

Tarea 02: Instalación Hadoop

Juan Manuel García Moyano
IABD 24/25
Informática y comunicaciones

Índice

1. Instalación del OpenJDK de Java.....	3
2. Configurar el usuario y la autenticación SSH sin contraseña.....	4
3. Descarga de Hadoop.....	8
4. Configurar las variables de entorno de Hadoop.....	9
5. Configuración del clúster Apache Hadoop: Modo Pseudo-Distribuido.....	12
Configurar NameNode y DataNode.....	12
Gestor Yarn.....	19

1. Instalación del OpenJDK de Java

Primero actualizo la lista de repositorios del sistema de Ubuntu.

```
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo apt update
[sudo] contraseña para juan:
Lo siento, pruebe otra vez.
[sudo] contraseña para juan:
Obj:1 http://es.archive.ubuntu.com/ubuntu jammy InRelease
Obj:2 http://security.ubuntu.com/ubuntu jammy-security InRelease
Obj:3 http://es.archive.ubuntu.com/ubuntu jammy-updates InRelease
Obj:4 http://es.archive.ubuntu.com/ubuntu jammy-backports InRelease
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
Se pueden actualizar 47 paquetes. Ejecute «apt list --upgradable» para verlos.
```

Instalo el OpenJDK v11 de Java.

```
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo apt install default-jdk
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
Se instalarán los siguientes paquetes adicionales:
  ca-certificates-java default-jdk-headless default-jre default-jre-headless
  fonts-dejavu-extra java-common libatk-wrapper-java libatk-wrapper-java-jni
  libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev
  libxdmcp-dev libxt-dev openjdk-11-jdk openjdk-11-jdk-headless openjdk-11-jre
  openjdk-11-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
Paquetes sugeridos:
  libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo
  openjdk-11-source visualvm fonts-ipafont-gothic fonts-ipafont-mincho
  fonts-wqy-microhei | fonts-wqy-zenhei
Se instalarán los siguientes paquetes NUEVOS:
  ca-certificates-java default-jdk default-jdk-headless default-jre
  default-jre-headless fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev
  libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-11-jdk
  openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev
  xorg-sgml-doctools xtrans-dev
0 actualizados, 24 nuevos se instalarán, 0 para eliminar y 47 no actualizados.
Se necesita descargar 122 MB de archivos.
Se utilizarán 275 MB de espacio de disco adicional después de esta operación.
```

Compruebo que se ha instalado correctamente, para ello compruebo la versión de Java.

```
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ java -version
openjdk version "11.0.24" 2024-07-16
OpenJDK Runtime Environment (build 11.0.24+8-post-Ubuntu-1ubuntu322.04)
OpenJDK 64-Bit Server VM (build 11.0.24+8-post-Ubuntu-1ubuntu322.04, mixed mode,
sharing)
juan@juan-Standard-PC-i440FX-PIIX-1996:~$
```

2. Configurar el usuario y la autenticación SSH sin contraseña.

Como no tengo SSH, instalo el paquete como server y cliente. Además del paquete pdsh para el cliente shell remoto multihilo que me permita ejecutar comandos en varios hosts en modo paralelo.


```
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo apt install openssh-server openssh-client pdsh
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias... Hecho
Leyendo la información de estado... Hecho
openssh-client ya está en su versión más reciente (1:8.9p1-3ubuntu0.10).
Se fijó openssh-client como instalado manualmente.
Se instalarán los siguientes paquetes adicionales:
  genders libgenders0 ncurses-term openssh-sftp-server
  ssh-import-id
Paquetes sugeridos:
  rdist molly-guard monkeysphere ssh_askpass
Se instalarán los siguientes paquetes NUEVOS:
  genders libgenders0 ncurses-term openssh-server
  openssh-sftp-server pdsh ssh-import-id
0 actualizados, 7 nuevos se instalarán, 0 para eliminar y 47 no actualizados.
Se necesita descargar 922 kB de archivos.
Se utilizarán 6.573 kB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] S
Des:1 http://es.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-sftp-server amd64 1:8.9p1-3ubuntu0.10 [38,9 kB]
Des:2 http://es.archive.ubuntu.com/ubuntu jammy-updates/main amd64 openssh-server amd64 1:8.9p1-3ubuntu0.10 [435 kB]
Des:3 http://es.archive.ubuntu.com/ubuntu jammy/universe amd64 libgenders0 amd64 1.22-1build4 [31,5 kB]
Des:4 http://es.archive.ubuntu.com/ubuntu jammy/universe amd64 genders amd64 1.22-1build4 [31,3 kB]
Des:5 http://es.archive.ubuntu.com/ubuntu jammy-updates/main amd64 ncurses-term all 6.3-2ubuntu0.1 [267 kB]
Des:6 http://es.archive.ubuntu.com/ubuntu jammy/universe amd64 pdsh amd64 2.31-3build2 [108 kB]
Des:7 http://es.archive.ubuntu.com/ubuntu jammy/main amd64 ssh-import-id all 5.11-0ubuntu1 [10,1 kB]
Descargados 922 kB en 0s (2.102 kB/s)
Preconfigurando paquetes ...
Seleccionando el paquete openssh-sftp-server previamente no seleccionado
```

Con el primer comando creo un nuevo usuario “hadoop” y con el segundo establezco la contraseña para este.

```
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo useradd -m -s /bin/bash hadoop
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo passwd hadoop
Nueva contraseña:
Vuelva a escribir la nueva contraseña:
passwd: contraseña actualizada correctamente
```

Añado al usuario “hadoop” al grupo “sudo”, para así poder ejecutar con este usuario el comando “sudo”.

```
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo usermod -aG sudo hadoop
```

Inicio sesión con el usuario “hadoop”.

```
juan@juan-Standard-PC-i440FX-PIIX-1996:~$ su - hadoop
Contraseña:
To run a command as administrator (user "root"), use "sudo <command>".
See "man sudo_root" for details.
```

Con este comando busco generar la clave pública y privada SSH. Cuando me pide la contraseña pulso ENTER para omitirla.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ ssh-keygen -t rsa
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
Created directory '/home/hadoop/.ssh'.
Enter passphrase (empty for no passphrase):
Enter same passphrase again:
Your identification has been saved in /home/hadoop/.ssh/id_rsa
Your public key has been saved in /home/hadoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:1ZMyi4qoKJOQp4uWHcm7R06atX6kFAwhvTESuDeITJs hadoop@juan-St
andard-PC-i440FX-PIIX-1996
The key's randomart image is:
+---[RSA 3072]---+
|.oo..          |
|..O=.          |
|+.=*          |
|. E. o    o + . |
| o... . S .    |
|o . = * o      |
|.++ % =        |
|B+ = = .       |
|Bo .+..        |
+-----[SHA256]-----+
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Primero verifico que la clave SSH ha sido generada. Con el segundo comando busco primero copiar la clave pública SSH “id_rsa.pub” en el archivo “authorized_keys” y luego le cambio el permiso por defecto a 600.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ ls ~/.ssh/
id_rsa id_rsa.pub
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ cat ~/.ssh/id_rsa.pub
>> ~/.ssh/authorized_keys
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ chmod 600 ~/.ssh/auth
orized_keys
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Ya terminada la configuración SSH sin contraseña, la verifico conectándome a la máquina local mediante el comando ssh.


```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:LHBKbVAbNkBoJ+rIQ+568J/mCAt/XEg
bPgETo2ghITs.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint
])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of k
nown hosts.
Welcome to Ubuntu 22.04.5 LTS (GNU/Linux 6.8.0-47-generic x86_64)

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

El mantenimiento de seguridad expandido para Applications está de
sactivado

Se pueden aplicar 45 actualizaciones de forma inmediata.
25 de estas son actualizaciones de seguridad estándares.
Para ver estas actualizaciones adicionales, ejecute: apt list --u
pgradable

Active ESM Apps para recibir futuras actualizaciones de seguridad
adicionales.
Vea https://ubuntu.com/esm o ejecute «sudo pro status»

25 updates could not be installed automatically. For more details
,
see /var/log/unattended-upgrades/unattended-upgrades.log

The programs included with the Ubuntu system are free software;
the exact distribution terms for each program are described in th
e
individual files in /usr/share/doc/*/copyright.

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted
```


3. Descarga de Hadoop

Descargo desde la web de Apache “hadoop-3.4.0.tar.gz” (descargo esta versión porque la que pone en el tutorial no está disponible).

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ wget https://dlcdn.ap
ache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0.tar.gz
--2024-10-25 19:52:34-- https://dlcdn.apache.org/hadoop/common/h
adoop-3.4.0/hadoop-3.4.0.tar.gz
Resolviendo dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132,
2a04:4e42::644
Conectando con dlcdn.apache.org (dlcdn.apache.org)[151.101.2.132]
:443... conectado.
Petición HTTP enviada, esperando respuesta... 200 OK
Longitud: 965537117 (921M) [application/x-gzip]
Guardando como: 'hadoop-3.4.0.tar.gz'

hadoop-3.4.0.tar 100%[=====>] 920,81M  24,7MB/s   en 31s

2024-10-25 19:53:06 (29,5 MB/s) - 'hadoop-3.4.0.tar.gz' guardado
[965537117/965537117]

hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Muestro que se ha descargado correctamente.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ ls
hadoop-3.4.0.tar.gz  hdfs  snap
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Extraigo (desempaquete) el paquete que recién me he descargado (no muestro el proceso porque me sale un chorizo y he intentado subir para mostrarlo pero se pierde el comando).

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ tar -xvzf hadoop-3.4.
0.tar.gz
```

Una vez desempaquetado, lo muevo al directorio “/usr/local/hadoop”.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo mv hadoop-3.4.0
/usr/local/hadoop
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Para finalizar este apartado, cambio la propiedad del directorio de instalación de hadoop “/usr/local/hadoop” al usuario “hadoop” y al grupo “hadoop”.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo chown -R hadoop:
hadoop /usr/local/hadoop
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

4. Configurar las variables de entorno de Hadoop

Abro el archivo “~/.bashrc” mediante el editor nano.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ nano ~/.bashrc
```

Añado las líneas al final del fichero “/home/hadoop/.bashrc” como se puede observar en la segunda captura.

```
GNU nano 6.2 /home/hadoop/.bashrc
# ~/.bashrc: executed by bash(1) for non-login shells.
# see /usr/share/doc/bash/examples/startup-files (in the package bash-doc)
# for examples
#
# If not running interactively, don't do anything
case $- in
  *(*) ;;
  *) return;;
esac
# don't put duplicate lines or lines starting with space in the history.
# See bash(1) for more options
HISTCONTROL=ignoreboth
# append to the history file, don't overwrite it
shopt -s histappend
# for setting history length see HISTSIZE and HISTFILESIZE in bash(1)
HISTSIZE=1000
HISTFILESIZE=2000
# check the window size after each command and, if necessary,
# update the values of LINES and COLUMNS.
shopt -s checkwinsize
# If set, the pattern "*" used in a pathname expansion context will
# match all files and zero or more directories and subdirectories.
#shopt -s globstar
# make less more friendly for non-text input files. see lesspipe(1)
```

```
# Hadoop environment variables
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export HADOOP_YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
export PDSH_RCMD_TYPE=ssh
```

Una vez guardado el fichero anterior, salgo del editor. Ahora ejecuto el primer comando para aplicar los nuevos cambios dentro del archivo “~/.bashrc”. Los siguientes 3 comandos muestran que las variables de entorno están correctamente.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ source ~/.bashrc
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ echo $JAVA_HOME
/usr/lib/jvm/java-11-openjdk-amd64
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ echo $HADOOP_HOME
/usr/local/hadoop
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ echo $HADOOP_OPTS
-Djava.library.path=/usr/local/hadoop/lib/native
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Abro el archivo “hadoop-env.sh”: **nano \$HADOOP_HOME/etc/hadoop/hadoop-env.sh**.
Descomento la línea que exporta la variable de entorno y cambio el valor al directorio de instalación de Java, tal y como está en la captura:

```
/usr/local/hadoop/etc/hadoop/hadoop-env.sh
###
# Generic settings for HADOOP
###

# Technically, the only required environment variable is JAVA_HOME
# All others are optional.  However, the defaults are probably not
# preferred.  Many sites configure these options outside of Hadoop
# such as in /etc/profile.d

# The java implementation to use.  By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
```

Compruebo la versión que me he instalado de Hadoop y que todo está correctamente:

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ hadoop version
Hadoop 3.4.0
Source code repository git@github.com:apache/hadoop.git -r bd8b77
f398f626bb7791783192ee7a5dfaeeec760
Compiled by root on 2024-03-04T06:35Z
Compiled on platform linux-x86_64
Compiled with protoc 3.21.12
From source with checksum f7fe694a3613358b38812ae9c31114e
This command was run using /usr/local/hadoop/share/hadoop/common/
hadoop-common-3.4.0.jar
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```


5. Configuración del clúster Apache Hadoop: Modo Pseudo-Distribuido

Configurar NameNode y DataNode

Una vez instalado Hadoop, configuro el NameNode y el DataNode para el clúster.

