



EDUCACIÓN

SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO®

Instituto Tecnológico de San Juan del Río

Departamento de Sistemas y Computación

"2019, +Año del Caudillo del Sur, Emiliano Zapata"

Instituto Tecnológico de San Juan del Río



Apache spark y hadoop

Luna González Rocio
Mendoza Trejo Jairo
Robles Padilla Oswaldo
Valencia Valencia Mauricio

Ing. Sistemas computacionales

San Juan del Río, Qro., 24 de Noviembre de 2020

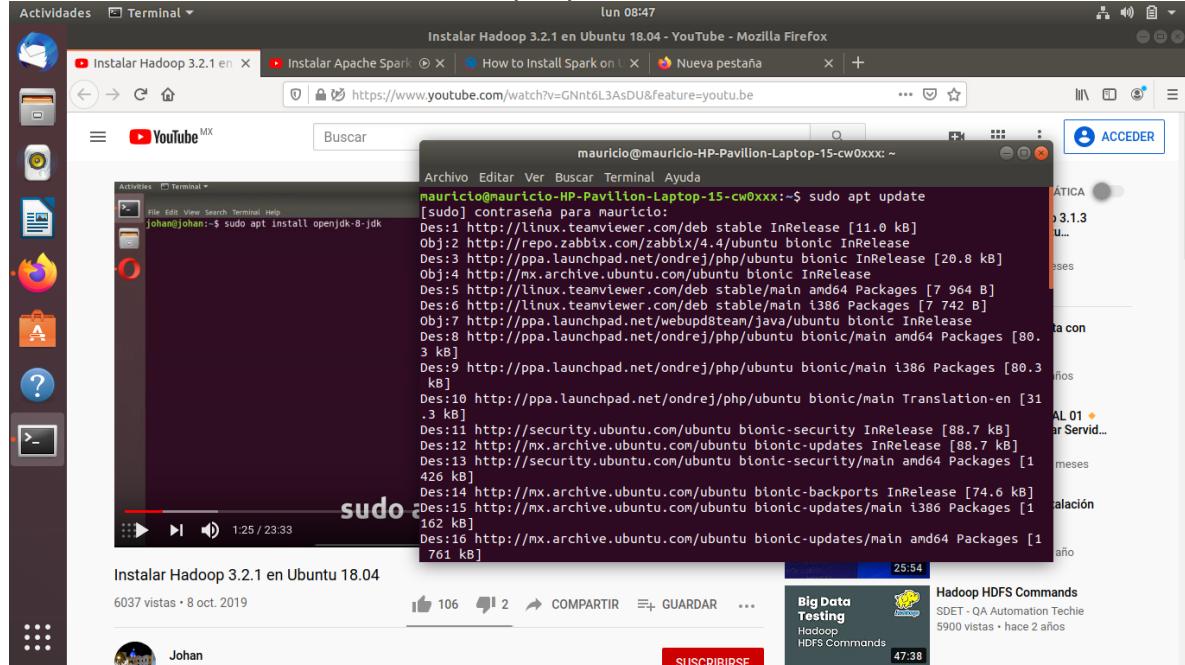


Av. Tecnológico # 2, C.P. 76800, San Juan del Río, Querétaro, México
Tels. 01 (427) 27 24118, 01 (427) 27 24178 Ext. 119 e-mail:
depto.sistemas@itsanjuan.edu.mx
www.tecnm.mx | www.itsanjuan.edu.mx

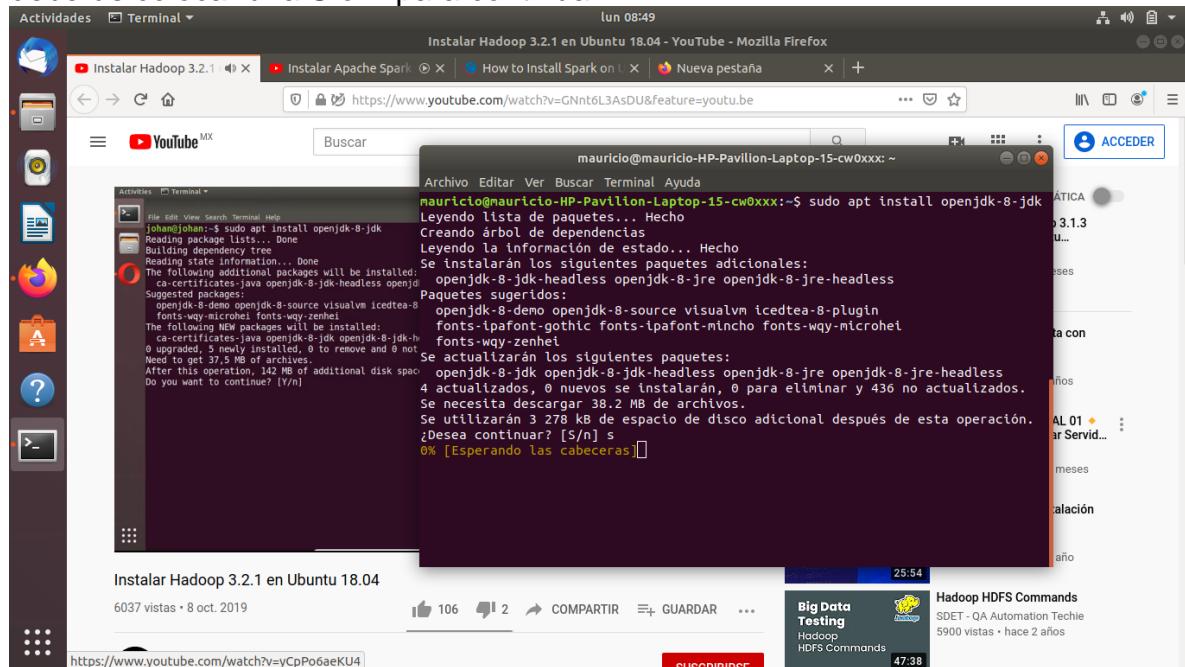


INSTALACIÓN DE APACHE SPARK Y HADOOP

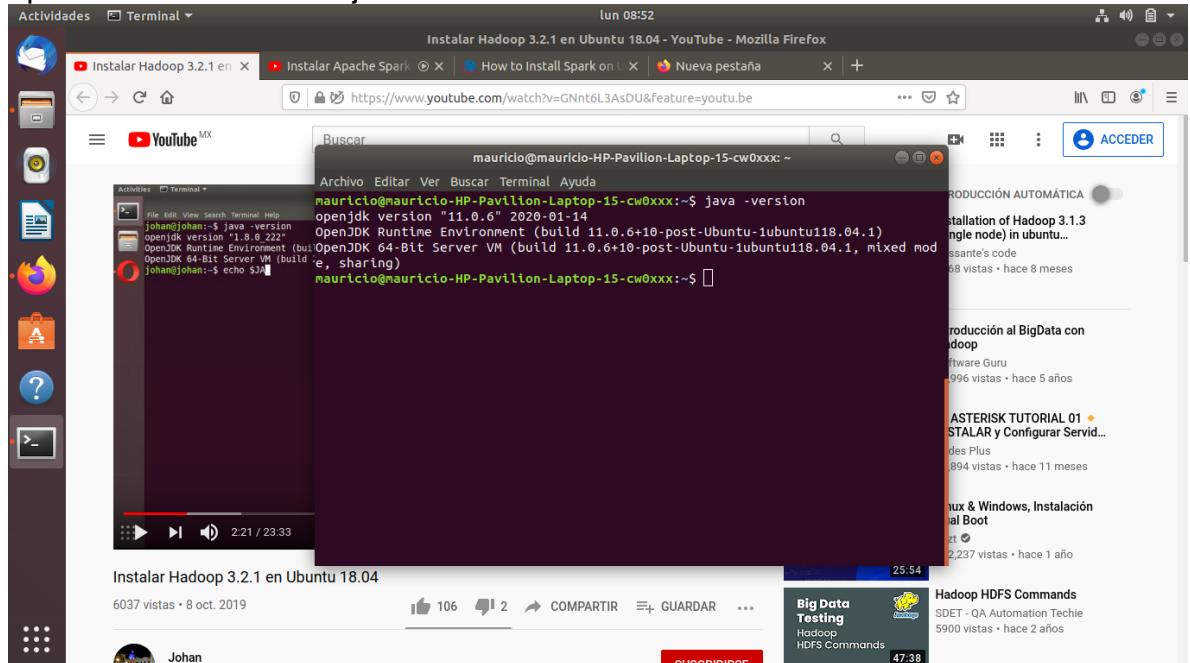
1. Para instalar apache spark primer paso tendremos que actualizar nuestro sistema usando el comando sudo apt update



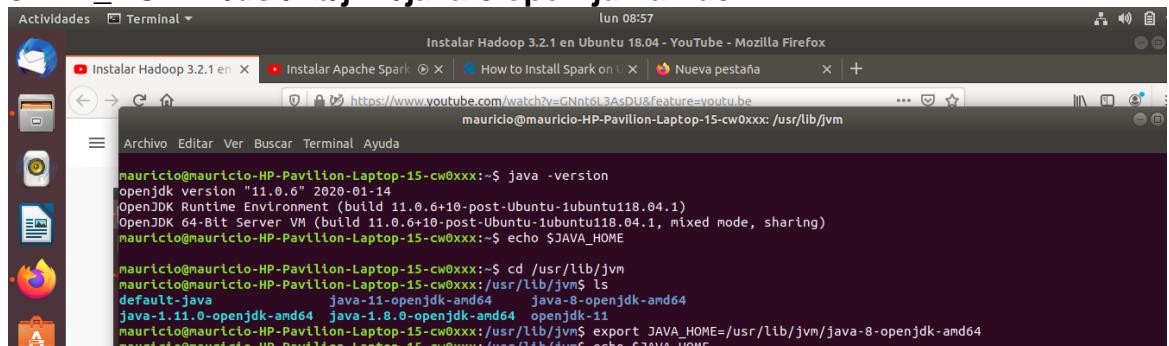
2. Una vez actualizado la terminal tendremos que instalar el java 8 ya que hadoop utiliza esta versión para funcionar correctamente para esto usaremos el comando **sudo apt install openjdk-8-jdk** aparecerá un mensaje donde deberás colocar una S o Y para continuar



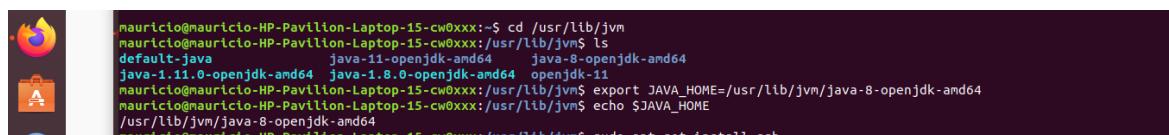
- Una vez instalado verificaremos que la instalación se realizara de manera correcta esto con el comando **java -version** si se instaló correctamente aparecerá la version de java instalada



- Verificar que java este en el patch con el comando **-\$ echo \$JAVA_HOME** debería estar vacío para poder agregarlo se usara lo siguiente **-\$ cd /usr/lib/jvm** una vez dentro del directorio listamos los archivos con ls ahora agregaremos el archivo al patch usando la ruta **export JAVA_HOME=/usr/lib/jvm/java-8.open-jdk-amd64**



- verificar con el comando **-\$echo \$JAVA_HOME** esta vez nos aparecerá listado de la siguiente manera



6. Como siguiente paso se instala ssh para el acceso de servidores remotos esto se hará con el comando **sudo apt-get install ssh**

```

Actividades Terminal lun 08:57
Instalar Hadoop 3.2.1 en Ubuntu 18.04 - YouTube - Mozilla Firefox
https://www.youtube.com/watch?v=GNnt6L3AsDU&feature=youtu.be
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx: /usr/lib/jvm

Archivo Editar Ver Buscar Terminal Ayuda
/usr/lib/jvm/java-8-openjdk-amd64
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/usr/lib/jvm$ sudo apt-get install ssh
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
Se instalarán los siguientes paquetes adicionales:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
Paquetes sugeridos:
  molly-guard monkeysphere rssh ssh-askpass
Se instalarán los siguientes paquetes NUEVOS:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
0 actualizados, 5 nuevos se instalarán, 0 para eliminar y 436 no actualizados.
Se necesita descargar 642 kB de archivos.
Se utilizarán 5 422 kB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] s
Des:1 http://mx.archive.ubuntu.com/ubuntu bionic/main amd64 openssh-sftp-server amd64 1:7.6p1-4ubuntu0.3 [45.6 kB]
Des:2 http://mx.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-server amd64 1:7.6p1-4ubuntu0.3 [333 kB]
Des:3 http://mx.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ssh all 1:7.6p1-4ubuntu0.3 [5 204 kB]
Des:4 http://mx.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ncurses-term all 6.1-1ubuntu1.18.04 [248 kB]
Des:5 http://mx.archive.ubuntu.com/ubuntu bionic-updates/main amd64 ssh-import-id all 5.7-0ubuntu1.1 [10.9 kB]
Descargados 642 kB en 3s (224 kB/s)
Preconfigurando paquetes ...
Seleccionando el paquete openssh-sftp-server previamente no seleccionado.
(Leyendo la base de datos ... 185295 ficheros o directorios instalados actualmente.)
Preconfigurando paquetes ...
Paquetes sugeridos:
  molly-guard monkeysphere rssh ssh-askpass
Se instalarán los siguientes paquetes adicionales:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
0 actualizados, 5 nuevos se instalarán, 0 para eliminar y 436 no actualizados.
Se necesita descargar 642 kB de archivos.
Se utilizarán 5 422 kB de espacio de disco adicional después de esta operación.
¿Desea continuar? [S/n] s
106 2 COMPARTE GUARDAR SUSCRIBIRSE
Fazt 25:54 302,237 vistas • hace 1 año
Big Data Testing Hadoop HDFS Commands
Hadoop HDFS Commands 47:38

```

7. Una vez descargado ssh deberemos instalar psdh para comandos remotos en paralelo con el comando **sudo apt-get install pdsh**

```

Actividades Terminal lun 09:00
Instalar Hadoop 3.2.1 en Ubuntu 18.04 - YouTube - Mozilla Firefox
https://www.youtube.com/watch?v=GNnt6L3AsDU&feature=youtu.be
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx: /usr/lib/jvm

Archivo Editar Ver Buscar Terminal Ayuda
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/usr/lib/jvm$ sudo apt-get install pdsh
Leyendo lista de paquetes... Hecho
Creando árbol de dependencias
Leyendo la información de estado... Hecho
Reading package list... done
Building dependency tree... done
Reading package information... done
The following additional packages will be installed:
  genders libgenders0
Paquetes sugeridos:
  genders libgenders0
Suggested packages:
  pdsh
  The following NEW packages will be installed:
    genders libgenders0 pdsh
  0 upgraded, 3 new
  0 actualizados, 3 nuevos se instalarán, 0 para eliminar y 436 no actualizados.
  Need to get 170 kB of archives.
  After this operation, 479 kB of additional space will be used.
  Do you want to continue? [S/n] s
  ¿Desea continuar? [S/n] s
106 2 COMPARTE GUARDAR SUSCRIBIRSE
Fazt 25:54 302,237 vistas • hace 1 año
Big Data Testing Hadoop HDFS Commands
Hadoop HDFS Commands 47:38

```

8. Cuando finalice iremos a nuestro navegador y entraremos a la página para <https://hadoop.apache.org/releases.html> descargar hadoop la version 3.2.1

Download

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
2.10.1	2020 Sep 21	source (checksum signature)	binary (checksum signature)	Announcement
3.1.4	2020 Aug 3	source (checksum signature)	binary (checksum signature)	Announcement
3.3.0	2020 Jul 14	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature)	Announcement
3.2.1	2019 Sep 22	source (checksum signature)	binary (checksum signature)	Announcement
2.9.2	2018 Nov 19	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:
<https://www.apache.org/dyn/closer.cgi/hadoop/common/hadoop-2.10.1/hadoop-2.10.1.tar.gz>

9. Seleccionar la opción y nos abrirá una ventana como la siguiente donde descaremos el archivo automáticamente después de seleccionar la descarga

Apache Download Mirrors - Mozilla Firefox

We suggest the following mirror site for your download:
<https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz>

Other mirror sites are suggested below.
It is essential that you verify the integrity of the download.
Please only use the backup mirrors to download KEYS files.

HTTP
<https://downloads.apache.org/hadoop/common/>

BACKUP SITES
Please only use the backup mirrors to download KEYS files.
<https://downloads.apache.org/hadoop/common/>

The full listing of mirror sites is also available.

RECOMMENDED MIRRORS

Abriendo hadoop-3.2.1.tar.gz

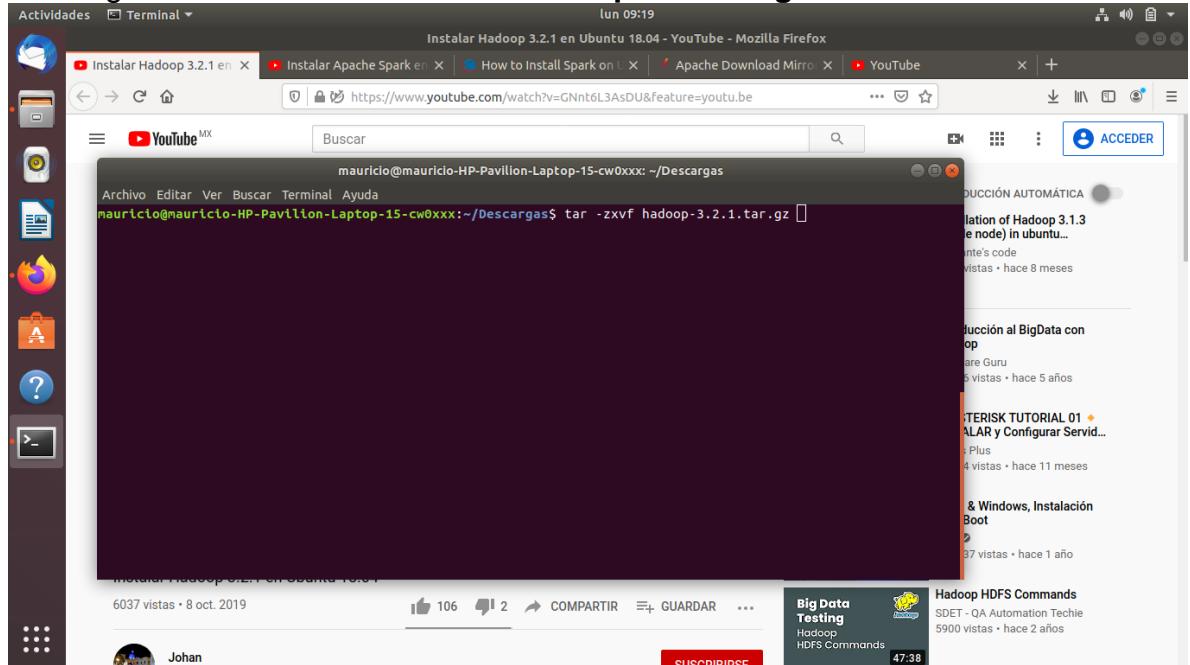
Ha elegido abrir:
 hadoop-3.2.1.tar.gz
el cual es un: archivador Gzip (343 MB)
de: <https://downloads.apache.org/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz>

¿Qué debería hacer Firefox con este archivo?

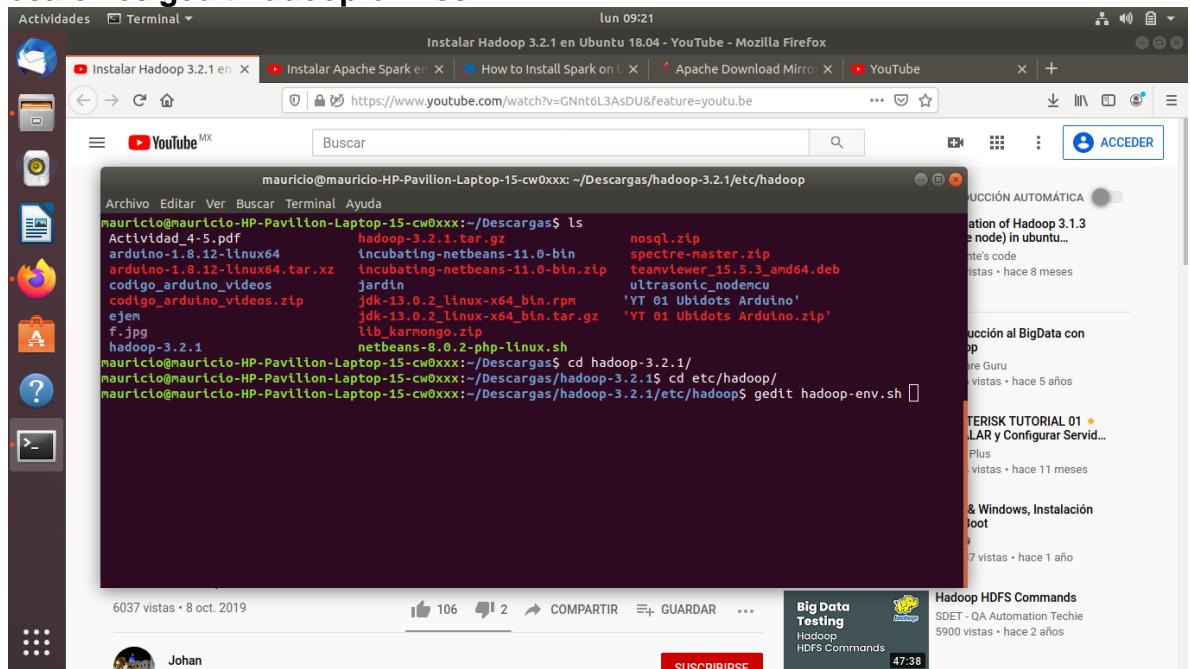
Abrir con **Gestor de archivadores (predeterminada)**
 Guardar archivo
 Repetir esta decisión de ahora en adelante para este tipo de archivos.

Aceptar **Cancelar**

10. Una vez descargado nos ubicamos en el directorio de descargas para ver el archivo que descargamos usando el comando CD descargas una vez ubicados en el directorio descomprimiremos el archivo que acabamos de descargar con el comando **tar -zxvf hadoop-3.2.1.tar.gz**



11. Al finalizar de descomprimir los archivos volvemos a listar el directorio con ls y veremos que ya existe una carpeta de hadoop ahora entraremos en la carpeta con **cd hadoop-3.2.1** tendremos que indicarle donde esta java para que pueda trabajar para eso usaremos la ruta **cd etc/hadoop/** una vez en esa ruta editaremos el archivo ssh mediante un editor de texto para eso usaremos **gedit hadoop-env.sh**



12. Nos abrirá el siguiente archivo donde podremos editar buscamos la palabra `expor` para localizar el `java_home` y colocar nuestra ruta “`usr/lib/jvm/java-8-openjdk-amd64`” y quitamos el comentario en el que se encuentra esa línea guardamos el archivo y salimos

```

# 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-8-openjdk-amd64"

# Location of Hadoop. By default, Hadoop will attempt to determine
# this location based upon its execution path.
#export HADOOP_HOME

# Location of Hadoop's configuration information. i.e., where this
# file is living. If this is not defined, Hadoop will attempt to
# locate it based upon its execution path.

# NOTE: It is recommend that this variable not be set here but in
# /etc/profile.d or equivalent. Some options (such as
# --config) may react strangely otherwise.
#export HADOOP_CONF_DIR=${HADOOP_HOME}/etc/hadoop

# The maximum amount of heap to use (Java -Xmx). If no unit
# is provided, it will be converted to MB. Daemons will
# prefer any Xmx setting in their respective _OPT variable.
# There is no default; the JVM will autoscale based upon machine
# memory size.
#export HADOOP_HEAPSIZE_MAX

# The minimum amount of heap to use (Java -Xms). If no unit
# is provided, it will be converted to MB. Daemons will
# prefer any Xms setting in their respective _OPT variable.
# There is no default; the JVM will autoscale based upon machine
# memory size.
#export HADOOP_HEAPSIZE_MIN

```

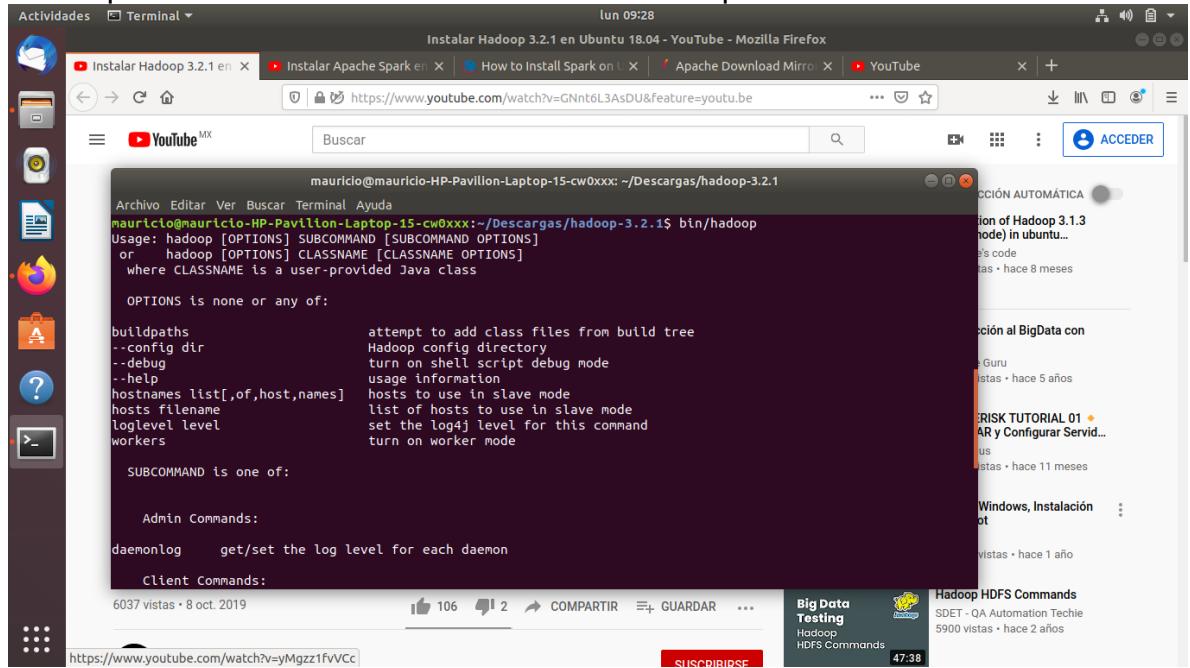
13. Regresamos a la terminal y regresamos a la carpeta principal de hadoop

```

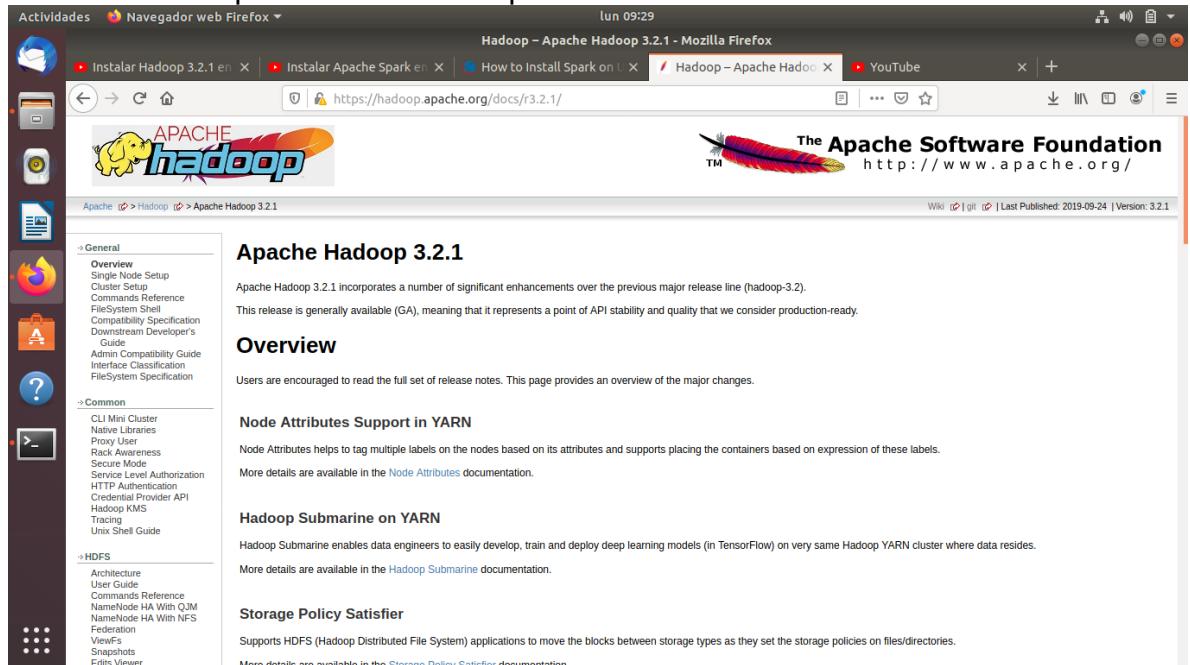
mauricio@mauricio-HP-Pavillion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1/etc/hadoop$ ls
Actividad_4-5.pdf          hadoop-3.2.1.tar.gz           nosql.zip
arduino-1.8.12-linux64      incubating-netbeans-11.0-bin   spectre-master.zip
arduino-1.8.12-linux64.tar.xz incubating-netbeans-11.0-bin.zip teamviewer_15.5.3_amd64.deb
codigo_arduino_videos         jardin                         ultrasonic_nodemcu
codigo_arduino_videos.zip    jdk-13.0.2_linux-x64_bin.rpm 'VT 01 Ubidots Arduino'
ejem                           jdk-13.0.2_linux-x64_bin.tar.gz 'VT 01 Ubidots Arduino.zip'
f.jpg                          lib_karmongo.zip
hadoop-3.2.1                 netbeans-8.0.2-php-linux.sh
mauricio@mauricio-HP-Pavillion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ cd hadoop-3.2.1/
mauricio@mauricio-HP-Pavillion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ cd etc/hadoop/
mauricio@mauricio-HP-Pavillion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1/etc/hadoop$ gedit hadoop-env.sh
mauricio@mauricio-HP-Pavillion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1/etc/hadoop$ 

