

GRADO

# INTRODUCCIÓN AL MANEJO DE DATOS MASIVOS CON HADOOP Y HERRAMIENTAS INSPIRADAS EN SQL

PRÁCTICA DE LA ASIGNATURA SISTEMAS DE BASES DE DATOS.

PREPARACIÓN DEL ENTORNO DE TRABAJO.

2021-2022

Versión 1.0

Dr. Agustín C. Caminero Herráez — Dr. Luis Grau Fernández

GRADO EN INGENIERÍA INFORMÁTICA

# 1 Contenido

- 1. Descripción.....2
- 2. Máquina virtual .....2
- 3. Contenedor.....7

## 1. Descripción

Para la práctica de esta asignatura, en el presente curso académico se van a ofrecer diversos entornos de trabajo, de entre los que el/la estudiante podrá elegir el que considere conveniente para la realización de la práctica. Solamente es necesario realizar la práctica en uno de los entornos propuestos.

Las opciones disponibles son las siguientes:

- Máquina virtual VirtualBox.
- Contenedor Docker.

Ambas posibilidades tienen las mismas características, si bien el contenedor Docker tiene menos requerimientos sobre el ordenador en el que se ejecute, especialmente sobre la memoria RAM.

Ambas posibilidades están basadas en los entornos oficiales de Cloudera, que han sido extendidos para instalar en ellos algunas herramientas de interés. De esta forma, trabajaremos con una distribución Hadoop comercial y ampliamente utilizada en numerosas instalaciones en producción. En nuestro caso, utilizaremos Jupyter notebooks para realizar nuestros desarrollos, herramienta que se encuentra instalada en el entorno de desarrollo.

En este documento se detalla la preparación de estos entornos de trabajo. Es necesario elegir uno de los dos para realizar la práctica.

## 2. Máquina virtual

En primer lugar veremos la máquina virtual con tecnología VirtualBox, basada en la máquina virtual oficial de Cloudera. Esta máquina virtual la descargaremos del siguiente enlace:

[https://unedo365-my.sharepoint.com/:u:/g/personal/accaminero\\_scc\\_uned\\_es/EXo\\_EwhGGONPrjPFxEdoK34Byzlzy7mv\\_E0exjRPu\\_JAg?e=UrC8Mh](https://unedo365-my.sharepoint.com/:u:/g/personal/accaminero_scc_uned_es/EXo_EwhGGONPrjPFxEdoK34Byzlzy7mv_E0exjRPu_JAg?e=UrC8Mh)

En dicho enlace descargaremos el fichero con extensión .ova que contiene la máquina virtual con las herramientas necesarias para realizar la práctica.