Para comenzar con la configuración de Hadoop, abro el fichero “\$HADOOP_HOME/etc/hadoop/core-site.xml” y añado las líneas que salen en la captura 2.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

```
GNU nano 6.2 /usr/local/hadoop/etc/hadoop/core-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>fs.defaultFS</name>
    <value>hdfs://0.0.0.0</value>
  </property>
</configuration>
```

Creo los directorios namenode y datanode que se utilizarán para el DataNode en el clúster. Después la propiedad de los directorios se los cambio al usuario Hadoop.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo mkdir -p /home/hadoop/hdfs/{namenode,datanode}
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo chown -R hadoop:hadoop /home/hadoop/hdfs
```

Abro el fichero `hdfs-site.xml` y le añado la configuración tal cual se puede ver en la segunda captura. Esas líneas me permite decirle a Hadoop el directorio que va a utilizar para el node de los datos y que la replicación va a ser en un nodo.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

```
GNU nano 6.2 /usr/local/hadoop/etc/hadoop/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>1</value>
  </property>

  <property>
    <name>dfs.name.dir</name>
    <value>file:///home/hadoop/hdfs/namenode</value>
  </property>

  <property>
    <name>dfs.data.dir</name>
    <value>file:///home/hadoop/hdfs/datanode</value>
  </property>
</configuration>
```

Una vez vez he terminado la configuración anterior, ejecuto el siguiente comando para formatear el sistema de archivos de Hadoop.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ hdfs namenode -format
```

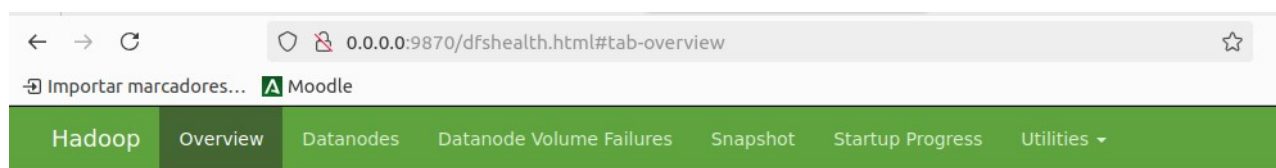

Esta sería la salida que me muestra.

```
2024-10-25 20:39:29,545 INFO metrics.TopMetrics: NNTop conf: dfs.
namenode.top.windows.minutes = 1,5,25
2024-10-25 20:39:29,562 INFO namenode.FSNamesystem: Retry cache o
n namenode is enabled
2024-10-25 20:39:29,562 INFO namenode.FSNamesystem: Retry cache w
ill use 0.03 of total heap and retry cache entry expiry time is 6
00000 millis
2024-10-25 20:39:29,579 INFO util.GSet: Computing capacity for ma
p NameNodeRetryCache
2024-10-25 20:39:29,579 INFO util.GSet: VM type          = 64-bit
2024-10-25 20:39:29,580 INFO util.GSet: 0.029999999329447746% max
memory 956 MB = 293.7 KB
2024-10-25 20:39:29,580 INFO util.GSet: capacity         = 2^15 = 32
768 entries
2024-10-25 20:39:29,689 INFO namenode.FSImage: Allocated new Bloc
kPoolId: BP-1151951731-127.0.1.1-1729881569669
2024-10-25 20:39:29,903 INFO common.Storage: Storage directory /h
ome/hadoop/hdfs/namenode has been successfully formatted.
2024-10-25 20:39:30,082 INFO namenode.FSImageFormatProtobuf: Savi
ng image file /home/hadoop/hdfs/namenode/current/fsimage.ckpt_000
00000000000000000 using no compression
2024-10-25 20:39:30,691 INFO namenode.FSImageFormatProtobuf: Imag
e file /home/hadoop/hdfs/namenode/current/fsimage.ckpt_0000000000
0000000000 of size 401 bytes saved in 0 seconds .
2024-10-25 20:39:30,810 INFO namenode.NNStorageRetentionManager:
Going to retain 1 images with txid >= 0
2024-10-25 20:39:30,835 INFO blockmanagement.DatanodeManager: Slo
w peers collection thread shutdown
2024-10-25 20:39:30,906 INFO namenode.FSNamesystem: Stopping serv
ices started for active state
2024-10-25 20:39:30,906 INFO namenode.FSNamesystem: Stopping serv
ices started for standby state
2024-10-25 20:39:30,924 INFO namenode.FSImage: FSImageSaver clean
checkpoint: txid=0 when meet shutdown.
2024-10-25 20:39:30,929 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at juan-Standard-PC-i440FX-P
IIX-1996/127.0.1.1
*****/
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Ahora inicio el NameNode y el DataNode, estos se ejecutarán en la dirección que configuré anteriormente en el fichero “core-site.xml”.

