|  |
| --- |
| Máster en Big Data |
| Tecnologías de Almacenamiento |
| 2. Hands-On: Uso de HDFS   * Albert Ripoll |

Índice

[1. Introducción 3](#_Toc96241371)

[2. Desplegar el nodo de Hadoop 3](#_Toc96241372)

[3. Uso de HDFS 4](#_Toc96241373)

[3.1. Explorar HDFS 4](#_Toc96241374)

[3.2. Insertar Archivos 4](#_Toc96241375)

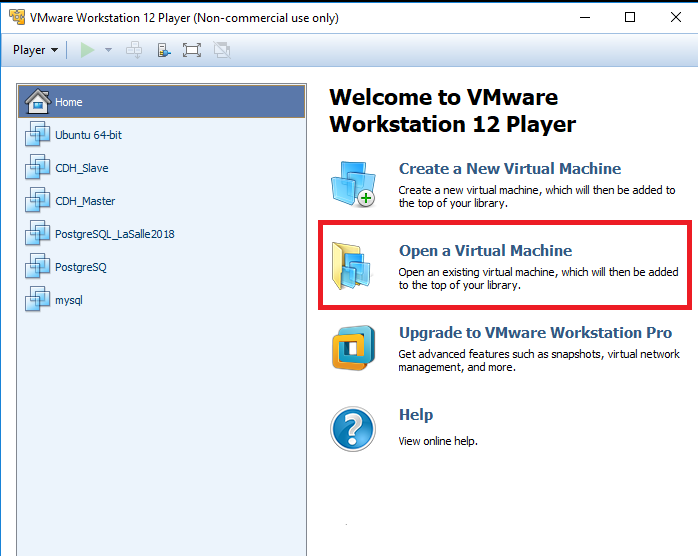
[3.3. Manipular Archivos 4](#_Toc96241376)

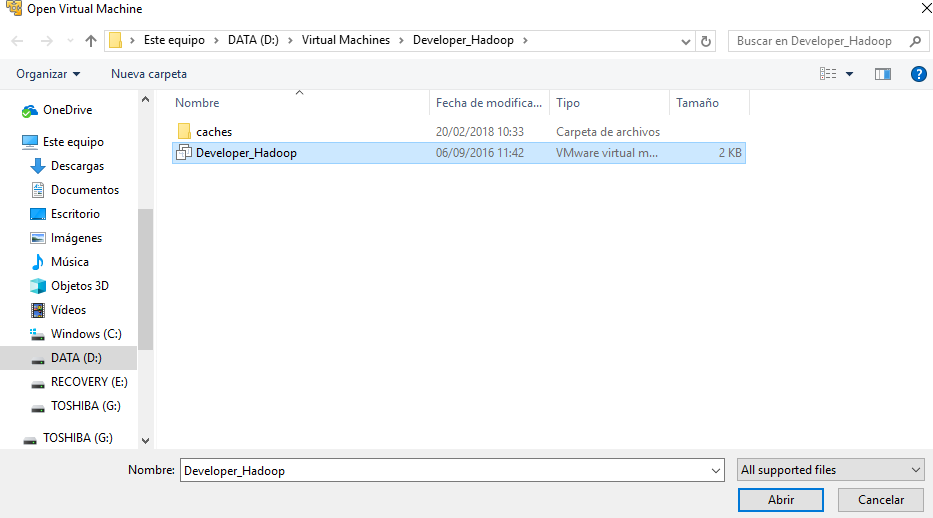
# Introducción

El objetivo de este Hands-On es ganar confianza con el uso de HDFS por línea de comandos.

# Desplegar el nodo de Hadoop

Junto con este enunciado se proporciona una instancia de clúster Hadoop pseudodistribuido. Descomprimir la carpeta e importar la máquina a VMWare como ya se ha hecho anteriormente.





# Uso de HDFS

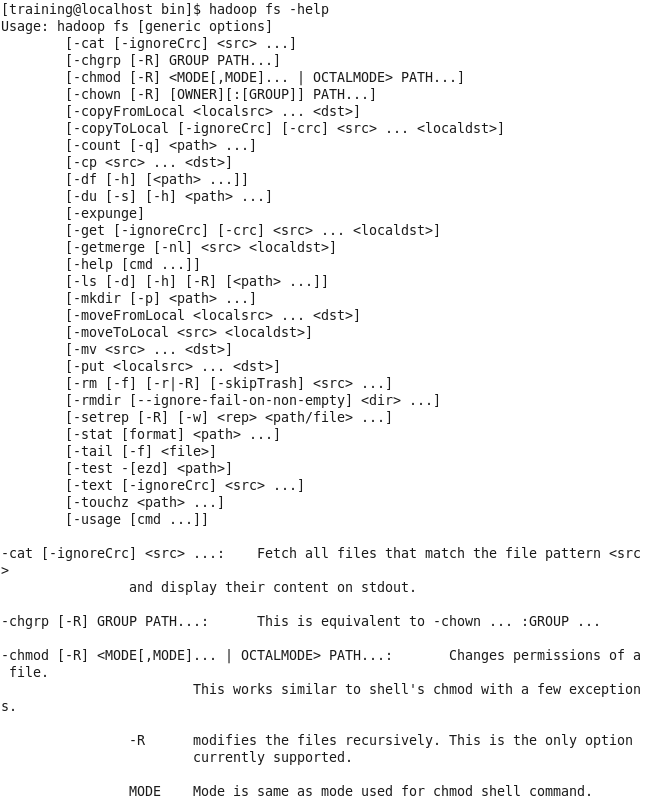
Todas las instrucciones deben introducirse mediante Shell de Linux.

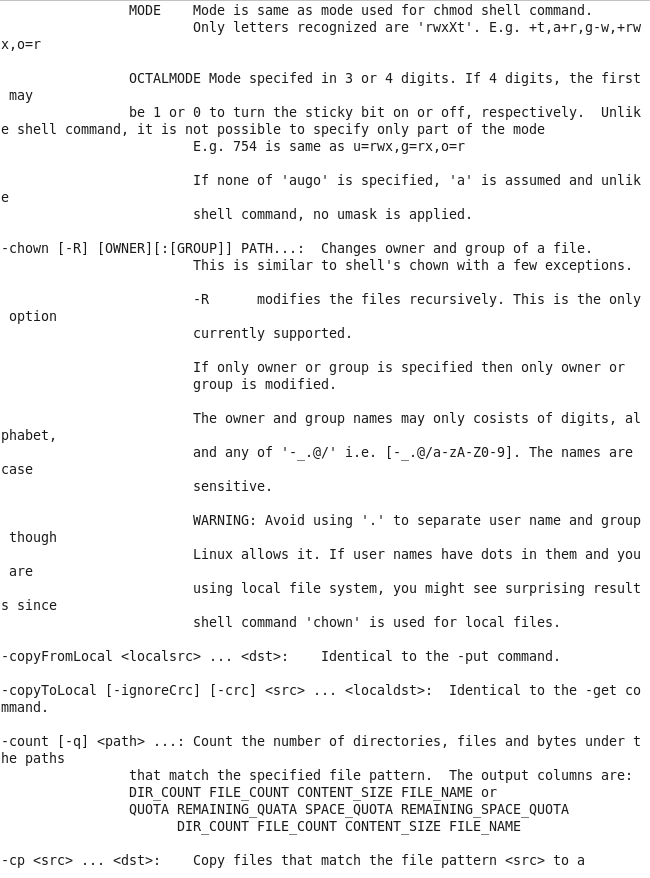
IMPORTANTE: Adjuntar el comando y captura de que los ficheros afectados por el comando se han creado/modificado/eliminado correctamente.