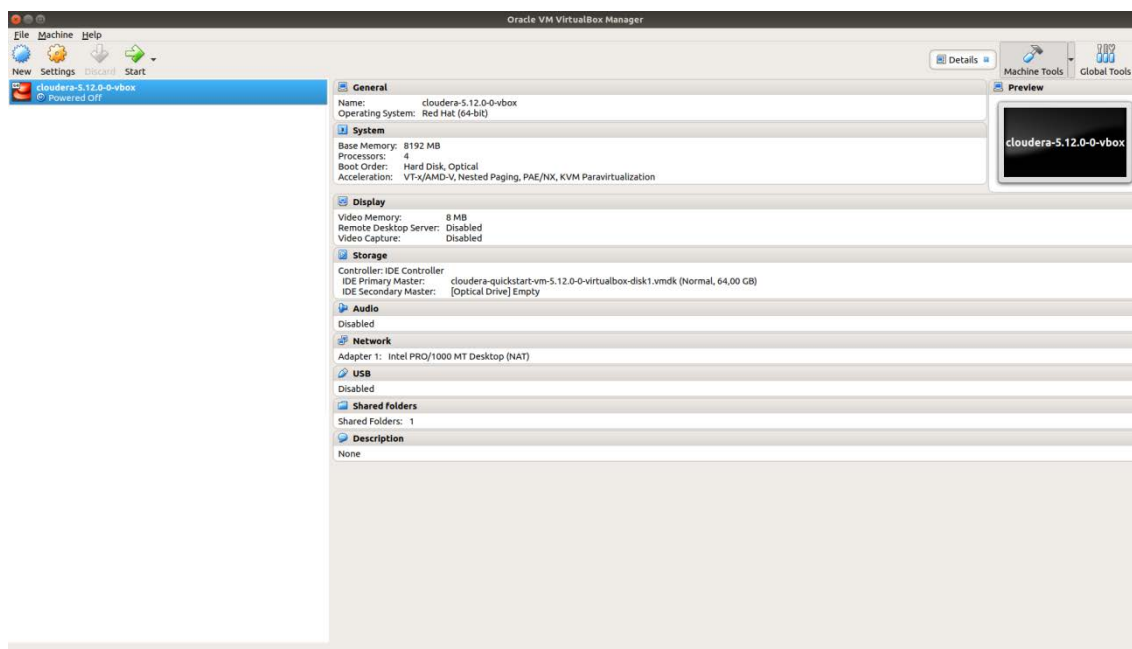
Para utilizar esta máquina virtual, deberemos tener el software VirtualBox descargado e instalado en nuestro ordenador. Dicho software se encuentra disponible desde el siguiente enlace:

## PREPARACIÓN DEL ENTORNO DE TRABAJO.

<https://www.virtualbox.org/>

Una vez tengamos VirtualBox instalado y la máquina virtual descargada y descomprimida en nuestro ordenador, deberemos importar la máquina virtual en nuestra instalación de VirtualBox. Para ello, ejecuta el programa VirtualBox, y en su ventana principal, haz click en “File”, seguido de “Import appliance”. Seguidamente aparecerá una ventana donde tendrás que navegar hasta la carpeta donde hayas descargado la máquina virtual, y seleccionar el fichero con extensión .ova. Una vez hayas importado correctamente la máquina virtual, ésta aparecerá a la izquierda en la ventana principal de VirtualBox, tal y como muestra la Figura 1.

Para arrancar la máquina virtual, selecciónala y clicla el botón “Start” que aparece en la parte de arriba de la ventana de VirtualBox. Una nueva ventana aparecerá, que será la del escritorio de la máquina virtual. Dependiendo de la potencia del ordenador donde se está ejecutando, este proceso puede tardar unos minutos. Una vez la máquina virtual esté funcionando, aparecerá una ventana del navegador de Internet de la máquina virtual, que se abre de forma automática al arrancar la máquina virtual, similar a la que muestra la Figura 2 <sup>1</sup>.



