

# Recomendaciones para desarrollar las prácticas 1-3:

el curso cuenta con una máquina virtual preparada sobre la distribución Ubuntu 20.04, lo cual puede diferir de las capturas mostradas en la plataforma.

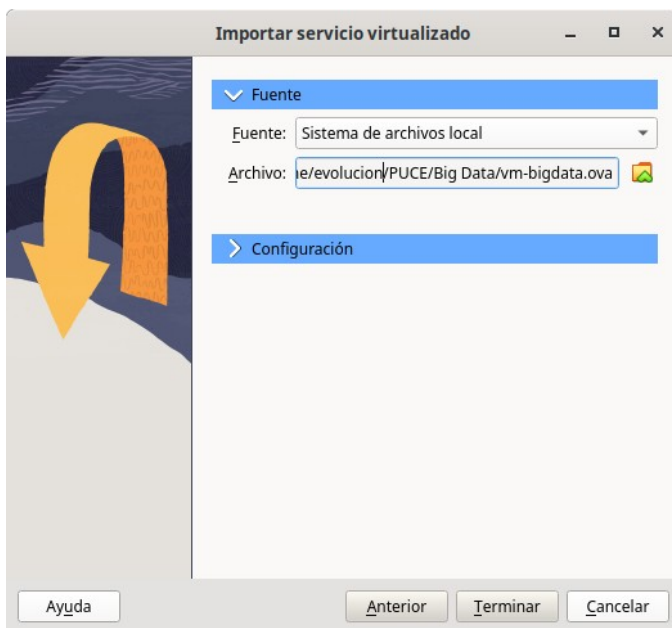
Usuario: bigdata  
Password: bigdata

Para interactuar de mejor manera con esta máquina puede instalar Guest Additions

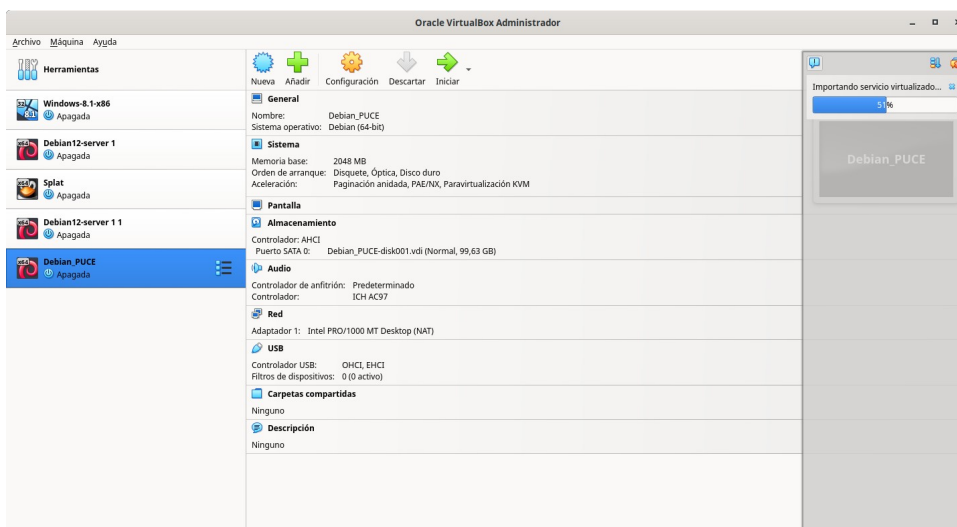
## 0. Importar la máquina virtual

Importe la máquina virtual desde :

Archivo → Importar Servicio Virtualizado



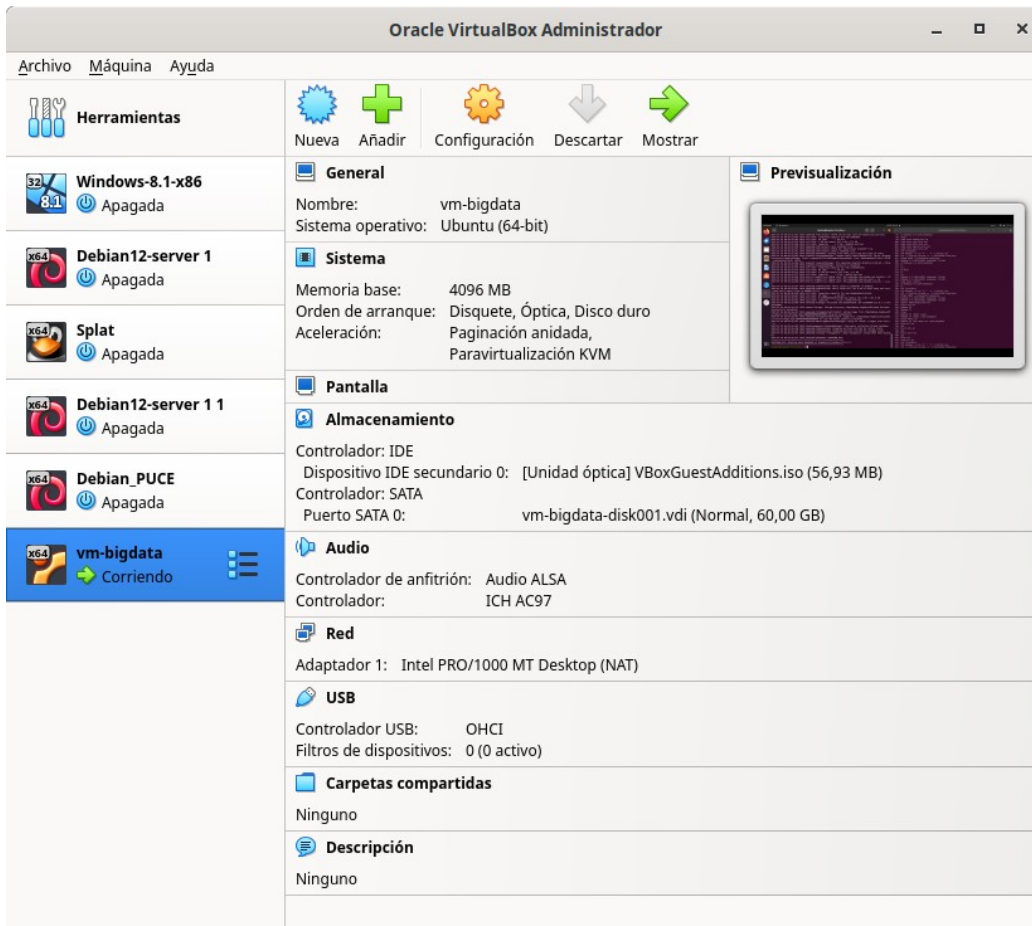
Espere a que termine el proceso



Una vez que termine el proceso debe verla listada dentro del menú de sus máquinas:

Arránquela y verifique que funcione:

Usuario: bigdata  
Password: bigdata



## 1. Guest Additions

Si desea tener una mejor resolución de pantalla de su máquina virtual habilite e instale Guest Additions.

Para ello en su máquina virtual vaya a:

Dispositivos → Insertar Imagen de CD de los complementos de invitado...

Al ser una distribución de Ubuntu, le preguntará si desea autoejecutar el volumen insertado.

Conteste que si.

La contraseña es: bigdata

Una vez que termine todo el proceso se presentará la siguiente ventana:

```
VirtualBox Guest Additions Installation
Verifying archive integrity... 100% MDS checksums are OK. All good.
Uncompressing VirtualBox 7.1.4 Guest Additions for Linux 100%
VirtualBox Guest Additions installer
VirtualBox Guest Additions: Starting.
VirtualBox Guest Additions: Setting up modules
VirtualBox Guest Additions: Building the VirtualBox Guest Additions kernel
modules. This may take a while.
VirtualBox Guest Additions: To build modules for other installed kernels, run
VirtualBox Guest Additions: /sbin/rcvboxadd quicksetup <version>
VirtualBox Guest Additions: or
VirtualBox Guest Additions: /sbin/rcvboxadd quicksetup all
VirtualBox Guest Additions: Building the modules for kernel 5.15.0-139-generic.
update-initramfs: Generating /boot/initrd.img-5.15.0-139-generic
VirtualBox Guest Additions: Running kernel modules will not be replaced until
the system is restarted or 'rcvboxadd reload' triggered
VirtualBox Guest Additions: reloading kernel modules and services
VirtualBox Guest Additions: kernel modules were not reloaded
VirtualBox Guest Additions: kernel modules and services were not reloaded
The log file /var/log/vboxadd-setup.log may contain further information.
Press Return to close this window...
```

Cuando finalice. Cierre la pantalla.

Al maximizar su ventana, la máquina virtual adoptará la resolución del espacio que se despliegue.

## 2. Actualizaciones.

Dado que Hadoop es un ecosistema compuesto por muchas librerías y dependencias, NO se recomienda actualizar el sistema operativo, ya que corre el riesgo de romperlo.

## 3. Formatear el namenode

Cuando use la máquina por primera vez DEBE formatear el namenode. Para ello ejecute:

```
bigdata@bigdata-VirtualBox:~$ hadoop-3.4.1/bin/hdfs namenode -format
```

Como salida verá algo similar a:

```
bigdata@bigdata-VirtualBox:~$ cd hadoop-3.4.1/
bin/          include/          libexec/          logs/           share/
etc/          lib/             licenses-binary/ sbin/
bigdata@bigdata-VirtualBox:~$ hadoop-3.4.1/bin/h
hadoop      hadoop.cmd  hdfs      hdfs.cmd
bigdata@bigdata-VirtualBox:~$ hadoop-3.4.1/bin/hdfs namenode -format
2025-07-29 00:59:14,613 INFO namenode.NameNode: STARTUP_MSG:
/*****
STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = bigdata-VirtualBox/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 3.4.1
STARTUP_MSG: classpath = /home/bigdata/hadoop-3.4.1/etc/hadoop:/home/bigdata/hadoop-3.4.1/share/h
adoop/common/lib/netty-codec-stomp-4.1.100.Final.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/
lib/kerby-pkix-2.0.3.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/slf4j-api-1.7.36.jar:/ho
me/bigdata/hadoop-3.4.1/share/hadoop/common/lib/jetty-http-9.4.53.v20231009.jar:/home/bigdata/hadoop
-3.4.1/share/hadoop/common/lib/jetty-util-9.4.53.v20231009.jar:/home/bigdata/hadoop-3.4.1/share/had
oop/common/lib/netty-handler-4.1.100.Final.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/cu
rator-framework-5.2.0.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/jackson-databind-2.12.7
.1.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/netty-transport-native-kqueue-4.1.100.Fina
l-osx-aarch_64.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/reload4j-1.2.22.jar:/home/bigd
ata/hadoop-3.4.1/share/hadoop/common/lib/jersey-servlet-1.19.4.jar:/home/bigdata/hadoop-3.4.1/share
/hadoop/common/lib/jetty-server-9.4.53.v20231009.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/
lib/kerby-config-2.0.3.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/netty-buffer-4.1.100.F
inal.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/kerb-util-2.0.3.jar:/home/bigdata/hadoop
-3.4.1/share/hadoop/common/lib/commons-collections-3.2.2.jar:/home/bigdata/hadoop-3.4.1/share/hadoo
n/common/lib/son-2.9.0.jar:/home/bigdata/hadoop-3.4.1/share/hadoop/common/lib/metrics-core-3.2.4.jar
```

.... mas líneas de terminal

Y terminará con:

```

2025-07-29 00:59:16,135 INFO common.Storage: Storage directory /tmp/hadoop-bigdata/dfs/name has been
successfully formatted.
2025-07-29 00:59:16,245 INFO namenode.FSImageFormatProtobuf: Saving image file /tmp/hadoop-bigdata/df
s/name/current/fsimage.ckpt_000000000000000000 using no compression
2025-07-29 00:59:16,367 INFO namenode.FSImageFormatProtobuf: Image file /tmp/hadoop-bigdata/dfs/name/
current/fsimage.ckpt_000000000000000000 of size 402 bytes saved in 0 seconds .
2025-07-29 00:59:16,391 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with txid >
= 0
2025-07-29 00:59:16,403 INFO blockmanagement.DatanodeManager: Slow peers collection thread shutdown
2025-07-29 00:59:16,448 INFO namenode.FSNamesystem: Stopping services started for active state
2025-07-29 00:59:16,449 INFO namenode.FSNamesystem: Stopping services started for standby state
2025-07-29 00:59:16,457 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdo
wn.
2025-07-29 00:59:16,457 INFO namenode.NameNode: SHUTDOWN_MSG:
/*****
SHUTDOWN_MSG: Shutting down NameNode at bigdata-VirtualBox/127.0.1.1
*****/
bigdata@bigdata-VirtualBox:~$

```

## 4. Iniciar los servicios de hdfs y yarn

Para poder desarrollar las prácticas, deben estar corriendo los servicios:

- hdfs
- yarn

Para ello puede ejecutar:

```

bigdata@bigdata-VirtualBox:~$ hadoop-3.4.1/sbin/start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as bigdata in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bigdata-VirtualBox]
Starting resourcemanager
Starting nodemanagers
bigdata@bigdata-VirtualBox:~$

```

## 5. Interacción con hdfs (Plataforma: 3.1 copiar datos a HDFS)

Primero listaremos los archivos disponibles en el sistema de ficheros de hadoop:

```

bigdata@bigdata-VirtualBox:~$ hadoop fs -ls
ls: `.`: No such file or directory
bigdata@bigdata-VirtualBox:~$

```

En este caso vemos que no existe ningún archivo. Copiemos uno desde nuestro sistema operativo Ubuntu al sistema de ficheros de hadoop. Para ello primero listemos los archivos en nuestra máquina:

```

bigdata@bigdata-VirtualBox:~$ ls -l
total 56
-rw-rw-r-- 1 bigdata bigdata 38 mar 4 15:25 datos1.txt
-rw-r--r-- 1 bigdata bigdata 38 mar 4 15:30 datos2.txt
-rw-rw-r-- 1 bigdata bigdata 0 mar 3 23:59 demo.csv
drwxr-xr-x 2 bigdata bigdata 4096 jul 9 22:46 Desktop
drwxr-xr-x 2 bigdata bigdata 4096 mar 4 17:28 Documents
drwxr-xr-x 2 bigdata bigdata 4096 mar 4 17:20 Downloads
drwxr-xr-x 11 bigdata bigdata 4096 mar 3 23:50 hadoop-3.4.1
drwxr-xr-x 2 bigdata bigdata 4096 mar 3 17:16 Music
-rw----- 1 bigdata bigdata 1 mar 3 22:20 nano.save
drwxr-xr-x 2 bigdata bigdata 4096 mar 3 17:16 Pictures
drwxr-xr-x 2 bigdata bigdata 4096 mar 3 17:16 Public
drwx----- 3 bigdata bigdata 4096 mar 3 17:25 snap
drwxr-xr-x 2 bigdata bigdata 4096 mar 3 17:16 Templates
drwxr-xr-x 2 bigdata bigdata 4096 mar 3 17:16 Videos
-rw-rw-r-- 1 bigdata bigdata 2513 mar 4 17:24 words-laodisea.txt
bigdata@bigdata-VirtualBox:~$

```

Vemos en este caso que existe el archivo words-laodisea.txt

Veamos su contenido con el comando cat:

```
bigdata@bigdata-VirtualBox:~$ cat words-laodisea.txt
La Odisea (en griego antiguo: Ὀδύσσεια, [Odýsseia]; en griego moderno: Οδύσσεια; en latín: Odyssea) es un poema épico griego compuesto por 24 cantos, atribuido al poeta griego Homero. Se cree que fue compuesta en el siglo VIII a. C. en los asentamientos que tenía Grecia en la costa oeste del Asia Menor (actual Turquía asiática). Según otros autores, la Odisea se completa en el siglo VII a. C. a partir de poemas que solo describían partes de la obra actual. Fue originalmente escrita en lo que se ha llamado dialecto homérico. Narra la vuelta a casa, tras la guerra de Troya, del héroe griego Odiseo (al modo latino, Ulises: Ὀδυσσεύς en griego; Vlixes en latín). Además de haber estado diez años fuera luchando, Odiseo tarda otros diez años en regresar a la isla de Ítaca, de la que era rey, período durante el cual su hijo Telémaco y su esposa Penélope han de tolerar en su palacio a los pretendientes que buscan desposarla (pues ya creían muerto a Odiseo), al mismo tiempo que consumen los bienes de la familia.

La mejor arma de Odiseo es su mētis o astucia.1 Gracias a su inteligencia – además de la ayuda provista por Palas Atenea, hija de Zeus Crónida – es capaz de escapar de los continuos problemas a los que ha de enfrentarse por designio de los dioses. Para esto, planea diversas artimañas, bien sean físicas – como pueden ser disfraces – o con audaces y engañosos discursos de los que se vale para conseguir sus objetivos.

El poema es, junto a la Iliada, uno de los primeros textos de la épica grecolatina y por tanto de la literatura occidental. Se cree que el poema original fue transmitido por vía oral durante siglos por aedos que recitaban el poema de memoria, alterándolo consciente o inconscientemente. Era transmitido en dialectos de la Antigua Grecia. Ya en el siglo IX a. C., con la reciente aparición del alfabeto, tanto la Odisea como la Iliada pudieron ser las primeras obras en ser transcritas, aunque la mayoría de la crítica se inclina por datarlas en el siglo VIII a. C. El texto homérico más antiguo que conocemos es la versión de Aristarco de Samotracia (siglo II a. C.). El poema está escrito usando una métrica llamada hexámetro dactílico. Cada línea de la Odisea original estaba formada por seis unidades o pies, siendo cada pie dáctilo o espondeo.2 Los distintos primeros cinco pies eran dáctilos y el último podía ser un espondeo o bien un troqueo. Los distintos pies se separan por cesuras o pausas.
```

