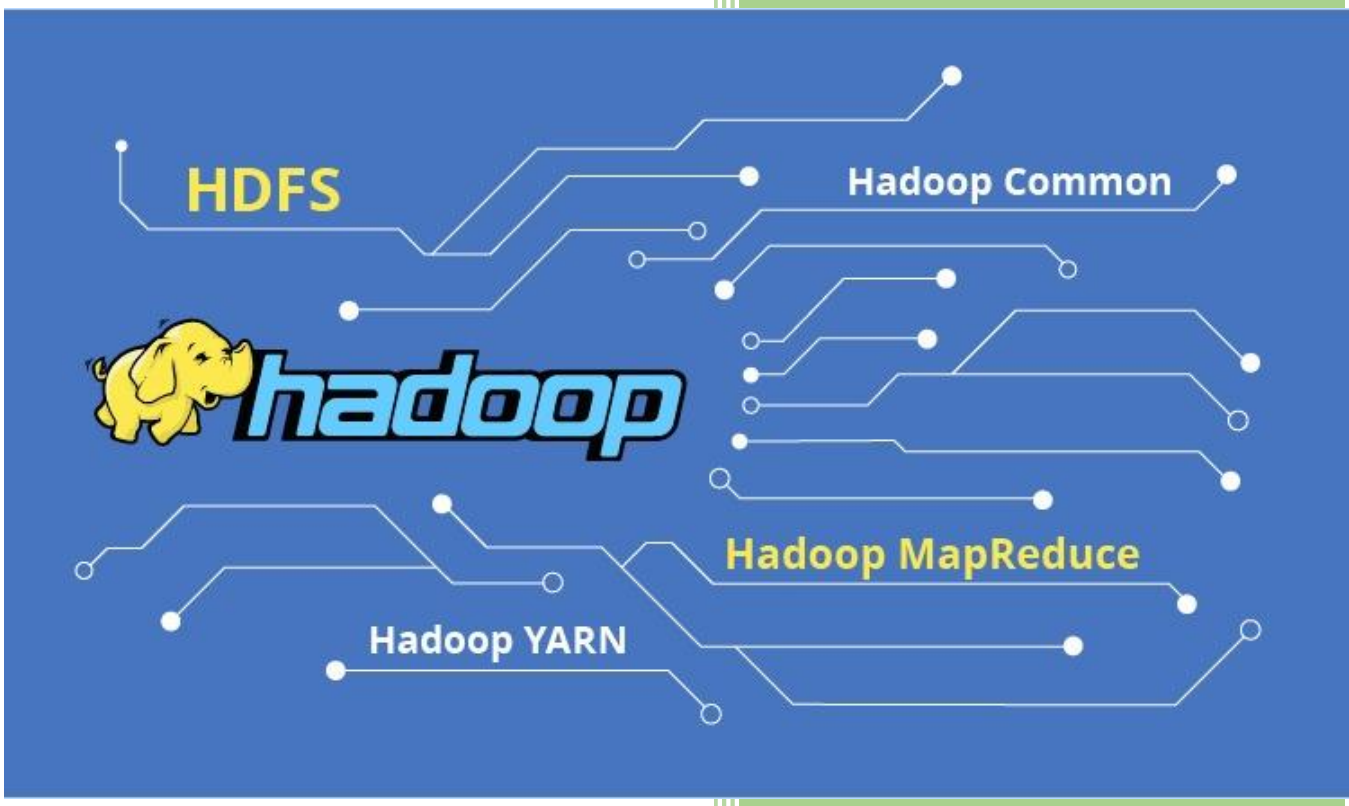
Hadoop with Hive & Nifi & Hbase in distributed mode & Tableau for Data Visualisation



Encadré par :

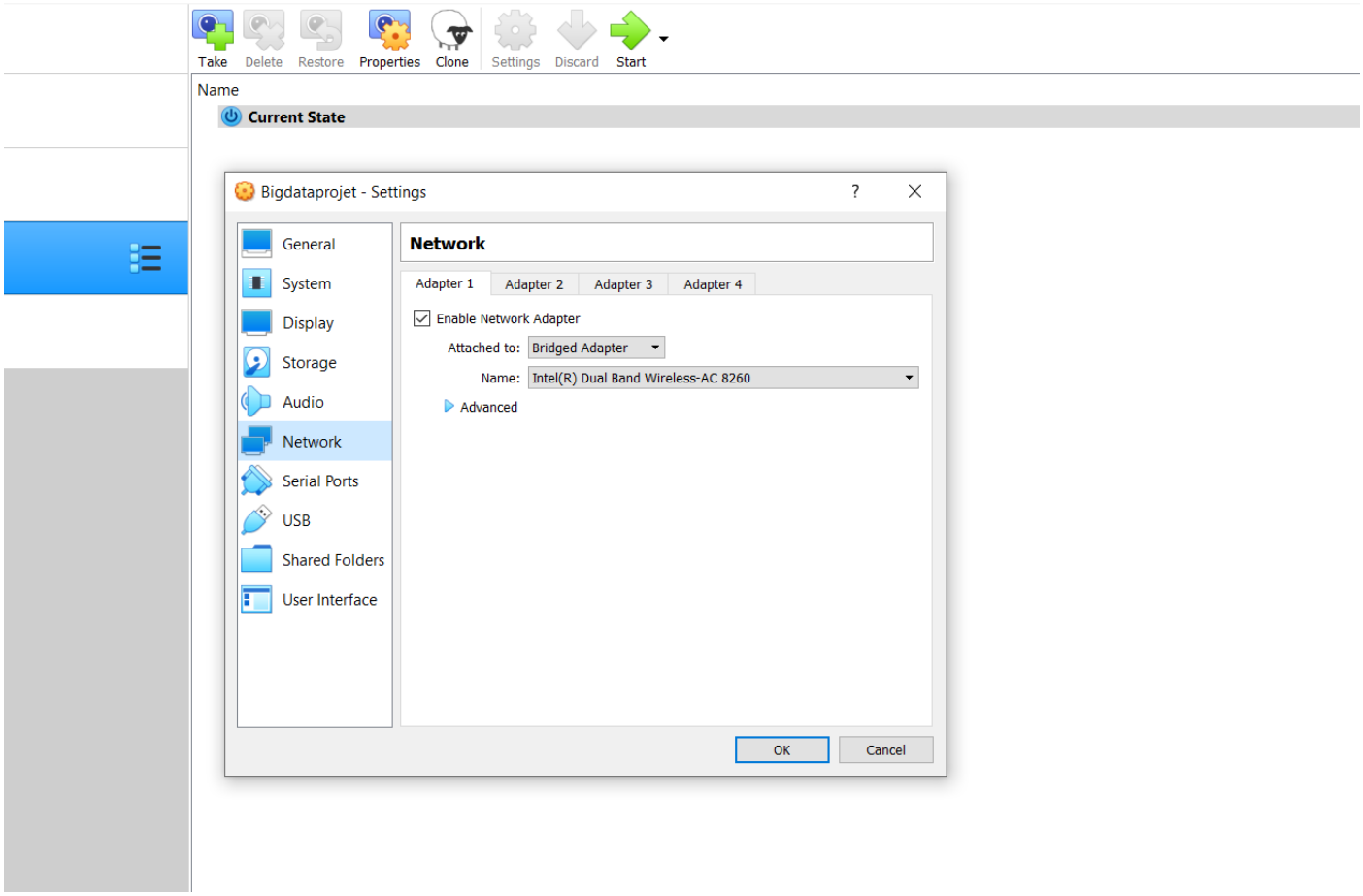
Prof. GHAZOUANI Mohamed

Réalisé par :

- **LAGHZAOUI Hind**
- **BERRADA Anas**
- **LAMHOUR Mohamed Akram**
- **MSALEK Mohamed**
- **KARAFI Youssef**

Setup Network:

YOU SHOULD CONFIGURE NETWORK AT BRIDGED ADAPTER

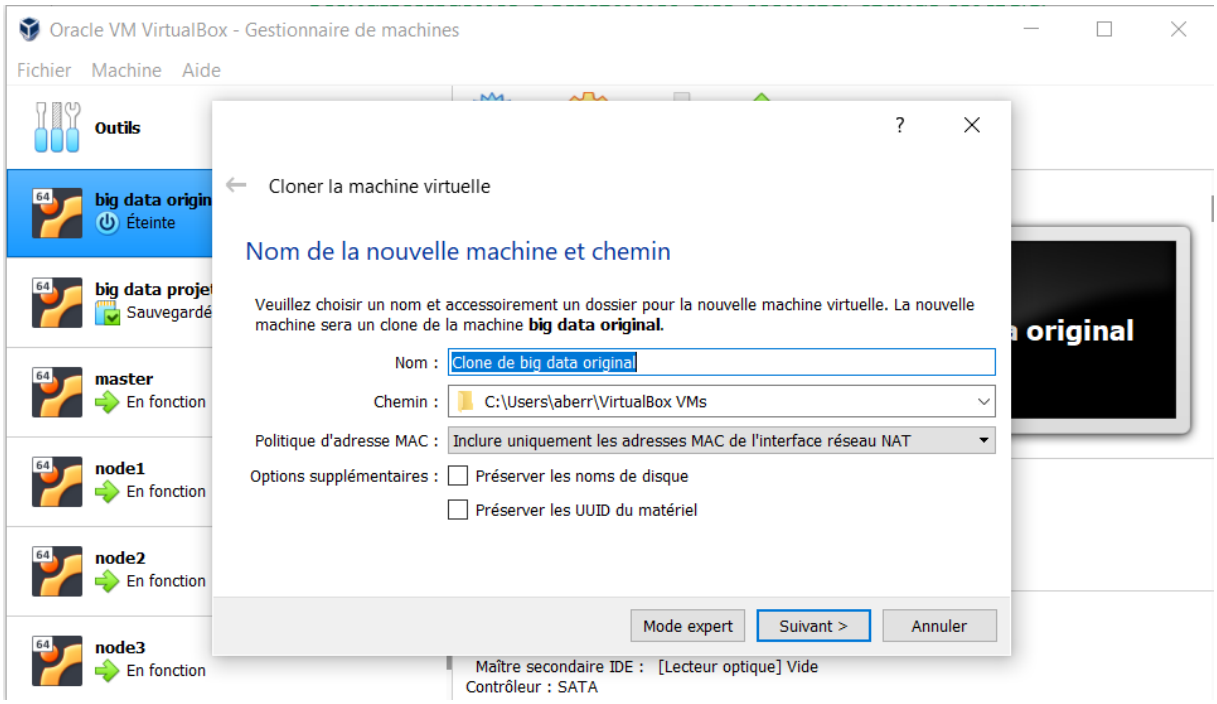


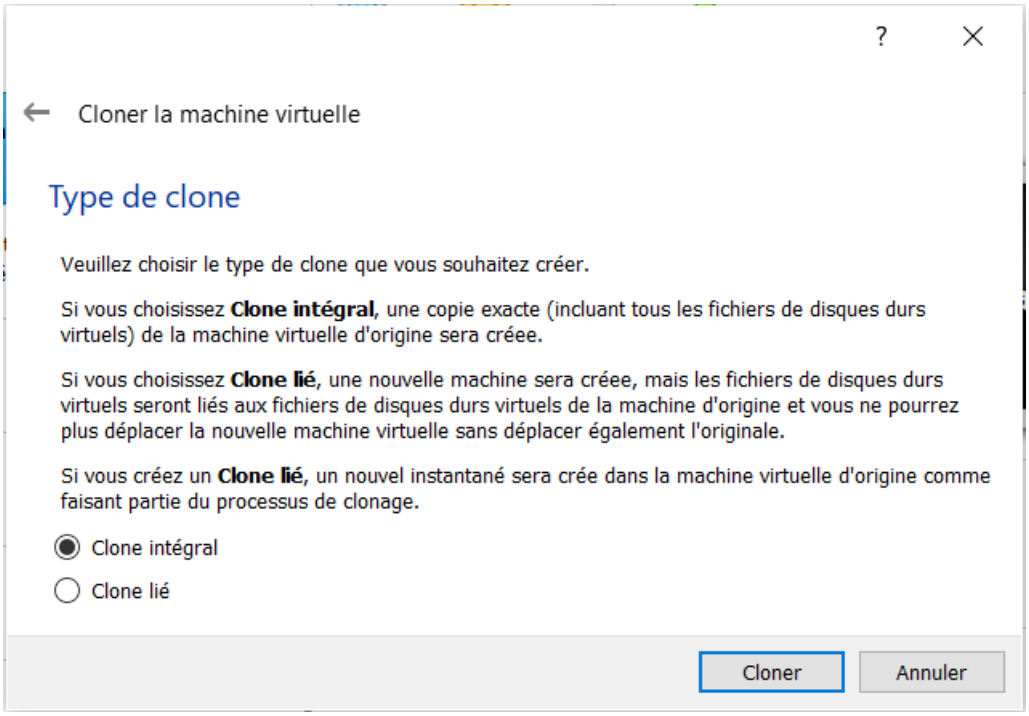
Multi node-cluster setup

WE WILL CREATE A 4-NODE CLUSTER SYSTEM (1MASTER, 3-SLAVE)

1. Create 4 Nodes

Create a master node then clone it using :





2. Install Vim editor

```
anas@master:~$ sudo apt install vim
[sudo] password for anas:
Reading package lists... Done
Building dependency tree
Reading state information... Done
vim is already the newest version (2:7.4.1689-3ubuntu1.5).
The following packages were automatically installed and are no longer required:
  linux-headers-4.15.0-45 linux-headers-4.15.0-45-generic
  linux-image-4.15.0-45-generic linux-modules-4.15.0-45-generic
  linux-modules-extra-4.15.0-45-generic snapd-login-service
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 0 not upgraded.
```

3. Change the hostname of all 4 systems

For each VM you have to specify the hostname accordingly

- Master

```
anas@master:~$ sudo vim /etc/hostname
```

```
master
~
~
~
~
```

- Node1

```
node1
~
~
~
~
```

- Node2

```
node2
~
~
~
~
```

- Node3

```
node3
~
~
~
~
```

Press **i** on the keyboard and write 'master' by deleting Ubuntu.

Press ESC on the keyboard

Save the configuration by **:wq**

4. Check the ip addresses for each node

Find the ip Address of all 3 systems and try to ping each other

- Master 192.168.0.153

```
anas@master:~$ ifconfig
enp0s3    Link encap:Ethernet  HWaddr 08:00:27:92:39:ac
          inet addr:192.168.0.153  Bcast:192.168.0.255  Mask:255.255.255.0
          inet6 addr: fe80::fd3b:2a6e:a5e6:e3f4/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:138164 errors:0 dropped:0 overruns:0 frame:0
          TX packets:23450 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:181836027 (181.8 MB)  TX bytes:2097737 (2.0 MB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:78 errors:0 dropped:0 overruns:0 frame:0
          TX packets:78 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:7542 (7.5 KB)  TX bytes:7542 (7.5 KB)
```

- Node1 192.168.0.164

```
anas@node1:~$ ifconfig
enp0s3    Link encap:Ethernet  HWaddr 08:00:27:bc:85:0f
          inet addr:192.168.0.164  Bcast:192.168.0.255  Mask:255.255.255.0
          inet6 addr: fe80::a9ec:765:a05c:6499/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:156360 errors:0 dropped:0 overruns:0 frame:0
          TX packets:29577 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:208834733 (208.8 MB)  TX bytes:2621391 (2.6 MB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:156 errors:0 dropped:0 overruns:0 frame:0
          TX packets:156 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:14635 (14.6 KB)  TX bytes:14635 (14.6 KB)
```

- Node2 192.168.0.181

```
anas@node2:~$ ifconfig
enp0s3    Link encap:Ethernet  HWaddr 08:00:27:b2:10:48
          inet addr:192.168.0.181  Bcast:192.168.0.255  Mask:255.255.255.0
          inet6 addr: fe80::d07f:46a9:56fc:23ce/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:23998 errors:0 dropped:0 overruns:0 frame:0
          TX packets:1247 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:10293800 (10.2 MB)  TX bytes:108734 (108.7 KB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:144 errors:0 dropped:0 overruns:0 frame:0
          TX packets:144 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:13457 (13.4 KB)  TX bytes:13457 (13.4 KB)
```

- Node3 192.168.0.154

```
anas@node3:~$ ifconfig
enp0s3    Link encap:Ethernet  HWaddr 08:00:27:20:ee:1b
          inet addr:192.168.0.154  Bcast:192.168.0.255  Mask:255.255.255.0
          inet6 addr: fe80::d1aa:afbf:61b2:a875/64 Scope:Link
          UP BROADCAST RUNNING MULTICAST  MTU:1500  Metric:1
          RX packets:8676 errors:0 dropped:0 overruns:0 frame:0
          TX packets:1462 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:11939110 (11.9 MB)  TX bytes:131432 (131.4 KB)

lo        Link encap:Local Loopback
          inet addr:127.0.0.1  Mask:255.0.0.0
          inet6 addr: ::1/128 Scope:Host
          UP LOOPBACK RUNNING  MTU:65536  Metric:1
          RX packets:114 errors:0 dropped:0 overruns:0 frame:0
          TX packets:114 errors:0 dropped:0 overruns:0 carrier:0
          collisions:0 txqueuelen:1000
          RX bytes:10153 (10.1 KB)  TX bytes:10153 (10.1 KB)
```

5. Update the hosts on all 4 nodes

Change the hosts file in /etc/hosts for Master and all the slaves(1-3):

```
anas@master:~$ sudo vim /etc/hosts
```

```
127.0.0.1       localhost
192.168.0.153   master
192.168.0.164   node1
192.168.0.181   node2
192.168.0.154   node3

# 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
~
```

6. Restart all the VM in order to reflect the changes.

```
anas@master:~$ reboot
```

7. Ping Each other using Hostname

- Master

```
anas@master:~$ ping node1
PING node1 (192.168.0.164) 56(84) bytes of data.
64 bytes from node1 (192.168.0.164): icmp_seq=1 ttl=64 time=2.68 ms
64 bytes from node1 (192.168.0.164): icmp_seq=2 ttl=64 time=1.72 ms
64 bytes from node1 (192.168.0.164): icmp_seq=3 ttl=64 time=1.95 ms
^Z
[1]+  Stopped                 ping node1
anas@master:~$ ping node2
PING node2 (192.168.0.181) 56(84) bytes of data.
64 bytes from node2 (192.168.0.181): icmp_seq=1 ttl=64 time=1.16 ms
64 bytes from node2 (192.168.0.181): icmp_seq=2 ttl=64 time=0.714 ms
64 bytes from node2 (192.168.0.181): icmp_seq=3 ttl=64 time=1.25 ms
64 bytes from node2 (192.168.0.181): icmp_seq=4 ttl=64 time=0.449 ms
^Z
[2]+  Stopped                 ping node2
anas@master:~$ ping node3
PING node3 (192.168.0.154) 56(84) bytes of data.
64 bytes from node3 (192.168.0.154): icmp_seq=1 ttl=64 time=1.28 ms
64 bytes from node3 (192.168.0.154): icmp_seq=2 ttl=64 time=1.01 ms
64 bytes from node3 (192.168.0.154): icmp_seq=3 ttl=64 time=0.777 ms
64 bytes from node3 (192.168.0.154): icmp_seq=4 ttl=64 time=0.954 ms
^Z
[3]+  Stopped                 ping node3
```

- Node1

```
çanas@node1:~$ ping master
PING master (192.168.0.153) 56(84) bytes of data.
64 bytes from master (192.168.0.153): icmp_seq=1 ttl=64 time=0.824 ms
64 bytes from master (192.168.0.153): icmp_seq=2 ttl=64 time=0.596 ms
64 bytes from master (192.168.0.153): icmp_seq=3 ttl=64 time=0.666 ms
64 bytes from master (192.168.0.153): icmp_seq=4 ttl=64 time=0.795 ms
64 bytes from master (192.168.0.153): icmp_seq=5 ttl=64 time=0.682 ms
64 bytes from master (192.168.0.153): icmp_seq=6 ttl=64 time=0.703 ms
^Z
[1]+  Stopped                 ping master
anas@node1:~$ ping node2
PING node2 (192.168.0.181) 56(84) bytes of data.
64 bytes from node2 (192.168.0.181): icmp_seq=1 ttl=64 time=1.27 ms
64 bytes from node2 (192.168.0.181): icmp_seq=2 ttl=64 time=0.646 ms
64 bytes from node2 (192.168.0.181): icmp_seq=3 ttl=64 time=0.735 ms
64 bytes from node2 (192.168.0.181): icmp_seq=4 ttl=64 time=0.642 ms
64 bytes from node2 (192.168.0.181): icmp_seq=5 ttl=64 time=0.817 ms
^C
--- node2 ping statistics ---
5 packets transmitted, 5 received, 0% packet loss, time 4051ms
rtt min/avg/max/mdev = 0.642/0.823/1.277/0.237 ms
anas@node1:~$ ping node3
PING node3 (192.168.0.154) 56(84) bytes of data.
64 bytes from node3 (192.168.0.154): icmp_seq=1 ttl=64 time=1.08 ms
64 bytes from node3 (192.168.0.154): icmp_seq=2 ttl=64 time=0.806 ms
64 bytes from node3 (192.168.0.154): icmp_seq=3 ttl=64 time=0.672 ms
64 bytes from node3 (192.168.0.154): icmp_seq=4 ttl=64 time=0.632 ms
^Z
[2]+  Stopped                 ping node3
```

- Node2

```
anas@node2:~$ ping master
PING master (192.168.0.153) 56(84) bytes of data.
64 bytes from master (192.168.0.153): icmp_seq=1 ttl=64 time=0.647 ms
64 bytes from master (192.168.0.153): icmp_seq=2 ttl=64 time=0.865 ms
64 bytes from master (192.168.0.153): icmp_seq=3 ttl=64 time=0.723 ms
^C
--- master ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2015ms
rtt min/avg/max/mdev = 0.647/0.745/0.865/0.090 ms
anas@node2:~$ ping node1
PING node1 (192.168.0.164) 56(84) bytes of data.
64 bytes from node1 (192.168.0.164): icmp_seq=1 ttl=64 time=0.800 ms
64 bytes from node1 (192.168.0.164): icmp_seq=2 ttl=64 time=0.811 ms
64 bytes from node1 (192.168.0.164): icmp_seq=3 ttl=64 time=0.786 ms
^C
--- node1 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2057ms
rtt min/avg/max/mdev = 0.786/0.799/0.811/0.010 ms
anas@node2:~$ ping node3
PING node3 (192.168.0.154) 56(84) bytes of data.
64 bytes from node3 (192.168.0.154): icmp_seq=1 ttl=64 time=1.26 ms
64 bytes from node3 (192.168.0.154): icmp_seq=2 ttl=64 time=0.912 ms
64 bytes from node3 (192.168.0.154): icmp_seq=3 ttl=64 time=0.656 ms
64 bytes from node3 (192.168.0.154): icmp_seq=4 ttl=64 time=0.861 ms
^C
--- node3 ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3010ms
rtt min/avg/max/mdev = 0.656/0.922/1.262/0.221 ms
anas@node2:~$
```

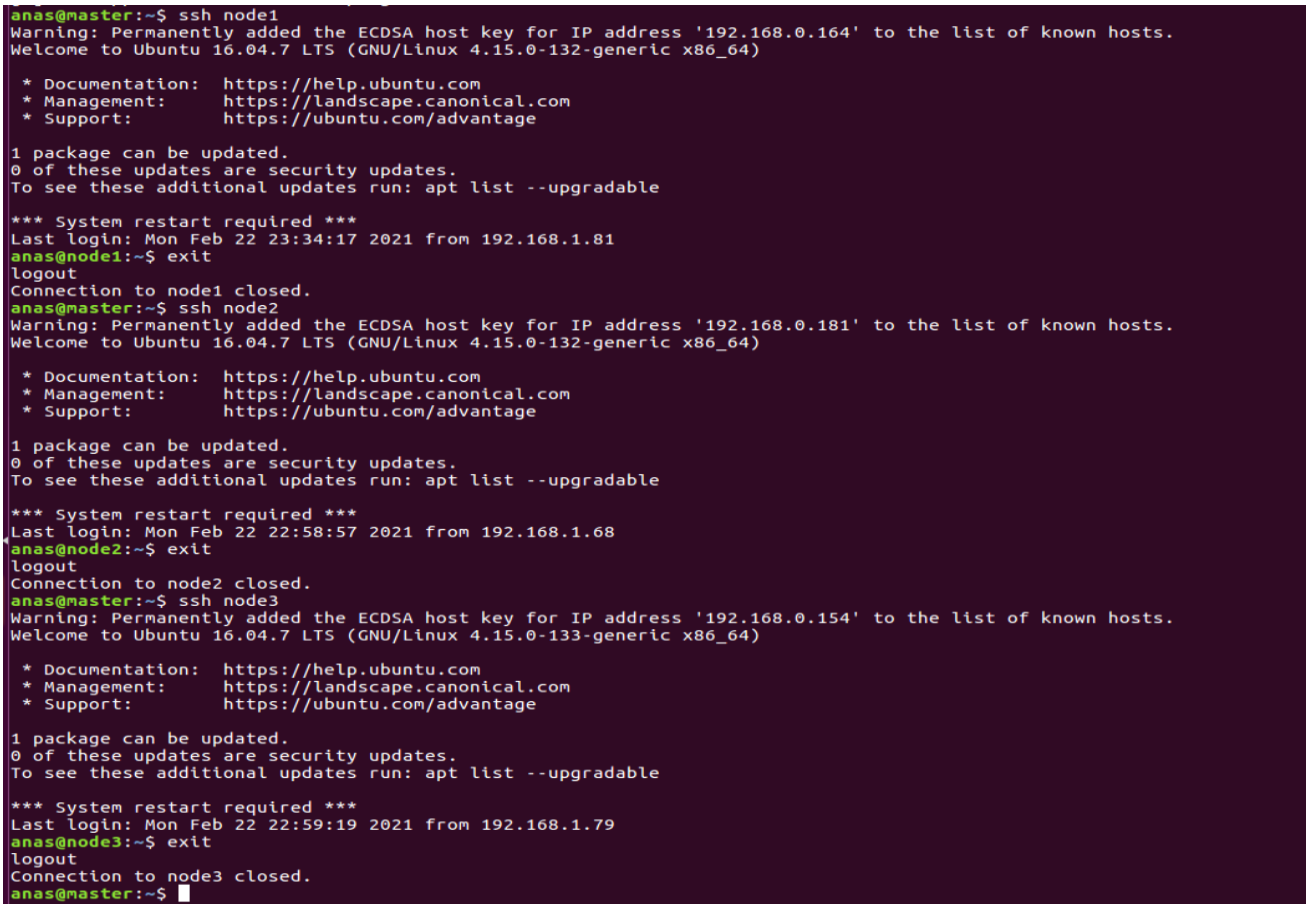
- Node3

```
anas@node3:~$ ping master
PING master (192.168.0.153) 56(84) bytes of data.
64 bytes from master (192.168.0.153): icmp_seq=1 ttl=64 time=0.834 ms
64 bytes from master (192.168.0.153): icmp_seq=2 ttl=64 time=0.766 ms
64 bytes from master (192.168.0.153): icmp_seq=3 ttl=64 time=0.813 ms
^C
--- master ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2007ms
rtt min/avg/max/mdev = 0.766/0.804/0.834/0.036 ms
anas@node3:~$ ping node2
PING node2 (192.168.0.181) 56(84) bytes of data.
64 bytes from node2 (192.168.0.181): icmp_seq=1 ttl=64 time=0.985 ms
64 bytes from node2 (192.168.0.181): icmp_seq=2 ttl=64 time=0.778 ms
64 bytes from node2 (192.168.0.181): icmp_seq=3 ttl=64 time=0.704 ms
^C
--- node2 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2008ms
rtt min/avg/max/mdev = 0.704/0.822/0.985/0.121 ms
anas@node3:~$ ping node1
PING node1 (192.168.0.164) 56(84) bytes of data.
64 bytes from node1 (192.168.0.164): icmp_seq=1 ttl=64 time=0.932 ms
64 bytes from node1 (192.168.0.164): icmp_seq=2 ttl=64 time=0.712 ms
64 bytes from node1 (192.168.0.164): icmp_seq=3 ttl=64 time=0.832 ms
^C
--- node1 ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2008ms
rtt min/avg/max/mdev = 0.712/0.825/0.932/0.092 ms
anas@node3:~$
```

8. Test SSH connectivity

Test the SSH connectivity by doing the following. It will ask for yes or no and you should type 'yes'. Perform SSH master/node1/node2/node3 on each of the node to verify the connectivity.

- Master



- Node1/Node2/Node3

Repeat this step for each node to verify the connectivity

It will ask for yes or no and you should type 'yes' We should be able to SSH master and SSH nodes without password prompt. If it asks for password while connecting to master or slave using SSH, there is something went wrong and you need to fix it before proceeding further.

9. Update core-site.xml(Master+ All Nodes)

- The changes you need to make to core-site.xml
  Change localhost to master

```
<?xml version="1.0" encoding="UTF-8"?>
<?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>fs.defaultFS</name>
<value>hdfs://master:9000</value>
</property>
</configuration>
```

10. Update hdfs-site.xml(Master + All Nodes)

- The changes you need to make to hdfs-site.xml
  a. Replication is set to 3
  b. Namenode configured only in master
  c. Datanode configured only in nodes

- Master only

```
anas@master:~/data/hadoop-2.7.3/etc/hadoop$ sudo vim hdfs-site.xml
```

```
<?xml version="1.0" encoding="UTF-8"?>
<?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>dfs.replication</name>
<value>3</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>/home/anas/data/hadoop_temp/hdfs/namenode</value>
</property>
</configuration>
```

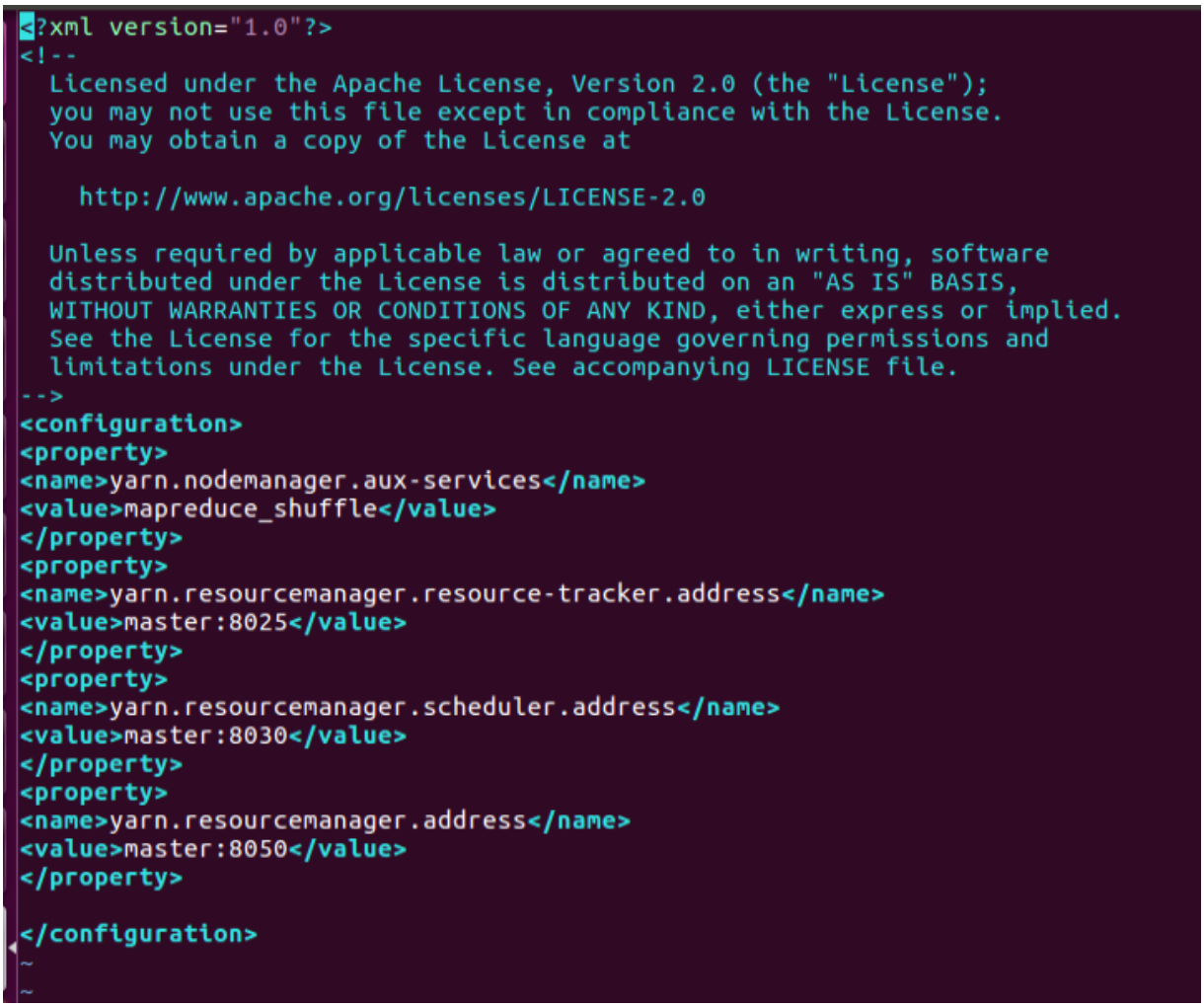
- Node1/Node2/Node3 only

```
anas@node1:~$  cd $HADOOP_HOME/etc/hadoop
anas@node1:~/data/hadoop-2.7.3/etc/hadoop$ sudo vim hdfs-site.xml
```

```
<?xml version="1.0" encoding="UTF-8"?>
<?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>dfs.replication</name>
<value>3</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/anas/data/hadoop_tmp/hdfs/datanode</value></property>
</configuration>
~
~
```

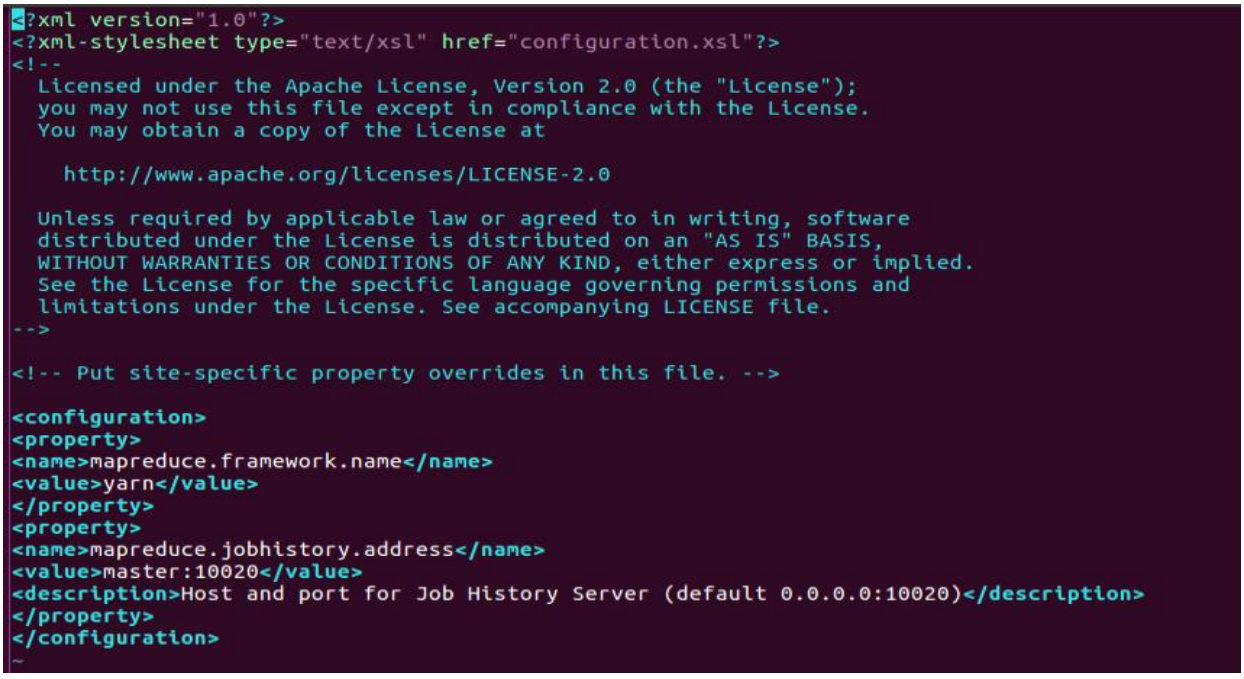
11. Update yarn-site.xml(Master + All Nodes)

```
anas@master:~/data/hadoop-2.7.3/etc/hadoop$ sudo vim yarn-site.xml
```



12. Update mapred-site.xml(Master + All Nodes)

```
anas@master:~/data/hadoop-2.7.3/etc/hadoop$ sudo vim mapred-site.xml
```



13. Update Master only Configuration

Update Masters and slaves file(Master Node only) If you see any entry related to localhost, feel free to delete it. This file is just helper file that are used by hadoop scripts to start appropriate services on master and nodes.

```
anas@master:~/data/hadoop-2.7.3/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-metrics.properties   mapred-site.xml.template
hadoop-policy.xml           masters
hdfs-site.xml               slaves
httpfs-env.sh               ssl-client.xml.example
httpfs-log4j.properties     ssl-server.xml.example
httpfs-signature.secret     yarn-env.cmd
httpfs-site.xml             yarn-env.sh
kms-acls.xml                yarn-site.xml
kms-env.sh
```

- Edit Slaves file

```
anas@master:~/data/hadoop-2.7.3/etc/hadoop$ vim slaves
node1
node2
node3
~
~
```

