MapReduce 簡介

Introduction to MapReduce

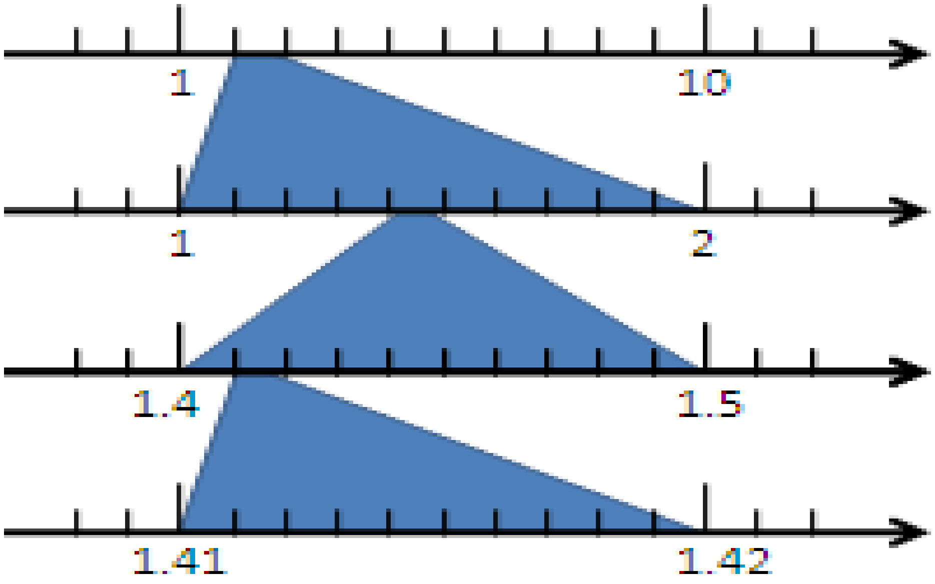
Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw

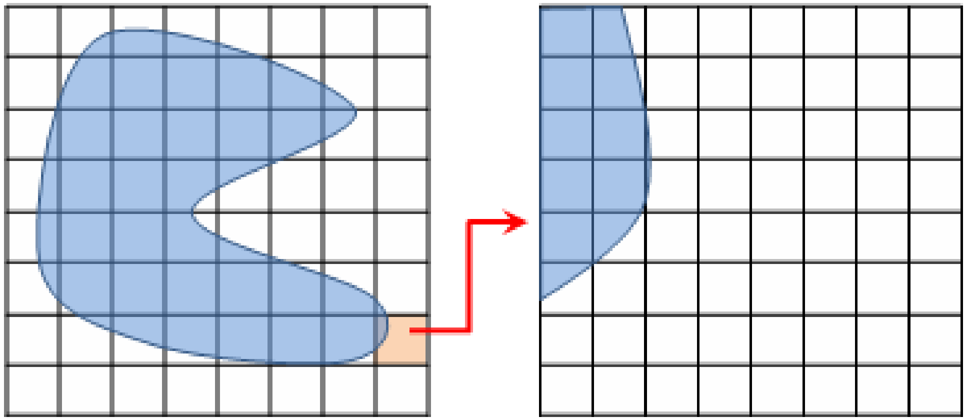
Divide and Conquer Algorithms
分而治之演算法

Example 1:

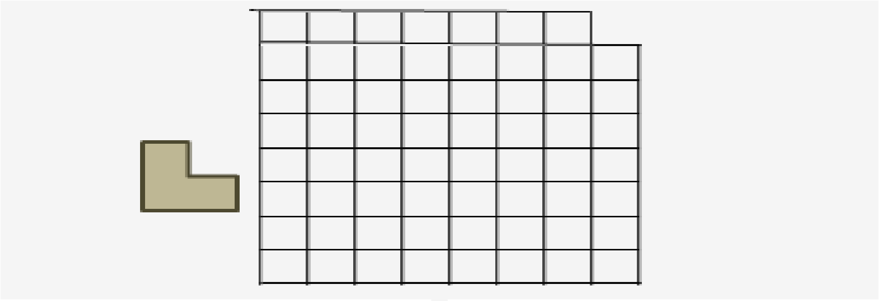


sqrt(2)

Example 2:



Example 3:



Example 4: The way to climb 5 steps stair within 2 steps each time. 眼前有五階樓梯，每次可踏上一階或踏上兩階，那麼爬完五階共有幾種踏法？

Ex：(1,1,1,1,1) or (1,2,1,1)

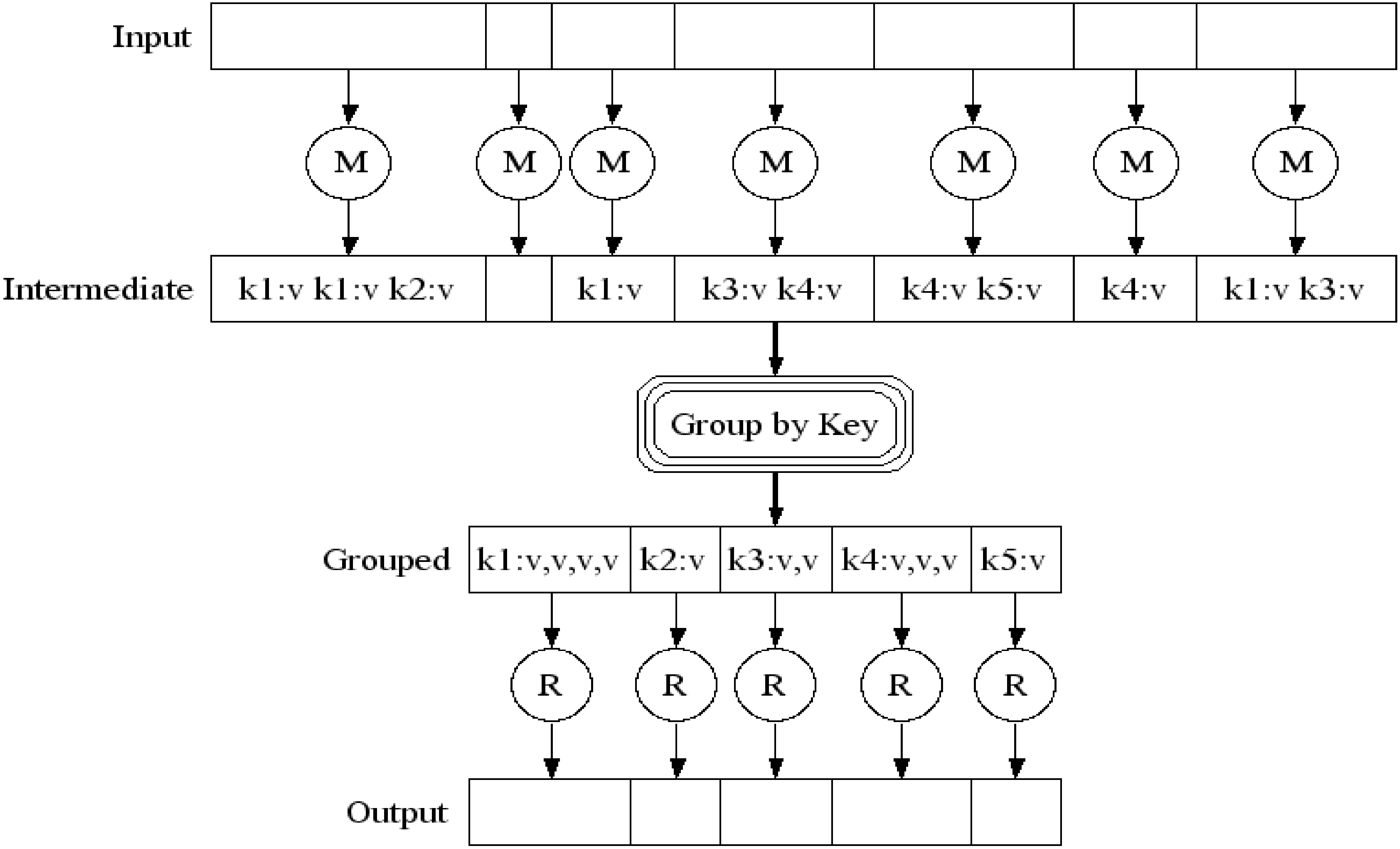
What is MapReduce ??
什麼是 MapReduce ??

- MapReduce 是 Google 申請的軟體專利，主要用來處理大量資料
- MapReduce is a patented software framework introduced by Google to support distributed computing on large data sets on clusters of computers.
- 啓發自函數編程中常用的 map 與 reduce 函數。
- The framework is inspired by map and reduce functions commonly used in functional programming, although their purpose in the MapReduce framework is not the same as their original forms
  - Map(...) : N → N
    - Ex. [1,2,3,4] – (*2) -> [2,4,6,8]
  - Reduce(...): N → 1
    - [1,2,3,4] - (sum) -> 10
- Logical view of MapReduce
    - Map(k1, v1) -> list(k2, v2)
    - Reduce(k2, list (v2)) -> list(k3, v3)

Source: http://en.wikipedia.org/wiki/MapReduce

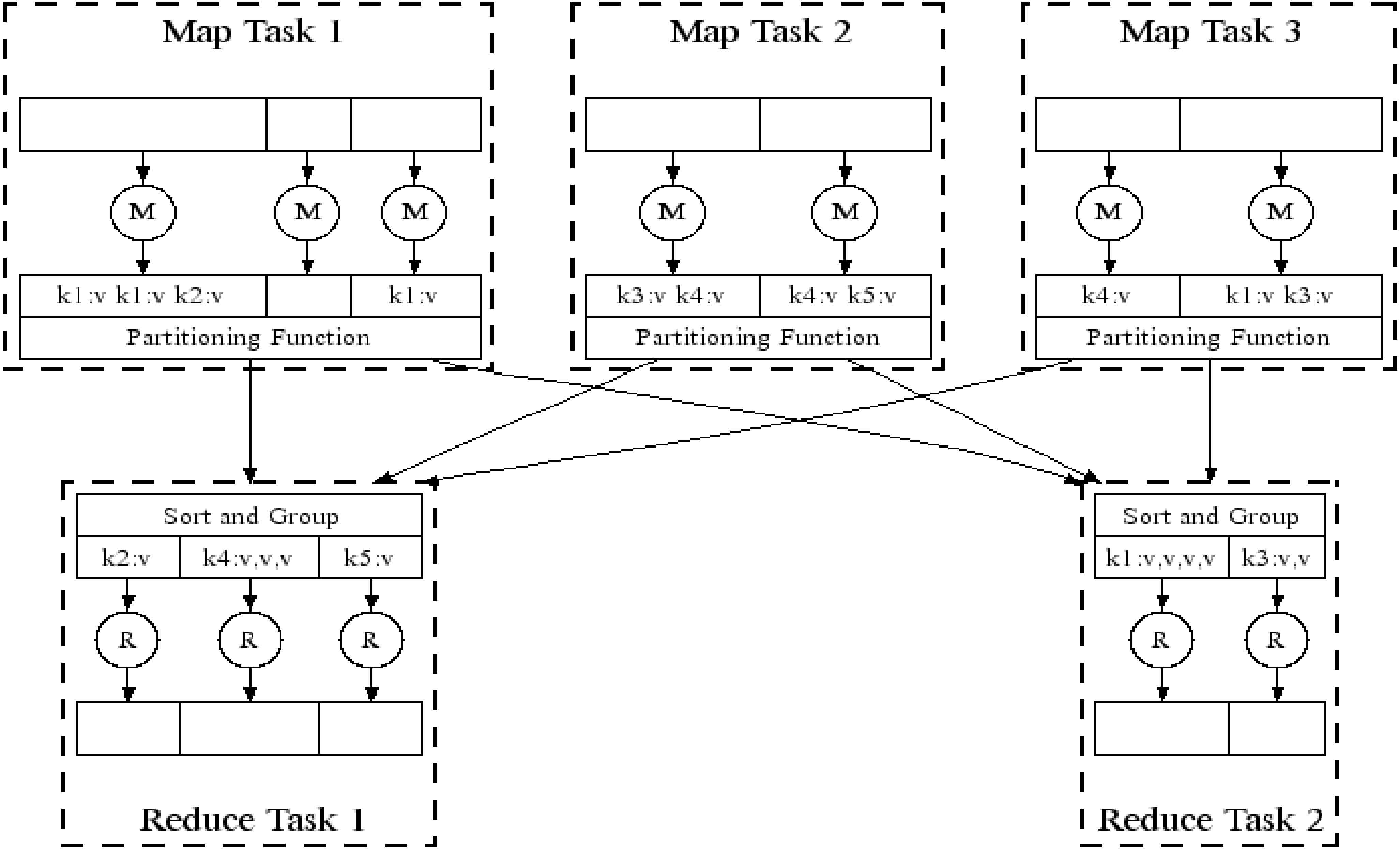
Google's MapReduce Diagram

Google 的 MapReduce 圖解

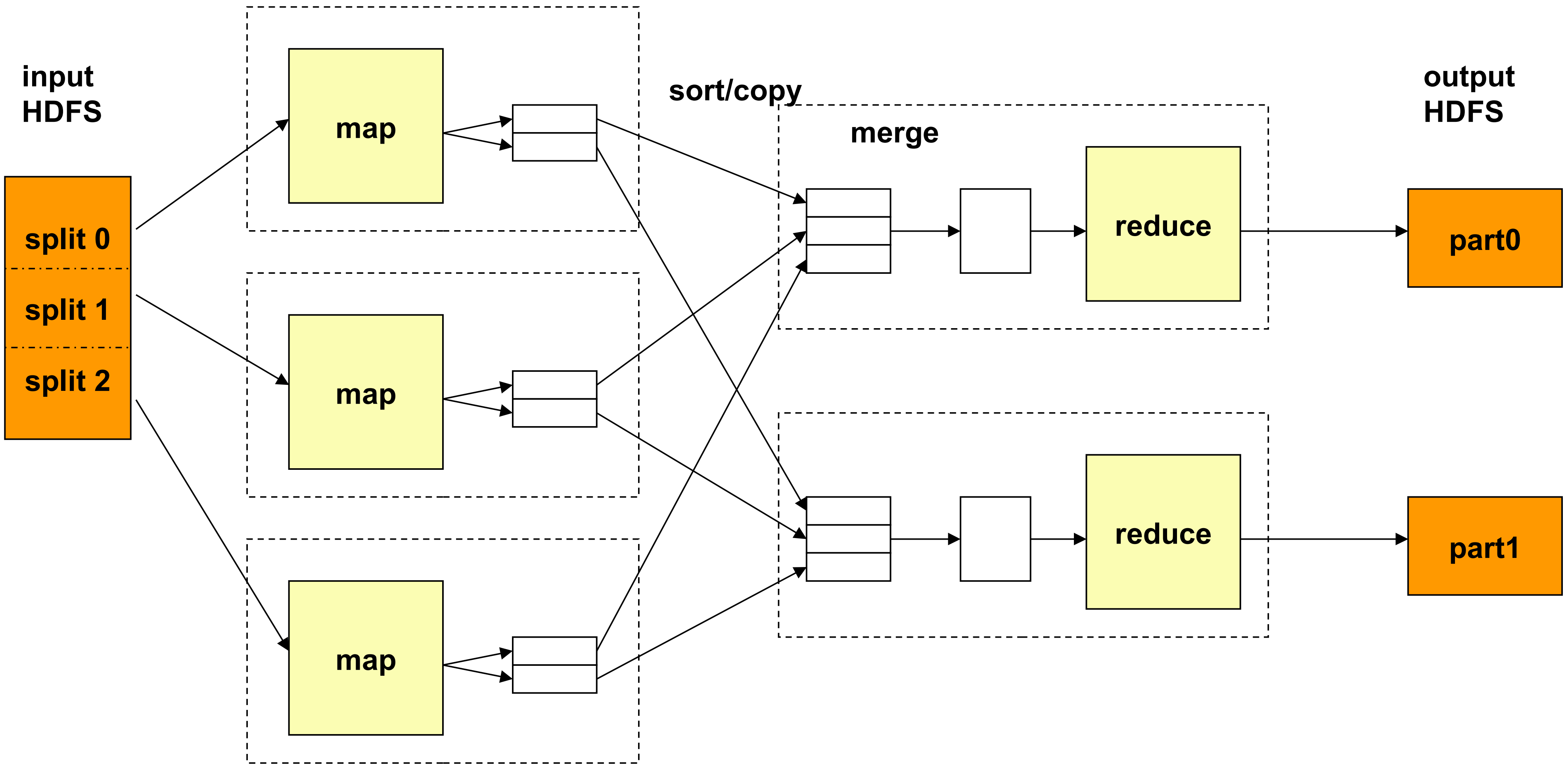


Google's MapReduce in Parallel

Google 的 MapReduce 平行版圖解



How does MapReduce work in Hadoop
Hadoop MapReduce 運作流程



JobTracker 跟 NameNode 取得需要運算的 blocks

JobTracker 選數個 TaskTracker 來作 Map 運算，產生些中間檔案

JobTracker 將中間檔案整合排序後，複製到需要的 TaskTracker 去

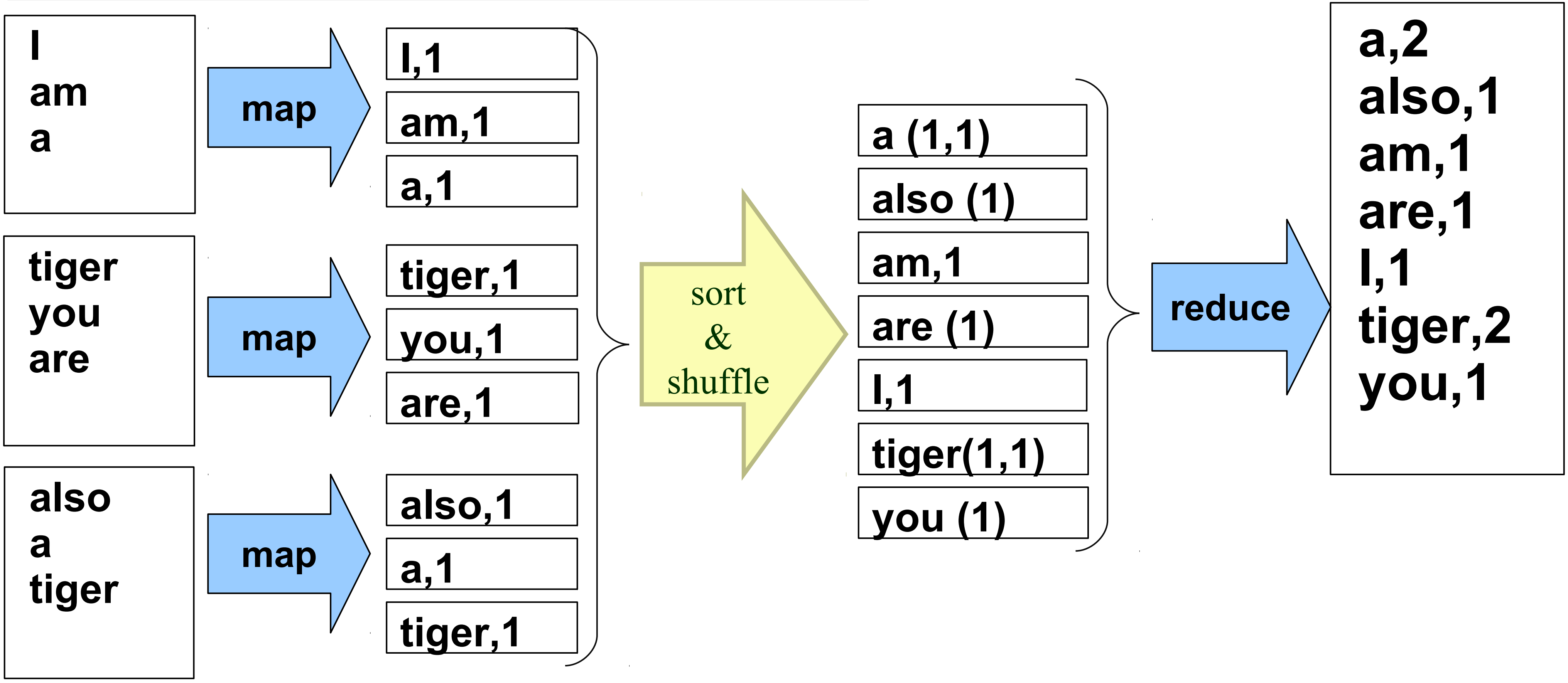
JobTracker 派遣 TaskTracker 作 reduce

reduce 完後通知 JobTracker 與 Namenode 以產生 output

MapReduce by Example (1)

MapReduce 運作實例 (1)

I am a tiger, you are also a tiger



I
am
a

map

I,1
am,1
a,1

tiger
you
are

map

tiger,1
you,1
are,1

also
a
tiger

map

also,1
a,1
tiger,1

sort
&
shuffle

a (1,1)
also (1)
am,1
are (1)
I,1
tiger(1,1)
you (1)

