

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



Encadré par :

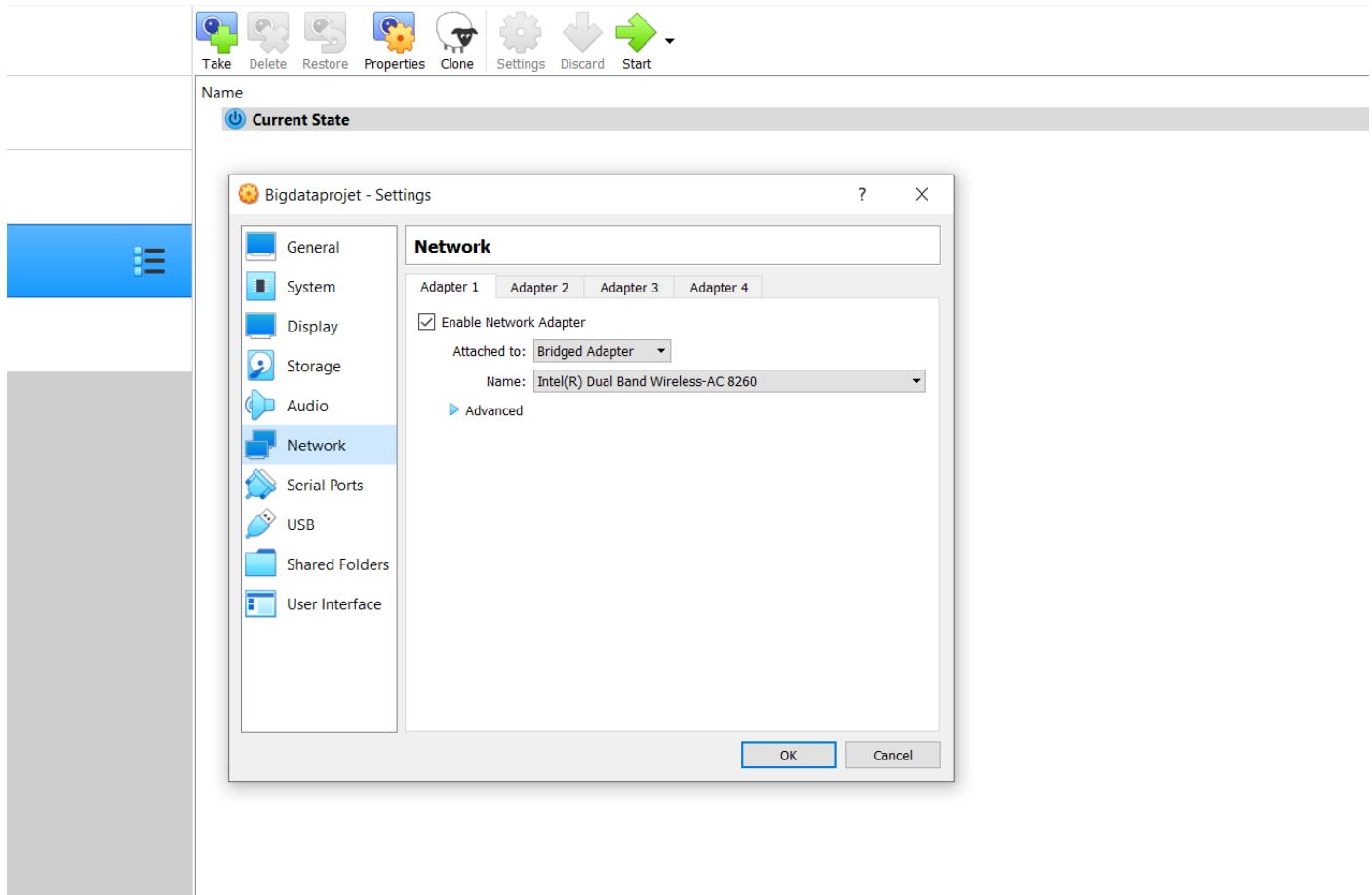
Prof. GHAZOUANI Mohamed

Réalisé par :

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

Setup Network:

YOU SHOULD CONFIGURE NETWORK AT BRIDGED ADAPTER

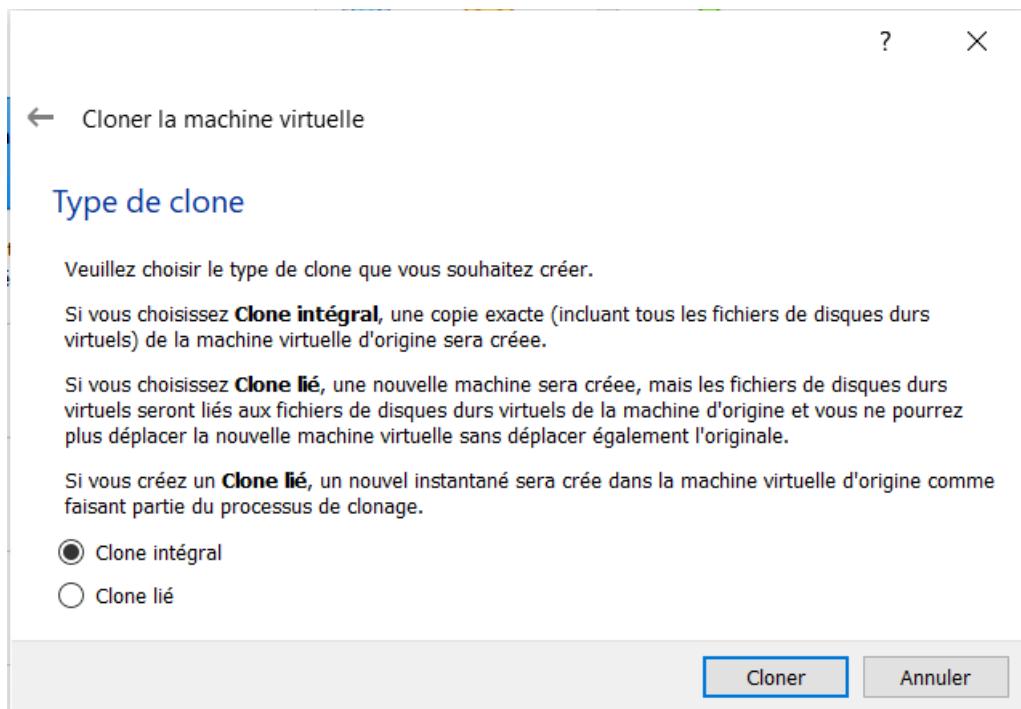
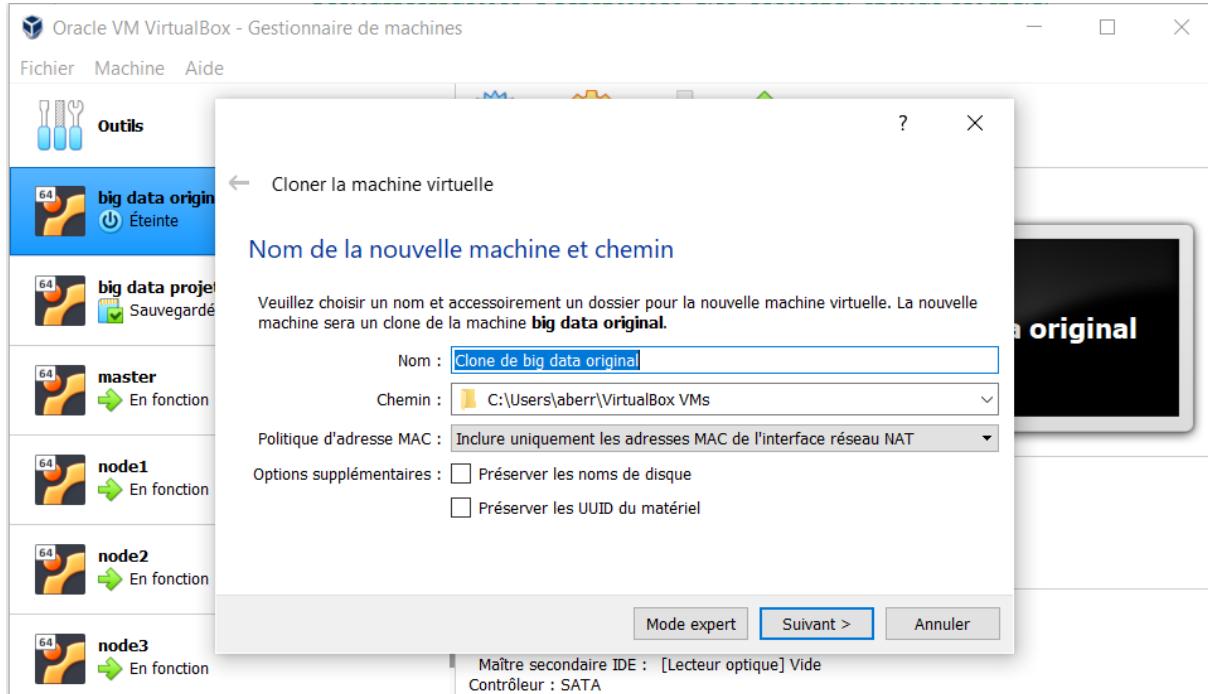


Multi node-cluster setup

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

```
anas@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

```
anas@master:~$ ssh node1
Warning: Permanently added the ECDSA host key for IP address '192.168.0.164' to the list of known hosts.
Welcome to Ubuntu 16.04.7 LTS (GNU/Linux 4.15.0-132-generic x86_64)

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

1 package can be updated.
0 of these updates are security updates.
To see these additional updates run: apt list --upgradable

*** System restart required ***
Last login: Mon Feb 22 23:34:17 2021 from 192.168.1.81
anas@node1:~$ exit
logout
Connection to node1 closed.
anas@master:~$ ssh node2
Warning: Permanently added the ECDSA host key for IP address '192.168.0.181' to the list of known hosts.
Welcome to Ubuntu 16.04.7 LTS (GNU/Linux 4.15.0-132-generic x86_64)

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

1 package can be updated.
0 of these updates are security updates.
To see these additional updates run: apt list --upgradable

*** System restart required ***
Last login: Mon Feb 22 22:58:57 2021 from 192.168.1.68
anas@node2:~$ exit
logout
Connection to node2 closed.
anas@master:~$ ssh node3
Warning: Permanently added the ECDSA host key for IP address '192.168.0.154' to the list of known hosts.
Welcome to Ubuntu 16.04.7 LTS (GNU/Linux 4.15.0-133-generic x86_64)

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

1 package can be updated.
0 of these updates are security updates.
To see these additional updates run: apt list --upgradable

*** System restart required ***
Last login: Mon Feb 22 22:59:19 2021 from 192.168.1.79
anas@node3:~$ exit
logout
Connection to node3 closed.
anas@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)

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

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



```
<?xml version="1.0" encoding="UTF-8"?>
<?xmlstylesheet 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"?>
<?xmlstylesheet 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
```



```
<?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>
<property>
<name>yarn.resourcemanager.resource-tracker.address</name>
<value>master:8025</value>
</property>
<property>
<name>yarn.resourcemanager.scheduler.address</name>
<value>master:8030</value>
</property>
<property>
<name>yarn.resourcemanager.address</name>
<value>master:8050</value>
</property>
</configuration>
~
```

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

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

```
<?xml version="1.0"?>
<xsl: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>mapreduce.framework.name</name>
<value>yarn</value>
</property>
<property>
<name>mapreduce.jobhistory.address</name>
<value>master:10020</value>
<description>Host and port for Job History Server (default 0.0.0.0:10020)</description>
</property>
</configuration>
```

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/anis/Desktop/hdfs
```

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

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

- c. Then type these commands:

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

```
sudo chmod 777 /home/anis/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. Review Yarn console:

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>

Cluster Metrics																
Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	Vcores Used	Vcores Total	Vcores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	
0	0	0	0	0	0 B	24 GB	0 B	0	24	0	3	0	0	0	0	

Scheduler Metrics		Scheduling Resource Type		Minimum Allocation				Maximum Allocation								
Capacity Scheduler		[MEMORY]		<memory:1024, vCores:1>				<memory:8192, vCores:32>								
Show 20 entries												Search:				
Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Mem Used	Mem Avail	Vcores Used	Vcores Avail	Version				
/default-rack	RUNNING	node1:37437	node1:8042	node1:8042	Wed Feb 24 13:59:57 +0100 2021		0	0 B	8 GB	0	8	2.7.3				
/default-rack	RUNNING	node2:34377	node2:8042	node2:8042	Wed Feb 24 13:59:56 +0100 2021		0	0 B	8 GB	0	8	2.7.3				
/default-rack	RUNNING	node3:35947	node3:8042	node3:8042	Wed Feb 24 13:59:57 +0100 2021		0	0 B	8 GB	0	8	2.7.3				

Showing 1 to 3 of 3 entries First Previous 1 Next Last



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

Screenshot of the HDFS健康检查页面 (localhost:50070/dfshealth.html#tab-datanode) showing Datanode Information.

The page has tabs: Overview, Datanodes (selected), 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 (192.168.0.164:50010)	1	In Service	8.78 GB	1.52 GB	6.23 GB	1.03 GB	0	1.52 GB (17.27%)	0	2.7.3
node2:50010 (192.168.0.181:50010)	1	In Service	8.78 GB	1.52 GB	6.21 GB	1.05 GB	0	1.52 GB (17.27%)	0	2.7.3
node3:50010 (192.168.0.154:50010)	2	In Service	8.78 GB	1.52 GB	5.79 GB	1.47 GB	0	1.52 GB (17.27%)	0	2.7.3

Decommissioning

Node	Last contact	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks In files under construction
------	--------------	-------------------------	------------------------------	--

Hadoop, 2016.



- 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.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	anas	supergroup	4.14 MB	3/31/2021, 2:51:55 AM	3	128 MB	new_york_data.csv

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	anas	supergroup	0 B	3/31/2021, 2:51:55 AM	0	0 B	USAccidents
drwxr-xr-x	anas	supergroup	0 B	4/1/2021, 12:11:14 PM	0	0 B	hbase
drwxr-xr-x	anas	supergroup	0 B	3/31/2021, 9:49:14 AM	0	0 B	test
drwxrwxrwx	anas	supergroup	0 B	3/31/2021, 3:23:26 AM	0	0 B	tmp
drwxr-xr-x	anas	supergroup	0 B	3/31/2021, 2:46:44 AM	0	0 B	user



master1 [En fonction] - Oracle VM VirtualBox

Namenode Information — Mozilla Firefox

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

Decommissioning

Node	Last contact	Under replicated blocks	Blocks with no live replicas	Under Replicated Blocks In files under construction
------	--------------	-------------------------	------------------------------	--

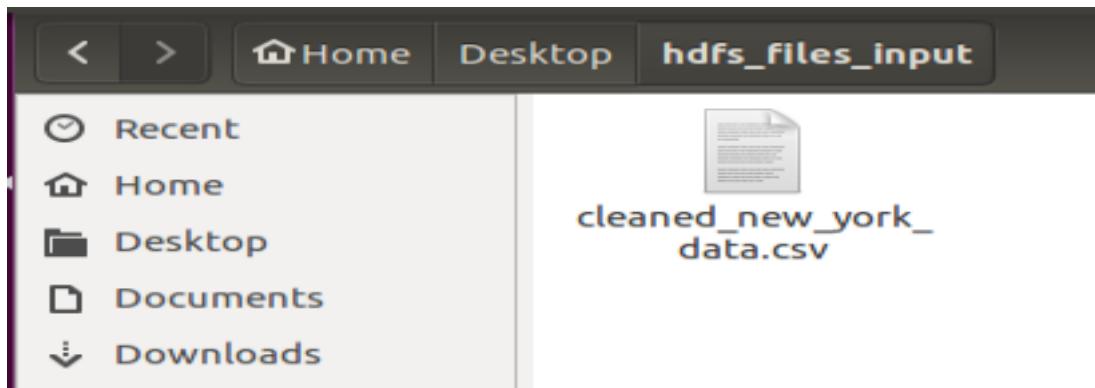
Hadoop, 2016.

Taper ici pour rechercher

FRA 11:55 ES 01/04/2021

Sending Dataset from Local to Hdfs with Nifi

FICHIER NIFI INPUT



NIFI GETFILE

The screenshot shows the Apache NiFi web interface. A 'Configure Processor' dialog is open for a 'GetFile' processor. The 'SETTINGS' tab is selected, showing the following properties:

Property	Value
Input Directory	/home/anas/Desktop/hdfs_files_input
File Filter	[\d]*
Batch Size	10
Keep Source File	false
Recurse Subdirectories	true
Polling Interval	0 sec
Ignore Hidden Files	true
Minimum File Age	0 sec
Maximum File Age	No value set
Minimum File Size	0 B
Maximum File Size	No value set

The 'SCHEDULING' tab is also visible. To the right, a 'PutHDFS' processor is shown in the flow, connected to the 'GetFile' processor. The 'PutHDFS' processor has the following settings:

Property	Value	Interval
In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min



NIFI PUTHDFS

The screenshot shows the NiFi Flow interface with a configuration dialog for a PutHDFS processor. The dialog is titled "Configure Processor" and has tabs for SETTINGS, SCHEDULING, PROPERTIES, and COMMENTS. The PROPERTIES tab is selected, showing configuration for required fields like Hadoop Configuration Resources, Kerberos Credentials Service, and Directory. The "Directory" field is set to "/USaccidents". Other properties include Conflict Resolution Strategy (set to "append"), Block Size, IO Buffer Size, Replication, Permissions umask, Remote Owner, and Remote Group. The "APPLY" button is visible at the bottom right of the dialog.

PUT FILE

The screenshot shows a basic NiFi flow. It starts with a "GetFile" processor (version 1.9.0) which reads from the "nifi-standard-nar" file. The output of "GetFile" is connected to a "Name success" placeholder node. This node then connects to a "PutHDFS" processor (version 1.9.0). The "PutHDFS" processor is configured to write to the "nifi-hadoop-nar" file. Both the "GetFile" and "PutHDFS" processors have a "5 min" timeout setting. The flow is visualized as a sequence of nodes connected by arrows, with the "Name success" node serving as a connector between the two main data handling components.



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
drwxrwx-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

```
master1 [En fonction] - Oracle VM VirtualBox
anas@master:/usr/local/hive/conf
hive> create table nyaccidents (id string , source string ,tmc float,severity int,start_time string,end_time string,start_lat float,start_lng float,distance float,number float,street st
ring,side varchar(1),city varchar(10),county varchar(10),state varchar(2),zipcode string,country varchar(4),timezone string , airport_code varchar(4),temperature float,wind_chill float,humidity float,pres
sure float, visibility float, wind_direction string , wind_speed float, precipitation float,weather_condition string,amenity string,bump string, crossing string, give_way string,junction string,no_exit st
ring,railway string,roundabout string,station string,stop string,traffic_calming string,traffic_signal string,turning_loop string,sunrise_sunset string,civil_twilight string, nautical_twilight string , as
trononical_twilight string ) row format delimited fields terminated by ',' lines terminated by '\n' stored as textfile;
```

DESC HIVE

master1 [En fonction] - Oracle VM VirtualBox

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

```
hive> desc nyaccidents;
OK
id                      string
source                  string
tmc                     float
severity                int
start_time              string
end_time                string
start_lat               float
start_lng               float
distance                float
description             string
number                  float
street                  string
side                    varchar(1)
city                    varchar(10)
county                 varchar(10)
state                   varchar(2)
zipcode                string
country                varchar(2)
timezone               string
airport_code            varchar(4)
temperature             float
wind_chill              float
humidity                float
pressure                float
visibility              float
wind_direction          string
wind_speed              float
precipitation           float
weather_condition        string
amenity                 string
bump                   string
crossing                string
give_way                string
junction               string
no_exit                 string
railway                 string
roundabout              string
station                 string
stop                   string
traffic_calming         string
traffic_signal           string
turning_loop             string
sunrise_sunset           string
civil_twilight           string
nautical_twilight        string
astronomical_twilight   string
Time taken: 0.198 seconds, Fetched: 46 row(s)
hive>
```



HIVE LOAD DATA

```
[master1:~] [En fonction] - Oracle VM VirtualBox  
anas@master:~/usr/local/hive/conf  
hive> load data local inpath '/home/anas/Desktop/new_york_data.csv' overwrite into table nyaccidents;  
Loading data to table new_york_accidents.nyaccidents  
Table new_york_accidents.nyaccidents stats: [numFiles=1, numRows=0, totalSize=4336523, rawDataSize=0]  
OK  
Time taken: 1.11 seconds  
hive>
```

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

The screenshot shows the Apache HBase master status interface running in Mozilla Firefox. The browser tabs include "master1 [En fonction] - Oracle VM VirtualBox", "Master: master — Mozilla Firefox", "Nodes of the cluster", "Browsing HDFS", and "Master: master". The main content area displays the following information:

- Region Servers:**

ServerName	Start time	Version	Requests Per Second	Num. Regions
master,16020,1617024933791	Mon Mar 29 15:35:33 CEST 2021	1.2.5	0	1
node1,16020,1617024930815	Mon Mar 29 15:35:30 CEST 2021	1.2.5	0	0
node2,16020,1617024930729	Mon Mar 29 15:35:30 CEST 2021	1.2.5	0	1
node3,16020,1617024930701	Mon Mar 29 15:35:30 CEST 2021	1.2.5	0	0
Total:4			0	2
- Backup Masters:**

ServerName	Port	Start Time

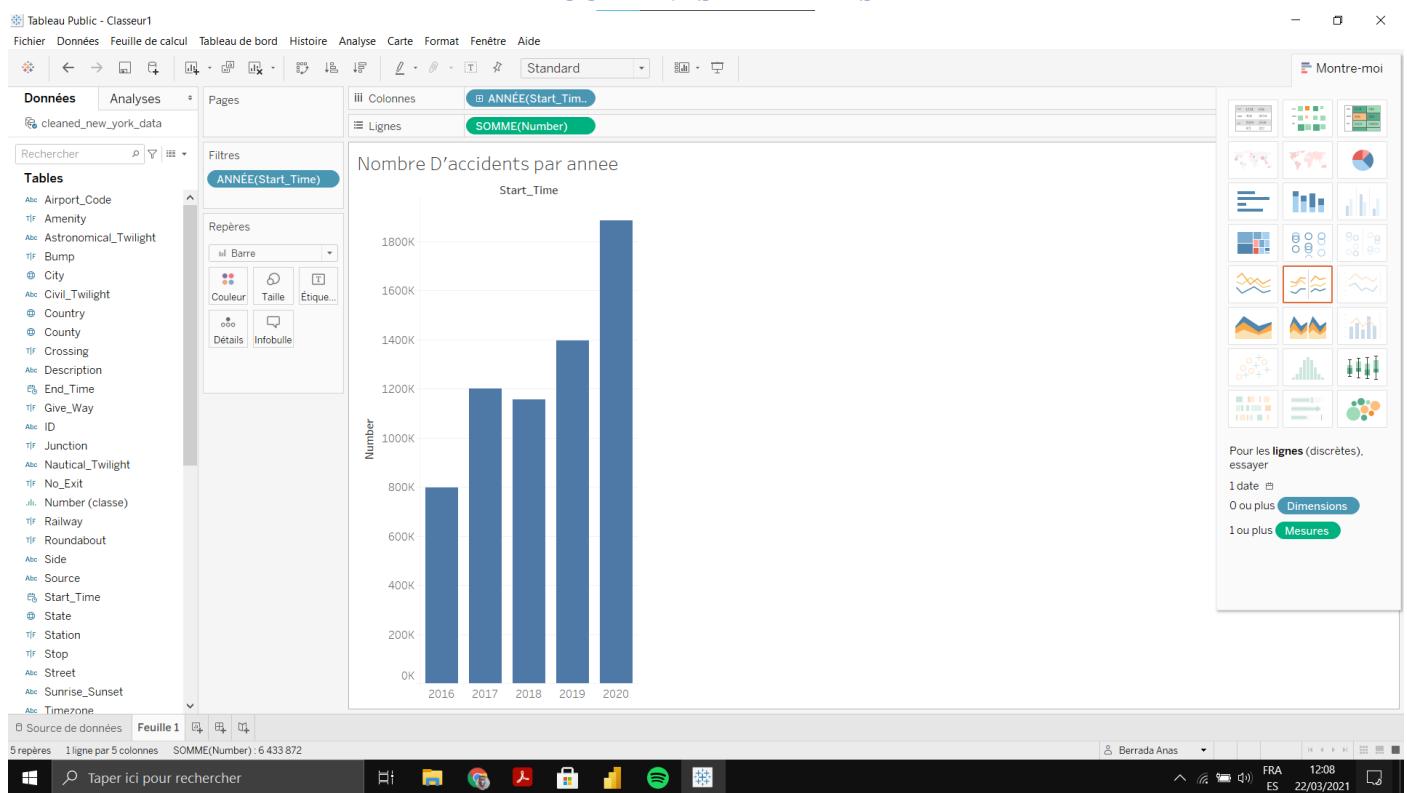
Total:0
- Tables:**

User Tables	System Tables	Snapshots

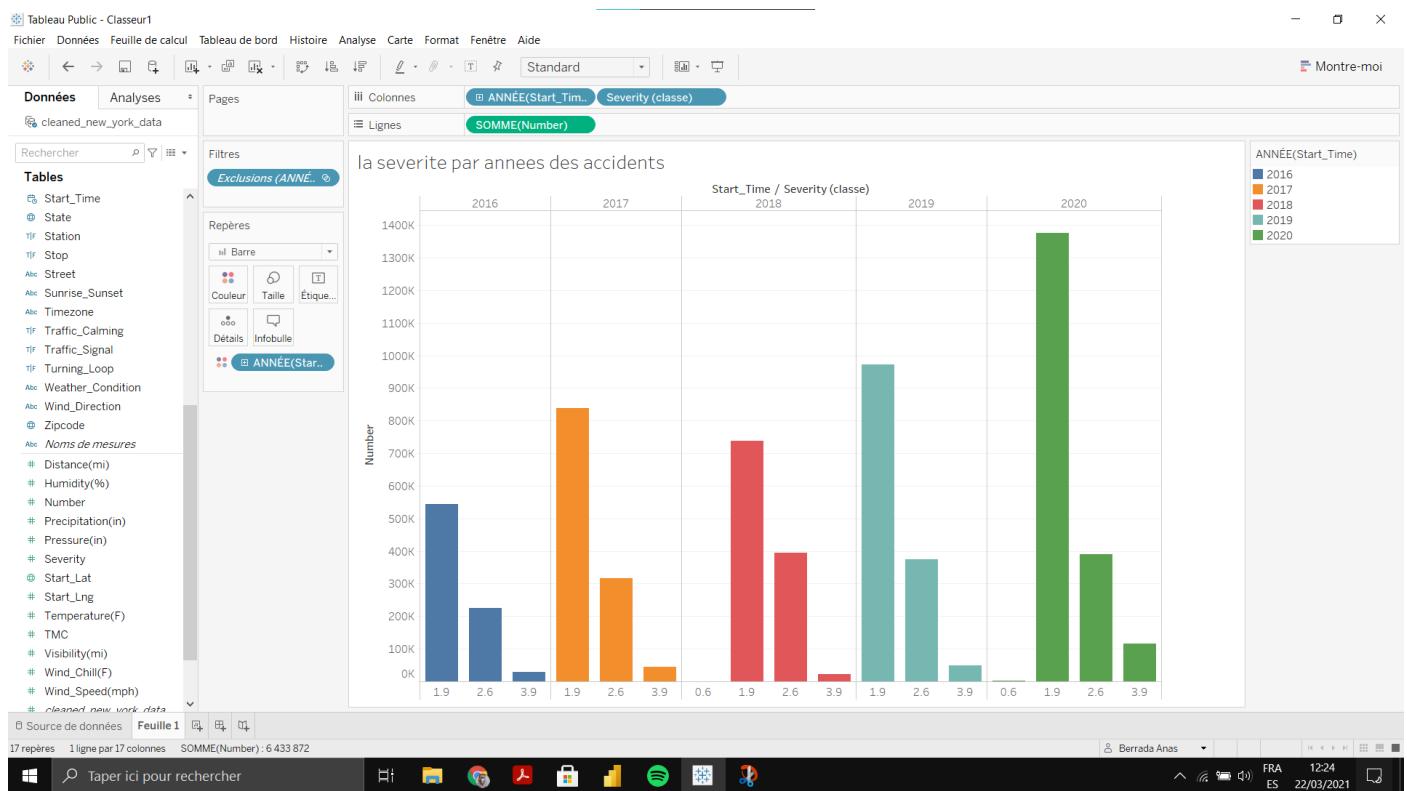
At the bottom of the browser window, there is a search bar with the placeholder "Taper ici pour rechercher" and a set of system icons.

Tableau for Data Visualisation (for example New York)

ACCIDENTS BY YEARS

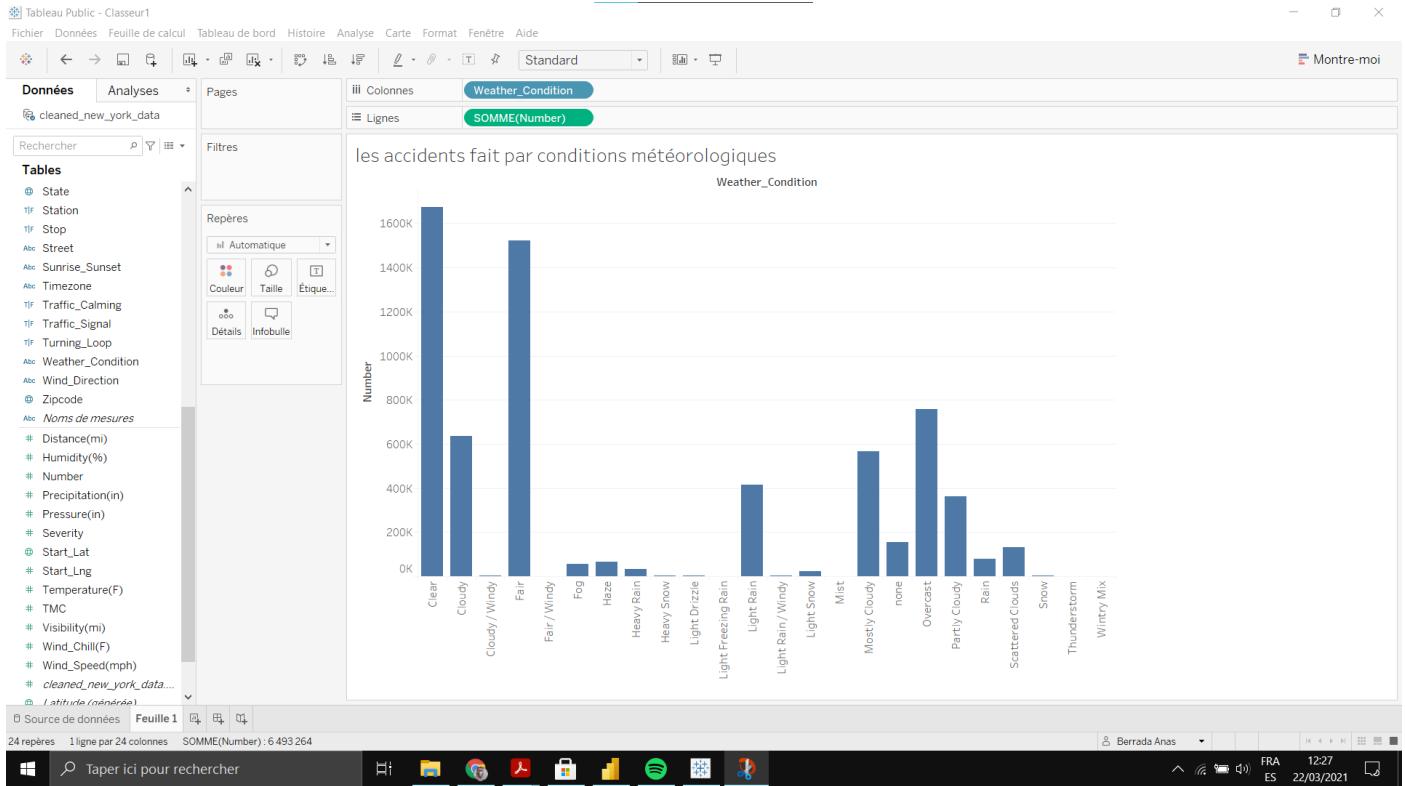


THE SEVERITY BY YEARS OF ACCIDENTS





ACCIDENTS MADE BY WEATHER CONDITIONS



NUMBER OF ACCIDENTS BY DAY AND NIGHT