Carguemos este archivo al sistema de ficheros de hadoop:

Para ello primero crearemos una carpeta llamada bigdata:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -mkdir /bigdata
```

Luego cargaremos el archivo words-laodisea.txt a esa carpeta:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -put words-laodisea.txt /bigdata
```

Para confirmar que el archivo ha sido subido listemos los ficheros en esa carpeta:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -ls /bigdata
Found 1 items
-rw-r--r--  1 bigdata supergroup      2513 2025-07-29 01:42 /bigdata/words-laodisea.txt
bigdata@bigdata-VirtualBox:~$
```

Podemos ver que el archivo se lista correctamente.

## 6. Copia un archivo desde HDFS

También podemos copiar un archivo desde hadoop a nuestra máquina local. Para ello primero crearemos un archivo vacío en hadoop y luego lo bajaremos a nuestra máquina.

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -touch /bigdata/words.txt
```

Para verificar que se creó correctamente listamos los ficheros de /bigdata en hadoop:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -ls /bigdata
Found 2 items
-rw-r--r--  1 bigdata supergroup      2513 2025-07-29 01:42 /bigdata/words-laodisea.txt
-rw-r--r--  1 bigdata supergroup         0 2025-07-29 01:50 /bigdata/words.txt
bigdata@bigdata-VirtualBox:~$
```

Ahora tenemos dos ficheros.

Descarguémoslo a nuestra máquina:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -copyToLocal /bigdata/words.txt
```

listemos los ficheros locales:

```
bigdata@bigdata-VirtualBox:~$ ls -l
total 56
-rw-rw-r-- 1 bigdata bigdata 38 mar  4 15:25 datos1.txt
-rw-r--r-- 1 bigdata bigdata 38 mar  4 15:30 datos2.txt
-rw-rw-r-- 1 bigdata bigdata  0 mar  3 23:59 demo.csv
drwxr-xr-x 2 bigdata bigdata 4096 jul  9 22:46 Desktop
drwxr-xr-x 2 bigdata bigdata 4096 mar  4 17:28 Documents
drwxr-xr-x 2 bigdata bigdata 4096 mar  4 17:20 Downloads
drwxr-xr-x 11 bigdata bigdata 4096 mar  3 23:50 hadoop-3.4.1
drwxr-xr-x 2 bigdata bigdata 4096 mar  3 17:16 Music
-rw----- 1 bigdata bigdata  1 mar  3 22:20 nano.save
drwxr-xr-x 2 bigdata bigdata 4096 mar  3 17:16 Pictures
drwxr-xr-x 2 bigdata bigdata 4096 mar  3 17:16 Public
drwx----- 3 bigdata bigdata 4096 mar  3 17:25 snap
drwxr-xr-x 2 bigdata bigdata 4096 mar  3 17:16 Templates
drwxr-xr-x 2 bigdata bigdata 4096 mar  3 17:16 Videos
-rw-rw-r-- 1 bigdata bigdata 2513 mar  4 17:24 words-laodisea.txt
-rw-r--r-- 1 bigdata bigdata  0 jul 29 01:54 words.txt
bigdata@bigdata-VirtualBox:~$
```

Vemos al final el fichero words.txt que fue descargado desde hadoop.

## 7. Eliminar archivos de HDFS

Eliminemos el archivo words.txt que creamos en hadoop:

```
bigdata@bigdata-VirtualBox:~$ hadoop fs -rm /bigdata/words.txt
Deleted /bigdata/words.txt
bigdata@bigdata-VirtualBox:~$ hadoop fs -ls /bigdata/
Found 1 items
-rw-r--r-- 1 bigdata supergroup 2513 2025-07-29 01:42 /bigdata/words-laodisea.txt
bigdata@bigdata-VirtualBox:~$
```

Como vemos en el listado de archivos ya no aparece en el sistema de ficheros de hadoop.