INSTALACION HADOOP

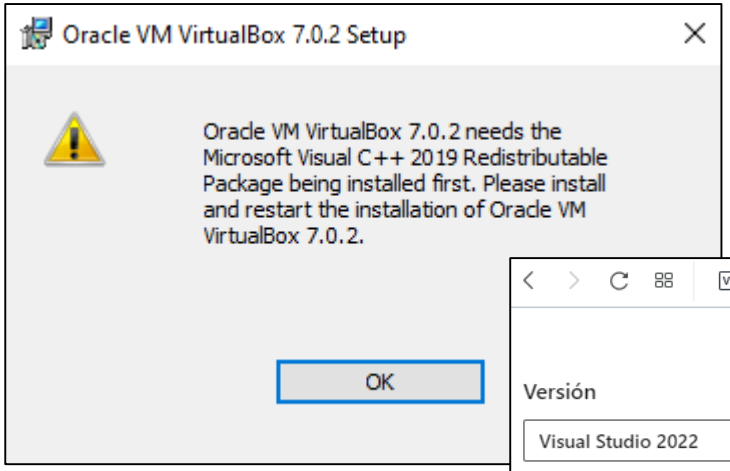
EDUARD LARA

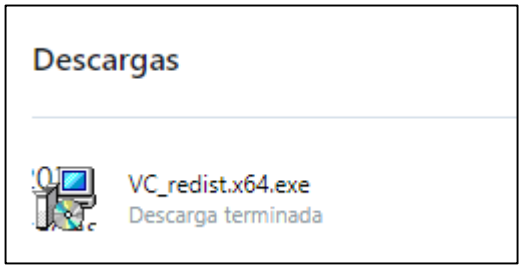
1. REQUISITOS

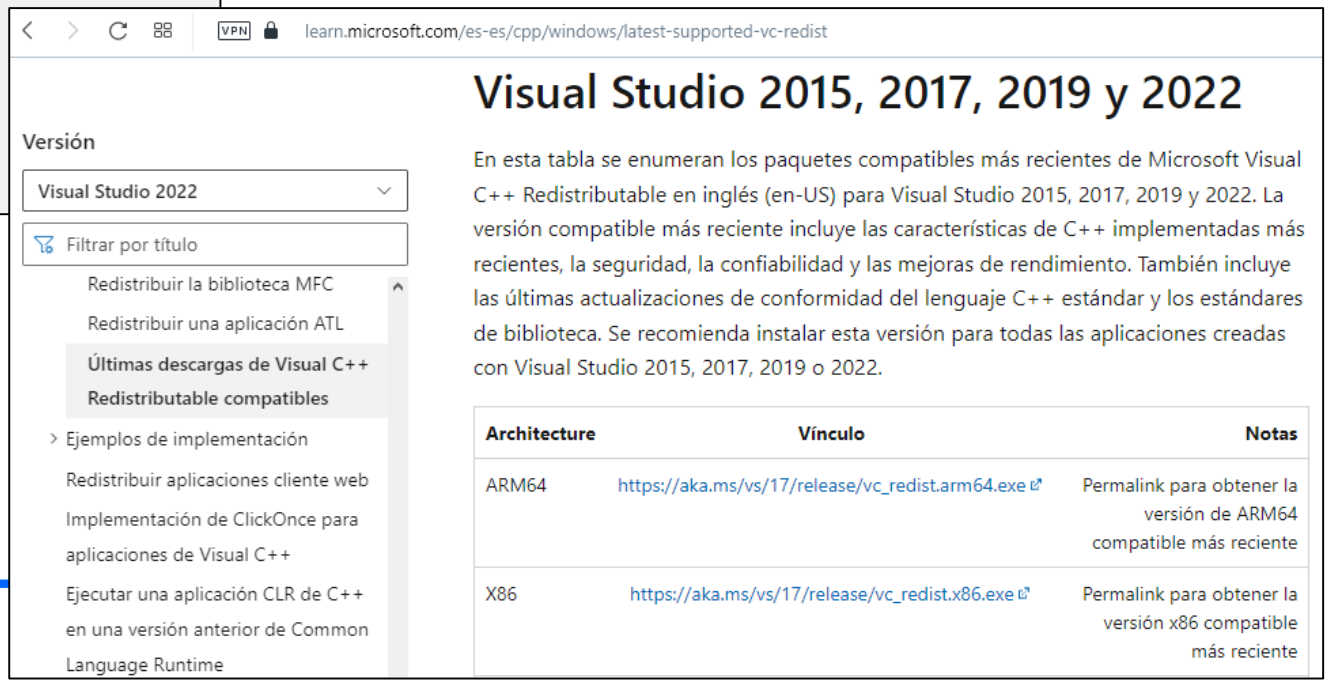
- PC 8 GB RAM
- Oracle Virtual Box
- Linux CentOS
- Hadoop

1. INSTALACION VIRTUAL BOX

Paso 1. Para instalar Virtual Box en Windows, se debe de tener instalado previamente Microsoft Visual C++ 2019 Redistributable. El fichero debe de tener un nombre similar a vc_redist_x64.exe

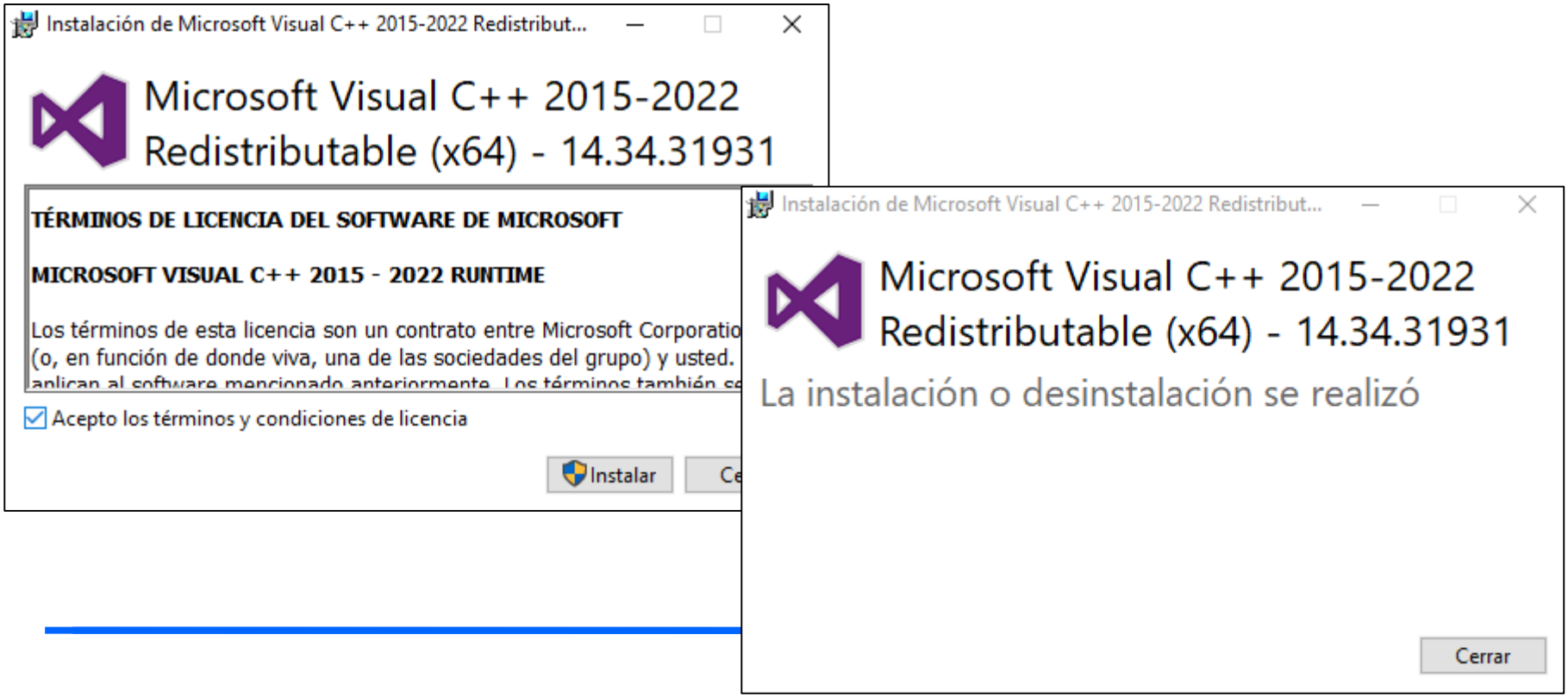






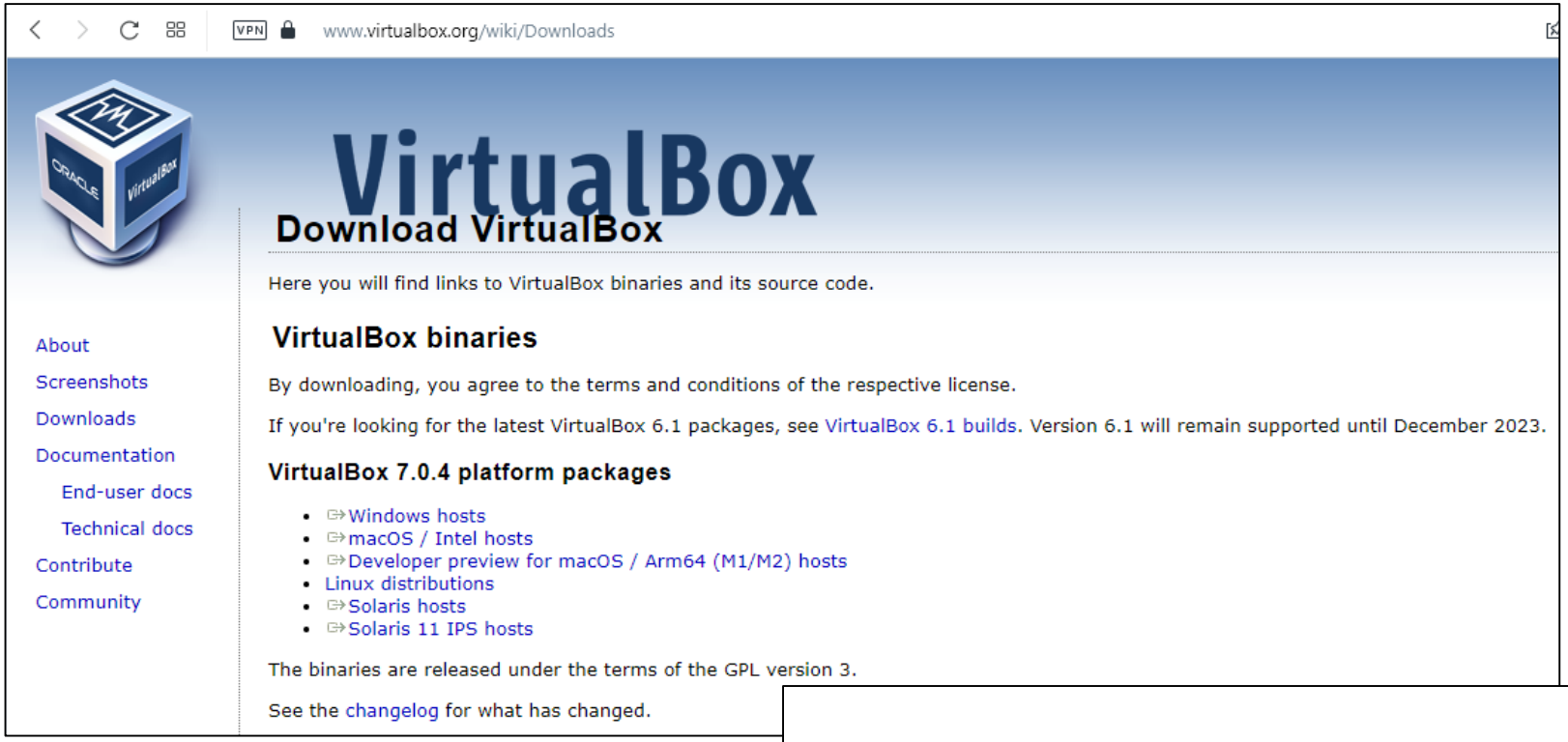
1. INSTALACION VIRTUAL BOX

Paso 2. Instalamos este librería. Microsoft Visual C++ Redistributable es una serie de archivos que se deben instalar en el sistema para que se puedan usar ciertos programas programados con Visual C++, entre ellos el Virtual Box



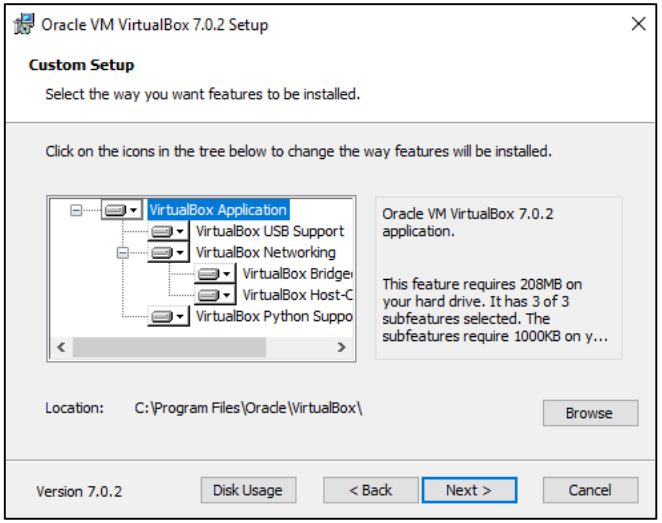
1. INSTALACION VIRTUAL BOX

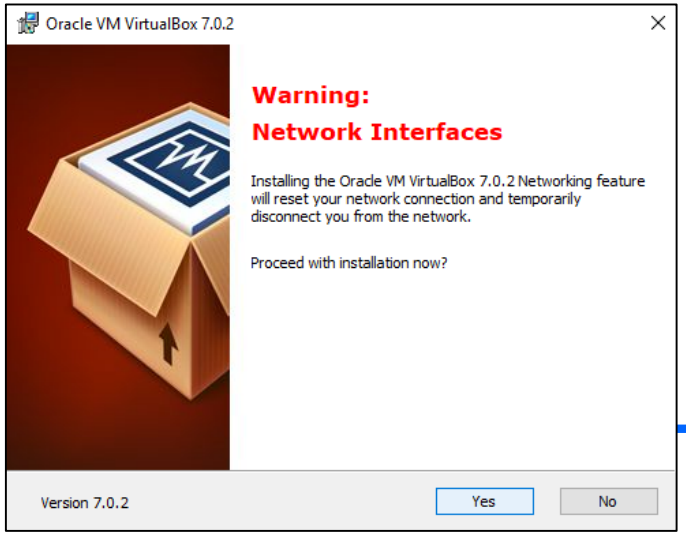
Paso 3. Para descargar virtual box, vamos a la página https://www.virtualbox.org/wiki/Downloads (la versión actual es la 7)

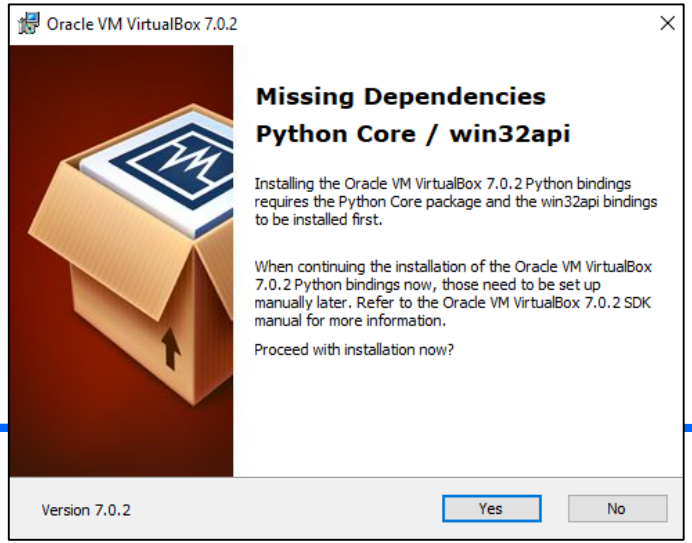


1. INSTALACION VIRTUAL BOX

Paso 4. Iniciamos la instalación de virtual box:

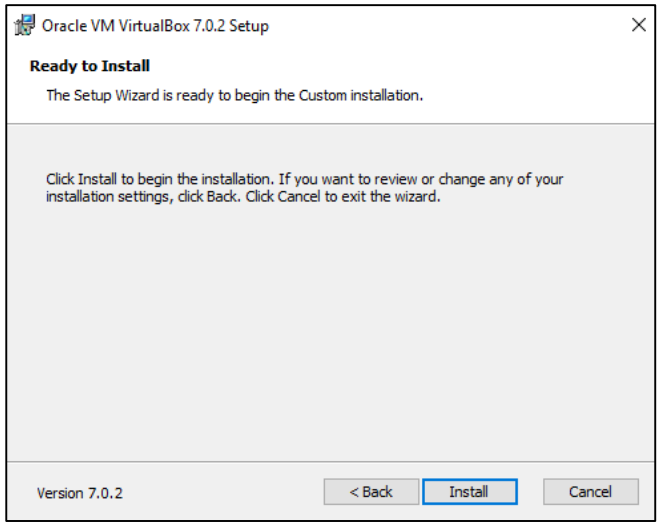


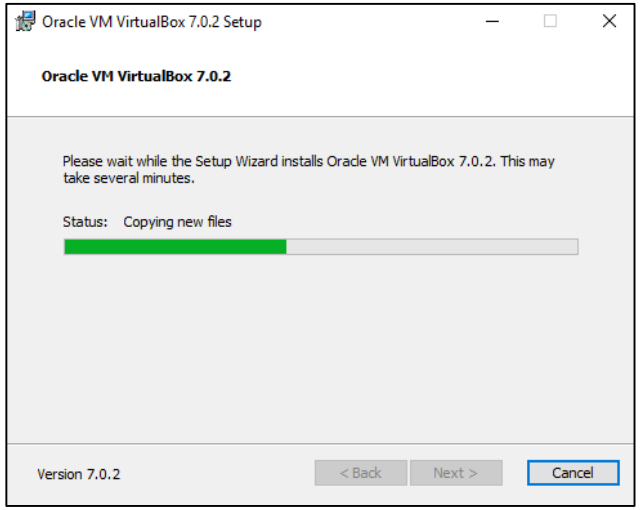


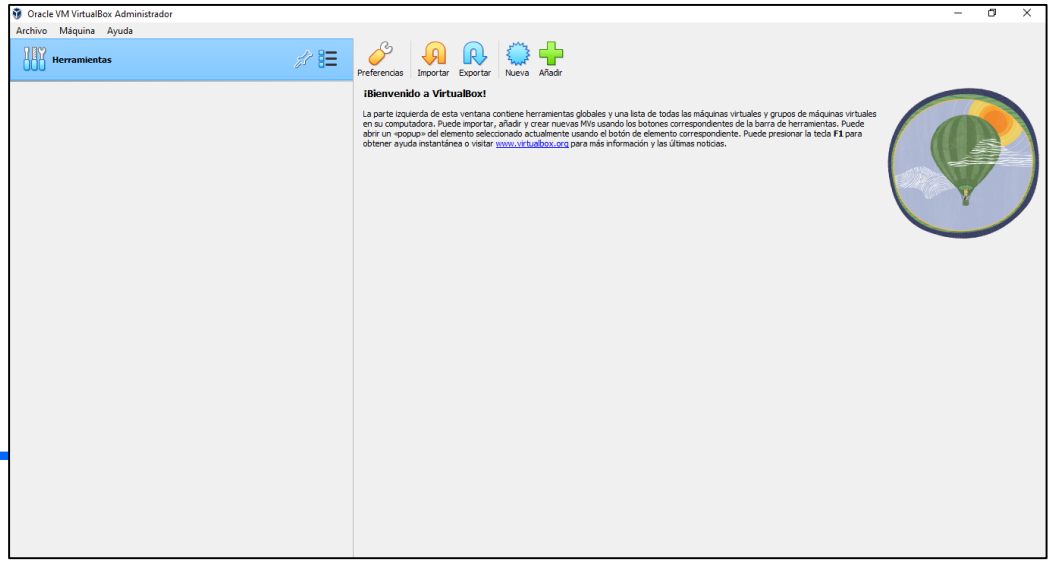


1. INSTALACION VIRTUAL BOX

Paso 5. Finalizamos la instalación e iniciamos el programa

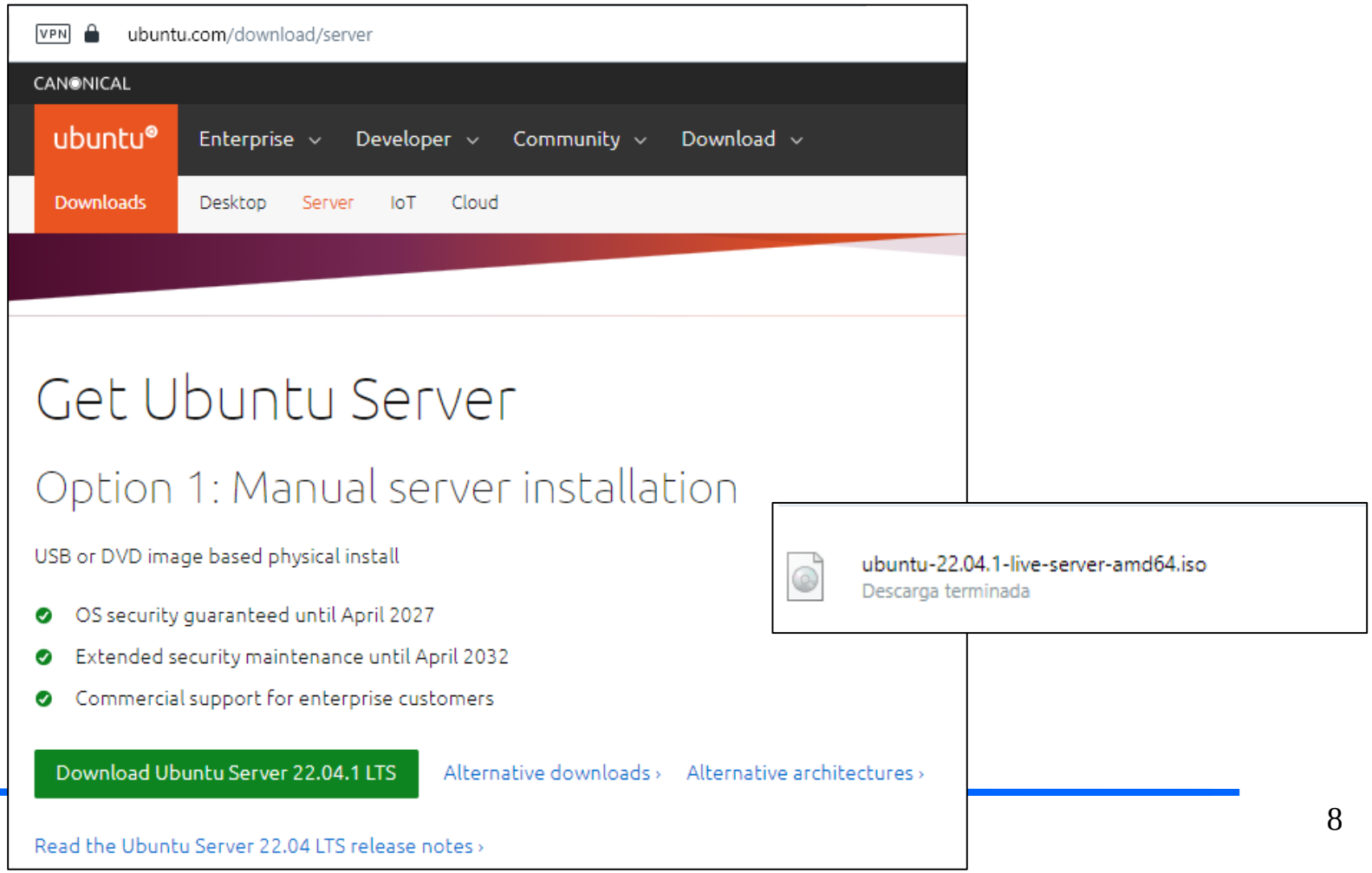






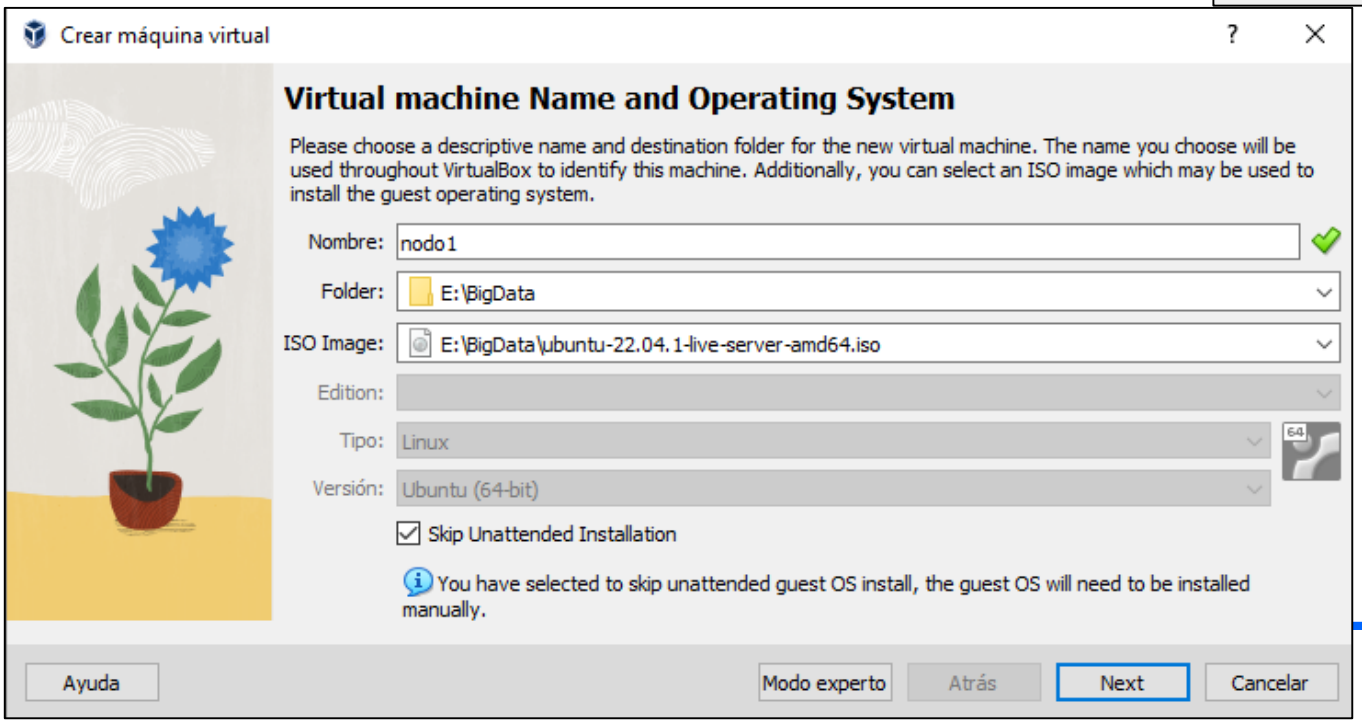
2. INSTALACION UBUNTU SERVER

Paso 1. Utilizaremos Ubuntu Server para crear la infraestructura hadoop. Vamos a la URL hhttps://ubuntu.com/download/server



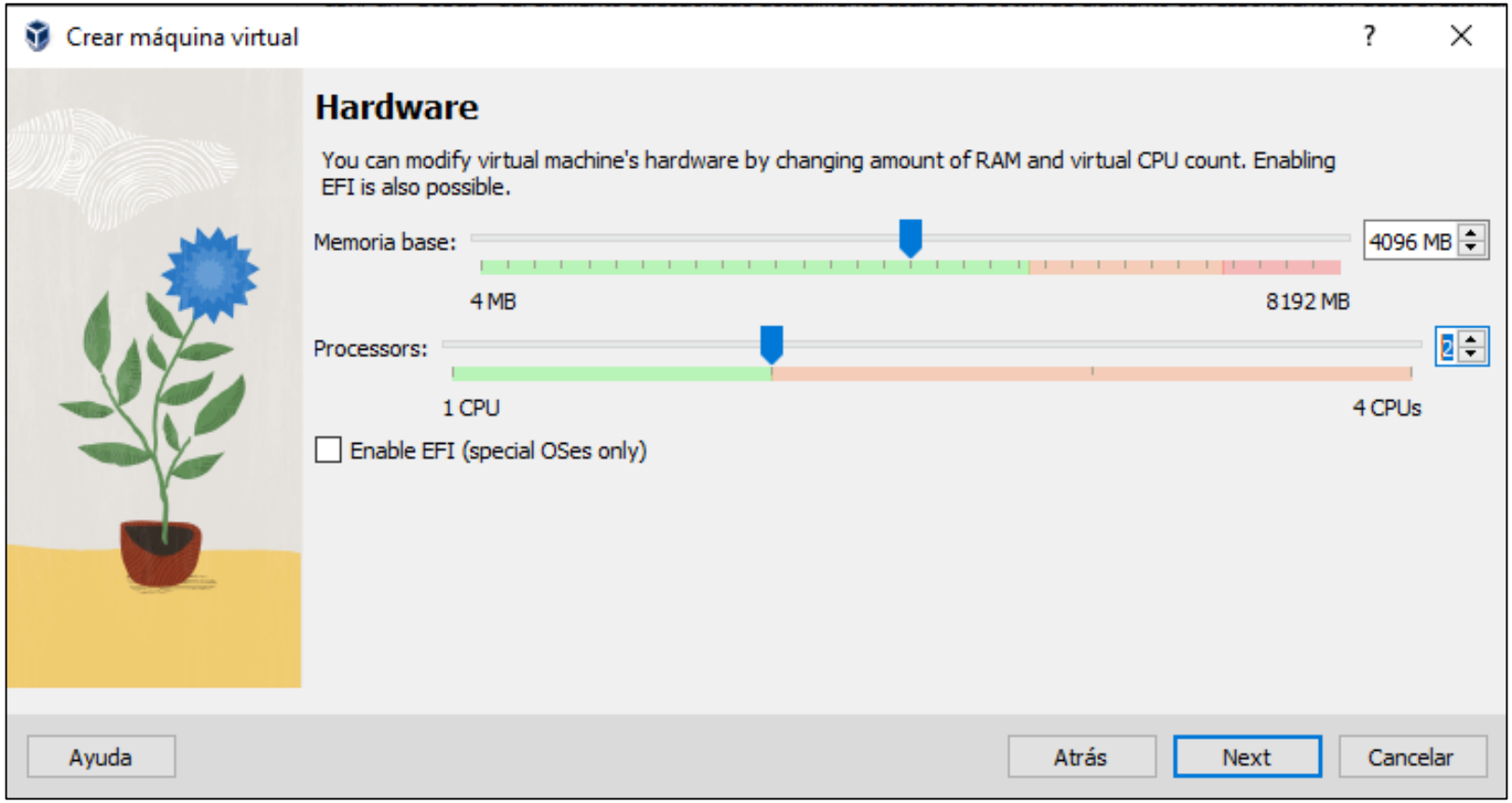
2. INSTALACION UBUNTU SERVER

Paso 2. Creamos una nueva maquina Virtual en Virtual Box haciendo clik en el botón Nueva. Le ponemos de nombre **nodo1**. Crearemos varios nodos que simularan varias maquinas físicas.



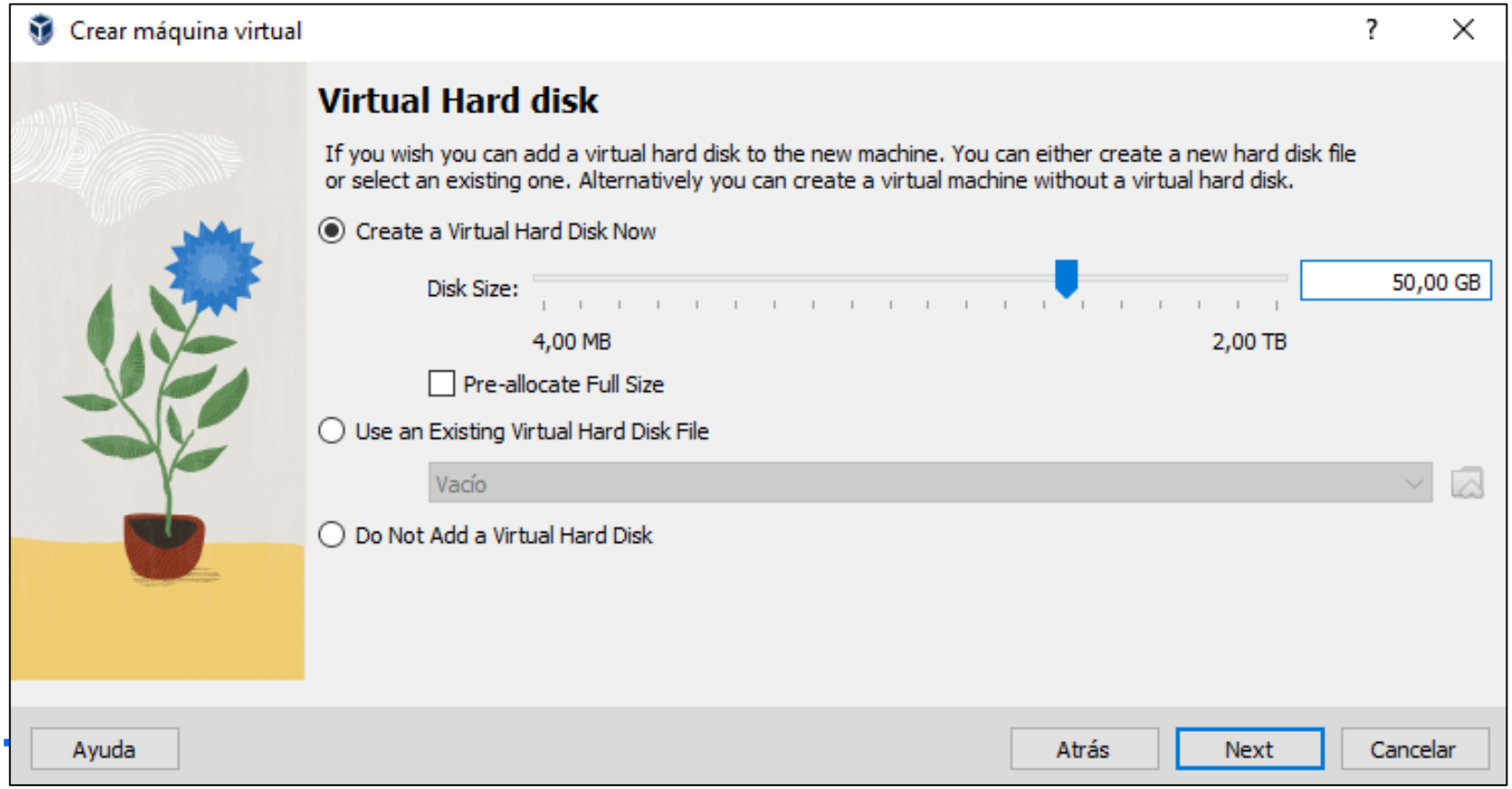
2. INSTALACION UBUNTU SERVER

Paso 3. Para la primera maquina que va a ser la maquina maestra vamos a asignarle un espacio de 4GB de RAM, y 2 CPUs. Para empezar la instalación y aprender hadoop, esto es suficiente



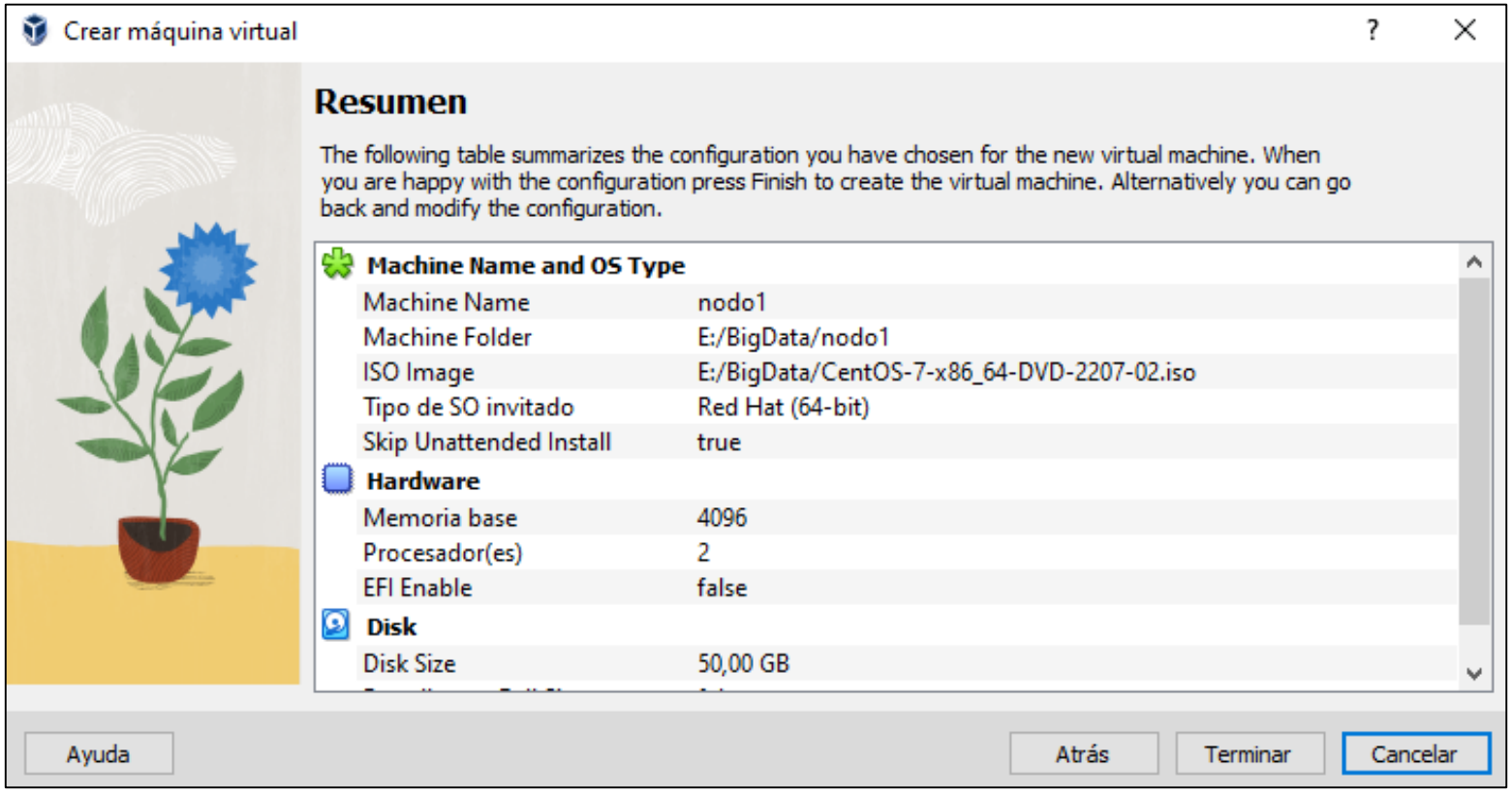
2. INSTALACION UBUNTU SERVER

Paso 4. Seleccionamos crear disco virtual, dejando el tamaño del disco duro a 50Gb (no significa que inicialmente vaya a usar todo ese espacio).



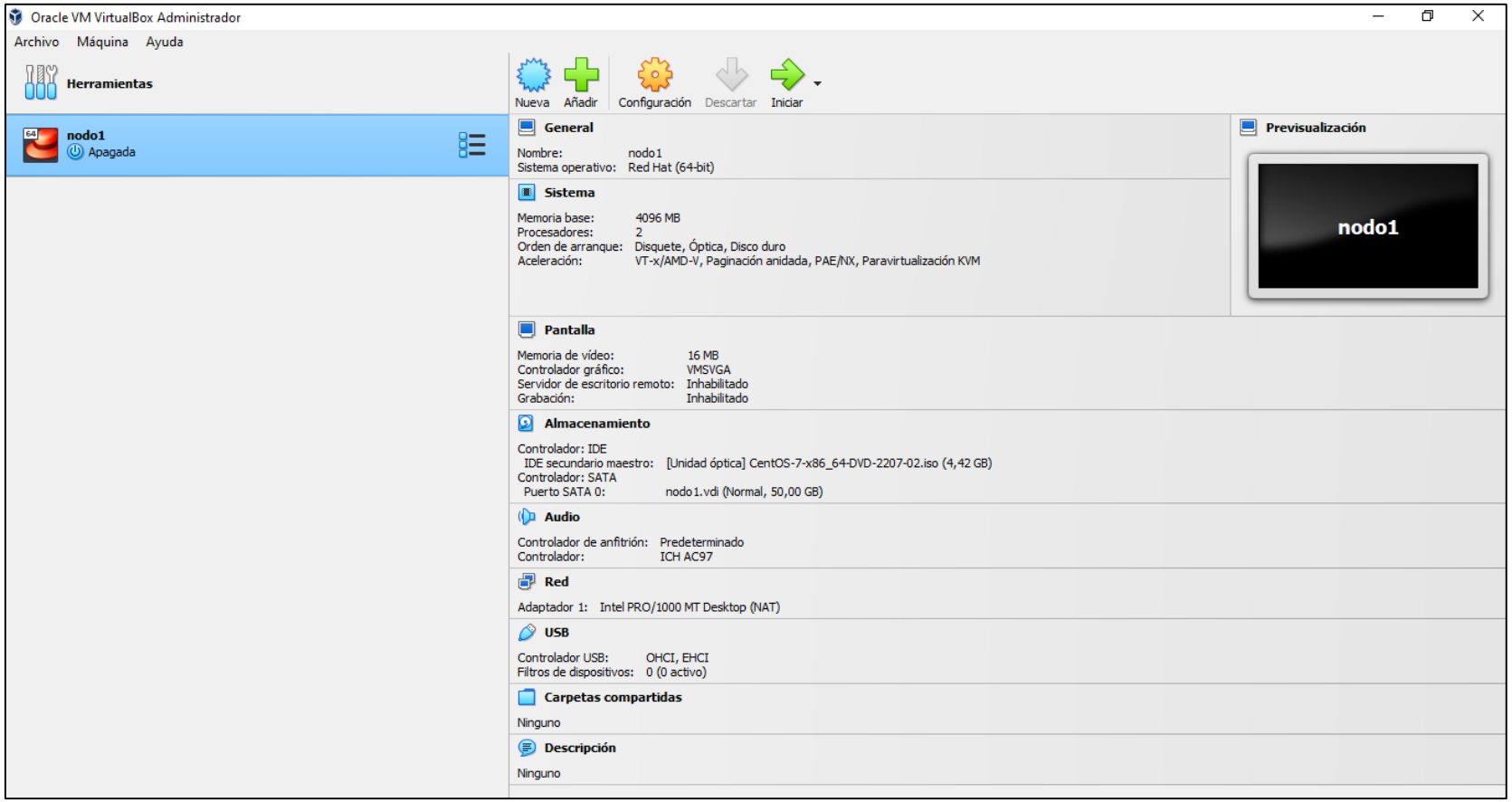
2. INSTALACION UBUNTU SERVER

Paso 5. Finalizamos el inicio de la instalación y obtenemos el resumen



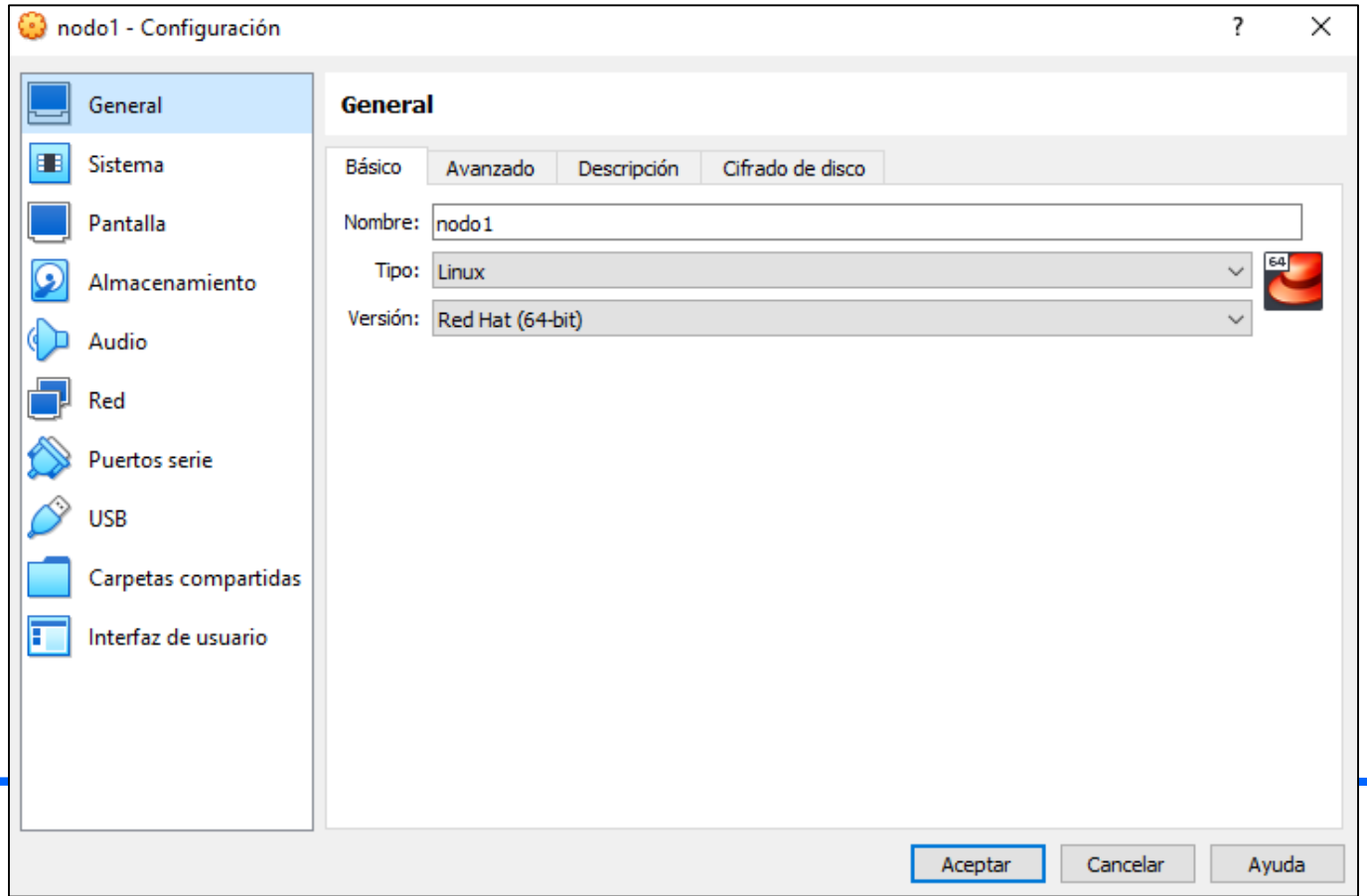
2. INSTALACION UBUNTU SERVER

Paso 6. Nos ha creado una maquina virtual con los requisitos indicados. Aquí nos pone el tipo de sistema operativo, la memoria base, la pantalla para la memoria de vídeo, el almacenamiento, etc



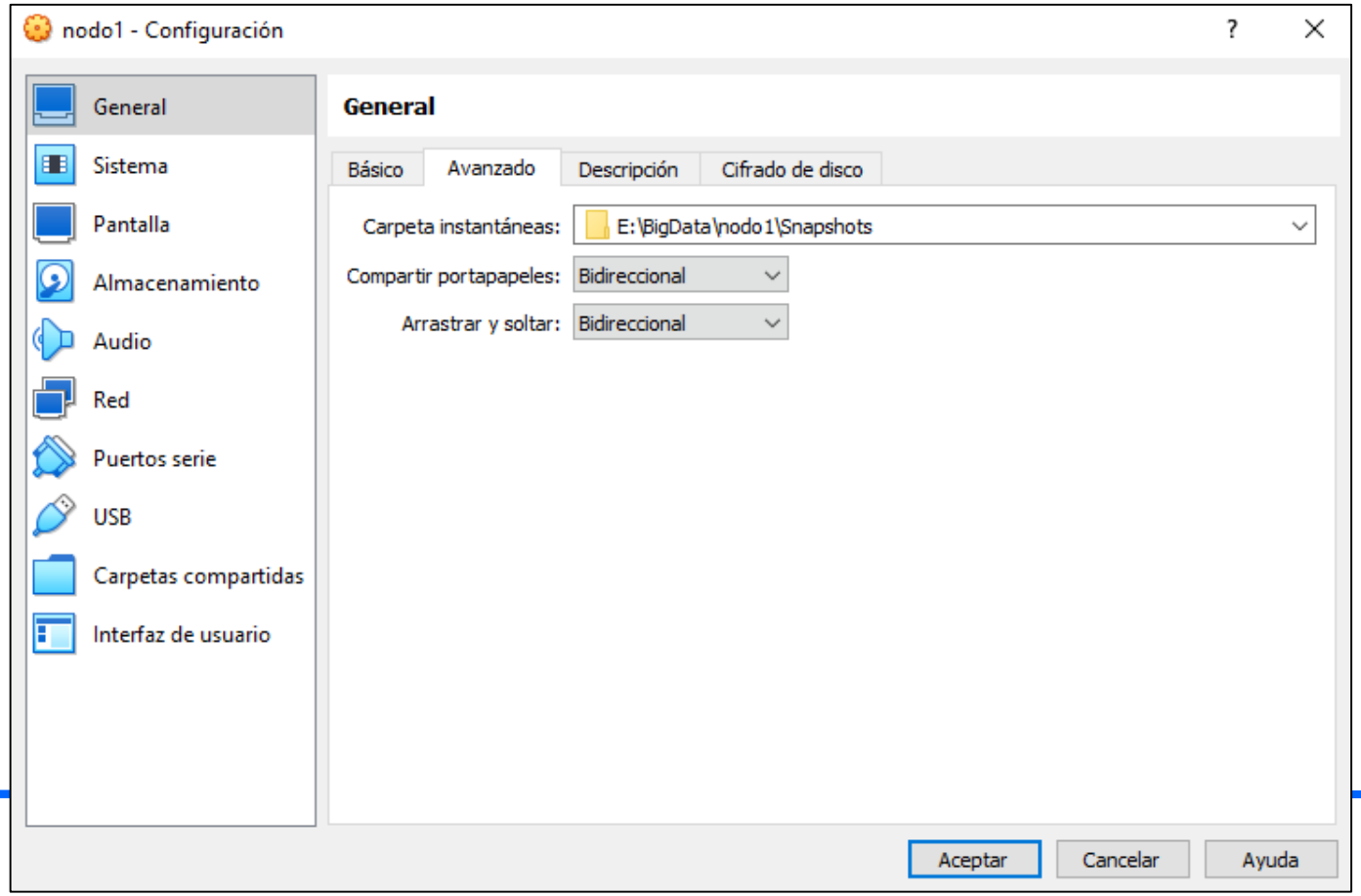
2. INSTALACION UBUNTU SERVER

Paso 7. Haremos algún pequeño cambio para que nuestra máquina sea más óptima. Para ello lo podemos hacer desde la pantalla general o ir a configuración del nodo1, donde nos aparece una pantalla con todas las características de la maquina



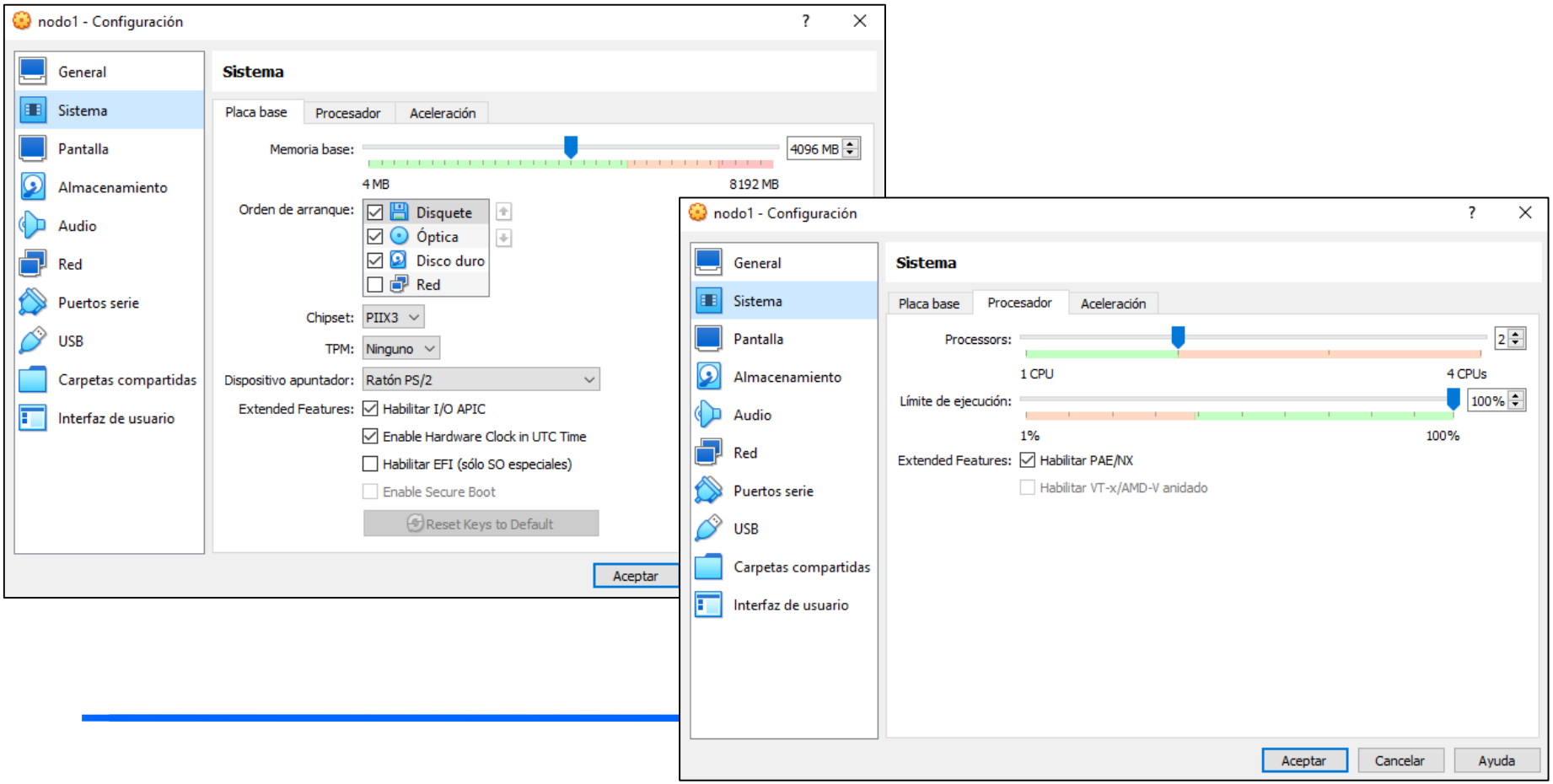
2. INSTALACION UBUNTU SERVER

Paso 8. Para trabajar mejor, vamos a General/Avanzado y pondremos compartir el portapapeles y arrastrar y soltar de forma Bidireccional, para que los copy&paste sean más óptimos.



2. INSTALACION UBUNTU SERVER

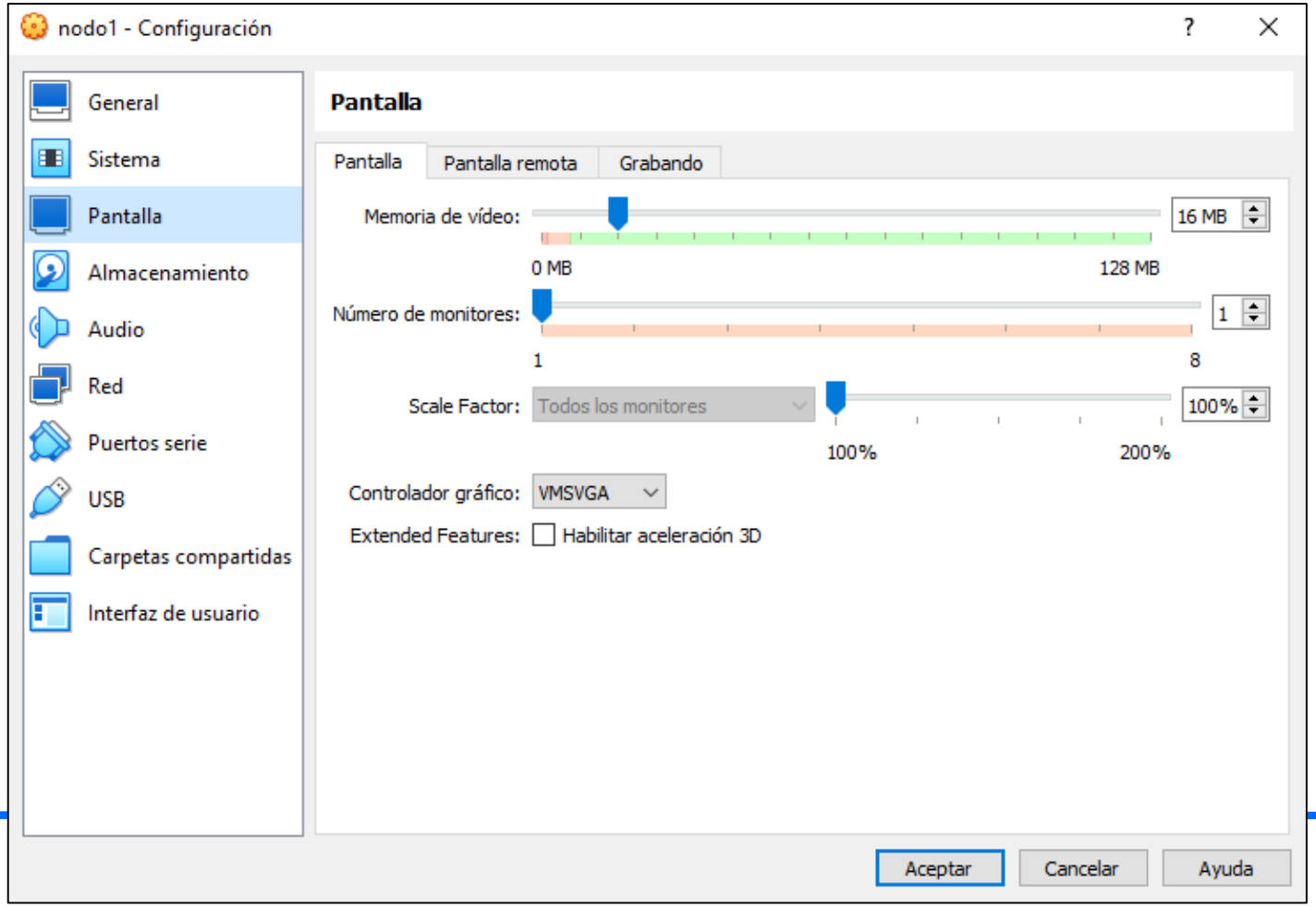
Paso 9. En Sistema/Placa Base comprobamos que efectivamente tenemos 4 GB y en procesador poner 2 aunque necesitaríamos 4 en teoría.



2. INSTALACION UBUNTU SERVER

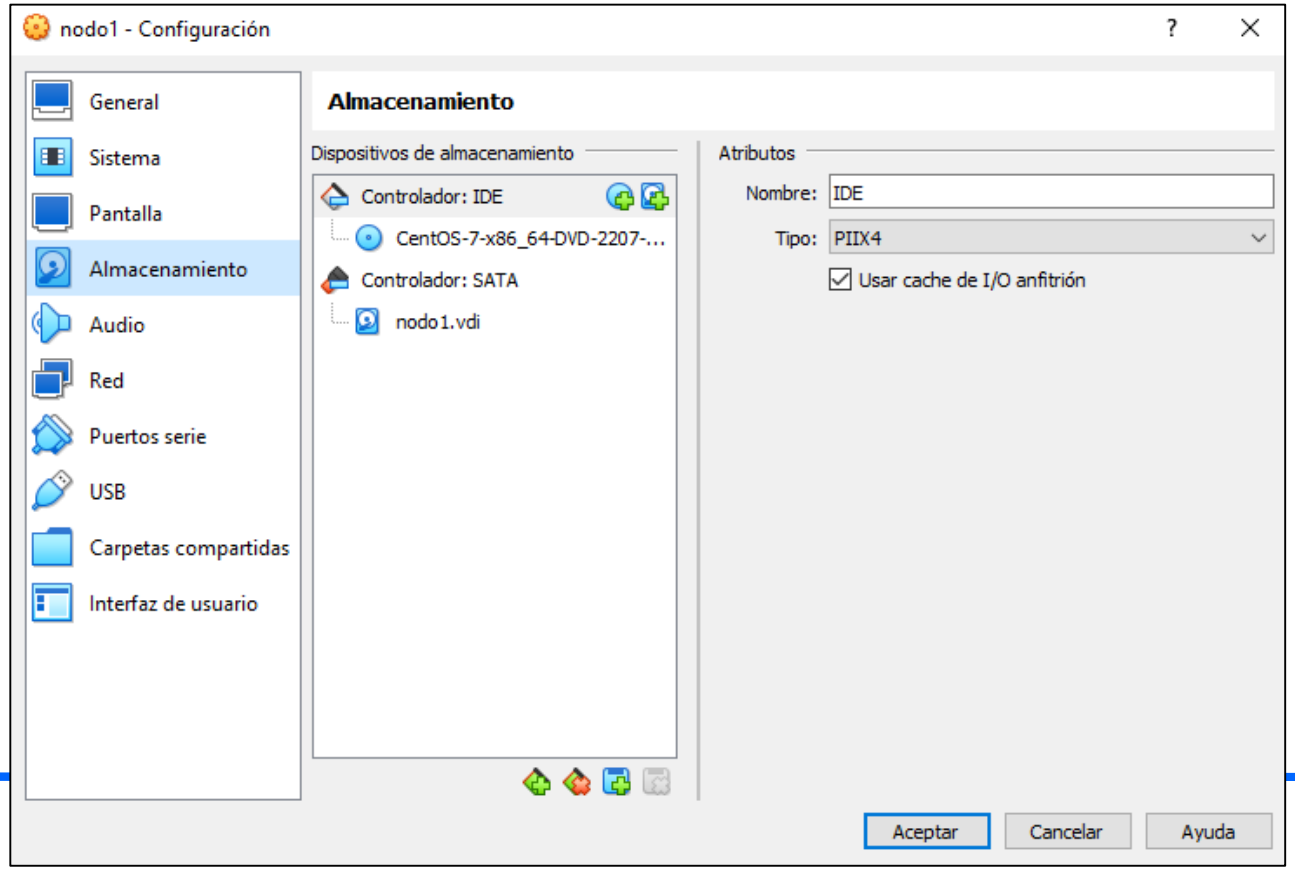
Paso 10. En la Pantalla podemos asignar una memoria un poquito más grande dependiendo de la tarjeta gráfica, pero no es muy importante

.



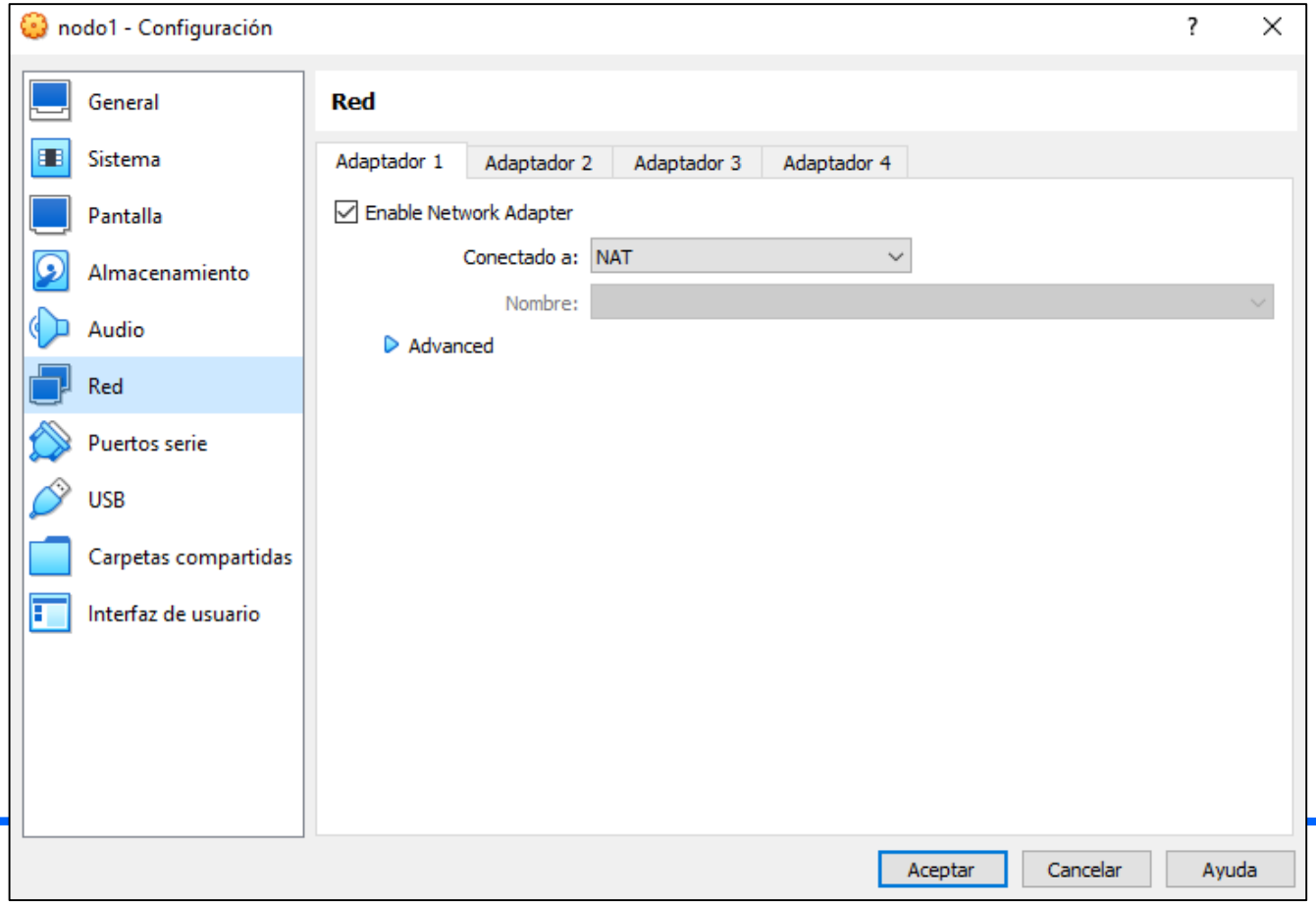
2. INSTALACION UBUNTU SERVER

Paso 11. En Almacenamiento tenemos una controladora de disco virtual SATA y una controladora IDE o con la unidad óptica o CDRom que contiene la imagen ISO del CentOS. Cuando arranquemos esta máquina virtual va a buscar el CD-ROM y va a intentar arrancar desde la ISO



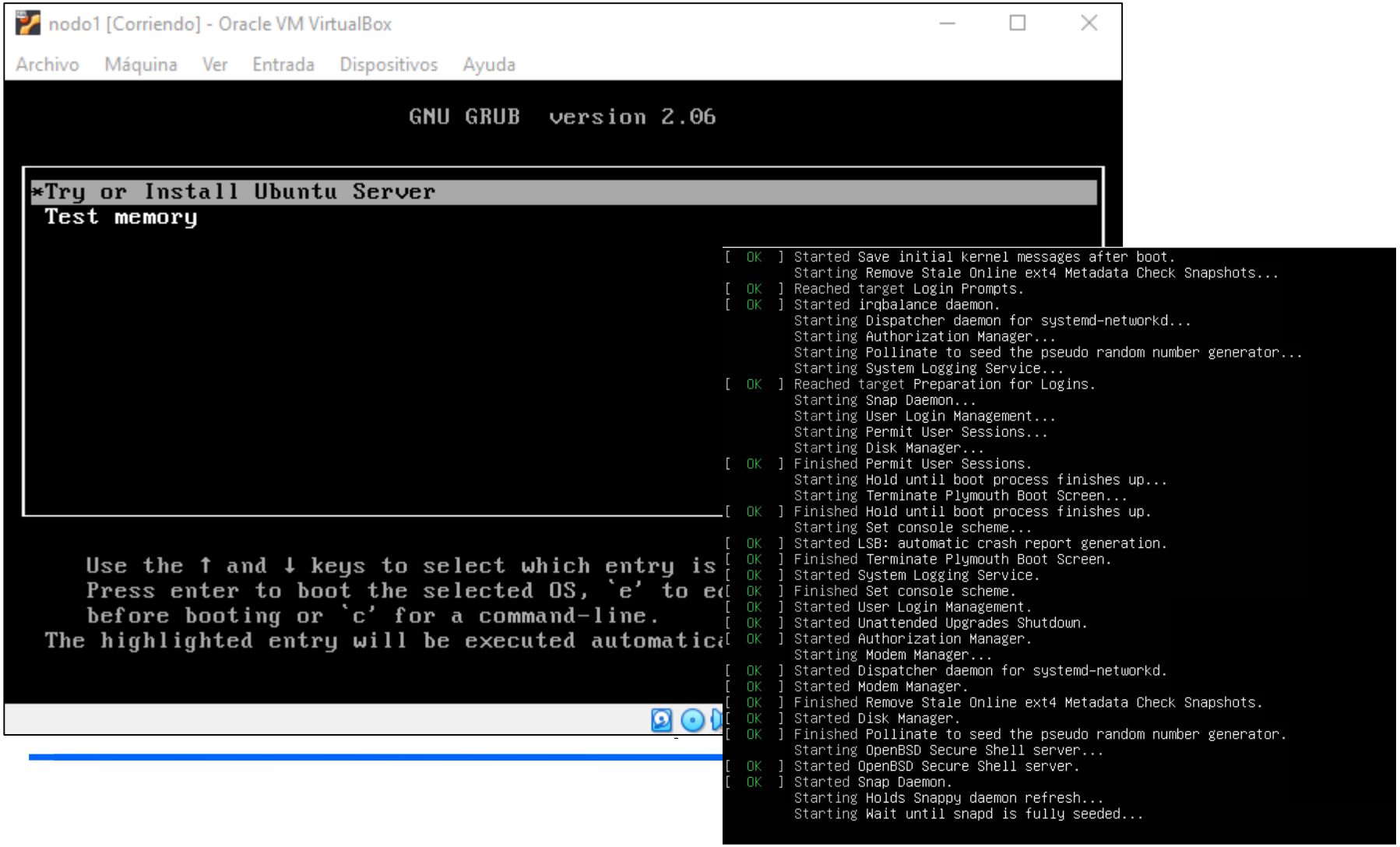
2. INSTALACION UBUNTU SERVER

Paso 12. En Red tenemos habilitada y configurada la red adaptador como NAT, es la forma de acceso a Internet
Aceptamos y guardamos las preferencias.



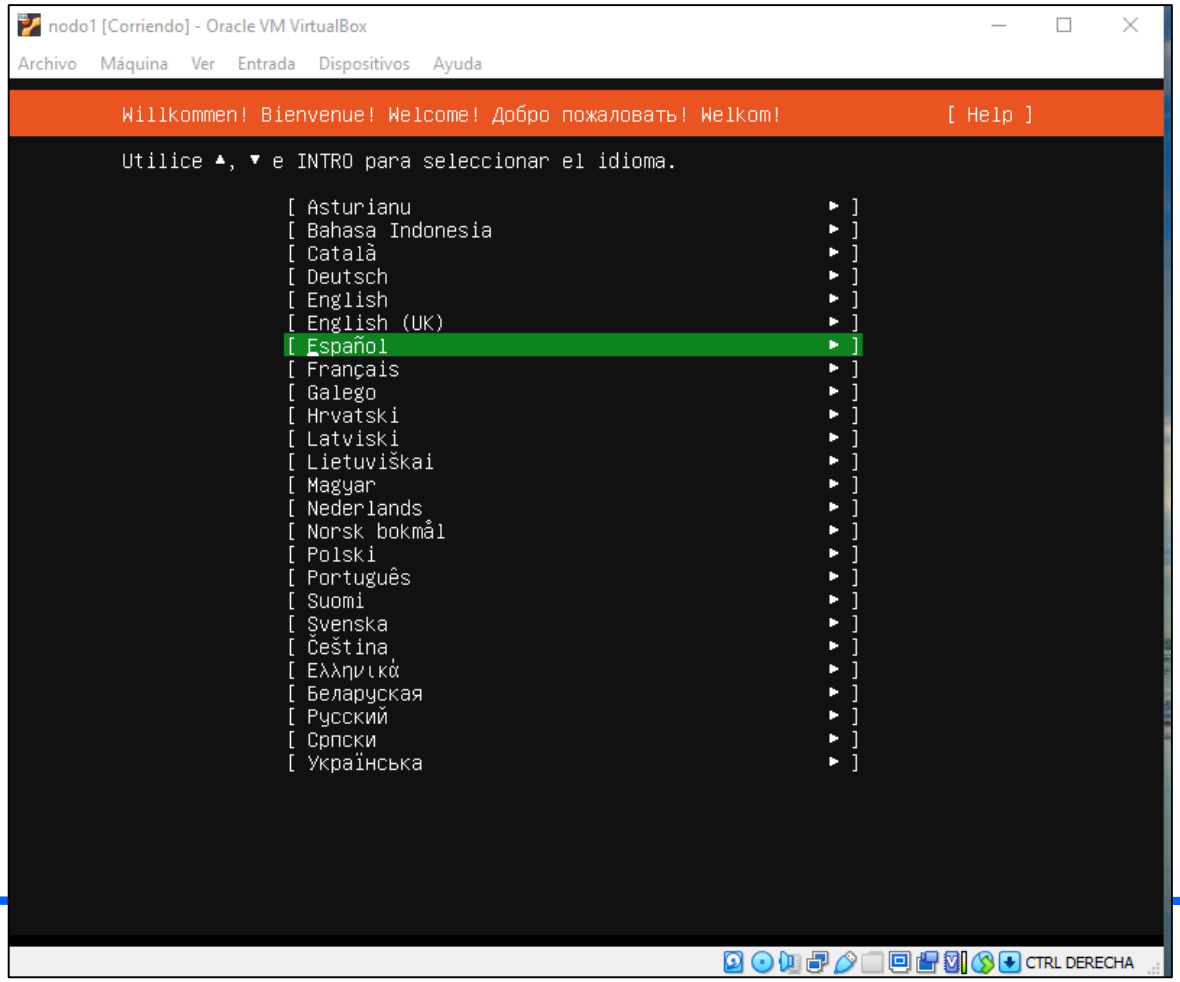
2. INSTALACION UBUNTU SERVER

Paso 13. Damos al botón Iniciar. Y seleccionamos la primera opción



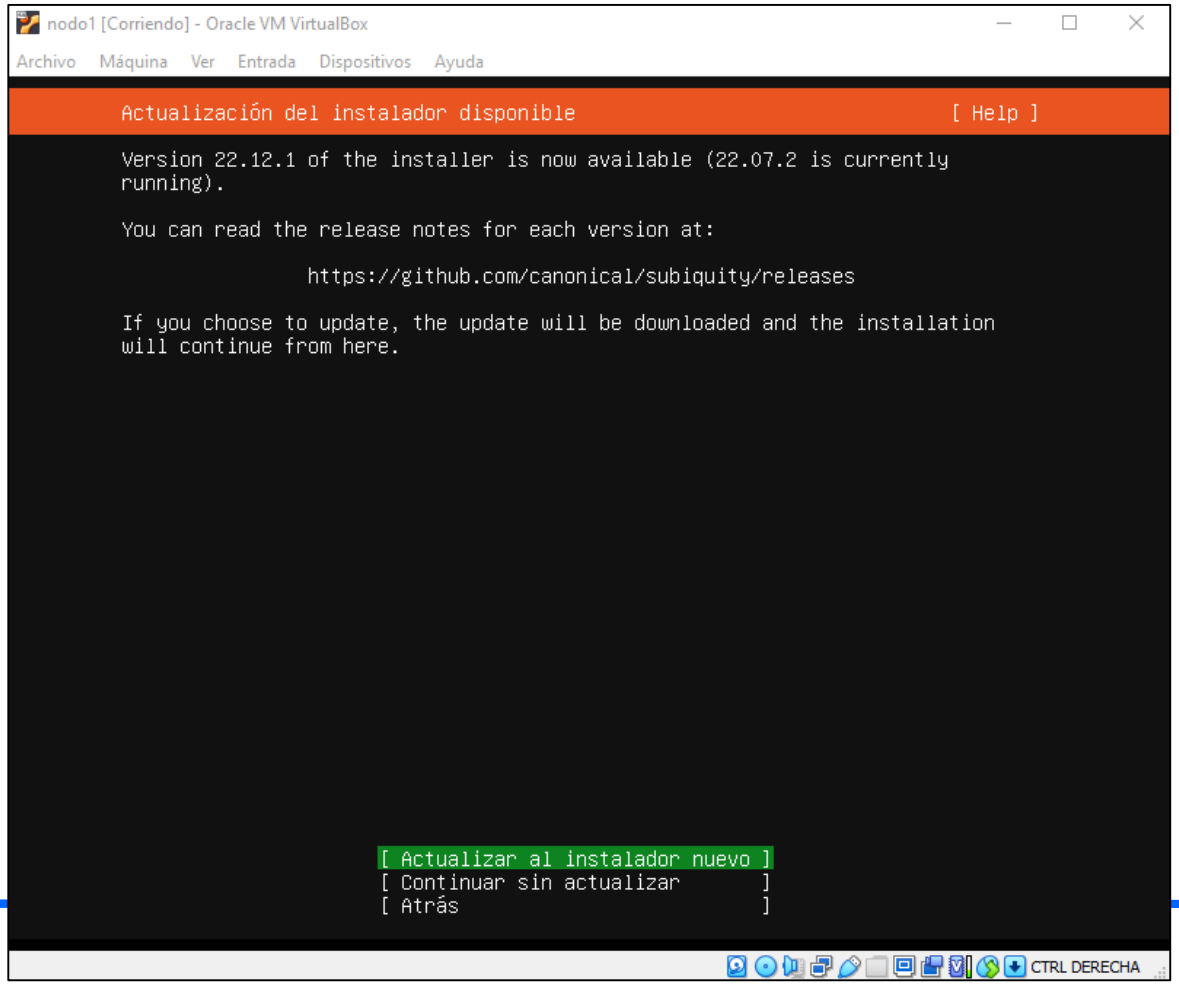
2. INSTALACION UBUNTU SERVER

Paso 14. La primera pantalla nos pide seleccionar el idioma que utilizaremos en el S.O.:



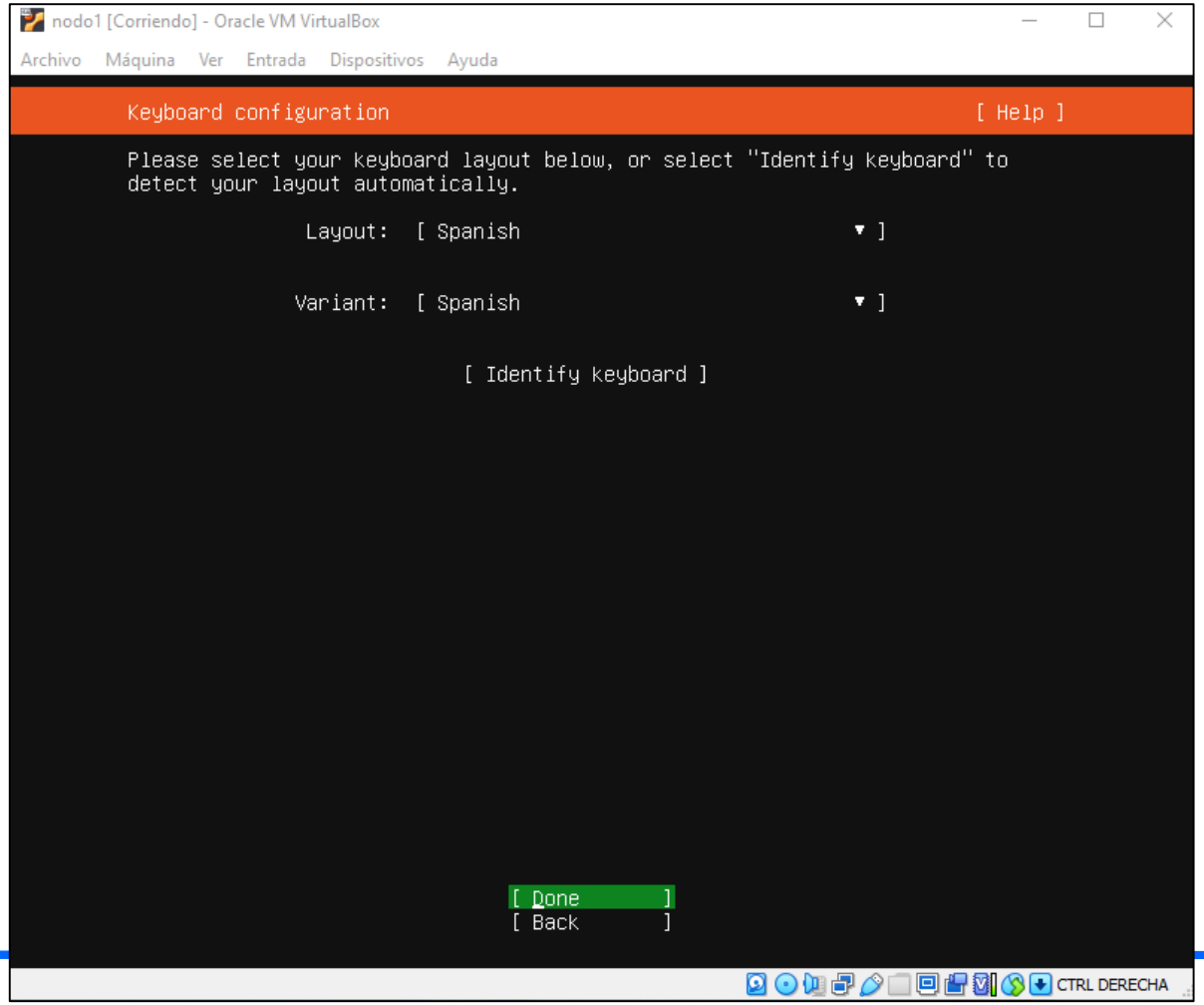
2. INSTALACION UBUNTU SERVER

Paso 15. En la siguiente pantalla podemos elegir actualizar el instalador a una versión mas nueva:



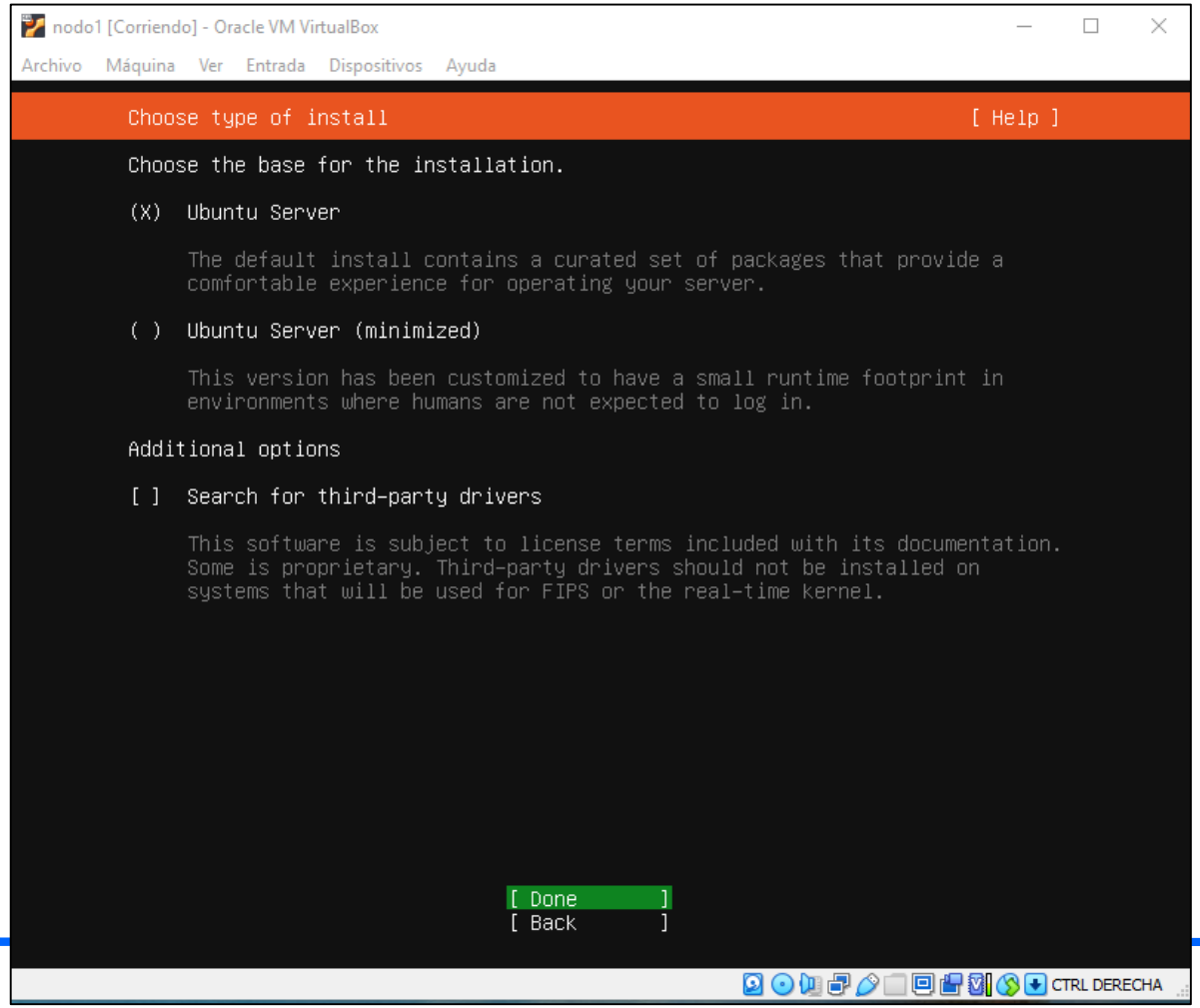
2. INSTALACION UBUNTU SERVER

Paso 16. Seleccionaremos el idioma del teclado:



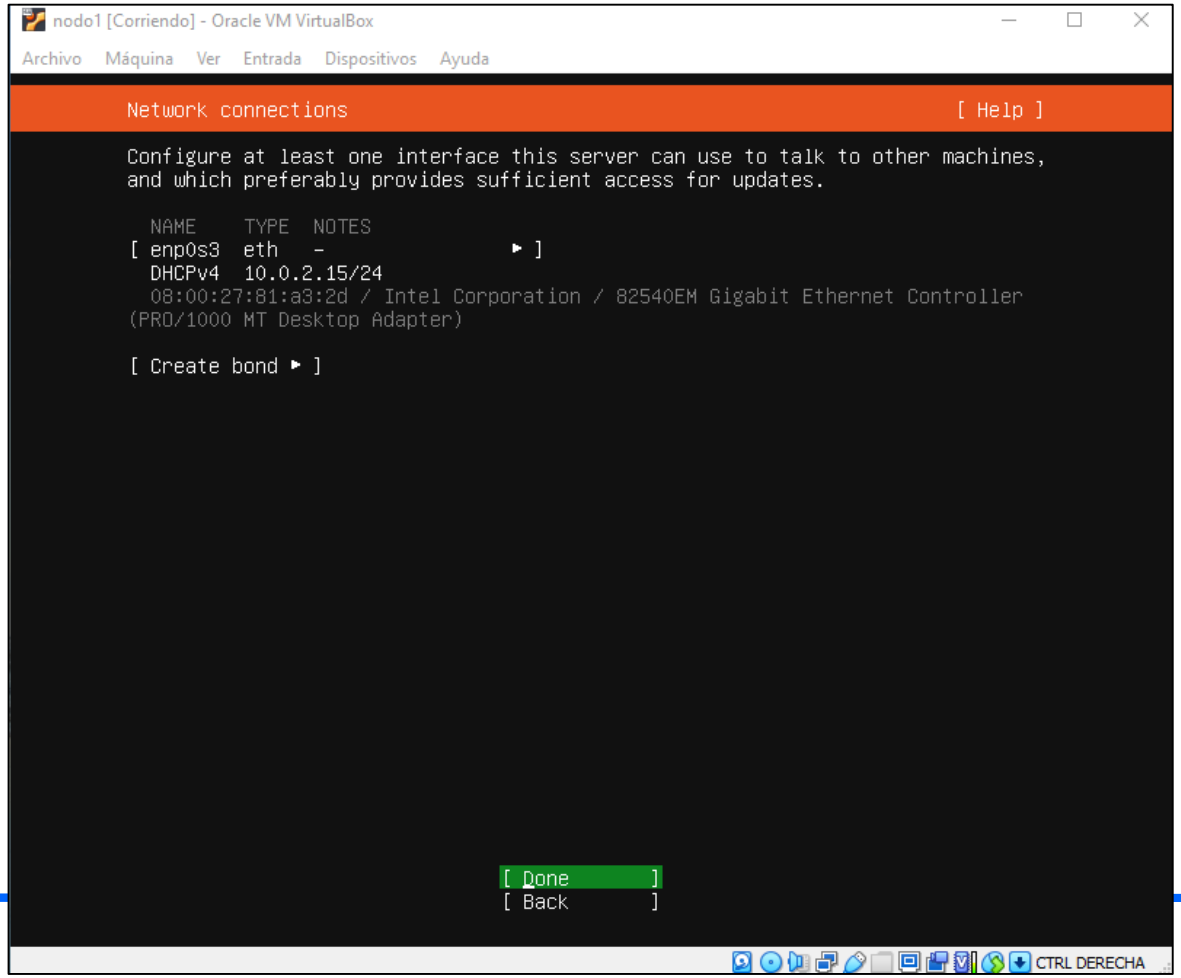
2. INSTALACION UBUNTU SERVER

Paso 17. Indicamos el tipo de instalación normal: Ubuntu Server:



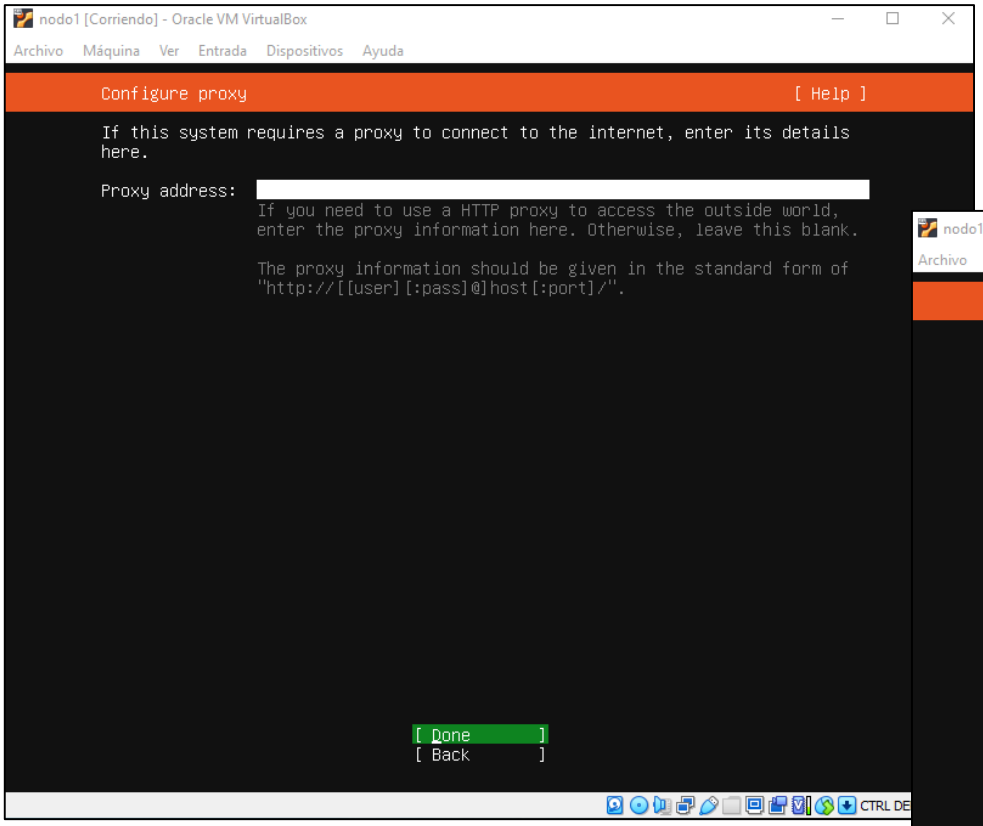
2. INSTALACION UBUNTU SERVER

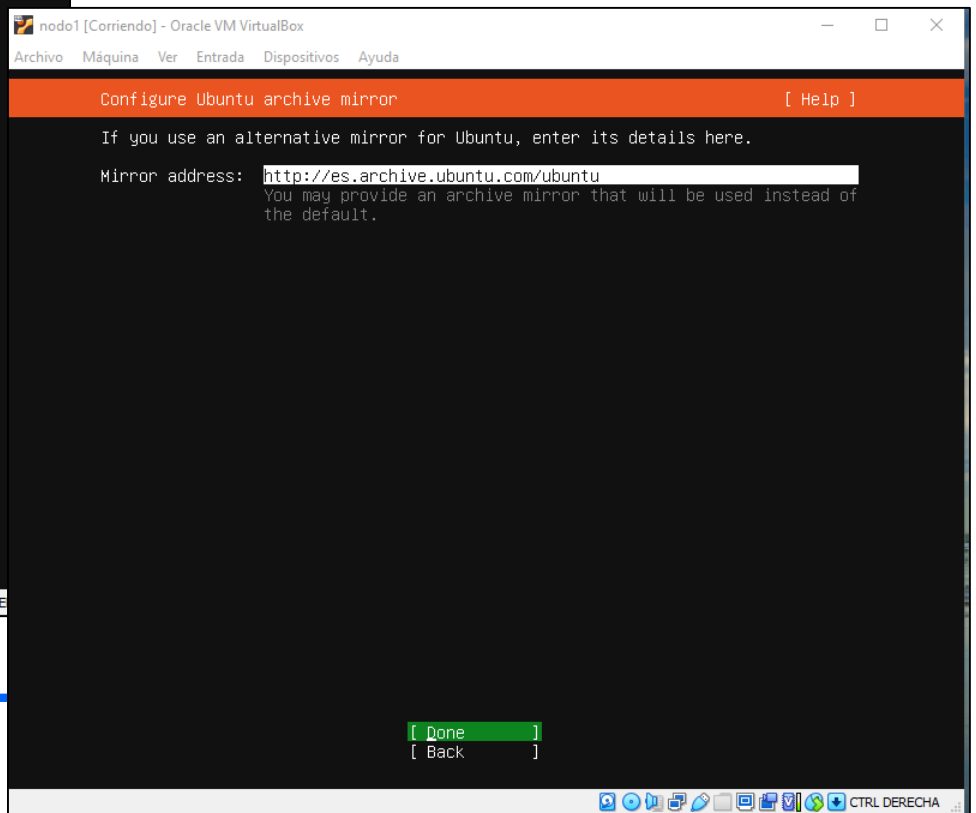
Paso 18. Aceptamos la selección de interficies del S.O. indicada y la IP que le ha proporcionado el NAT de virtual Box



2. INSTALACION UBUNTU SERVER

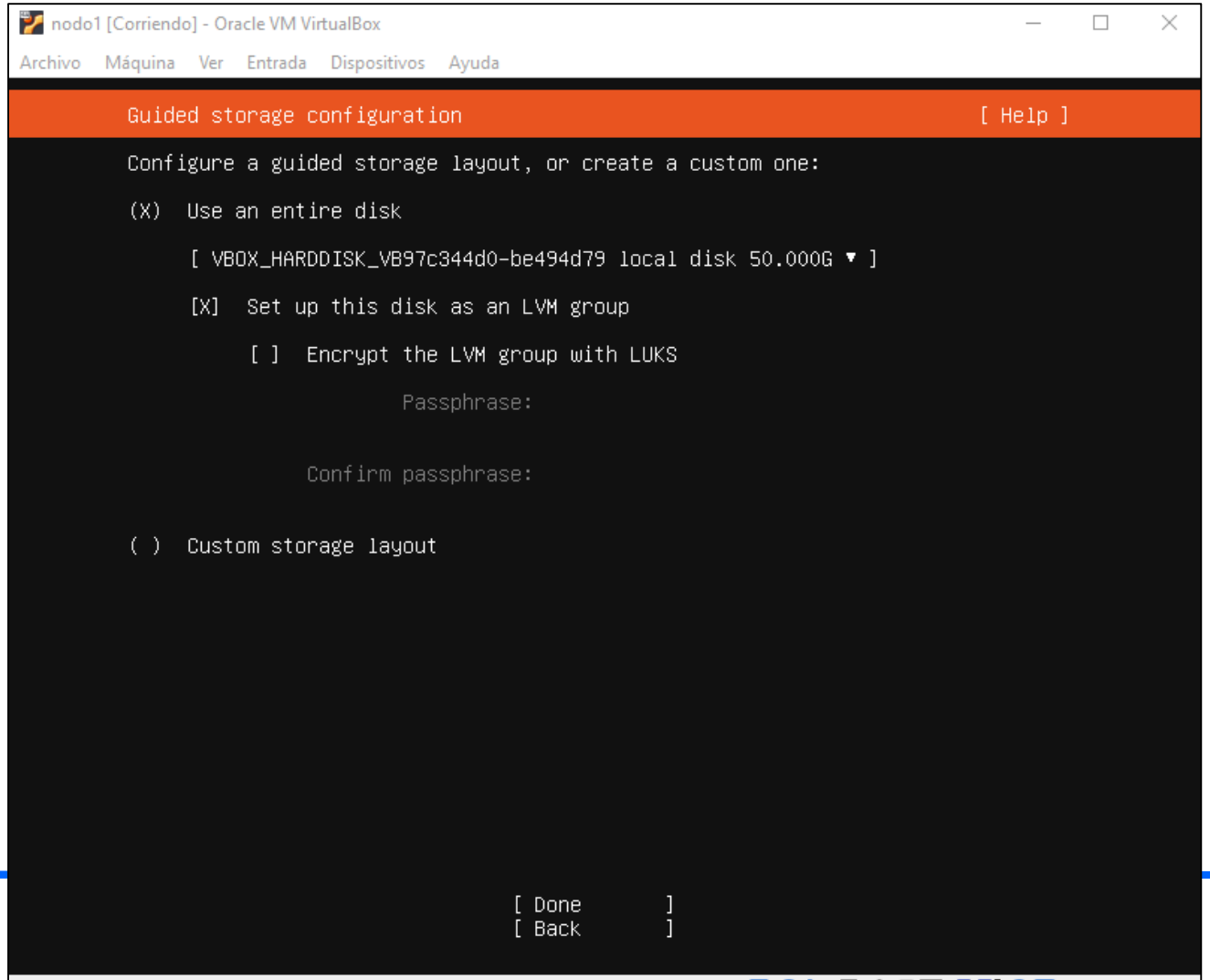
Paso 19. Saltamos el proxy server y dejamos el servidor mirror de Ubuntu por defecto que nos propone:





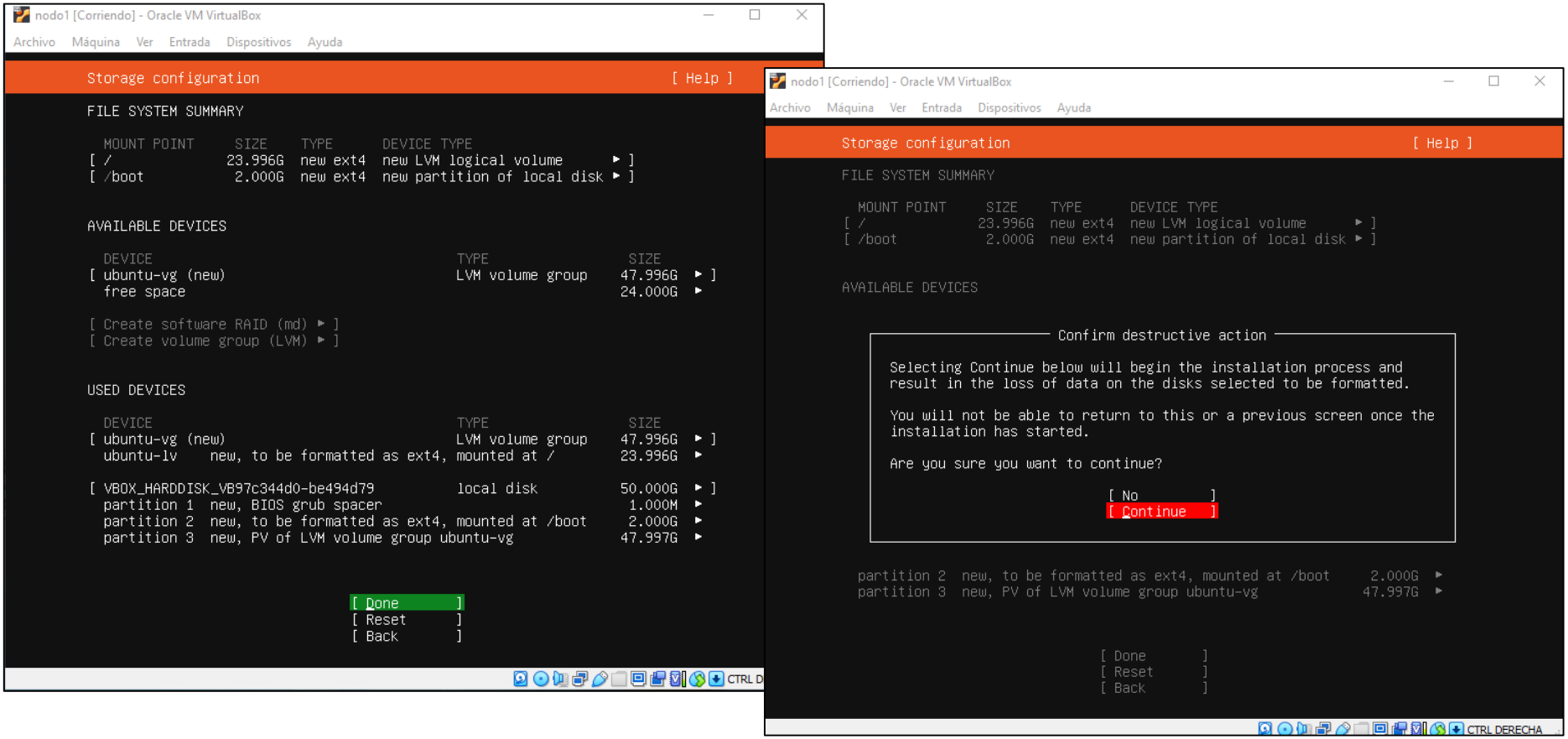
2. INSTALACION UBUNTU SERVER

Paso 20. Indicamos que queremos utilizar el espacio del disco entero



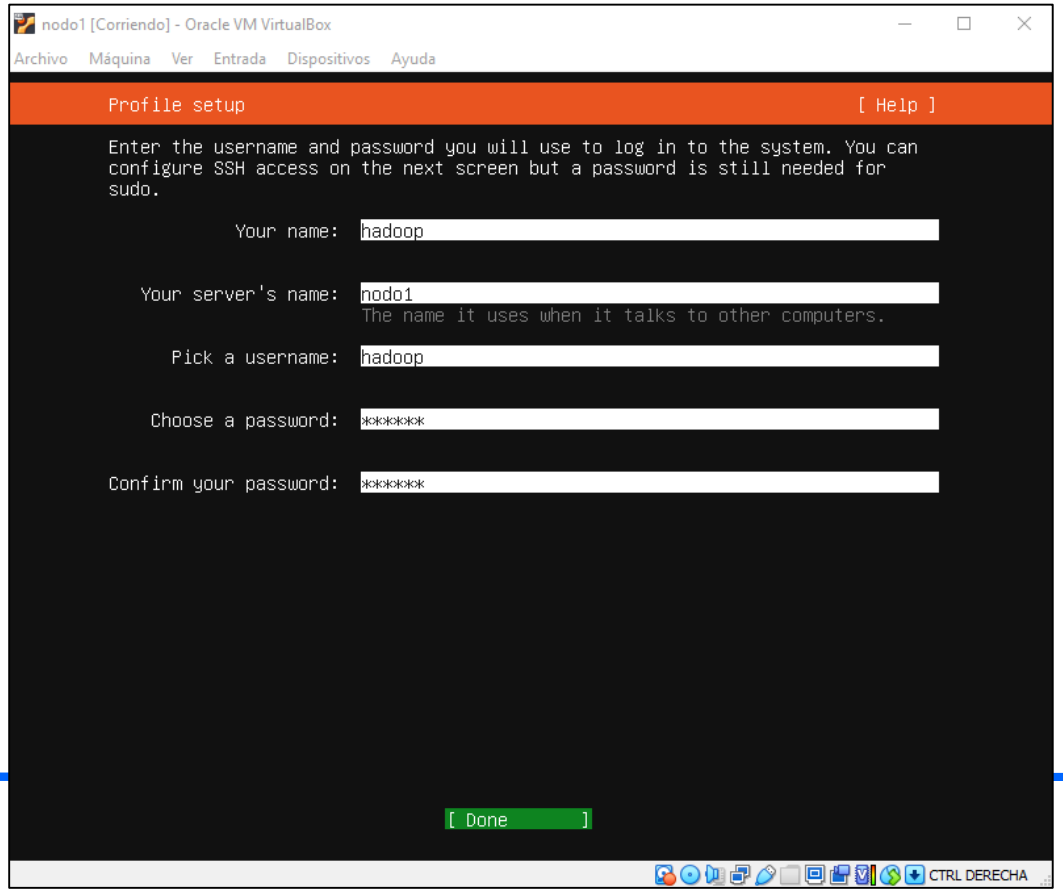
2. INSTALACION UBUNTU SERVER

Paso 21. Obtenemos un resumen del sistema de ficheros. A la confirmación de que se formateara el disco duro y que se perderán los datos indicamos la opción de continuar:



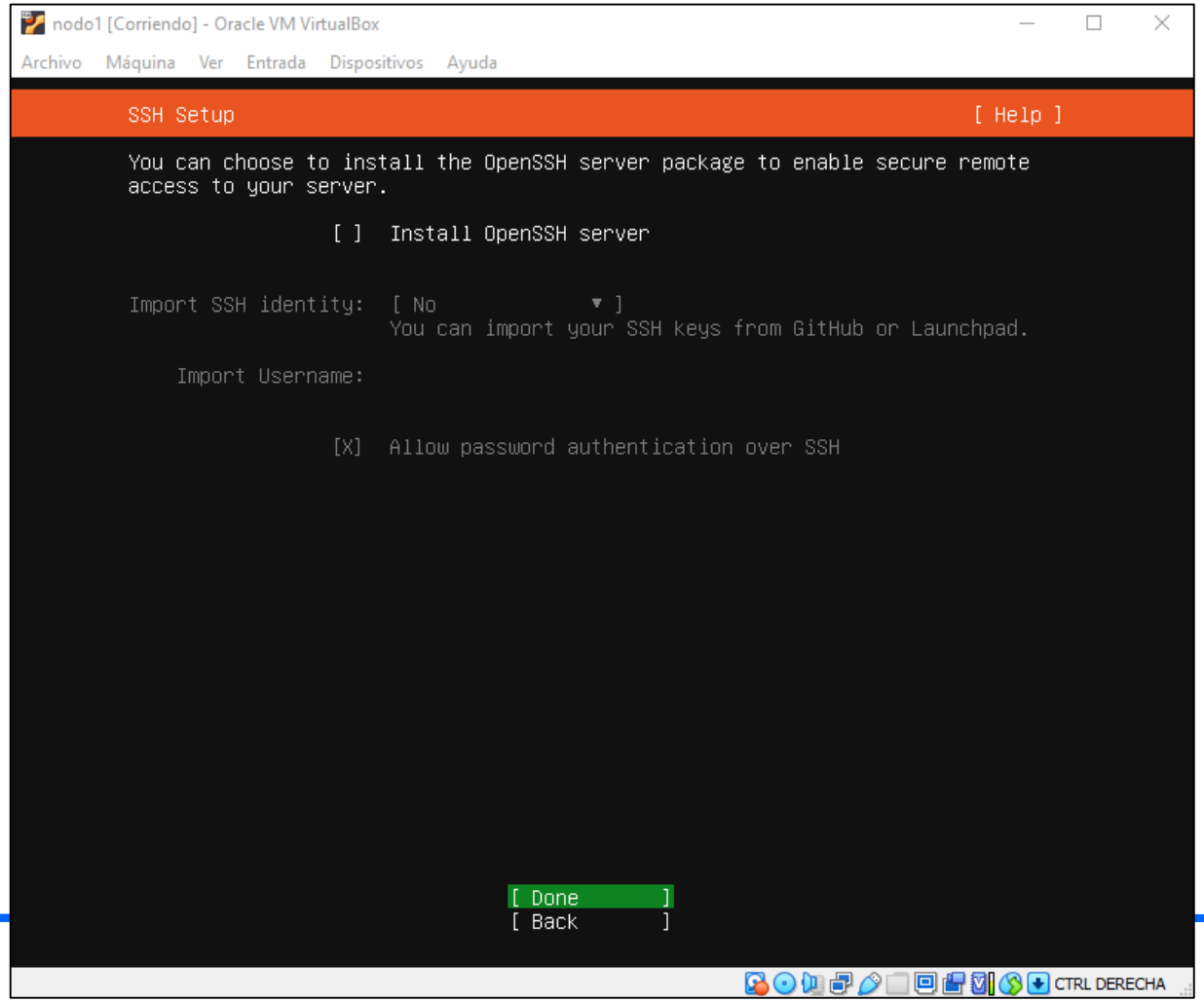
2. INSTALACION UBUNTU SERVER

Paso 22. Creamos un usuario, de nombre hadoop con contraseña hadoop. No es aconsejable trabajar con el cluster hadoop con el usuario root. En un entorno de real de producción no se debe utilizar el usuario root porque representa un problema de seguridad. El nombre del servidor nodo1



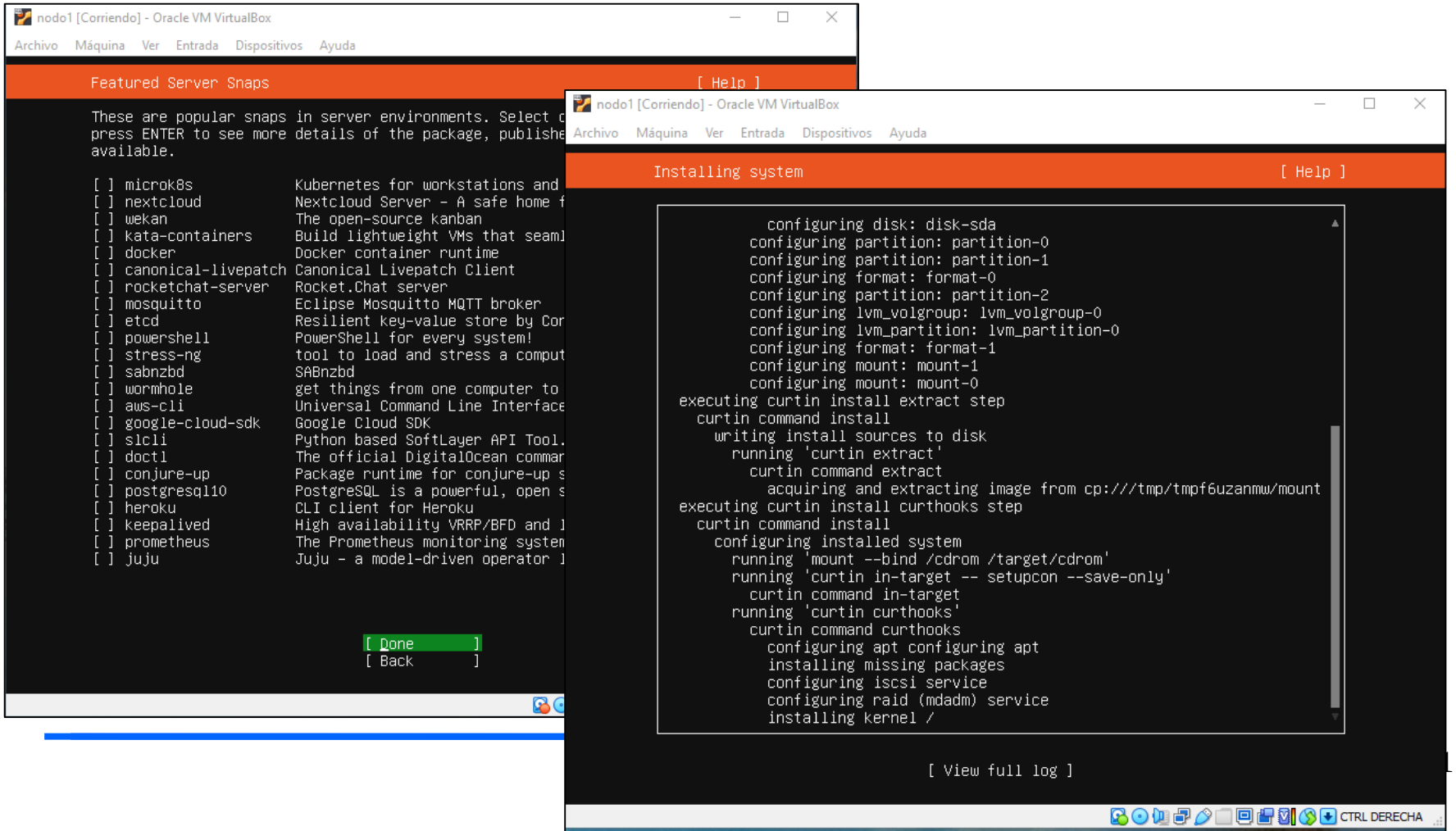
2. INSTALACION UBUNTU SERVER

Paso 23. Podemos instalar el servidor SSH aquí o sino más adelante



2. INSTALACION UBUNTU SERVER

Paso 24. En principio no instalamos ninguna característica mas de servidor. Empieza la instalación de kernel de Linux Ubuntu:



2. INSTALACION UBUNTU SERVER

Paso 25. Finalmente acaba y reiniciamos. Nos logamos con el usuario hadoop y password hadoop:

```
Install complete!                                                    [ He

      configuring installed system
        running 'mount --bind /cdrom /target/cdrom'
        running 'curtin in-target -- setupcon --save-only'
          curtin command in-target
        running 'curtin curthooks'
          curtin command curthooks
            configuring apt configuring apt
            installing missing packages
            configuring iscsi service
            configuring raid (mdadm) service
            installing kernel
            setting up swap
            apply networking config
            writing etc/fstab
            configuring multipath
            updating packages on target system
            configuring pollinate user-agent on target
            updating initramfs configuration
            configuring target system bootloader
            installing grub to target devices
final system configuration
  configuring cloud-init
  calculating extra packages to install
  downloading and installing security updates
    curtin command in-target
  restoring apt configuration
    curtin command in-target
subiquity/Late/run

                              [ View full log ]
                              [ Reboot Now    ]
```

```
nodo11 login: hadoop
Password: _
```

```
hadoop@nodo11:~$ pwd
/home/hadoop
hadoop@nodo11:~$ _
```

2. INSTALACION UBUNTU SERVER

Paso 26. Ejecutamos los siguientes comandos:

sudo apt update

sudo apt install xserver-xorg-core xserver-xorg-input-all -y

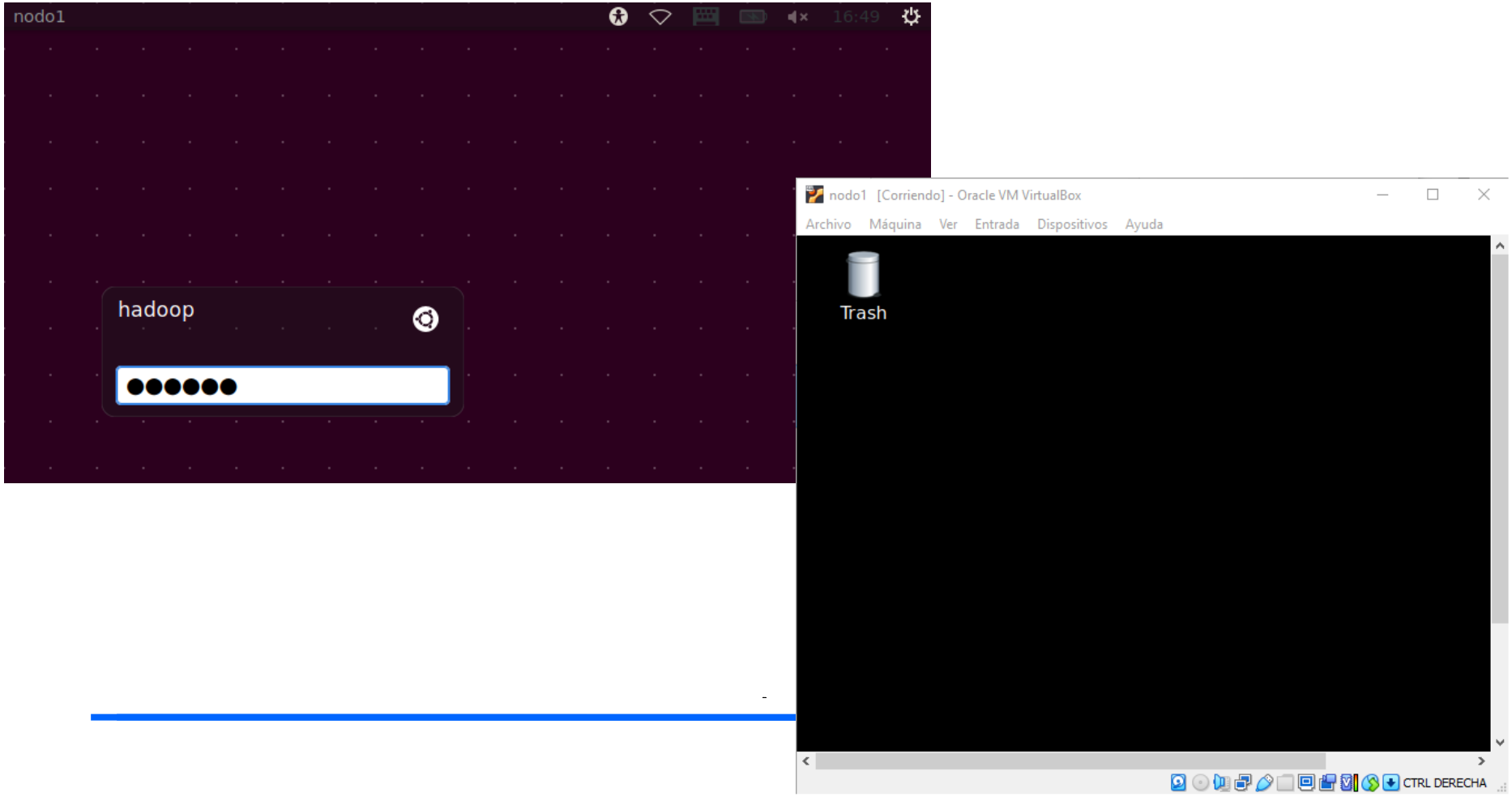
sudo apt install xorg lxde-core lxde-icon-theme –y

```
hadoop@nodo1:~$ sudo apt install xserver-xorg-core xserver-xorg-input-all -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
xserver-xorg-input-all is already the newest version (1:7.7+23ubuntu2).
xserver-xorg-core is already the newest version (2:21.1.3-2ubuntu2.5).
0 upgraded, 0 newly installed, 0 to remove and 64 not upgraded.
hadoop@nodo1:~$
```

```
hadoop@nodo1:~$ sudo apt install xorg lxde-core lxde-icon-theme -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
xorg is already the newest version (1:7.7+23ubuntu2).
lxde-core is already the newest version (11).
lxde-icon-theme is already the newest version (0.5.1-2.1).
0 upgraded, 0 newly installed, 0 to remove and 64 not upgraded.
hadoop@nodo1:~$ _
```

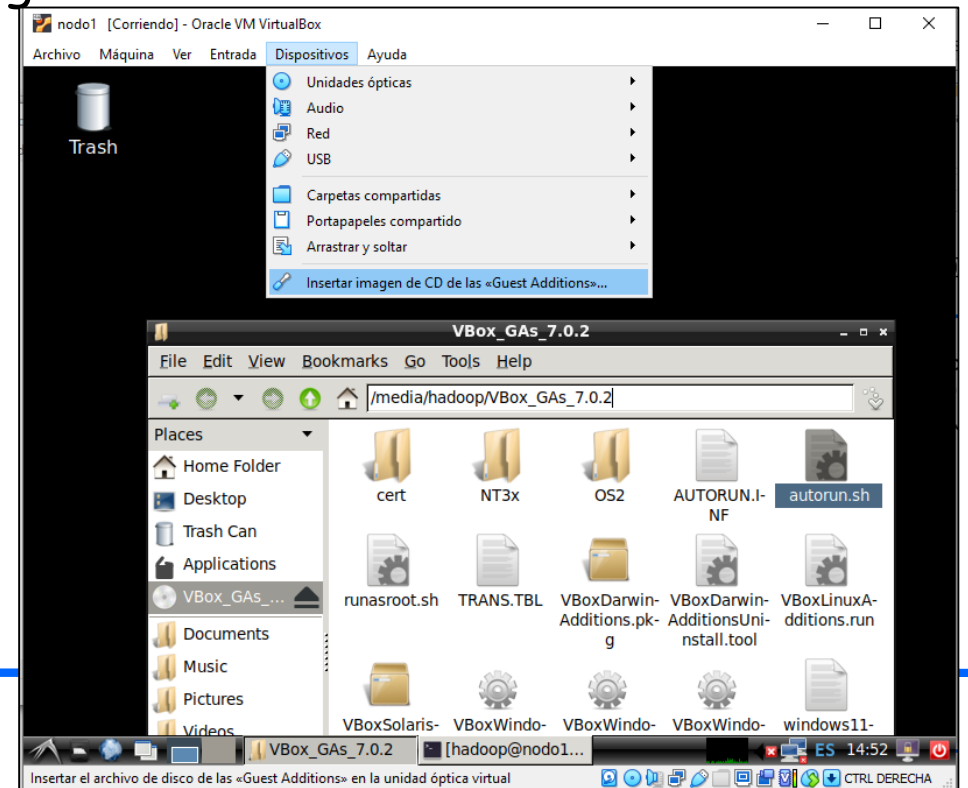
2. INSTALACION UBUNTU SERVER

Paso 27. Reiniciamos el equipo y observamos la mínima parte grafico del S.O. que nos ayudara a interaccionar con el nodo:



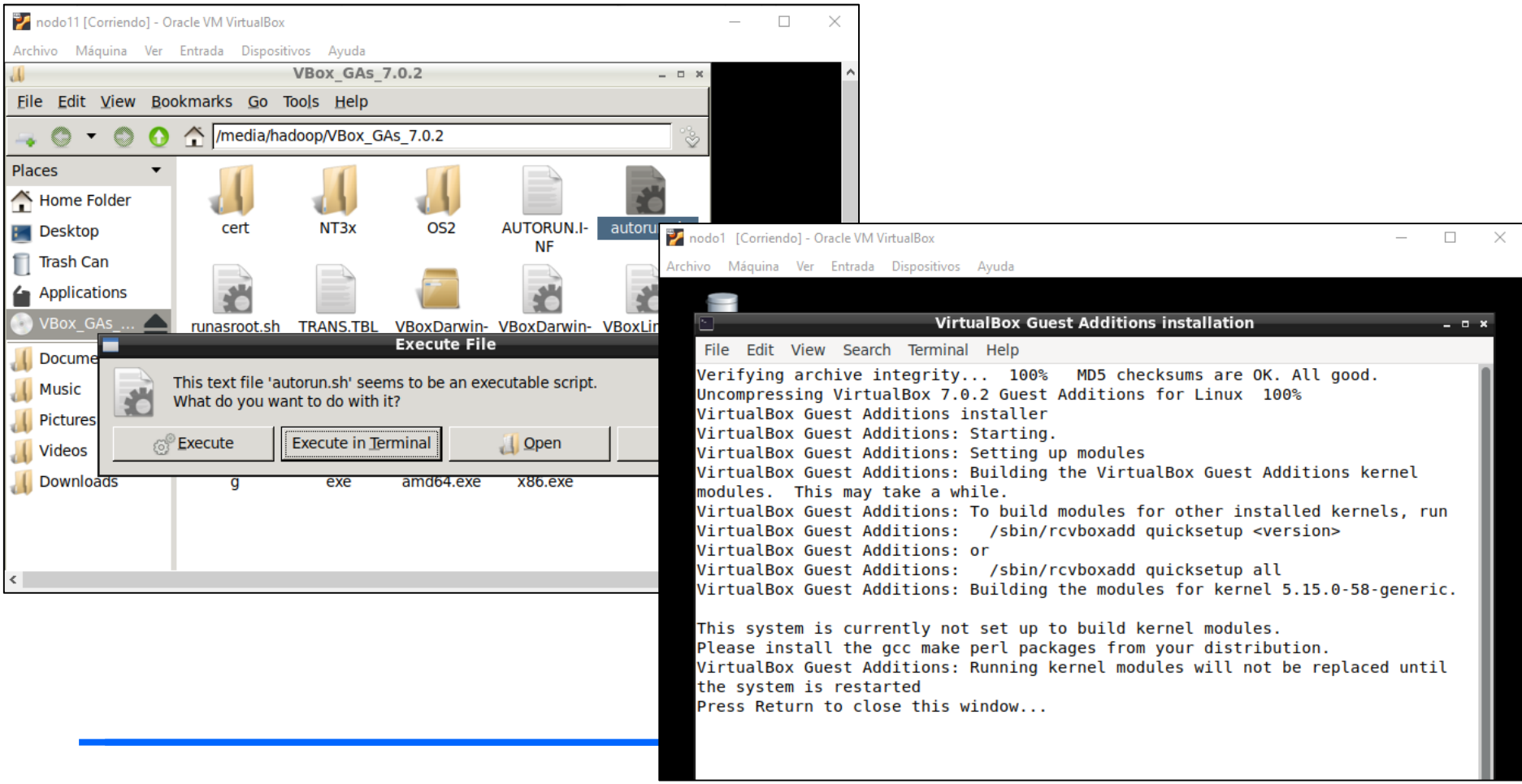
2. INSTALACION UBUNTU SERVER

Paso 28. Instalaremos las Guest Additions. Es un componente de VirtualBox que nos permite movernos libremente con el ratón entre la máquina virtual y nuestro PC (evitando tener que pulsar la tecla Ctrl de la derecha) y también para poder tener una mejor resolución y una mejor gestión de la pantalla. Vamos a Dispositivos y hacemos click en Insertar Imagen de CD de las Guest Additions".



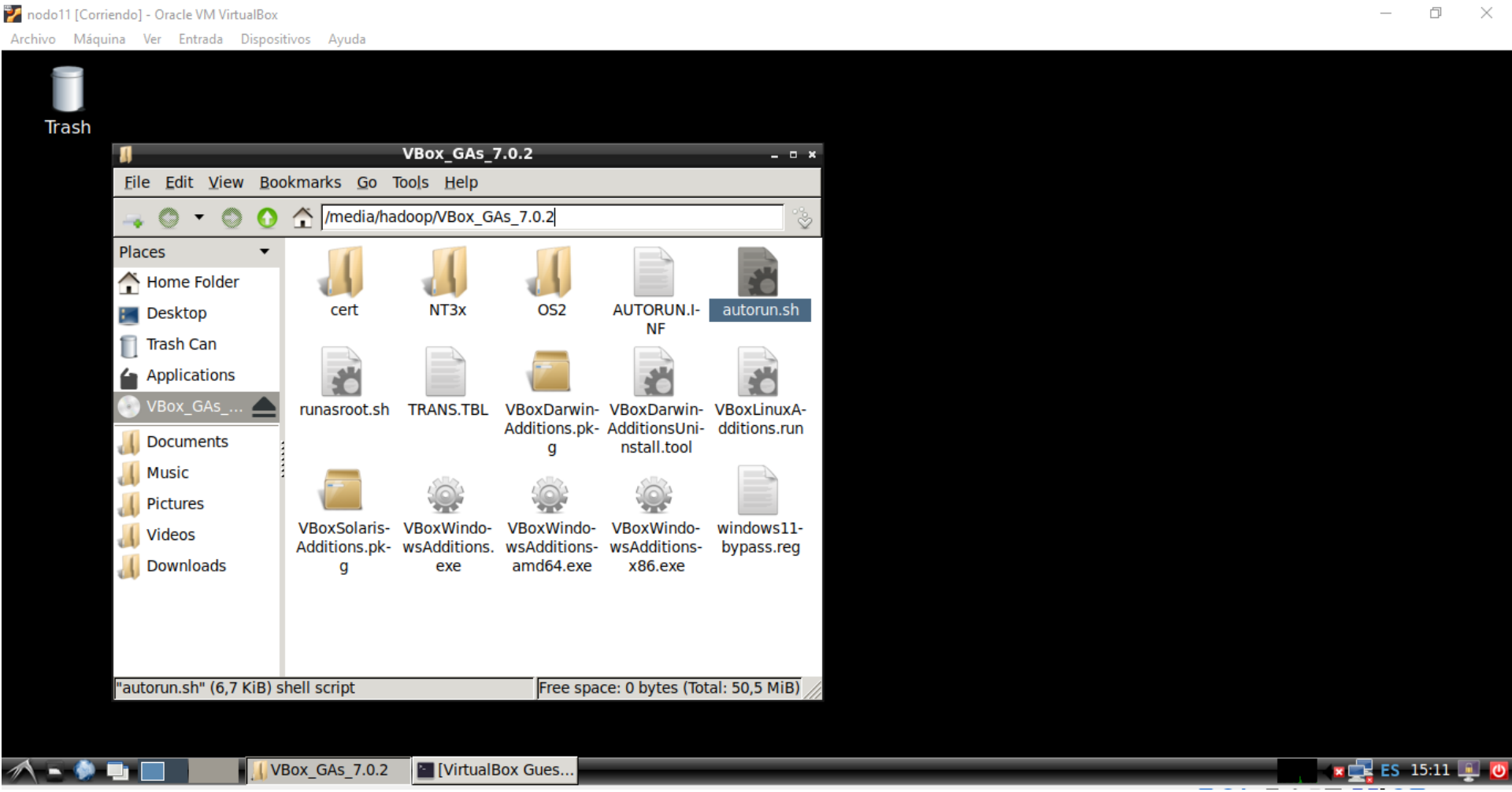
2. INSTALACION UBUNTU SERVER

Paso 29. Instalaremos el paquete bzip2 y ejecutamos autorun.sh.



2. INSTALACION UBUNTU SERVER

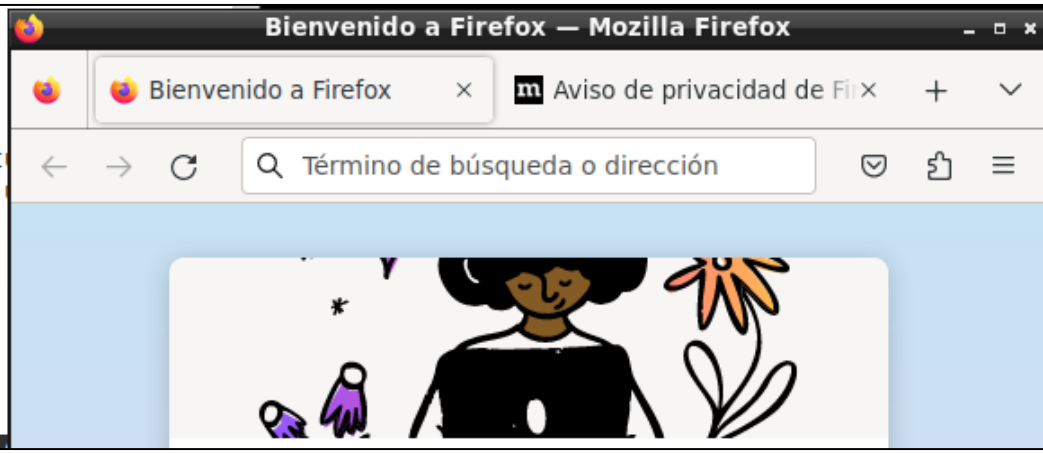
Paso 30. Finalmente la maquina virtual nodo1 ocupará toda la pantalla



2. INSTALACION UBUNTU SERVER

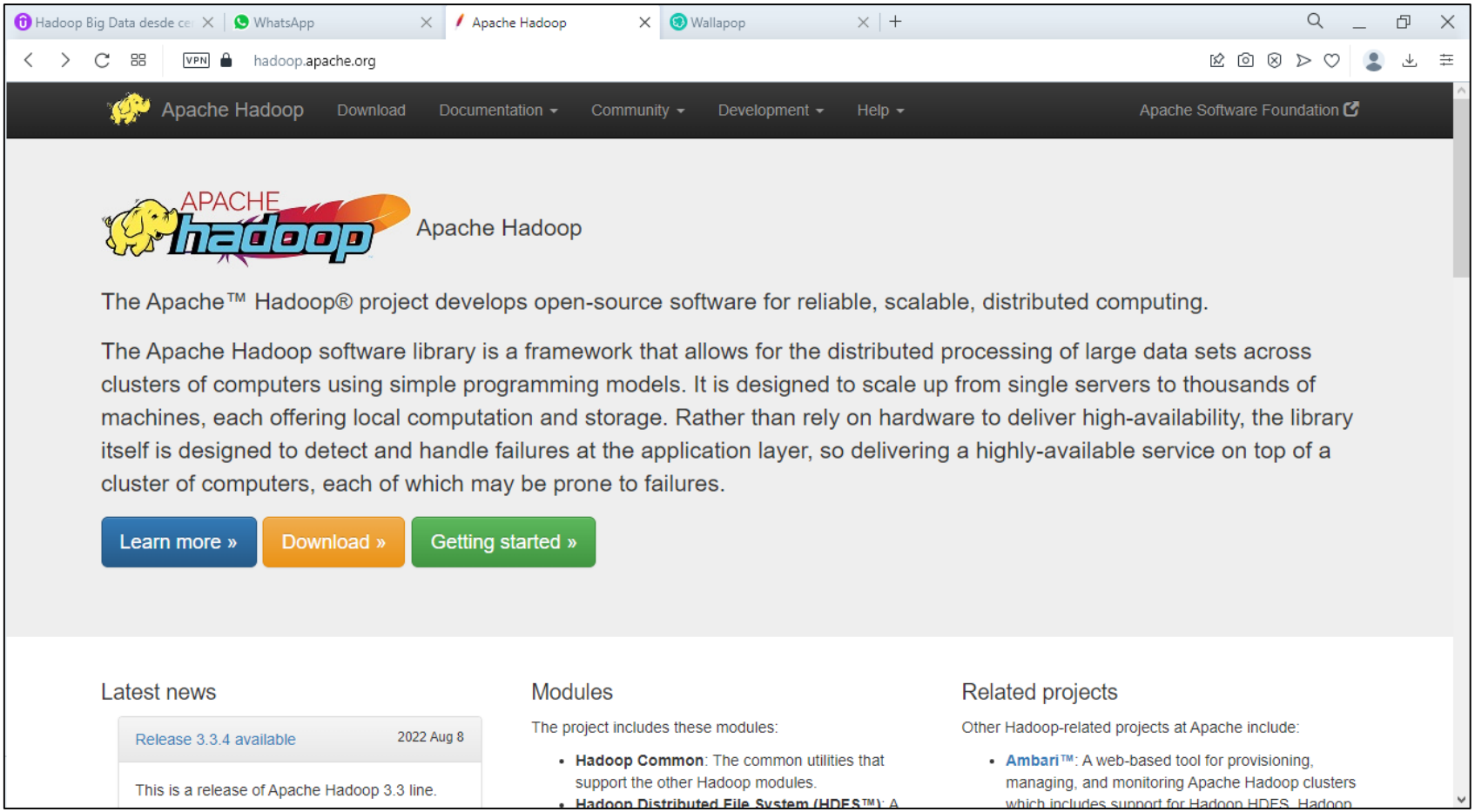
Paso 31. Por ultimo instalamos el navegador Firefox y lo ejecutamos:

```
hadoop@nodo11:~$ sudo apt install firefox
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
firefox is already the newest version (1:1snap1-0ubunt
0 upgraded, 0 newly installed, 0 to remove and 96 not
hadoop@nodo11:~$ firefox&
[1] 58098
/usr/bin/firefox: 12: xdg-settings: not found
hadoop@nodo11:~$
hadoop@nodo11:~$
hadoop@nodo11:~$
hadoop@nodo11:~$
```



3. INSTALACION HADOOP

Paso 1. Vamos a la página hadoop https://hadoop.apache.org
Hadoop es un proyecto de los múltiples que tiene Apache.



3. INSTALACION HADOOP

Paso 2. Hadoop está compuesto realmente de:

- Un core con los módulos básicos de hadoop
- Un sistema de ficheros HDFS
- Hadoop Yarn es la versión más moderna
- Hadoop MapReduce

Modules

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.

3. INSTALACION HADOOP

Paso 3. Los diferentes proyectos alrededor de hadoop:

- Ambari
- Avro
- Cassandra
- Chukwa
- HBase
- Hive
- Mahout
- Ozone
- Pig
- Spark
- Submarine
- Tez
- ZooKeeper

Related projects

Other Hadoop-related projects at Apache include:

- **Ambari™**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Avro™**: A data serialization system.
- **Cassandra™**: A scalable multi-master database with no single points of failure.
- **Chukwa™**: A data collection system for managing large distributed systems.
- **HBase™**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive™**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Mahout™**: A Scalable machine learning and data mining library.
- **Ozone™**: A scalable, redundant, and distributed object store for Hadoop.
- **Pig™**: A high-level data-flow language and execution framework for parallel computation.
- **Spark™**: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Submarine**: A unified AI platform which allows engineers and data scientists to run Machine Learning and Deep Learning workload in distributed cluster.
- **Tez™**: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
- **ZooKeeper™**: A high-performance coordination service for distributed applications.

3. INSTALACION HADOOP

Paso 4. Descargamos el binario de la ultima versión de hadoop la 3.3.4 → hadoop-3.3.4.tar.gz

hadoop.apache.org/releases.html

Apache Hadoop | Download | Documentation | Community | Development | Help | Apache Software Foundation

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
3.3.4	2022 Aug 8	source (checksum signature)	binary (checksum signature) binary-aarch64 (checksum signature
3.2.4	2022 Jul 22	source (checksum signature)	binary (checksum signature)
2.10.2	2022 May 31	source (checksum signature)	binary (checksum signature)

nodo1 [Corriendo] - Oracle VM VirtualBox

Archivo Máquina Ver Entrada Dispositivos Ayuda

Apache Downloads — Mozilla Firefox

Apache Downloads

https://www.apache.org/dyn/closer.cgi/hadoop/comm

News | About | Make a D

hadoop-3.3.4.tar.gz
Completada — 663 MB

Mostrar todas las descargas

"THE APACHE WAY"

Projects | People | Community | License | Sponsors

We suggest the following site for your download:

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

Alternate download locations are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (`.asc` file) or a hash (`.md5` or `.sha*` file).

HTTP

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

Apache Downloa... | [VBox_GAs_7.0.2] | ES 12:11

3. INSTALACION HADOOP

Paso 5. Desde un terminal y ejecutamos los siguientes comandos para descomprimir los ficheros:

```
hadoop@nodo1 (10.0.2.15) - byobu
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:~$ pwd
/home/hadoop
hadoop@nodo1:~$ ls
Desktop  Documents  Downloads  Music  Pictures  Public  snap  Templates
hadoop@nodo1:~$ cd Downloads/
hadoop@nodo1:~/Downloads$ ls
hadoop-3.3.4.tar.gz
hadoop@nodo1:~/Downloads$ sudo tar xvf hadoop-3.3.4.tar.gz
```

```
hadoop-3.3.4/share/doc/hadoop/hadoop-hdfs-nfs/images/bg.jpg
hadoop-3.3.4/share/doc/hadoop/hadoop-hdfs-nfs/images/newwindow.png
hadoop-3.3.4/share/doc/hadoop/hadoop-hdfs-nfs/images/h3.jpg
hadoop@nodo1:~/Downloads$ ls
hadoop-3.3.4  hadoop-3.3.4.tar.gz
hadoop@nodo1:~/Downloads$ mv hadoop-3.3.4 hadoop
hadoop@nodo1:~/Downloads$ ls
hadoop  hadoop-3.3.4.tar.gz
hadoop@nodo1:~/Downloads$
```

3. INSTALACION HADOOP

Paso 6. Movemos la carpeta a hadoop a /opt, usando sudo de usuario root. Cambiamos el propietario de esta carpeta /opt/hadoop al usuario hadoop mediante el comando chown:

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:~/Downloads$ ls
hadoop  hadoop-3.3.4.tar.gz
hadoop@nodo1:~/Downloads$ mv hadoop /opt
mv: cannot move 'hadoop' to '/opt/hadoop': Permission denied
1 hadoop@nodo1:~/Downloads$ sudo mv hadoop /opt
hadoop@nodo1:~/Downloads$ ls /opt
hadoop
hadoop@nodo1:~/Downloads$ ls
hadoop-3.3.4.tar.gz
hadoop@nodo1:~/Downloads$ chown hadoop /opt/hadoop/
chown: changing ownership of '/opt/hadoop/': Operation not permitted
1 hadoop@nodo1:~/Downloads$ sudo chown hadoop /opt/hadoop/
[sudo] password for hadoop:
hadoop@nodo1:~/Downloads$ ls /opt
hadoop
hadoop@nodo1:~/Downloads$ ls -l /opt
total 4
drwxr-xr-x 10 hadoop 1024 4096 jul 29 13:44 hadoop
hadoop@nodo1:~/Downloads$
```

3. INSTALACION HADOOP

Paso 7. Examinando las recomendaciones, para ir sobreseguro instalaremos la versión 8 jdk de java

Hadoop Java Versions

Creado por Akira Ajisaka, modificado por última vez en oct 19, 2020

```
ads$ java -version
d, but can be installed with:
k-11-jre-headless  # version 11.0.17+8-1ubuntu2~22.04,

t-jre              # version 2:1.11-72build2
k-17-jre-headless  # version 17.0.5+8-2ubuntu1~22.04
k-18-jre-headless  # version 18.0.2+9-2~22.04
k-19-jre-headless  # version 19.0.1+10-1ubuntu1~22.04
k-8-jre-headless   # version 8u352-ga-1~22.04
loads$ sudo apt install openjdk-8-jdk-headless
. Done
e... Done
on... Done
l packages will be installed:
java-common openjdk-8-jre-headless

-demo openjdk-8-source fonts-dejavu-extra
fonts-ipafont-mincho fonts-wqy-microhei
s-indic
ges will be installed:
java-common openjdk-8-jdk-headless
s
talled, 0 to remove and 96 not upgraded.
archives.
44 MB of additional disk space will be used.
? [S/n]
```

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:~/Downloads$ java -version
sudo apt install openjdk-11-jre-headless  # version 11.0.17+8-1ubuntu2~22.04
hadoop@nodo1:~/Downloads$ java -version
Command 'java' not found, but can be installed with:
sudo apt install default-jre              # version 2:1.11-72build2, or
sudo apt install openjdk-11-jre-headless  # version 11.0.17+8-1ubuntu2~22.04
sudo apt install openjdk-17-jre-headless  # version 17.0.5+8-2ubuntu1~22.04
sudo apt install openjdk-18-jre-headless  # version 18.0.2+9-2~22.04
127 hadoop@nodo1:~/Downloads$ sudo apt install openjdk-8-jre
[sudo] password for hadoop:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java
  libatk-wrapper-java-jni openjdk-8-jre-headless
Suggested packages:
  default-jre fonts-ipafont-gothic fonts-ipafont-mincho fonts-wqy-microhei
  fonts-wqy-zenhei fonts-indic
The following NEW packages will be installed:
  ca-certificates-java fonts-dejavu-extra java-common libatk-wrapper-java
Need to get 33,0 MB of archives.
After this operation, 112 MB of additional disk space will be used.
Do you want to continue? [S/n] s
```

```
127 hadoop@nodo1:~/Downloads$ java -version
openjdk version "1.8.0_352"
OpenJDK Runtime Environment (build 1.8.0_352-8u352-ga-1~22.04-b08)
OpenJDK 64-Bit Server VM (build 25.352-b08, mixed mode)
hadoop@nodo1:~/Downloads$
```

3. INSTALACION HADOOP

Paso 8. Si vamos al directorio hadoop podemos ver que tenemos una serie de directorios.

