Jorge Adrián Padilla Velasco
A00570894
31 de octubre de 2018

Práctica 3.1

Introducción

MapReduce es un modelo de programación para dar soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras y al commodity computing. El nombre del framework está inspirado en los nombres de dos importantes métodos, macros o funciones en programación funcional: Map y Reduce.

MapReduce ha sido adoptado mundialmente, ya que existe una implementación OpenSource denominada Hadoop.

MapReduce se emplea en la resolución práctica de algunos algoritmos susceptibles de ser paralelizados. No obstante, MapReduce no es la solución para cualquier problema, de la misma forma que cualquier problema no puede ser resuelto eficientemente por MapReduce. Por regla general se abordan problemas con datasets de gran tamaño, alcanzando los petabytes de tamaño. Es por esta razón por la que este framework suele ejecutarse en sistema de archivos distribuidos (HDFS).

Hadoop es un framework de software que soporta aplicaciones distribuidas bajo una licencia libre. Permite a las aplicaciones trabajar con miles de nodos y petabytes de datos. Hadoop se inspiró en los documentos Google para MapReduce y Google File System (GFS).

La función Map recibe como parámetros un par de (clave, valor) y devuelve una lista de pares. Esta función se encarga del mapeo y se aplica a cada elemento de la entrada de datos, por lo que se obtendrá una lista de pares por cada llamada a la función Map. Después se agrupan todos los pares con la misma clave de todas las listas, creando un grupo por cada una de las diferentes claves generadas. No hay requisito de que el tipo de datos para la entrada coincida con la salida y no es necesario que las claves de salida sean únicas.

Map (clave1, valor1) –> lista (clave2, valor2)

La función Reduce se aplica en paralelo para cada grupo creado por la función Map().La función Reduce se llama una vez para cada clave única de la salida de la función Map. Junto con esta clave, se pasa una lista de todos los valores asociados con la clave para que pueda realizar alguna fusión para producir un conjunto más pequeño de los valores.

Reduce (clave2, lista(valor2)) –> lista(valor2)

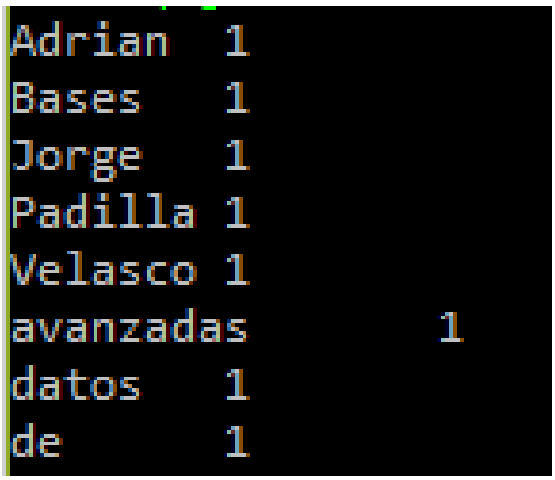
Desarrollo

En esta práctica, tuve problemas al momento de correr los ejemplos de los diferentes tutoriales que intenté seguir de Hadoop, tras instalarlo. Intenté cosas diferentes y en todas me marcaba errores extraños que no supe como arreglar.

Así que lo que hice fue pedirle ayuda a mi compañero Dominic Márquez. Él sí pudo correr los ejemplos y realizar la parte práctica necesaria para esta práctica. Así que con su ayuda, para la primera demostración utilicé un archivo de texto, el cual contenía las siguientes palabras:

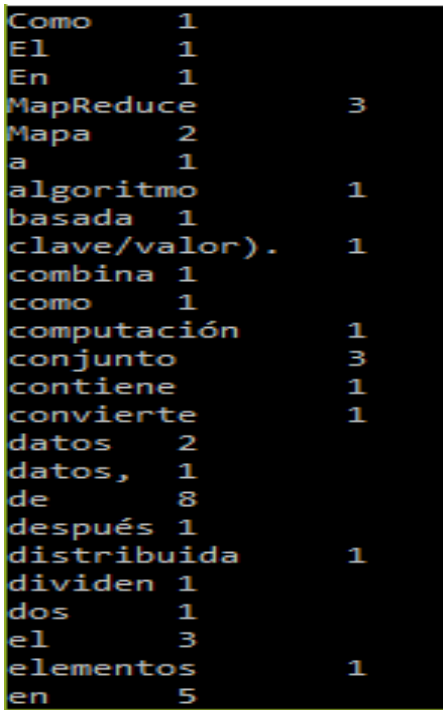
"Jorge Adrian Padilla Velasco
Bases de datos avanzadas"

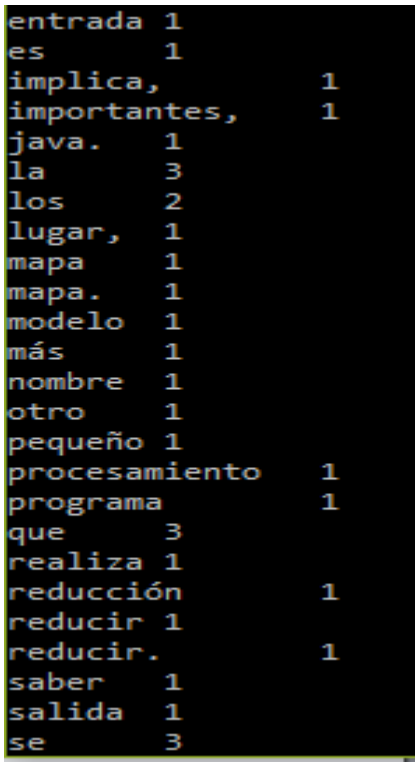
Obteniendo el siguiente resultado:

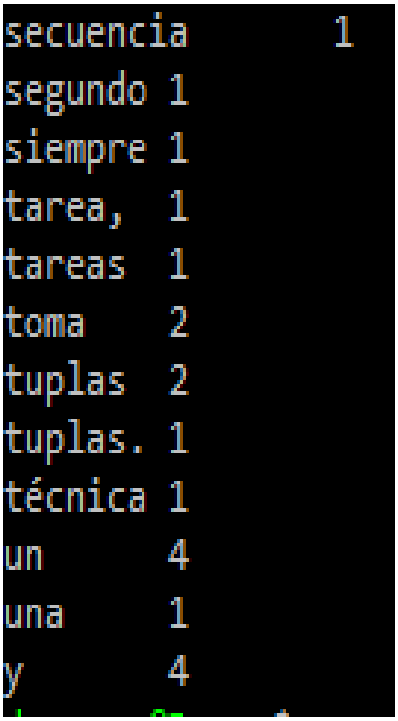


Para la segunda demostración utilice un archivo de texto, el cual contenía la definición de MapReduce, obtenida de la paǵina de *TutorialsPoint*.

Obteniendo el siguiente resultado:







Conclusiones

En la primera demostración, no se encontraron palabras repetidas, ya que es un archivo pequeño y simple que yo mismo creé. Por otra parte, para la segunda demostración se utilizó un archivo de texto más extenso. En este caso, podemos ver que se encontraron varias palabras duplicadas.

Referencias

https://es.wikipedia.org/wiki/MapReduce

https://es.wikipedia.org/wiki/Apache_Hadoop

http://blogs.solidq.com/es/big-data/que-es-mapreduce/

https://www.tutorialspoint.com/es/hadoop/hadoop_mapreduce.htm