Ejemplos pair Rdds:

Ahora vamos a realizar un análisis al estilo de escritura de Shakespeare, para ello nos interesa conocer las palabras más repetidas en sus obras. Muestra por pantalla las 100 palabras que más veces aparecen en las obras de Shakespeare, junto con la frecuencia de aparición de cada una, ordenadas descendientemente (de mayor a menor frecuencia de aparición).

val logs=sc.textFile("shakespeare/\*")

val logs2=logs.flatMap(line => line.split(" "))

val logs3=logs2.map(word => (word, 1)).reduceByKey(\_+\_)

val logs4=logs3.map(word => word.swap).sortByKey(false).map(value => value.swap)

El análisis anterior no es todo lo completo que una situación real nos puede requerir, pues entre las palabras más repetidas encontramos: pronombres, artículos, proposiciones...; es decir, palabras que no nos aportan información alguna. A este tipo de palabras se les conoce como palabras vacías (StopWords), y nos interesa eliminarlas de nuestro análisis.

Para la realización de este ejercicio vamos a necesitar un dataset que contenga "stop words" en inglés (pues, como es lógico, los libros de Shakespeare están en este idioma). Copia el archivo "stop-word-list.csv" en la máquina virtual y cópialo al HDFS (en la ruta por defecto).

También tenemos que tener en cuenta que, para hacer nuestro análisis, no nos interesan ni las líneas que solo contengan espacios en blanco ni las palabras que estén formadas por una única letra. Tampoco haremos distinción entre mayúsculas ni minúsculas, por lo que, por ejemplo, "Hello" y "hello" deberán ser tomadas como la misma palabra en el análisis.

El método para cambiar una línea a minúsculas es: toLowerCase

La expresión regular que toma únicamente caracteres alfabéticos es: "[^a-zA-Z]+"

Un archivo CSV está delimitado por comas.

No nos interesa mostrar las ocurrencias de una única letra. Por lo que deberemos filtrar el RDD final.

El resultado de este ejercicio debería ser parecido a:

