INSTALACION HADOOP LINUX

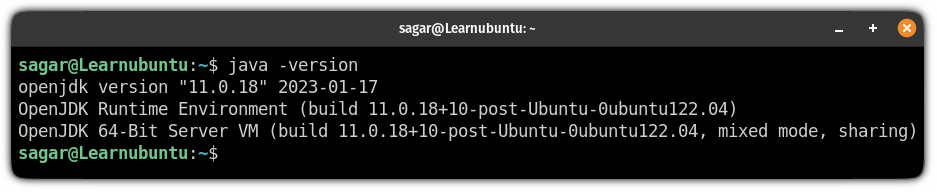
### Step 1: Installing Java on Ubuntu

To install java on Ubuntu, all you have to do is execute the following command:

sudo apt install default-jdk default-jre -y

To verify the installation, check the java version on your system:

java -version

[](https://learnubuntu.com/content/images/2023/03/check-the-java-version-on-ubuntu.png)

### Step 2: Create a user for Hadoop and configure SSH

First, [create a new user](https://learnubuntu.com/add-delete-users/) named hadoop:

sudo adduser hadoop

To [enable superuser privileges to the new user](https://learnubuntu.com/add-sudo-user/), add it to the sudo group:

sudo usermod -aG sudo hadoop

Once done, [switch to the user](https://learnubuntu.com/change-user/) hadoop:

sudo su - hadoop

Next, install the OpenSSH server and client:

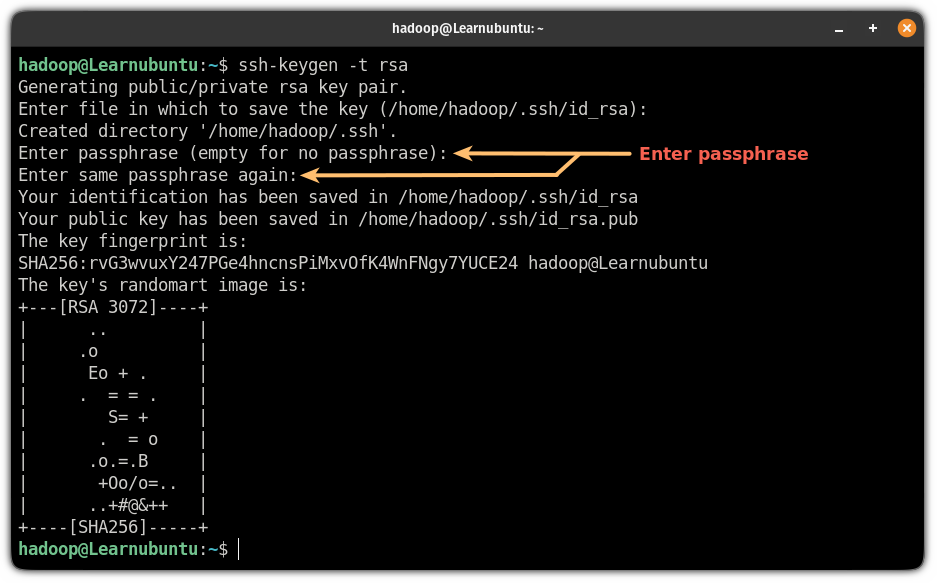
sudo apt install openssh-server openssh-client -y

Now, use the following command to generate private and public keys:

ssh-keygen -t rsa

Here, it will ask you:

* Where to save the key (hit enter to save it inside your home directory)
* Create passphrase for keys (leave blank for no passphrase)

[](https://learnubuntu.com/content/images/2023/03/generate-public-and-private-keys-for-SSH-on-Ubuntu.png)

Now, add the public key to authorized\_keys:

cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

Use the chmod command to change the file permissions of authorized\_keys:

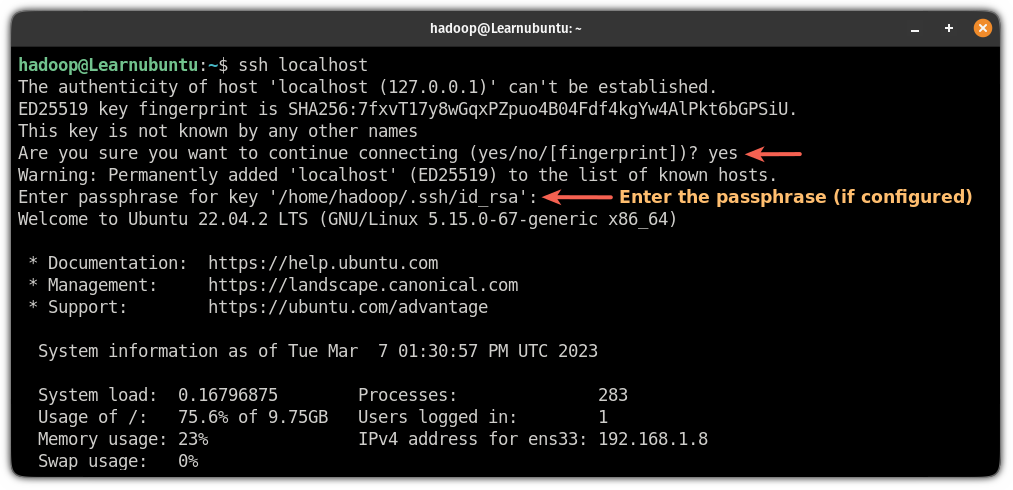
sudo chmod 640 ~/.ssh/authorized\_keys

Finally, verify the SSH configuration:

ssh localhost

(si no funciona: **sudo service ssh start** o **sudo systemctl start ssh)**

If you have not configured the password, all you have to do is type yes and hit enter if you added a passphrase for the keys, it will ask you to enter here:

[](https://learnubuntu.com/content/images/2023/03/connect-localhost-via-SSH.png)

### Step 3: Download and install Apache Hadoop on Ubuntu

If you have created a user for Hadoop, first, log in as the hadoop user:

sudo su - hadoop

Now, [visit the download page for Apache Hadoop](https://downloads.apache.org/hadoop/common/stable/?ref=learnubuntu.com) and copy the link for the most recent stable release.

While writing, its 3.3.4 so I will be using [the wget command](https://learnubuntu.com/install-wget/) to download this release:

wget https://downloads.apache.org/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz

Once you are done with the download, extract the file using the following command:

tar -xvzf hadoop-3.3.4.tar.gz

Next, move the extracted file to the /usr/local/hadoop using the following command:

sudo mv hadoop-3.3.4 /usr/local/hadoop

Now, create a directory to store logs:

sudo mkdir /usr/local/hadoop/logs

Finally, change the ownership of the /usr/local/hadoop to the user hadoop:

sudo chown -R hadoop:hadoop /usr/local/hadoop

### Step 4: Configure Hadoop on Ubuntu

Here, I will walk you through the configuration of the Hadoop environment variable.

First, open the .bashrc file using the following command:

sudo nano ~/.bashrc

[Jump to the end of the line in the nano text editor](https://linuxhandbook.com/beginning-end-file-nano/?ref=learnubuntu.com) by pressing Alt + / and paste the following lines:

export HADOOP\_HOME=/usr/local/hadoop

export HADOOP\_INSTALL=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

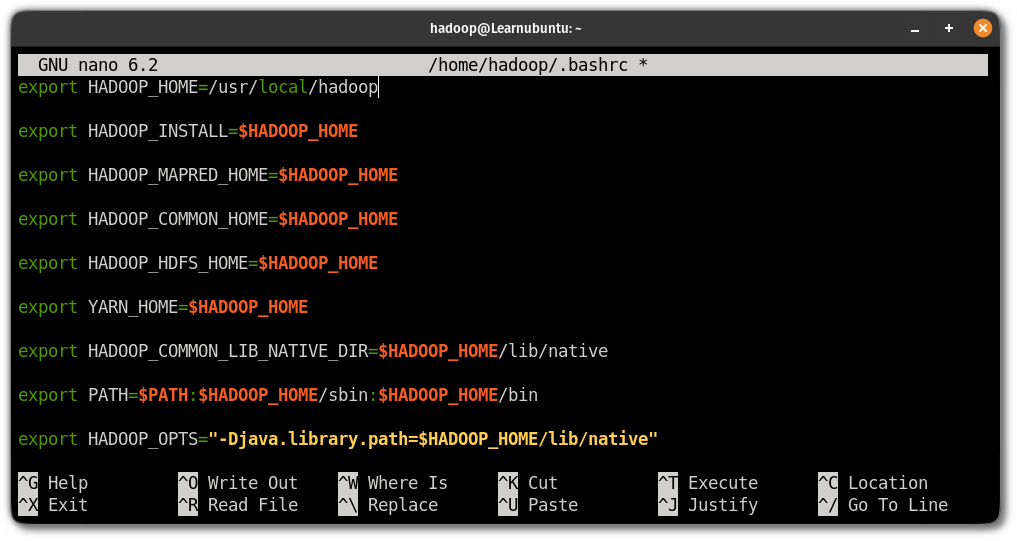
export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export 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"

[](https://learnubuntu.com/content/images/2023/03/configure-Hadoop-environment-variable-in-ubuntu.png)

[Save changes and exit from the nano](https://linuxhandbook.com/nano-save-exit/?ref=learnubuntu.com) text editor.

To enable the changes, source the .bashrc file:

source ~/.bashrc

### Step 5: Configure java environment variables

To use Hadoop, you are required to enable its core functions which include YARN, HDFS, MapReduce, and Hadoop-related project settings.

To do that, you will have to define java environment variables in hadoop-env.sh file.

#### Edit the hadoop-env.sh file

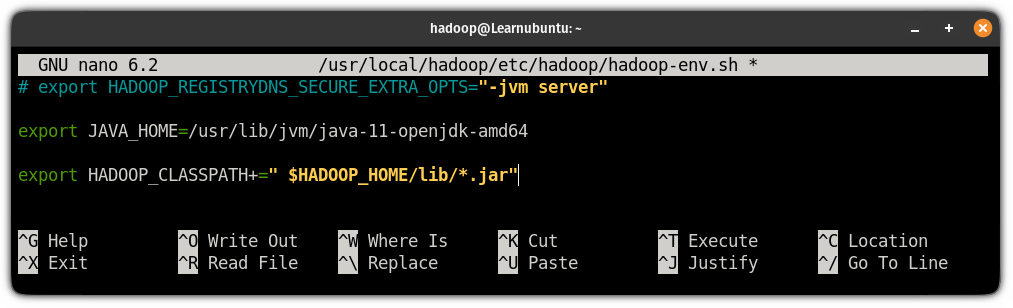
First, open the hadoop-env.sh file:

sudo nano $HADOOP\_HOME/etc/hadoop/hadoop-env.sh

Press Alt + / to jump to the end of the file and paste the following lines in the file to add the path of the Java:

export JAVA\_HOME=/usr/lib/jvm/java-11-openjdk-amd64

export HADOOP\_CLASSPATH+=" $HADOOP\_HOME/lib/\*.jar"

[](https://learnubuntu.com/content/images/2023/03/add-java-path-in-the-Hadoop-env-file.png)

Save changes and exit from the text editor.

Next, change your current working directory to /usr/local/hadoop/lib:

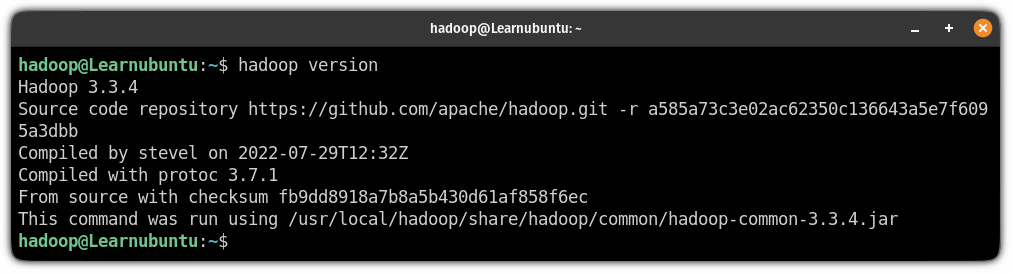
cd /usr/local/hadoop/lib

Here, download the javax activation file:

sudo wget https://jcenter.bintray.com/javax/activation/javax.activation-api/1.2.0/javax.activation-api-1.2.0.jar

Once done, check the Hadoop version in Ubuntu:

hadoop version

[](https://learnubuntu.com/content/images/2023/03/check-the-installed-version-of-hadoop.png)

Next, you will have to edit the core-site.xml file to specify the URL for the name node.

#### Edit the core-site.xml file

First, open the core-site.xml file using the following command:

sudo nano $HADOOP\_HOME/etc/hadoop/core-site.xml

And add the following lines in between <configuration>  </configuration>:

<configuration>

<property>

<name>fs.default.name</name>

<value>hdfs://0.0.0.0:9000</value>

<description>The default file system URI</description>

</property>

</configuration>

Save the changes and exit from the text editor.

Next, create a directory to store node metadata using the following command:

sudo mkdir -p /home/hadoop/hdfs/{namenode,datanode}

And change the ownership of the created directory to the hadoop user:

sudo chown -R hadoop:hadoop /home/hadoop/hdfs

#### Edit the hdfs-site.xml configuration file

By configuring the hdfs-site.xml file, you will define the location for storing node metadata, fs-image file.

So first open the configuration file:

sudo nano $HADOOP\_HOME/etc/hadoop/hdfs-site.xml

And paste the following line in between <configuration> ... </configuration>:

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

Save changes and exit from the hdfs-site.xml file.

#### Edit the mapred-site.xml file

By editing the mapred-site.xml file, you can define the MapReduce values.

To do that, first, open the configuration file using the following command:

sudo nano $HADOOP\_HOME/etc/hadoop/mapred-site.xml

And paste the following line in between <configuration> ... </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>

Save and exit from the nano text editor.

#### Edit the yarn-site.xml file

This is the last configuration file that needs to be edited to use the Hadoop service.

The purpose of editing this file is to define the YARN settings.

First, open the configuration file:

sudo nano $HADOOP\_HOME/etc/hadoop/yarn-site.xml

Paste the following in between <configuration> ... </configuration>:

<property>

<description>The hostname of the RM.</description>

<name>yarn.resourcemanager.hostname</name>

<value>localhost</value>

</property>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.env-whitelist</name>

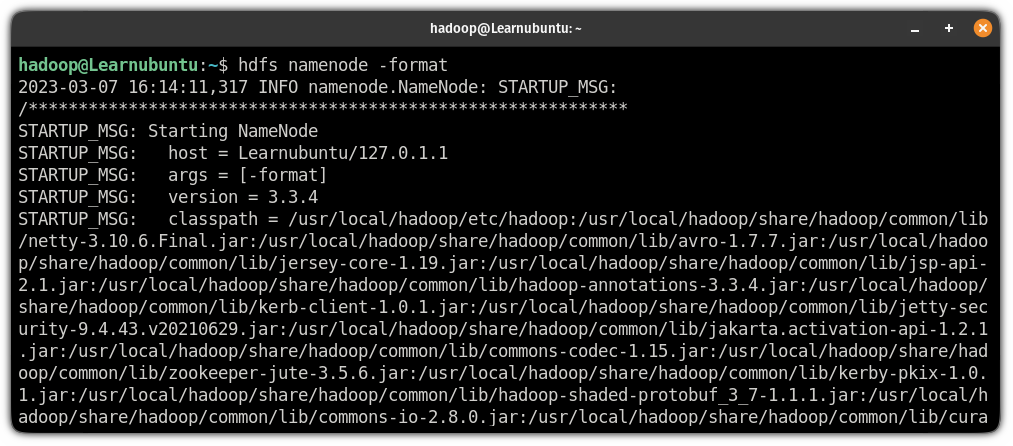
<value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PREPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_HOME,PATH,LANG,TZ,HADOOP\_MAPRED\_HOME</value>

</property>

Save changes and exit from the config file.

Finally, use the following command to validate the Hadoop configuration and to format the HDFS NameNode:

hdfs namenode -format

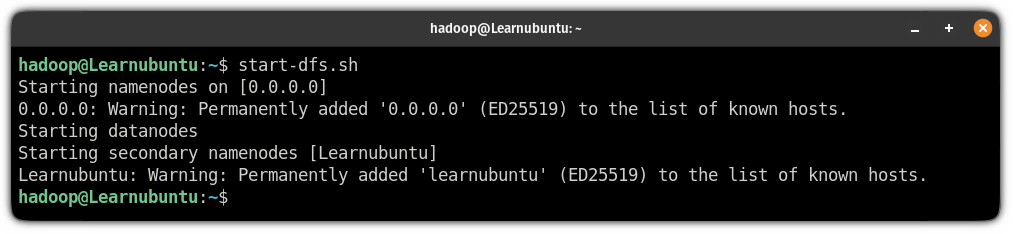
[](https://learnubuntu.com/content/images/2023/03/validate-the-Hadoop-configuration-in-Ubuntu.png)

### Step 6: Start the Hadoop cluster

To start the Hadoop cluster, you will have to start the previously configured nodes.

