Alumno: **Bru Montes, Israel**

Edición **Octubre 2024 – Grupo A** a 23/12/2024

# 1. Ejercicio 1

(7,0 ptos) Dado un dataset que contenga entradas con la forma “persona;tienda;gasto”, crea un programa llamado mediaGastadoPersonaTienda que para cada persona indique su gasto medio por tienda, siguiendo el formato persona;tienda;gastomedio. Ejemplo:

Entrada Salida

Alice;Tienda1;50 Alice;Tienda1;75

Alice;Tienda2;20 Alice;Tienda2;20

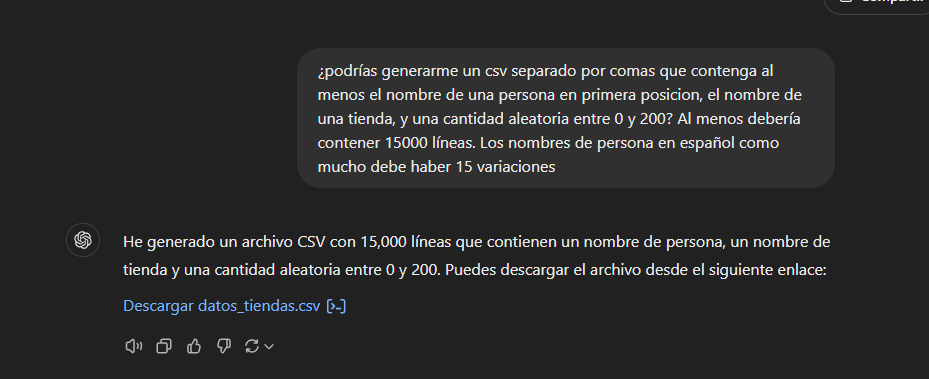
Bob;Tienda1;30 Bob;Tienda1;25

Alice;Tienda1;100

Bob;Tienda1;2

Generamos con ChatGPT un archivo csv para probar con las condiciones que se muestran en la siguiente captura:





Se genera este archivo para tener más datos.

En el archivo mapper, vamos a reducir esta lista a la 1a columna una combinación de “nombre-tienda” después el valor de compra.

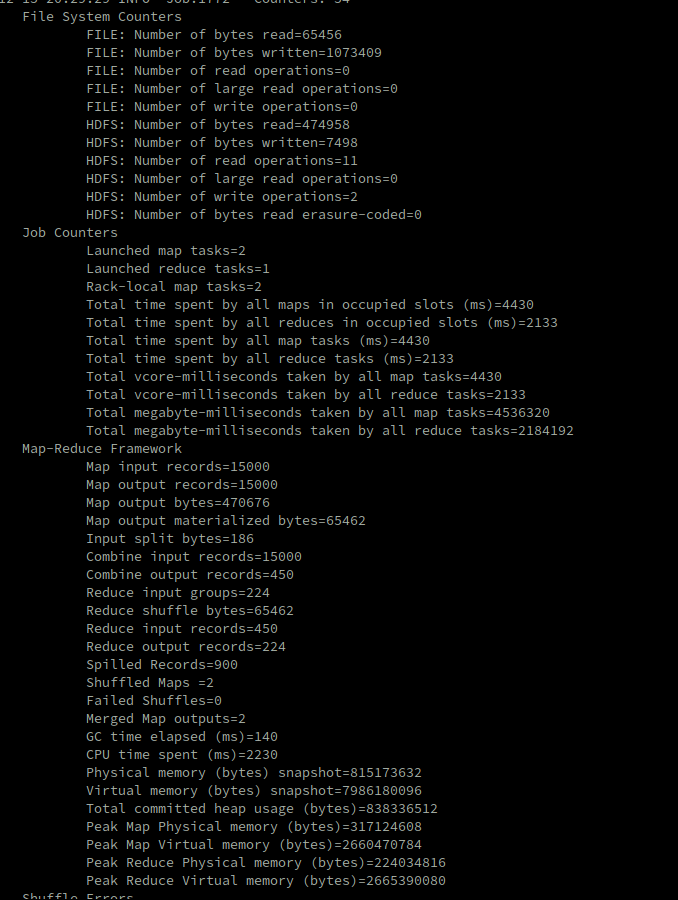
Con el resultado del mapper, el combiner nos va a construir una lista para esa clave cliente-tienda con todo el gasto que ha realizado en la tienda, separado por comas.

En el reducer, obtendremos todas las claves cliente-tienda y con leyenda la lista de gasto obtenida del combiner haremos la media y la volcaremos en el fichero de salida de datos.

[root@namenode hadoop]# cd input\_files/

root@namenode input\_files]# hadoop jar ../share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar -files mapper\_PTG.py,reducer\_PTG.py,combiner\_PTG.py -mapper mapper\_PTG.py -combiner combiner\_PTG.py -reducer reducer\_PTG.py -input dataset\_PTG.csv -output ptg\_out.txt

Y obtenemos como resultado obtenemos los siguientes datos en el terminal:



2. (7,0 ptos) Para este ejercicio utilizarán el fichero de entrada cite75\_99.txt que puede ser descargado del National Bureau of Economic Research (NBER) de EEUU (<http://www.nber.org/patents/>). Una descripción detallada de este fichero puede encontrarse en:

Hall, B. H., A. B. Jaffe, and M. Trajtenberg (2001). "The NBER Patent Citation Data File: Lessons, Insights and Methodological Tools." NBER Working Paper 8498.

Este fichero contiene citas de patentes emitidas entre 1975 y 1990 en los EEUU. Es un fichero CSV (comma-separated values) con más de 16,5 millones de filas, y las primeras líneas son como sigue:

"CITING","CITED"

3858241,956203

3858241,1324234

3858241,3398406

3858241,3557384

3858241,3634889

3858242,1515701

3858242,3319261

3858242,3668705

La estrategia a seguir sería, con el archivo Mapper invertir las filas, es decir, la 2a columna que contiene la patente referenciada, ponerla la primera, y la nueva patente que cita (1a columna) ponerla la 2a.

En combiner, como obtenemos ya la lista ordenada por las patentes referenciadas o citadas, agruparemos todas las patentes (lista) que citan una patente separadas por comas.

En el reducer, obtendremos todas las agrupaciones de combiner para una patente y las uniremos en una única lista de que patentes las nombran.

Ejecutamos los siguientes comandos:

[root@namenode hadoop]# cd input\_files/

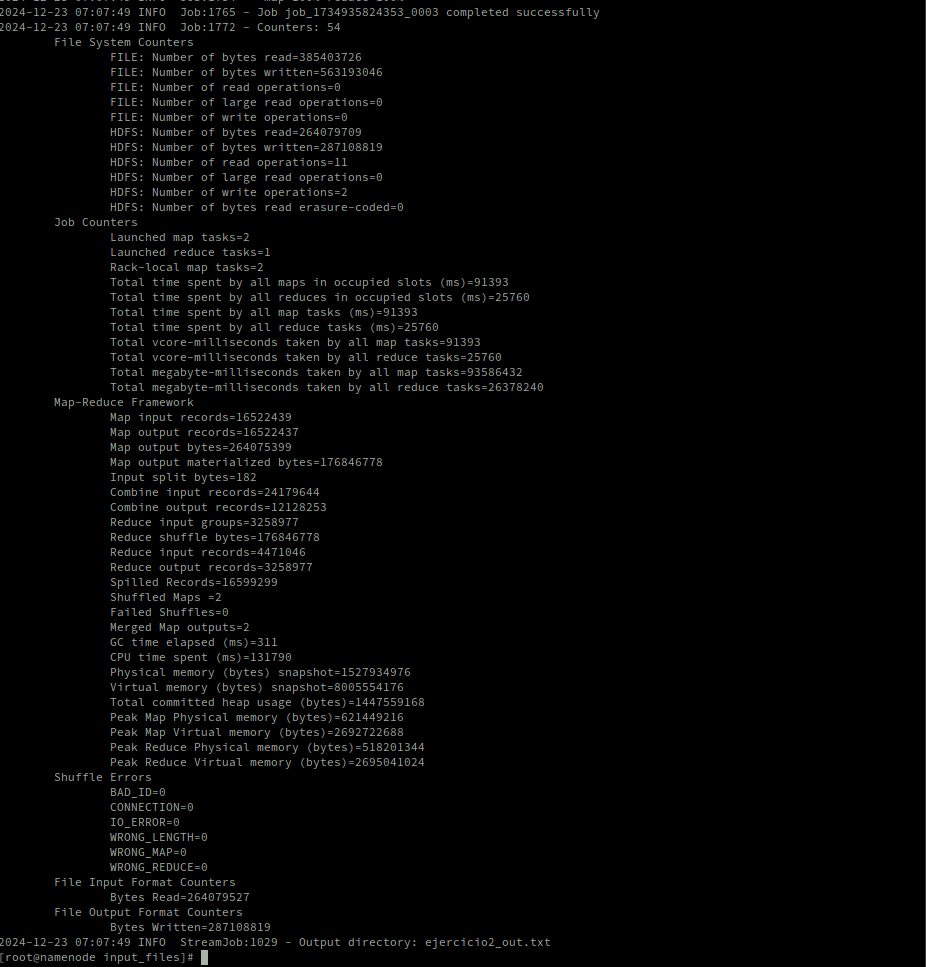
Copiamos a nuestra carpeta shared el archivo cite75\_99.txt y luego subimos en hdfs el archivo de datos para poder ejecutar.