```
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ start-dfs.sh
Starting namenodes on [0.0.0.0]
Starting datanodes
Starting secondary namenodes [juan-Standard-PC-i440FX-PIIX-1996]
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$
```

Verifico que todo está funcionando correctamente. Para ello abro un navegador web y en la URL pongo la IP configurada (0.0.0.0) y el puerto que por defecto es 9870.



Overview '0.0.0.0:8020' (✓active)

Started:	Wed Oct 30 19:19:48 +0100 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760
Compiled:	Mon Mar 04 07:35:00 +0100 2024 by root from (HEAD detached at release-3.4.0-RC3)
Cluster ID:	CID-6788ff86-aa0e-4279-bf2f-3ed8fb185931
Block Pool ID:	BP-1151951731-127.0.1.1-1729881569669

Summary

Security is off.

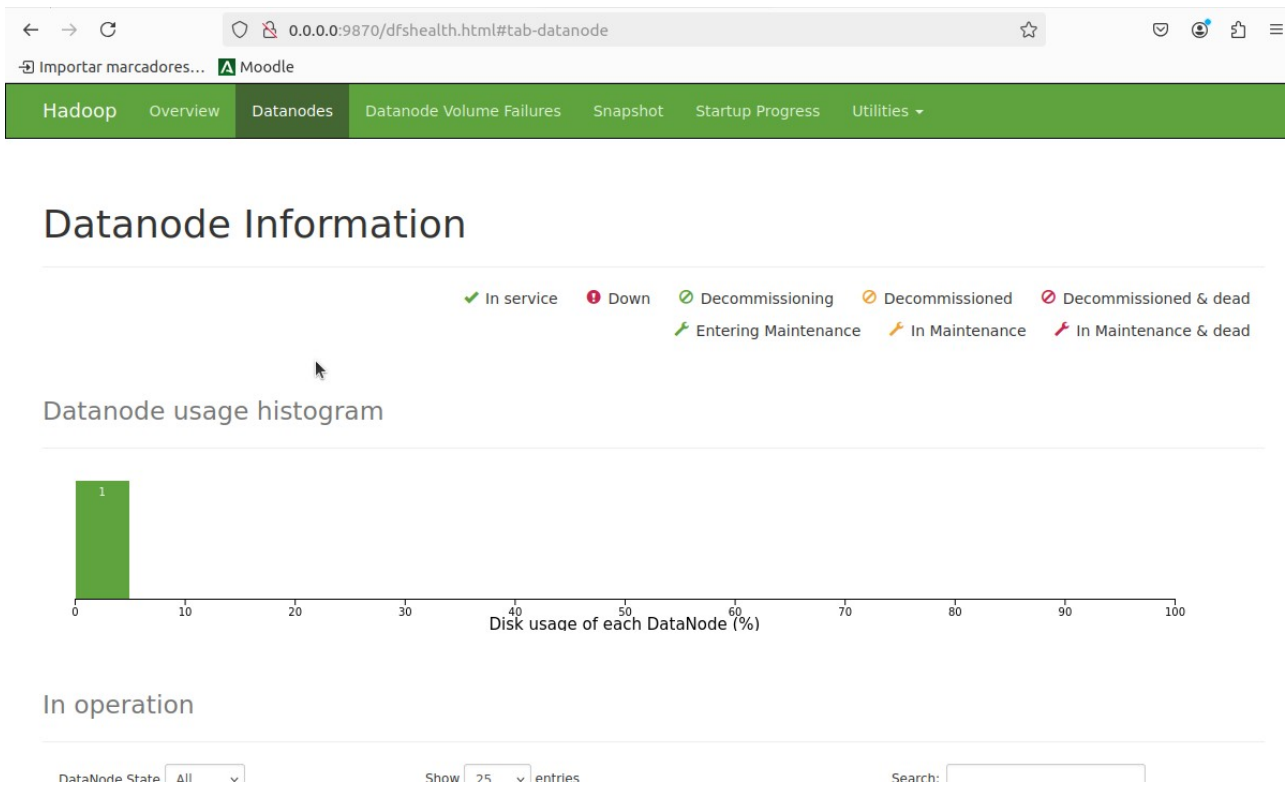
Safemode is off.

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

Heap Memory used 60.93 MB of 100 MB Heap Memory. Max Heap Memory is 956 MB.

Non Heap Memory used 54.51 MB of 57.5 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

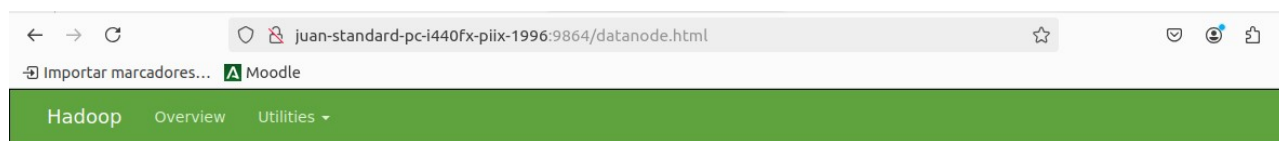
Hago clic en Datanodes y así podemos cual está activo. En la segunda captura se observa que el nodo está funcionando en el puerto 9866.



Para continuar pulso en la dirección HTTP Address que se ve en la captura.



Una vez pulso el enlace anterior, puedo observar que el DataNode se está ejecutando en el directorio configurado “/home/hadoop/hdfs/datanode”.



DataNode on juan-Standard-PC-i440FX-PIIX-1996:9866

Cluster ID:	CID-6788ff86-aa0e-4279-bf2f-3ed8fb185931
Started:	Wed Oct 30 19:19:53 +0100 2024
Version:	3.4.0, rbd8b77f398f626bb7791783192ee7a5dfaeec760

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)
0.0.0.0:8020	active	BP-1151951731-127.0.1.1-1729881569669	RUNNING	0s	0s	6 minutes	0 B (128 MB)

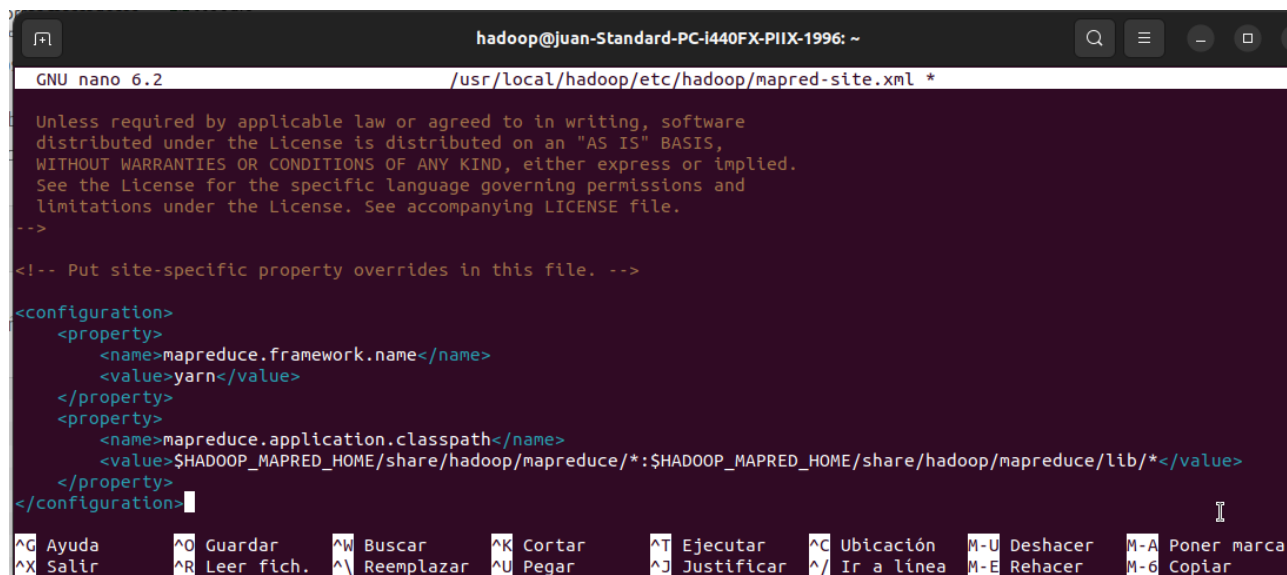
Volume Information

Ahora voy a ejecutar MapReduce en el gestor Yarn (gestor de recursos y gestor de nodos).

Gestor Yarn

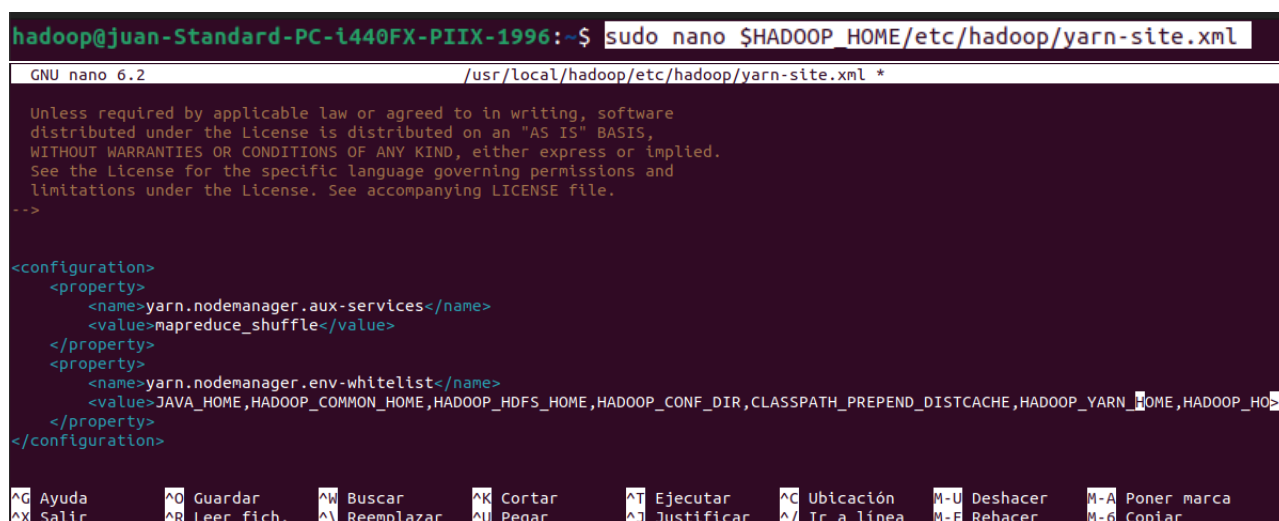
Para ejecutar MapReduce en Yarn en modo pseudo-distribuido, realizo la siguiente configuración.

Primero abro el archivo “mapred-site.xml” utilizando el editor nano como se puede ver en la captura y añado las líneas de configuración.



```
hadoop@juan-Standard-PC-i440FX-PIIX-1996: ~  
GNU nano 6.2 /usr/local/hadoop/etc/hadoop/mapred-site.xml *  
  
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.application.classpath</name>  
    <value>$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/*:$HADOOP_MAPRED_HOME/share/hadoop/mapreduce/lib/*</value>  
  </property>  
</configuration>  
  
^G Ayuda      ^O Guardar    ^W Buscar     ^K Cortar     ^T Ejecutar   ^C Ubicación  M-U Deshacer  M-A Poner marca  
^X Salir      ^R Leer fich. ^L Reemplazar ^U Pegar      ^J Justificar ^/ Ir a línea  M-E Rehacer  M-6 Copiar
```

Ahora abro el fichero yarn-site.xml, de nuevo, con el editor nano y le añado la configuración, como se muestra en las capturas.



```
hadoop@juan-Standard-PC-i440FX-PIIX-1996: ~$ sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml  
GNU nano 6.2 /usr/local/hadoop/etc/hadoop/yarn-site.xml *  
  
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.env-whitelist</name>  
    <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PREPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_HO</value>  
  </property>  
</configuration>  
  
^G Ayuda      ^O Guardar    ^W Buscar     ^K Cortar     ^T Ejecutar   ^C Ubicación  M-U Deshacer  M-A Poner marca  
^X Salir      ^R Leer fich. ^L Reemplazar ^U Pegar      ^J Justificar ^/ Ir a línea  M-E Rehacer  M-6 Copiar
```



```

GNU nano 6.2 /usr/local/hadoop/etc/hadoop/yarn-site.xml *
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.env-whitelist</name>
    <value>HADOOP_HOME,PATH,LANG,TZ,HADOOP_MAPRED_HOME</value>
  </property>
</configuration>

```

Ahora ejecuto el siguiente comando para iniciar los demonios Yarn y puedo comprobar que se ha realizado correctamente.