reduce

a,2
also,1
am,1
are,1
I,1
tiger,2
you,1

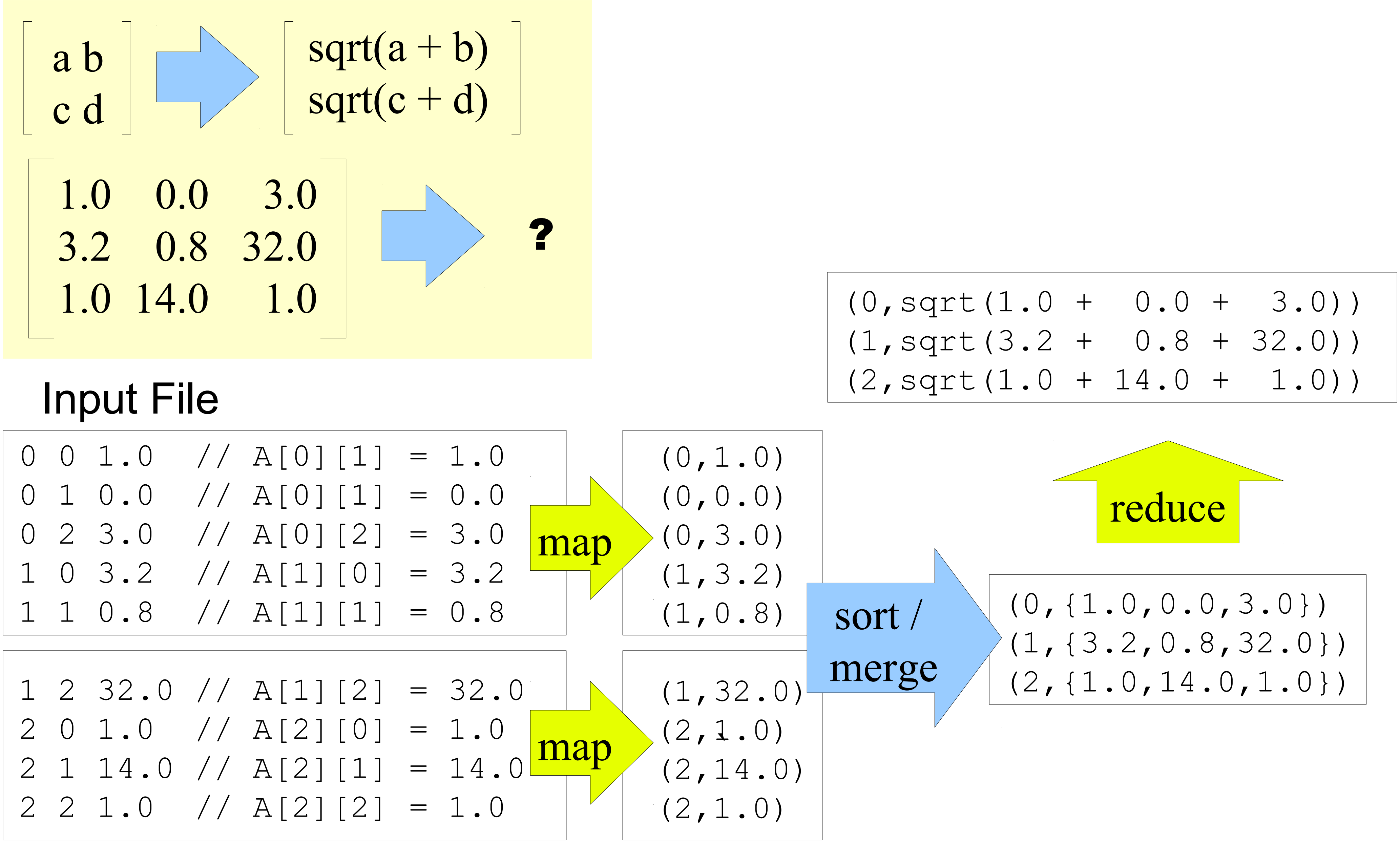
JobTracker 先選了三個 Tracker 做 map

Map 結束後， hadoop 進行中間資料的重組與排序

JobTracker 再選一個 TaskTracker 作 reduce

MapReduce by Example (2)

MapReduce 運作實例 (2)



$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \Rightarrow \begin{bmatrix} \mathrm{sqrt}(a + b) \\ \mathrm{sqrt}(c + d) \end{bmatrix}$$

$$\begin{bmatrix} 1.0 & 0.0 & 3.0 \\ 3.2 & 0.8 & 32.0 \\ 1.0 & 14.0 & 1.0 \end{bmatrix} \Rightarrow ?$$

Input File

```
0 0 1.0   // A[0][1] = 1.0
0 1 0.0   // A[0][1] = 0.0
0 2 3.0   // A[0][2] = 3.0
1 0 3.2   // A[1][0] = 3.2
1 1 0.8   // A[1][1] = 0.8
```

```
1 2 32.0 // A[1][2] = 32.0
2 0 1.0  // A[2][0] = 1.0
2 1 14.0 // A[2][1] = 14.0
2 2 1.0  // A[2][2] = 1.0
```

```
(0,1.0)
(0,0.0)
(0,3.0)
(1,3.2)
(1,0.8)
```

```
(1,32.0)
(2,1.0)
(2,14.0)
(2,1.0)
```

```
(0,{1.0,0.0,3.0})
(1,{3.2,0.8,32.0})
(2,{1.0,14.0,1.0})
```

```
(0,sqrt(1.0 +  0.0 +  3.0))
(1,sqrt(3.2 +  0.8 + 32.0))
(2,sqrt(1.0 + 14.0 +  1.0))
```

MapReduce is suitable to ....

MapReduce 合適用於 ....

- 大規模資料集
- Large Data Set
- 可拆解
- Parallelization

- Text tokenization
- Indexing and Search
- Data mining
- machine learning
- …

- http://www.dbms2.com/2008/08/26/known-applications-of-mapreduce/
- http://wiki.apache.org/hadoop/PoweredBy