs dfs -put /home/hduser/Documents/code/poeme.txt /
2020-11-17 11:39:54,447 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: '/home/hduser/Documents/code/poeme.txt': No such file or directory
hduser@riali-VirtualBox:~/Documents/code$ bin/hdfs dfs -put /home/hduser/Documents/code/poeme.txt /
2020-11-17 11:40:51,377 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
put: '/poeme.txt': File exists
hduser@riali-VirtualBox:~/Documents/code$ bin/hdfs dfs -ls /
2020-11-17 11:41:23,131 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r-- 1 hduser supergroup 1670 2020-11-17 00:11 /poeme.txt
hduser@riali-VirtualBox:~/Documents/code$ cd /home/hduser/Documents/code/
hduser@riali-VirtualBox:~/Documents/code$ hadoop jar wcount.jar org.hadoop.wordcount.WCount /poeme.txt /results
2020-11-17 11:42:48,083 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2020-11-17 11:42:49,654 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2020-11-17 11:42:50,836 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2020-11-17 11:42:50,836 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2020-11-17 11:43:53,112 INFO input.FileInputFormat: Total input files to process
```

- `hadoop fs -ls /results`
- `hadoop fs -cat /results/part-r-00000`

```
hduser@riali-VirtualBox:~/Documents/code$ hadoop fs -ls /results
2020-11-17 11:43:38,675 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 hduser supergroup 0 2020-11-17 11:42 /results/_SUCCESS
-rw-r--r-- 1 hduser supergroup 2823 2020-11-17 11:42 /results/part-r-00000
hduser@riali-VirtualBox:~/Documents/code$ hadoop fs -cat /results/part-r-00000
2020-11-17 11:44:24,582 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2020-11-17 11:44:26,078 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localhostTrusted = false, remoteHostTrusted = false
a 6 occurrences.
adorant 1 occurrences.
ailes 1 occurrences.
aima 1 occurrences.
amour 1 occurrences.
au 11 occurrences.
bas 1 occurrences.
belle 1 occurrences.
bles 1 occurrences.
bras 1 occurrences.
bretagne 1 occurrences.
brula 1 occurrences.
celle 1 occurrences.
celui 20 occurrences.
cette 1 occurrences.
chancelle 1 occurrences.
```