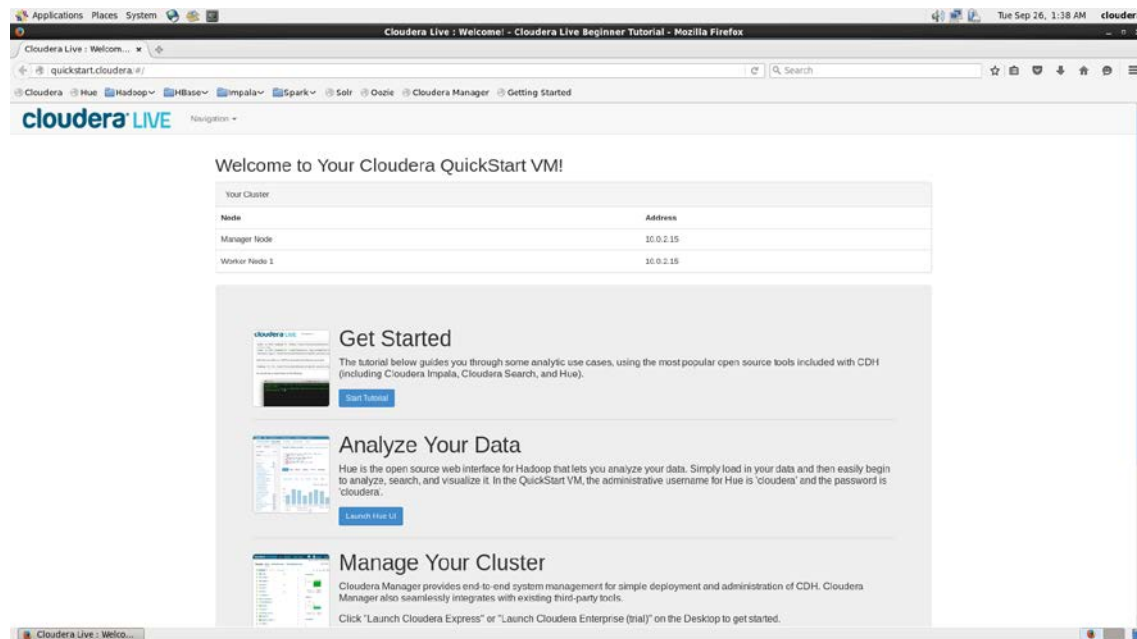
**Figura 1. Ventana de VirtualBox.**

Para apagar la máquina virtual, podemos hacerlo clicando en “System” → “Shut Down”.

---

<sup>1</sup> Se recomienda NO actualizar el software de la máquina virtual cuando lo pida.

## PREPARACIÓN DEL ENTORNO DE TRABAJO.

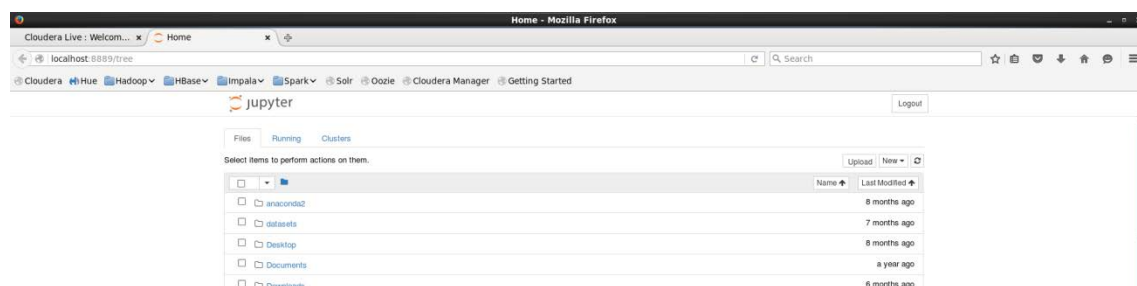


**Figura 2. Máquina virtual iniciada.**

Una vez tengamos la máquina virtual iniciada, podremos abrir un terminal y ejecutar la siguiente orden para arrancar el servidor Jupyter notebooks:

```
$ /home/cloudera/anaconda2/bin/jupyter notebook
```

Tras lo cual se abrirá una ventana del navegador donde el servidor Jupyter notebooks nos permitirá comenzar a trabajar, y que mostrará los contenidos del directorio actual que actuará como directorio raíz para el servidor jupyter, en nuestro caso el directorio home (ver Figura 3).



**Figura 3. Jupyter notebooks iniciado en la máquina virtual.**

Con el fin de que el trabajo con la máquina virtual sea lo más fácil posible, una de las tareas a realizar es definir una carpeta para que la máquina anfitrión (nuestro ordenador) y la máquina virtual compartan una carpeta. De esta forma, podremos transferir ficheros entre ambas máquinas de forma sencilla.

