Informe de prácticas: MapReduce

Introducción al Big Data Analytics(BDA)

Máster Universitario en Inteligencia Computacional e Internet de las Cosas

Universidad de Córdoba, EPSC

2022/2023

Autor:

Antonio Gómez Giménez (i72gogia@uco.es)

Índice:

1. Ejercicio A:

Queremos conocer la frecuencia de acceso a los distintos recursos (sean páginas php o archivos).

Teniendo en cuenta que los datos se encontrarán en un dataset dentro de un csv, habría que realizar los siguientes paso:

1. Extracción de los datos del csv y filtrado de aquellos valores que nos interesan. En este caso, teniendo en cuenta el problema presente, los datos a escoger serán la “URL”. Una vez tengamos todos los datos se dividirán teniendo en cuenta los rdd o particiones (depende de la tecnología usada), para su posterior mapeo.
2. Mapeo de los datos en clave-valor. Una vez tenemos los datos para cada nodo, se realiza el mapeo de los mismos donde se realiza un agregate, es decir, para cada par clave-valor se le incrementará en uno.
3. Shuffling y reduce. Tras realizar el mapeo, se realizará el shuffling y reduce, donde todas las combinaciones clave-valor se agruparán dando como resultado el total de veces que se ha accedido a una página o a un archivo.
4. Como esto no es lo que se pide, ya que se nos pide la frecuencia. Se realiza un nuevo map reduce. Donde primero se cambiará la clave valor, siendo ahora la clave el número de veces que se accede a un recurso.
5. Gracias al apartado anterior, realizamos una suma de todos los valores (key) para así poder calcular la frecuencia para cada uno. Finalmente para cada clave valor, se le realiza para cada key la división por el total calculado anteriormente, de esta forma, obtendremos como resultado el par clave valor pedido por el problema. La frecuencia de acceso a los distintos recursos (la frecuencia de cada URL).

2. Ejercicio B:

Para aquellas solicitudes a páginas php respondidas exitosamente, queremos saber cuántos accesos únicos (distintos clientes) ha tenido cada página.

Teniendo en cuenta que los datos se encontrarán en un dataset dentro de un csv, habría que realizar los siguientes paso:

1. Extracción de los datos del csv y filtrado de aquellos valores que nos interesan. En este caso, teniendo en cuenta el problema presente, los datos a escoger serán la “IP” y la “URL”, pero de la URL solo aquellas que satisfacen solicitudes exitosas, es decir, códigos 20x. Una vez tengamos todos los datos se dividirán teniendo en cuenta los rdd o particiones (depende de la tecnología usada), para su posterior mapeo.
2. Mapeo de los datos en clave-valor. Una vez tenemos los datos para cada nodo, se realiza el mapeo de los mismos donde se realiza un agregate, es decir, para cada par clave-valor se le incrementará en uno.
3. Shuffling y reduce. Tras realizar el mapeo, se realizará el shuffling y reduce, donde todas las combinaciones clave-valor se agruparán dando como resultado el total de veces que ha accedido una ip a una determinada página.
4. Como esto no es lo que se pide. Se realiza un nuevo map reduce. Donde primero se eliminan los valores que no nos interesan, en concreto el conteo de veces que una ip accede a una página. (También eliminaremos los ip, para poder agrupar así en por páginas)
5. Una vez tenemos los nuevos datos las páginas (repetidas para cada acceso de ip, por el caso anterior). Realizamos un map, donde se realiza un agregate, es decir, para cada par clave-valor se le incrementará en uno.
6. Shuffling y reduce. Tras realizar el mapeo, se realizará el shuffling y reduce, donde todas las combinaciones clave-valor se agruparán dando como resultado la cantidad de accesos de ip distintas para cada página.

3. Ejercicio C:

Para cada cliente, queremos saber a cuántas páginas php distintas ha accedido a lo largo del tiempo.

Este caso es similar al anterior, en vez de contar las ip como únicas, se cuenta como únicas las páginas php.

Teniendo en cuenta que los datos se encontrarán en un dataset dentro de un csv, habría que realizar los siguientes paso:

1. Extracción de los datos del csv y filtrado de aquellos valores que nos interesan. En este caso, teniendo en cuenta el problema presente, los datos a escoger serán la “IP” y la “URL”, pero de la URL solo aquellas que satisfacen solicitudes exitosas, es decir, códigos 20x. Una vez tengamos todos los datos se dividirán teniendo en cuenta los rdd o particiones (depende de la tecnología usada), para su posterior mapeo.
2. Mapeo de los datos en clave-valor. Una vez tenemos los datos para cada nodo, se realiza el mapeo de los mismos donde se realiza un agregate, es decir, para cada par clave-valor se le incrementará en uno.
3. Shuffling y reduce. Tras realizar el mapeo, se realizará el shuffling y reduce, donde todas las combinaciones clave-valor se agruparán dando como resultado el total de veces que ha accedido una ip a una determinada página.
4. Como esto no es lo que se pide. Se realiza un nuevo map reduce. Donde primero se eliminan los valores que no nos interesan, en concreto el conteo de veces que una ip accede a una página. (También eliminaremos los URL, para poder agrupar así en por IP)
5. Una vez tenemos los nuevos datos las ip (repetidas para cada acceso de ip, por el caso anterior). Realizamos un map, donde se realiza un agregate, es decir, para cada par clave-valor se le incrementará en uno.
6. Shuffling y reduce. Tras realizar el mapeo, se realizará el shuffling y reduce, donde todas las combinaciones clave-valor se agruparán dando como resultado la cantidad de páginas php distintas para cada ip.

4. Ejercicio D:

Queremos conocer la frecuencia de acceso de cada cliente a recursos de nuestro servidor.

Teniendo en cuenta que los datos se encontrarán en un dataset dentro de un csv, habría que realizar los siguientes paso:

1. Extracción de los datos del csv y filtrado de aquellos valores que nos interesan. En este caso, teniendo en cuenta el problema presente, los datos a escoger serán la “IP” y la “URL”. Una vez tengamos todos los datos se dividirán teniendo en cuenta los rdd o particiones (depende de la tecnología usada), para su posterior mapeo.
2. Mapeo de los datos en clave-valor. Una vez tenemos los datos para cada nodo, se realiza el mapeo de los mismos donde se realiza un agregate, es decir, para cada par clave-valor se le incrementará en uno.
3. Shuffling y reduce. Tras realizar el mapeo, se realizará el shuffling y reduce, donde todas las combinaciones clave-valor se agruparán dando como resultado el total de veces que ha accedido una ip a una determinada página.
4. Como esto no es lo que se pide, ya que se nos pide la frecuencia. Se realiza un nuevo map reduce. Donde primero se cambiará la clave valor, siendo ahora la clave el número de veces que se accede al servidor.
5. Gracias al apartado anterior, realizamos una suma de todos los valores (key) para así poder calcular la frecuencia para cada uno. Finalmente para cada clave valor, se le realiza para cada key la división por el total calculado anteriormente, de esta forma, obtendremos como resultado el par clave valor pedido por el problema. Es decir, la frecuencia de acceso de cada cliente a recursos de nuestro servidor (la frecuencia de cada ip a cada recurso).