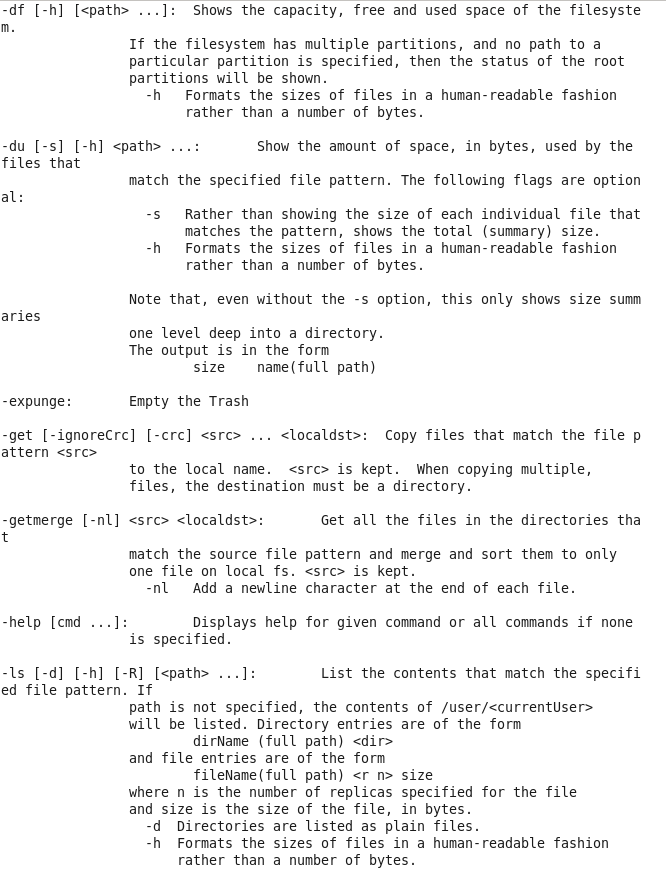
## Explorar HDFS

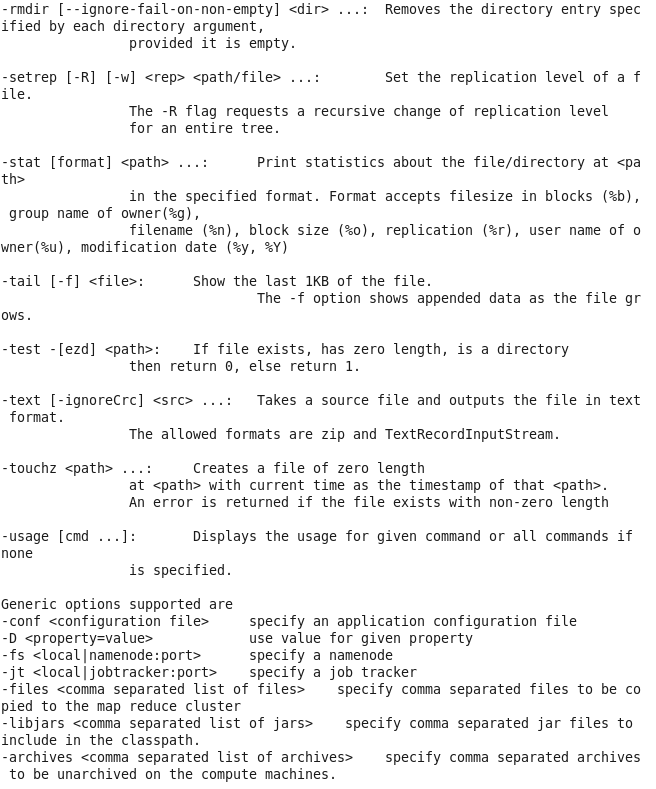
Para explorar HDFS (Hadoop Distributed File System) usamos el comando “hadoop fs”.

1. Muestra el comando de ayuda de Hadoop

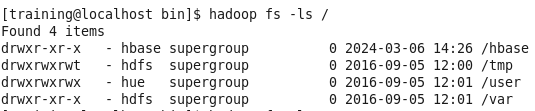






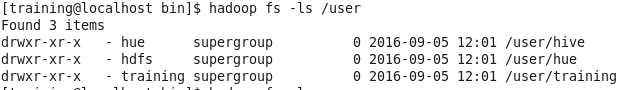


1. Listar el contenido de la raíz



Vemos que dentro de Hadoop lo que hay son 4 carpetas: hbase, tmp, usr, var creadas en las fechas y horas listadas.

1. Listar el contenido del directorio */user*



Accedemos a la carpeta user y vemos lo que hay. Hay 3 carpetas: hive, hue, training creadas en las fechas y horas listadas.

## Insertar Archivos

1. En el directorio */home/training/training\_materials/developer/data* del filesystem local descomprimir el archivo s*hakespeare.tar.gz* e insertar en HDFS el contenido de la carpeta descomprimida en */user/training/Shakespeare*

Para descomprimir el archivo shakespeare.tar.gz en el directorio local:

• Primero accedemos al directorio que nos ha dicho el enunciado con cd (*change directory*).

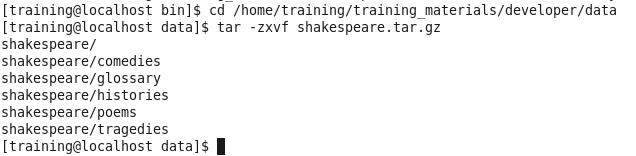
• Después usamos la función *tar* que es la que se utiliza para manipular archivos comprimidos. Su nombre proviene de “tape archive” de cuando las copias eran con cintas magnéticas.

-z: Indica a tar que descomprima el archivo utilizando gzip.

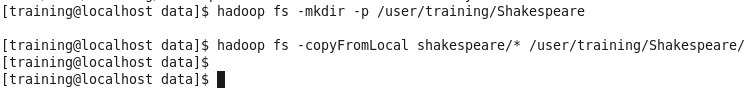
-x: Le indica a tar que extraiga el contenido del archivo. (como si fuera cntrl+x; cortar)

-v: Hace que tar imprima información detallada sobre el proceso en la consola (es lo que se conoce como modo verbose).

-f: Indica el nombre del archivo con el que se trabajará que es el *shakespeare.tar.gz*



• Finalmente copiamos el contenido descomprimido a HDFS. En primer lugar creamos un directorio (*-mkdir*) en /user/training/Shakespeare. La opción “-p” indica a Hadoop que cree todos los directorios necesarios en la ruta especificada, incluso si algunos de los directorios intermedios no existen. Es decir, si la carpeta de training dentro de user no existe, la opción “-p” hará que Hadoop cree todos los directorios intermedios automàticamente. En segundo lugar, copiamos del local “-copyFromLocal” toda (\*) la carpeta de shakespeare en la carpeta “Shakespeare” que acabamos de crear.

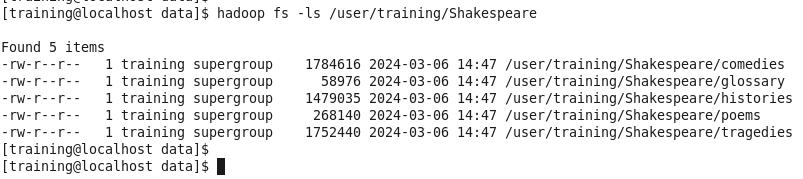


• Otra forma de hacerlo seria con la función “put”

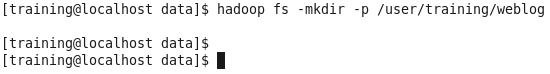


1. Listar el directorio donde se ha realizado la importación

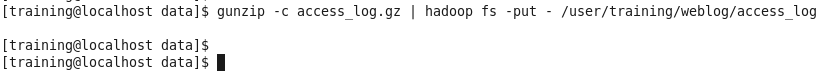
Se usa la función “ls” para listar el contenido de dicho directorio.



1. Crear un directorio llamado *weblog* en HDFS en la ruta siguiente: */user/training*

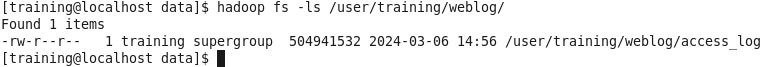


1. Importar el archivo *access\_log.gz*. (gunzip –c descomprimirá el archivo y volcará todo el contenido a la salida estándar y – en hadoop leerá los datos de la salida estándar)



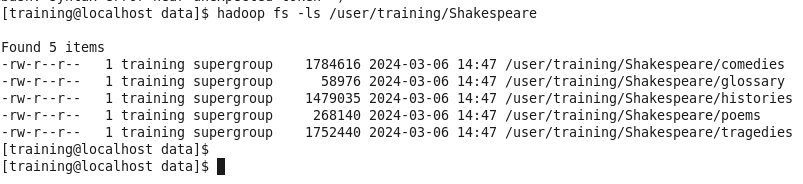
Utilizamos gunzip -c para descomprimir el archivo access\_log.gz y luego hadoop fs -put - para leer desde la salida estándar (-) y colocar los datos en HDFS en la ruta /user/training/weblog/access\_log.

Verificamos que que el archivo access\_log se haya importado correctamente a HDFS leyendo la carpeta weblog:

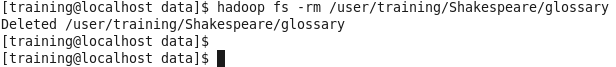


## Manipular Archivos

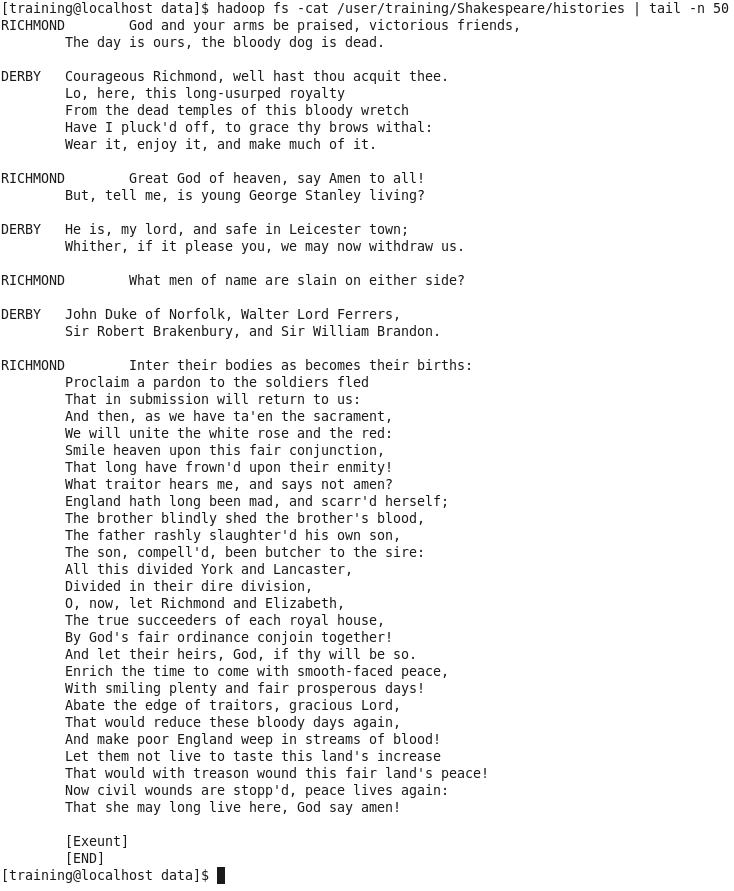
1. Listar el contenido de la carpeta *Shakespeare*



