



**Universidad
Internacional
de Valencia**

MÁSTER EN BIG DATA Y DATA SCIENCE

03MBID Procesamiento de datos masivos

CURSO 2021-2022

**Programa 2: 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:

Viniendo del primer programa, con este tardé media hora en ver lo que estaba mal.

Descripción del defecto:

El problema en este programa es que el combiner sólo manda al reducer los clientes que han comprado en más de 3 tiendas según los datos que llegan al propio combiner.

Si por la forma en la que está el programa estructurado llega una parte del dataset donde un cliente en particular compra en dos establecimientos a un combiner y al otro combiner llega otra parte del dataset donde este mismo cliente compra en otras dos tiendas distintas, estos datos nunca se emitirían al reducer.

Para solucionar este problema, podríamos dejar el combiner de tal forma que sirva únicamente para eliminar duplicados. Lo explico con un ejemplo.

Si nos llega a un combiner:

Bob HyM

Bob AhorraMás

Bob HyM

El combiner sólo emitiría al reducer:

Bob Hym

Bob Ahorramás.

Y el reducer se encargaría de aplicar el filtro y de emitir los clientes que hayan comprado en más de 2 tiendas distintas.