0

Javier Plaza Sisqués

Universidad europea  Grandes Volúmenes de Datos

Actividad 02 - Hadoop

# 

# Contenido

[0](#_Toc157856966)

[Contenido 1](#_Toc157856967)

[1. Obtén una imagen de Hadoop para Docker: Puedes buscar en Docker Hub una imagen de Hadoop que alguien más haya creado. 2](#_Toc157856968)

[a. Iniciar y verificar el contenedor de Hadoop 2](#_Toc157856969)

[2. Practica Hadoop 3](#_Toc157856970)

[a. Crear un nuevo directorio en HDFS 3](#_Toc157856971)

[b. Copiar los archivos al contenedor 3](#_Toc157856972)

[c. Listar los archivos en el directorio HDFS creado 3](#_Toc157856973)

[d. Leer el contenido de los archivos (Sustituye localfile.txt son los archivos descargados): 4](#_Toc157856974)

[3. Operación map-reduce 4](#_Toc157856975)

# Obtén una imagen de Hadoop para Docker: Puedes buscar en Docker Hub una imagen de Hadoop que alguien más haya creado.

## Iniciar y verificar el contenedor de Hadoop

Para verificar que nuestra instalación de Hadoop se ha realizado correctamente deberemos de ejecutar el siguiente comando:

Texto

Descripción generada automáticamente

Como podemos ver hemos instalado la versión 3.2.1 de Hadoop.

Podemos comprobar que funciona correctamente utilizando un comando sencillo como listar el directorio raíz dentro del sistema a de Hadoop, para ello hemos usado los siguiente:

Texto

Descripción generada automáticamente

También podemos verificar la instalación conectándonos al nodo vía web:

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

# Practica Hadoop

## Crear un nuevo directorio en HDFS

Para poder crear un nuevo directorio en Hadoop simplemente nos conectaremos al contenedor correspondiente y ejecutaremos el comando mkdir seguido de la carpeta a crear:

Texto

Descripción generada automáticamente

## Copiar los archivos al contenedor

Para copiar los archivos que se han descargado para la práctica deberemos de usar el comando put dentro de Hadoop, podemos ver un ejemplo en la siguiente captura:

Texto

Descripción generada automáticamente con confianza media

Cabe destacar que los datos fueron descargados previamente y guardados en la carpeta /datos.

## Listar los archivos en el directorio HDFS creado

Una vez todos los datos hayan sido copiados deberemos de listar la carpeta contenedora de los mismos para verificar su correcto copiado, podemos verlo en la siguiente caputra:

Texto

Descripción generada automáticamente

## Leer el contenido de los archivos (Sustituye localfile.txt son los archivos descargados):

Por último, podremos leer los datos utilizando el comando cat, a continuación se muestra una captura con el contenido de uno de los csv:

Texto

Descripción generada automáticamente

# Operación map-reduce

La operación map-reduce que se ha utilizado ha sido para contar el número de pesos que se repiten en el archivo players.csv. En primer lugar, se obtenido el contenido del fichero csv, una vez se ha almacenado en una variable se divide en líneas haciendo un Split mediante el salto de carro (\n). Una vez tengamos divididas las líneas simplemente se ejecuta un bucle foreach que recorre cada una de las mismas. Por cada línea se realiza otro Split usando como separador el carácter “,”. Obtenemos la altura y la almacenamos para ser contada.

Texto

Descripción generada automáticamente

La segunda parte consta simplemente de obtener los valores anteriores e ir sumando para obtener el número de veces que se repite. Con esto conseguimos obtener el número de personas que tienen el mismo peso.