- Below masters file does not exists by default. It gets created the files

```
anas@master:~/data/hadoop-2.7.3/etc/hadoop$ vim masters
master
~
~
```

In Hadoop 3:

In this part you will not find the files: "Salves" and "Masters" but you will find a file named: "Workers" then in this file you should edit it like:

node1
node2
node3

And master is by default

Note: You don't need to configure them in slave nodes

14. Recreate Namenode folder(Master Only)

a. **First you have to delete the older hdfs file.**

```
sudo rm -rf /home/anas/Desktop/hdfs
```

b. **Then create another file that contains hdfs file like:**

```
sudo mkdir -p /home/anas/data/hadoop_tmp/hdfs/namenode
```

c. **Then type these commands:**

```
sudo chown anas:hadoop -R /home/anas/data/hadoop_tmp/

sudo chmod 777 /home/anas/data /hadoop tmp/hdfs/namenode
```

d. **Change the path in hdfs-site.xml:**

```
<?xml version="1.0" encoding="UTF-8"?>
<?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>dfs.replication</name>
<value>3</value>
</property>
<property>
<name>dfs.namenode.name.dir</name>
<value>/home/anas/data/hadoop_temp/hdfs/namenode</value>
</property>
</configuration>
```

15. Recreate Datanode folder(All Nodes Only)

a. **First you have to delete the older hdfs file.**

```
sudo rm -rf /home/anas/Desktop/hdfs
```

b. **Then create another file that contains hdfs file like:**

```
sudo mkdir -p /home/anas/data/hadoop_tmp/hdfs/datanode
```

c. **Then type these commands:**

```
sudo chown anas:hadoop -R /home/anas/data/hadoop_tmp/

sudo chmod 777 /home/anas/data /hadoop tmp/hdfs/datanode
```

d. Change the path in hdfs-site.xml:

```
<?xml version="1.0" encoding="UTF-8"?>
<?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>dfs.replication</name>
<value>3</value>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>/home/anas/data/hadoop_tmp/hdfs/datanode</value></property>
</configuration>
~
```

16. Format the Namenode(Master only)

Before starting the cluster, we need to format the Namenode. Use the following command only on master node:

```
hdfs namenode -format
```

17. Start the DFS & Yarn (Master Only)

```
anas@master:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [master]
master: starting namenode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-namenode-master.out
node3: starting datanode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-datanode-node3.out
node2: starting datanode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-datanode-node2.out
node1: starting datanode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-datanode-node1.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-secondarynamenode-master.out
starting yarn daemons
starting resourcemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-resourcemanager-master.out
node3: starting nodemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-nodemanager-node3.out
node1: starting nodemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-nodemanager-node1.out
node2: starting nodemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-nodemanager-node2.out
anas@master:~$
```

You should observe that it tries to start data node on slave nodes one by one. Once it is started, Do a JPS on Master and slaves.

- JPS on Master node

```
anas@master:~/data/hadoop-2.7.3/etc/hadoop$ start-dfs.sh && start-yarn.sh
Starting namenodes on [master]
master: Warning: Permanently added the ECDSA host key for IP address '192.168.0.172' to the list of known hosts.
master: starting namenode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-namenode-master.out
node1: starting datanode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-datanode-node1.out
node3: starting datanode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-datanode-node3.out
node2: starting datanode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-datanode-node2.out
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/anas/data/hadoop-2.7.3/logs/hadoop-anas-secondarynamenode-master.out
starting yarn daemons
starting resourcemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-resourcemanager-master.out
node1: starting nodemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-nodemanager-node1.out
node3: starting nodemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-nodemanager-node3.out
node2: starting nodemanager, logging to /home/anas/data/hadoop-2.7.3/logs/yarn-anas-nodemanager-node2.out
anas@master:~/data/hadoop-2.7.3/etc/hadoop$ jps
26033 SecondaryNameNode
25809 NameNode
26443 Jps
26187 ResourceManager
anas@master:~/data/hadoop-2.7.3/etc/hadoop$
```

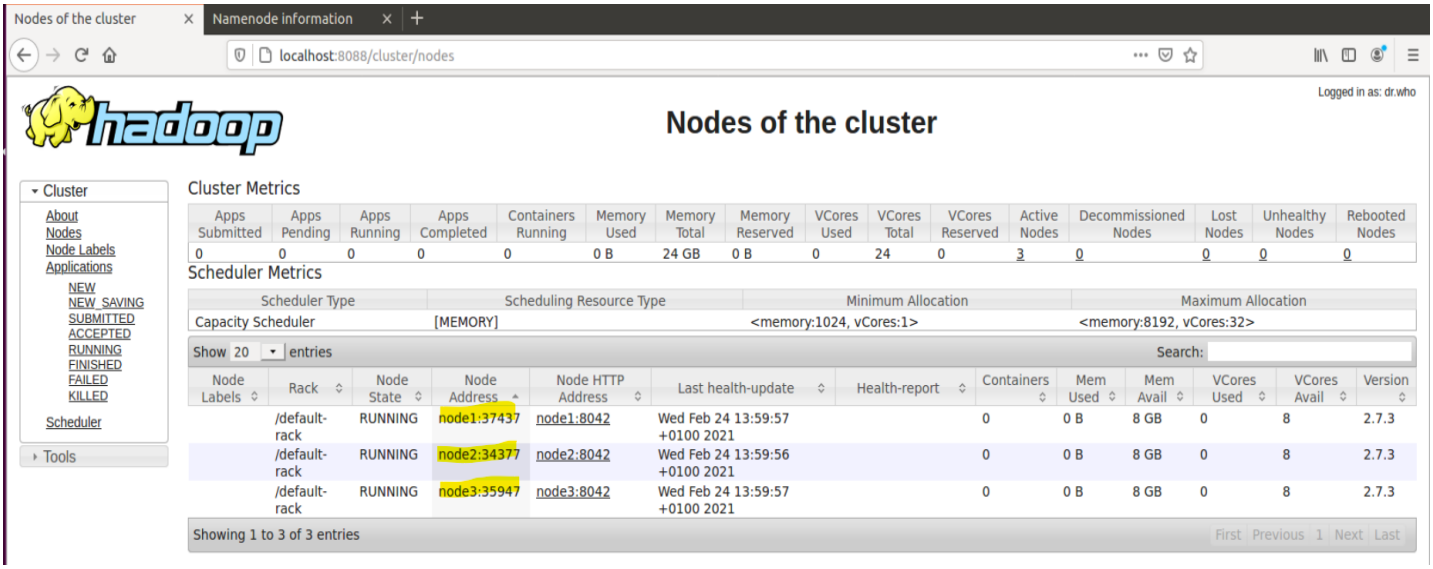
- JPS on slave nodes(node1 et node2 et node3)

```
anas@node2:~$ jps
28465 Jps
28338 NodeManager
28213 DataNode
anas@node2:~$
```

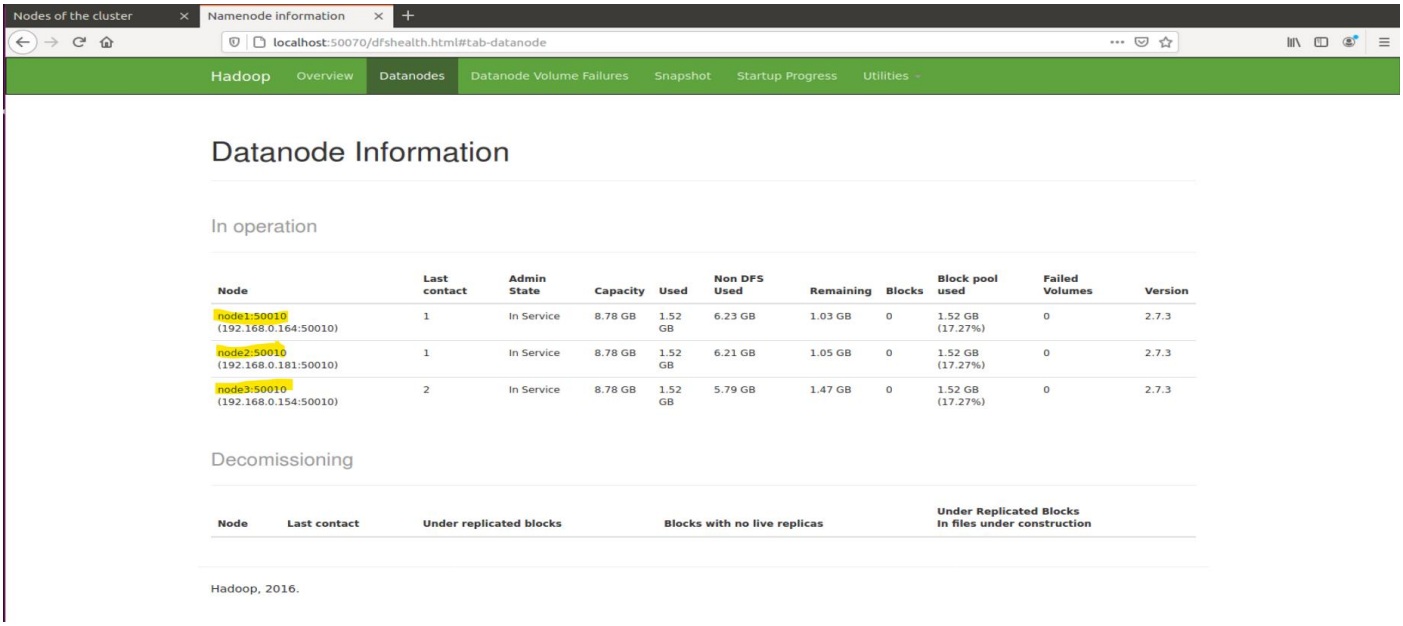
18. <u>Review Yarn console:</u>

If all the services started successfully on all nodes, then you should see all of your nodes listed under Yarn nodes. You can hit the following url on your browser and verify that:

- http://master:8088/cluster/nodes



- http://master:50070 # can show live node count and info about each live node.



- You can also get the report of your cluster by issuing the below commands

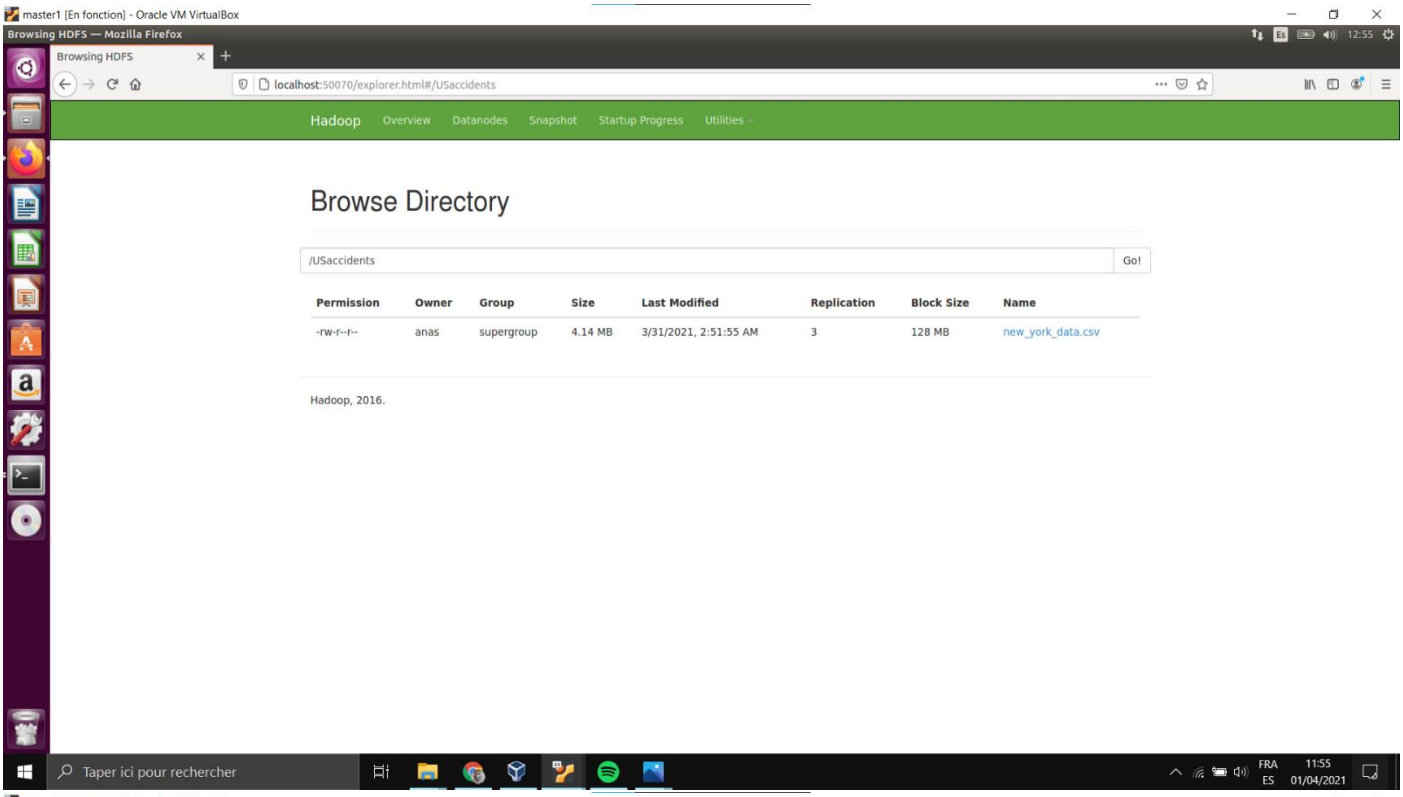
```
anas@master:~$ hdfs dfsadmin -report
```

