

Instalación de Hadoop single node

1. Instalamos java:

1.1. Abrimos la terminal

1.2. Nos logeamos como root: su -

1.3. Buscamos qué versiones hay del jdk de java disponibles: dnf search openjdk-devel

```
[root@localhost ~]# dnf search openjdk-devel
Last metadata expiration check: 0:29:02 ago on Thu 18 Nov 2021 10:57:57 AM CET.
===== Name Matched: openjdk-devel =====
java-1.8.0-openjdk-devel.x86_64 : OpenJDK 8 Development Environment
java-11-openjdk-devel.x86_64 : OpenJDK 11 Development Environment
java-17-openjdk-devel.x86_64 : OpenJDK 17 Development Environment
[root@localhost ~]#
```

En mi caso la última versión es la 17

1.4. Instalamos la versión 17 de java: sudo dnf install java-17.0-openjdk-devel

```
[root@localhost ~]# sudo dnf install java-17-openjdk-devel
```

1.5. Indicamos que la versión de java que queremos utilizar es la que acabamos de instalar, para ello: sudo alternatives --config java

Seleccionamos la opción en la que aparece la versión de java que instalamos, en mi caso la 1.

```
[root@localhost ~]# sudo alternatives --config java
There is 1 program that provides 'java'.

  Selection    Command
  -----
*+ 1          java-17-openjdk.x86_64 (/usr/lib/jvm/java-17-openjdk-17.0.1.0.12-2.el8_5.x86_64/bin/java)

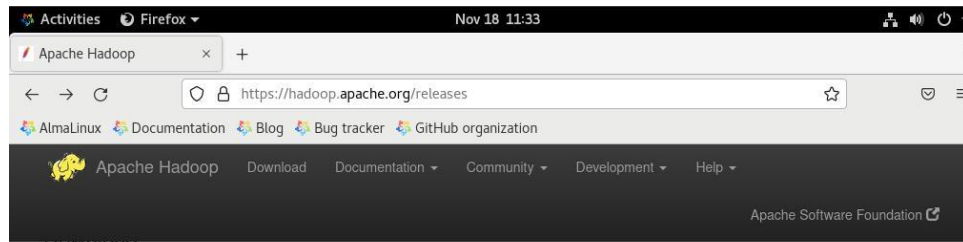
Enter to keep the current selection[+], or type selection number: 1
[root@localhost ~]#
```

1.6. Comprobamos que se instaló correctamente: java -version

```
[root@localhost ~]# java -version
openjdk version "17.0.1" 2021-10-19 LTS
OpenJDK Runtime Environment 21.9 (build 17.0.1+12-LTS)
OpenJDK 64-Bit Server VM 21.9 (build 17.0.1+12-LTS, mixed mode, sharing)
[root@localhost ~]#
```

2. Instalamos Hadoop:

2.1. Buscamos las versiones de Hadoop: <https://hadoop.apache.org/releases.html>



Hadoop is released as source code tarballs with corresponding binary tarballs for convenience. The downloads are distributed via mirror sites and should be checked for tampering using GPG or SHA-512.

Version	Release date	Source download	Binary download	Release notes
3.3.1	2021 Jun 15	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)	Announcement
3.2.2	2021 Jan 9	source (checksum signature)	binary (checksum signature)	Announcement
2.10.1	2020 Sep 21	source (checksum signature)	binary (checksum signature)	Announcement

To verify Hadoop releases using GPG:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the signature file `hadoop-X.Y.Z-src.tar.gz.asc` from [Apache](#).
3. Download the [Hadoop KEYS](#) file.
4. `gpg --import KEYS`
5. `gpg --verify hadoop-X.Y.Z-src.tar.gz.asc`

To perform a quick check using SHA-512:

1. Download the release `hadoop-X.Y.Z-src.tar.gz` from a [mirror site](#).
2. Download the checksum `hadoop-X.Y.Z-src.tar.gz.sha512` or `hadoop-X.Y.Z-src.tar.gz.mds` from [Apache](#).
3. `shasum -a 512 hadoop-X.Y.Z-src.tar.gz`