[root@namenode input\_files]# hdfs dfs -put cite75\_99.txt

Ejecutamos el siguiente comando para ejecutar tanto el mapper como el reducer y volcamos la información en el archivo salida1.txt

[root@namenode input\_files]# hadoop jar ../share/hadoop/tools/lib/hadoop-streaming-3.4.1.jar -files mapper\_Cita.py,combiner\_Cita.py,reducer\_Cita.py -mapper mapper\_Cita.py -combiner combiner\_Cita.py -reducer reducer\_Cita.py -input cite75\_99.txt -output ejercicio2\_out.txt

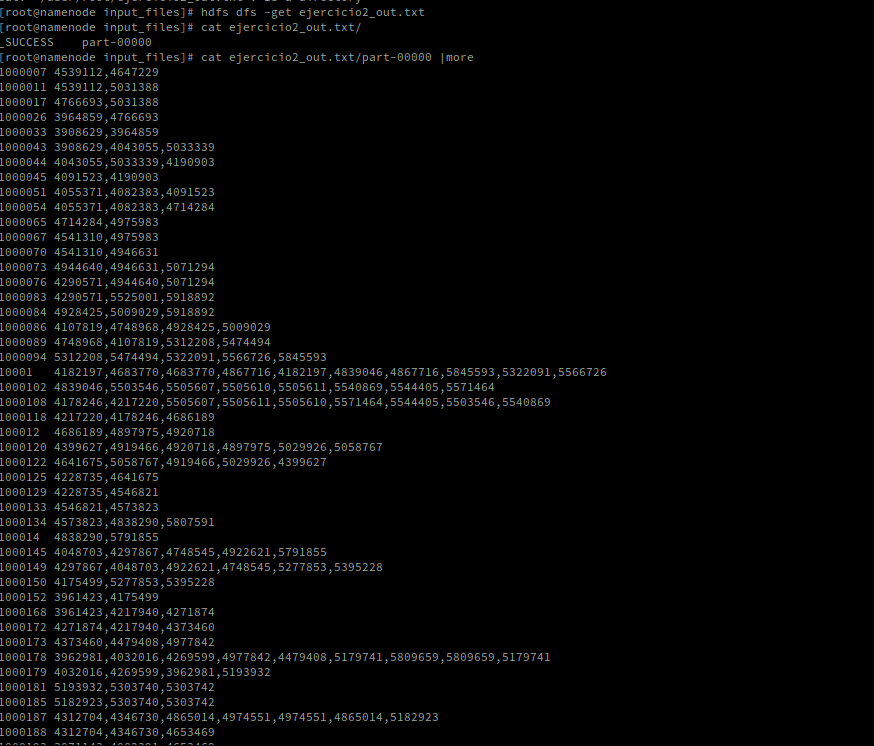
Y el resultado que obtenemos es el siguiente:



Nos bajamos de hdfs el archivo de salida para poder examinarlo (también con hdfs no es necesario bajarlo a la carpeta “input\_files”).

[root@namenode input\_files]# hdfs dfs -get ejercicio2\_out.txt .

[root@namenode input\_files]# cat ejercicio2\_out.txt/part-00000 |more



3. (6,0 ptos) Dado el siguiente programa Big Data que tiene defectos de diseño, se debe entender cuál es el defecto y hacer un informe que contenga: (1) Nombre y apellidos, (2) Tiempo empleado por el alumno en entender cuál es el defecto, y (3) Descripción del defecto. Todos los archivos del programa se encuentran en la carpeta: “PersonasQueCompranEnMuchasTiendas”

Nombre y apellidos: Israel Bru Montes

Tiempo: 10min.

Descripción del defecto:

El defecto se encuentra en el combiner, ya que por cada cliente cuenta las tiendas que ha visitado, pero realmente puede ocurrir que tanto mapper como combiner esten segmentados en diferente nodo, por lo que esta forma de “contar” no sería sobre el global de datos si no que sobre la información parcial que contiene el nodo o la segmentación de datos que procesa el combiner; este tipo de operaciones hay que hacerlas en el reducer, ya que obtiene la salida de todos los combiner.