



# 深入解析雲端大量資料分析技術

Part 3 : Deep Dive into Data Science Technologies

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



Powered by DRBL

# Open Source Mapping of Google Core Technologies

## Google 三大關鍵技術對應的自由軟體

### BigTable

A huge key-value datastore

HBase, Hypertable  
Cassandra, ....

### MapReduce

To parallel process data

Hadoop MapReduce API  
Sphere MapReduce API, ...

### Google File System

To store petabytes of data

Hadoop Distributed File System (HDFS)  
Sector Distributed File System

更多不同語言的 MapReduce API 實作：

<http://trac.nchc.org.tw/grid/intertrac/wiki%3Ajazz/09-04-14%23MapReduce>

其他值得觀察的分散式檔案系統：

- IBM GPFS - <http://www-03.ibm.com/systems/software/gpfs/>
- Lustre - <http://www.lustre.org/>
- Ceph - <http://ceph.newdream.net/>

# Building PaaS with Open Source

## 用自由軟體打造 PaaS 雲端服務

應用軟體 Application  
Social Computing, Enterprise, ISV, ...

eyeOS, Nutch, ICAS,  
X-RIME, ...

程式語言 Programming  
Web 2.0 介面, Mashups, Workflows, ...

Hadoop (MapReduce),  
Sector/Sphere, AppScale

控制管理 Control  
Qos Negotiation, Admission Control,  
Pricing, SLA Management, Metering...

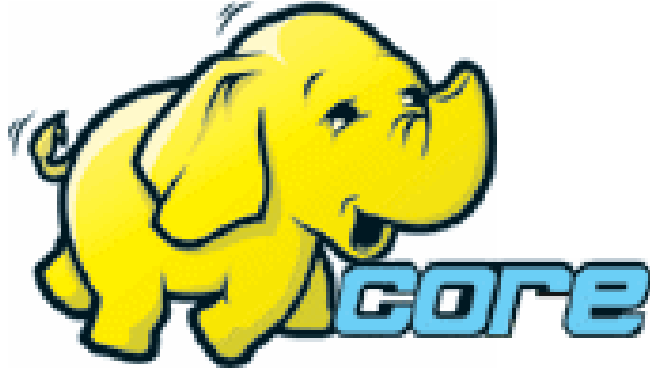
OpenNebula, Enomaly,  
Eucalyptus, OpenQRM, ...

虛擬化 Virtualization  
VM, VM management and Deployment

Xen, KVM, VirtualBox,  
QEMU, OpenVZ, ...

硬體設施 Hardware  
Infrastructure: Computer, Storage, Network

# Hadoop

- <http://hadoop.apache.org>
  - Hadoop 是 Apache Top Level 開發專案
  - **Hadoop is Apache Top Level Project**
  - 目前主要由 Yahoo! 資助、開發與運用
  - **Major sponsor is Yahoo!**
  - 創始者是 Doug Cutting，參考 Google Filesystem
  - **Developed by Doug Cutting, Reference from Google Filesystem**
  - 以 Java 開發，提供 HDFS 與 MapReduce API。
  - **Written by Java, it provides HDFS and MapReduce API**
  - 2006 年使用在 Yahoo 內部服務中
  - **Used in Yahoo since year 2006**
  - 已佈署於上千個節點。
  - **It had been deploy to 4000+ nodes in Yahoo**
  - 處理 Petabyte 等級資料量。
  - **Design to process dataset in Petabyte**
- 
- Facebook、Last.fm  
、Joost are also  
powered by Hadoop**

# Sector / Sphere

- <http://sector.sourceforge.net/>
- 由美國資料探勘中心研發的自由軟體專案。
- **Developed by National Center for Data Mining, USA**
- 採用 C/C++ 語言撰寫，因此效能較 Hadoop 更好。
- **Written by C/C++, so performance is better than Hadoop**
- 提供「類似」Google File System 與 MapReduce 的機制
- **Provide file system similar to Google File System and MapReduce API**
- 基於UDT高效率網路協定來加速資料傳輸效率
- **Based on UDT which enhance the network performance**
- Open Cloud Testbed有提供測試環境，並開發Ma1Stone效能評比軟體
- **Open Cloud Consortium provide Open Cloud Testbed and develop Ma1Stone toolkit for benchmark**

**Sector-Sphere**

National Center for Data Mining  
University of Illinois at Chicago



Open Data Group

<http://www.opendatagroup.com/>

# Why should we learn Hadoop ?

## 為何需要學習 **Hadoop ??**

[Search Jobs](#) [Browse Jobs](#) [Local Jobs](#) [Salaries](#) [Employment Trends](#)

**simply**hired<sup>®</sup>

job search made simple

Employment Trends

Xen, Hyper-V, Hadoop

Tip: You can compare trends by separating them with commas.

Xen, Hyper-v, Hadoop Trends



### Xen, Hyper-v, Hadoop Job Trends

This graph displays the percentage of jobs with your search terms anywhere in the job listing. Since November 2008, the following has occurred:

- [Xen jobs](#) increased 141%
- [Hyper-v jobs](#) increased 551%
- [Hadoop jobs](#) did not change or there is no data available

## 1. Data Explore

資訊大爆炸

## 2. Data Mining Tool

方便作資料探勘的工作

## 3. Looking for Jobs

好找工作 !!

# Comparison between Google and Hadoop

## Google 與 Hadoop 的比較表

<b>Develop Group</b>	Google	Apache
<b>Sponsor</b>	Google	Yahoo, Amazon
<b>Algorithm Method</b>	MapReduce	MapReduce
<b>Resource</b>	open document	open source
<b>File System (MapReduce)</b>	GFS	HDFS
<b>Storage System (for structure data)</b>	big-table	HBase
<b>Search Engine</b>	Google	Nutch
<b>OS</b>	Linux	Linux / GPL

# What is Hadoop ?

用一句話解釋 **Hadoop** 是什麼??

***Hadoop is a **software platform** that lets one easily write and run applications that **process vast amounts of data.*****

*Hadoop* 是一個讓使用者簡易撰寫並執行**處理海量資料**應用程式的**軟體平台**。

亦可以想像成一個**處理海量資料的生產線**，只須學會定義 **Map** 跟 **Reduce** 工作站該做哪些事情。



# *Features of Hadoop ...*

## **Hadoop 這套軟體的特色是 ...**

- **海量 Vast Amounts of Data**
  - 擁有儲存與處理大量資料的能力
  - Capability to **STORE** and **PROCESS** vast amounts of data.
- **經濟 Cost Efficiency**
  - 可以用在由一般 PC 所架設的叢集環境內
  - Based on large clusters built of **commodity hardware**.
- **效率 Parallel Performance**
  - 透過分散式檔案系統的幫助，以致得到快速的回應
  - With the help of HDFS, Hadoop **have better performance**.
- **可靠 Robustness**
  - 當某節點發生錯誤，能即時自動取得備份資料及佈署運算資源
  - Robustness to add and remove computing and storage resource without shutdown entire system.

# *Founder of Hadoop – Doug Cutting*

**Hadoop** 這套軟體的創辦人 **Doug Cutting**

Doug Cutting Talks About The Founding Of Hadoop

clouderahadoop

9 部影片

編輯訂閱項目

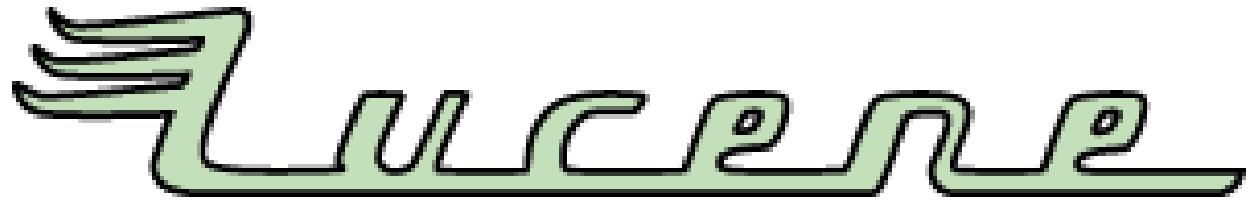


Doug Cutting Talks About The Founding Of Hadoop

<http://www.youtube.com/watch?v=qxC4urJOchs>

# History of Hadoop ... 2002~2004

## Hadoop 這套軟體的歷史源起 ... 2002~2004



- Lucene

- <http://lucene.apache.org/>
- 用Java 設計的高效能文件索引引擎API
- a high-performance, full-featured **text search engine library** written entirely in **Java**.
- 索引文件中的每一字，讓搜尋的效率比傳統逐字比較還要高的多
- Lucene create an **inverse index** of every word i n different documents. It enhance performance of text searching.

# History of Hadoop ... 2002~2004

## Hadoop 這套軟體的歷史源起 ... 2002~2004

- Nutch



- <http://nutch.apache.org/>
- Nutch 是基於開放原始碼所開發的網站搜尋引擎
- Nutch is open source web-search software.
- 利用 Lucene 函式庫開發
- It builds on Lucene and Solr, adding web-specifics, such as a crawler, a link-graph database, parsers for HTML and other document formats, etc.



# Three Gifts from Google ....

## 來自 Google 的三個禮物 ....

- Nutch 後來遇到儲存大量網站資料的瓶頸
- Nutch encounter storage issue
- Google 在一些會議分享他們的三大關鍵技術
- Google shared their design of web-search engine
  - SOSP 2003 : “The Google File System”
  - <http://labs.google.com/papers/gfs.html>
  - OSDI 2004 : “MapReduce : Simplified Data Processing on Large Cluster”
  - <http://labs.google.com/papers/mapreduce.html>
  - OSDI 2006 : “Bigtable: A Distributed Storage System for Structured Data”
  - <http://labs.google.com/papers/bigtable-osdi06.pdf>



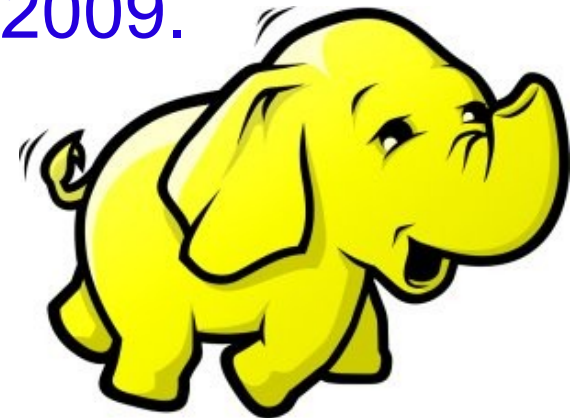
# ***History of Hadoop ... 2004 ~ Now***

## ***Hadoop 這套軟體的歷史源起 ... 2004 ~ Now***

- Dong Cutting reference from Google's publication
- Added DFS & MapReduce implement to Nutch
- According to **user feedback** on the mail list of Nutch ....
- Hadoop became separated project **since Nutch 0.8**
- Nutch DFS → Hadoop Distributed File System (HDFS)
- **Yahoo** hire Dong Cutting to build a team of web search engine at **year 2006**.
  - Only **14 team members** (engineers, clusters, users, etc.)
- Dong Cutting joined Cloudera at year 2009.

**YAHOO!**

 **cloudera**



# Who Use Hadoop ??

## 有哪些公司在用 **Hadoop** 這套軟體 ??

- **Yahoo** is the key contributor currently.
- **IBM** and **Google** teach Hadoop in universities ...
- [http://www.google.com/intl/en/press/pressrel/20071008\\_ibm\\_univ.html](http://www.google.com/intl/en/press/pressrel/20071008_ibm_univ.html)
- **The New York Times** used **100 Amazon EC2 instances** and a Hadoop application to process **4TB of raw image TIFF data** (stored in S3) into **11 million finished PDFs** in the space of **24 hours** at a computation cost of about **\$240** (not including bandwidth)
  - from <http://en.wikipedia.org/wiki/Hadoop>
- <http://wiki.apache.org/hadoop/AmazonEC2>
- <http://wiki.apache.org/hadoop/PoweredBy>
  - A9.com
  - ADSDAQ by Contextweb
  - EHarmony
  - Facebook
  - Fox Interactive Media
  - IBM
  - ImageShack
  - ISI
  - Joost
  - Last.fm
  - Powerset
  - The New York Times
  - Rackspace
  - Veoh
  - Metaweb

# Performance improvement of Hadoop

## Hadoop 過去幾年的效能改進 (from Yahoo)

年份	日期	節點數	耗時 ( 小時 )
2006	四月	188	47.9
2006	五月	500	42
2006	十一月	20	1.8
2006	十一月	100	3.3
2006	十一月	500	5.2
2006	十一月	900	7.8
2007	七月	20	1.2
2007	七月	100	1.3
2007	七月	500	2
2007	七月	900	2.5

Sort benchmark, every nodes with terabytes data.



# *Hadoop in production run ....*

## 商業運轉中的 **Hadoop** 應用 ....

- **February 19, 2008**
- **Yahoo! Launches World's Largest Hadoop Production Application**
- <http://developer.yahoo.net/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html>

Number of links between pages in the index	roughly 1 trillion links
Size of output	over 300 TB, compressed!
Number of cores used to run single Map-Reduce job	over 10,000
Raw disk used in the production cluster	over 5 Petabytes

# *Hadoop in production run ....*

## 商業運轉中的 **Hadoop** 應用 ....

- **September 30, 2008**
- **Scaling Hadoop to 4000 nodes at Yahoo!**
- [http://developer.yahoo.net/blogs/hadoop/2008/09/scaling\\_hadoop\\_to\\_4000\\_nodes\\_a.html](http://developer.yahoo.net/blogs/hadoop/2008/09/scaling_hadoop_to_4000_nodes_a.html)

<b>Total Nodes</b>	<b>4000</b>
<b>Total cores</b>	<b>30000</b>
<b>Data</b>	<b>16PB</b>

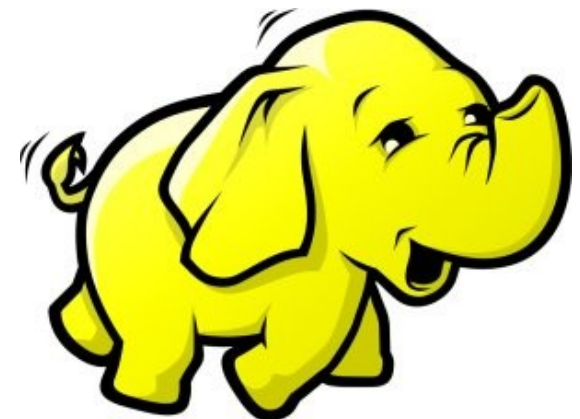
	<b>500-node cluster</b>		<b>4000-node cluster</b>	
	<b>write</b>	<b>read</b>	<b>write</b>	<b>read</b>
<b>number of files</b>	990	990	14,000	14,000
<b>file size (MB)</b>	320	320	360	360
<b>total MB processes</b>	316,800	316,800	5,040,000	5,040,000
<b>tasks per node</b>	2	2	4	4
<b>avg. throughput (MB/s)</b>	<b>5.8</b>	<b>18</b>	<b>40</b>	<b>66</b>



# Hadoop 專業術語

*Introduction to Hadoop Terminology*

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



# Two Key Elements of Operating System

## 作業系統兩大關鍵組成元素

**Scheduler**  
程序排程



**File System**  
檔案系統



# Terminologies of Hadoop

## Hadoop 文件中的專業術語

- Job
  - 任務
- Task
  - 小工作
- JobTracker
  - 任務分派者
- TaskTracker
  - 小工作的執行者
- Client
  - 發起任務的客戶端
- Map
  - 應對
- Reduce
  - 總和



- Namenode
  - 名稱節點
- Datanode
  - 資料節點
- Namespace
  - 名稱空間
- Replication
  - 副本
- Blocks
  - 檔案區塊 (64M)
- Metadata
  - 屬性資料



# *Two Key Roles of HDFS*

## **HDFS** 軟體架構的兩種關鍵角色

### 名稱節點 **NameNode**

- **Master Node**
- **Manage NameSpace of HDFS**
- **Control Permission of Read and Write**
- **Define the policy of Replication**
- **Audit and Record the NameSpace**
- **Single Point of Failure**

### 資料節點 **DataNode**

- **Worker Nodes**
- **Perform operation of Read and Write**
- **Execute the request of Replication**
- **Multiple Nodes**

# *Two Key Roles of Job Scheduler*

## *程序排程的兩種關鍵角色*

### *JobTracker*

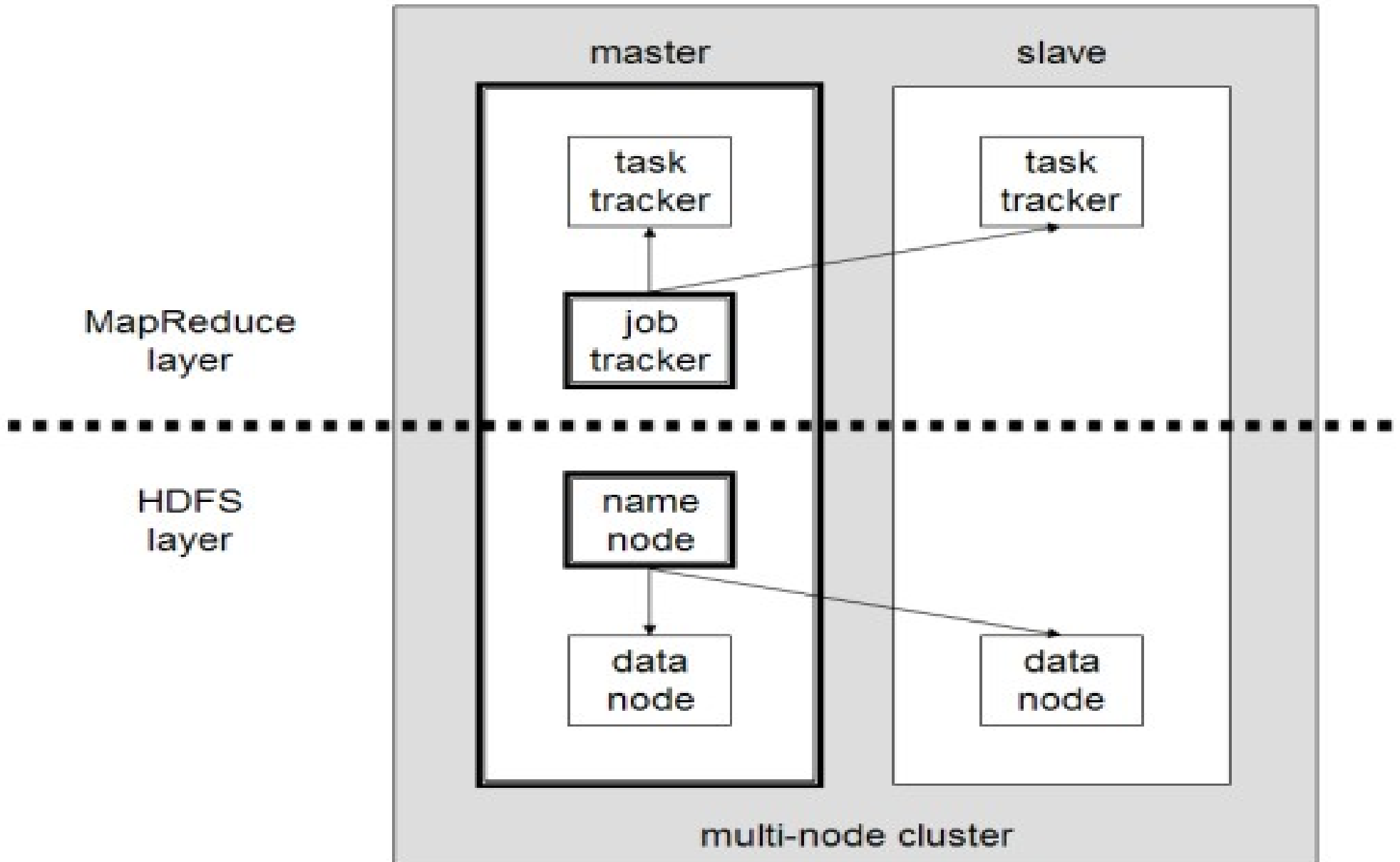
- **Master Node**
- **Receive Jobs from Hadoop Clients**
- **Assigned Tasks to TaskTrackers**
- **Define Job Queuing Policy, Priority and Error Handling**
- **Single Point of Failure**

### *TaskTracker*

- **Worker Nodes**
- **Excute Mapper and Reducer Tasks**
- **Save Results and report task status**
- **Multiple Nodes**