```

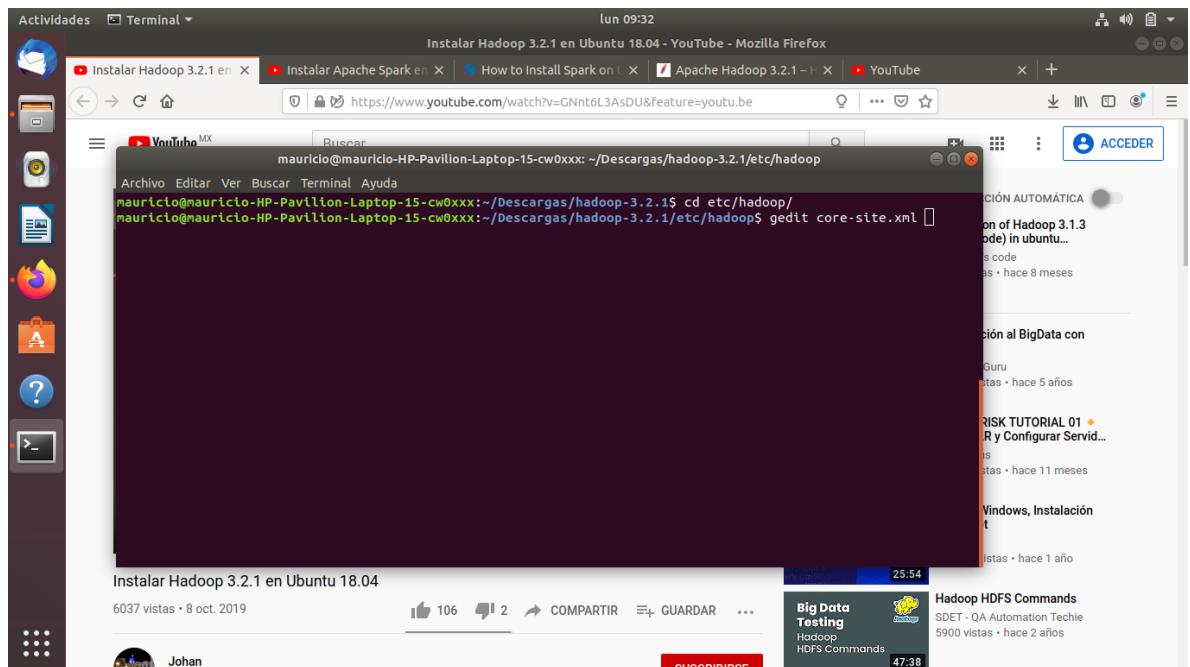
14. Una vez en la carpeta principal de hadoop colocaremos el comando bin /hadoop nos mostrara la documentación de hadoop



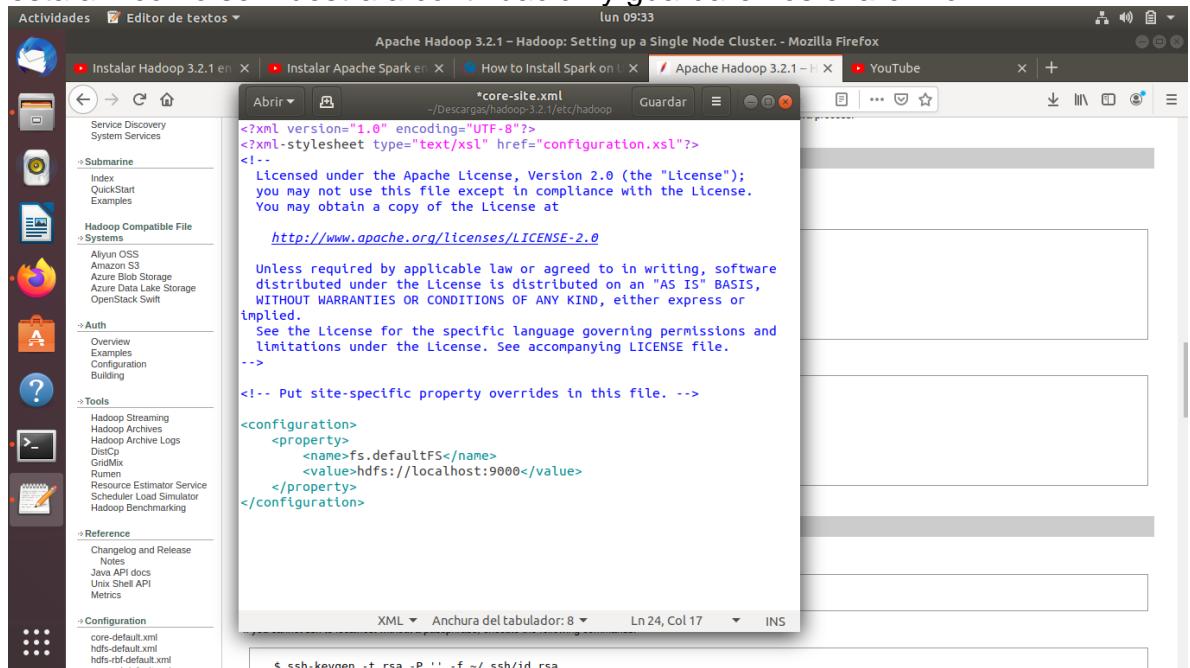
15. Una vez que nos mostró la documentación entraremos en la página de <https://hadoop.apache.org/docs/r3.2.1/> donde nos mostrara la documentación para instalar hadoop correctamente



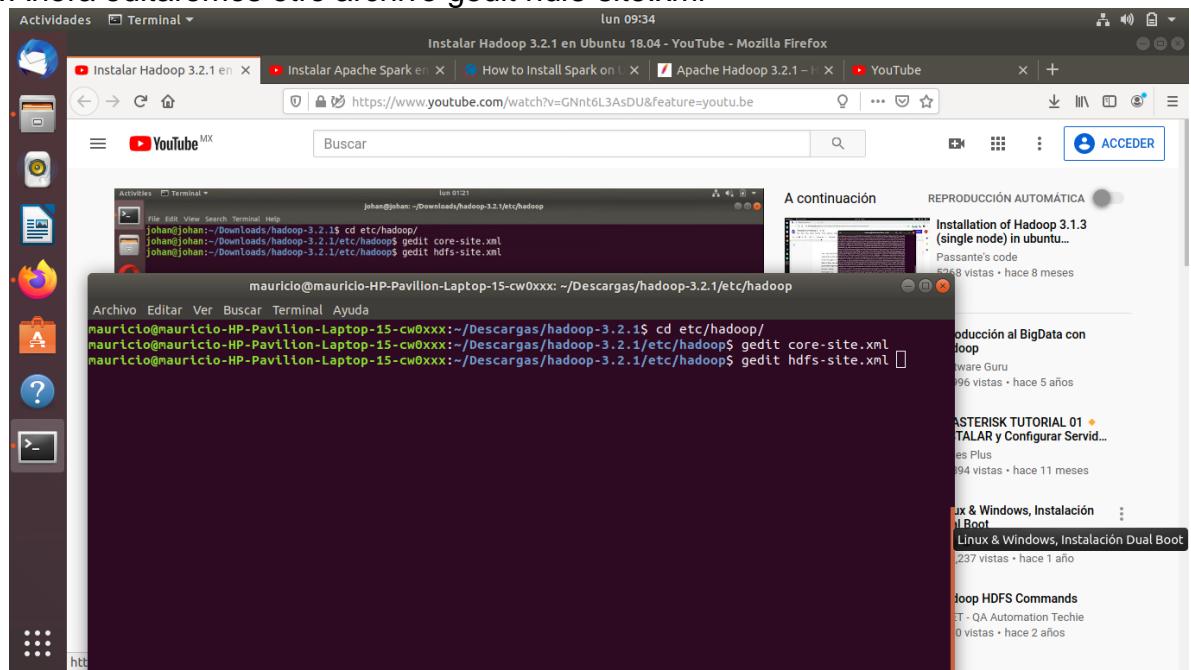
16. Entraremos al directorio **cd etc/hadoop** y editaremos el archivo core con el comando **gedit core-site.xml**



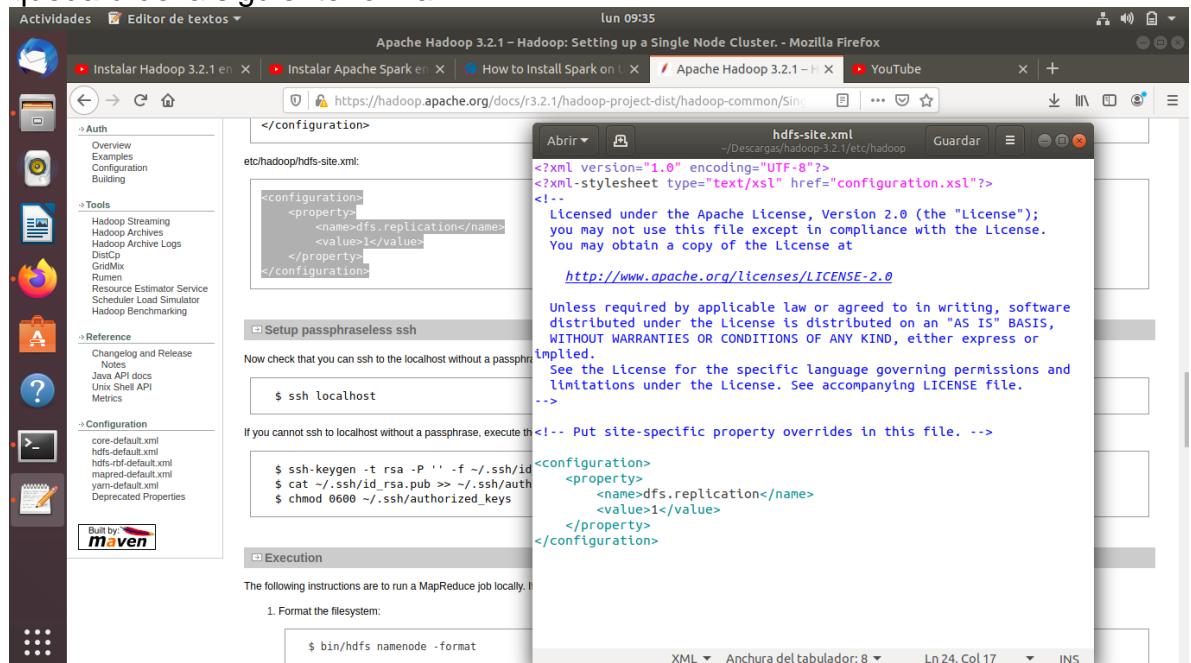
17. En la página vamos al archivo de core y copiamos y pegamos el código que está ahí como se muestra a continuación y guardaremos el archivo



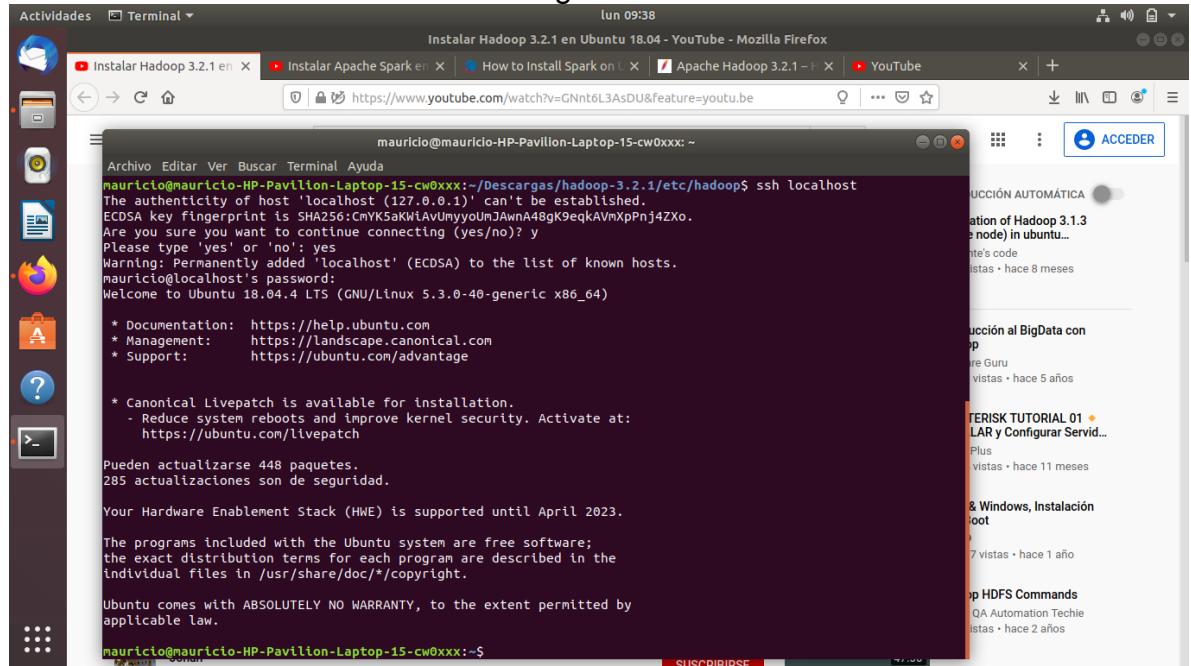
18. Ahora editaremos otro archivo gedit hdfs-site.xml



19. Y como en el anterior paso se editará usando el código de la página y quedará de la siguiente forma



20. Ahora entraremos al local usando el siguiente comando ssh localhost



```
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:ChyK5aKwIAVUmyoUnjAwnA48gk9eqKAVmXpPnj4ZXo.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
mauricio@localhost's password:
Welcome to Ubuntu 18.04.4 LTS (GNU/Linux 5.3.0-40-generic x86_64)

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

 * Canonical Livepatch is available for installation.
   - Reduce system reboots and improve kernel security. Activate at:
     https://ubuntu.com/livepatch

Pueden actualizarse 448 paquetes.
285 actualizaciones son de seguridad.

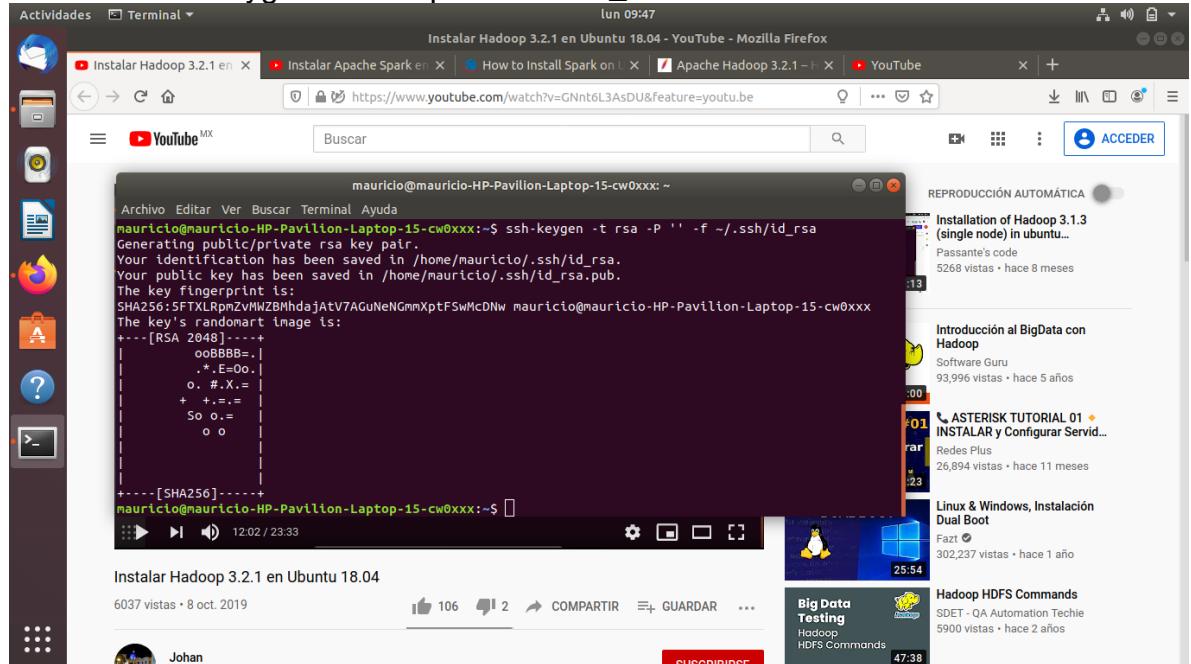
Your Hardware Enablement Stack (HWE) is supported until April 2023.

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
applicable law.

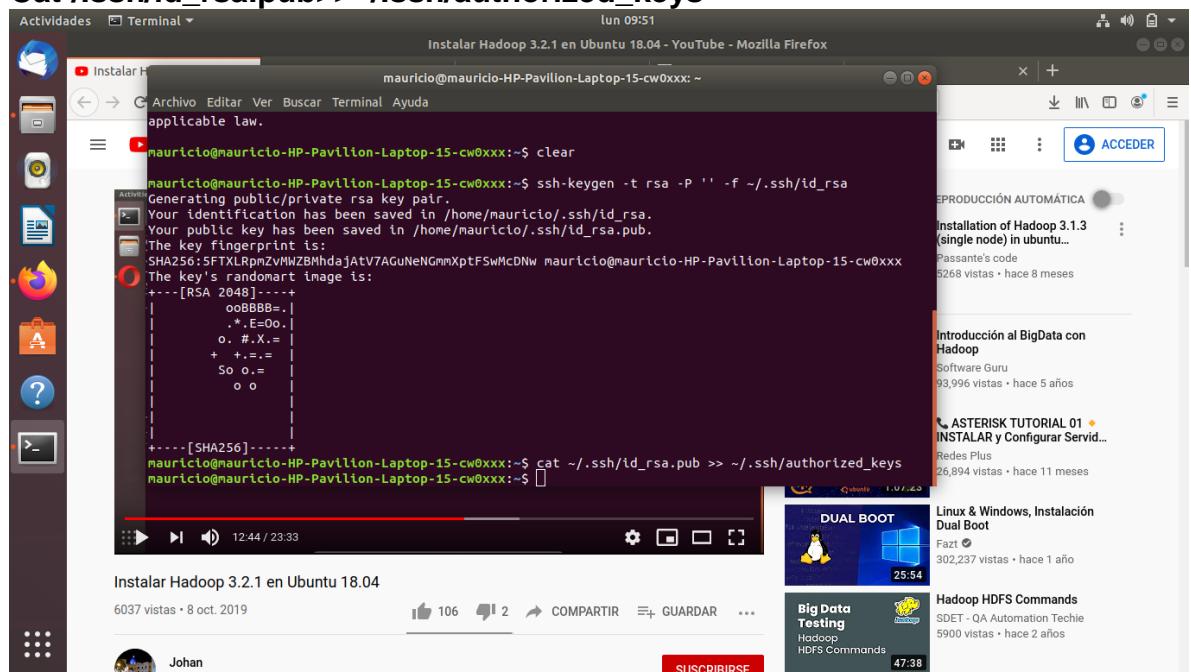
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$
```

21. Después de entrar al local host generaremos una llave con el siguiente comando ssh-keygen -t rsa -p “-f ./ssh/td_rsa



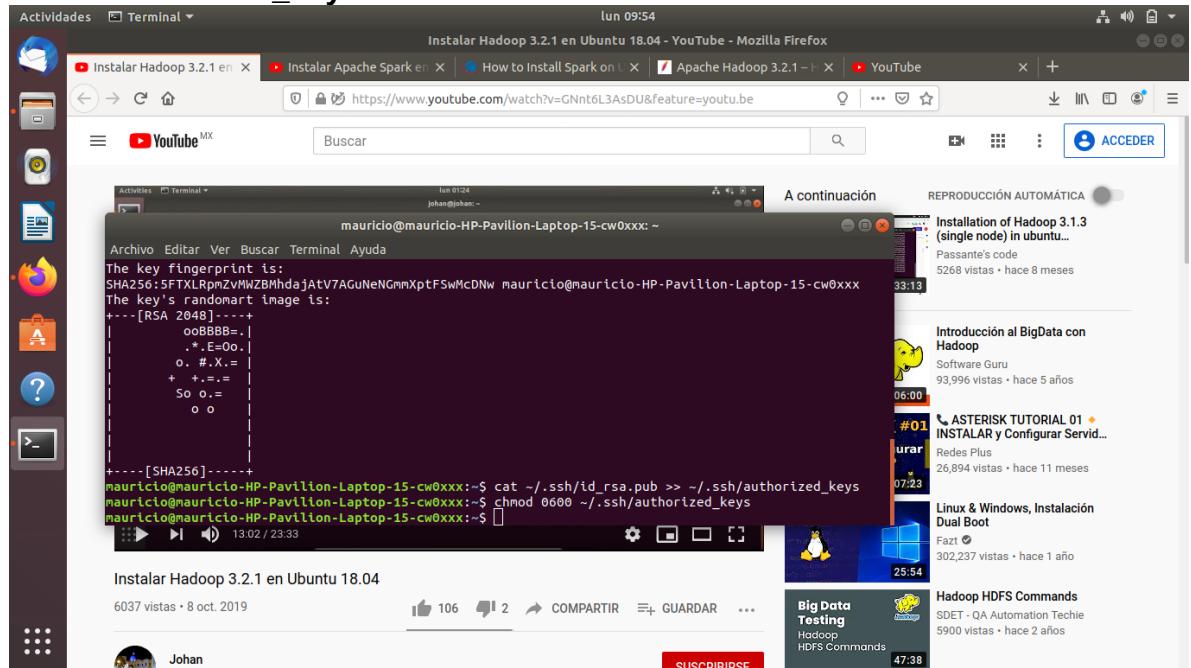
```
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ ssh-keygen -t rsa -P '' -f ./ssh/id_rsa
Generating public/private rsa key pair.
Your identification has been saved in /home/mauricio/.ssh/id_rsa.
Your public key has been saved in /home/mauricio/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:SFTXLRpzVzMWZBMMdajAtV7AGuNeNGmmXptFswMcNw mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx
The key's randomart image is:
+---[RSA 2048]----+
| ooBBBB=|
| .*.E=O=|
| o. #.X=|
| + .+=.=|
| So o.=|
| o o |
| |
| +---[SHA256]----+
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$
```

22. Usaremos el siguiente comando para continuar con la llave
Cat-./ssh/id_rsa.pub>>~/ssh/authorized_keys



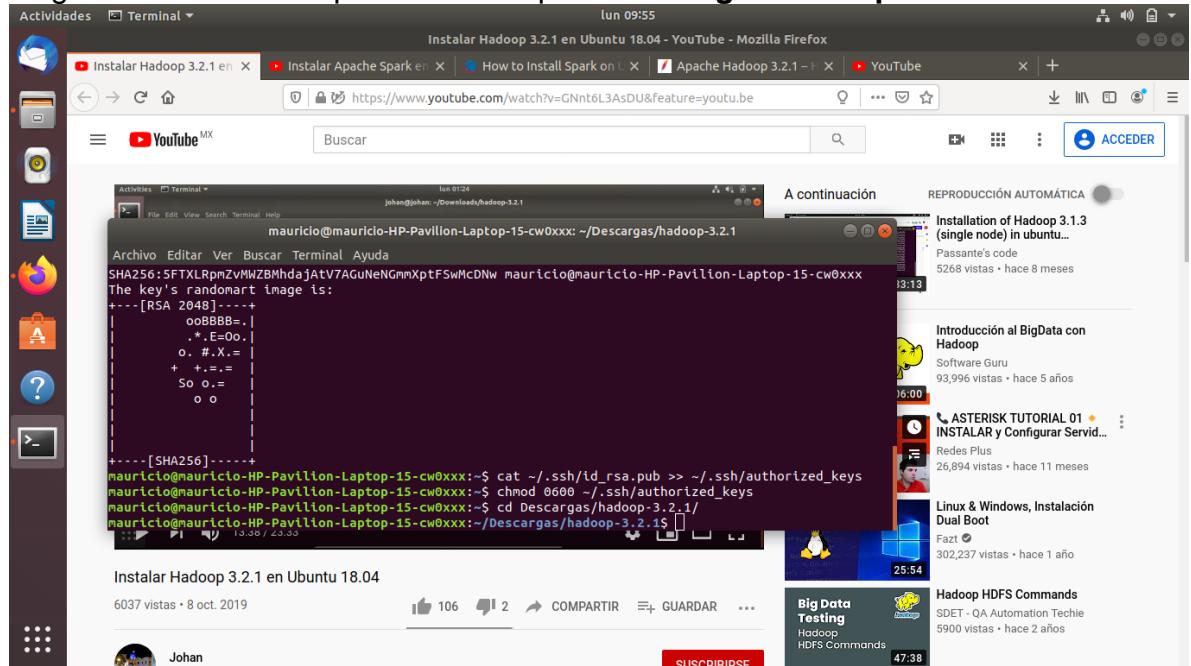
A screenshot of a Linux desktop environment. On the left is a dock with icons for various applications. In the center is a terminal window titled "Instalar Hadoop 3.2.1 en Ubuntu 18.04 - YouTube - Mozilla Firefox". The terminal shows the command "cat ./ssh/id_rsa.pub >> ~/ssh/authorized_keys" being run. To the right of the terminal is a YouTube video player for the video "Instalar Hadoop 3.2.1 en Ubuntu 18.04" by "Johan". The video has 6037 views and was uploaded on October 8th, 2019. Below the video player is a sidebar with other recommended videos.

23. Autorizaremos la llave con el comando chmod 0600
~/.ssh/authorized_keys

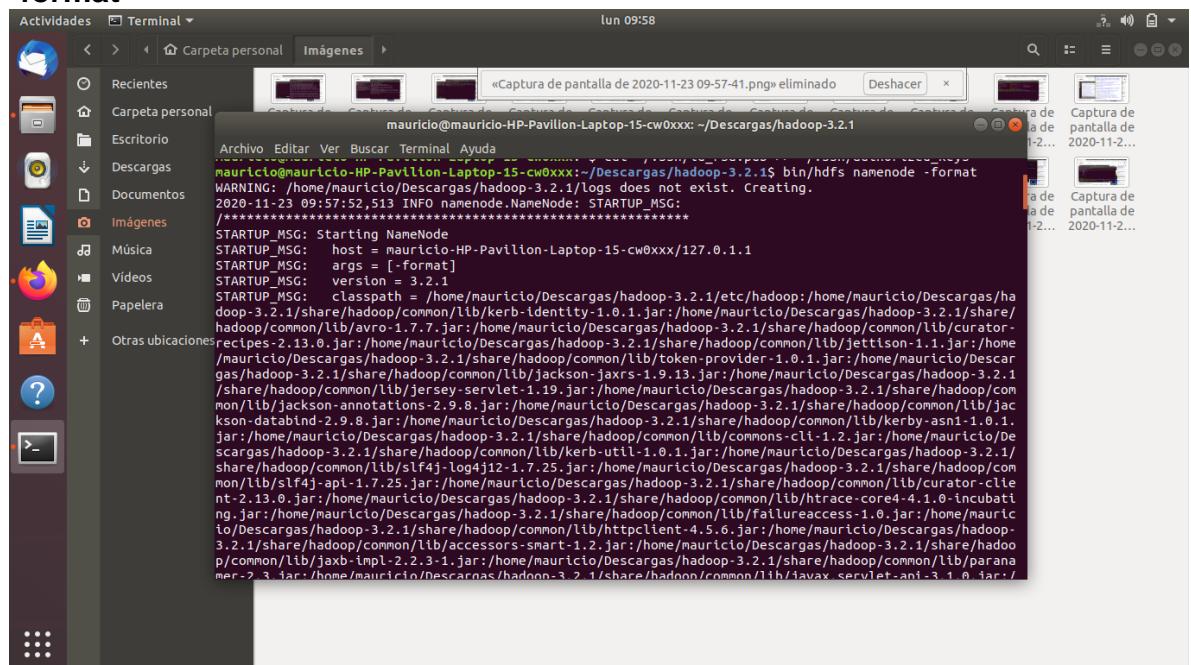


A screenshot of a Linux desktop environment, similar to the previous one. It shows a terminal window with the same command "cat ./ssh/id_rsa.pub >> ~/ssh/authorized_keys" and the addition of "chmod 0600 ~./ssh/authorized_keys". To the right is the same YouTube video player for the Hadoop installation guide. The video has 6037 views and was uploaded on October 8th, 2019. The sidebar also includes the same recommended videos as the first screenshot.

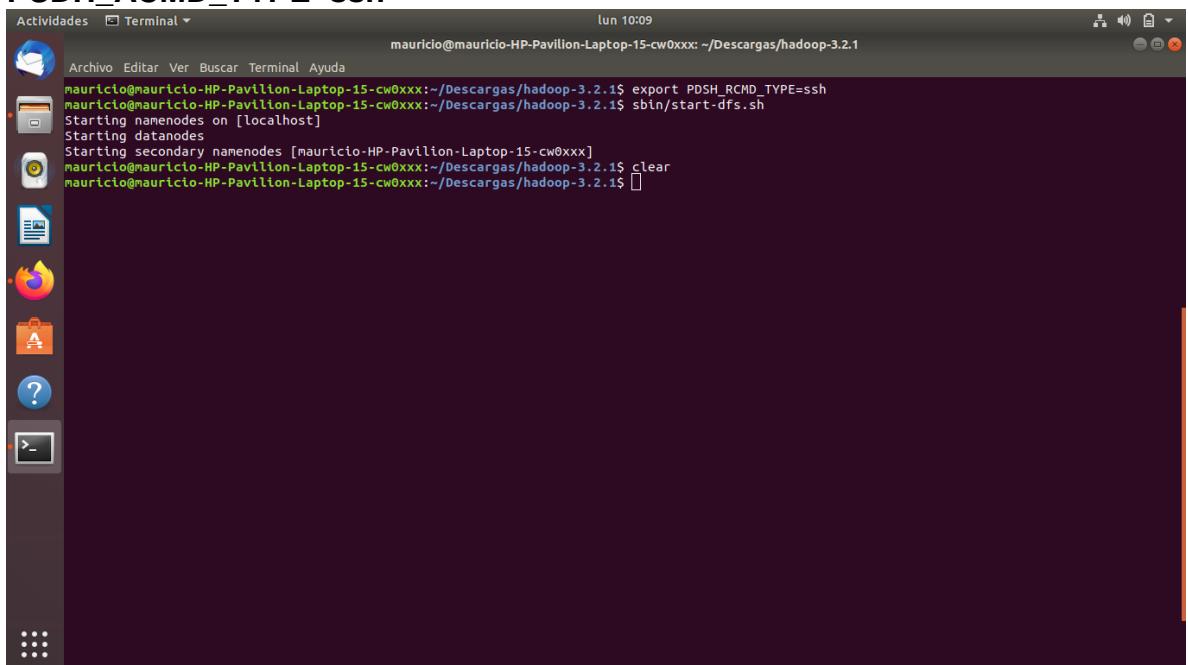
24. Regresaremos a la carpeta de hadoop **cd descargas/hadoop3.2.1**



25. Le daremos formato a nuestro disco usando el comando **bin/hdfs namenode -format**

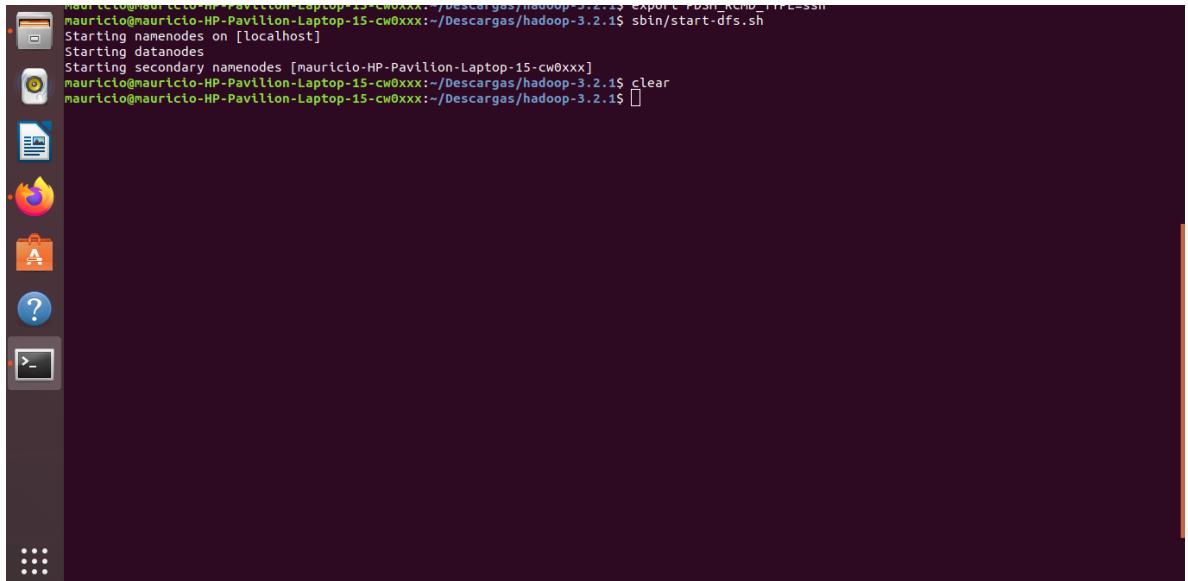


26. Usaremos el siguiente comando para seguir configurando el formato **export PSDH_ACMD_TYPE=ssh**



```
Actividades Terminal lun 10:09
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ export PSDH_ACMD_TYPE=ssh
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [mauricio-HP-Pavilion-Laptop-15-cw0xxx]
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ clear
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$
```

El siguiente paso es crear los demonios con el comando **sbin/start.dfs.sh** y con esto iniciaran los demonios del local host



```
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ sbin/start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [mauricio-HP-Pavilion-Laptop-15-cw0xxx]
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ clear
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$
```

27. Ahora en el navegador entraremos al local host de hadoop **localhost9870**

The screenshot shows a Firefox browser window with the title "Namenode Information - Mozilla Firefox". The address bar shows "localhost:9870/dfshealth.html#tab-overview". The main content area has a green header bar with tabs: "Hadoop" (selected), "Overview", "Datanodes", "Datanode Volume Failures", "Snapshot", "Startup Progress", and "Utilities". Below the header, there's a table with the following data:

Started:	Mon Nov 23 10:05:15 -0600 2020
Version:	3.2.1, rb3cbbb467e22ea829b3808f4b7b01d07e0bf3842
Compiled:	Tue Sep 10 10:56:00 -0500 2019 by rohithsharmaks from branch-3.2.1
Cluster ID:	CID-adff28c4-5797-4b81-8679-30742ef04446
Block Pool ID:	BP-1347129731-127.0.1.1-1606147073637

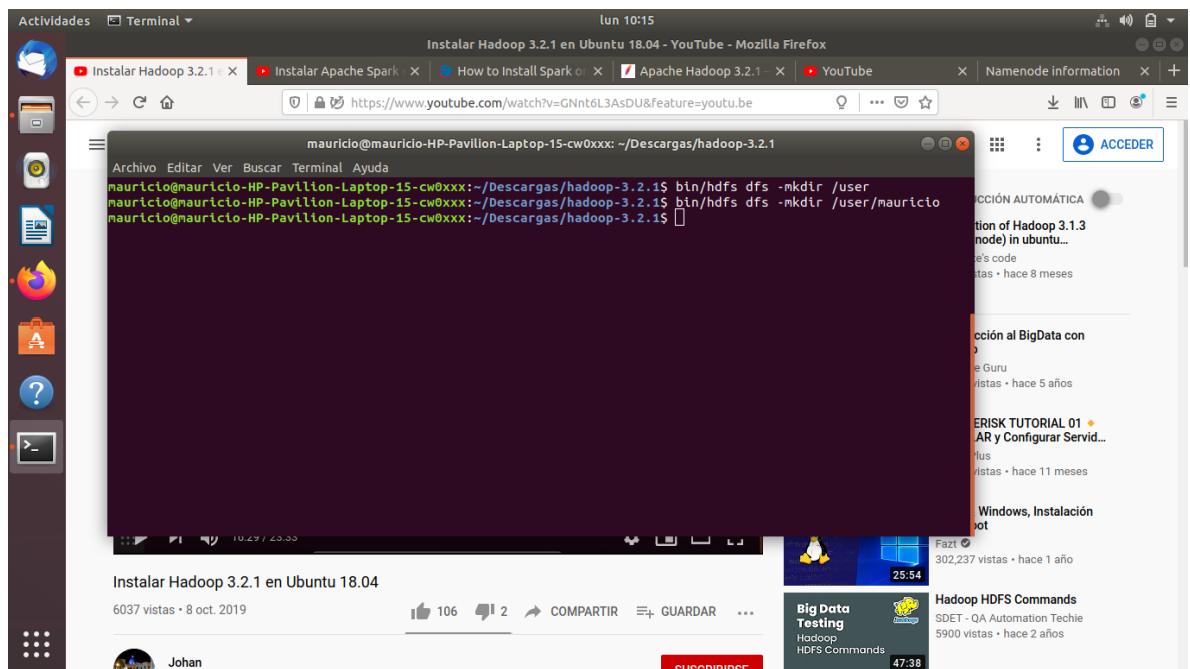
Below the table, there's a section titled "Summary" with the following text:

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 101.12 MB of 292.5 MB Heap Memory. Max Heap Memory is 2.38 GB.
Non Heap Memory used 46.9 MB of 48.46 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

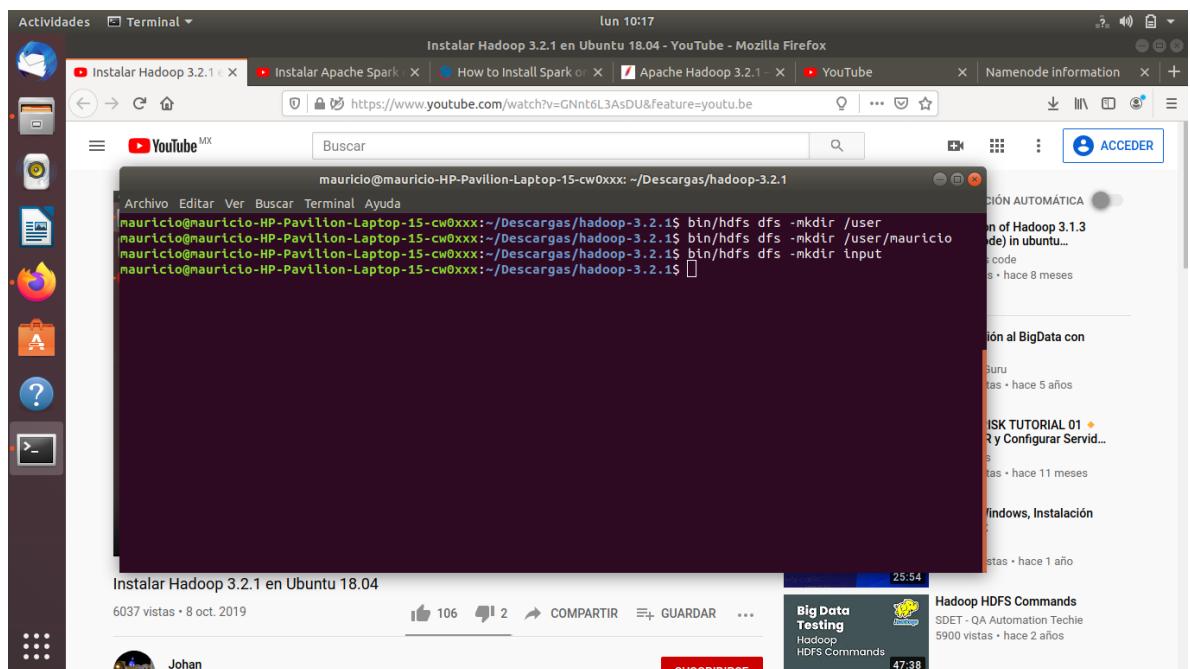
28. Ahora crearemos las carpetas del sistema usando el comando **bin/hdfs dfs -mkdir /user**

The screenshot shows a terminal window with the title "Terminal". The command "mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1\$ bin/hdfs dfs -mkdir /user" is being typed and executed. The terminal output shows the creation of the directory "/user".

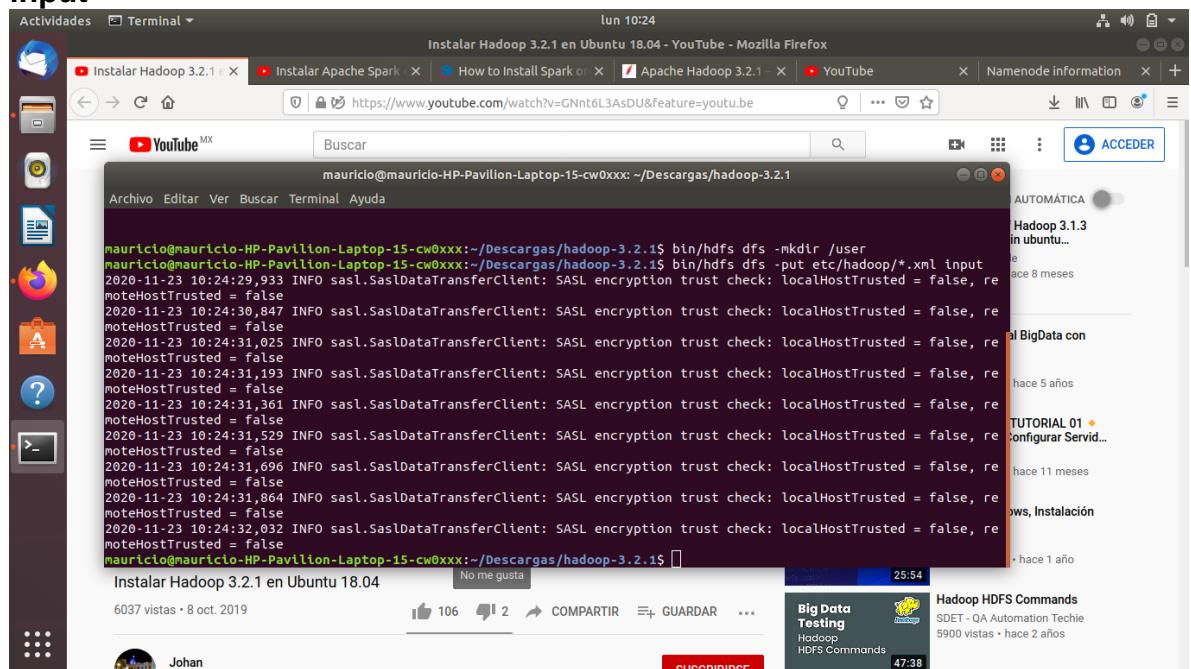
29. La siguiente carpeta que crearemos será **bin/hdfs dfs –mkdir /mauricio**



30. El siguiente paso es crear la ruta en el directorio con el comando **bin/ dfs – mkdir input**

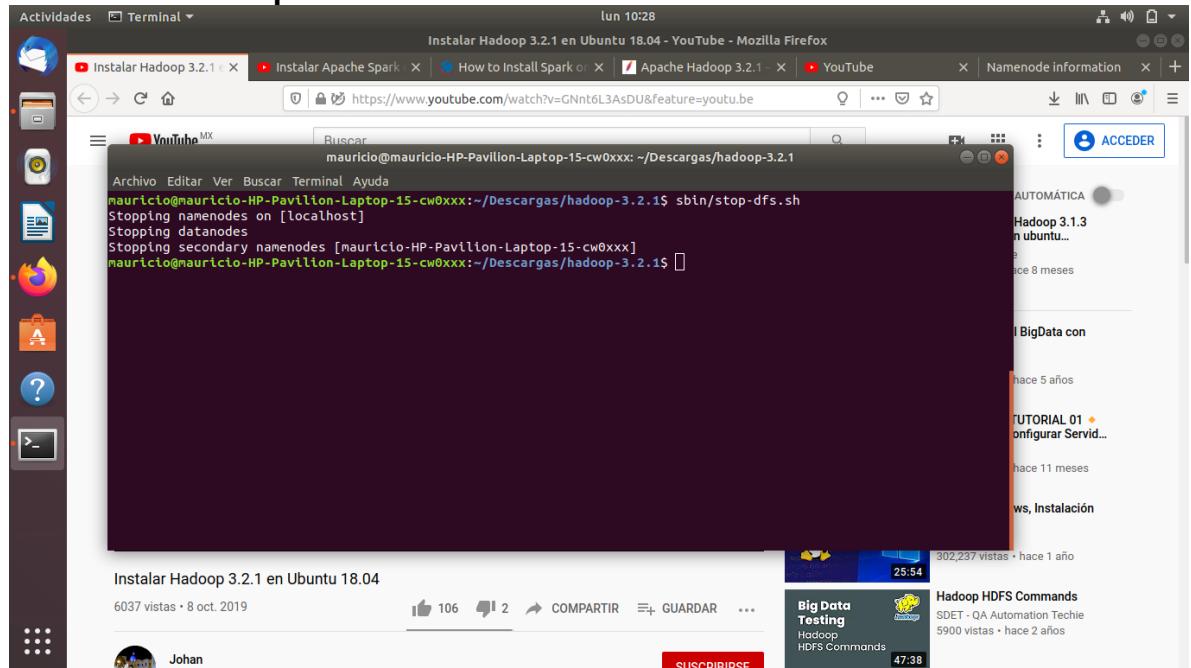


31. El siguiente paso es usar el comando `bin/hdfs dfs -put etc/hadoop/*.xml input`



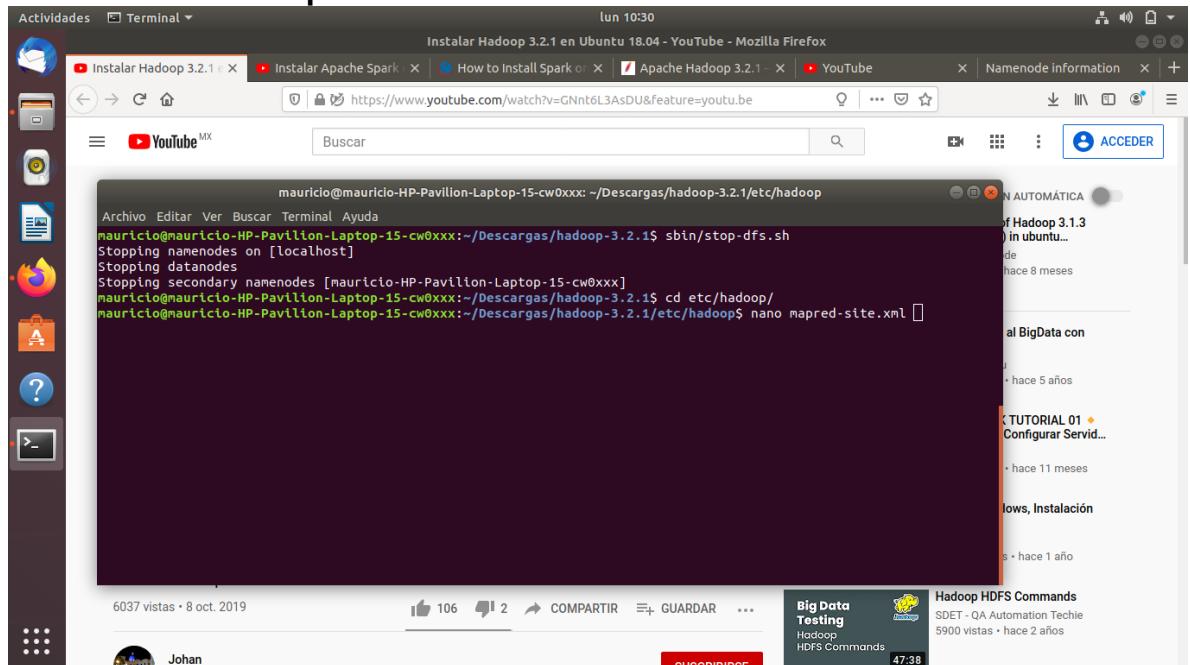
A screenshot of a Linux desktop environment showing a terminal window. The terminal window title is "Instalar Hadoop 3.2.1 en Ubuntu 18.04 - YouTube - Mozilla Firefox". The terminal content shows the command "mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1\$ bin/hdfs dfs -mkdir /user mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1\$ bin/hdfs dfs -put etc/hadoop/*.xml input" being run. The output of the command is displayed below the command line. The terminal window is overlaid on a YouTube video player for a tutorial on installing Hadoop 3.2.1 in Ubuntu 18.04.