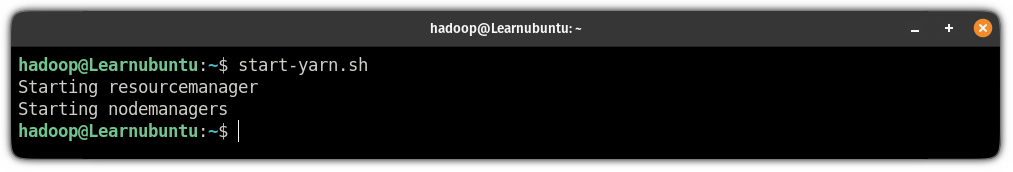
So let's start with starting the NameNode and DataNode:

start-dfs.sh

[](https://learnubuntu.com/content/images/2023/03/start-the-NameNode-and-DataNode-in-Hadoop-cluster-in-ubuntu.png)

Next, start the node manager and resource manager:

start-yarn.sh

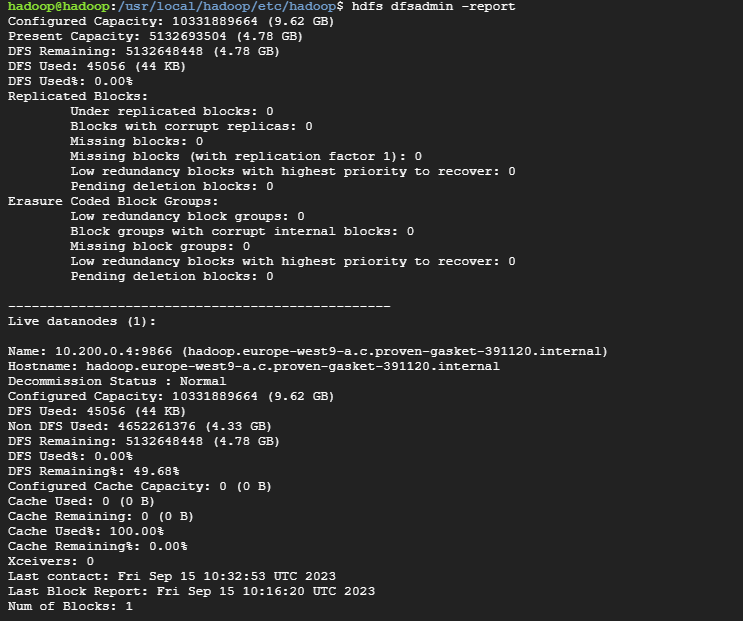
[](https://learnubuntu.com/content/images/2023/03/start-the-node-manager-and-resource-manager-to-start-the-Hadoop-cluster-in-ubuntu.png)

To verify whether the services are running as intended, use the following command:

jps

Comprobamos el reporte de hdfs:

**hdfs dfsadmin -report**



Creamos un directorio en hdfs:

hdfs dfs -mkdir /user

Comprobamos:

hdfs dfs -ls /user/

Si la instalación es local, podemos ver el monitor de recursos en:

<http://localhost:8088/cluster/nodes>

mapred streaming -files mapper.py,reducer.py -mapper "python3 mapper.py" -reducer "python3 reducer.py" -input hdfs:///user/input.txt -output hdfs:///user/salida

hdfs dfs -ls /user/salida/

Found 2 items

-rw-r--r-- 1 hadoop supergroup 0 2023-09-15 10:41 /user/salida/\_SUCCESS

-rw-r--r-- 1 hadoop supergroup 172 2023-09-15 10:41 /user/salida/part-00000

hdfs dfs -cat /user/salida/part-00000

Cada 1

Este 2

Las 1

de 3

documento 2

documento. 1

documentos. 1

ejemplo 2

en 1

es 2

los 1

otro 1

palabras 1

palabras. 1

pueden 1

repetirse 1

texto. 1

tiene 1

un 2

varias 1

<https://learnubuntu.com/install-hadoop/#google_vignette>