2_Ejecutando un MapReduce: wordcount

En este ejercicio simplemente ejecutaremos un Job consistente en la ejecución del wordcount en MapReduce sobre el dataset shakespeare. Por simplicidad, los ficheros .class y el jar ya están creados.

Como hemos comentado, wordcount cuenta el número de palabras distintas que hay en un texto dado.

Pasos a ejecutar

1. Copiar en la ruta "/home/cloudera/ejercicios" la carpeta "wordcount" y su contenido.
2. Comprobar que se han copiado correctamente
3. Examinar el contenido de los tres ficheros java para asegurarnos de que están correctos.
    a. Prestar atención los parámetros de entrada de cada clase, los tipos de datos de entrada, salida e intermedios, etc.
4. La carpeta wordcount, como hemos visto, ya contiene los javas compilados y el jar creado, por lo que solo tenemos que ejecutar el submit del job hadoop usando nuestro fichero JAR para contar las ocurrencias de palabras contenidas en nuestra carpeta "shakespeare". Nuestro jar contiene las clases java compiladas dentro de un paquete llamado "solutions", por eso se le llama de este modo.hadoop
    a. "hadoop jar wordcount/wc.jar solution.WordCount shakespeare /user/cloudera/wordcounts"
5. Una vez ejecutado, probamos a ejecutarlo nuevamente
    a. ¿Qué ocurre?

Ya existe

6. Comprobamos el resultado de nuestro MapReduce
    a. "hadoop fs -ls /user/cloudera/wordcounts"
7. Como solo hemos usado un reduce, vemos que solo hay un archivo de salida
    a. "/user/cloudera/wordcounts/part-r-00000"
8. Observamos el contenido del fichero
    a. "hadoop fs -cat /user/cloudera/wordcounts/part-r-00000 | less"
    b. Escribiendo la letra "q" salimos del comando less
9. Volvemos a ejecutar el job de nuevo
    a. "hadoop jar wc.jar solution.WordCount shakespeare/poems /user/cloudera/pwords"
10. Borramos la salida producida por nuestros jobs
    a. "hadoop fs -rm -r /user/cloudera/wordcounts /user/cloudera/pwords"
11. Ejecutamos nuevamente nuestro job
    a. "hadoop jar wc.jar solution.WordCount shakespeare /user/cloudera/count2"
12. Mientras se ejecuta, en otro terminal ejecutamos lo siguiente, para ver la lista de Jobs que se están ejecutando
    a. "mapred job -list"
13. Si conocemos la id de un job, lo podemos matar. Recordemos que cerrando un terminal no se mata el job. Para ello, ejecutamos en otra terminal lo siguiente
    a. "mapred job -kill jobid"
14. Si no te ha dado tiempo, prueba a ejecutar el job otra vez cambiándolo de nombre y prueba nuevamente.