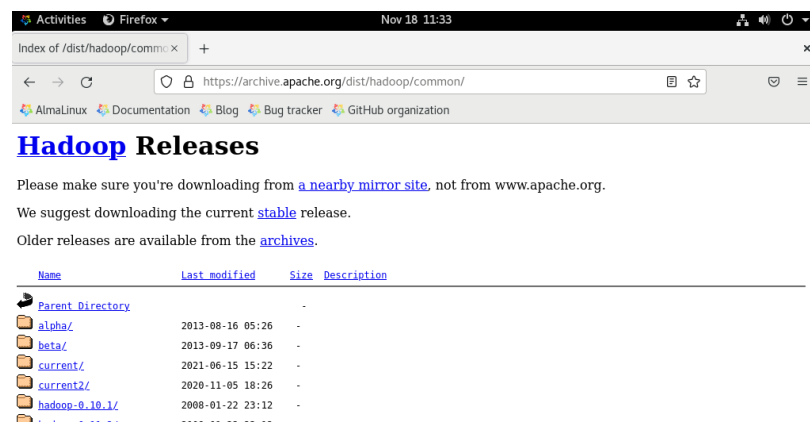
All previous releases of Hadoop are available from the [Apache release archive](#) site.

Many third parties distribute products that include Apache Hadoop and related tools. Some of these are listed on the [Distributions wiki page](#).

License

2.2. Se recomienda instalar los binarios de una versión estable, por ejemplo la 3.3.0

2.2.1. Clickeamos en “Apache relase archive”. Tendremos:



2.2.2. Entramos en la carpeta de la que vamos a descargar:

Activities Firefox Nov 18 13:16

Index of /dist/hadoop/commo x +

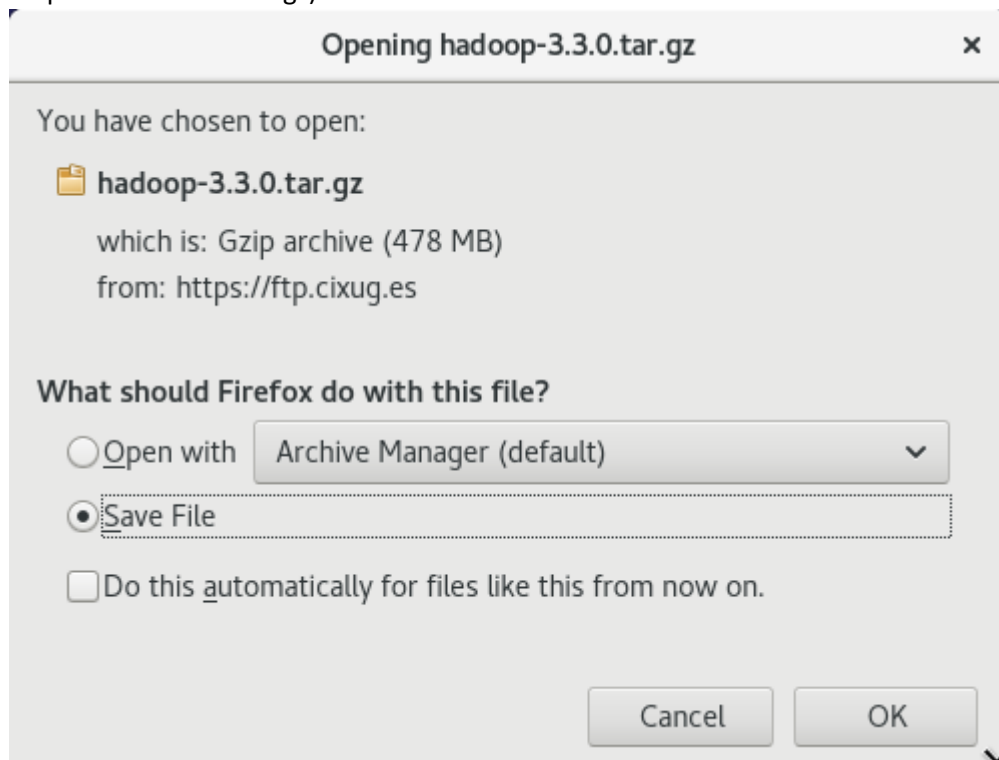
https://archive.apache.org/dist/hadoop/common/hadoop-3.3.0/

AlmaLinux Documentation Blog Bug tracker GitHub organization

Index of /dist/hadoop/common/hadoop-3.3.0

Name	Last modified	Size	Description
Parent Directory	-	-	-
CHANGELOG.md	2020-07-15 17:05	376K	
CHANGELOG.md.asc	2020-07-15 17:05	819	
CHANGELOG.md.sha512	2020-07-15 17:05	153	
RELEASENOTES.md	2020-07-15 17:05	26K	
RELEASENOTES.md.asc	2020-07-15 17:05	819	
RELEASENOTES.md.sha512	2020-07-15 17:05	156	
hadoop-3.3.0-aarch64.tar.gz	2020-07-15 17:19	478M	
hadoop-3.3.0-aarch64.tar.gz.asc	2020-07-15 17:19	819	
hadoop-3.3.0-aarch64.tar.gz.sha512	2020-07-15 17:19	168	
hadoop-3.3.0-rat.txt	2020-07-15 17:05	2.0M	
hadoop-3.3.0-rat.txt.asc	2020-07-15 17:05	819	
hadoop-3.3.0-rat.txt.sha512	2020-07-15 17:05	161	
hadoop-3.3.0-site.tar.gz	2020-07-15 17:33	40M	
hadoop-3.3.0-site.tar.gz.asc	2020-07-15 17:33	819	
hadoop-3.3.0-site.tar.gz.sha512	2020-07-15 17:33	165	
hadoop-3.3.0-src.tar.gz	2020-07-15 17:05	32M	
hadoop-3.3.0-src.tar.gz.asc	2020-07-15 17:05	819	
hadoop-3.3.0-src.tar.gz.sha512	2020-07-15 17:05	164	
hadoop-3.3.0.tar.gz	2020-07-15 17:30	478M	
hadoop-3.3.0.tar.gz.asc	2020-07-15 17:30	819	
hadoop-3.3.0.tar.gz.sha512	2020-07-15 17:30	160	

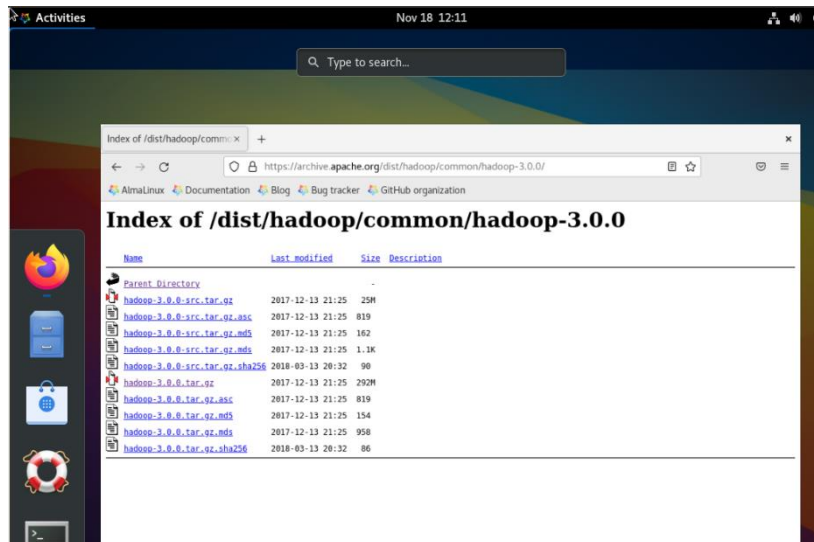
2.2.3. Descargamos los binarios, para ello pulsamos en la opción .tar.gz (IMPORTANTE: no pulsar en la -src.tar.gz):



