Hadoop

x

Jupyter

Big Data Aplicado

Alejandro Cruz Aguilar

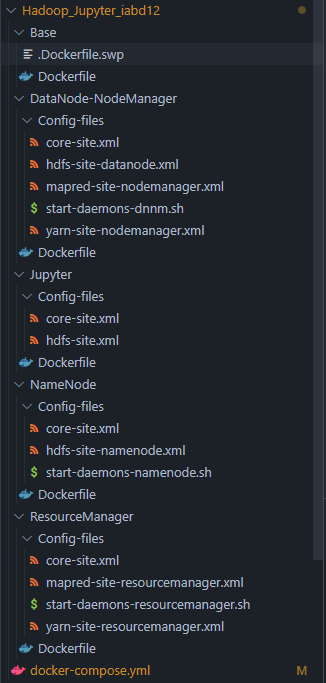
*Índice*

[Preparación del Entorno 2](#_Toc475207713)

## Preparación del Entorno

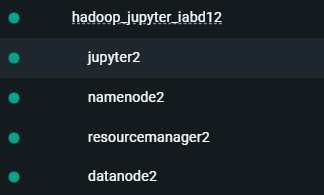
Antes de crear el cluster, necesitaremos lo siguiente:

* Una imagen BASE, de la cual el resto de los contenedores van a funcionar. Este debe tener instalado el HADOOP.
* Una imagen de JUPYTER, el cual tendrá instalada las librearías básicas, como Python, PIP...
* Una imagen de DATANODE; NAMENODE; RESOURCEMANAGER. Estos, para poder subir el archivo mediante HADOOP.
* Por último, un docker-compose, el cual se encargará de enlazar todos estos contenedores.



## Uso de Hadoop mediante en consola

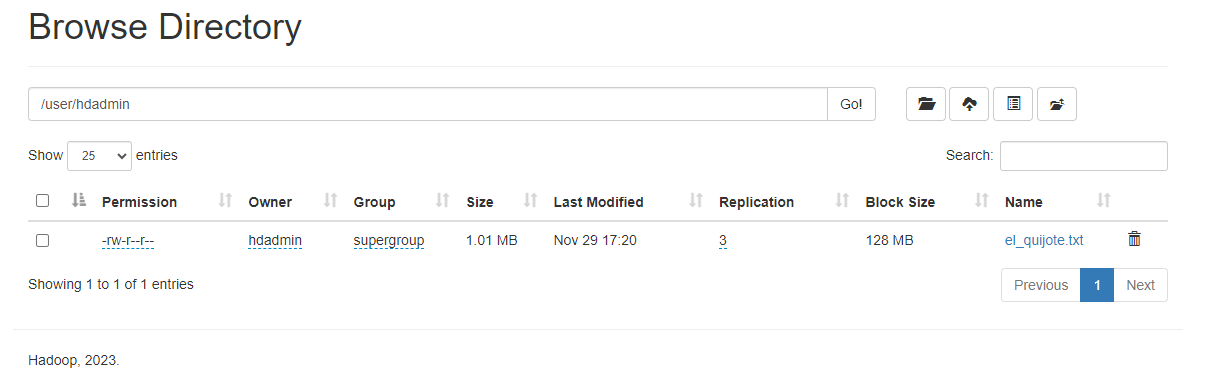
Ahora que hemos creado el cluster con todo lo que necesitamos, (jupyter, namenode, datanode, resourcemanager), podemos continuar subiendo el archivo “el\_quijote.txt” a hadoop.



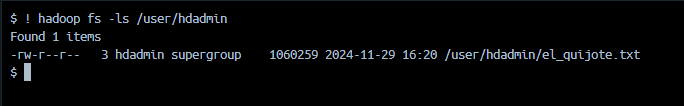
La sintaxis del comando es “hdfs dfs –put <ruta\_archivo> <ruta\_hadoop>”



Ahora podemos comprobar desde el mismo namenode, que el archivo a sido subido sin ningun problema.

Tambien, pro si queremos asegurarnos, podemos comprobarlo mediante la consola.

Con la sintaxis “!hadoop fs –ls /user/hdadmin”, nos muestra el archivo y sus permisos.



## Creación de los scripts para contar

Ya que hemos subido el archivo a hadoop, tendremos que contar cuanta cantidad hay de cada palabra. Para ello, necesitaremos un mapreduce, lo cual lo podemos dividir en 2 partes, el mapper, y el reducer.

El código del mapper es el siguiente.



Y el código del reducer es este:



