



財團法人國家實驗研究院

國家高速網路與計算中心

NATIONAL CENTER FOR HIGH-PERFORMANCE COMPUTING

Map Reduce 介紹



王耀聰 陳威宇 楊順發

jazz@nchc.org.tw

waue@nchc.org.tw

shunfa@nchc.org.tw

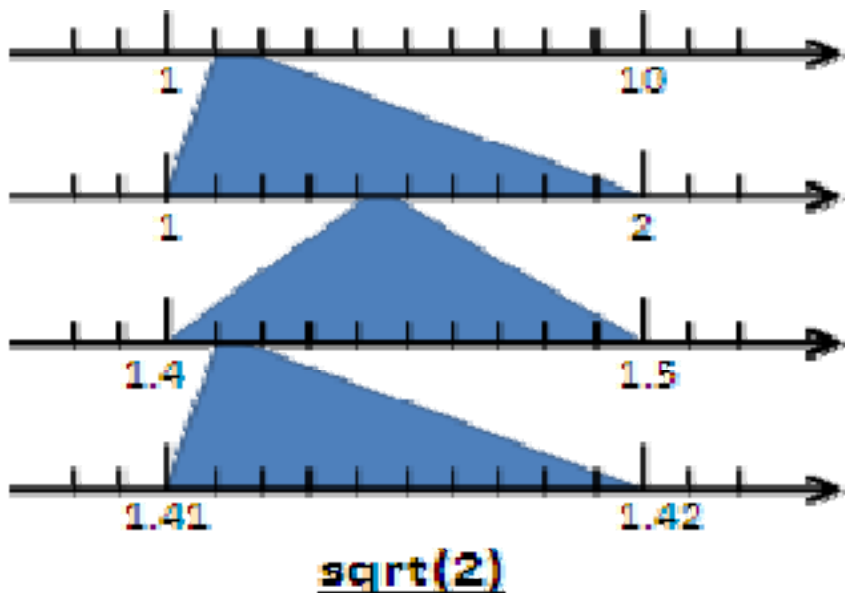
國家高速網路與計算中心(NCHC)



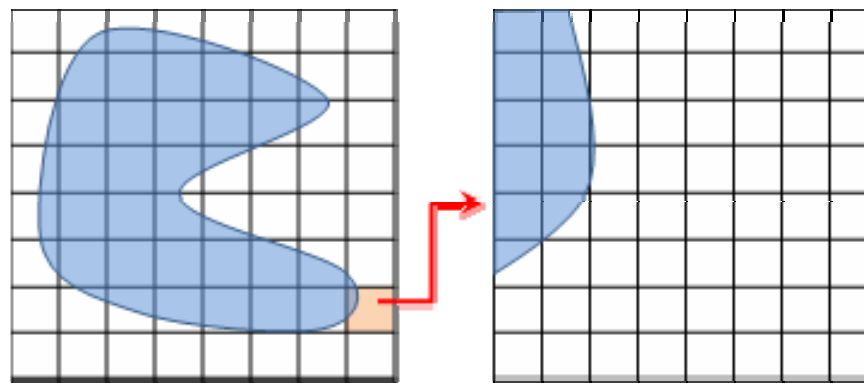
自由軟體實驗室

Divide and Conquer

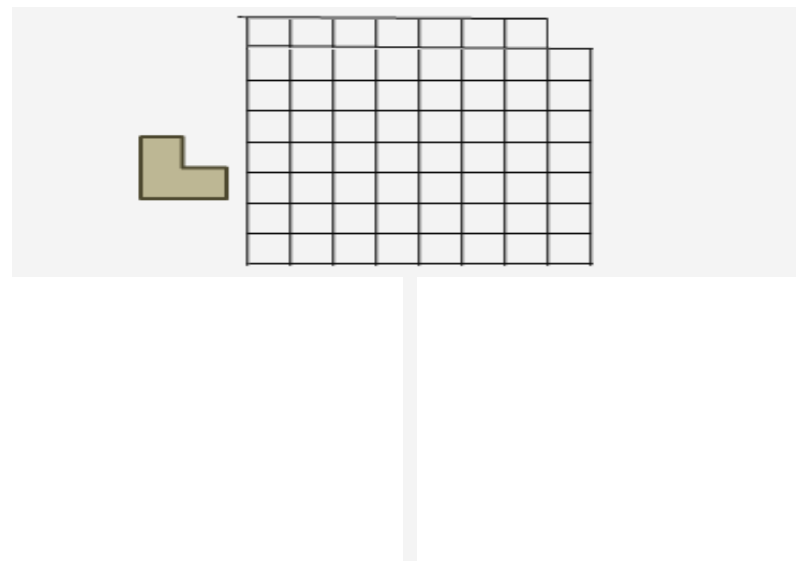
範例一：十分逼近法



範例二：方格法求面積



範例三：鋪滿 L 形磁磚



範例四：

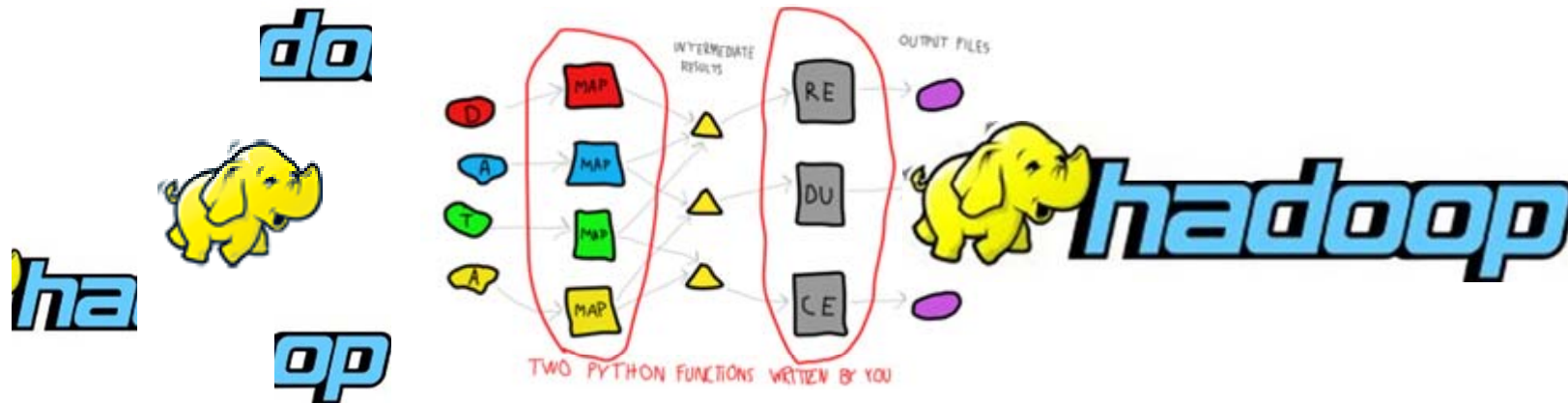
眼前有五階樓梯，每次可踏上一階或踏上兩階，那麼爬完五階共有幾種踏法？

Ex : (1,1,1,1,1) or (1,2,1,1)