Para que sea posible esta compartición, además de para otras tareas, es recomendable la instalación de un software llamado "VirtualBox Guest Additions" en el sistema operativo de la máquina virtual. Para ello, en primer lugar, con la

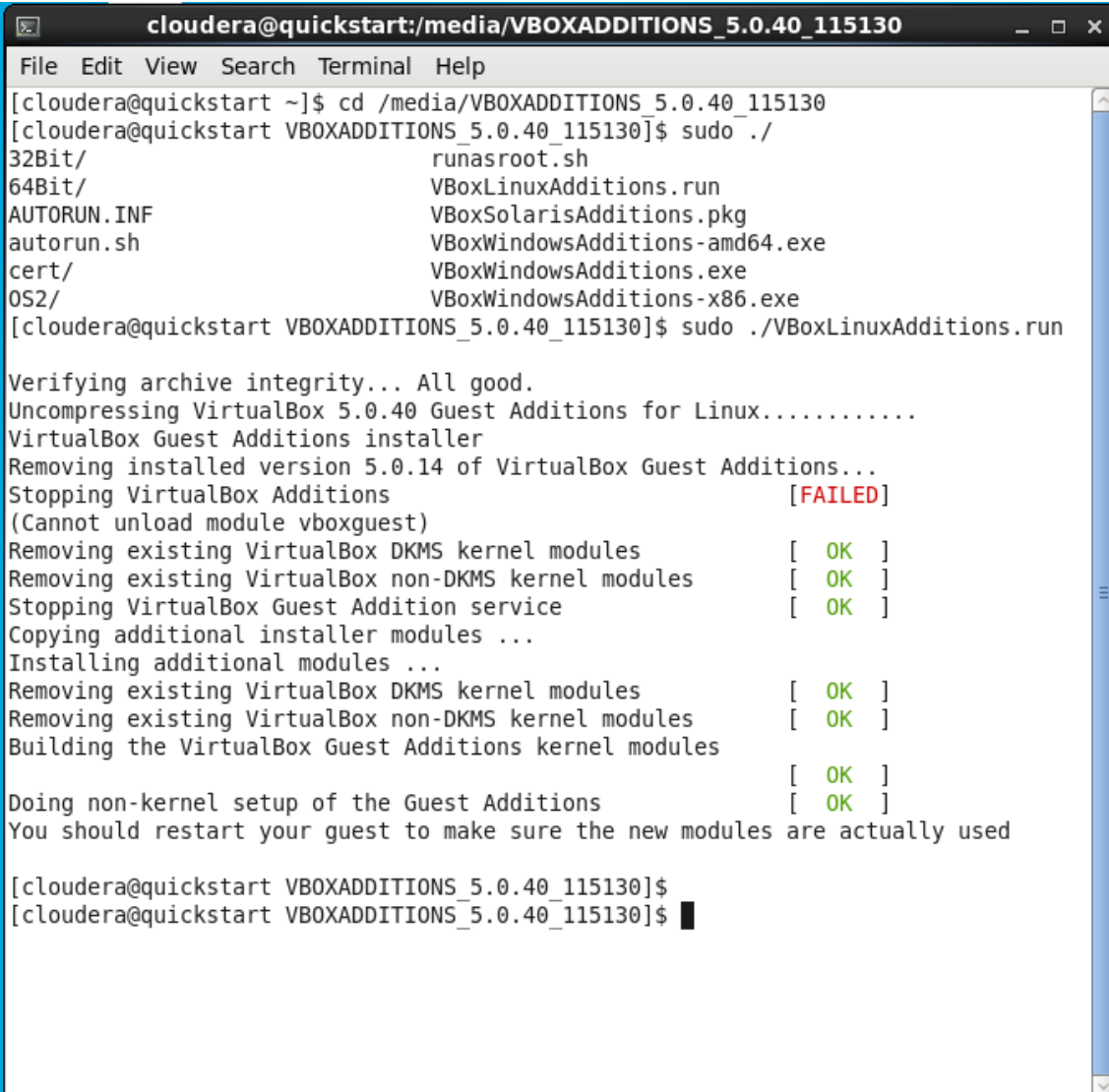
## PREPARACIÓN DEL ENTORNO DE TRABAJO.

