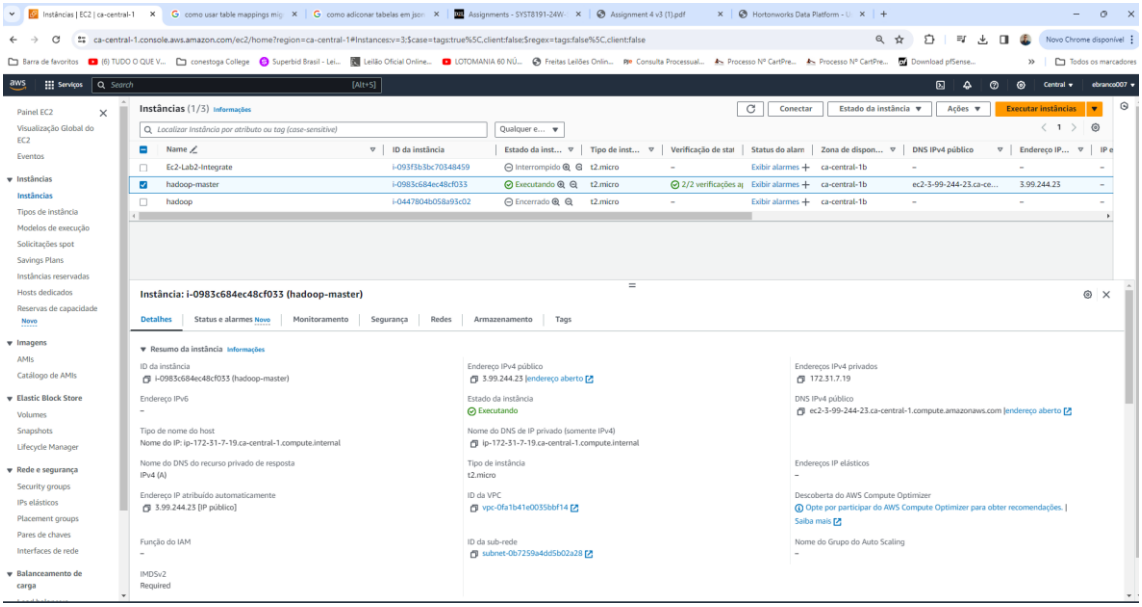
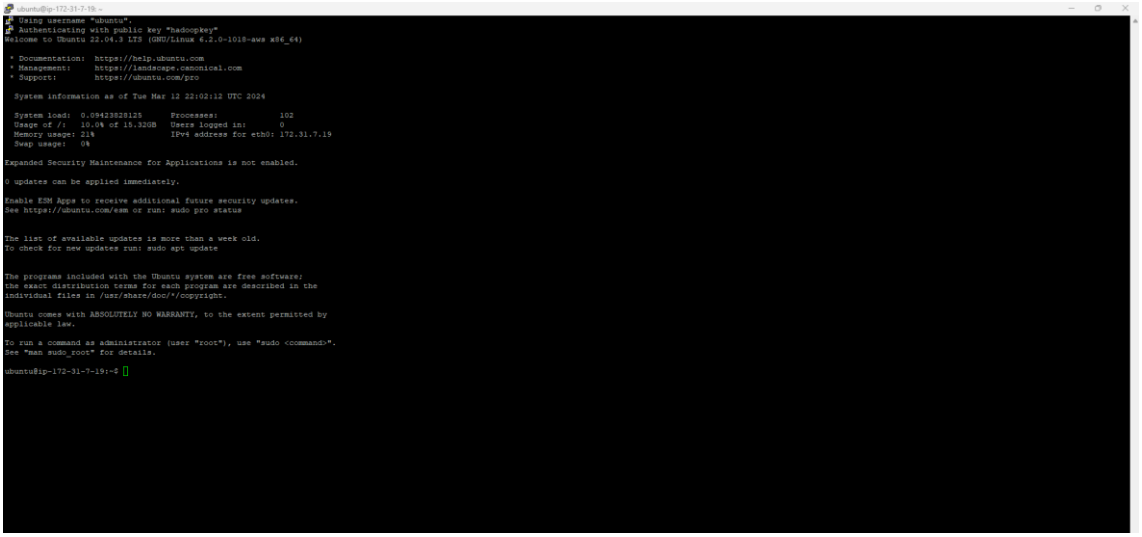
Deploy Hadoop on AWS

1- Deploy Ec2 instance on AWS.

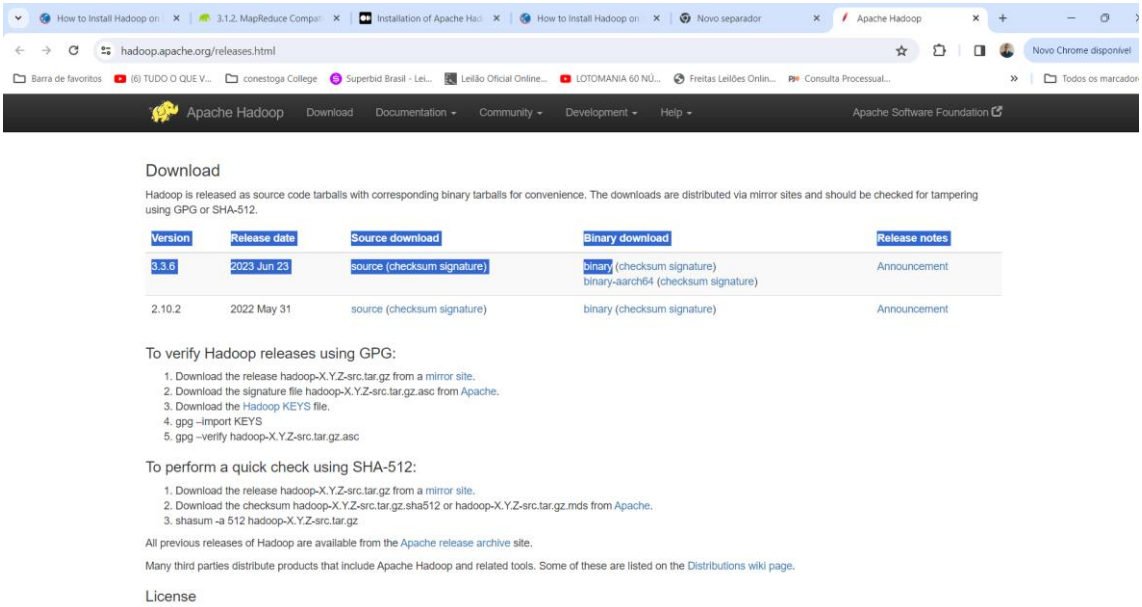


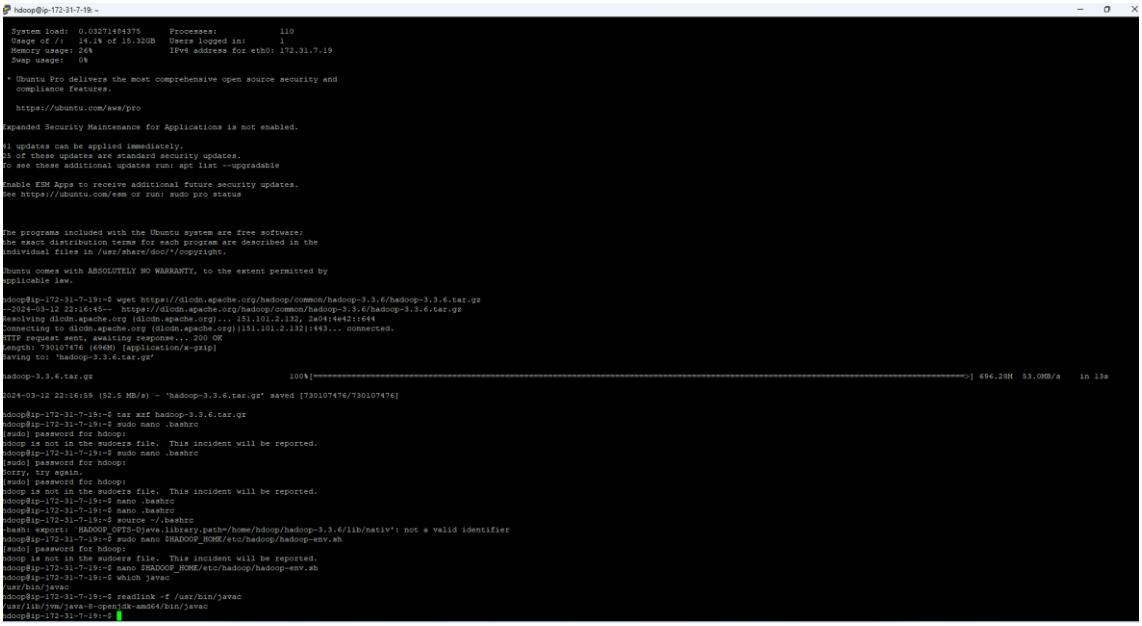
Connect instance using putty.



After updating the system, I installed openjdk-8-jdk -y; after creating the hope user and generating keys, I gave chmod 600.

Go to this website to take the address to download the binary version.



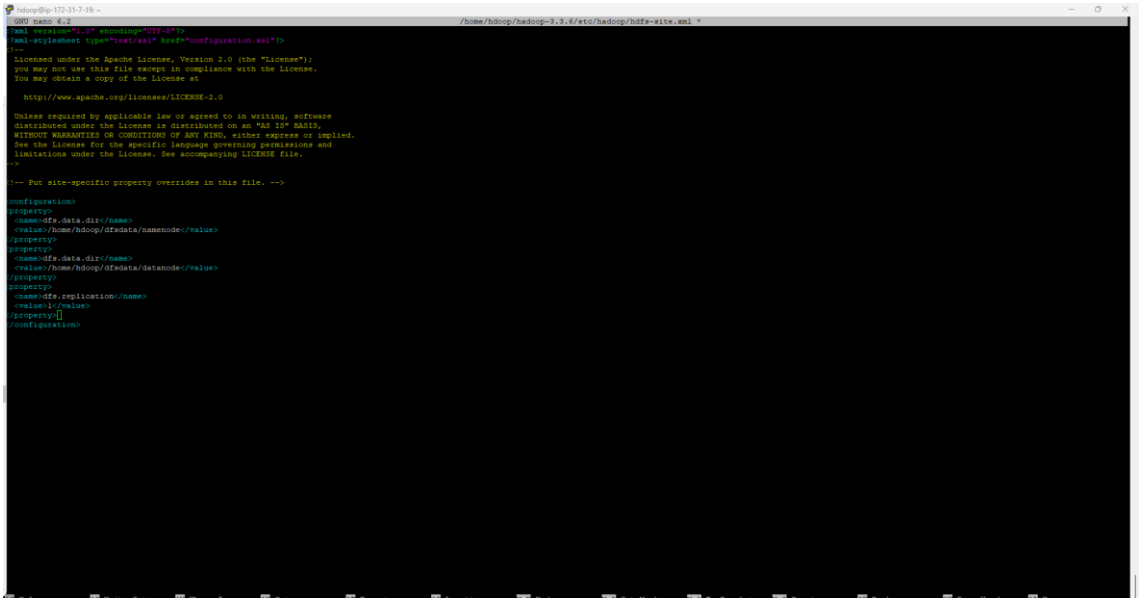


In this step – I just edit file and add this line.

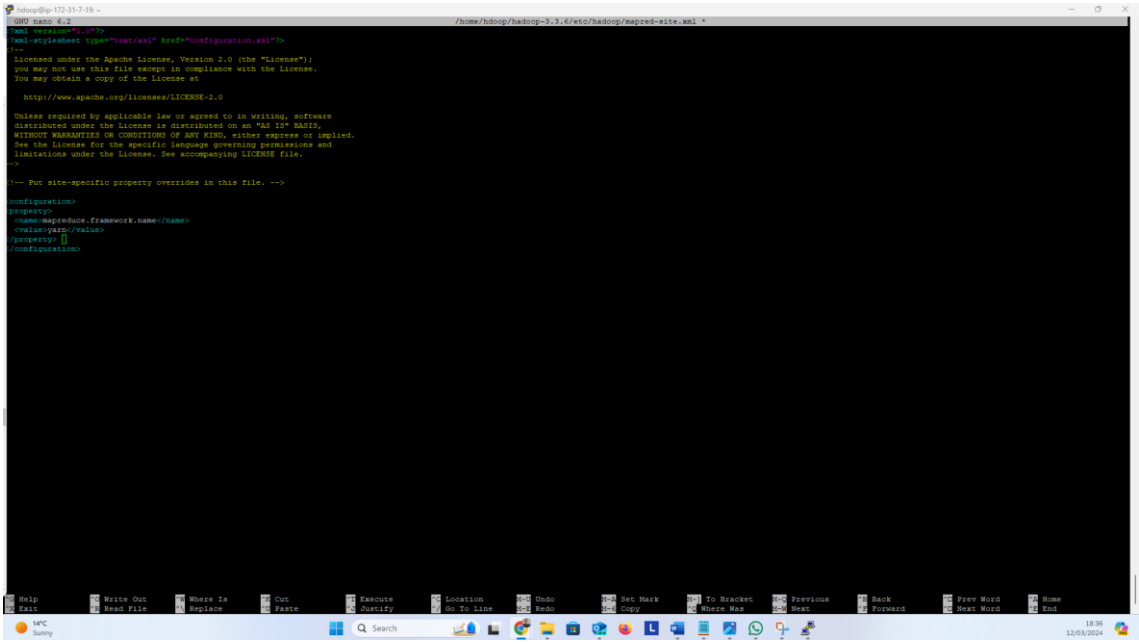
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

In this step, I edit core-site.xml file. And put address my instance aws.

This step I edit hdfs-site.xml file. Configure namenode and datanode storage;

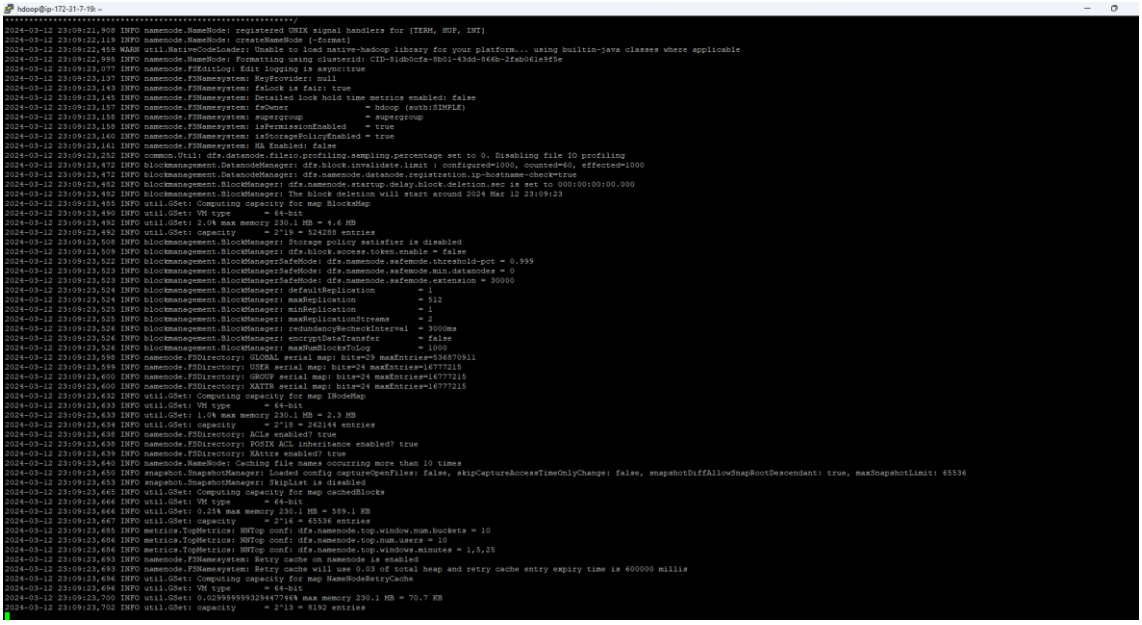


This step edit mapred-site.xml. to define mapreduce values.



In this step, edit yarn-site.xml its content node manager, resource manager, containrs and application master. So I put ip address instance.

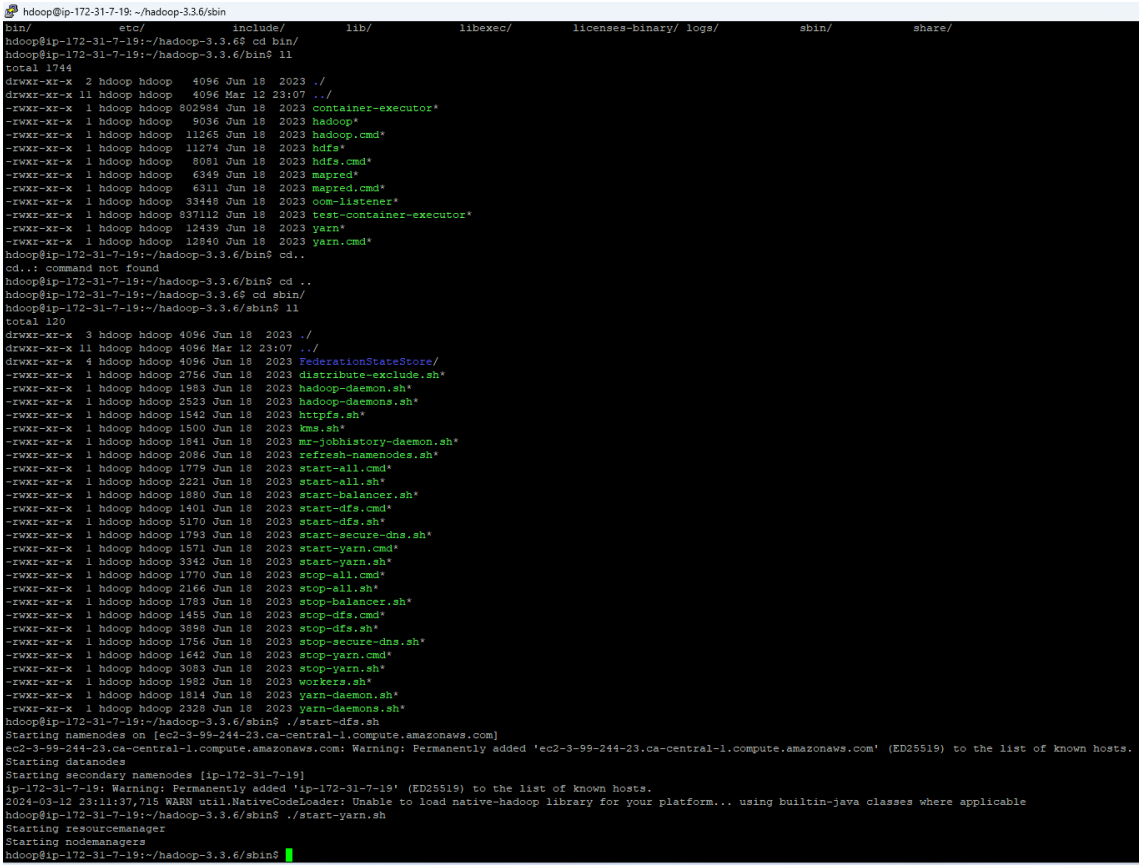
Format hdfs namenode – it´s important before starting Hadoop services for the first time.



Access directory. So, I started the service with the command ./start-dfs.sh

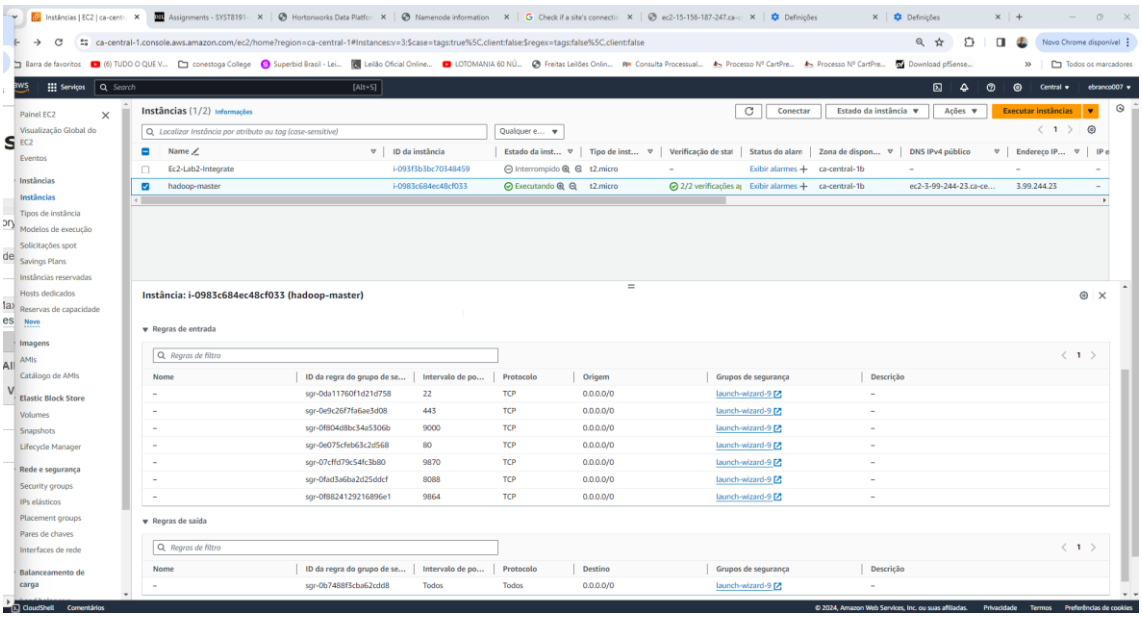
```
hdoop@ip-172-31-7-19: ~/hadoop-3.3.6/sbin
drwxr-xr-x  3 hdoop hdoop  4096 Jun 18  2023 sbin/
drwxr-xr-x  4 hdoop hdoop  4096 Jun 18  2023 share/
hdoop@ip-172-31-7-19:~/hadoop-3.3.6$ cd
bin/            etc/            include/        lib/            libexec/        licenses-binary/ logs/           sbin/           share/
hdoop@ip-172-31-7-19:~/hadoop-3.3.6$ cd bin/
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ ll
total 1744
drwxr-xr-x  2 hdoop hdoop   4096 Jun 18  2023 ./
drwxr-xr-x 11 hdoop hdoop   4096 Mar 12 23:07 ../
-rwxr-xr-x  1 hdoop hdoop 802984 Jun 18  2023 container-executor*
-rwxr-xr-x  1 hdoop hdoop   9036 Jun 18  2023 hadoop*
-rwxr-xr-x  1 hdoop hdoop  11265 Jun 18  2023 hadoop.cmd*
-rwxr-xr-x  1 hdoop hdoop  11274 Jun 18  2023 hdfs*
-rwxr-xr-x  1 hdoop hdoop   8081 Jun 18  2023 hdfs.cmd*
-rwxr-xr-x  1 hdoop hdoop   6349 Jun 18  2023 mapred*
-rwxr-xr-x  1 hdoop hdoop   6311 Jun 18  2023 mapred.cmd*
-rwxr-xr-x  1 hdoop hdoop  33448 Jun 18  2023 oom-listener*
-rwxr-xr-x  1 hdoop hdoop 837112 Jun 18  2023 test-container-executor*
-rwxr-xr-x  1 hdoop hdoop  12439 Jun 18  2023 yarn*
-rwxr-xr-x  1 hdoop hdoop  12840 Jun 18  2023 yarn.cmd*
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd..
cd..: command not found
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd ..
hdoop@ip-172-31-7-19:~/hadoop-3.3.6$ cd sbin/
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$ ll
total 120
drwxr-xr-x  3 hdoop hdoop 4096 Jun 18  2023 ./
drwxr-xr-x 11 hdoop hdoop 4096 Mar 12 23:07 ../
drwxr-xr-x  4 hdoop hdoop 4096 Jun 18  2023 FederationStateStore/
-rwxr-xr-x  1 hdoop hdoop 2756 Jun 18  2023 distribute-exclude.sh*
-rwxr-xr-x  1 hdoop hdoop 1983 Jun 18  2023 hadoop-daemon.sh*
-rwxr-xr-x  1 hdoop hdoop 2523 Jun 18  2023 hadoop-daemons.sh*
-rwxr-xr-x  1 hdoop hdoop 1542 Jun 18  2023 httpfs.sh*
-rwxr-xr-x  1 hdoop hdoop 1500 Jun 18  2023 kms.sh*
-rwxr-xr-x  1 hdoop hdoop 1841 Jun 18  2023 mr-jobhistory-daemon.sh*
-rwxr-xr-x  1 hdoop hdoop 2086 Jun 18  2023 refresh-namenodes.sh*
-rwxr-xr-x  1 hdoop hdoop 1779 Jun 18  2023 start-all.cmd*
-rwxr-xr-x  1 hdoop hdoop 2221 Jun 18  2023 start-all.sh*
-rwxr-xr-x  1 hdoop hdoop 1880 Jun 18  2023 start-balancer.sh*
-rwxr-xr-x  1 hdoop hdoop 1401 Jun 18  2023 start-dfs.cmd*
-rwxr-xr-x  1 hdoop hdoop 5170 Jun 18  2023 start-dfs.sh*
-rwxr-xr-x  1 hdoop hdoop 1793 Jun 18  2023 start-secure-dns.sh*
-rwxr-xr-x  1 hdoop hdoop 1571 Jun 18  2023 start-yarn.cmd*
-rwxr-xr-x  1 hdoop hdoop 3342 Jun 18  2023 start-yarn.sh*
-rwxr-xr-x  1 hdoop hdoop 1770 Jun 18  2023 stop-all.cmd*
-rwxr-xr-x  1 hdoop hdoop 2166 Jun 18  2023 stop-all.sh*
-rwxr-xr-x  1 hdoop hdoop 1783 Jun 18  2023 stop-balancer.sh*
-rwxr-xr-x  1 hdoop hdoop 1455 Jun 18  2023 stop-dfs.cmd*
-rwxr-xr-x  1 hdoop hdoop 3898 Jun 18  2023 stop-dfs.sh*
-rwxr-xr-x  1 hdoop hdoop 1756 Jun 18  2023 stop-secure-dns.sh*
-rwxr-xr-x  1 hdoop hdoop 1642 Jun 18  2023 stop-yarn.cmd*
-rwxr-xr-x  1 hdoop hdoop 3083 Jun 18  2023 stop-yarn.sh*
-rwxr-xr-x  1 hdoop hdoop 1982 Jun 18  2023 workers.sh*
-rwxr-xr-x  1 hdoop hdoop 1814 Jun 18  2023 yarn-daemon.sh*
-rwxr-xr-x  1 hdoop hdoop 2328 Jun 18  2023 yarn-daemons.sh*
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$ ./start-dfs.sh
Starting namenodes on [ec2-3-99-244-23.ca-central-1.compute.amazonaws.com]
ec2-3-99-244-23.ca-central-1.compute.amazonaws.com: Warning: Permanently added 'ec2-3-99-244-23.ca-central-1.compute.amazonaws.com' (ED25519) to the list of known hosts.
Starting datanodes
Starting secondary namenodes [ip-172-31-7-19]
ip-172-31-7-19: Warning: Permanently added 'ip-172-31-7-19' (ED25519) to the list of known hosts.
2024-03-12 23:11:37,715 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$
```

Starting yarn. Using command ./start-yarn.sh

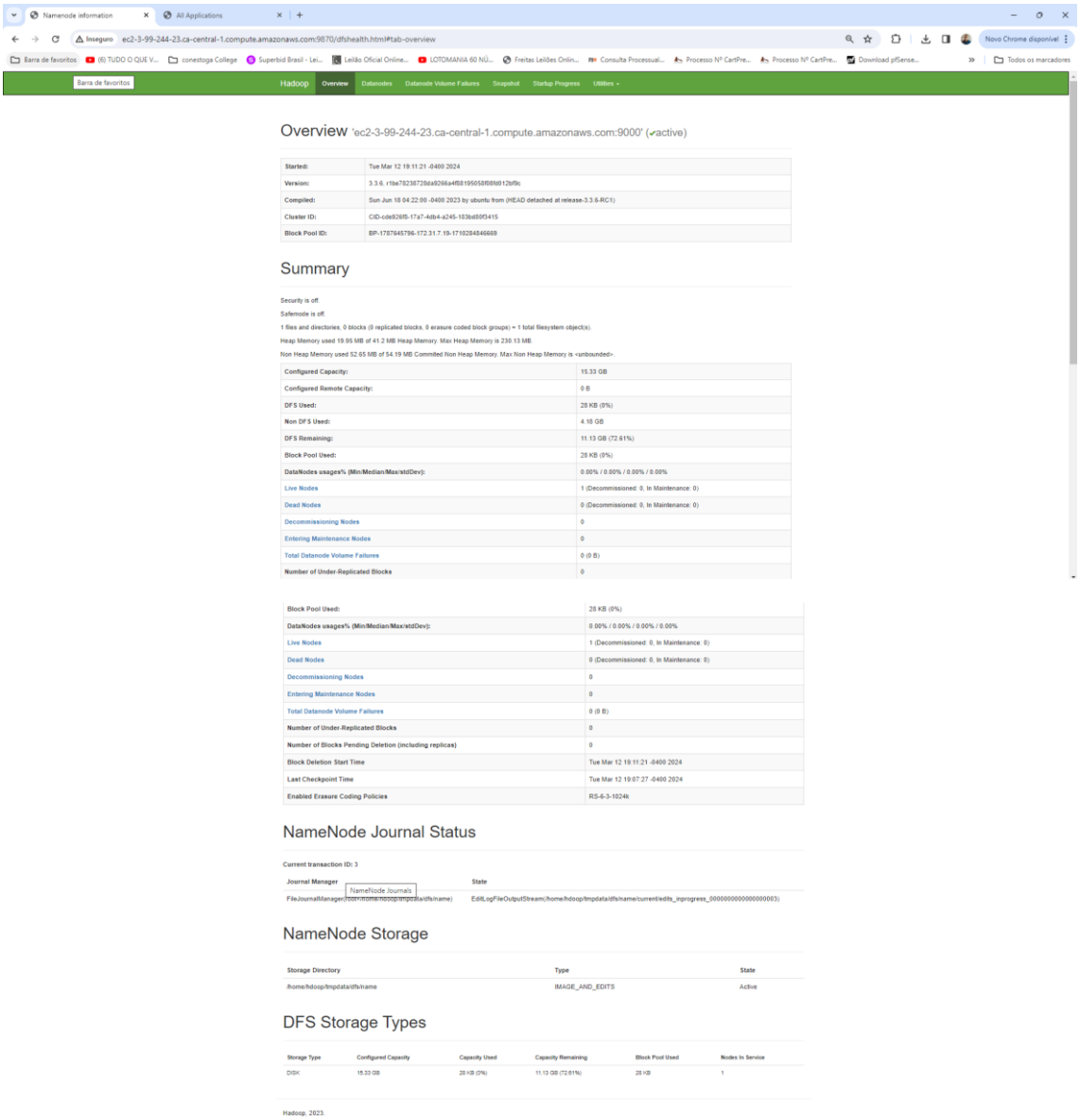


```
hdoop@ip-172-31-7-19: ~/hadoop-3.3.6/sbin
bin/            etc/            include/        lib/            libexec/        licenses-binary/ logs/           sbin/           share/
hdoop@ip-172-31-7-19:~/hadoop-3.3.6$ cd bin/
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ ll
total 1744
drwxr-xr-x  2 hdoop hdoop   4096 Jun 18  2023 ./
drwxr-xr-x 11 hdoop hdoop   4096 Mar 12 23:07 ../
-rwxr-xr-x  1 hdoop hdoop 802984 Jun 18  2023 container-executor*
-rwxr-xr-x  1 hdoop hdoop   9036 Jun 18  2023 hadoop*
-rwxr-xr-x  1 hdoop hdoop  11265 Jun 18  2023 hadoop.cmd*
-rwxr-xr-x  1 hdoop hdoop  11274 Jun 18  2023 hdfs*
-rwxr-xr-x  1 hdoop hdoop   8081 Jun 18  2023 hdfs.cmd*
-rwxr-xr-x  1 hdoop hdoop   6349 Jun 18  2023 mapred*
-rwxr-xr-x  1 hdoop hdoop   6311 Jun 18  2023 mapred.cmd*
-rwxr-xr-x  1 hdoop hdoop  33448 Jun 18  2023 oom-listener*
-rwxr-xr-x  1 hdoop hdoop 837112 Jun 18  2023 test-container-executor*
-rwxr-xr-x  1 hdoop hdoop  12439 Jun 18  2023 yarn*
-rwxr-xr-x  1 hdoop hdoop  12840 Jun 18  2023 yarn.cmd*
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd..
cd..: command not found
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd ..
hdoop@ip-172-31-7-19:~/hadoop-3.3.6$ cd sbin/
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$ ll
total 120
drwxr-xr-x  3 hdoop hdoop 4096 Jun 18  2023 ./
drwxr-xr-x 11 hdoop hdoop 4096 Mar 12 23:07 ../
drwxr-xr-x  4 hdoop hdoop 4096 Jun 18  2023 FederationStateStore/
-rwxr-xr-x  1 hdoop hdoop 2756 Jun 18  2023 distribute-exclude.sh*
-rwxr-xr-x  1 hdoop hdoop 1983 Jun 18  2023 hadoop-daemon.sh*
-rwxr-xr-x  1 hdoop hdoop 2523 Jun 18  2023 hadoop-daemons.sh*
-rwxr-xr-x  1 hdoop hdoop 1542 Jun 18  2023 httpfs.sh*
-rwxr-xr-x  1 hdoop hdoop 1500 Jun 18  2023 kms.sh*
-rwxr-xr-x  1 hdoop hdoop 1841 Jun 18  2023 mr-jobhistory-daemon.sh*
-rwxr-xr-x  1 hdoop hdoop 2086 Jun 18  2023 refresh-namenodes.sh*
-rwxr-xr-x  1 hdoop hdoop 1779 Jun 18  2023 start-all.cmd*
-rwxr-xr-x  1 hdoop hdoop 2221 Jun 18  2023 start-all.sh*
-rwxr-xr-x  1 hdoop hdoop 1880 Jun 18  2023 start-balancer.sh*
-rwxr-xr-x  1 hdoop hdoop 1401 Jun 18  2023 start-dfs.cmd*
-rwxr-xr-x  1 hdoop hdoop 5170 Jun 18  2023 start-dfs.sh*
-rwxr-xr-x  1 hdoop hdoop 1793 Jun 18  2023 start-secure-dns.sh*
-rwxr-xr-x  1 hdoop hdoop 1571 Jun 18  2023 start-yarn.cmd*
-rwxr-xr-x  1 hdoop hdoop 3342 Jun 18  2023 start-yarn.sh*
-rwxr-xr-x  1 hdoop hdoop 1770 Jun 18  2023 stop-all.cmd*
-rwxr-xr-x  1 hdoop hdoop 2166 Jun 18  2023 stop-all.sh*
-rwxr-xr-x  1 hdoop hdoop 1783 Jun 18  2023 stop-balancer.sh*
-rwxr-xr-x  1 hdoop hdoop 1455 Jun 18  2023 stop-dfs.cmd*
-rwxr-xr-x  1 hdoop hdoop 3898 Jun 18  2023 stop-dfs.sh*
-rwxr-xr-x  1 hdoop hdoop 1756 Jun 18  2023 stop-secure-dns.sh*
-rwxr-xr-x  1 hdoop hdoop 1642 Jun 18  2023 stop-yarn.cmd*
-rwxr-xr-x  1 hdoop hdoop 3083 Jun 18  2023 stop-yarn.sh*
-rwxr-xr-x  1 hdoop hdoop 1982 Jun 18  2023 workers.sh*
-rwxr-xr-x  1 hdoop hdoop 1814 Jun 18  2023 yarn-daemon.sh*
-rwxr-xr-x  1 hdoop hdoop 2328 Jun 18  2023 yarn-daemons.sh*
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$ ./start-dfs.sh
Starting namenodes on [ec2-3-99-244-23.ca-central-1.compute.amazonaws.com]
ec2-3-99-244-23.ca-central-1.compute.amazonaws.com: Warning: Permanently added 'ec2-3-99-244-23.ca-central-1.compute.amazonaws.com' (ED25519) to the list of known hosts.
Starting datanodes
Starting secondary namenodes [ip-172-31-7-19]
ip-172-31-7-19: Warning: Permanently added 'ip-172-31-7-19' (ED25519) to the list of known hosts.
2024-03-12 23:11:37,715 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$
```

Open ports in security groups



Access Hadoop



Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

Overview 'ec2-3-99-244-23.ca-central-1.compute.amazonaws.com:9000' (active)

Started:	Tue Mar 12 19:11:21 -0400 2024
Version:	3.3.6, r1be7823872da92664af881950580fd012bf9c
Compiled:	Sun Jun 18 04:22:00 -0400 2023 by ubuntu from (HEAD detached at release-3.3.6-RC1)
Cluster ID:	CID-cde926f5-17a7-4db4-a245-183bd90f3415
Block Pool ID:	BP-1787645796-172.31.7.19-1710284846669

Summary

Security is off.

Safemode is off.

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

Heap Memory used 19.95 MB of 41.2 MB Heap Memory. Max Heap Memory is 230.13 MB.

Non Heap Memory used 52.65 MB of 54.19 MB Commited Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	15.33 GB
Configured Remote Capacity:	0 B
DFS Used:	28 KB (0%)
Non DFS Used:	4.18 GB
DFS Remaining:	11.13 GB (72.61%)
Block Pool Used:	28 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0

Block Pool Used:	28 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion (including replicas)	0
Block Deletion Start Time	Tue Mar 12 19:11:21 -0400 2024
Last Checkpoint Time	Tue Mar 12 19:07:27 -0400 2024
Enabled Erasure Coding Policies	RS-6-3-1024k

NameNode Journal Status

Current transaction ID: 3

Journal Manager	State
FileJournalManager(root=/home/hdoop/tmpdata/dfs/name)	EditLogFileOutputStream(/home/hdoop/tmpdata/dfs/name/current/edits_inprogress_0000000000000000003)

NameNode Journals

NameNode Storage

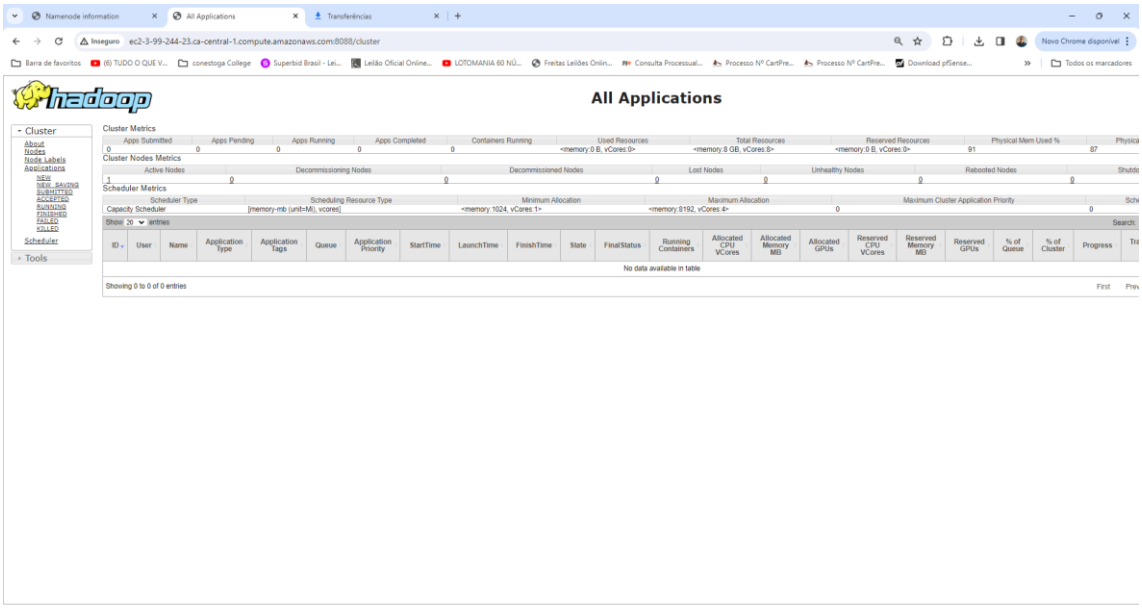
Storage Directory	Type	State
/home/hdoop/tmpdata/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

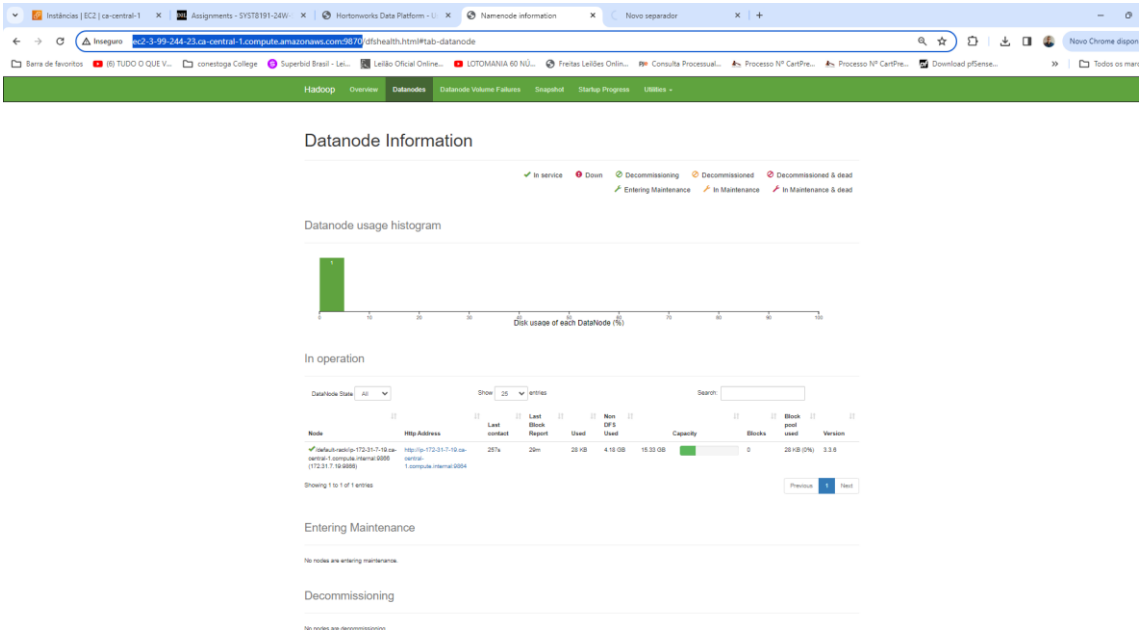
Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	15.33 GB	28 KB (0%)	11.13 GB (72.61%)	28 KB	1

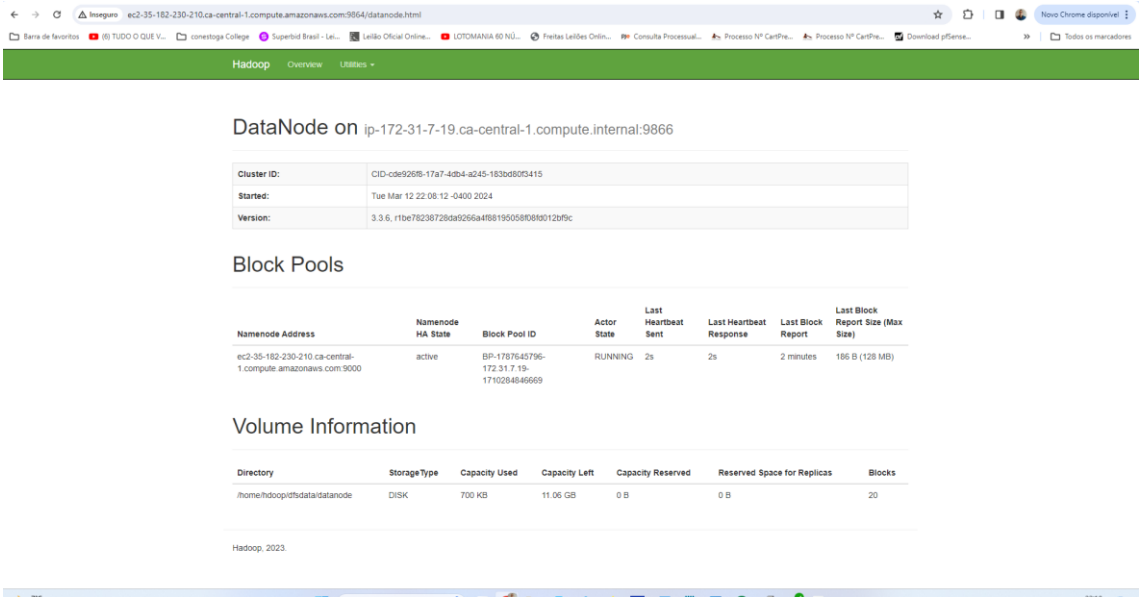
Hadoop, 2023.

Accessing web Hadoop



datanode





Node running;