Map Reduce 起源

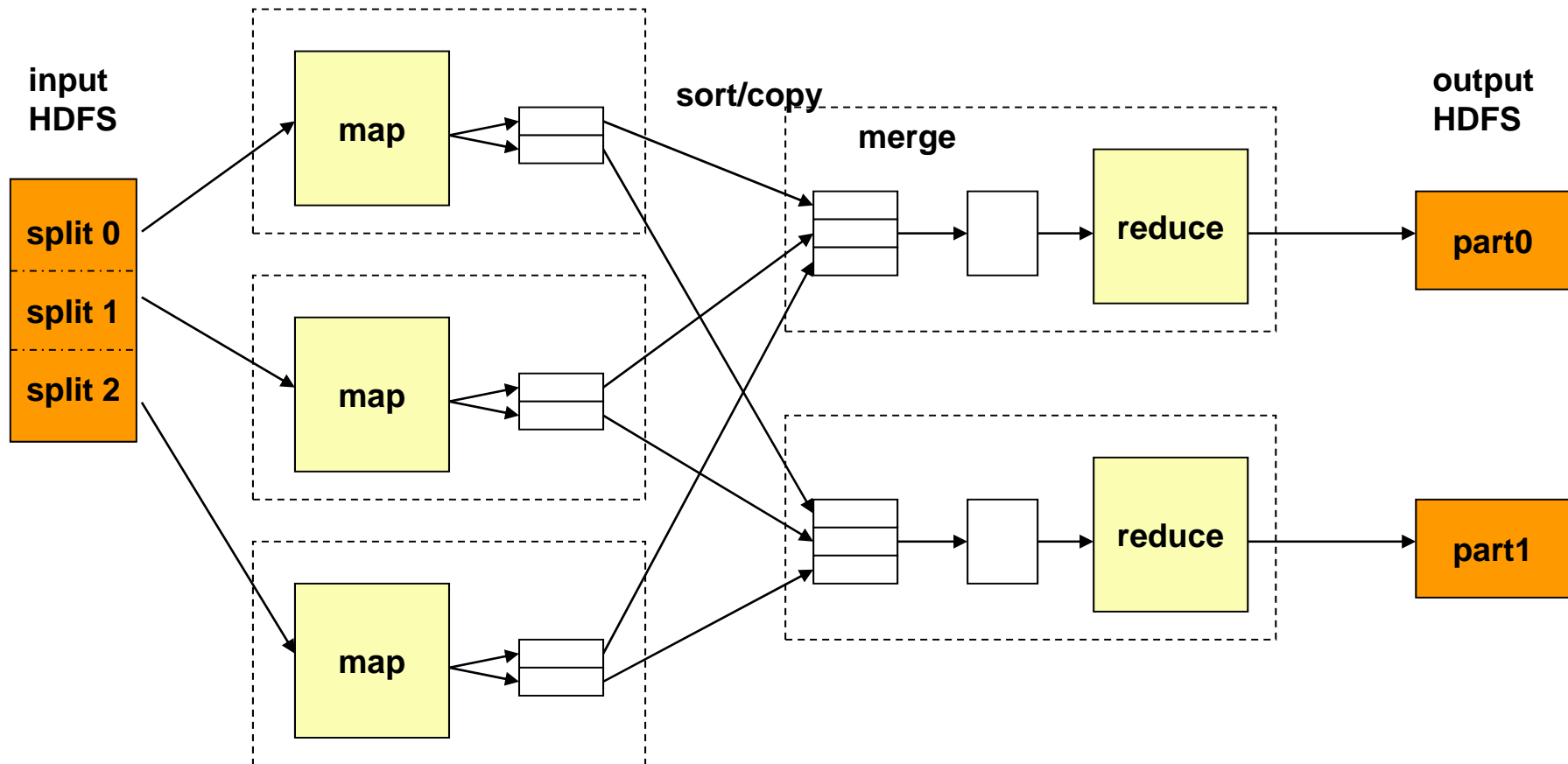
- Functional Programming : Map Reduce
 - map(...) :
 - [1,2,3,4] - (*2) -> [2,4,6,8]
 - reduce(...):
 - [1,2,3,4] - (sum) -> 10
- 演算法 (Algorithms) :
 - Divide and Conquer
 - 分而治之
- 在程式設計的軟體架構內，適合使用在大規模數據的運算中

Hadoop MapReduce 定義



Hadoop Map/Reduce是一個易於使用的軟體平台，以MapReduce為基礎的應用程序，能夠運作在由上千台PC所組成的大型叢集上，並以一種可靠容錯的方式平行處理上P級別的資料集。