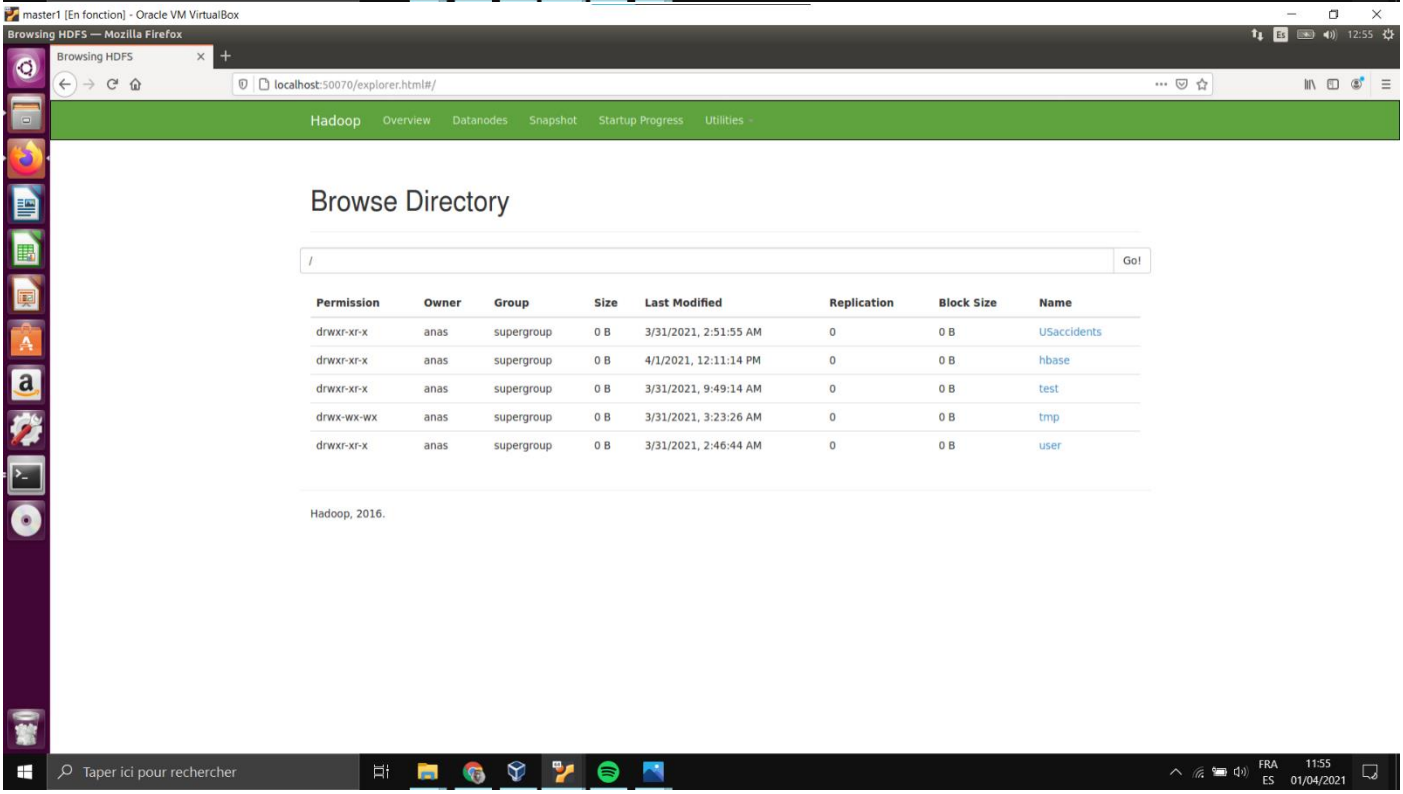
```
anas@master:~$ hdfs dfsadmin -report
Configured Capacity: 28275376128 (26.33 GB)
Present Capacity: 8699142144 (8.10 GB)
DFS Remaining: 3814756352 (3.55 GB)
DFS Used: 4884385792 (4.55 GB)
DFS Used%: 56.15%
Under replicated blocks: 0
Blocks with corrupt replicas: 0
Missing blocks: 0
Missing blocks (with replication factor 1): 0

-------------------------------------------------
Live datanodes (3):

Name: 192.168.0.154:50010 (node3)
Hostname: node3
Decommission Status : Normal
Configured Capacity: 9425125376 (8.78 GB)
DFS Used: 1628131328 (1.52 GB)
Non DFS Used: 6221590528 (5.79 GB)
DFS Remaining: 1575403520 (1.47 GB)
DFS Used%: 17.27%
DFS Remaining%: 16.71%
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: Wed Feb 24 14:39:07 CET 2021


Name: 192.168.0.164:50010 (node1)
Hostname: node1
Decommission Status : Normal
Configured Capacity: 9425125376 (8.78 GB)
DFS Used: 1628131328 (1.52 GB)
Non DFS Used: 6686531584 (6.23 GB)
DFS Remaining: 1110462464 (1.03 GB)
DFS Used%: 17.27%
DFS Remaining%: 11.78%
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: Wed Feb 24 14:39:07 CET 2021


Name: 192.168.0.181:50010 (node2)
Hostname: node2
Decommission Status : Normal
Configured Capacity: 9425125376 (8.78 GB)
DFS Used: 1628123136 (1.52 GB)
Non DFS Used: 6668111872 (6.21 GB)
DFS Remaining: 1128890368 (1.05 GB)
DFS Used%: 17.27%
DFS Remaining%: 11.98%
Configured Cache Capacity: 0 (0 B)
```

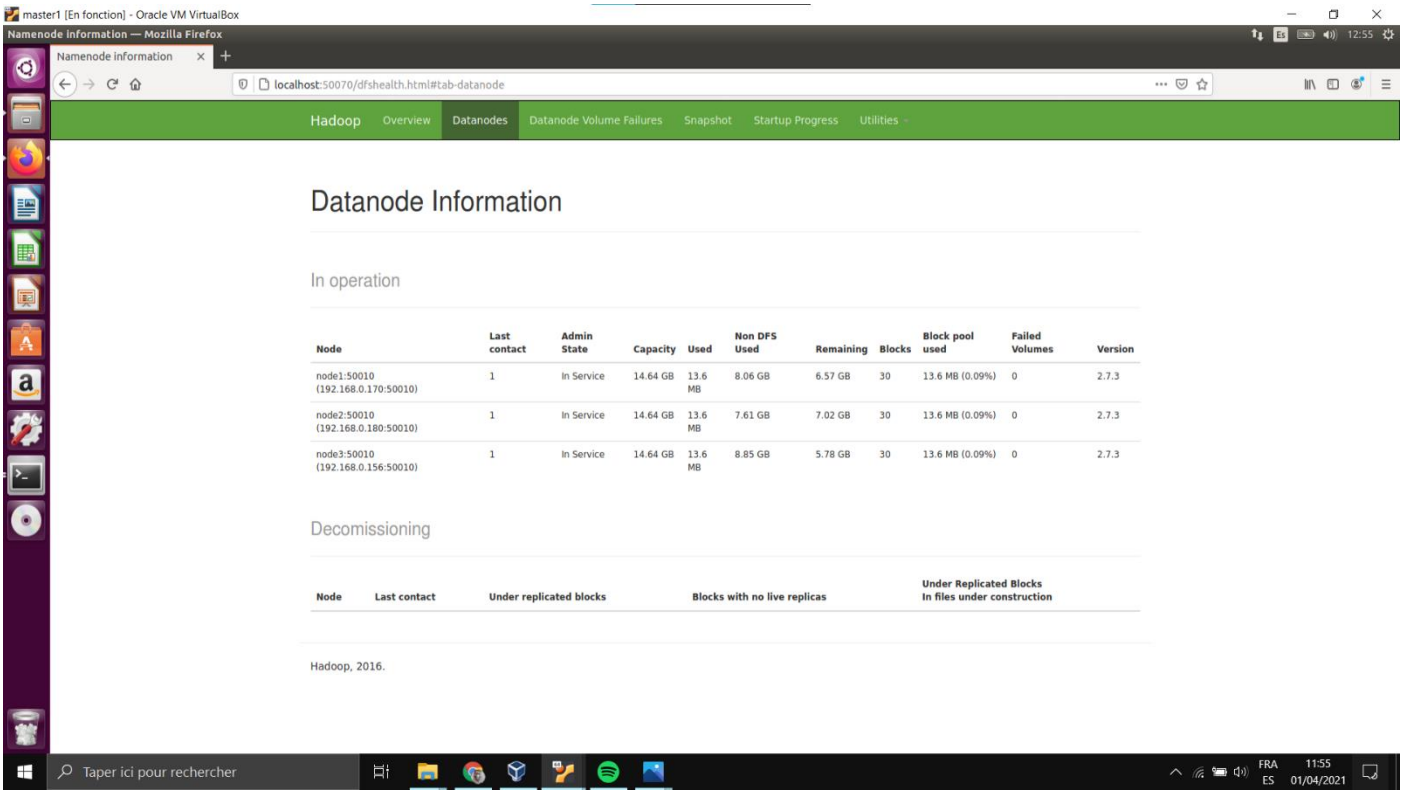
19. TP of a us-accidents dataset:

- https://www.kaggle.com/sobhanmoosavi/us-accidents

In this step we find that our partition blocks are not allocated because nothing was transferred into hdfs.