# *Different Roles of Hadoop Architecture*

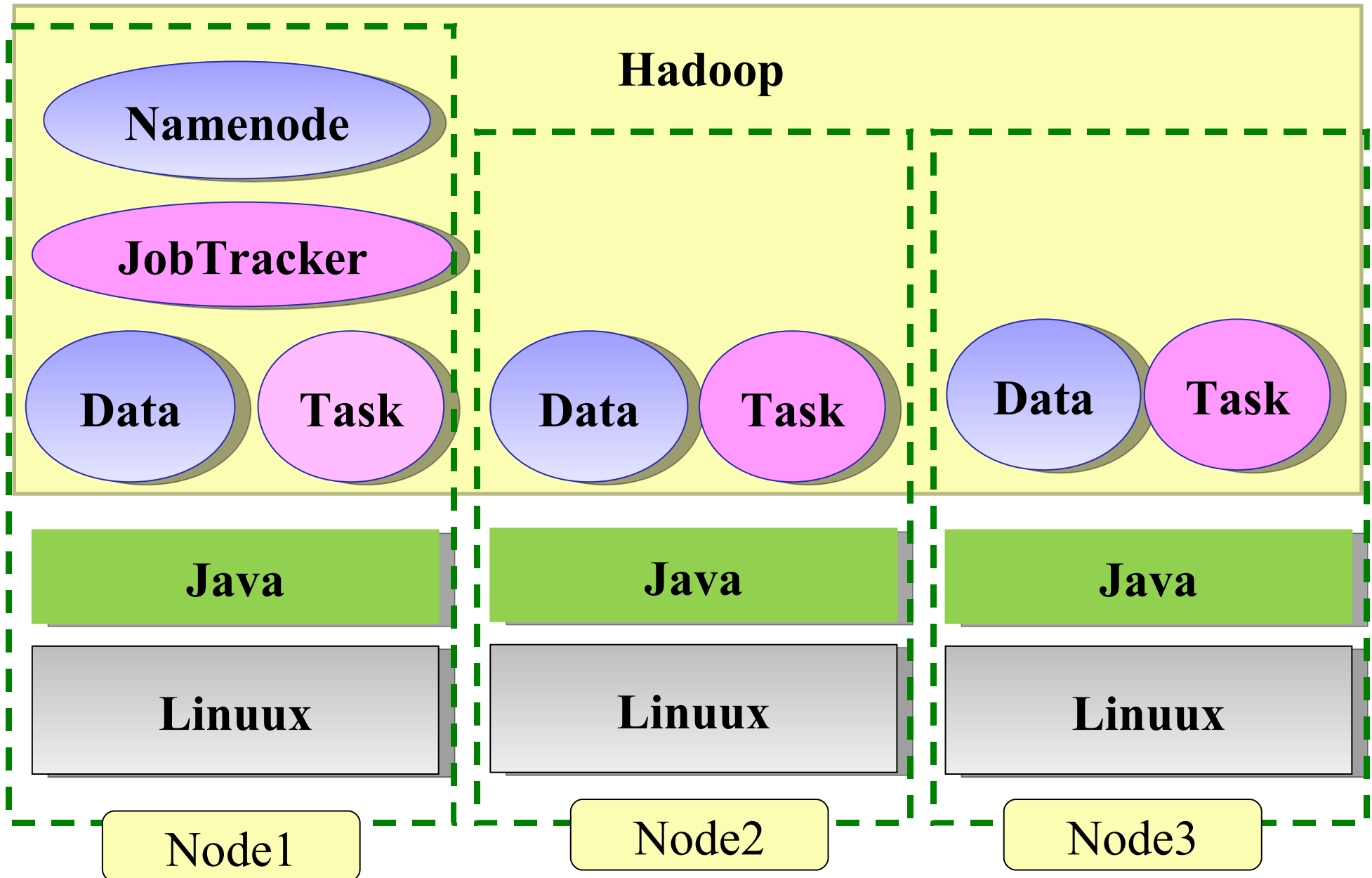
## **Hadoop** 軟體架構中的不同角色





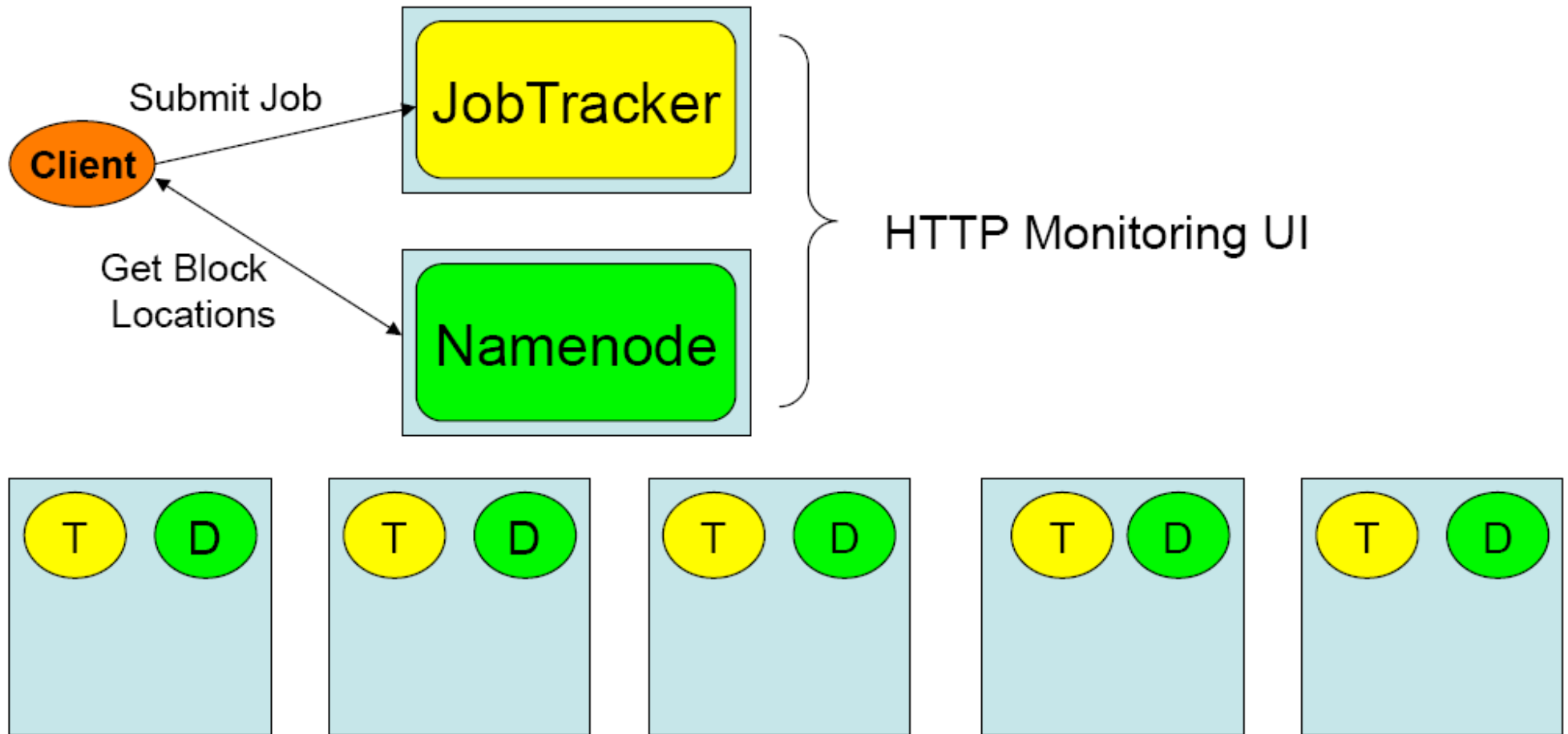
# Distributed Operating System of Hadoop

**Hadoop** 建構成一個分散式作業系統



# About Hadoop Client ...

## 不在雲裡的 **Hadoop Client**

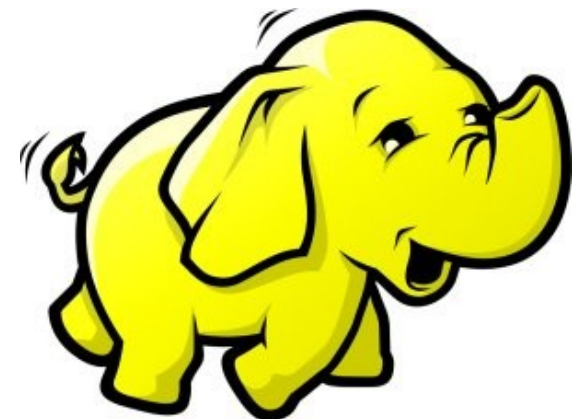




# HDFS 簡介

*Introduction to Hadoop Distributed File System*

**Jazz Wang**  
**Yao-Tsung Wang**  
**jazz@nchc.org.tw**



# ***What is HDFS ??***

## ***什麼是 HDFS ??***

- **Hadoop Distributed File System**

- 實現類似 Google File System 分散式檔案系統
- Reference from Google File System.
- 一個易於擴充的分散式檔案系統，目的為對大量資料進行分析
- **A scalable distributed file system for large data analysis .**
- 運作於廉價的普通硬體上，又可以提供容錯功能
- **based on commodity hardware with high fault-tolerant.**
- 給大量的用戶提供總體性能較高的服務
- **It have better overall performance to serve large amount of users.**

# *Features of HDFS ...*

## **HDFS 的特色是 ...**

- **硬體錯誤容忍能力 Fault Tolerance**
  - 硬體錯誤是正常而非異常
  - Failure is the norm rather than exception
  - 自動恢復或故障排除
  - automatic recovery or report failure
- **串流式的資料存取 Streaming data access**
  - 批次處理多於用戶交互處理
  - Batch processing rather than interactive user access.
  - 高 Throughput 而非低 Latency
  - High aggregate data bandwidth (throughput)

# Features of HDFS ...

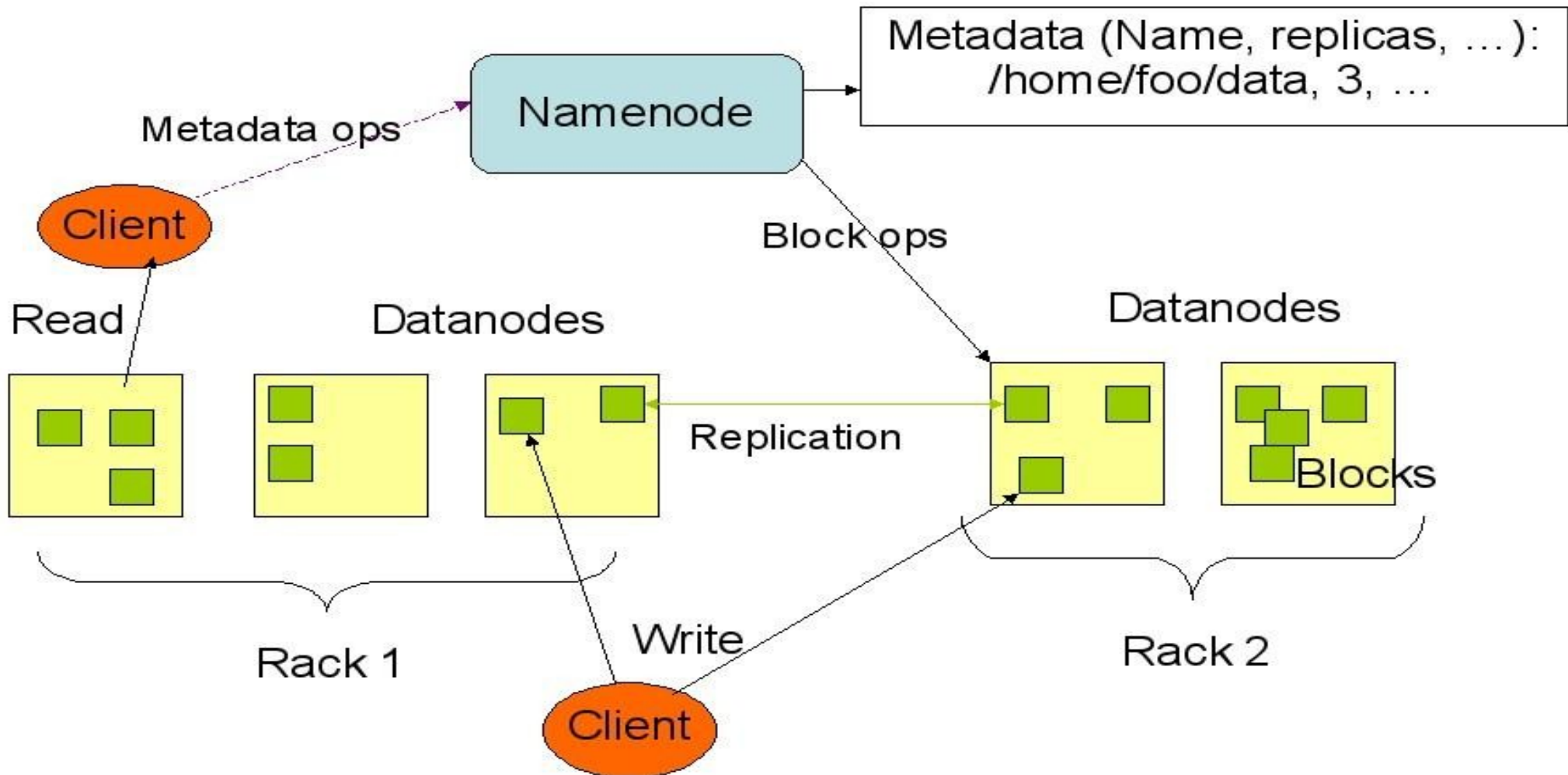
## HDFS 的特色是 ...

