CDAP – Práctica 9
Map/reduce

Apartado A (ejemplo)

Se proporcionan dos ficheros ejecutables en python **join1_mapper.py** y **join1_reducer.py** y dos ficheros de ejemplo **join1_FileA.txt** y **join1_FileB.txt**

Se puede probar la ejecución de estos ficheros mediante la siguiente orden en el terminal linux:

```
$ cat join1_File*.txt | ./join1_mapper.py | sort | ./join1_reducer.py
```

El programa **join1_mapper.py** distingue si la clave de entrada es única, en cuyo caso es la identidad, o si contiene una fecha, en cuyo caso realiza una operación de filtrado pasando la fecha al valor.

El programa **join1_reducer.py** recibe los datos generados por el programa anterior en orden alfabético, y genera los datos agrupados. Como se puede comprobar, la salida del reductor es:

```
Apr-04 able 13 991
Dec-15 able 100 991
Jan-01 able 5 991
Feb-02 about 3 11
Mar-03 about 8 11
Feb-22 actor 3 22
Feb-23 burger 5 15
Mar-08 burger 2 15
```

Apartado B (obligatorio)

Para realizar este apartado, se proporcionan una serie de ficheros que contienen datos de ventas de novelas ofrecidas por librerías:

- Los ficheros **join2_book?.txt** consisten en listas de novelas y, para cada uno, la librería que lo ofrece.
- Los ficheros **join2_num?.txt** contienen también listas de novelas y, para cada novela, el número de ejemplares que se han vendido.

El objetivo de este apartado es implementar una tarea map/reduce que proporcione respuesta a la siguiente pregunta:

¿Cuál ha sido el número total de novelas vendidas (a través de cualquier librería) de las novelas que han sido vendidas por FNAC y por Amazon? En función de esa repuesta, ¿cuál de las dos librerías hace una mejor selección de los libros que ofrece?

La implementación se puede realizar a partir de los programas proporcionados en el apartado anterior, o se puede usar cualquier otro lenguaje de

programación. Los programas pueden ser implementados y depurados usando el terminal linux. Una vez hecho esto, deben ser probados en hadoop. Para comprobar que esto ha sido hecho, se deben adjuntar a la práctica capturas de pantalla que muestren que los ficheros han sido incorporados al sistema HDFS (en el navegador localhost:9870 → Utilities → Browse the file system ...) y una captura de la salida del terminal en el que se han ejecutado los programas.

El resultado de la ejecución debe dar una salida cuyas primeras líneas se pueden ver a continuación para Amazon:

```
Una_Aventura_Impactante 52233
Una_Aventura_Impactante 0
Una_Aventura_Increíble 49741
Una_Aventura_Magnífica 54153
Una_Aventura_Magnífica 0
Una_Aventura_Magnífica 0
Una_Aventura_Mortal 51676
Una_Aventura_Mortal 0
Una_Aventura_Mortal 0
```

NOTA1: el mapper para esta tarea es sencillo. Una vez implementado, se puede comprobar su funcionamiento en el terminal:

```
$ cat join2_*.txt | ./join2_mapper.py | sort
```

NOTA2: el reducer será un poco más complejo, pero no debemos de perder de vista que a su entrada los datos van a estar ordenados alfabéticamente.

Aspectos a tener en cuenta:

- Se valorará que el código esté libre de errores y warnings y bien comentado y estructurado.
- El programa deberá ser subido al ejercicio de faitic (junto con cualquier otro fichero que se desee añadir) en formato .zip o .tar.gz. El nombre del fichero será **PR09bGR#.zip** ó **PR09bGR#.tar.gz** (# es el número del grupo). **Debe ser subido por todos los miembros del grupo de prácticas**.

Apartado C (optativo)

Para realizar este apartado, se facilita un fichero que contiene información sobre las ventas realizadas en una cadena de grandes almacenes en el mes de enero de 2012. Cada línea del fichero **`purchases.txt`** contiene los siguientes campos: fecha, hora, ciudad, sección, importe, medio de pago.

Se pide implementar los programas map/reduce que permitan responder las siguientes preguntas:

- ¿Cuál es el medio de pago más utilizado para la compra de *Computers*?
- Para cada medio de pago, ¿cuál es la sección que realiza el mayor importe de ventas?

Se debe adjuntar un pequeño documento pdf justificando brevemente la decisión que se ha tomado sobre el contenido de los campos <clave,valor> y explicando brevemente la implementación y los resultados. Aspectos a tener en cuenta:

- En el informe se valorará su calidad y las conclusiones a que se llegan. Además, se valorará el uso de gráficas que muestren los resultados obtenidos.
- El informe pdf junto con los programas deberán ser enviados por correo electrónico en formato .zip o .tar.gz. El nombre del fichero será **PR09c.zip** ó **PR09c.tar.gz**.
- **El apartado es optativo y será evaluado de manera individual.**