MÁSTER EN BIG DATA Y DATA SCIENCE

03MBID Procesamiento de datos masivos

CURSO 2021-2022

Programa 1: Detectar problemas en un programa con el modelo de procesamiento MapReduce

Hecho por el estudiante Carlos de la Morena Coco

Tiempo empleado en entender el defecto del programa:

Un par de horas más o menos, más tiempo del que me gustaría reconocer la verdad.

Descripción del defecto:

El problema está en que no se tiene en cuenta que el mapper y el reducer pueden hacerse en distintas particiones en el caso de que tengamos muchos datos en el dataset.
En el caso de que la entrada sea muy grande (o le especifiquemos a hadoop que queremos que actúe de esta manera), diferentes particiones van a procesar los datos en paralelo en el mapper y en el combiner.
Pondré un ejemplo para explicarlo mejor.
Supongamos que tenemos el siguiente dataset:
Alice 10
Alice 20
Alice 10
Alice 5
Bob 5
Bob 25
Bob 25
Bob 10

Y supongamos que se analizarán en paralelo en dos particiones de forma que una partición reciba:
Alice 10 Alice 20 Bob 5 Alice 5

Y la otra partición:
Alice 10 Bob 25 Bob 25 Bob 10

En este caso, el combiner de la primera partición enviaría al reducer:
Alice 10 1
Alice 20 1
Bob 5 1
Alice 5 1
Y el de la segunda partición:
Alice 10 1
Bob 25 2
Bob 10

Y el reducer daría como respuesta que el gasto más común de Alice es 10 (una vez), y el de Bob es 25 (dos veces), lo cual no es cierto, ya que en este ejemplo Alice realiza el pago de 10 dos veces.

Para corregir este problema, había que añadir un apartado en el reducer que se dedique a sumar aquellos conteos de gastos por persona que llegan desde en combiner, para ya posteriormente hacer la comparación y devolver el resultado correspondiente.