- MAP REDUCE FEST -

José Marcos Barreto
CI: 3819407-9
mbarreto@fing.edu.uy

Compilación

En todas las carpetas con los códigos se encuentra un archivo llamado “compile.sh”. Al ejecutar dicho archivo, compila todas las clases encontradas en “src” y genera el jar correspondiente, copiándolo a la misma carpeta que el archivo “compile.sh”.

Challenge 1

Básicamente se siguieron todos los pasos indicados en el pdf y se obtuvo como resultado un archivo “Friends3.txt” con la suma de palabras definidas en el archivo “Friends.txt” obtenida desde Facebook.

Challenge 2

1.1.2

Question 1:

1. *What does Reduce input groups mean? What can we do for changing this value?*

R - Son la cantidad de claves de entrada para la función reduce. Sería el conjunto de keys obtenida como resultado de aplicar la función map. Para cambiar ese valor se debería cambiar la funcion map a utilizar.

2. *What does Map input records mean? and, Map output records?*

R - Son la cantidad de registros de entrada para aplicarle la función MAP, el output es la cantidad de datos de grupos (key,value) de salida de la función map. En este caso, pueden haber varias parejas con mismo key (que conforman un grupo). La cantidad de keys diferentes es el *reduce input groups*.

3. *In your opinion how are Map input records and Map output records related?*

R - La función map utilizada es una función que mapea “*Map input records*” a “*Map output records*”. Esos dos conjuntos están relacionados por la función Map utilizada.

Question 2:

1. Which are the counters that let verify that the `combiner` worked?

   R – Se puede chequear el “combine input records”, “combine output records”. Se puede apreciar como se reduce el “reduce input groups”.

2. How can we estimate the advantage of using a `combiner`? Give its values when using and not using the combiner and justify your answer.

   R – Se puede estimar al verificar que se puedan tener muchos elementos en un mismo grupo como resultado de aplicar la operación map. Combiene utilizar un combiner por ejemplo en el ejemplo de conteo de palabras, que van a haber muchos resultados key,value de la operación map con la misma key y con valor 1. El combiner en este caso sirve para agregar resultados en memoria, obteniendo en vez de muchas parejas (key_i,1) obtener una con valor (key_i, sum(values)). Básicamente, cuando los "map output records" son mucho mayor que el "reduce input groups".
   No tiene sentido utilizar un combiner cuando los grupos "reduce input groups" y el "map output records" son valores similares, esto es, no hay muchas claves repetidas como resultado de aplicar la función map.

1.2 In-mapper combining

Question 3:

1. Se leyó el libro y se realizaron las modificaciones pertinentes.
2. Make sure that the new implementation works correctly. Which are the counters that let easily verify that everything worked fine? Which are the expected values?

R – El "combine input records" tiene valor 0, se puede verificar además que el "map output records" en este caso es igual al "reduce input groups". Esto significa que se generó un grupo por clave, ese grupo es el key de la palabra y el value es la cantidad de ocurrencias. El valor esperado era que se reduzcan la diferencia entre grupos y records, sin ver un cambio en los contadores del combinador pues esto es resuelto por el mapper. El resultado en este caso del mapper es optimo.

1.3 Analysing public data

Question 4: Se realizó la implementación indicada y es enviada adjunta en el mismo paquete.

Question 5: Se adjunta también el código que implementa esta solución. Para resolver el problema, se envia por cada valor una pareja "SUM" con valor el valor de la transacción. De esta manera el reducer recibirá todas las parejas SUM y samara sus valores, obteniendo la suma de todas las transacciones obteniendo así la suma total independiente de la categoria.

Question 6:

3- El código es modificado (enviado como un proyecto aparte), el reducer es el que calcula el promedio por categoría.

4 – No, porque el reducer calcula un resultado en base al grupo de keys generado por los mappers. Si se usa el reducer como combiner, se mezclaran valores de montos con promedios.

5 – La implementación fue realizada utilizando las indicaciones de la letra. Se implementó una clase que contiene un par "suma", "cantidad". Este valor es utilizado para realizar sumas parciales por categoría.

Question 7:

Inicialmente, compacté todas las planillas de los diferentes meses en un único archivo, concatenando el nombre del mes al final de cada línea, de la siguiente forma:

cat BedfordBoroughCouncil_April2010.csv | sed 's/$/:ABRIL/' >> BedfordBoroughCouncil.csv

Posteriormente, utilizando como *key* el nombre del mes, sumé todos los montos de todas las transacciones para dicho mes. Utilicé el Reducer como Combiner pues al ser sumas no afecta el resultado y mejora la eficiencia.

Question 8:

Implementé un *WritableComparable* para poder utilizar una clave combinada, que en este caso será "Mes" "Categoria".

Se modifican el Combiner y el Reducer para soportar esto, pero las clases son las mismas que en el ejercicio 1.3.5.

Challenge 3

Los datos de prueba se obtuvieron del siguiente link (http://www.clearbits.net/creators/146-stack-exchange-data-dump/contents). De donde se realizaron pruebas con corpus chicos por su gran tamaño.
Se obtuvo del libro el código del Mapper y del Reducer, se utiliza además el Reducer como Combiner para optimizar el algoritmo.

Para reducir el tamaño del corpus, se trabajó con una parte de archivos chicos descargados (los más chicos descomprimidos pesan en el entorno de los 2GB). Se realizó "cat stackoverflow-posts.xml > stackoverflow-posts-small.xml" y cortando la ejecución del cat luego de unos segundos, obteniendo así un archivo que tiene los primeros posts.

Se obtuvo un índice invertido como resultado de la ejecución del algoritmo map-reduce. Parte de la misma es:
http://en.wikipedia.org/w/index.php?title=accelerated_c%2b%2b&action=edit&redlink=1 81989
http://en.wikipedia.org/w/index.php?title=chromium&printable=yes 62020
http://en.wikipedia.org/w/index.php?title=html&oldid=248853176 256998
http://en.wikipedia.org/w/index.php?title=special%3asearch&search=list+of+companies&ns0=1&fulltext=search 227347
http://en.wikipedia.org/w/index.php?title=transactional_ntfs&oldid=233375400 150835