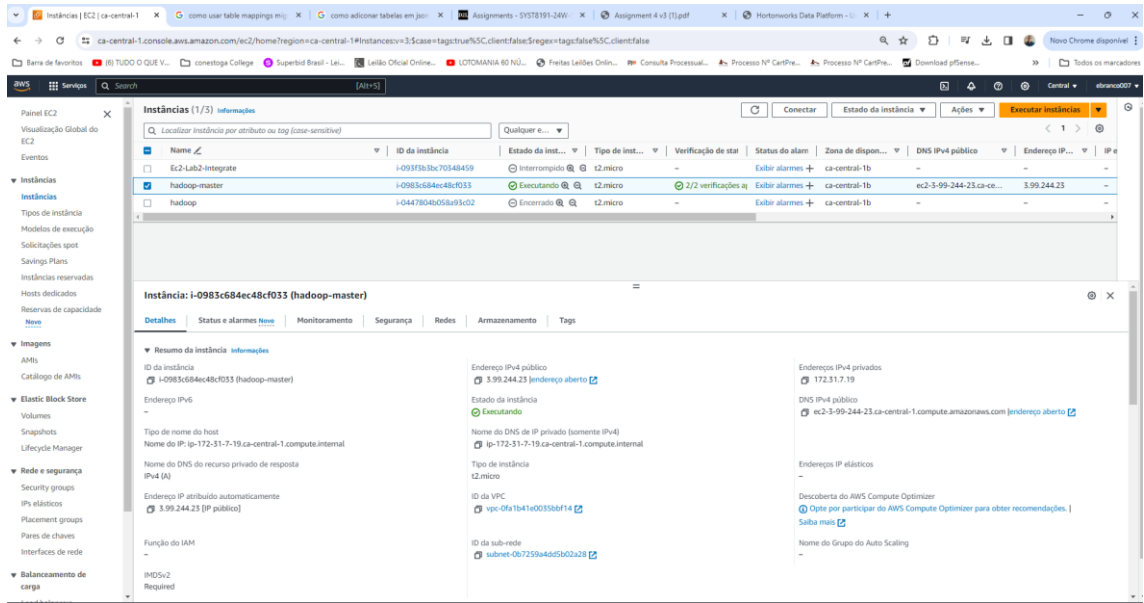
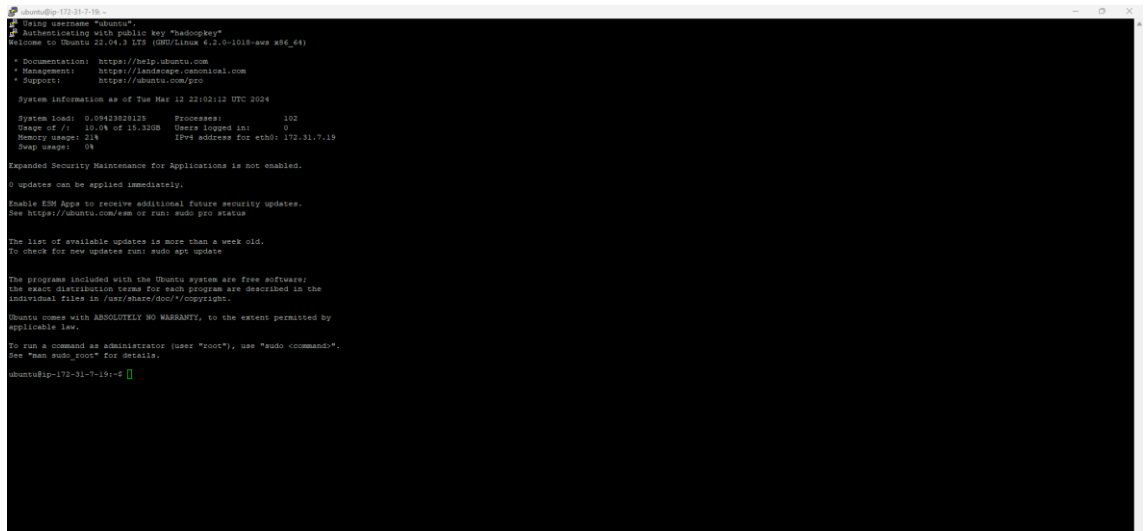


Deploy Hadoop on AWS

1- Deploy Ec2 instance on AWS.

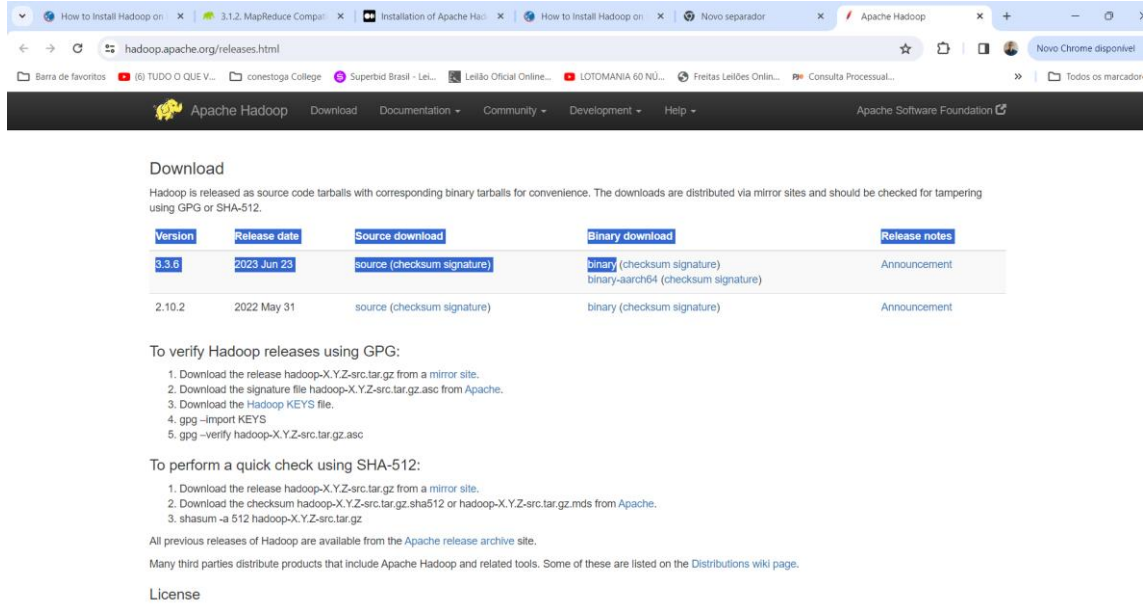


Connect instance using putty.



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.



The screenshot shows the Apache Hadoop website's releases page. The page title is "Download" and it provides information about downloading Hadoop releases. It includes a table with columns for Version, Release date, Source download, Binary download, and Release notes. The table lists two versions: 3.3.6 (released 2023 Jun 23) and 2.10.2 (released 2022 May 31). For version 3.3.6, the source download is "source (checksum signature)" and the binary download is "binary (checksum signature)" and "binary-aarch64 (checksum signature)". The release notes for 3.3.6 are "Announcement". For version 2.10.2, the source download is "source (checksum signature)" and the binary download is "binary (checksum signature)". The release notes for 2.10.2 are "Announcement". Below the table, there are instructions on how to verify Hadoop releases using GPG and SHA-512, and a note about third-party distributors.

Version	Release date	Source download	Binary download	Release notes
3.3.6	2023 Jun 23	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)	Announcement
2.10.2	2022 May 31	source (checksum signature)	binary (checksum signature)	Announcement

```
hadoop@ip-172-31-7-19:~$
System load: 0.027/4.943/7.0      Processes: 110
Mem usage: 14.4k of 16.300M      Swap usage: 0k
Distro: rockylinux/8.7
* Ubuntu Pro delivers the most comprehensive open source security and
  compliance features.
  https://ubuntu.com/esm/pro
Expanded Security Maintenance for Applications is not enabled.
All updates can be applied immediately.
25 of these updates are standard security updates.
To see these additional updates run apt list --upgradable
Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*/copyright.
Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
available law.

hadoop@ip-172-31-7-19:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
2024-03-12 22:16:45-- https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 204:442:1:644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 70107476 (66M) [application/x-gzip]
Saving to: 'hadoop-3.3.6.tar.gz'

hadoop-3.3.6.tar.gz 100%[#####] 696.23M 53.00M/s in 13s

2024-03-12 22:16:59 (52.5 MB/s) - 'hadoop-3.3.6.tar.gz' saved [780107476/780107476]

hadoop@ip-172-31-7-19:~$ tar xzf hadoop-3.3.6.tar.gz
hadoop@ip-172-31-7-19:~$ sudo nano .bashrc
[sudo] password for hadoop:
hadoop@ip-172-31-7-19:~$ sudo nano .bashrc
[sudo] password for hadoop:
hadoop@ip-172-31-7-19:~$ sudo nano .bashrc
[sudo] password for hadoop:
hadoop@ip-172-31-7-19:~$ source ~/.bashrc
bash: export: `HADOOP_OPTS=java.library.path=/home/hadoop/hadoop-3.3.6/lib/native`: not a valid identifier
hadoop@ip-172-31-7-19:~$ sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
[sudo] password for hadoop:
hadoop@ip-172-31-7-19:~$ nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
hadoop@ip-172-31-7-19:~$ which javac
/usr/bin/javac
hadoop@ip-172-31-7-19:~$ readlink -f /usr/bin/javac
/usr/lib/jvm/java-8-openjdk-amd64/bin/javac
hadoop@ip-172-31-7-19:~$
```

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

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



```
hadoop@ip-172-31-7-19: ~
hadoop 6.2 /home/hadoop/hadoop-3.3.6/etc/hadoop/hadoop-env.sh

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

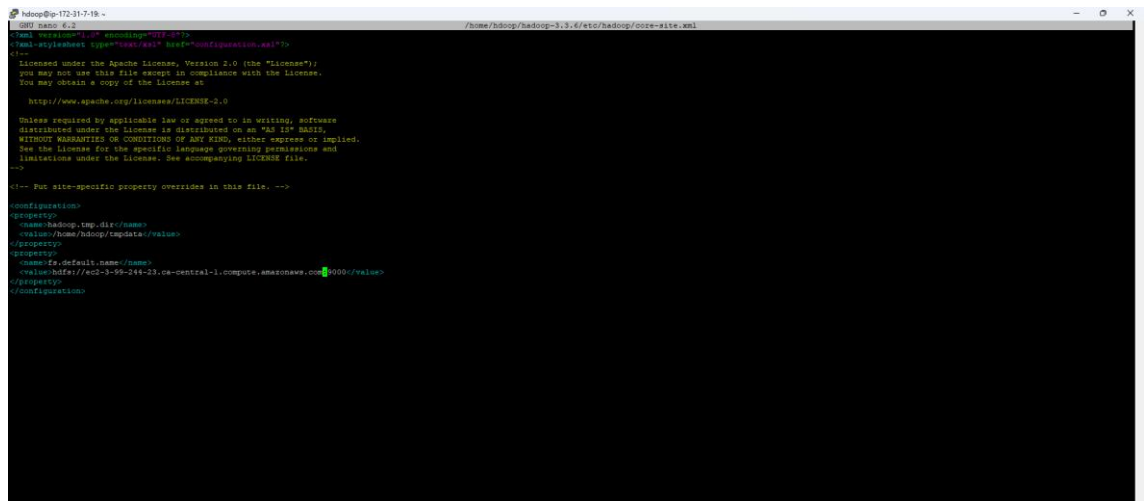
Unless required by applicable law or agreed to in writing, software
distributed under the License is distributed on an "AS IS" BASIS,
WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
See the License for the specific language governing permissions and
limitations under the License.

# Set Hadoop-specific environment variables here.

##
## THIS FILE ACTS AS THE MASTER FILE FOR ALL HADOOP PROJECTS.
## SETTINGS HERE WILL BE READ BY ALL HADOOP COMMANDS, THEREFORE,
## ONE CAN USE THIS FILE TO SET JAVA, HDFS, AND YARN/SOFSX
## CONFIGURATION OPTIONS INSTEAD OF xxx-env.sh.
##
## Precedence rules!
## [yarn-env.sh|hdfs-env.sh] > hadoop-env.sh > hard-coded defaults
##
## (HADO_AYS|HDFS_AYS) > HADOOP_AYS > hard-coded defaults
##
## Many of the options here are built from the perspective that users
## may want to provide OVERRIDING values on the command line.
## For example:
##
##export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
## JAVA_HOME=/usr/java/testing/java-8
##
## Therefore, the vast majority (BUT NOT ALL) of these defaults
## are configured for substitution and not export. If export
## is preferable, modify this file accordingly.
##
###
### Generic settings for HADOOP
###
## Technically, the only required environment variable is JAVA_HOME.
## All others are optional. However, the defaults are probably not
## preferred. Many sites configure these options outside of Hadoop.
## even as in /etc/profile.d
##
## The Java implementation to use. By default, this environment
## variable is REQUIRED on ALL platforms except OS X!
## export JAVA_HOME=
##
## Location of Hadoop. By default, Hadoop will attempt to determine
## this location based upon its execution path.
## export HADOOP_HOME=/usr/lib/jvm/java-8-openjdk-amd64/

##
## Help Write Out Where Is Out Execute Location Read All Lines Set Mark To Bracket Previous Back Forward First Word Next Word Home
## Exit Read File Read File Replace Paste Quit Go To Line Redo Copy Where Mas Next Next Next Word End
```

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



```
hadoop@ip-172-31-7-19: ~
hadoop 6.2 /home/hadoop/hadoop-3.3.6/etc/hadoop/core-site.xml

<?xml version="1.0" encoding="UTF-8"?>
<configuration>
  <property>
    <name>fs.defaultFS</name>
    <value>s3a://aws-logs-911111111111-us-east-1:bucket:aws-logs-911111111111-us-east-1</value>
  </property>
  </configuration>
</xml>

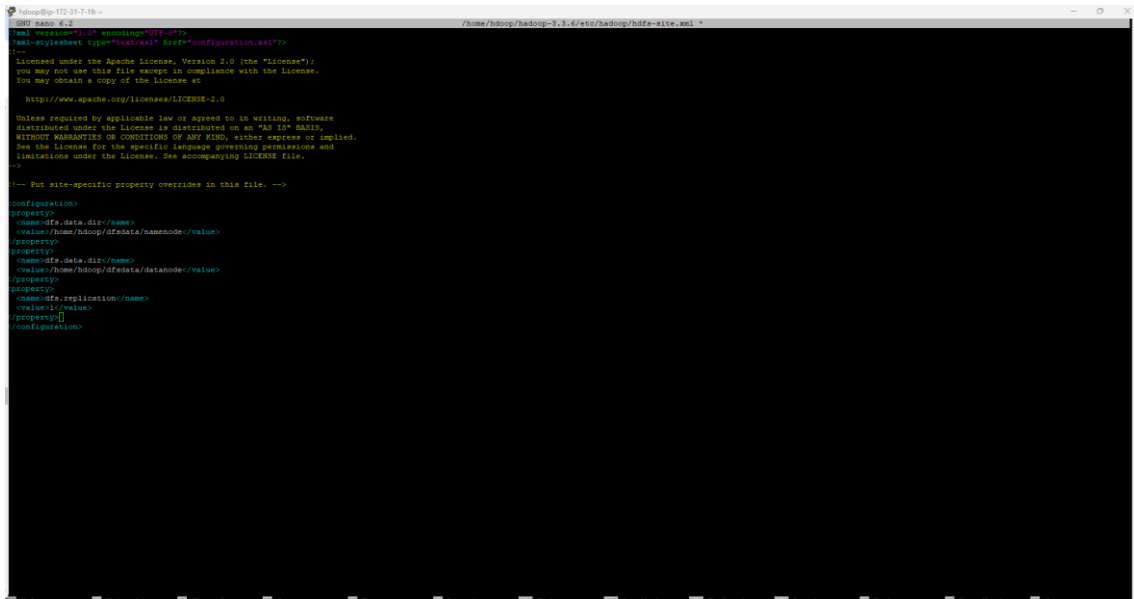
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at
http://www.apache.org/licenses/LICENSE-2.0

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>hadoop.tmp.dir</name>
    <value>/home/hadoop/tmpdata</value>
  </property>
  <property>
    <name>fs.defaultFS</name>
    <value>hdfs://ec2-3-66-244-23-ca-central-1.compute.amazonaws.com:9000/</value>
  </property>
</configuration>
```

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



```
hadoop@ip-172-31-7-19: ~$ cat /etc/hadoop/hdfs-site.xml
<?xml version="1.0" encoding="UTF-8"?>
<!--
 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.data.dir</name>
<value>/home/hadoop/dfsdata/namenode</value>
</property>
<property>
<name>dfs.data.dir</name>
<value>/home/hadoop/dfsdata/datanode</value>
</property>
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
</configuration>
```

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



```
hadoop@ip-172-31-7-19: ~$ cat /etc/hadoop/mapred-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.
 -->
<!-- Put site-specific property overrides in this file. -->
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
</property>
</configuration>
```

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

```
hadoop@ip-172-31-7-19 ~
└─$ cat /etc/hadoop/yarn-site.xml
<!--
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.nodemanager.aux-service.mapreduce_shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
<name>yarn.resourcemanager.hostname</name>
<value>ec2-3-93-244-23.ca-central-1.compute.amazonaws.com</value>
</property>
<property>
<name>yarn.acl.enable</name>
<value>false</value>
</property>
<property>
<name>yarn.nodemanager.env-whitelist</name>
<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>

```

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

```
hadoop@ip-172-31-7-19 ~
└─$ hdfs namenode -format
2024-03-12 23:09:21,908 INFO namenode.NameNode: registered UNIX signal handlers for [TERM, SIGHUP, INT]
2024-03-12 23:09:22,119 INFO namenode.NameNode: createNameNode [-format]
2024-03-12 23:09:22,458 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-03-12 23:09:22,985 INFO namenode.NameNode: Formatting using clusterid: CID-81b0dcfa-b901-43dd-96db-2fab01e5fde
2024-03-12 23:09:23,077 INFO namenode.FSUtilLog: Edit logging is asynchronous
2024-03-12 23:09:23,143 INFO namenode.FSNamesystem: fslock is fail: true
2024-03-12 23:09:23,143 INFO namenode.FSNamesystem: Detailed lock hold time metrics enabled: false
2024-03-12 23:09:23,157 INFO namenode.FSNamesystem: fsOwner = hadoop (auth:SIMPLE)
2024-03-12 23:09:23,158 INFO namenode.FSNamesystem: supergroup = supergroup
2024-03-12 23:09:23,158 INFO namenode.FSNamesystem: isHadoopConfEnabled = true
2024-03-12 23:09:23,160 INFO namenode.FSNamesystem: isStoragePolicyEnabled = true
2024-03-12 23:09:23,161 INFO namenode.FSNamesystem: HA Enabled: false
2024-03-12 23:09:23,252 INFO common.Util: dfs.datanode.fileio.profiling.sampling.percentage set to 0. Disabling file IO profiling
2024-03-12 23:09:23,472 INFO blockmanagement.DataNodeManager: dfs.block.invalidate.limit: configured=1000, counted=0, effective=1000
2024-03-12 23:09:23,482 INFO blockmanagement.BlockManager: dfs.namenode.startup.delay.block.deletion.sec is set to 000:00:00:00:00.000
2024-03-12 23:09:23,485 INFO blockmanagement.BlockManager: The block deletion will start around 2024 Mar 12 23:09:23
2024-03-12 23:09:23,485 INFO util.GSet: Computing capacity for map BlockMap
2024-03-12 23:09:23,490 INFO util.GSet: VM type = 64-bit
2024-03-12 23:09:23,496 INFO util.GSet: 2.0B max memory 230.1 MB = 4.6 MB
2024-03-12 23:09:23,492 INFO util.GSet: capacity = 2^18 = 524288 entries
2024-03-12 23:09:23,508 INFO blockmanagement.BlockManager: Storage policy serializer is disabled
2024-03-12 23:09:23,522 INFO blockmanagement.BlockManagerSafeMode: dfs.block.access.token.enable = false
2024-03-12 23:09:23,522 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safeMode.thresholdPct = 0.999
2024-03-12 23:09:23,523 INFO blockmanagement.BlockManagerSafeMode: dfs.namenode.safeMode.min.datanodes = 0
2024-03-12 23:09:23,524 INFO blockmanagement.BlockManager: defaultReplication = 1
2024-03-12 23:09:23,524 INFO blockmanagement.BlockManager: maxReplication = 512
2024-03-12 23:09:23,524 INFO blockmanagement.BlockManager: minReplication = 1
2024-03-12 23:09:23,524 INFO blockmanagement.BlockManager: maxReplicationStreams = 2
2024-03-12 23:09:23,524 INFO blockmanagement.BlockManager: redundancyCheckInterval = 300ms
2024-03-12 23:09:23,524 INFO blockmanagement.BlockManager: corruptDataTransfer = false
2024-03-12 23:09:23,524 INFO blockmanagement.BlockManager: maxBlockLogToLog = 1000
2024-03-12 23:09:23,524 INFO namenode.FSDirectory: GLOBAL serial map: bit=29 maxEntries=546070911
2024-03-12 23:09:23,524 INFO namenode.FSDirectory: USER serial map: bit=24 maxEntries=16777215
2024-03-12 23:09:23,600 INFO namenode.FSDirectory: GROUP serial map: bit=24 maxEntries=16777215
2024-03-12 23:09:23,600 INFO namenode.FSDirectory: MIXED serial map: bit=24 maxEntries=16777215
2024-03-12 23:09:23,630 INFO util.GSet: Computing capacity for map IBlockMap
2024-03-12 23:09:23,633 INFO util.GSet: VM type = 64-bit
2024-03-12 23:09:23,633 INFO util.GSet: 1.0B max memory 230.1 MB = 2.3 MB
2024-03-12 23:09:23,633 INFO util.GSet: capacity = 2^18 = 262144 entries
2024-03-12 23:09:23,636 INFO namenode.FSDirectory: ACLs enabled? true
2024-03-12 23:09:23,636 INFO namenode.FSDirectory: POSIX ACL inheritance enabled? true
2024-03-12 23:09:23,640 INFO namenode.NameNode: Caching file name occurring more than 10 times
2024-03-12 23:09:23,650 INFO snapshott.SnapshottManager: Loaded config captureOpenFiles: false, skipCaptureAccessTimeOnlyChang: false, snapshottDiffLocalSnapRootDependant: true, maxSnapshottLimit: 65536
2024-03-12 23:09:23,665 INFO util.GSet: Computing capacity for map cachedBlocks
2024-03-12 23:09:23,667 INFO util.GSet: VM type = 64-bit
2024-03-12 23:09:23,667 INFO util.GSet: 0.25B max memory 230.1 MB = 59.1 MB
2024-03-12 23:09:23,667 INFO util.GSet: capacity = 2^16 = 65536 entries
2024-03-12 23:09:23,686 INFO metrics.TopMetrics: Hadoop conf: dfs.namenode.top.window.num.buckets = 10
2024-03-12 23:09:23,686 INFO metrics.TopMetrics: Hadoop conf: dfs.namenode.top.num.users = 10
2024-03-12 23:09:23,686 INFO metrics.TopMetrics: Hadoop conf: dfs.namenode.top.window.minutes = 1.5,25
2024-03-12 23:09:23,686 INFO namenode.FSNamesystem: Retry cache on namenode is enabled
2024-03-12 23:09:23,690 INFO namenode.FSNamesystem: Retry cache will use 0.03 of total heap and retry cache entry expiry time is 600000 millis
2024-03-12 23:09:23,696 INFO util.GSet: Computing capacity for map NameNodeRetryCache
2024-03-12 23:09:23,696 INFO util.GSet: VM type = 64-bit
2024-03-12 23:09:23,700 INFO util.GSet: 0.2999999923477169 max memory 230.1 MB = 70.7 MB
2024-03-12 23:09:23,700 INFO util.GSet: capacity = 2^13 = 8192 entries