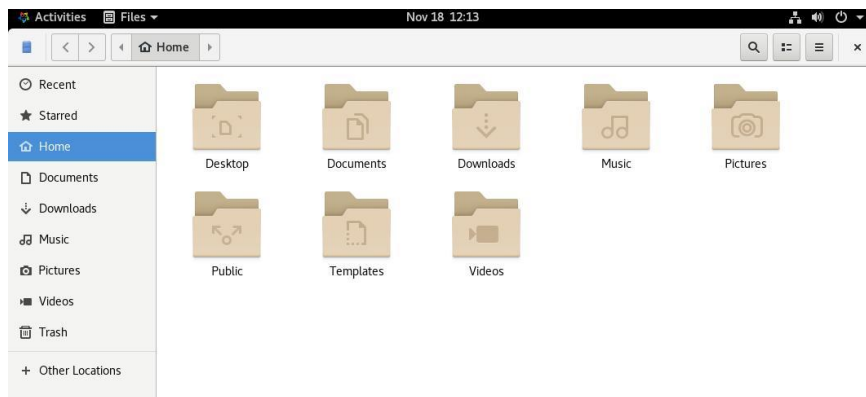
Nos da la opción de abrir o guardar. Le damos a guardar.

2.3. El archivo se descargó en la carpeta Descargas (o Download si estáis en inglés). Para verlo (notar que estos pasos pueden variar dependiendo de la configuración de gnome):

2.3.1. Ponemos el ratón en la parte izquierda superior. Tendremos:

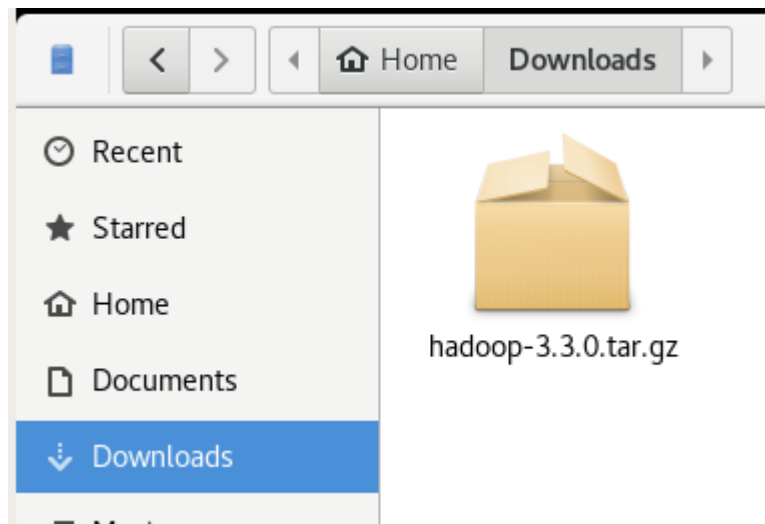


2.3.2. Ha salido un nuevo menú en la parte izquierda. Pulsamos en el archivador Azul, tendremos:



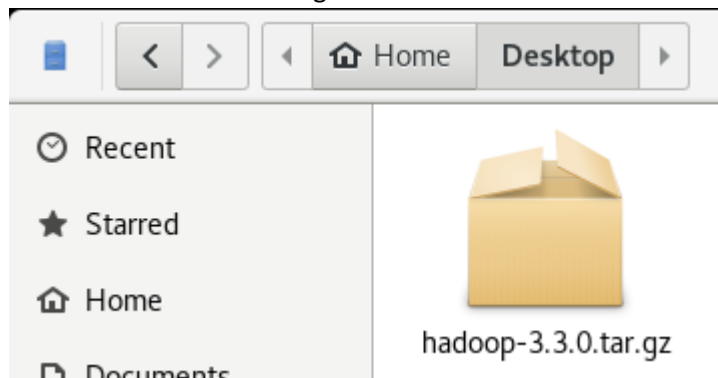
Desde aquí ya nos podemos situar en las carpetas Descargas (Downloads), Escritorio (Desktop, etc)

2.3.3. Entramos en descargas:



2.4. Cortamos el archivo y lo pegamos en la carpeta escritorio

2.5. Movemos el archivo .tar.gz al escritorio:



2.6. Lo descomprimos:

```
cd /home/moranjesus/Desktop
tar xzf hadoop-3.3.0.tar.gz
```

```
[root@localhost ~]# cd /home/moranjesus/Desktop/
[root@localhost Desktop]# tar xzf hadoop-3.3.0.tar.gz
[root@localhost Desktop]#
```

2.7. Movemos la carpeta a /usr/local, para ello: mv hadoop-3-3-0 /usr/local

```
[root@localhost Desktop]# mv hadoop-3.3.0 /usr/local/
```

Puede que tarde unos minutos

2.8. Indicamos que el propietario sea nuestro usuario:

```
sudo chown -R moranjesus:moranjesus /usr/local/hadoop-3.3.0
[root@localhost Desktop]# sudo chown -R moranjesus:moranjesus /usr/local/hadoop-3.3.0
[root@localhost Desktop]#
```

3. Habilitamos ssh:

3.1. Salimos de root de la terminal: exit:

```
[root@localhost Desktop]# exit
logout
[moranjesus@localhost ~]$
```

3.2. Generamos una clave ssh para nuestro usuario: ssh-keygen -t rsa -P ""

```
[moranjesus@localhost ~]$ ssh-keygen -t rsa -P ""
```

Nos pide un archivo en el que guardar, le damos enter:

```
[moranjesus@localhost ~]$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/moranjesus/.ssh/id_rsa):
Created directory '/home/moranjesus/.ssh'.
Your identification has been saved in /home/moranjesus/.ssh/id_rsa.
Your public key has been saved in /home/moranjesus/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:Xyj/9dwc6UAzYURbGT4m6bvxpEsHC4ZUS2/wCpSo6YI moranjesus@localhost.localdomain
The key's randomart image is:
+---[RSA 2048]---+
|      . . . + . o  oo|
|      . . . o * = .  |
|      o  o . X +     |
|      o  . o . = + .  |
|      . . S . . + *   |
| E . .   + . . o * .  |
|      .      o * . =   |
|      . . . X + o     |
|      . . . + . o =   |
+-----[SHA256]-----+
[moranjesus@localhost ~]$ █
```

- 3.3. Indicamos que se pueden realizar conexiones vía ssh con la anterior clave:

```
cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
```

```
[moranjesus@localhost ~]$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys █
```

- 3.4. Eliminamos los permisos que tengamos en authorized_keys: chmod -R go= ~/.ssh

```
[moranjesus@localhost ~]$ chmod -R go= ~/.ssh
[moranjesus@localhost ~]$
```

- 3.5. Indicamos que la capeta .ssh pertenezca a nuestro usuario:

```
chown -R moranjesus:moranjesus ~/.ssh
[moranjesus@localhost ~]$ chown -R moranjesus:moranjesus ~/.ssh/
[moranjesus@localhost ~]$
```

- 3.6. Hacemos una conexión vía ssh para que se registre en hosts conocidos: ssh localhost

```
[moranjesus@localhost ~]$ ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
ECDSA key fingerprint is SHA256:/0zIQ4YJCYqeDVONRqLFdltUAtaqkdzLLrhaToCntME.
ECDSA key fingerprint is MD5:08:86:79:18:63:27:1f:c4:de:81:87:b2:d0:64:13:0b.
Are you sure you want to continue connecting (yes/no)?
```

Indicamos que yes

```
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Last login: Tue Nov 20 18:57:00 2018
[moranjesus@localhost ~]$
```

- 3.7. Salimos de la conexión: exit

```
[moranjesus@localhost ~]$ exit
logout
Connection to localhost closed.
[moranjesus@localhost ~]$
```

4. Actualizamos las variables de entorno:

- 4.1. Buscamos donde está java instalado: which javac

```
[moranjesus@localhost ~]$ which javac
/usr/bin/javac
[moranjesus@localhost ~]$ █
```

