1_Descubriendo HDFS

1. Lo primero que tenemos que hacer es habilitar la transferencia de ficheros entre vuestra máquina virtual y vuestro host. En las settings del VMWare asociadas a la VMCloudera que tenéis, habilitad la opción shared folders y añadid una carpeta en vuestros host Windows desde donde copiaremos los archivos necesarios para trabajar.
2. Una vez creada, arrancad la MV. La carpeta en Windows compartida estará en la ruta "/mnt/hgfs"
3. Abre un terminal en la MV y cámbiale el idioma a español
4. Escribe el siguiente comando
    a. `Hadoop fs`
    b. Verás un mensaje describiendo todos los comandos posibles que contiene la FsShell
5. Escribe
    a. `Hadoop fs -ls /`
    b. Esto muestra el contenido del directorio root en HDFS
6. Crea un directorio local de trabajo
    a. Por ejemplo /home/cloudera/ejercicios

    ```
    mkdir ejercicios
    rmdir ejercios (si está vacio)
    rm -r → borrar en cascada
    cp origen destino
    ```

    b. Copia el archivo "`shakespeare.tar.gz`" en ese directorio

    ```
    cp origen destino
    ```

    c. Descomprime el archivo
        i. `"tar zxvf shakespeare.tar.gz"`
    d. Copia la carpeta que acabas de descomprimir en HDFS en la ruta /user/cloudera/shakespeare. Tu home a partir de ahora será /user/cloudera
        i. `"hadoop fs -put shakespeare /user/cloudera/shakespeare"`
7. Lista el contenido de tu home en HDFS para comprobar que se ha copiado la carpeta shakespeare.
        i. "`hadoop fs -ls /user/cloudera`"
8. Observa el contenido de la carpeta shakespeare en hdfs
        i. "`hadoop fs -ls /user/cloudera/shakespeare`"
9. Borra la subcarpeta "glossary" de la carpeta shakespeare en HDFS
    a. "`hadoop fs -rm /user/cloudera/shakespeare/glossary`"
10. Comprueba que se ha borrado
11. Lista las primeras 50 últimas líneas de la subcarpeta "histories". Puedes usar los comando "cat" y "tail" (head -> primeras)
    a. "`hadoop fs -cat /user/cloudera/shakespeare/histories | tail -n 50`"
12. Copia al Sistema de ficheros local de tu MV el fichero "poems" en la ruta /home/cloudera/ejercicios/shakespeare/shakepoems.txt

    a. “hadoop fs -get /user/cloudera/shakespeare/poems /home/cloudera/ejercicios/shakespeare/shakepoems.txt”
13. Muestra las últimas líneas de shakepoems.txt copiado en tu local por pantalla
    tail 5 shakepoems.txt
14. Si has terminado muy rápido, juega un poco con los comandos disponibles en la Shell de hadoop fs. Para ello introduce “hadoop fs” y observa las posibilidades.

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.
    a. “hadoop jar 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.