```

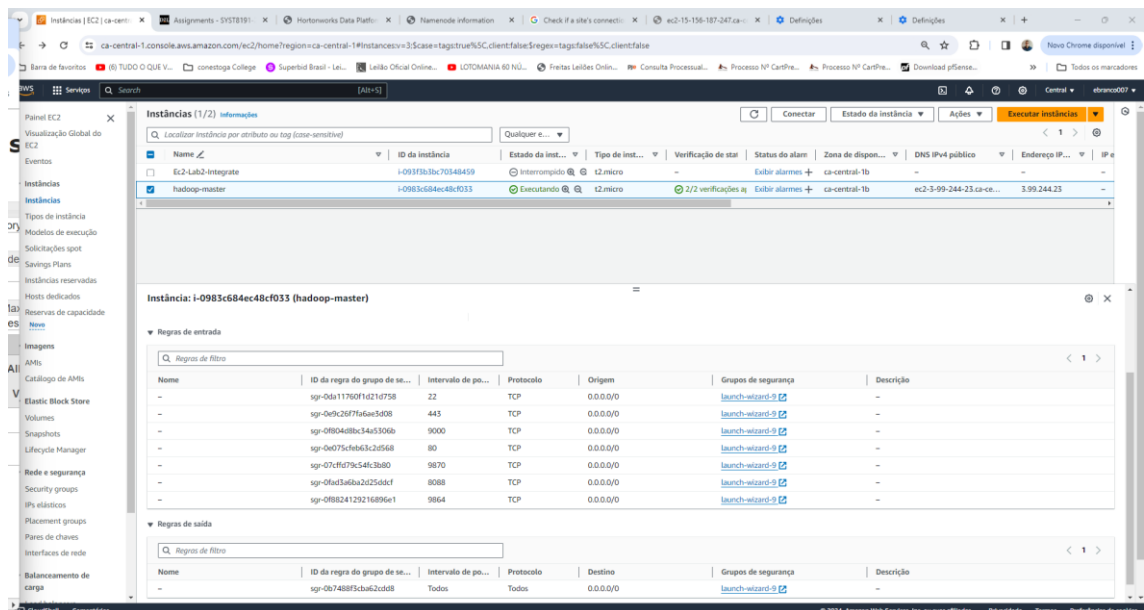
Access directory. So, I started the service with the command ./start-dfs.sh

```
hdoo@ip-172-31-7-19:~/hadoop-3.3.6/sbin
drwxr-xr-x 4 hdoop hdoop 4096 Jun 18 2023 sbin/
drwxr-xr-x 4 hdoop hdoop 4096 Jun 18 2023 share/
hdoo@ip-172-31-7-19:~/hadoop-3.3.6$ cd
bin/          etc/          include/      lib/          libexec/     licenses-binary/ logs/        sbin/        share/
hdoo@ip-172-31-7-19:~/hadoop-3.3.6$ cd bin/
hdoo@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 11263 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 39448 Jun 18 2023 oom-lsrecr*
-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*
hdoo@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd ..
cd: no command not found
hdoo@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd ..
hdoo@ip-172-31-7-19:~/hadoop-3.3.6$ cd sbin/
hdoo@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 1933 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 1799 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 1485 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 1592 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*
hdoo@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
hdoo@ip-172-31-7-19:~/hadoop-3.3.6/sbin$
```