máquina virtual encendida, clicamos en “Devices” → “Insert Guest Additions CD Image”. Esto sirve para montar en la máquina virtual una imagen de CD que contiene el software a instalar. Tras esto, desde un terminal, cambiamos al directorio donde ha sido montado dicho CD y ejecutamos la siguiente orden:

```
$ sudo ./VBoxLinuxAdditions.run
```

Dando el resultado que se muestra en la Figura 4. Tras reiniciar la máquina virtual, las guest additions estarán instaladas y en funcionamiento.

Tras instalar las guest additions, ahora deberemos configurar una carpeta compartida entre la máquina anfitrión y la máquina virtual. Para ello, debemos modificar las características (“Settings”) de la máquina virtual. Esto se puede hacer, entre otras opciones, con la máquina virtual encendida clicando en “Machine” → “Settings” (como se muestra en Figura 5)



```
cloudera@quickstart:/media/VBOXADDITIONS_5.0.40_115130
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ cd /media/VBOXADDITIONS_5.0.40_115130
[cloudera@quickstart VBOXADDITIONS_5.0.40_115130]$ sudo ./
32Bit/                               runasroot.sh
64Bit/                               VBoxLinuxAdditions.run
AUTORUN.INF                          VBoxSolarisAdditions.pkg
autorun.sh                           VBoxWindowsAdditions-amd64.exe
cert/                                VBoxWindowsAdditions.exe
OS2/                                 VBoxWindowsAdditions-x86.exe
[cloudera@quickstart VBOXADDITIONS_5.0.40_115130]$ sudo ./VBoxLinuxAdditions.run

Verifying archive integrity... All good.
Uncompressing VirtualBox 5.0.40 Guest Additions for Linux.....
VirtualBox Guest Additions installer
Removing installed version 5.0.14 of VirtualBox Guest Additions...
Stopping VirtualBox Additions [FAILED]
(Cannot unload module vboxguest)
Removing existing VirtualBox DKMS kernel modules [ OK ]
Removing existing VirtualBox non-DKMS kernel modules [ OK ]
Stopping VirtualBox Guest Addition service [ OK ]
Copying additional installer modules ...
Installing additional modules ...
Removing existing VirtualBox DKMS kernel modules [ OK ]
Removing existing VirtualBox non-DKMS kernel modules [ OK ]
Building the VirtualBox Guest Additions kernel modules [ OK ]
Doing non-kernel setup of the Guest Additions [ OK ]
You should restart your guest to make sure the new modules are actually used

[cloudera@quickstart VBOXADDITIONS_5.0.40_115130]$
[cloudera@quickstart VBOXADDITIONS_5.0.40_115130]$ █
```

Figura 4. Instalación de VirtualBox Guest Additions.

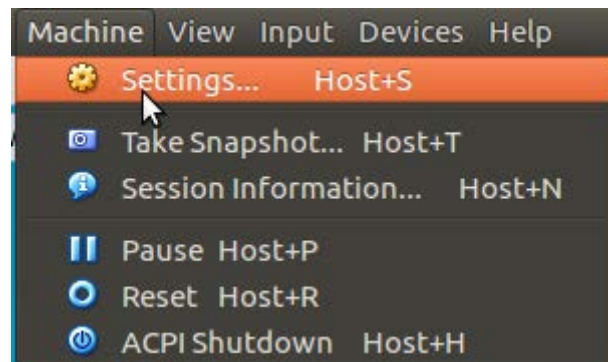


Figura 5. Settings.

En la ventana de "Settings", seleccionamos el apartado "Shared folders" y clicamos en el icono de la derecha para añadir una nueva carpeta, lo que muestra la ventana que se presenta en la Figura 6. Ahora tendremos que insertar la ruta de la carpeta de la máquina anfitrión que deseemos compartir, un nombre para dicha carpeta compartida, y definir el resto de opciones de compartición que se indican.