Hadoop-MapReduce 運作流程



JobTracker跟NameNode取得需要運算的blocks

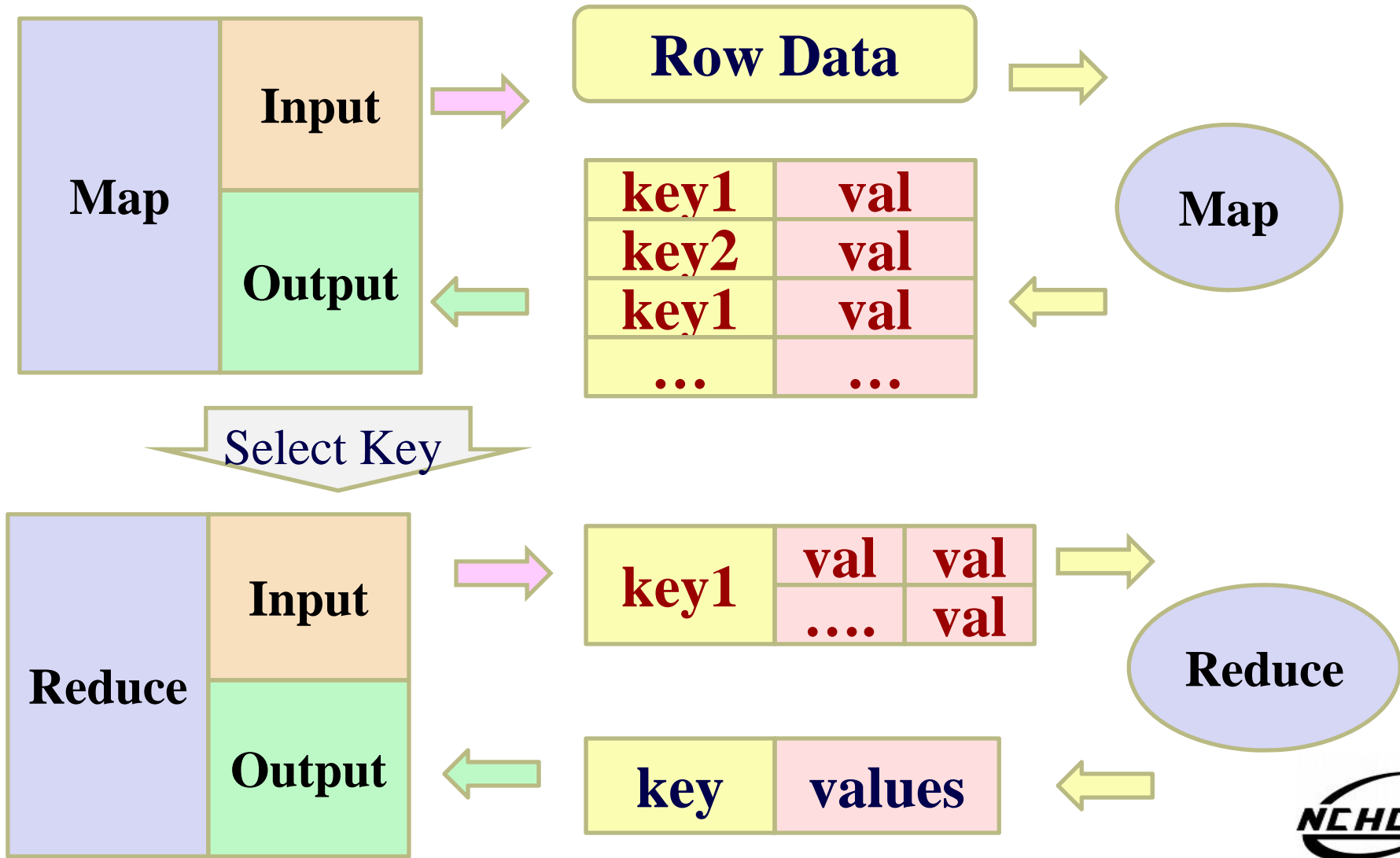
JobTracker選數個TaskTracker來作Map運算，產生些中間檔案