32. El siguiente paso es detener los demonios y para esto se requiere el siguiente comando `sbin/stop-dfs.sh`

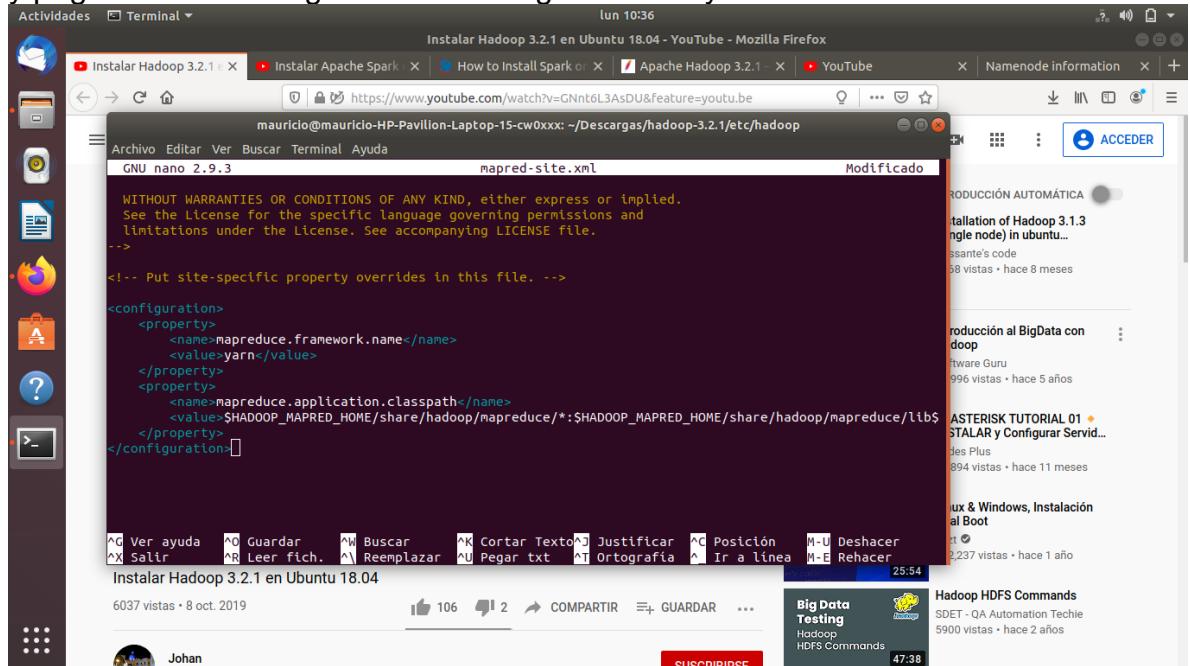


A screenshot of a Linux desktop environment showing a terminal window. The terminal window title is "Instalar Hadoop 3.2.1 en Ubuntu 18.04 - YouTube - Mozilla Firefox". The terminal content shows the command "mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1\$ sbin/stop-dfs.sh" being run. The output of the command is displayed below the command line. The terminal window is overlaid on a YouTube video player for a tutorial on installing Hadoop 3.2.1 in Ubuntu 18.04.

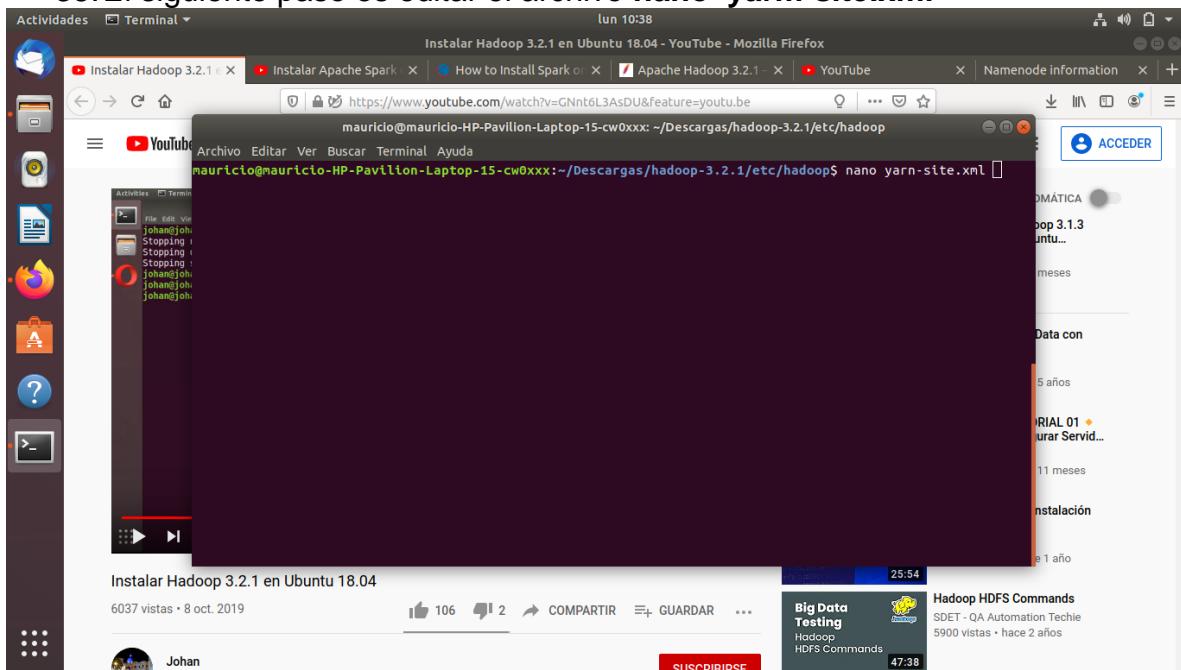
33. Ahora terminaremos de configurar java usando los siguientes comandos, pero primero tendremos que entrar en la carpeta `cd etc/hadoop/` y usaremos el comando `nano mapred-site.xml`



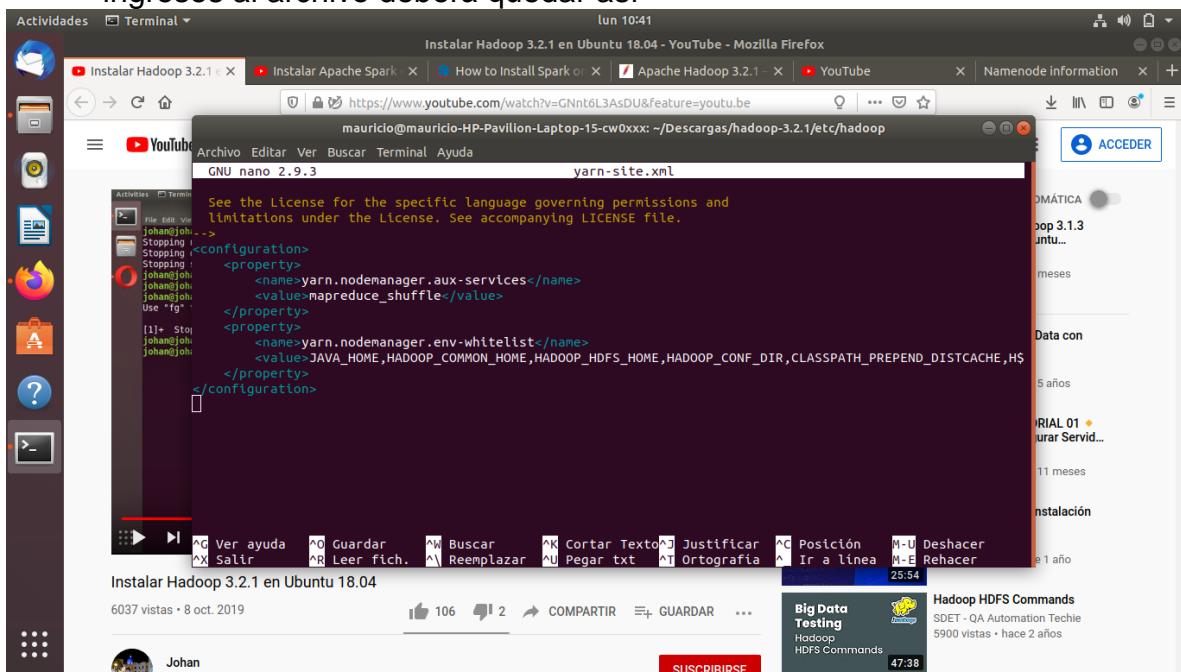
34. En la página de hadoop encontraremos la configuración de mapred copiamos y pegaremos de la siguiente manera guardamos y salimos



35. El siguiente paso es editar el archivo **nano yarn-site.xml**



36. Este código se encuentra en la página <https://hadoop.apache.org> cuando ingreses al archivo deberá quedar así



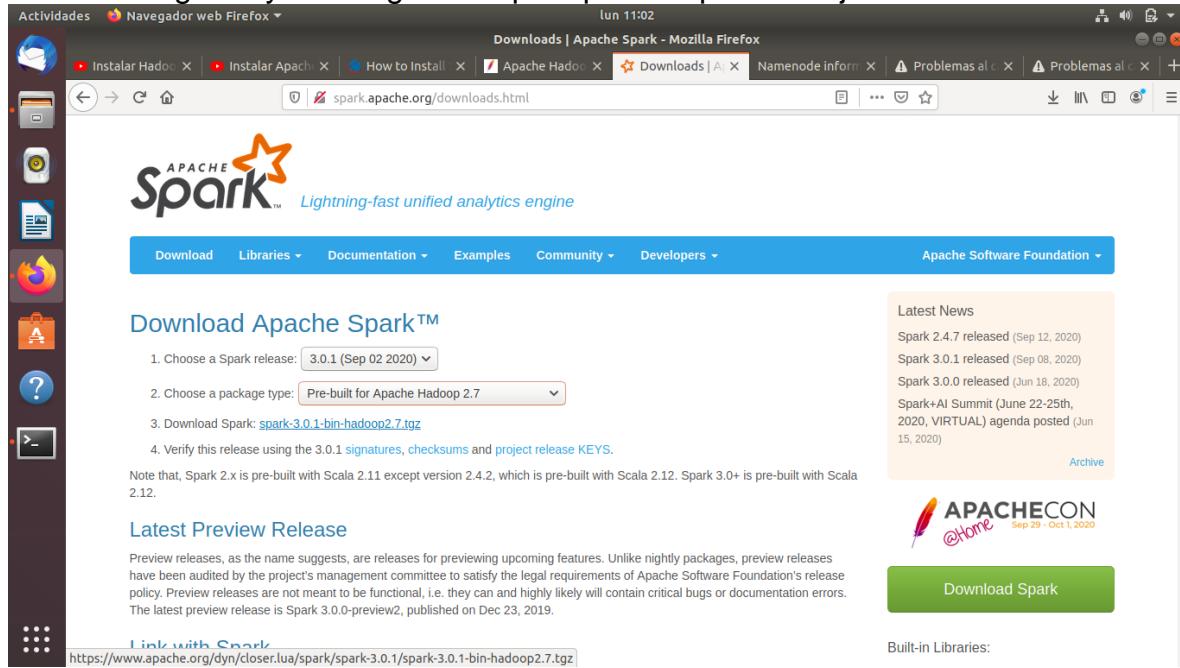
37. Ahora regresamos a la carpeta principal de badoo e iniciamos el yarn con el comando **sbin/Start-yarn.sh**

The screenshot shows a Linux desktop environment with a dark theme. Two terminal windows are open, both titled "Terminal". The top terminal window shows the command "mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx: ~/Descargas/hadoop-3.2.1\$ sbin/start-yarn.sh" being run, with the output "Starting resourcemanager" and "Starting nodemanagers". The bottom terminal window shows the command "mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx: ~/Descargas/hadoop-3.2.1\$ sbin/start-yarn.sh" being run, with the output "Starting resourcemanager", "Starting nodemanagers", "Stopping namenodes", "Starting namenodes on [localhost]", "Starting datanodes", "Starting secondary namenodes [mauricio-HP-Pavilion-Laptop-15-cw0xxx]", and "Use *fg* to re". In the background, a YouTube video player is visible, showing a video titled "Instalar Hadoop 3.2.1 en Ubuntu 18.04 - YouTube" by "Johan". The video has 6037 views and was uploaded on October 8, 2019.

```
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx: ~/Descargas/hadoop-3.2.1$ nano yarn-site.xml
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1/etc/hadoop$ cd ../..
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ sbin/start-yarn.sh
Starting resourcemanager
Starting nodemanagers
Stopping namenodes
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [mauricio-HP-Pavilion-Laptop-15-cw0xxx]
Use *fg* to re
Stopping nodemanagers
[1]- Stopped Stopping resourcemanager
johangohan:[~/Descargas/hadoop-3.2.1]$ mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas/hadoop-3.2.1$ sbin/stop-yarn.sh
Starting resou
Starting nodem
Starting namen
Starting datan
Starting secon
johangohan:[~/Descargas/hadoop-3.2.1$ start-all.cmd
start-all.sh
start-balanc
start-dfs.sh
johangohan:[~/Descargas/hadoop-3.2.1$ stop-yarn.sh
Stopping nodem
Stopping resou
johangohan:[~/Descargas/hadoop-3.2.1$
```

INSTALACION DE SPARK

38.Como primer punto en nuestro sistema operativo nos fuimos a nuestro navegador y descargamos Spark para su próxima ejecución.



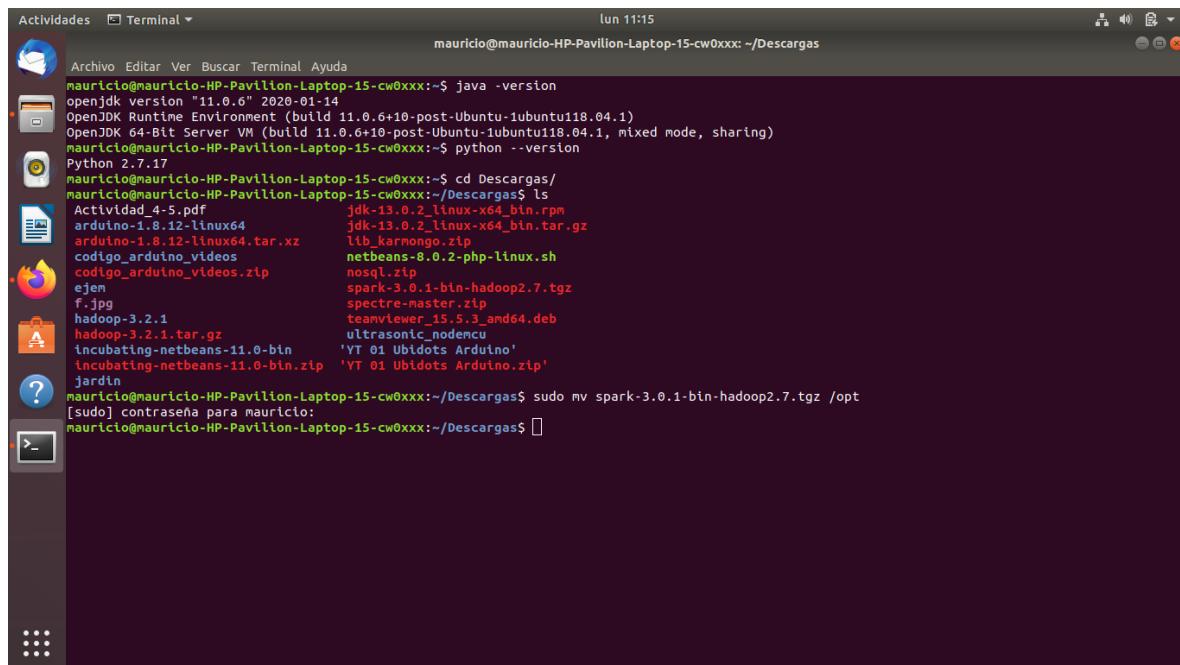
The screenshot shows a Firefox browser window with the URL spark.apache.org/downloads.html. The page displays the Apache Spark logo and the tagline "Lightning-fast unified analytics engine". It features a navigation bar with links for Download, Libraries, Documentation, Examples, Community, Developers, and Apache Software Foundation. Below the navigation bar, there's a section titled "Download Apache Spark™" with a numbered list of steps:

1. Choose a Spark release: 3.0.1 (Sep 02 2020)
2. Choose a package type: Pre-built for Apache Hadoop 2.7
3. Download Spark: [spark-3.0.1-bin-hadoop2.7.tgz](https://www.apache.org/dyn/closer.lua/spark/spark-3.0.1/spark-3.0.1-bin-hadoop2.7.tgz)
4. Verify this release using the 3.0.1 signatures, checksums and project release KEYS.

Note that, Spark 2.x is pre-built with Scala 2.11 except version 2.4.2, which is pre-built with Scala 2.12. Spark 3.0+ is pre-built with Scala 2.12.

On the right side of the page, there's a "Latest News" box listing recent releases and a "Download Spark" button. A sidebar on the left contains links for "Link with Spark" and "Built-in Libraries".

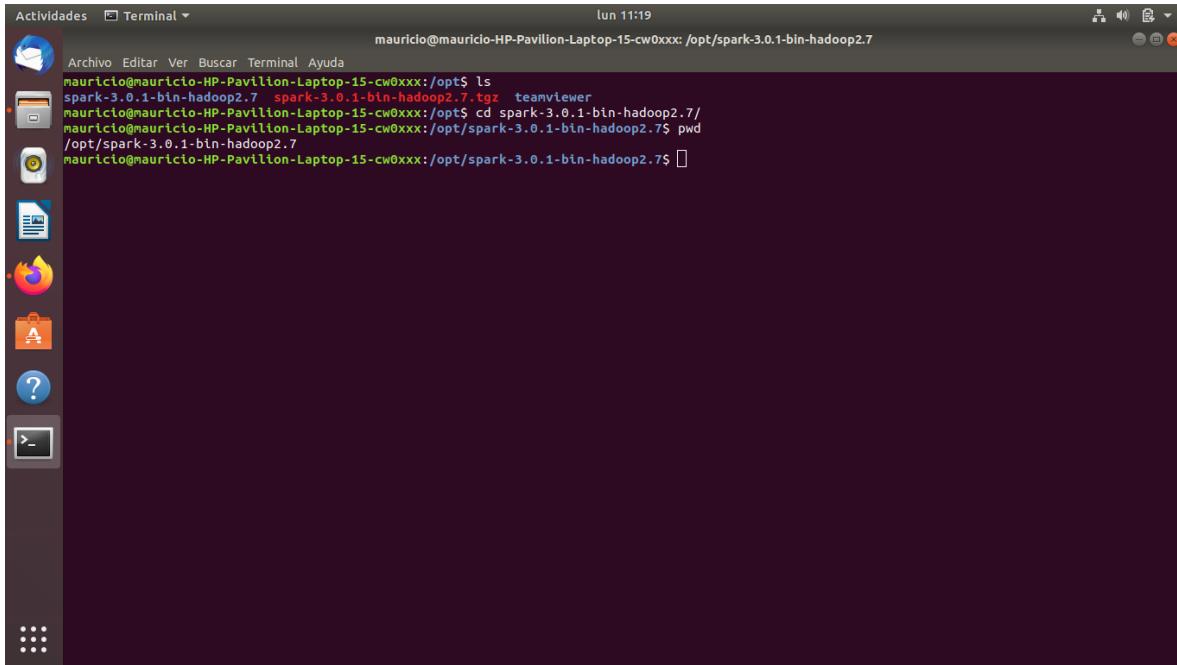
39.Seguidamente una vez descargado accedimos a nuestra consola y buscamos el archivo en descargas y pusimos el comando **sudo mvspark-3.0.3-bin-hadoop2.7.tgz /opt**



The screenshot shows a terminal window with the command `sudo mvspark-3.0.3-bin-hadoop2.7.tgz /opt` being run. The terminal output shows the user navigating to the Downloads directory, listing files, and then executing the command. The terminal window is part of a desktop environment with a dark theme, and the background shows a file manager interface.

```
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas$ java -version
openjdk version "11.0.6" 2020-01-14
OpenJDK Runtime Environment (build 11.0.6+10-post-Ubuntu-1ubuntu118.04.1)
OpenJDK 64-Bit Server VM (build 11.0.6+10-post-Ubuntu-1ubuntu118.04.1, mixed mode, sharing)
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas$ python --version
Python 2.7.17
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas$ cd Descargas/
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas$ ls
Actividadad_4-5.pdf          jdk-13.0_2_linux-x64_bin.rpm
arduino-1.8.12-linux64        jdk-13.0_2_linux-x64_bin.tar.gz
arduino-1.8.12-linux64.tar.xz lib_karmongo.zip
codigo_arduino_videos         netbeans-8.0.2-php-linux.sh
codigo_arduino_videos.zip     nosql.zip
ejen                           spark-3.0.1-bin-hadoop2.7.tgz
f.jpg                          spectre-master.zip
hadoop-3.2.1                  teamviewer_15.5.3_amd64.deb
hadoop-3.2.1.tar.gz           ultrasonic_nodemcu
incubating-netbeans-11.0-bin   'YT 01 Ubidots Arduino'
incubating-netbeans-11.0-bin.zip 'YT 01 Ubidots Arduino.zip'
jardin                         jardn
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas$ sudo mv spark-3.0.1-bin-hadoop2.7.tgz /opt
[sudo] contraseña para mauricio:
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~/Descargas$
```

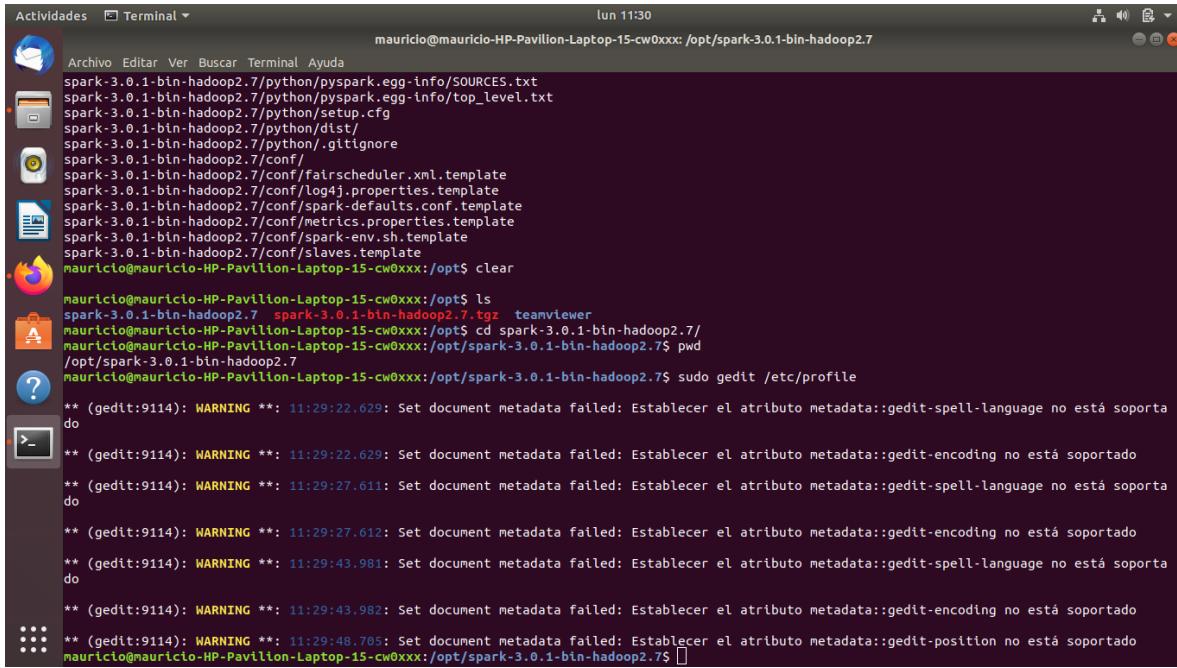
40. Una vez accedido a la carpeta opt pondremos el comando **cd spark-3.0.1-bin-hadoop2.7/** seguidamente nos saltara a otra línea pondremos **pwd** e iniciamos



```
Actividades Terminal
lun 11:19
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7

Archivo Editar Ver Buscar Terminal Ayuda
spark-3.0.1-bin-hadoop2.7 spark-3.0.1-bin-hadoop2.7.tgz teamviewer
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt$ cd spark-3.0.1-bin-hadoop2.7/
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7$ pwd
/opt/spark-3.0.1-bin-hadoop2.7
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7$
```

41. Despues de esta parte insertaremos el comando **sudo gedit /etc/profile** y damos enter y empezara a cargar la instalación nos mandara advertencias de los paquetes que carga pero se instala correctamente.



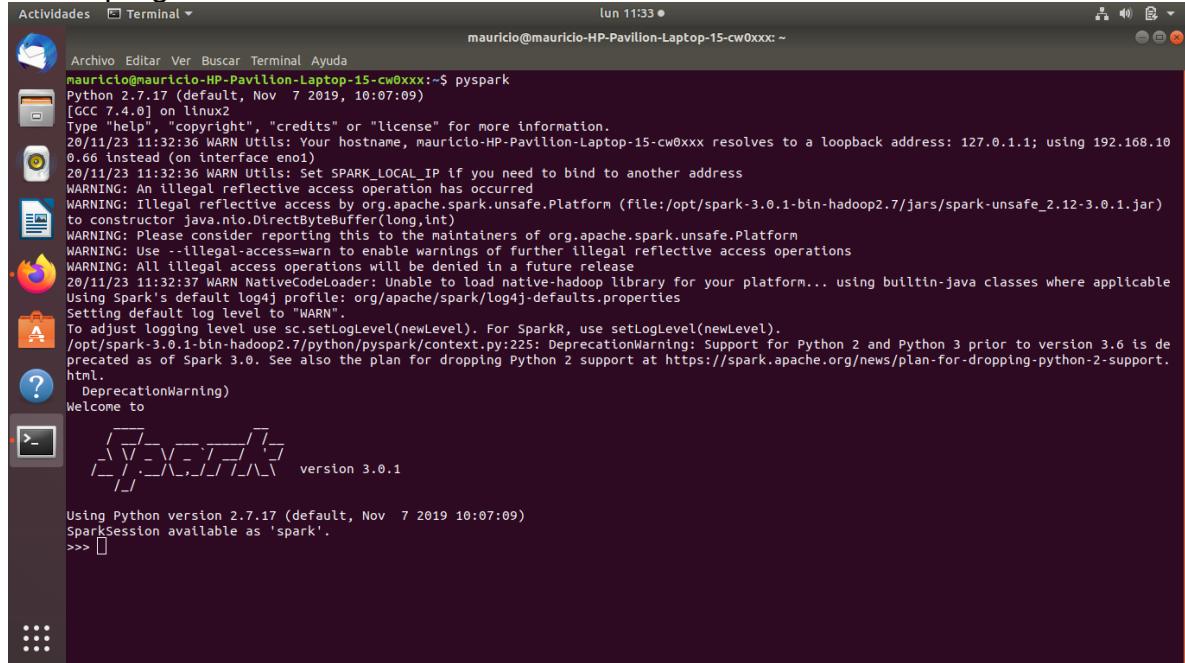
```
Actividades Terminal
lun 11:30
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7

Archivo Editar Ver Buscar Terminal Ayuda
spark-3.0.1-bin-hadoop2.7/python/pyspark.egg-info/SOURCES.txt
spark-3.0.1-bin-hadoop2.7/python/pyspark.egg-info/top_level.txt
spark-3.0.1-bin-hadoop2.7/python/setup.cfg
spark-3.0.1-bin-hadoop2.7/python/dist/
spark-3.0.1-bin-hadoop2.7/python/.gitignore
spark-3.0.1-bin-hadoop2.7/conf/
spark-3.0.1-bin-hadoop2.7/conf/fairscheduler.xml.template
spark-3.0.1-bin-hadoop2.7/conf/log4j.properties.template
spark-3.0.1-bin-hadoop2.7/conf/spark-defaults.conf.template
spark-3.0.1-bin-hadoop2.7/conf/metrics.properties.template
spark-3.0.1-bin-hadoop2.7/conf/spark-env.sh.template
spark-3.0.1-bin-hadoop2.7/conf/slaves.template
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt$ clear

mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt$ ls
spark-3.0.1-bin-hadoop2.7 spark-3.0.1-bin-hadoop2.7.tgz teamviewer
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt$ cd spark-3.0.1-bin-hadoop2.7/
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7$ pwd
/opt/spark-3.0.1-bin-hadoop2.7
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7$ sudo gedit /etc/profile

** (gedit:9114): WARNING **: 11:29:22.629: Set document metadata failed: Establecer el atributo metadata::gedit-spell-language no está soportado
** (gedit:9114): WARNING **: 11:29:22.629: Set document metadata failed: Establecer el atributo metadata::gedit-encoding no está soportado
** (gedit:9114): WARNING **: 11:29:27.611: Set document metadata failed: Establecer el atributo metadata::gedit-spell-language no está soportado
** (gedit:9114): WARNING **: 11:29:27.612: Set document metadata failed: Establecer el atributo metadata::gedit-encoding no está soportado
** (gedit:9114): WARNING **: 11:29:43.981: Set document metadata failed: Establecer el atributo metadata::gedit-spell-language no está soportado
** (gedit:9114): WARNING **: 11:29:43.982: Set document metadata failed: Establecer el atributo metadata::gedit-encoding no está soportado
** (gedit:9114): WARNING **: 11:29:48.705: Set document metadata failed: Establecer el atributo metadata::gedit-position no está soportado
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7$
```

42. Una vez instalado en nuestra consola de comando digitaremos **pyspark** y nos arrojara la información de spark la versión entre otras cualidades del programa.



