EJERCICIOS DE SPARK

## Recuperar el libro de Moby Dick del proyecto Gutenberg: <http://www.gutenberg.org/cache/epub/2701/pg2701.txt>

El libro se encuentra en la siguiente dirección: <http://www.gutenberg.org/files/2701/2701-0.txt>

Descargamos el fichero de la web y lo incluimos en la carpeta del ejercicio:

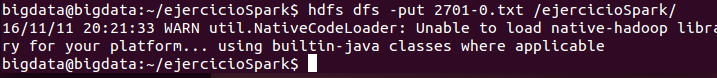




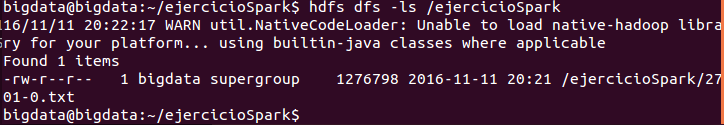
## Crear una carpeta ejercicioSpark en HDFS

## 

## Subir el fichero a HDFS a la carpeta anterior



## Comprobar que el libro está correctamente subido

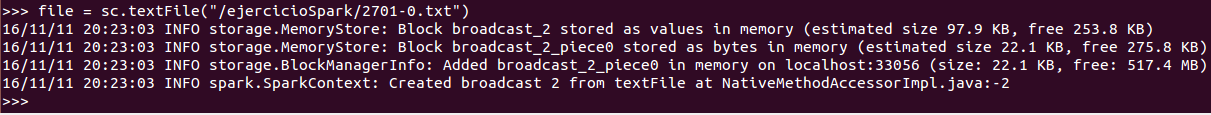


## Acceder a pyspark

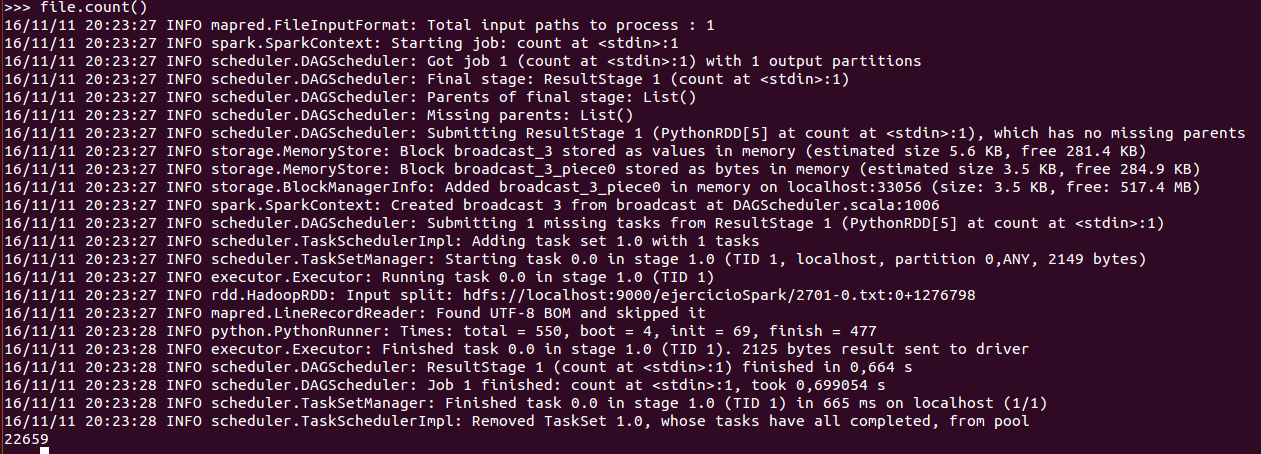


## Indicar los comandos y el resultado de contar el número de líneas que tiene el fichero.

Cargamos el fichero en un RDD:



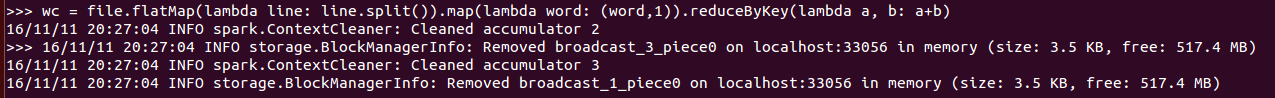
Contamos las líneas con la función count():



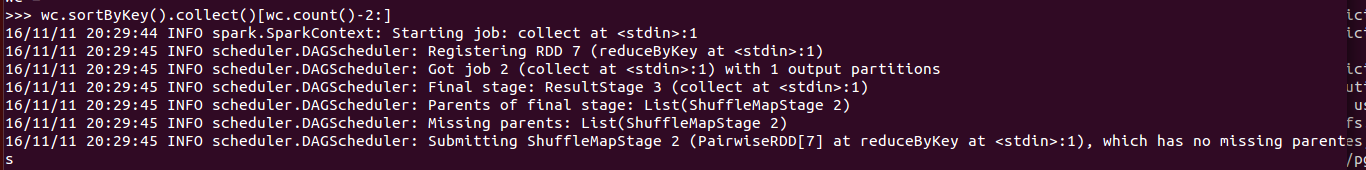
Hay 22.659 líneas en el fichero.

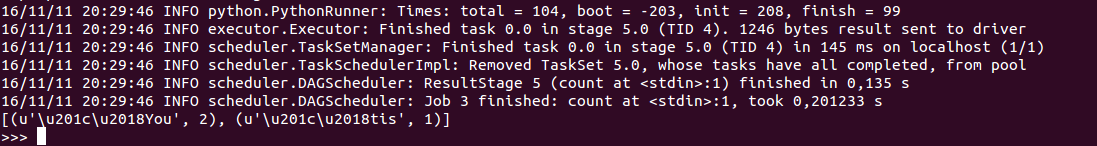
## Ejecutar un word count e indicar ordenadas alfabéticamente las dos últimas palabras que aparecen y el número de repeticiones

Guardamos el contador de palabras en un RDD:

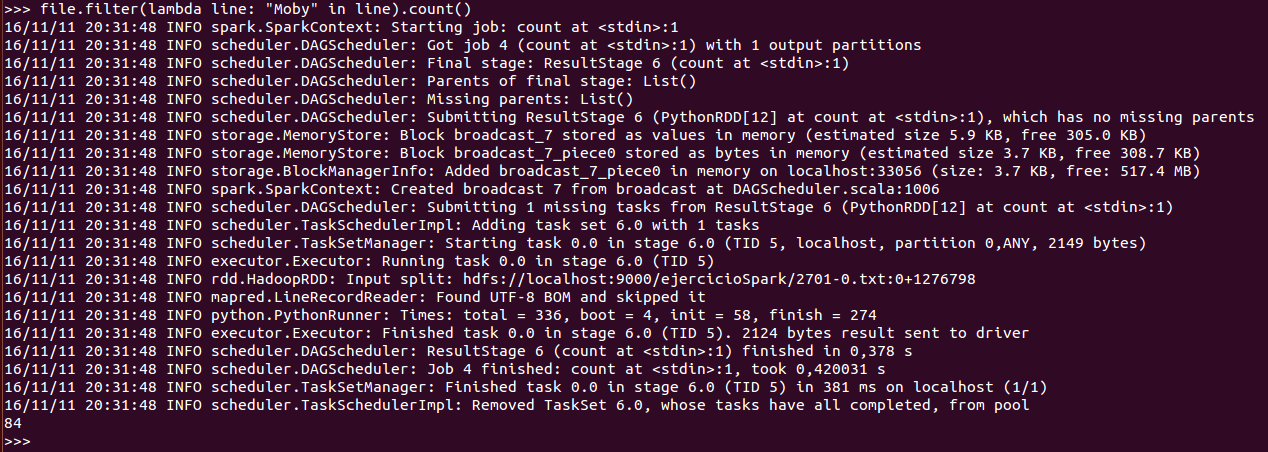


Ordenamos alfabéticamente:

`



## Indicar las instrucciones y el valor devuelto del número de líneas en las que aparece la palabra Moby en el libro



## Describe con tus palabras las diferencias de usar Spark frente a Map Reduce

Spark se ejecuta en memoria mientras que MapReduce se ejecuta en los discos duros. Esto hace que Spark ejecute los procesos mucho más rápido que MapReduce.

No obstante, esto también hace que MapReduce sea más escalable que Spark, ya que este último necesita que los datos quepan en la memoria.