1. Borra el archivo *glossary*



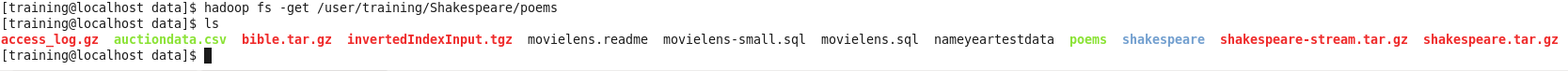
1. Muestra las últimas 50 líneas del archivo *histories*
2. hadoop fs -cat /user/training/Shakespeare/histories: Esto imprime el contenido del archivo histories ubicado en el directorio /user/training/Shakespeare en HDFS. hadoop fs -cat simplemente muestra el contenido del archivo en la salida estándar. | tail -n 50: Utiliza el operador de tubería (|) para pasar la salida del comando anterior (las líneas completas del archivo histories) al comando tail -n 50. Este último comando muestra las últimas 50 líneas de su entrada.



1. Descarga el archivo *poems* en el filesystem local

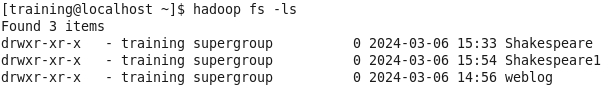
Para coger archivo de Hadoop se usa el -get.

Usamos “ls” para leer (listar) lo que tenemos en el archivo local. No ponemos “hadoop fs” al inicio porque ahora este archivo no está en hadoop sino que está en nuestra computadora local ya que lo acabamos de traer. En la salida vemos que aparece el archivo “poems” en verde. Los distintos colores son para indicar distintos formatos de archivos.



# Aprendizajes

## Ejercicio b) Listar directorio de donde se ha realizado la importación



Comandos para escribir:

Hadoop fs → Se actúa en hadoop

-ls → List directory contents (listar el contenido que hay)

Respuesta en pantalla:

• En la primera columna hay los permisos (RWX).

d → Indica que es un directorio.

r → Readable (permiso sobre el archivo que indica se puede leer)

w → Writeable (permiso sobre el archivo que indica se puede escribir)

x → Executable (permiso sobre el archivo que indica se puede ejecutar)

Las siguientes 3 letras son los permisos del usuario dueño del archivo (rwx). Las siguientes 3 letras los permisos del grupo al que pertenece el usuario (r-x). Y las 3 siguientes letras es para cualquier otro grupo (r-x)

De esa forma en este caso el usuario dueño puede leer, escribir y ejecutar sus propios archivos. El grupo de usuarios puede leer y ejecutar pero no (-) escribir. Los otros usuarios de otro grupo puede lee pueden leer y ejecutar pero no (-) escribir

• En la segunda columna hay el nombre de usuario y el grupo.

Training -> es como se llama nuestro usuario

Supergroup -> es como se llama nuestro grupo de usuario

• En la tercera columna hay el nombre de usuario y el grupo.

0 → Cuando es un directorio

Otro número -> El tamaño en bytes que tiene el archivo

• En la cuarta columna hay la fecha que se creó la carpeta en formato yyyy-mm-dd hh-mm.

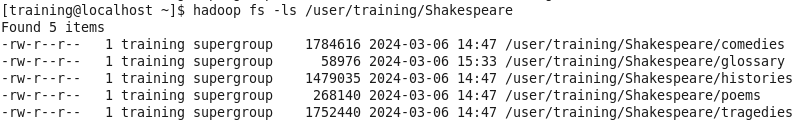
• En la quinta columna hay el nombre de la carpeta o archivo.

Comandos para escribir:

Hadoop fs → Se actúa en hadoop

-ls → List directory contents (listar el contenido que hay)

/user/training/Shakespeare → Ruta de carpetas que nos lleva a la carpeta de Shakespeare



Respuesta en pantalla:

Es del mismo estilo que la respuesta anterior.

## Ejercicio c) Listar directorio de donde se ha realizado la importación