- 4.2. Buscamos el enlace al que nos lleva: readlink -f /usr/bin/javac

```
[moranjesus@localhost ~]$ readlink -f /usr/bin/javac
/usr/lib/jvm/java-17-openjdk-17.0.1.0.12-2.el8_5.x86_64/bin/javac
[moranjesus@localhost ~]$ █
```

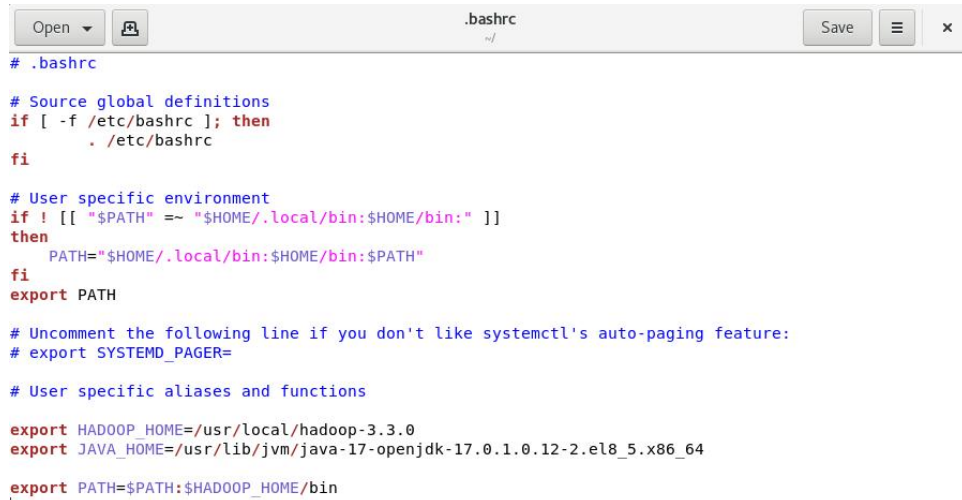
Copiamos la ubicación de nuestro java, es decir, lo que nos muestra antes de "/bin/javac", es decir: /usr/lib/jvm/java-17-openjdk-17.0.1.12-2.el8_5.x86_64

- 4.3. Editamos el archivo bashrc, para ello: gedit ~/.bashrc

```
[moranjesus@localhost ~]$ gedit ~/.bashrc
```

- 4.4. Añadimos:

```
export HADOOP_HOME=/usr/local/hadoop-3.3.0
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-17.0.1.12-2.el8_5.x86_64
export PATH=$PATH:$HADOOP_HOME/bin
```



```
# .bashrc

# Source global definitions
if [ -f /etc/bashrc ]; then
    . /etc/bashrc
fi

# User specific environment
if ! [[ "$PATH" =~ "$HOME/.local/bin:$HOME/bin:" ]]
then
    PATH="$HOME/.local/bin:$HOME/bin:$PATH"
fi
export PATH

# Uncomment the following line if you don't like systemctl's auto-paging feature:
# export SYSTEMD_PAGER=

# User specific aliases and functions

export HADOOP_HOME=/usr/local/hadoop-3.3.0
export JAVA_HOME=/usr/lib/jvm/java-17-openjdk-17.0.1.12-2.el8_5.x86_64

export PATH=$PATH:$HADOOP_HOME/bin
```

- 4.5. Pulsamos en guardar (botón Save) y cerramos la aplicación gedit

- 4.6. Desde la terminal indicamos que actualice la configuración: source ~/.bashrc

```
[moranjesus@localhost ~]$ source ~/.bashrc
```

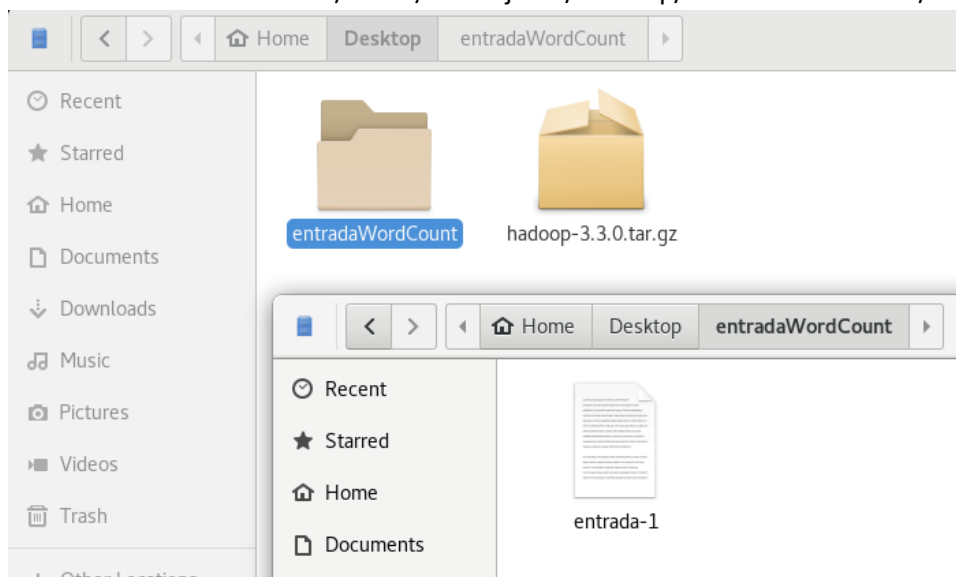
5. Comprobamos que Hadoop se instaló bien: hadoop version

```
[moranjesus@localhost ~]$ hadoop version
Hadoop 3.3.0
Source code repository https://gitbox.apache.org/repos/asf/hadoop.git -r aa96f1871bfd858f9bac59cf2a81ec470da649a
f
Compiled by brahma on 2020-07-06T18:44Z
Compiled with protoc 3.7.1
From source with checksum 5dc29b802d6ccd77b262ef9d04d19c4
This command was run using /usr/local/hadoop-3.3.0/share/hadoop/common/hadoop-common-3.3.0.jar
[moranjesus@localhost ~]$
```

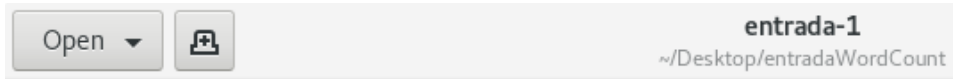
6. Creamos una carpeta llamada entradaWordCount, para ello:

```
mkdir /home/moranjesus/Desktop/entradaWordCount
```

7. Creamos un archivo: touch /home/moranjesus/Desktop/entradaWordCount/entrada-1



8. Damos doble click al archivo y le añadimos varias frases:



Esto es una línea de prueba
segunda línea de prueba
Podemos incluir las líneas que queramos
esta es la última línea

(IMPORTANTE: nunca dejar una línea vacía porque sino los programas pueden que os fallen)

9. Ejecutamos un programa MapReduce de ejemplo para que nos cuente las veces que aparece cada palabra:

```
hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar wordcount /home/moranjesus/Desktop/entradaWordCount /home/moranjesus/Desktop/salidaWordCount
```

```
[moranjesus@localhost ~]$ hadoop jar $HADOOP_HOME/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.0.jar wordcount /home/moranjesus/Desktop/entradaWordCount /home/moranjesus/Desktop/salidaWordCount
```

```
File System Counters
  FILE: Number of bytes read=604866
  FILE: Number of bytes written=1366458
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
Map-Reduce Framework
  Map input records=4
  Map output records=21
  Map output bytes=200
  Map output materialized bytes=193
  Input split bytes=121
  Combine input records=21
  Combine output records=16
  Reduce input groups=16
  Reduce shuffle bytes=193
  Reduce input records=16
  Reduce output records=16
  Spilled Records=32
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=86
  Total committed heap usage (bytes)=331489280
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
```

10. Vemos el archivo de salida. Para ello doble click en la carpeta salida SalidaWordCount y el archivo part-r-00000