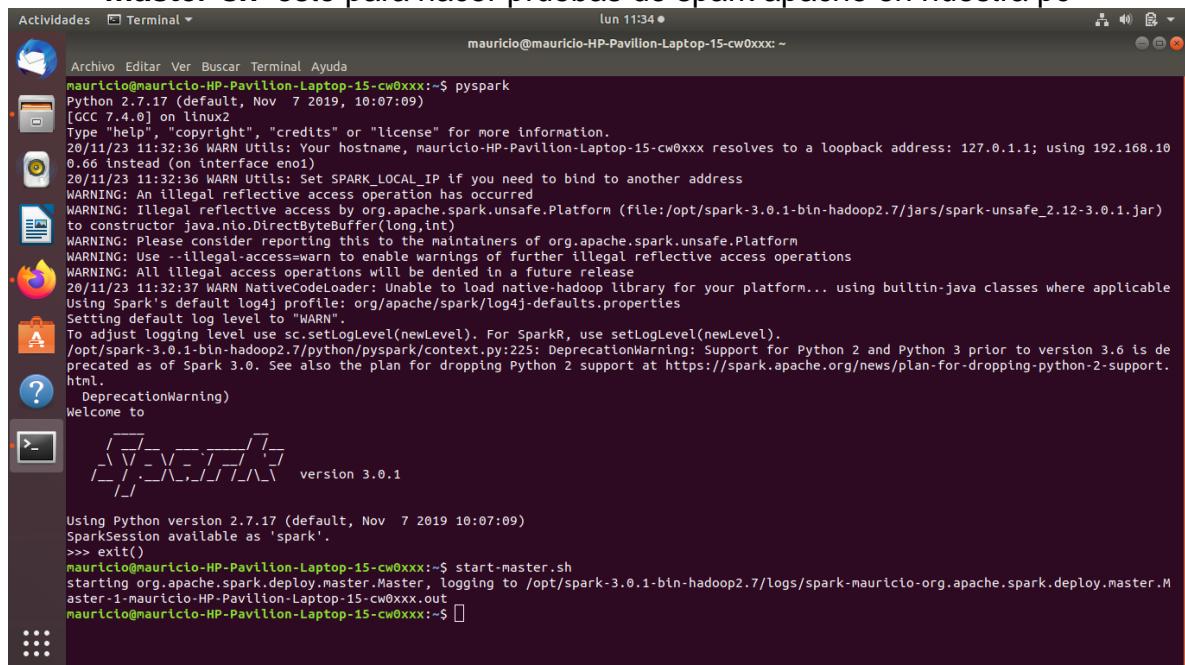
```

Actividades Terminal lun 11:33 ●
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ pyspark
Python 2.7.17 (default, Nov  7 2019, 10:07:09)
[GCC 7.4.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
20/11/23 11:32:36 WARN Utils: Your hostname, mauricio-HP-Pavilion-Laptop-15-cw0xxx resolves to a loopback address: 127.0.1.1; using 192.168.10.0.66 instead (on interface eno1)
20/11/23 11:32:36 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark-3.0.1-bin-hadoop2.7/jars/spark-unsafe_2.12-3.0.1.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/11/23 11:32:37 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
/opt/spark-3.0.1-bin-hadoop2.7/python/pyspark/context.py:225: DeprecationWarning: Support for Python 2 and Python 3 prior to version 3.6 is deprecated as of Spark 3.0. See also the plan for dropping Python 2 support at https://spark.apache.org/news/plan-for-dropping-python-2-support.html.
  DeprecationWarning)
Welcome to
   __|  / _ \
  / \_| | \| |
 /  \_\  / \_\_/
 /_ \_\_\_ \_\_\_
 version 3.0.1

Using Python version 2.7.17 (default, Nov  7 2019 10:07:09)
SparkSession available as 'spark'.
>>> []

```

43. Una vez hecho esto nos iremos a otra línea en la cual escribiremos **start-master.sh** esto para hacer pruebas de spark apache en nuestra pc



```

Actividades Terminal lun 11:34 ●
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ pyspark
Python 2.7.17 (default, Nov  7 2019, 10:07:09)
[GCC 7.4.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
20/11/23 11:32:36 WARN Utils: Your hostname, mauricio-HP-Pavilion-Laptop-15-cw0xxx resolves to a loopback address: 127.0.1.1; using 192.168.10.0.66 instead (on interface eno1)
20/11/23 11:32:36 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark-3.0.1-bin-hadoop2.7/jars/spark-unsafe_2.12-3.0.1.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/11/23 11:32:37 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
/opt/spark-3.0.1-bin-hadoop2.7/python/pyspark/context.py:225: DeprecationWarning: Support for Python 2 and Python 3 prior to version 3.6 is deprecated as of Spark 3.0. See also the plan for dropping Python 2 support at https://spark.apache.org/news/plan-for-dropping-python-2-support.html.
  DeprecationWarning)
Welcome to
   __|  / _ \
  / \_| | \| |
 /  \_\  / \_\_/
 /_ \_\_\_ \_\_\_
 version 3.0.1

Using Python version 2.7.17 (default, Nov  7 2019 10:07:09)
SparkSession available as 'spark'.
>>> exit()
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark-3.0.1-bin-hadoop2.7/logs/spark-mauricio-org.apache.spark.deploy.master.Master-1-mauricio-HP-Pavilion-Laptop-15-cw0xxx.out
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ []

```

44. Una vez hecho ese comando nos dirigiremos a nuestro buscador y pondremos **localhost** para ver que efectivamente tenga conexión en donde nos arrojara spark y sus apartados que lo conforman

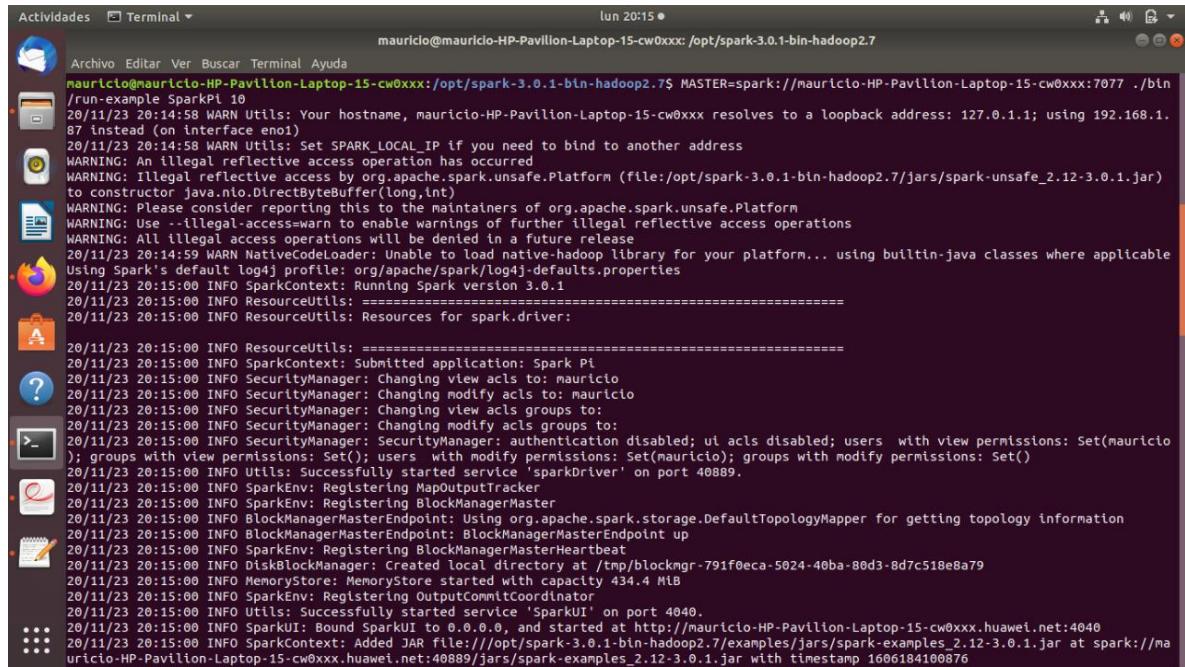
45. Regresamos a la consola y escribiremos **start-slave.sh spark://Mauricio-HP-pavilion-laptop-15-cw0xxx:7077** y seguidamente iniciara en el localhost el programa a ejecutar

```

Actividades Terminal lun 11:37
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ pyspark
Python 2.7.17 (default, Nov  7 2019, 10:07:09)
[GCC 7.4.0] on linux2
Type "help", "copyright", "credits" or "license" for more information.
20/11/23 11:32:36 WARN Utils: Your hostname, mauricio-HP-Pavilion-Laptop-15-cw0xxx resolves to a loopback address: 127.0.1.1; using 192.168.10.0_66 instead (on interface eno1)
20/11/23 11:32:36 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark-3.0.1-bin-hadoop2.7/jars/spark-unsafe_2.12-3.0.1.jar)
to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/11/23 11:32:37 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
/opt/spark-3.0.1-bin-hadoop2.7/python/pyspark/context.py:225: DeprecationWarning: Support for Python 2 and Python 3 prior to version 3.6 is deprecated as of Spark 3.0. See also the plan for dropping Python 2 support at https://spark.apache.org/news/plan-for-dropping-python-2-support.html.
  DeprecationWarning)
Welcome to
    / \   / \
   /   \ /   \
  /     \ /     \
 /       \ /       \
          version 3.0.1

Using Python version 2.7.17 (default, Nov  7 2019 10:07:09)
SparkSession available as 'spark'.
>>> exit()
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ start-master.sh
starting org.apache.spark.deploy.master.Master, logging to /opt/spark-3.0.1-bin-hadoop2.7/logs/spark-mauricio-org.apache.spark.deploy.master.Master-1-mauricio-HP-Pavilion-Laptop-15-cw0xxx.out
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ start-slave.sh spark://mauricio-HP-Pavilion-Laptop-15-cw0xxx:7077
starting org.apache.spark.deploy.worker.Worker, logging to /opt/spark-3.0.1-bin-hadoop2.7/logs/spark-mauricio-org.apache.spark.deploy.worker.Worker-1-mauricio-HP-Pavilion-Laptop-15-cw0xxx.out
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:~$ 
```

46. Una vez tomado ejemplo que se desea correr se mete en consola en este caso será **MASTER-spark://Mauricio-HP-pavilon-laptop-15-cw0xxx:7077**
./bin/run-example SparkPL y seguidamente se comienza a correr y lo visualizaremos q no marque error.

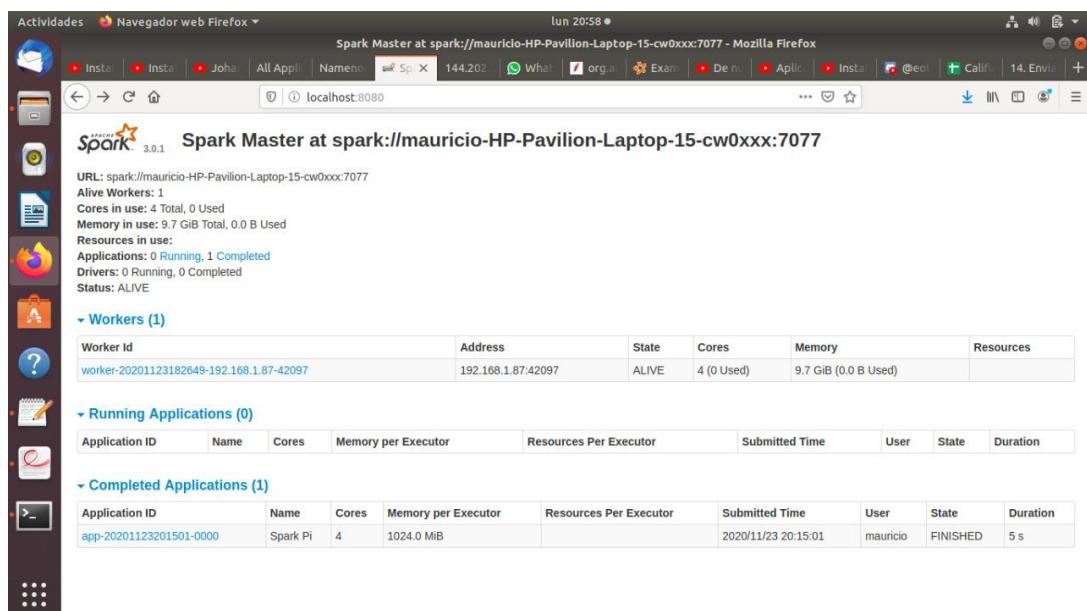


```

Actividades Terminal lun 20:15 •
mauricio@mauricio-HP-Pavilion-Laptop-15-cw0xxx:/opt/spark-3.0.1-bin-hadoop2.7$ ./run-example SparkPL 10
20/11/23 20:14:58 WARN Utils: Your hostname, mauricio-HP-Pavilion-Laptop-15-cw0xxx resolves to a loopback address: 127.0.1.1; using 192.168.1.87 instead (on interface eno1)
20/11/23 20:14:58 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark-3.0.1-bin-hadoop2.7/jars/spark-unsafe_2.12-3.0.1.jar)
to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use -illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denoted in a future release
20/11/23 20:14:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
20/11/23 20:15:00 INFO SparkContext: Running Spark version 3.0.1
20/11/23 20:15:00 INFO ResourceUtils: ======
20/11/23 20:15:00 INFO ResourceUtils: Resources for spark.driver:
20/11/23 20:15:00 INFO ResourceUtils: ======
20/11/23 20:15:00 INFO SparkContext: Submitted application: Spark Pi
20/11/23 20:15:00 INFO SecurityManager: Changing view acls to: mauricio
20/11/23 20:15:00 INFO SecurityManager: Changing modify acls to: mauricio
20/11/23 20:15:00 INFO SecurityManager: Changing view acls groups to:
20/11/23 20:15:00 INFO SecurityManager: Changing modify acls groups to:
20/11/23 20:15:00 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(mauricio); groups with view permissions: Set(); users with modify permissions: Set(mauricio); groups with modify permissions: Set()
20/11/23 20:15:00 INFO Utils: Successfully started service 'sparkDriver' on port 40889.
20/11/23 20:15:00 INFO SparkEnv: Registering MapOutputTracker
20/11/23 20:15:00 INFO SparkEnv: Registering BlockManagerMaster
20/11/23 20:15:00 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
20/11/23 20:15:00 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
20/11/23 20:15:00 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
20/11/23 20:15:00 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-791f0eca-5024-40ba-80d3-8d7c518e8a79
20/11/23 20:15:00 INFO MemoryStore: MemoryStore started with capacity 434.4 MB
20/11/23 20:15:00 INFO SparkEnv: Registering OutputCommitCoordinator
20/11/23 20:15:00 INFO Utils: Successfully started service 'SparkUI' on port 4040.
20/11/23 20:15:00 INFO SparkUi: Bound SparkUI to 0.0.0.0, and started at http://mauricio-HP-Pavilion-Laptop-15-cw0xxx.huawei.net:4040
20/11/23 20:15:00 INFO SparkContext: Added JAR file:///opt/spark-3.0.1-bin-hadoop2.7/examples/jars/spark-examples_2.12-3.0.1.jar at spark://mauricio-HP-Pavilion-Laptop-15-cw0xxx.huawei.net:40889/jars/spark-examples_2.12-3.0.1.jar with timestamp 1606184160876

```

47. Y finalmente volvemos a nuestro navegador y colocamos Localhost y nos aparecerá en workers y en **completed Applications** la aplicación ejemplo finalizado con éxito



Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-202011231201501-0000	Spark Pi	4	1024.0 MiB		2020/11/23 20:15:01	mauricio	FINISHED	5 s