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:/opt/hadoop$ pwd
/opt/hadoop
hadoop@nodo1:/opt/hadoop$ ls -l
total 116
drwxr-xr-x 2 1024 1024  4096 jul 29 13:44 bin
drwxr-xr-x 3 1024 1024  4096 jul 29 12:35 etc
drwxr-xr-x 2 1024 1024  4096 jul 29 13:44 include
drwxr-xr-x 3 1024 1024  4096 jul 29 13:44 lib
drwxr-xr-x 4 1024 1024  4096 jul 29 13:44 libexec
-rw-rw-r-- 1 1024 1024 24707 jul 28 20:30 LICENSE-binary
drwxr-xr-x 2 1024 1024  4096 jul 29 13:44 licenses-binary
-rw-rw-r-- 1 1024 1024 15217 jul 16 18:20 LICENSE.txt
-rw-rw-r-- 1 1024 1024 29473 jul 16 18:20 NOTICE-binary
-rw-rw-r-- 1 1024 1024  1541 abr 22  2022 NOTICE.txt
-rw-rw-r-- 1 1024 1024   175 abr 22  2022 README.txt
drwxr-xr-x 3 1024 1024  4096 jul 29 12:35 sbin
drwxr-xr-x 4 1024 1024  4096 jul 29 14:21 share
hadoop@nodo1:/opt/hadoop$
```

3. INSTALACION HADOOP

bin: Contiene básicamente una serie de scripts y de comandos que nos van a permitir lanzar y trabajar con los procesos hadoop: hadoop (gestión de hadoop), hdfs (para la parte de los datos), mapred y yarn (para la parte de los procesos).

etc: contiene los ficheros de configuración de hadoop. Ficheros XML Properties etc que son los que vamos a ir modificando

lib: contiene librerías nativas para hacer la compilación más rápida.

libexec: contiene una serie de ficheros de configuración extra

sbin: contiene scripts binarios de ayuda que permiten arrancar, parar tareas hadoop (arrancar y parar HFS, arrancar y parar yarn)

share: Contiene toda la paquetería de Hadoop: librerías, ejemplos hay distintos componentes y scripts para hacer pruebas.

3. INSTALACION HADOOP

Paso 9. Añadimos la variables de entorno PATH i JAVA_HOME en el fichero .bashrc del usuario hadoop:

```
File  Edit  View  Search  Terminal  Help
/home/hadoop
hadoop@nodo1:~$ ls -al
total 92
drwxr-x--- 15 hadoop hadoop 4096 ene  2 09:26 .
drwxr-xr-x  3 root   root   4096 ene  1 19:38 ..
-rw-------  1 hadoop hadoop  139 ene  1 20:27 .bash_history
-rw-r--r--  1 hadoop hadoop  220 ene  6  2022 .bash_logout
-rw-r--r--  1 hadoop hadoop 3868 ene  2 09:09 .bashrc
drwx------ 10 hadoop hadoop 4096 ene  2 14:11 .cache
drwx------ 14 hadoop hadoop 4096 ene  2 10:07 .config
drwxrwxr-x  2 hadoop hadoop 4096 ene  1 20:33 Desktop
-rw-r--r--  1 hadoop hadoop   23 ene  1 20:33 .dmrc
drwxr-xr-x  2 hadoop hadoop 4096 ene  1 20:33 Documents
drwxr-xr-x  2 hadoop hadoop 4096 ene  2 12:41 Downloads
drwxrwxr-x  3 hadoop hadoop 4096 ene  1 20:33 .local
drwxr-xr-x  2 hadoop hadoop 4096 ene  1 20:33 Music
drwxr-xr-x  2 hadoop hadoop 4096 ene  1 20:33 Pictures
-rw-r--r--  1 hadoop hadoop  807 ene  6  2022 .profile
drwxr-xr-x  2 hadoop hadoop 4096 ene  1 20:33 Public
drwx------  3 hadoop hadoop 4096 ene  2 09:26 snap
drwx------  2 hadoop hadoop 4096 ene  1 19:38 .ssh
-rw-r--r--  1 hadoop hadoop    0 ene  1 20:15 .sudo_as_admin_successful
drwxr-xr-x  2 hadoop hadoop 4096 ene  1 20:33 Templates
drwxr-xr-x  2 hadoop hadoop 4096 ene  1 20:33 Videos
-rw-------  1 hadoop hadoop   50 ene  2 09:09 .Xauthority
-rw-------  1 hadoop hadoop  756 ene  2 09:09 .xsession-errors
-rw-------  1 hadoop hadoop  756 ene  1 20:33 .xsession-errors.old
hadoop@nodo1:~$
```

3. INSTALACION HADOOP

Paso 10. Actualizamos la variables de entorno PATH, de manera que estén en el path los directorios bin y sbin de hadoop. Hacemos **nano .bashrc**

```
nodo1 [Corriendo] - Oracle VM VirtualBox
Archivo  Máquina  Ver  Entrada  Dispositivos  Ayuda
hadoop@nodo1 (10.0.2.15) - byobu
File  Edit  View  Search  Terminal  Help
  GNU nano 6.2                          .bashrc *
# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi
[ -r /home/hadoop/.config/byobu/prompt ] && . /home/hadoop/.config/byobu/prompt

export HADOOP_HOME=/opt/hadoop
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin

^G Help      ^O Write Out  ^W Where Is   ^K Cut       ^T Execute    ^C Location
^X Exit      ^R Read File  ^\ Replace    ^U Paste     ^J Justify    ^/ Go To Li
u  22.04 0:-*                55! 8h18m 0.72 2x2.5GHz 3.8G25% 24G48% 2023-01-02 18
```

3. INSTALACION HADOOP

Paso 11. Recargamos las variables de entorno de bashrc y al ejecutar hadoop versión vemos que ahora la variable JAVA_HOME no esta definida. Buscamos mediante el comando update-alternatives –list java la ubicación del jdk de java

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:~$ hadoop version
ERROR: JAVA_HOME is not set and could not be found.
1 hadoop@nodo1:~$ nano .bashrc
hadoop@nodo1:~$ . ./bashrc
bash: ./bashrc: No such file or directory
1 hadoop@nodo1:~$ nano .bashrc
hadoop@nodo1:~$ . ./.bashrc
hadoop@nodo1:~$ hadoop version
ERROR: JAVA_HOME is not set and could not be found.
1 hadoop@nodo1:~$ update-alternatives --list java
/usr/lib/jvm/java-8-openjdk-amd64/jre/bin/java
hadoop@nodo1:~$
```

3. INSTALACION HADOOP

Paso 12. Agregamos el path anterior a la variable JAVA_HOME en .bashrc. Recargamos y vemos que ahora si coge

```
export HADOOP_HOME=/opt/hadoop
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:$HADOOP_HOME/bin:$HADOOP_HOME/sbin
```

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:~$ nano .bashrc
hadoop@nodo1:~$ . ./.bashrc
hadoop@nodo1:~$ hadoop version
Hadoop 3.3.4
Source code repository https://github.com/apache/hadoop.git -r a585a73c3e02ac62350c1366
43a5e7f6095a3dbb
Compiled by stevel on 2022-07-29T12:32Z
Compiled with protoc 3.7.1
From source with checksum fb9dd8918a7b8a5b430d61af858f6ec
This command was run using /opt/hadoop/share/hadoop/common/hadoop-common-3.3.4.jar
hadoop@nodo1:~$
```

```
hadoop@nodo1:/$ hadoop version
Hadoop 2.10.2
Subversion Unknown -r 965fd380006fa78b2315668fbc7eb432e1d8200f
Compiled by ubuntu on 2022-05-24T22:35Z
Compiled with protoc 2.5.0
From source with checksum d3ab737f7788f05d467784f0a86573fe
This command was run using /opt/hadoop/share/hadoop/common/hadoop-common-2.10.2.
jar
hadoop@nodo1:/$
```

3. INSTALACION HADOOP

Paso 13. Para indicar la variable de entorno JAVA_HOME para todos los usuarios de Linux debemos indicar lo mismo en el fichero /etc/.profile → **sudo nano /etc/.profile**
Para cargar podemos salir y logarnos de nuevo o ejecutar **source /etc/.profile** para aplicar los cambios inmediatamente en nuestro actual shell

```
File  Edit  View  Search  Terminal  Help
  GNU nano 6.2                    /etc/.profile
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$JAVA_HOME/bin:$PATH
```

```
hadoop@nodo1:/opt/hadoop/sbin$ sudo nano /etc/.profile
[sudo] password for hadoop:
hadoop@nodo1:/opt/hadoop/sbin$ source /etc/.profile
hadoop@nodo1:/opt/hadoop/sbin$
```

4. COMPROBACION FUNCIONAMIENTO HADOOP

Paso 1. Ejecutaremos Hadoop en modo standalone: utilizaremos un pequeño paquete de ejemplos que viene dentro del propio Hadoop para ejecutar un programa MapReduce (es un proceso que permite ejecutar múltiples hilos de un determinado recurso).

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:~$ pwd
/home/hadoop
hadoop@nodo1:~$ cd /opt/hadoop/
hadoop@nodo1:/opt/hadoop$ ls -l
total 116
drwxr-xr-x 2 1024 1024  4096 jul 29 13:44 bin
drwxr-xr-x 3 1024 1024  4096 jul 29 12:35 etc
drwxr-xr-x 2 1024 1024  4096 jul 29 13:44 include
drwxr-xr-x 3 1024 1024  4096 jul 29 13:44 lib
drwxr-xr-x 4 1024 1024  4096 jul 29 13:44 libexec
-rw-rw-r-- 1 1024 1024 24707 jul 28 20:30 LICENSE-binary
drwxr-xr-x 2 1024 1024  4096 jul 29 13:44 licenses-binary
-rw-rw-r-- 1 1024 1024 15217 jul 16 18:20 LICENSE.txt
-rw-rw-r-- 1 1024 1024 29473 jul 16 18:20 NOTICE-binary
-rw-rw-r-- 1 1024 1024  1541 abr 22  2022 NOTICE.txt
-rw-rw-r-- 1 1024 1024   175 abr 22  2022 README.txt
drwxr-xr-x 3 1024 1024  4096 jul 29 12:35 sbin
drwxr-xr-x 4 1024 1024  4096 jul 29 14:21 share
hadoop@nodo1:/opt/hadoop$ cd share
hadoop@nodo1:/opt/hadoop/share$ cd hadoop/
hadoop@nodo1:/opt/hadoop/share/hadoop$ cd mapreduce/
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$
```

4. COMPROBACION FUNCIONAMIENTO HADOOP

Paso 2. Aquí están los ejemplos que vamos a más utilizar. Tenemos una serie de librerías entre la que destaca hadoop-mapreduce-examples, que es la que vamos a utilizar para ver un ejemplo de cómo funciona mapreduce

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ pwd
/opt/hadoop/share/hadoop/mapreduce
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ ls -l
total 5304
-rw-r--r-- 1 1024 1024  590752 jul 29 13:22 hadoop-mapreduce-client-app-3.3.4.jar
-rw-r--r-- 1 1024 1024  805750 jul 29 13:22 hadoop-mapreduce-client-common-3.3.4.jar
-rw-r--r-- 1 1024 1024 1636329 jul 29 13:22 hadoop-mapreduce-client-core-3.3.4.jar
-rw-r--r-- 1 1024 1024  181707 jul 29 13:22 hadoop-mapreduce-client-hs-3.3.4.jar
-rw-r--r-- 1 1024 1024    9966 jul 29 13:22 hadoop-mapreduce-client-hs-plugins-3.3.4.ja
r
-rw-r--r-- 1 1024 1024   49783 jul 29 13:22 hadoop-mapreduce-client-jobclient-3.3.4.jar
-rw-r--r-- 1 1024 1024 1658927 jul 29 13:22 hadoop-mapreduce-client-jobclient-3.3.4-tes
ts.jar
-rw-r--r-- 1 1024 1024   90704 jul 29 13:22 hadoop-mapreduce-client-nativetask-3.3.4.ja
r
-rw-r--r-- 1 1024 1024   62093 jul 29 13:22 hadoop-mapreduce-client-shuffle-3.3.4.jar
-rw-r--r-- 1 1024 1024   22263 jul 29 13:22 hadoop-mapreduce-client-uploader-3.3.4.jar
-rw-r--r-- 1 1024 1024  280990 jul 29 13:22 hadoop-mapreduce-examples-3.3.4.jar
drwxr-xr-x 2 1024 1024    4096 jul 29 13:44 jdiff
drwxr-xr-x 2 1024 1024    4096 jul 29 13:44 lib-examples
drwxr-xr-x 2 1024 1024    4096 jul 29 13:44 sources
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$
```

4. COMPROBACION FUNCIONAMIENTO HADOOP

Paso 3. Para ver el contenido que tiene hadoop-mapreduce-examples, debemos instalar el comando jar

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ jar tf hadoop-mapreduce-examples-3.3.4
.jar
Command 'jar' not found, but can be installed with:
sudo apt install default-jdk              # version 2:1.11-72build2, or
sudo apt install openjdk-11-jdk-headless  # version 11.0.17+8-1ubuntu2~22.04
sudo apt install fastjar                  # version 2:0.98-7
sudo apt install openjdk-17-jdk-headless  # version 17.0.5+8-2ubuntu1~22.04
sudo apt install openjdk-18-jdk-headless  # version 18.0.2+9-2~22.04
sudo apt install openjdk-19-jdk-headless  # version 19.0.1+10-1ubuntu1~22.04
sudo apt install openjdk-8-jdk-headless   # version 8u352-ga-1~22.04
127 hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ sudo apt install openjdk-8-jdk-hea
dless
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
Suggested packages:
  openjdk-8-demo openjdk-8-source
The following NEW packages will be installed:
  openjdk-8-jdk-headless
```

4. COMPROBACION FUNCIONAMIENTO HADOOP

Paso 4. Al mostrar el contenido de hadoop-mapreduce-examples. nos sale una lista de determinados comandos o opciones que hay dentro de este paquete. Utilizaremos el comando grep que permite buscar cadenas dentro de los ficheros.

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ jar tf hadoop-mapreduce-examples-3.3.4
.jar
org/apache/hadoop/examples/terasort/
org/apache/hadoop/examples/pi/
org/apache/hadoop/examples/pi/math/
org/apache/hadoop/examples/terasort/TeraChecksum$ChecksumMapper.class
org/apache/hadoop/examples/terasort/TeraScheduler$Host.class
org/apache/hadoop/examples/terasort/TeraSort$SimplePartitioner.class
org/apache/hadoop/examples/terasort/TeraSortConfigKeys.class
org/apache/hadoop/examples/terasort/TeraChecksum$ChecksumReducer.class
org/apache/hadoop/examples/terasort/GenSort.class
org/apache/hadoop/examples/terasort/TeraGen$RangeInputFormat.class
org/apache/hadoop/examples/terasort/Unsigned16.class
org/apache/hadoop/examples/terasort/TeraGen$SortGenMapper.class
org/apache/hadoop/examples/terasort/TeraValidate$ValidateMapper.class
org/apache/hadoop/examples/terasort/TeraGen$RangeInputFormat$RangeRecordReader.class
org/apache/hadoop/examples/terasort/TeraSort$TotalOrderPartitioner$LeafTrieNode.class
org/apache/hadoop/examples/terasort/TeraSort$TotalOrderPartitioner$TrieNode.class
org/apache/hadoop/examples/terasort/TeraSort.class
```

```
org/apache/hadoop/examples/dancing/Sudoku$RowConstraint.class
org/apache/hadoop/examples/dancing/DistributedPentomino$PentMap.class
org/apache/hadoop/examples/BaileyBorweinPlouffe$BbpReducer$1.class
org/apache/hadoop/examples/Grep.class
org/apache/hadoop/examples/AggregateWordCount.class
org/apache/hadoop/examples/RandomTextWriter.class
```

4. COMPROBACION FUNCIONAMIENTO HADOOP

Paso 5. Prepararemos el entorno para la prueba. Crearemos el directorio /tmp/entrada y aquí copiaremos los ficheros XML de configuración de hadoop/etc con el objetivo de tener datos para buscar algo. Buscaremos entre estos ficheros todas las palabras que empiecen por la cadena "kms".

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ mkdir /tmp/entrada
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ cp /opt/hadoop/etc/hadoop/*.xml /tmp/e
ntrada
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$ ls /tmp/entrada
capacity-scheduler.xml  hdfs-rbf-site.xml  kms-acls.xml     yarn-site.xml
core-site.xml           hdfs-site.xml      kms-site.xml
hadoop-policy.xml       httpfs-site.xml    mapred-site.xml
hadoop@nodo1:/opt/hadoop/share/hadoop/mapreduce$
```

4. COMPROBACION FUNCIONAMIENTO HADOOP

Paso 6. Con el siguiente comando hadoop buscaremos los ficheros que contengan palabras que empiecen por kms_

hadoop jar hadoop-mapreduce-examples-3.3.4.jar grep /tmp/entrada /tmp/salida 'kms[a-z.]+'

Nos encontramos que en la versión descargada 3.3.4, este comando falla. No genera la carpeta de salida /tmp/salida

```
        at org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.jav
a:327)
        at org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitt
er.java:200)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1571)
        at org.apache.hadoop.mapreduce.Job$11.run(Job.java:1568)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInforma
tion.java:1878)
        at org.apache.hadoop.mapreduce.Job.submit(Job.java:1568)
        at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1589)
        at org.apache.hadoop.examples.Grep.run(Grep.java:94)
        at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:81)
        at org.apache.hadoop.examples.Grep.main(Grep.java:103)
        at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
        at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.
Caused by: java.io.IOException: Input path does not exist: file:/opt/hadoop/shar
e/hadoop/mapreduce/grep-temp-1024778084
        at org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedL
istStatus(FileInputFormat.java:313)
        ... 29 more
255 hadoop@nodo11:/opt/hadoop/share/hadoop/mapreduce$ ls /tmp
```

4. COMPROBACION FUNCIONAMIENTO HADOOP

Paso 7. En el resto de versiones actualmente descargables (3.2.4 y 2.7.2), se genera la carpeta /tmp/salida correctamente

En el directorio /tmp/salida, si todo es correcto tendré dos ficheros:

- _SUCCESS → indica que ha sido correcta la consulta
- part-r-0000 → Indica el numero de veces que se repite el texto buscado

```
hadoop@nodo11:/opt/hadoop/share/hadoop/mapreduce$ ls /tmp/salida
part-r-00000  _SUCCESS
hadoop@nodo11:/opt/hadoop/share/hadoop/mapreduce$ ls -l /tmp/salida/
total 4
-rw-r--r-- 1 hadoop hadoop 11 feb  9 23:58 part-r-00000
-rw-r--r-- 1 hadoop hadoop  0 feb  9 23:58 _SUCCESS
hadoop@nodo11:/opt/hadoop/share/hadoop/mapreduce$ cat /tmp/salida/part-r-00000
9       kms.acl.
hadoop@nodo11:/opt/hadoop/share/hadoop/mapreduce$ hadoop jar hadoop-mapreduce-examples-3.2.4.jar
grep /tmp/entrada /tmp/salida 'kms[a-z.]+'
```

5. INSTALACIÓN SSH

Paso 1. El comando ssh-keygen crea las claves publica y privada. Genera en el subdirectorio .ssh dos ficheros con estas claves

```
File  Edit  View  Search  Terminal  Help
hadoop@nodo1:~$ pwd
/home/hadoop
hadoop@nodo1:~$ ssh-keygen
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoop/.ssh/id_rsa):
/home/hadoop/.ssh/id_rsa already exists.
Overwrite (y/n)? y
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:Q35NeoVlcJC3pKQXXl2jl64Up/rvrHurkdStkRevrnQ hadoop@nodo1
The key's randomart image is:
+---[RSA 3072]----+
|            o++oo|
|            +== +|
|         .  +oO.* |
|        o  .++.O.+|
|         S o.o+oo+|
|          o .+ o= |
|            ..+E  |
|            ..oo. |
|             .*O=.|
+----[SHA256]-----+
hadoop@nodo1:~$
```

```
hadoop@nodo1:~$ ls -l .ssh
total 8
-rw------- 1 hadoop hadoop    0 ene  1 19:38 authorized_keys
-rw------- 1 hadoop hadoop 2602 ene  4 13:12 id_rsa
-rw-r--r-- 1 hadoop hadoop  566 ene  4 13:12 id_rsa.pub
hadoop@nodo1:~$
```

5. INSTALACIÓN SSH

Paso 2. Al final pone hadoop nodo 1 dentro de la clave publica

```
hadoop@nodo1:~$ cat .ssh/id_rsa.pub
ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABgQDO8RRSrbN4WXBSIpG4VN1M6fjdlR6H2wLp80utW/VU
2fy0SQV9o+1j1VNw2zbLJ0X+aFKtAkHgyA+DWT4XgUqHuLbeTNETAu5OjN/B4LBbOSHjuUWQvnCJkgXy
SBzabd1AKa5GgivpKiGh/7f4QwZBj2C7QIaQpeuBTK+nbvDoLxfO4h9eZvfV7NyaKBAbXQP7jrnVdAkY
Z0l6nhC3A73X6VHERQLgTROQ7b3CQwxN0r3gB29+JIVteykSsGUi/0YkoMl34FgEhjuoFtk9+6G7Qf9b
54oBO71cxRyo3OR+CXKWXFv5VSkKi3PZTvdGP3+1T/5mvYGHfSKvbwteVqXe3o2Q46UY/T7wNND9et67
LSDRaAfwoh9Gu6rNV8kJcntJGH8PH7Vu0pQC8DJHg+kuCLf+UVcmJCeSAmgXQ/dVwIHB7oP2QM+7iM3i
Mq5HQfb/CyxSac7JIrnpViMVsulL0P8vcrQ3WZvHEzJxkFrMgJqCOXkU64somgHsn/bkWhU= hadoop@
nodo1
hadoop@nodo1:~$
```

```
hadoop@nodo1:~$ cd .ssh
hadoop@nodo1:~/.ssh$ ls
authorized_keys  id_rsa  id_rsa.pub
hadoop@nodo1:~/.ssh$ cp id_rsa.pub authorized_keys
hadoop@nodo1:~/.ssh$ ls
authorized_keys  id_rsa  id_rsa.pub
hadoop@nodo1:~/.ssh$
```

5. INSTALACIÓN SSH

Paso 3. Nos conectamos a la misma maquina con ssh nodo1 (simulamos la conexión desde fuera). Nos pide si vamos a configurar esta maquina como servidor autorizado. No pide contraseña para entrar,. Me puedo conectar a un nodo remoto sin tener que entra las credenciales
Con exit vamos al nodo original

```
hadoop@nodo1:~/.ssh$ ssh nodo1
The authenticity of host 'nodo1 (127.0.1.1)' can't be established.
ED25519 key fingerprint is SHA256:hVU04B988kL7av27mS8wa6TURtu3DKrXPsOwhAowUD8.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'nodo1' (ED25519) to the list of known hosts.
Welcome to Ubuntu 22.04.1 LTS (GNU/Linux 5.15.0-56-generic x86_64)

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

  System information as of mié 04 ene 2023 13:29:39 UTC

  System load:  0.3896484375       Processes:               185
  Usage of /:   50.7% of 23.45GB   Users logged in:         1
  Memory usage: 12%                IPv4 address for enp0s3: 10.0.2.15
  Swap usage:   0%


56 updates can be applied immediately.
To see these additional updates run: apt list --upgradable


Last login: Sun Jan  1 19:48:55 2023
hadoop@nodo1:~$
```

```
hadoop@nodo1:~$ exit
logout
Connection to nodo1 closed.
hadoop@nodo1:~/.ssh$
```