Sistemas BD: procesamiento batch

Sistemas de Big Data

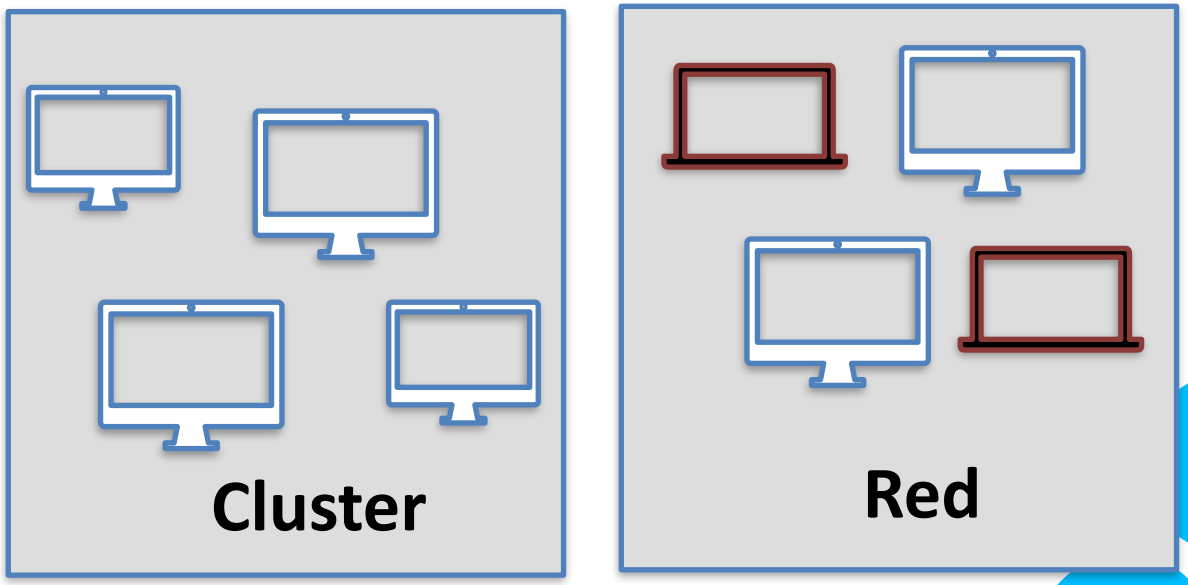
Ricardo García Ródenas
Ricardo.Garcia@uclm.es

Sistemas de Big_Data

Curso Especialización Inteligencia_Artificial y Big_Data

Procesamiento batch

- Recolección y almacenamiento datos **antes** de procesamiento
- Computación distribuida: **MapReduce**



MapReduce

Google 2004

- Máquinas procesamiento MapReduce
- Google File System (GFS)
- BigTable

MapReduce

Apache Software Foundation

- Hadoop MapReduce
- Hadoop YARN
- Hadoop Distributed File System (HDFS) y Hbase

MapReduce

Apache Software Foundation

- Entorno programación
- Escalabilidad
- Tolerante a fallos

MapReduce

Datos

- Registros < **clave**, **valor** >

$$< k, v >$$

MapReduce

Etapas

- Etapa Map

  Función

  map(k,v)

- Etapa Shuffle

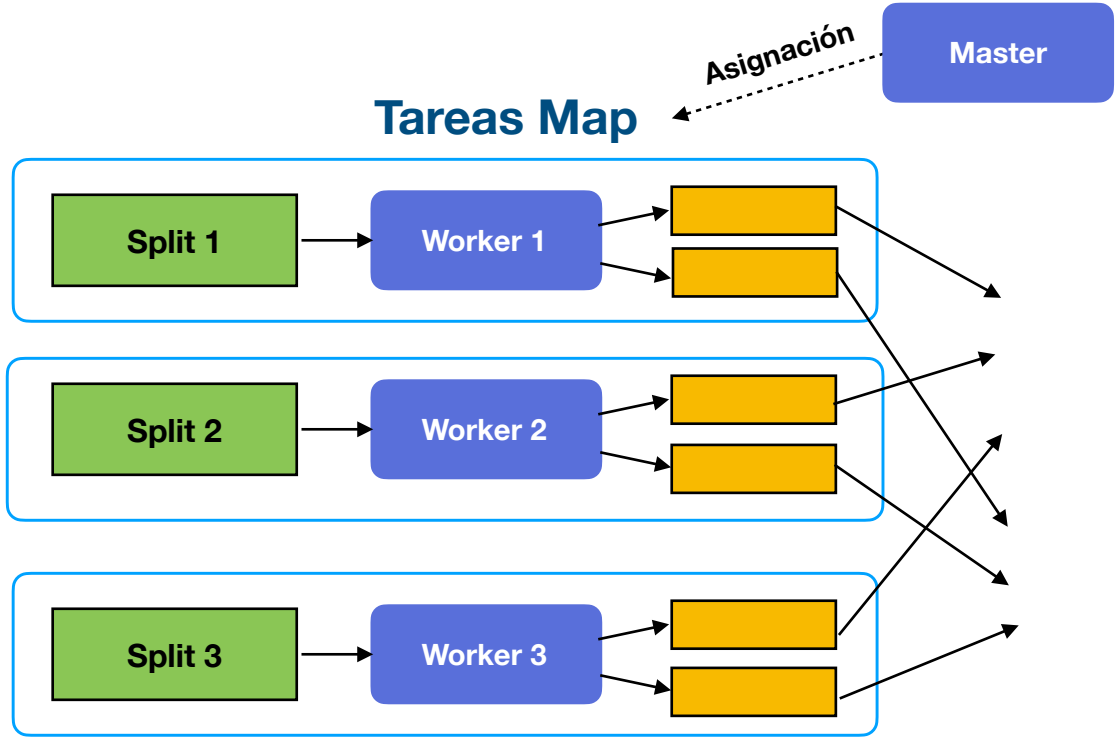
- Etapa Reduce

  Función

  reduce(k,list(v))

MapReduce

Etapa Map



Input

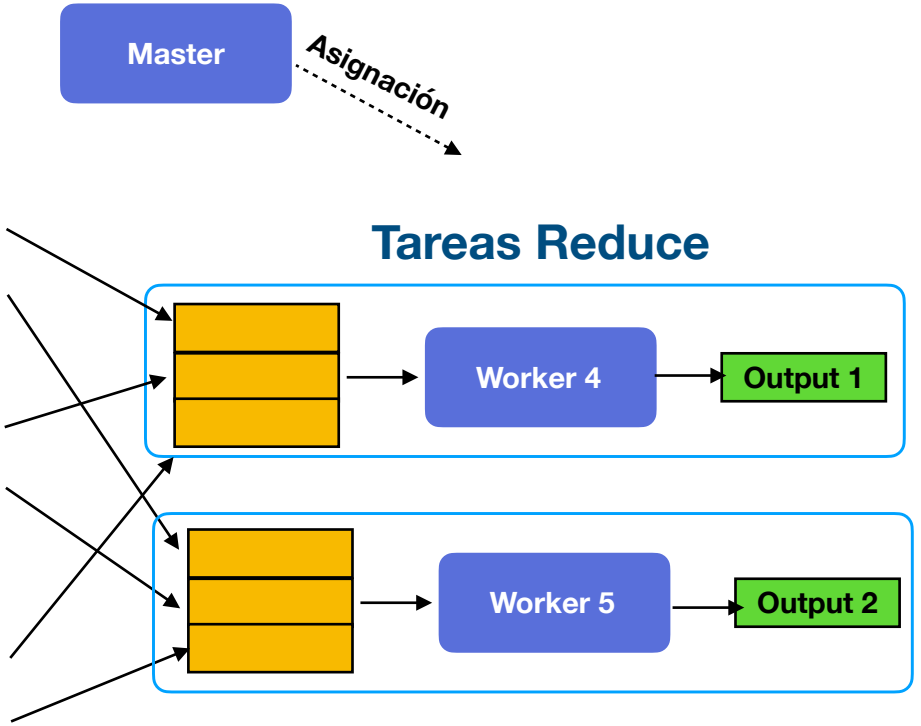
$< k_1, v_1 >$

Datos intermedios

$< k_2, v_2 >$

MapReduce

Etapa Reduce

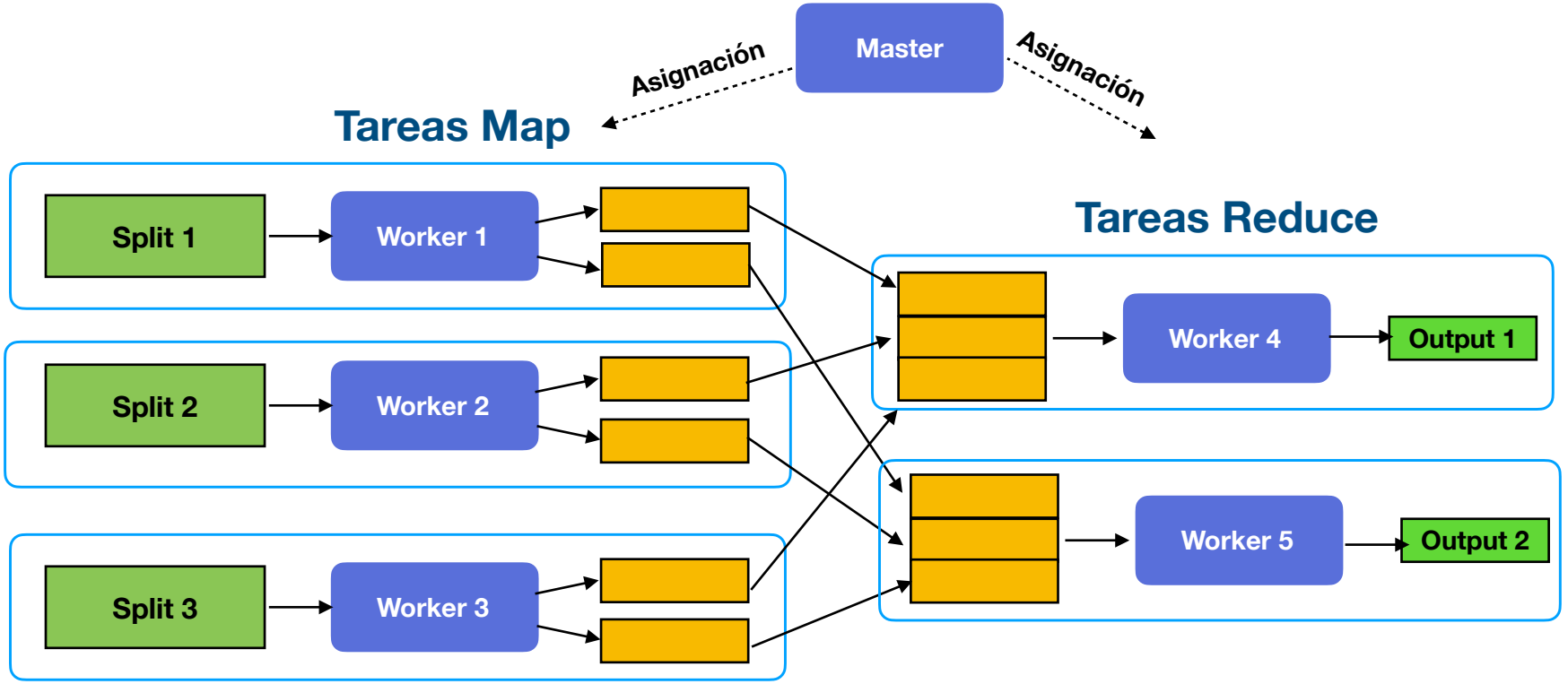


Datos intermedios

$< k_2, [v_2] >$

Output

$< k_3, v_3 >$



Input

$< k_1, v_1 >$

Datos intermedios

$< k_2, v_2 >$ $< k_2, [v_2] >$

Output

$< k_3, v_3 >$

Sistemas BD procesamiento batch

MapReduce

Sistemas de Big Data

Ricardo García Ródenas
Ricardo.Garcia@uclm.es

Sistemas de Big_Data

Curso Especialización Inteligencia_Artificial y Big_Data