Práctica 1

Descripción

Esta práctica se puntuará. La nota de la misma formará parte de la nota final de la asignatura. El alumno deberá escribir el ćodigo de un programa Java que responda a los requerimientos.

Tiempo

La práctica deberá entregarse en un plazo de dos semanas. Consultar fecha de entrega en portal aula virtual.

Enunciado

Es frecuente que la carga de datos en HDFS se realice en bruto o a partir de datos en streaming, lo que a veces conduce a tener una gran cantidad de ficheros “pequeños”. En la sesión teórica vimos que eso puede ser ineficiente o perjudicial para su gestión en Hadoop por lo querríamos desarrollar un sencillo programa capaz de transformar un conjunto grande de ficheros pequeños en un único fichero grande que los contenga todos.

- La herramienta compactará varios ficheros presentes en HDFS en un único fichero también en HDFS.
- El fichero de destino deberá ser un `SequenceFile` y estará comprimido.
- La clave del SequenceFile será la ruta del fichero original y el valor el contenido de dicho fichero.
- No será necesario que los ficheros estén concatenados en ningún orden particular.
- La herramienta se invocará como:

```
hadoop jar utad-utils.jar utad.hdfs.Compactador <glob-ficheros-orig> <fichero-dest>
```

- <*glob-ficheros-orig*>: Es un patrón que sigue la sintaxis de globs y que captura todos los ficheros que deben ser compactados.
- <*fichero-dest*>: Es una ruta, relativa o absoluta, al fichero de destino. El fichero no debe existir con antelación. Si lo hace, el programa devolverá un error indicándolo.

Ej.:

```
$ hadoop jar utad-utils.jar utad.hdfs.Compactador \
          /user/cloudera/rawdata/*/*.dat \
          /user/cloudera/datos.dat
```