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

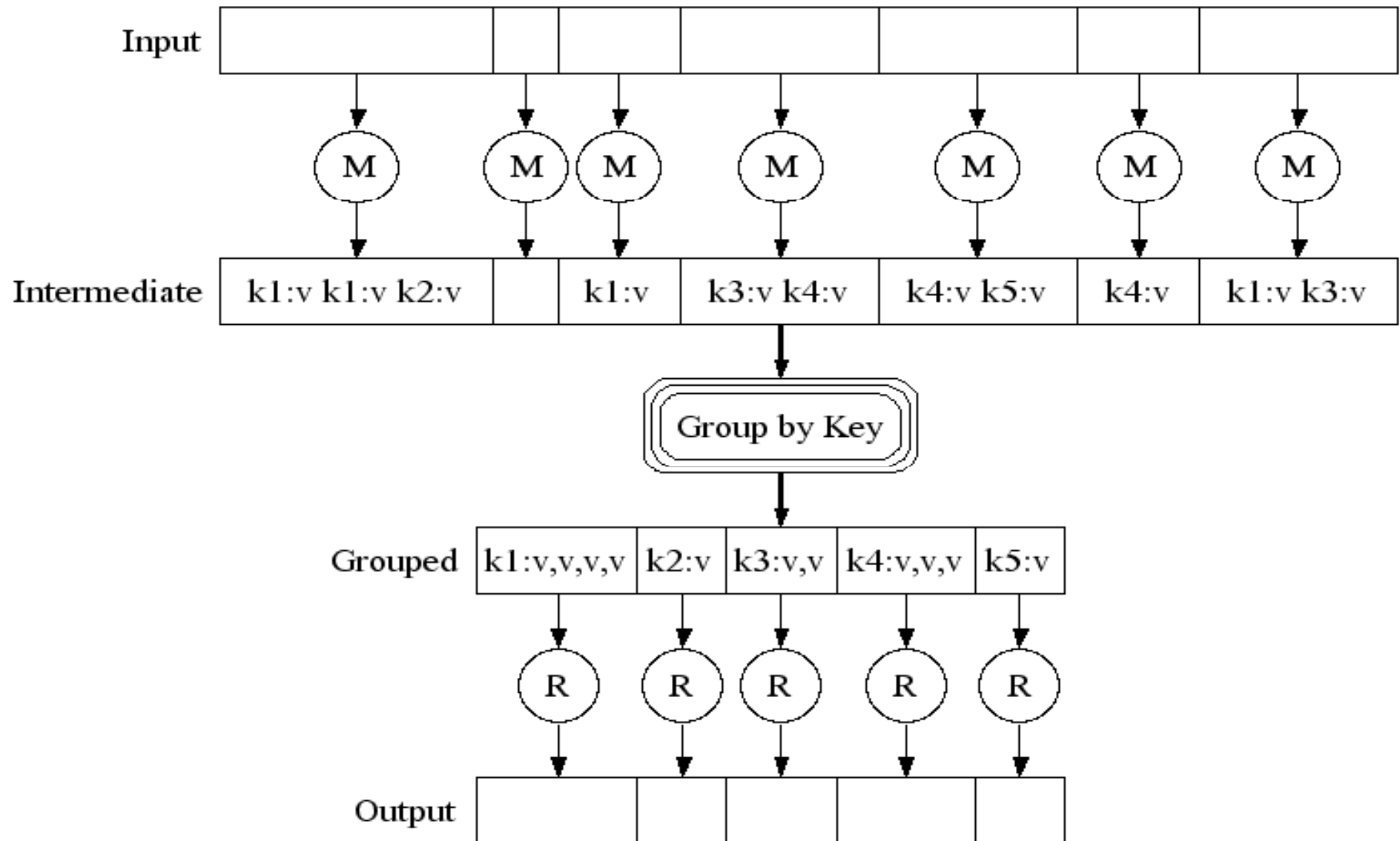
JobTracker派遣TaskTracker作reduce

reduce完後通知JobTracker與NameNode以產生output

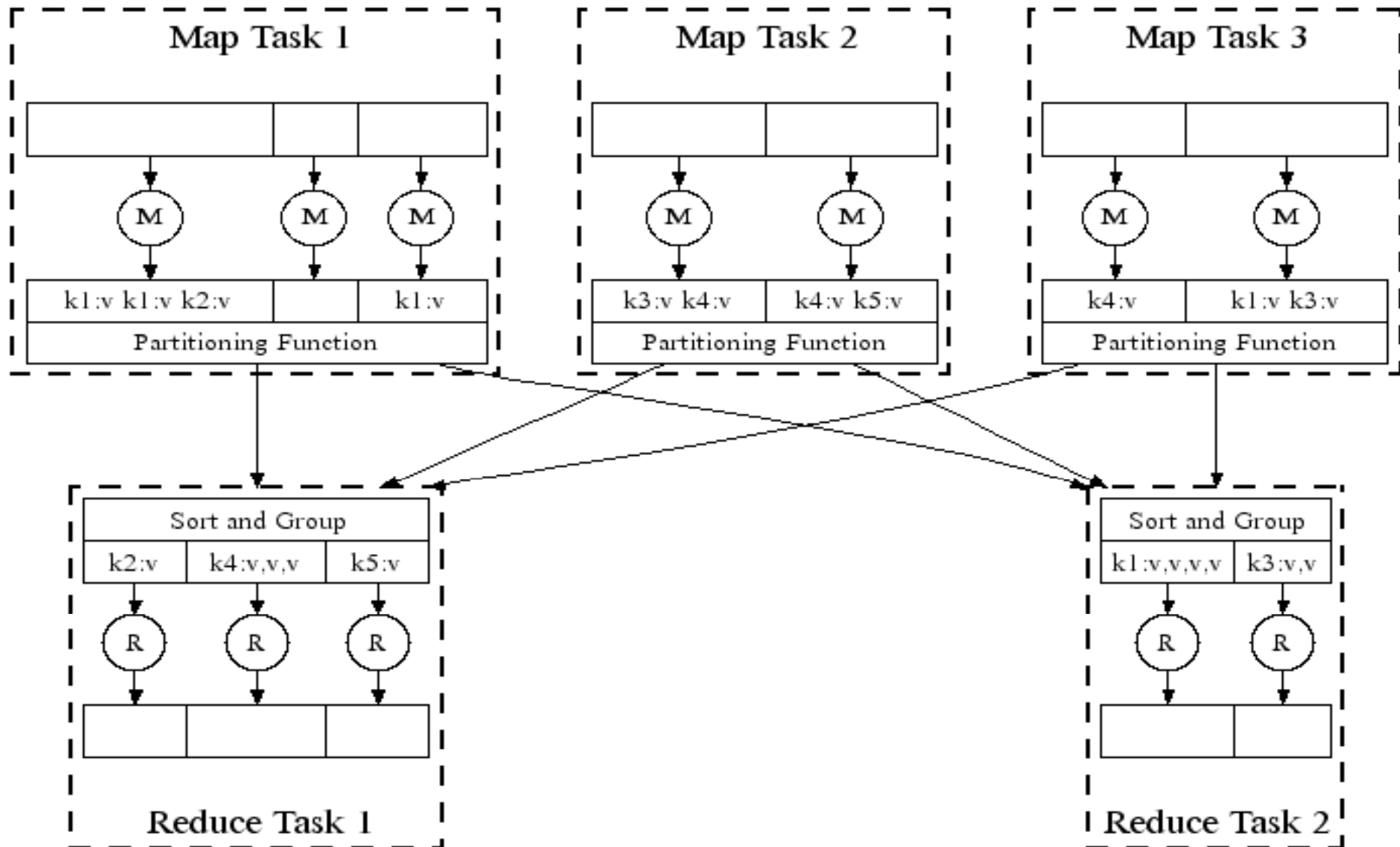
MapReduce 與 $\langle \text{Key}, \text{Value} \rangle$