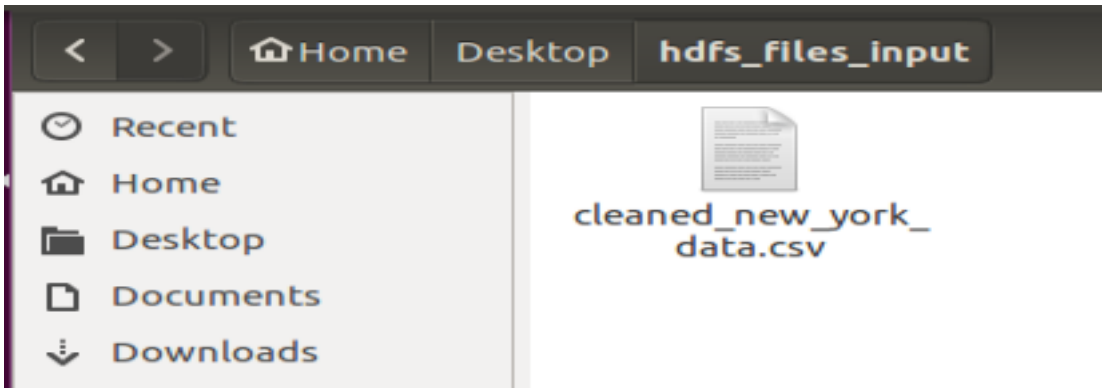




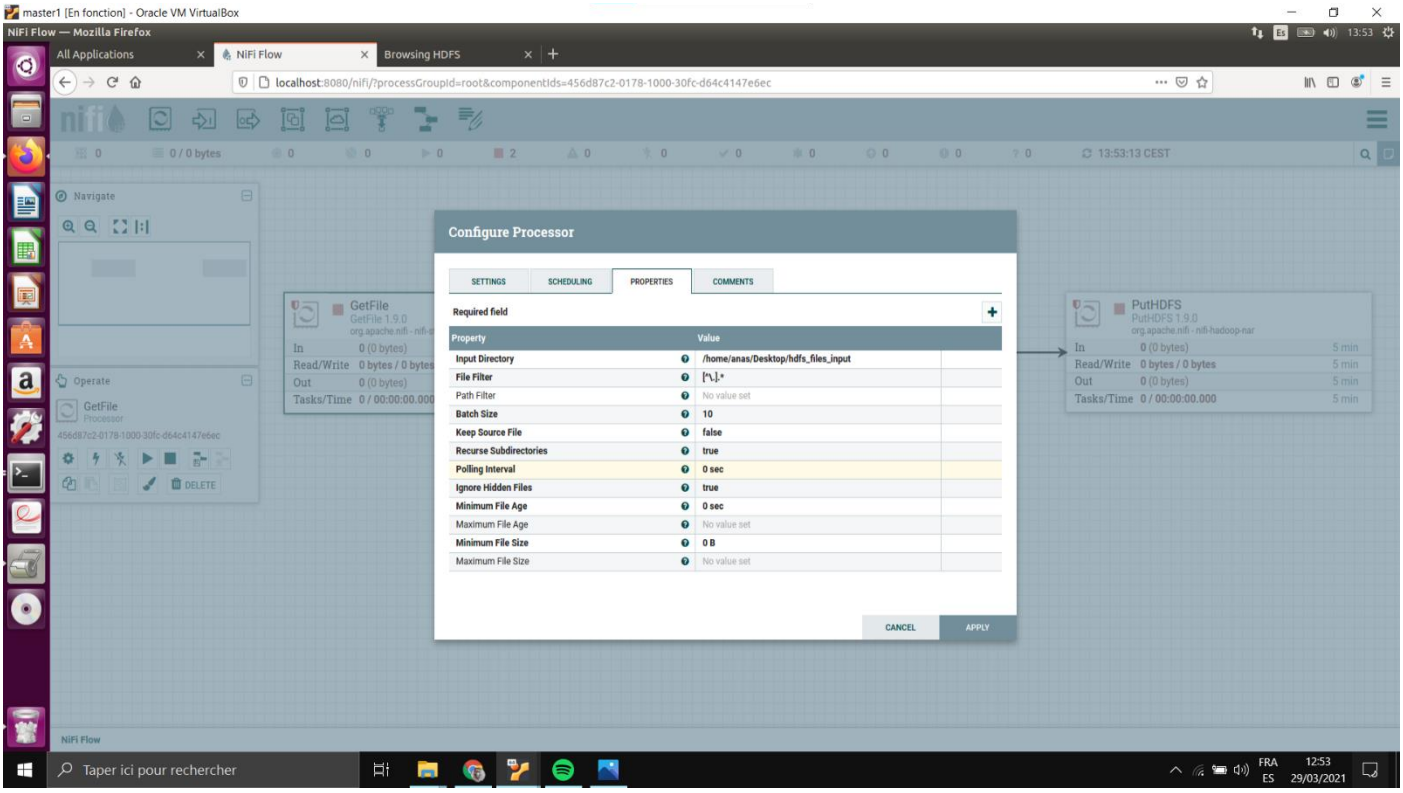

master1 [En fonction] - Oracle VM VirtualBox
Namenode information — Mozilla Firefox
Namenode information
localhost:50070/dfshealth.html#tab-datanode
Hadoop
Overview
Datanodes
Datanode Volume Failures
Snapshot
Startup Progress
Utilities
Datanode Information
In operation
Node
Last contact
Admin State
Capacity
Used
Non DFS Used
Remaining
Blocks
Block pool used
Failed Volumes
Version
node1:50010 (192.168.0.170:50010)
1
In Service
14.64 GB
13.6 MB
8.06 GB
6.57 GB
30
13.6 MB (0.09%)
0
2.7.3
node2:50010 (192.168.0.180:50010)
1
In Service
14.64 GB
13.6 MB
7.61 GB
7.02 GB
30
13.6 MB (0.09%)
0
2.7.3
node3:50010 (192.168.0.156:50010)
1
In Service
14.64 GB
13.6 MB
8.85 GB
5.78 GB
30
13.6 MB (0.09%)
0
2.7.3
Decomissioning
Node
Last contact
Under replicated blocks
Blocks with no live replicas
Under Replicated Blocks In files under construction
Hadoop, 2016.
Taper ici pour rechercher
FRA ES
11:55 01/04/2021

Sending Dataset from Local to Hdfs with Nifi

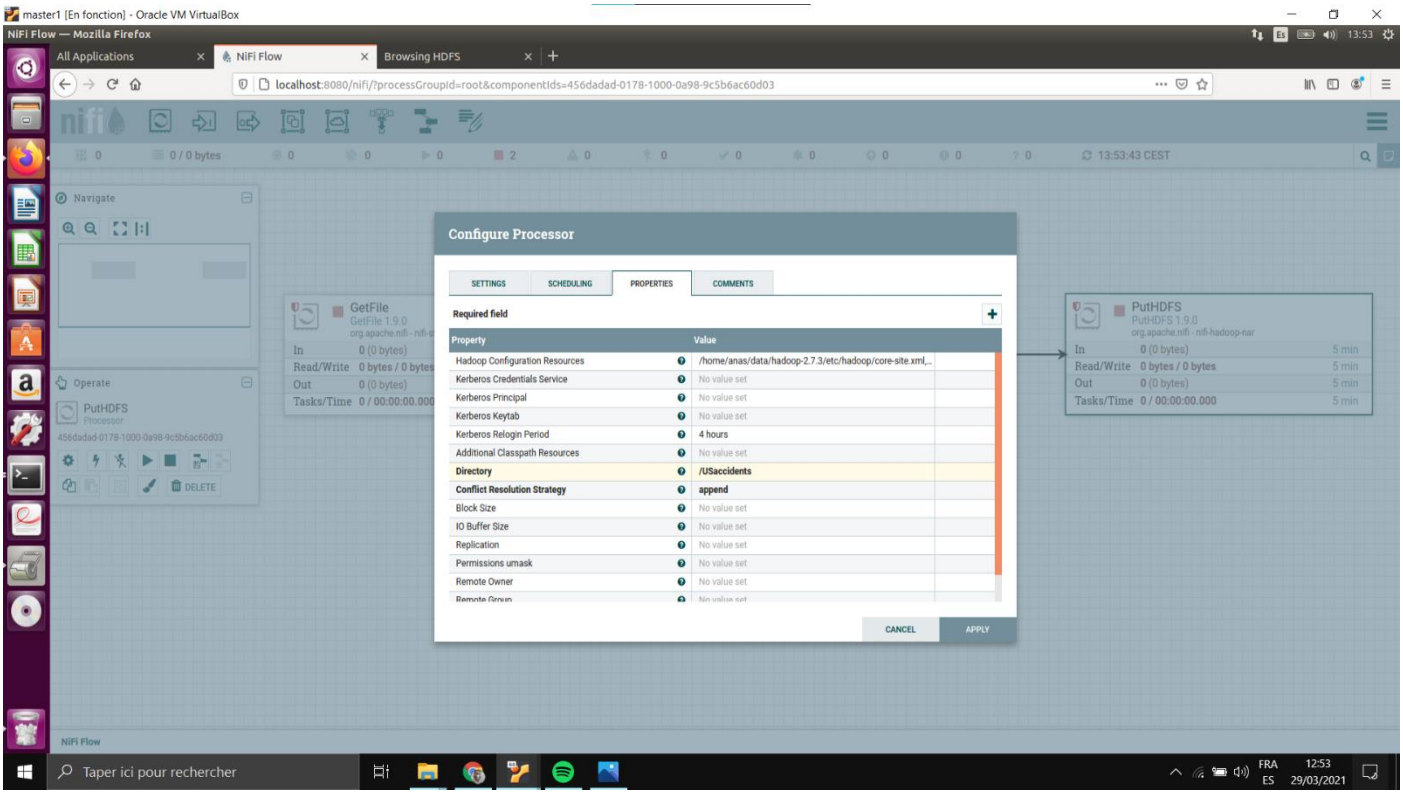
FICHIER NIFI INPUT



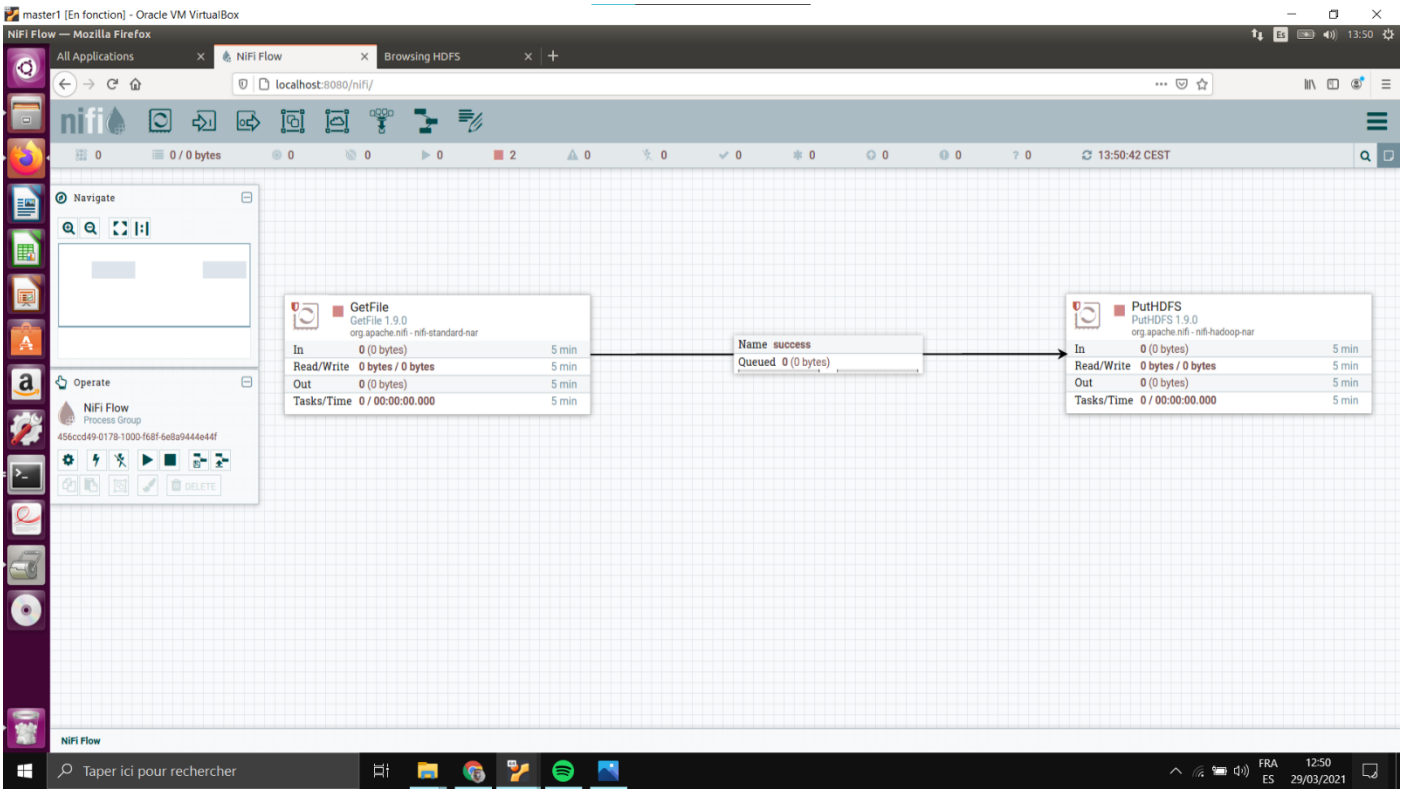
NIFI GETFILE



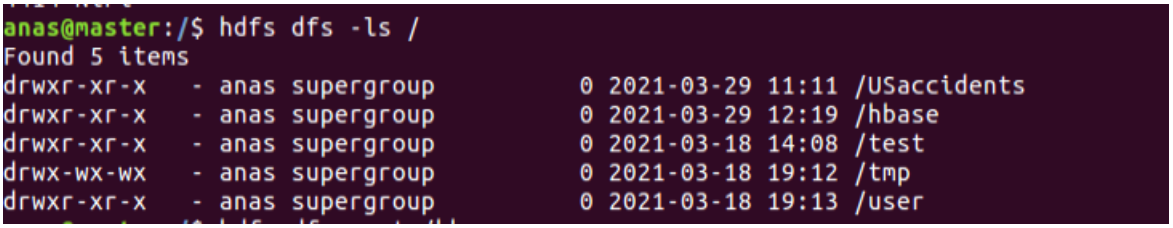
NIFI PUTHDFS



PUT FILE

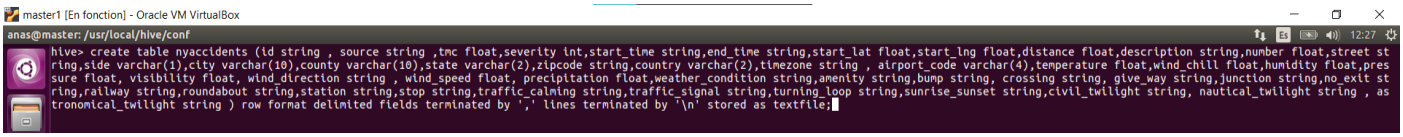


HDFS LS

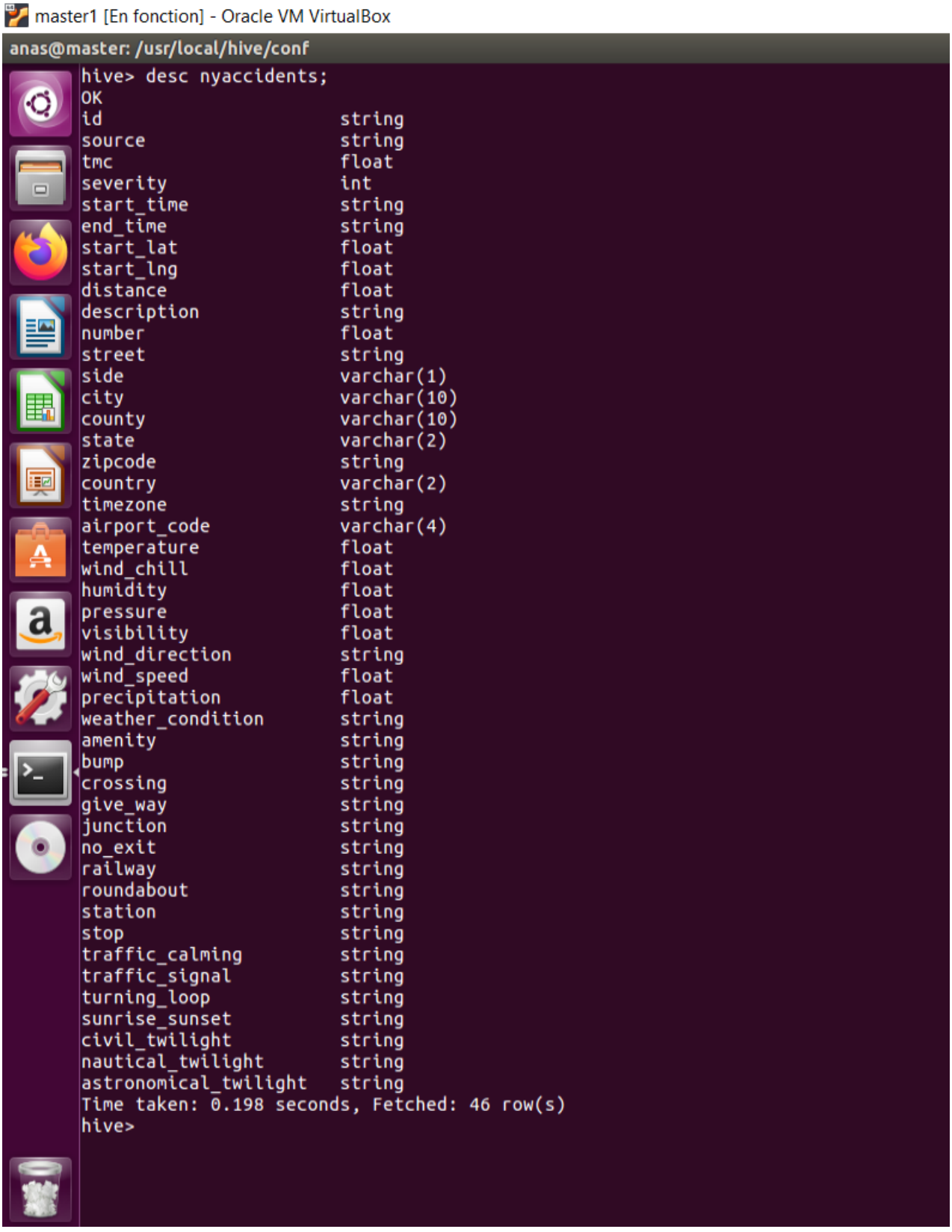

anas@master:/$ hdfs dfs -ls /
Found 5 items
drwxr-xr-x - anas supergroup 0 2021-03-29 11:11 /USaccidents
drwxr-xr-x - anas supergroup 0 2021-03-29 12:19 /hbase
drwxr-xr-x - anas supergroup 0 2021-03-18 14:08 /test
drwx-wx-wx - anas supergroup 0 2021-03-18 19:12 /tmp
drwxr-xr-x - anas supergroup 0 2021-03-18 19:13 /user

Creating Table and Loading Dataset from Local with Hive

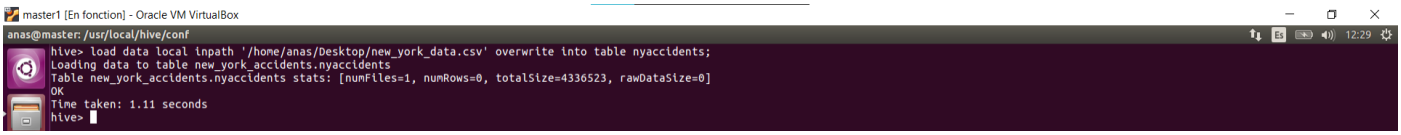
CREATION DATABASE HIVE



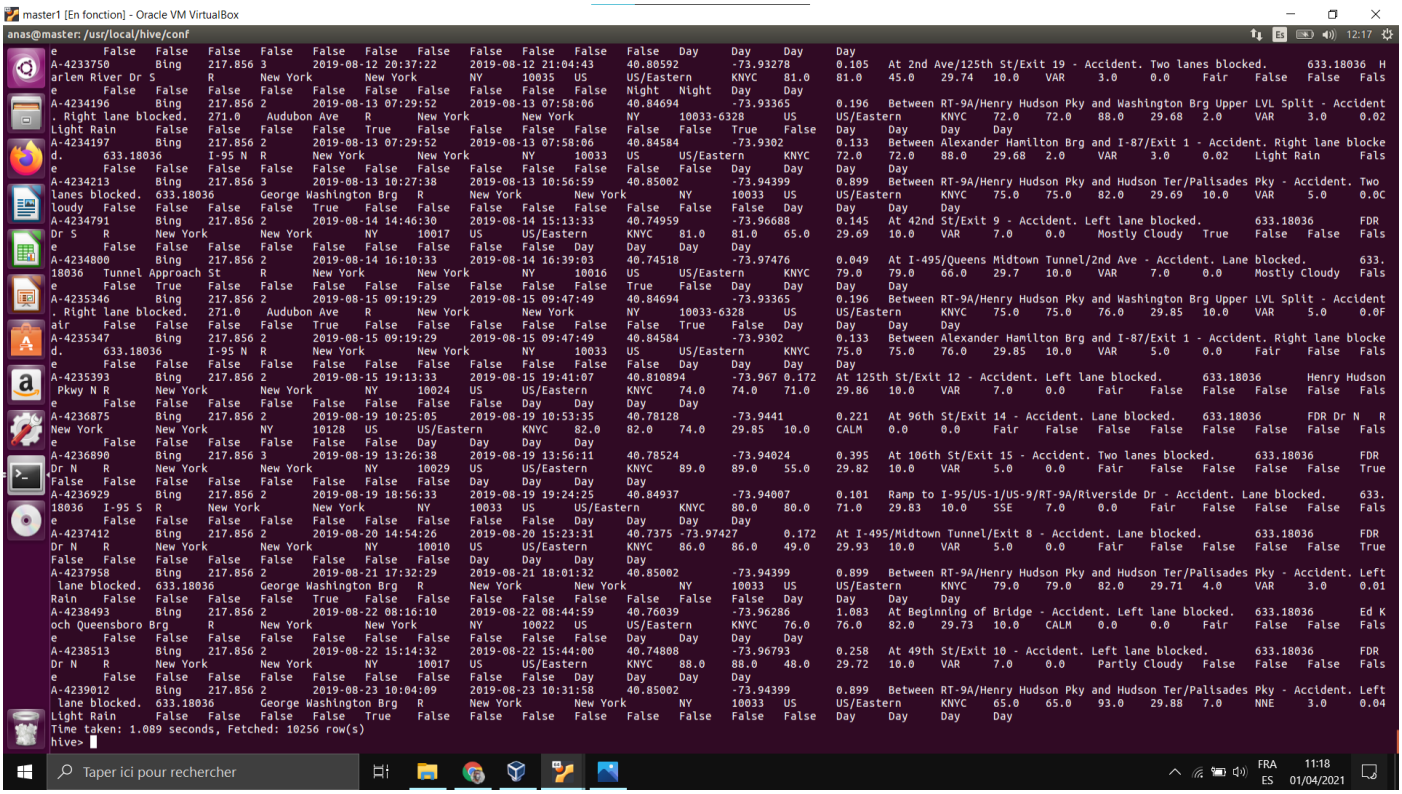
DESC HIVE



HIVE LOAD DATA



HIVE DATASET



Hbase on Hadoop in Distributed mode

HBASE-SITE.XML CONFIG

anas@master: /usr/local/hbase/conf

```
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
/**
 *
 * Licensed to the Apache Software Foundation (ASF) under one
 * or more contributor license agreements.  See the NOTICE file
 * distributed with this work for additional information
 * regarding copyright ownership.  The ASF licenses this file
 * to you 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.
 */
-->
<configuration>
<property>
<name>hbase.rootdir</name>
<value>hdfs://master:9000/hbase</value>
</property>
//Here you have to set the path where you want HBase to store its built in zookeeper files.
<property>
<name>hbase.zookeeper.property.dataDir</name>
<value>/home/anas/data/zookeeper</value>
</property>
<property>
<name>hbase.cluster.distributed</name>
<value>true</value>
</property>
<property>
  <name>hbase.zookeeper.quorum</name>
  <value>master:2181,node1:2181,node2:2181,node3:2181</value>
</property>
</configuration>
~
```

COPY HBASE WITH SCP FROM MASTER TO ALL NODES

(WITHOUT "#")

```
anas@master:/usr/local$ #scp hbase node1:/usr/local/
anas@master:/usr/local$ #scp hbase node2:/usr/local/
anas@master:/usr/local$ #scp hbase node3:/usr/local/
```

MASTER "JPS"

```
anas@master:~$ jps
5330 Jps
3046 NameNode
3430 ResourceManager
4007 HMaster
3944 HQuorumPeer
4392 RunJar
3272 SecondaryNameNode
4174 HRegionServer
```

ALL NODES "JPS"

--NODE1--

```
anas@node1:~$ jps
3318 NodeManager
3606 HRegionServer
3498 HQuorumPeer
3180 DataNode
3822 Jps
```

--NODE2—

```
anas@node2:~$ jps
2593 NodeManager
3447 HRegionServer
2460 DataNode
3342 HQuorumPeer
4543 Jps
```

--NODE3—

```
anas@node3:~$ jps
4290 Jps
3270 HQuorumPeer
2745 DataNode
3371 HRegionServer
2895 NodeManager
```

MASTER-STATUS #BASESTATS

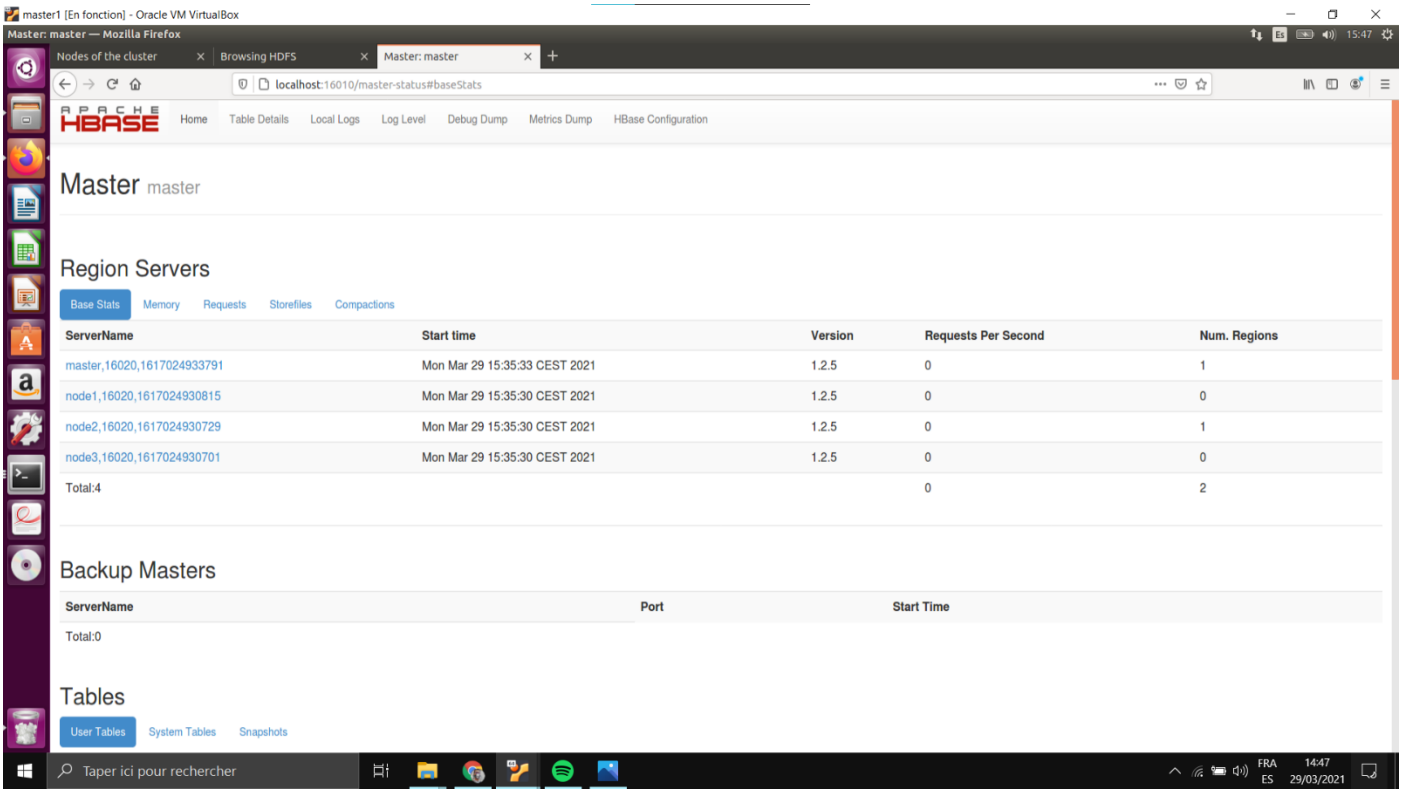
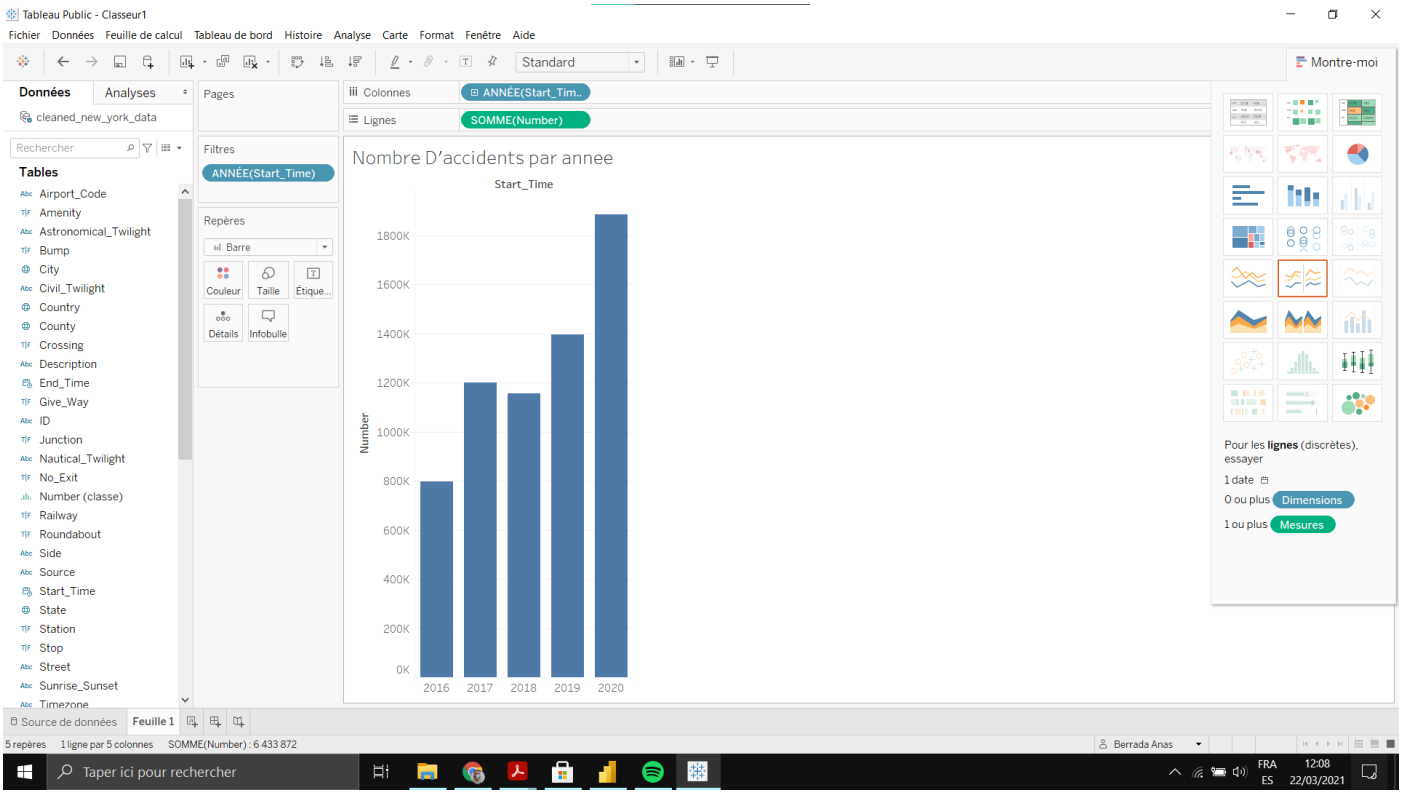
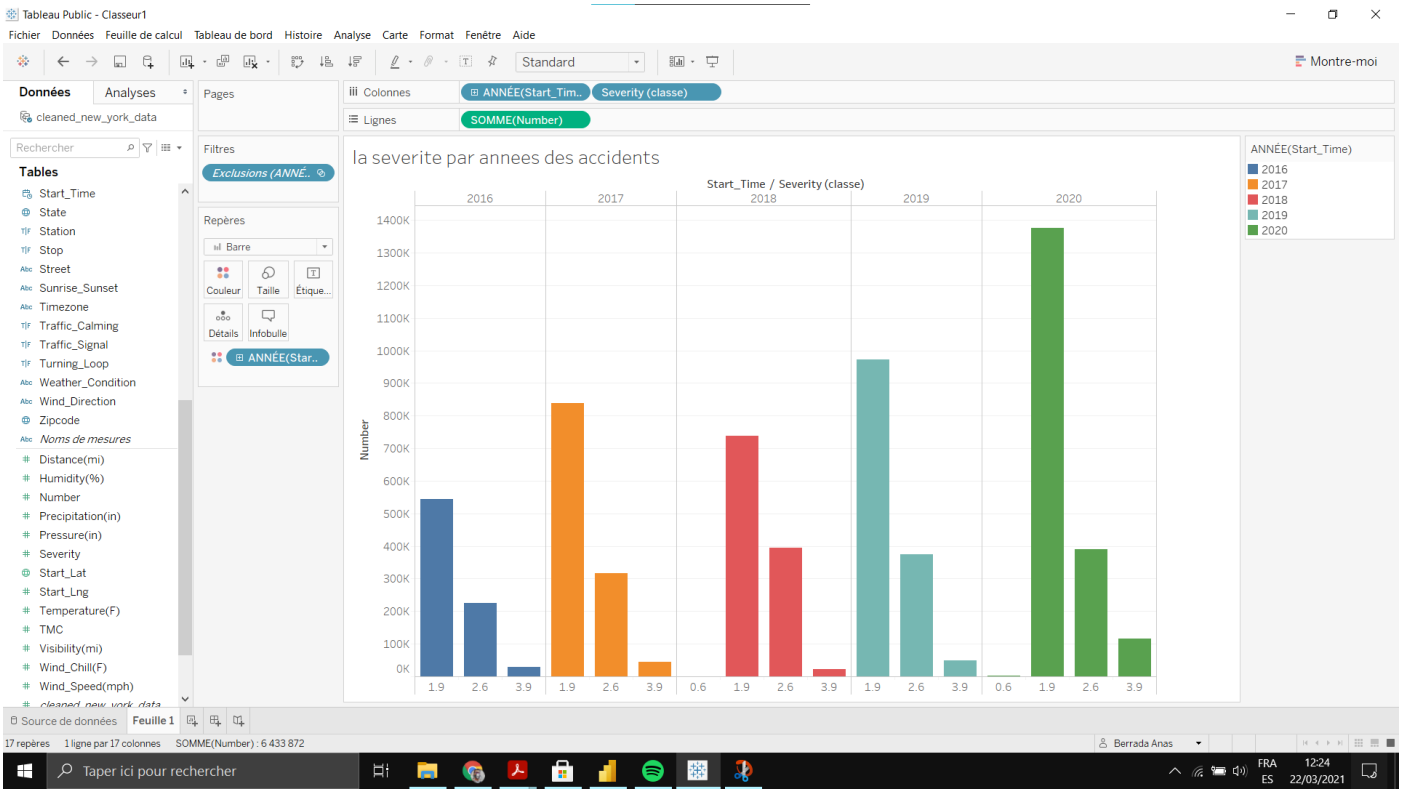


Tableau for Data Visualisation (for example New York)

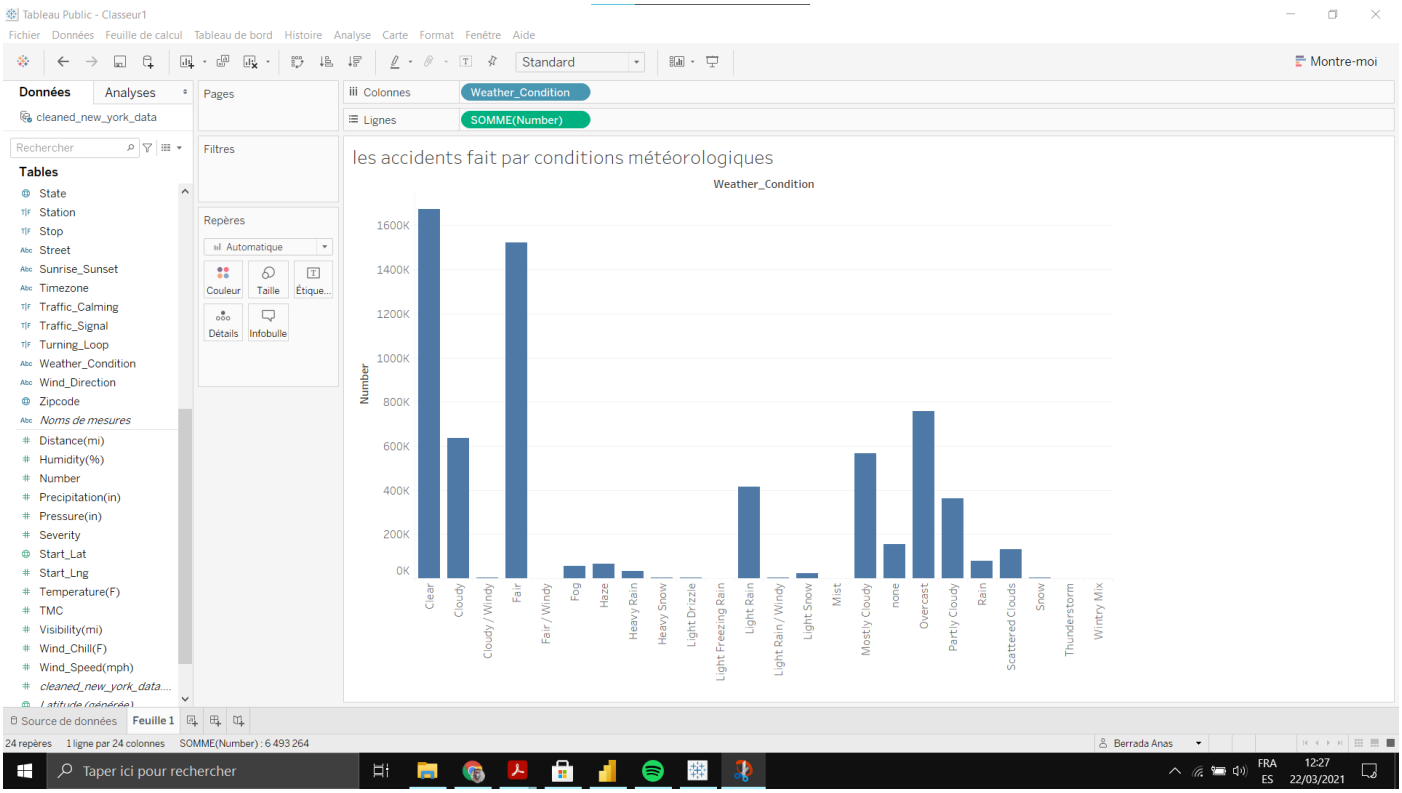
ACCIDENTS BY YEARS



THE SEVERITY BY YEARS OF ACCIDENTS



ACCIDENTS MADE BY WEATHER CONDITIONS



NUMBER OF ACCIDENTS BY DAY AND NIGHT