```

hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hadoop@juan-Standard-PC-i440FX-PIIX-1996:~$

```

Abro un navegador web y compruebo que el ResourceManager está ejecutandose en el puerto 8088. Lo que se ve en la captura, es la interfaz web del Gestor de Recursos Hadoop. Se puede monitorizar todos los procesos en ejecución dentro del clúster.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running
0	0	0	0	0

Cluster Nodes Metrics

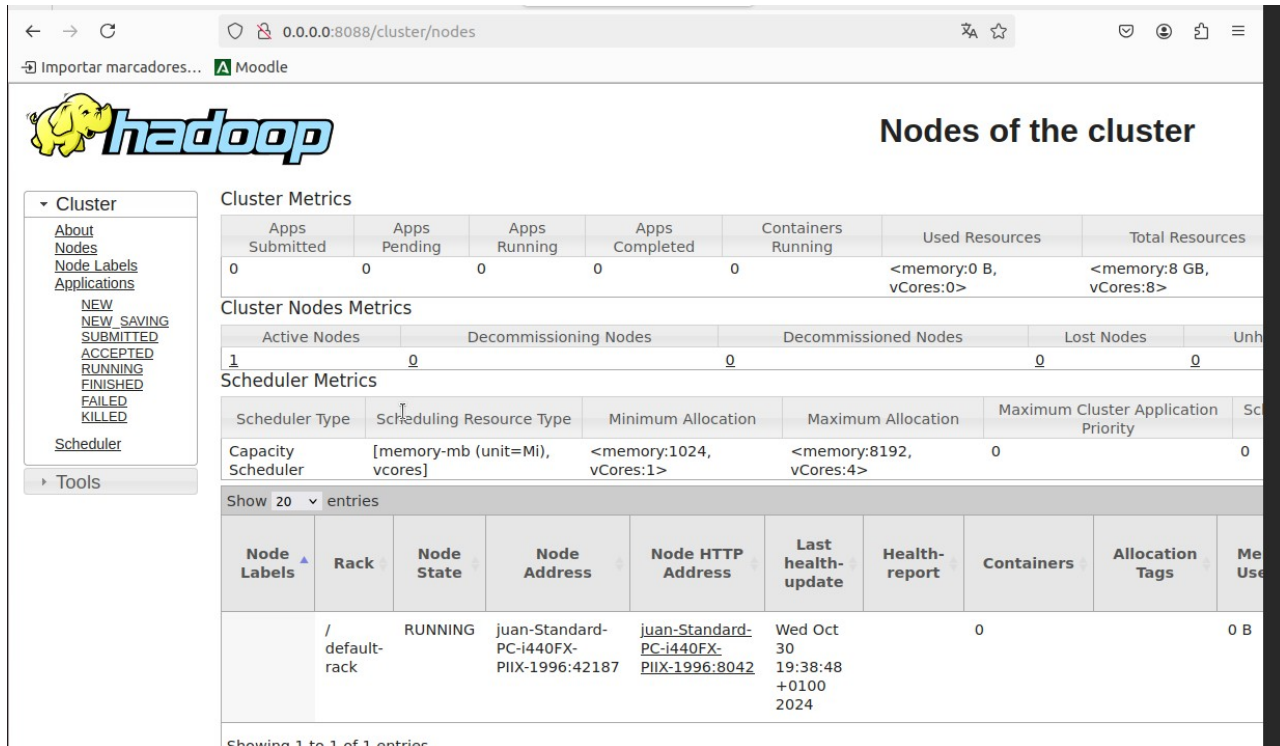
Active Nodes	Decommissioning Nodes	Decommissioned Nodes
1	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation
Capacity Scheduler	[memory-mb (unit=Mb), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:1>

Showing 0 to 0 of 0 entries

En el menú de la derecha, pulso en Nodes. Aquí observo el nodo que se está ejecutando actualmente en el clúster Hadoop.



The screenshot shows the Hadoop cluster management interface. The main title is "Nodes of the cluster". The interface includes a sidebar with navigation links, cluster metrics, cluster nodes metrics, scheduler metrics, and a table of active nodes.

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources
0	0	0	0	0	<memory:0 B, vCores:0>	<memory:8 GB, vCores:8>

Cluster Nodes Metrics

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

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:8192, vCores:4>	0	0

Showing 20 entries

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Memory Used
/	default-rack	RUNNING	juan-Standard-PC-i440FX-PIIX-1996:42187	juan-Standard-PC-i440FX-PIIX-1996:8042	Wed Oct 30 19:38:48 +0100 2024		0		0 B

Showing 1 to 1 of 1 entries