MapReduce 圖解

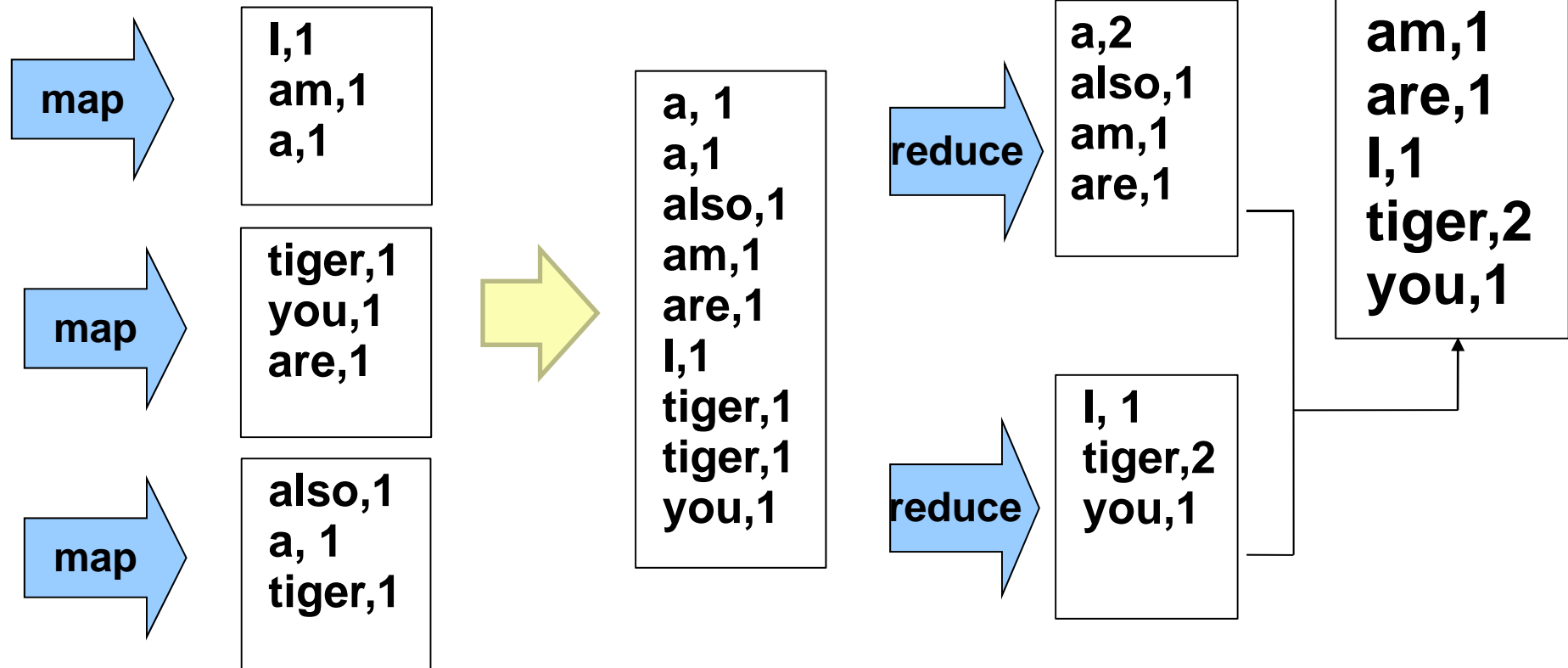


MapReduce in Parallel



範例

I am a tiger, you are also a tiger



JobTracker先選了三個
Tracker做map

Map結束後，hadoop進行
中間資料的整理與排序

JobTracker再選兩個
TaskTracker作reduce

Hadoop適用於..

- 大規模資料集
- 可拆解
 - 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>

Hadoop Applications (1)

- Adobe
 - use Hadoop and HBase in several areas from **social services** to structured data storage and processing for **internal use**.
- Adknowledge - Ad network
 - used to build the recommender system for **behavioral targeting**, plus other **clickstream analytics**
- Alibaba
 - processing **sorts of business data** dumped out of database and joining them together. These data will then be fed into **iSearch**, our vertical search engine.
- AOL
 - We use hadoop for variety of things ranging from **ETL style processing** and **statistics generation** to running advanced algorithms for doing **behavioral analysis**

Hadoop Applications (2)

- Baidu - the leading Chinese language search engine
 - Hadoop used to analyze the **log of search and do some mining** work on web page database
- Contextweb - ADSDAQ Ad Exchange
 - use Hadoop to store ad serving log and use it as a source for **Ad optimizations/Analytics/reporting/machine learning**.
- Detikcom - Indonesia's largest news portal
 - use hadoop, pig and hbase to analyze **search log, generate Most View News,**
 - generate top **wordcloud**, and analyze all of our **logs**

Hadoop Applications (3)

- DropFire
 - generate **Pig Latin** scripts that describe structural and semantic conversions between data contexts
 - use Hadoop to **execute these scripts** for production-level deployments
- Facebook
 - use Hadoop to store copies of internal log and dimension data sources
 - use it as a source for reporting/analytics and machine learning.
- Freestylers - Image retrieval engine
 - use Hadoop 影像處理
- Hosting Habitat
 - 取得所有clients的軟體資訊
 - 分析並告知clients 未安裝或未更新的軟體

Hadoop Applications (4)

- IBM
 - Blue Cloud Computing Clusters
- ICCS
 - 用 Hadoop and Nutch to crawl Blog posts 並分析之
- IIT, Hyderabad
 - We use hadoop 資訊檢索與提取
- Journey Dynamics
 - 用 Hadoop MapReduce 分析 billions of lines of GPS data 並產生交通路線資訊.
- Krugle
 - 用 Hadoop and Nutch 建構 原始碼搜尋引擎

Hadoop Applications (5)

- SEDNS - Security Enhanced DNS Group
 - 收集全世界的 DNS 以探索網路分散式內容.
- Technical analysis and Stock Research
 - 分析股票資訊
- University of Maryland
 - 用Hadoop 執行 machine translation, language modeling, bioinformatics, email analysis, and image processing 相關研究
- University of Nebraska Lincoln, Research Computing Facility
 - 用Hadoop跑約200TB的CMS經驗分析
 - 緊湊渺子線圈（CMS，Compact Muon Solenoid）為瑞士歐洲核子研究組織CERN的大型強子對撞器計劃的兩大通用型粒子偵測器中的一個。

Hadoop Applications (6)

- PARC
 - Used Hadoop to analyze Wikipedia conflicts
- Search Wikia
 - A project to help develop open source social search tools
- Yahoo!
 - Used to support research for Ad Systems and Web Search
 - 使用Hadoop平台來發現發送垃圾郵件的殭屍網絡
- 趨勢科技
 - 過濾像是釣魚網站或惡意連結的網頁內容