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

```
hadoop@ip-172-31-7-19: ~/hadoop-3.3.6/sbin
bin/          etc/          include/     lib/         libexec/    licenses-binary/ logs/        sbin/        share/
hadoop@ip-172-31-7-19:~/hadoop-3.3.6$ cd bin/
hadoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ ll
total 1744
drwxr-xr-x 2 hadoop hadoop  4096 Jun 18 2023 ./
drwxr-xr-x 11 hadoop hadoop  4096 Mar 12 23:07 ../
-rwxr-xr-x 1 hadoop hadoop 802984 Jun 18 2023 container-executor*
-rwxr-xr-x 1 hadoop hadoop  9036 Jun 18 2023 hadoop*
-rwxr-xr-x 1 hadoop hadoop 11265 Jun 18 2023 hadoop.cmd*
-rwxr-xr-x 1 hadoop hadoop 11274 Jun 18 2023 hdfs*
-rwxr-xr-x 1 hadoop hadoop  8081 Jun 18 2023 hdfs.cmd*
-rwxr-xr-x 1 hadoop hadoop  6349 Jun 18 2023 mapred*
-rwxr-xr-x 1 hadoop hadoop  6311 Jun 18 2023 mapred.cmd*
-rwxr-xr-x 1 hadoop hadoop 33448 Jun 18 2023 oom-listener*
-rwxr-xr-x 1 hadoop hadoop 837112 Jun 18 2023 test-container-executor*
-rwxr-xr-x 1 hadoop hadoop 12439 Jun 18 2023 yarn*
-rwxr-xr-x 1 hadoop hadoop 12840 Jun 18 2023 yarn.cmd*
hadoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd..
cd..: command not found
hadoop@ip-172-31-7-19:~/hadoop-3.3.6/bin$ cd ..
hadoop@ip-172-31-7-19:~/hadoop-3.3.6$ cd sbin/
hadoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$ ll
total 120
drwxr-xr-x 3 hadoop hadoop  4096 Jun 18 2023 ./
drwxr-xr-x 11 hadoop hadoop  4096 Mar 12 23:07 ../
drwxr-xr-x 4 hadoop hadoop  4096 Jun 18 2023 FederationStateStore/
-rwxr-xr-x 1 hadoop hadoop 2786 Jun 18 2023 distribute-exclude.sh*
-rwxr-xr-x 1 hadoop hadoop 1983 Jun 18 2023 hadoop-daemon.sh*
-rwxr-xr-x 1 hadoop hadoop 2523 Jun 18 2023 hadoop-daemons.sh*
-rwxr-xr-x 1 hadoop hadoop 1542 Jun 18 2023 httpfs.sh*
-rwxr-xr-x 1 hadoop hadoop 1500 Jun 18 2023 kms.sh*
-rwxr-xr-x 1 hadoop hadoop 1841 Jun 18 2023 mr-jobhistory-daemon.sh*
-rwxr-xr-x 1 hadoop hadoop 2086 Jun 18 2023 refresh-namenodes.sh*
-rwxr-xr-x 1 hadoop hadoop 1779 Jun 18 2023 start-all.cmd*
-rwxr-xr-x 1 hadoop hadoop 2221 Jun 18 2023 start-all.sh*
-rwxr-xr-x 1 hadoop hadoop 1880 Jun 18 2023 start-balancer.sh*
-rwxr-xr-x 1 hadoop hadoop 1401 Jun 18 2023 start-dfs.cmd*
-rwxr-xr-x 1 hadoop hadoop 5170 Jun 18 2023 start-dfs.sh*
-rwxr-xr-x 1 hadoop hadoop 1793 Jun 18 2023 start-secure-dns.sh*
-rwxr-xr-x 1 hadoop hadoop 1871 Jun 18 2023 start-yarn.cmd*
-rwxr-xr-x 1 hadoop hadoop 3342 Jun 18 2023 start-yarn.sh*
-rwxr-xr-x 1 hadoop hadoop 1770 Jun 18 2023 stop-all.cmd*
-rwxr-xr-x 1 hadoop hadoop 2166 Jun 18 2023 stop-all.sh*
-rwxr-xr-x 1 hadoop hadoop 1783 Jun 18 2023 stop-balancer.sh*
-rwxr-xr-x 1 hadoop hadoop 1485 Jun 18 2023 stop-dfs.cmd*
-rwxr-xr-x 1 hadoop hadoop 3898 Jun 18 2023 stop-dfs.sh*
-rwxr-xr-x 1 hadoop hadoop 1756 Jun 18 2023 stop-secure-dns.sh*
-rwxr-xr-x 1 hadoop hadoop 1642 Jun 18 2023 stop-yarn.cmd*
-rwxr-xr-x 1 hadoop hadoop 3083 Jun 18 2023 stop-yarn.sh*
-rwxr-xr-x 1 hadoop hadoop 1982 Jun 18 2023 workers.sh*
-rwxr-xr-x 1 hadoop hadoop 1814 Jun 18 2023 yarn-daemon.sh*
-rwxr-xr-x 1 hadoop hadoop 2328 Jun 18 2023 yarn-daemons.sh*
hadoop@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
hadoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@ip-172-31-7-19:~/hadoop-3.3.6/sbin$
```

Open ports in security groups



Access Hadoop

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-f16e782387736a02664468195958086912a9fc
Completed:	Sun Jun 16 04:22:00 -0400 2023 by ubuntu from (HEAD detached at release-3.3.6.RC1)
Cluster ID:	CD-c683265-1747-4684-a245-183bd909415
Block Pool ID:	BP-1781945796-172.31.7.19-1718234846689

Summary

Security is off
 Submitters is off
 1 file 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 62.55 MB of 54.19 MB Committed 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 (72.81%)
Block Pool Used:	28 KB (0%)
DataNodes usage%, (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 usage%, (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	NameNode Journal	State
FileJournalManager	file:///home/hadoop/data/dfs/name	EdtLgFileOutputStream/home/hadoop/data/dfs/name/current/edts_inprogress_000000000000000000000000

NameNode Storage

Storage Directory	Type	State
home/hadoop/tmp/data/dfs/name	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DDK	15.33 GB	28 KB (0%)	11.13 GB (72.81%)	28 KB	1

Hadoop: 2023.

Accessing web Hadoop

The screenshot shows the Hadoop web interface with the title "All Applications". On the left, there is a navigation menu with options like "Cluster", "Nodes", "Applications", "Scheduler", "Capacity Scheduler", "Jobs", "Logs", "Tools", and "Help". The main content area displays several summary metrics:

- Cluster Metrics:** Apps Submitted: 0, Apps Pending: 0, Apps Running: 0, Apps Completed: 0, Containers Running: 0, Used Resources: <memory 0 B, vCores 0>, Total Resources: <memory 8 GB, vCores 8>, Reserved Resources: <memory 0 B, vCores 0>, Physical Mem Used %: 91, Physical Mem Available %: 87.
- Cluster Nodes Metrics:** Active Nodes: 0, Decommissioning Nodes: 0, Decommissioned Nodes: 0, Lost Nodes: 0, Unhealthy nodes: 0, Rebooted Nodes: 0, Stalled: 0.
- Scheduler Metrics:** Scheduler Type: memory-emb (unit=MB, vcores), Scheduling Resource Type: <memory 1024, vCores 1>, Minimum Allocation: <memory 8192, vCores 4>, Maximum Allocation: 0, Maximum Cluster Application Priority: 0.
- Capacity Scheduler:** Showing 0 to 0 of 0 entries.

At the bottom, there is a table header for applications with columns: ID, User, Name, Application Type, Application Tags, Queue, Application Priority, StartTime, LaunchTime, FinishTime, State, FinalStatus, Running Containers, Allocated CPU V-Cores, Allocated Memory MB, Allocated GPUs, Reserved CPU V-Cores, Reserved Memory MB, Reserved GPUs, % of Queue, % of Cluster, Progress, and Time. The table currently contains no data.

datanode

The screenshot shows the "Datanode Information" page in the Hadoop web interface. At the top, there is a legend for node states: In service (green check), Down (red dot), Decommissioning (yellow circle), Decommissioned (orange circle), Decommissioned & dead (red circle), Entering Maintenance (green check), In Maintenance (orange check), and In Maintenance & dead (red check).

Below the legend is a "Datanode usage histogram" showing a single green bar representing the disk usage of each Datanode, with the x-axis labeled "Disk usage of each Datanode (%)" ranging from 0 to 100.

The "In operation" section contains a table with the following data:

Node	Http Address	Last contact	Last Block Report	Used	Non DFS Used	Capacity	Blocks	Block pool used	Version
✓	http://172.31.7.10:50070	25%	20m	28 KB	4.18 GB	15.33 GB	0	28 KB (0%)	3.18

Below the table, there are sections for "Entering Maintenance" (No nodes are entering maintenance) and "Decommissioning" (No nodes are decommissioning).

ec2-35-182-230-210.ca-central-1.compute.amazonaws.com/9964/datanode.html

DataNode on ip-172-31-7-19.ca-central-1.compute.internal:9866

Cluster ID:	CID-cde9269b-17a7-4db4-a245-183bd80f5415
Started:	Tue Mar 12 22:08:12 -0400 2024
Version:	3.3.6, r1be78238728da9266a4f8819505808012bf9c

Block Pools

namenode Address	namenode HA State	Block Pool ID	Actor State	Last Heartbeat Sent	Last Heartbeat Response	Last Block Report	Last Block Report Size (Max Size)
ec2-35-182-230-210.ca-central-1.compute.amazonaws.com:9000	active	BP-1767645796-172.31.7.19-1710284846669	RUNNING	2s	2s	2 minutes	186 B (128 MB)

Volume Information

Directory	Storage Type	Capacity Used	Capacity Left	Capacity Reserved	Reserved Space for Replicas	Blocks
/home/hadoop/dfs/data/datanode	DISK	700 KB	11.06 GB	0 B	0 B	20

Hadoop, 2023.

Node running;

Instâncias | EC2 | ca-central-1 | Assignments - SY3T8191-3800 | Hortonworks Data Platform - U | Namenode information | Nodes of the cluster

ec2-3-99-244-23.ca-central-1.compute.amazonaws.com:8050/cluster/nodes

logged in as: di=who

Nodes of the cluster

- Cluster
- About
- Nodes
- Node Labels
- Applications
- Jobs
- Jobs #RUNNING
- JOBS SUBMITTED
- ACCEPTED
- Running
- FINISHED
- FAILED
- SKIPPED
- Scheduler
- Tools

Cluster Metrics		Apps Submitted		Apps Pending		Apps Running		Apps Completed		Containers Running		Used Resources		Total Resources		Reserved Resources		Physical Mem Used %		Physical VCores Used %	
0		0		0		0		0		0		<memory 0 B, vCores 0>		<memory 0 B, vCores 0>		<memory 0 B, vCores 0>		91		87	

Cluster Nodes Metrics		Active Nodes		Decommissioning Nodes		Decommissioned Nodes		Last Nodes		Unhealthy Nodes		Rebooted Nodes		Shutdown Nodes	
1		0		0		0		0		0		0		0	

Scheduler Metrics		Scheduler Type		Scheduling Resource Type		Minimum Allocation		Maximum Allocation		Maximum Cluster Application Priority		Scheduler Busy %	
Capacity Scheduler		(memory-mb, vcores)		<memory 1024, vCores 1>		<memory 8192, vCores 4>		0		Maximum Cluster Application Priority		0	

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health update	Health report	Containers	Allocation Tags	Mem Used	Mem Avail	Phys Mem Used %	VCores Used	VCores Avail	Phys VCores Used %	Version
default-rack		RUNNING	ip-172-31-7-19.ca-central-1.compute.internal.9866	ip-172-31-7-19.ca-central-1.compute.internal.9866	Tue Mar 12 23:52:50 -0400 2024		0		0 B	8 GB	91	0	8	91	3.3.6

Showing 1 to 1 of 1 entries

First Previous **1** Next Last