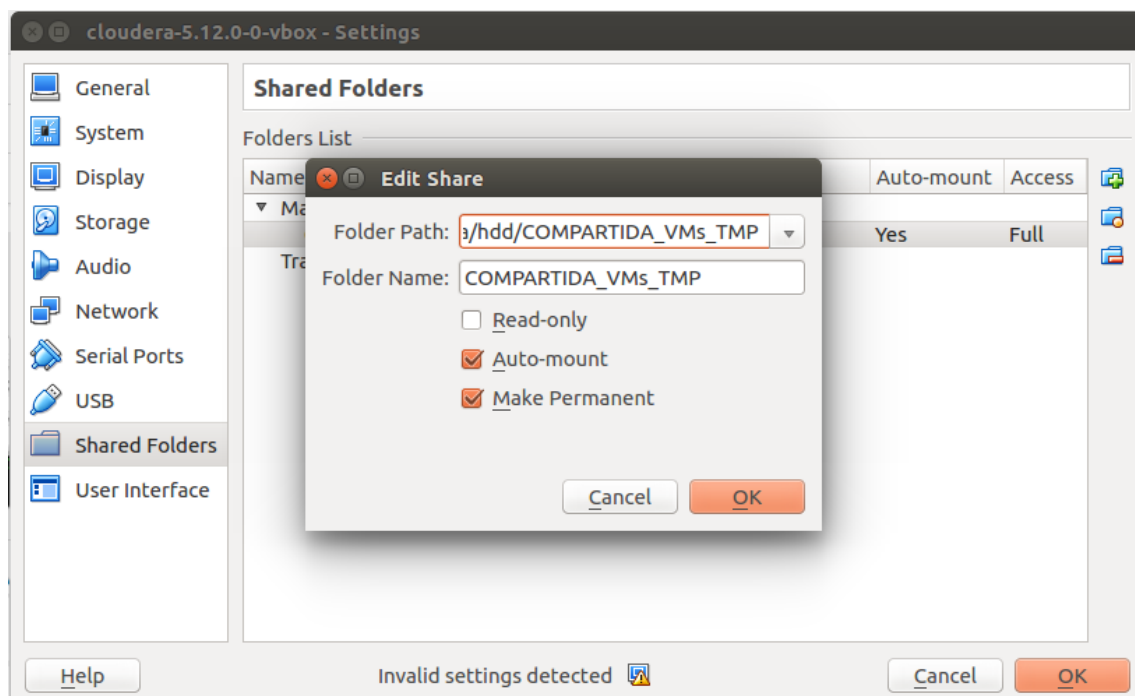


Figura 6. Carpeta compartida.

Para finalizar con este procedimiento, tenemos que incluir al usuario "cloudera" en el grupo "vboxsf" con la siguiente orden.

```
$ sudo usermod -a -G vboxsf cloudera
```



Tras cerrar la sesión del usuario actual e iniciar otra nueva, podremos acceder a la carpeta compartida a través del acceso directo que aparece en el escritorio de la máquina virtual, como se muestra en la Figura 7.

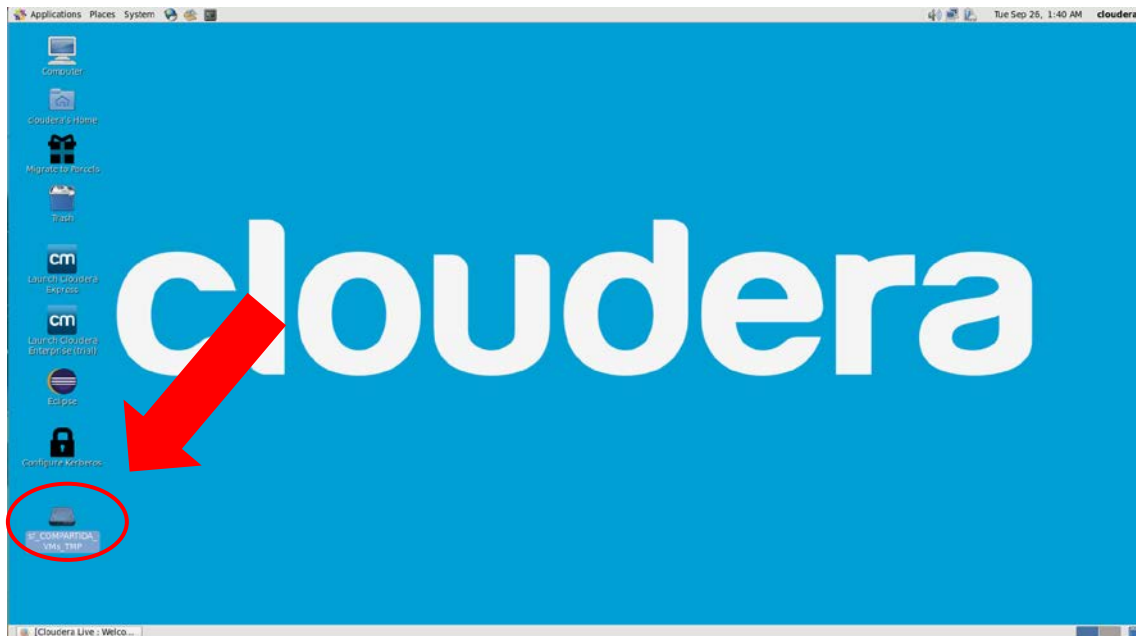


Figura 7. Acceso directo a carpeta compartida.

### 3. Contenedor

Un entorno similar al anterior pero con menores requerimientos computacionales está basado en contenedores Docker. Estos contenedores permiten el despliegue de aplicaciones de forma ligera ya que no incluyen un sistema operativo completo (al contrario que la máquina virtual).

Docker es una herramienta muy completa y sus detalles de implementación y funcionamiento quedan fuera del ámbito de esta asignatura. Existen en Internet gran cantidad de materiales para que el/la estudiante que esté interesado pueda profundizar en ella si lo desea.

Para utilizar el contenedor Docker para esta práctica, lo primero es instalar el programa Docker, siguiendo las instrucciones que se encuentran en el siguiente enlace según el sistema operativo que tenga el ordenador donde se va a instalar:

<https://docs.docker.com/install/>



## PREPARACIÓN DEL ENTORNO DE TRABAJO.

También es necesario instalar la herramienta Docker-compose, siguiendo las instrucciones del siguiente enlace:

<https://docs.docker.com/compose/install/>

Una vez tengamos estos programas instalados, deberemos descargar el archivo `mids-cloudera-hadoop.zip` disponible en el curso virtual. Una vez descomprimido, contiene una serie de archivos que definen la creación del entorno de trabajo. Uno de estos archivos debe ser editado, es el archivo `docker-compose.yml`, en su última línea debemos sustituir `<CARPETA LOCAL>` por la ruta de una carpeta de nuestro sistema de archivos. Esta carpeta local será la carpeta donde se almacenarán los notebooks que creemos. En los videos se utiliza la carpeta `/media/notebooks`.

En este momento, ya estamos listos para crear el contenedor para realizar la práctica. Para ello, desde un terminal de línea de comandos, debemos cambiarnos al directorio donde se encuentre el archivo `docker-compose.yml` y a continuación ejecutamos la siguiente orden:

```
docker-compose up -d
```

En un ordenador con sistema operativo Ubuntu, la ejecución de esta orden muestra el resultado mostrado en la [Figura 8](#). La primera vez que se ejecute esta orden en un ordenador, se mostrará el proceso de descarga del contenedor, proceso que puede tardar dependiendo de la conexión de red que estemos utilizando.

```
$ docker-compose up -d
Starting midsclouderahadoop_quickstart.cloudera_1 ...
Starting midsclouderahadoop_quickstart.cloudera_1 ... done
```

**Figura 8. Resultado de docker-compose.**

Tras esto, podremos ver el contenedor recién creado si ejecutamos la orden siguiente, cuyo resultado se observa en la [Figura 9](#).

```
docker ps
```

```
$ docker ps
CONTAINER ID   IMAGE                                COMMAND                  CREATED        STATUS        PORTS
a42b85de163   ankitharwan/mids-cloudera-hadoop:latest  "bash -c '/root/star..."  8 days ago    Up 26 minutes   0.0.0.0:7188->7188/tcp, 0.0.0.0:8032->8032/tcp, 0.0.0.0:8042->8042/tcp, 0.0.0.0:8088->8088/tcp, 0.0.0.0:8889->8889/tcp, 0.0.0.0:8983->8983/tcp, 0.0.0.0:10000->10000/tcp, 0.0.0.0:10020->10020/tcp, 0.0.0.0:11000->11000/tcp, 0.0.0.0:19888->19888/tcp, 0.0.0.0:10070->10070/tcp, 0.0.0.0:10075->10075/tcp, 0.0.0.0:10010->10010/tcp, 0.0.0.0:8022->22/tcp, 0.0.0.0:8887->8887/tcp   midsclouderahadoop_quickstart.cloudera_1
```

**Figura 9. Docker ps**

Esta orden muestra, entre otra información, el identificador del contenedor. Este identificador se puede utilizar para conectarnos a él por línea de comandos, con la orden siguiente, cuyo resultado se muestra en la [Figura 10](#) (hay que notar que no es necesario copiar el identificador completo del contenedor, basta con los primeros dígitos):

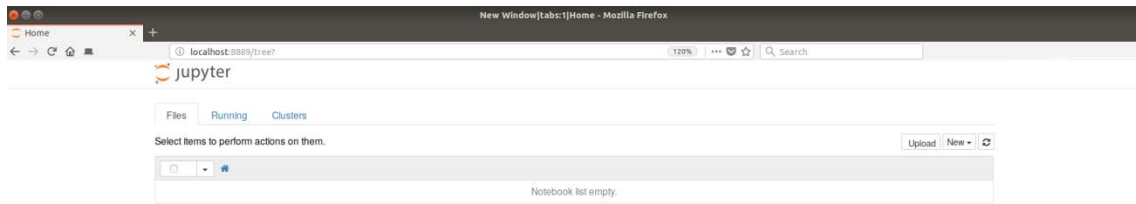
## PREPARACIÓN DEL ENTORNO DE TRABAJO.

```
docker exec -i -t <id del contenedor> /bin/bash
```

```
$ docker exec -i -t a56 /bin/bash
[root@quickstart /]#
```

**Figura 10. Conexión al contenedor.**

Estando el contenedor en ejecución, abriremos un navegador e introduciremos la dirección <http://localhost:8889>, lo que nos llevará al servidor de Jupyter notebooks (ver Figura 11), donde podremos empezar a trabajar. Como se ve en la Figura 11 no muestra nada, debido a que la carpeta local que hemos indicado en el fichero docker-compose.yml se encuentra vacía.



**Figura 11. Servidor Jupyter notebooks arrancado.**

Para parar el contenedor, ejecutamos la orden:

```
docker stop <id del contenedor>
```

Para volver a ejecutar el mismo contenedor, sin crear otro nuevo, ejecutamos la orden:

```
docker start <id del contenedor>
```

## 4.Recomendaciones

En el caso de que el/la estudiante no tenga experiencia previa con máquinas virtuales o contenedores, es recomendable no dejar el despliegue del entorno de trabajo (bien con la máquina virtual o bien con el contenedor) y la realización de la práctica para el final del período de entrega.