- **大規模資料集 Large data sets and files**
  - 支援 Petabytes 等級的磁碟空間
  - Support Petabytes size
- **一致性模型 Coherency Model**
  - 一次寫入，多次存取 Write-once-read-many
  - 簡化一致性處理問題 This assumption simplifies coherency
- **在地運算 Data Locality**
  - 到資料的節點上計算 > 將資料從遠端複製過來計算
  - “move compute to data” > “move data to compute”
- **異質平台移植性 Heterogeneous**
  - 即使硬體不同也可移植、擴充
  - HDFS could be deployed on different hardware

# How HDFS manage data ...

## HDFS 如何管理資料 ...

HDFS Architecture



# How does HDFS work ...

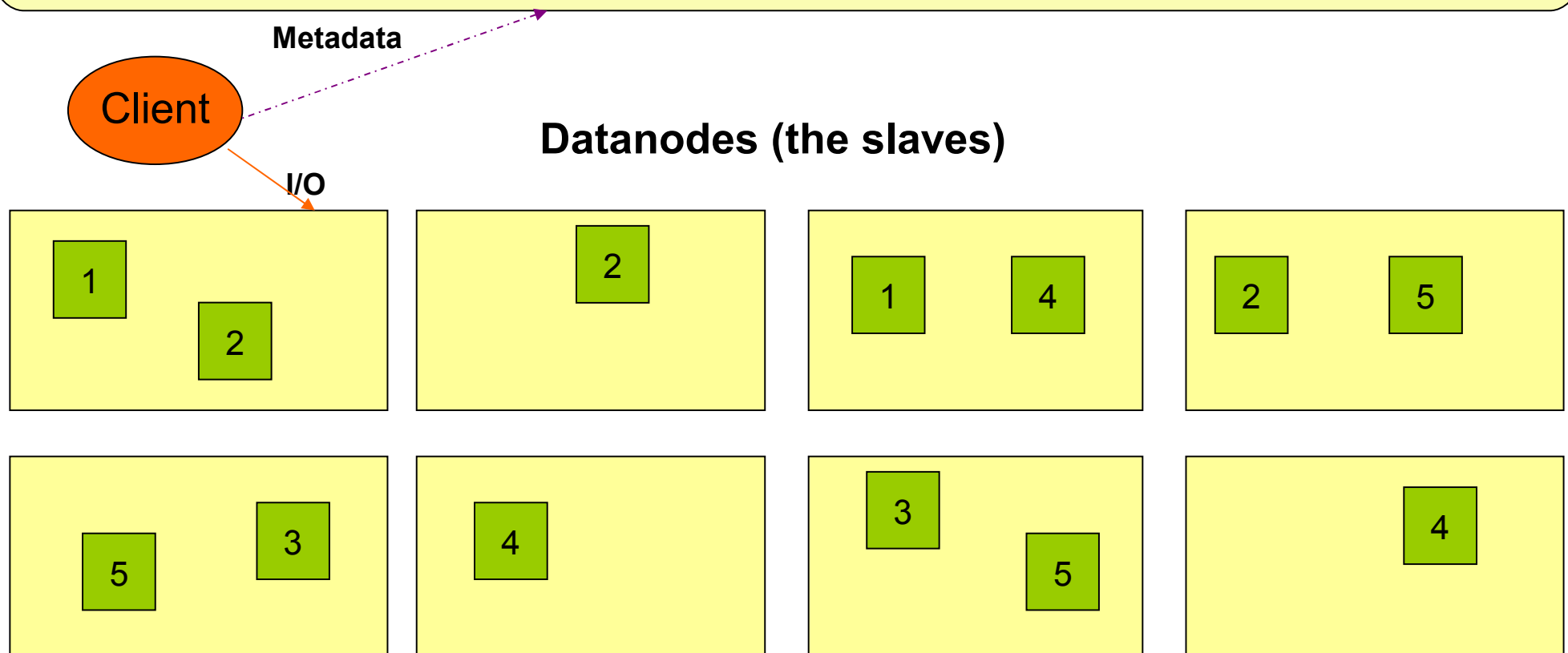
## HDFS 如何運作 ...

Namenode (the master)

Path and Filename – Replication , blocks

name:/users/joeYahoo/myFile - copies:2, blocks:{1,3}

name:/users/bobYahoo/someData.gzip, copies:3, blocks:{2,4,5}

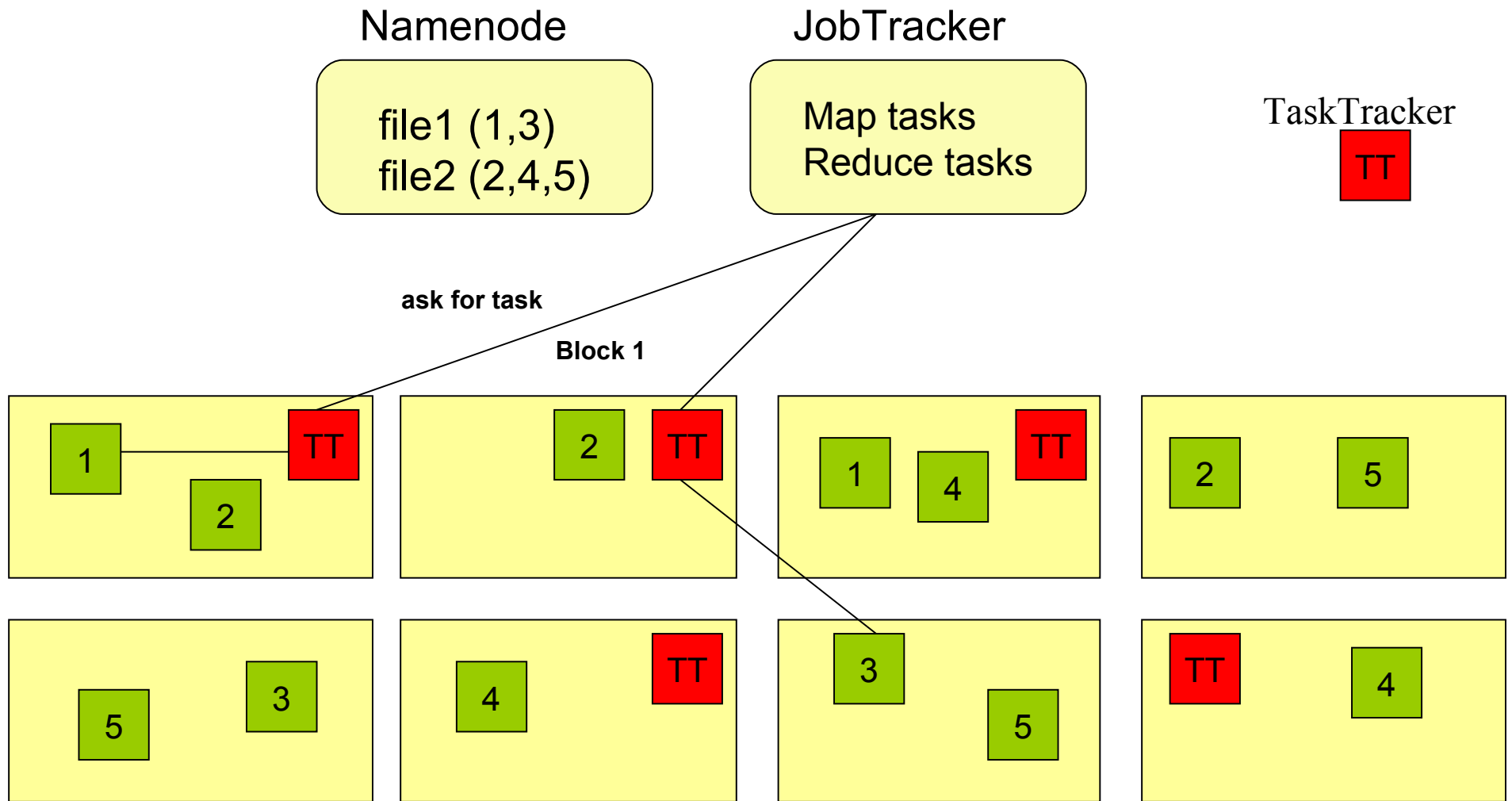




# About Data Locality ...

## HDFS 如何達成在地運算 ...

- Increase reliability and read bandwidth
  - robustness : read replication while found any failure
  - High read bandwidth : distribute read ( but increase write bottleneck )



# About Fault Tolerance ...

## HDFS 如何達成容錯機制 ...

資料崩毀  
Data Corrupt

網路或資料  
節點失效  
Network Fault  
DataNode Fault

名稱節點錯誤  
NameNode Fault

- 資料完整性 Data integrity
  - checked with CRC32
  - 用副本取代出錯資料
  - Replace corrupt block with replication one
- Heartbeat
  - Datanode send **heartbeat** to Namenode
- Metadata
  - FSImage、Editlog 為核心印象檔及日誌檔
  - FSImage – core file system mapping image
  - Editlog – like. SQL transaction log
  - 多份儲存，當名稱節點故障時可以手動復原
  - Multiple backups of FSImage and Editlog
  - Manually recovery while NameNode Fault

# Coherency Model and Performance of HDFS

## HDFS 的一致性機制與效能 ...

- **檔案一致性機制 Coherency model of files**
  - 刪除檔案 \ 新增寫入檔案 \ 讀取檔案皆由名稱節點負責
  - NameNode handle the operation of write, read and delete.
- **巨量空間及效能機制 Large Data Set and Performance**
  - 預設每個區塊大小以 64MB 為單位
  - By default, the block size is 64MB
  - 大區塊可提高存取效率
  - Bigger block size will enhance read performance
  - 檔案有可能大過一顆磁碟
  - Single file stored on HDFS might be larger than single physical disk of DataNode.
  - 區塊均勻散佈各節點以分散讀取流量
  - Fully distributed blocks increase throughput of reading.

# POSIX like HDFS commands

## 與 **POSIX** 相似的操作指令 ...

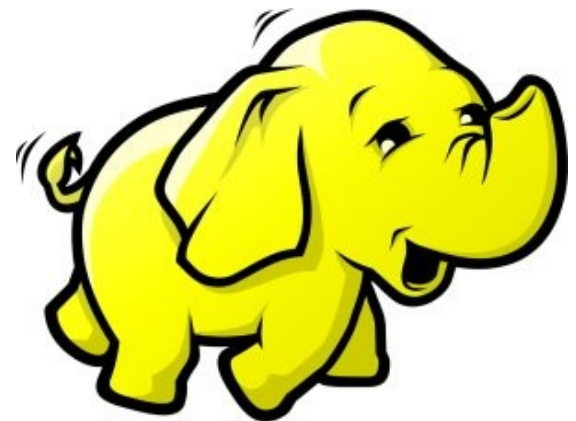
```
jazz@hadoop:~$ hadoop fs
Usage: java FsShell
    [-ls <path>]
    [-lsr <path>]
    [-du <path>]
    [-dus <path>]
    [-count[-q] <path>]
    [-mv <src> <dst>]
    [-cp <src> <dst>]
    [-rm <path>]
    [-rmr <path>]
    [-expunge]
    [-put <localsrc> ... <dst>]
    [-copyFromLocal <localsrc> ... <dst>]
    [-moveFromLocal <localsrc> ... <dst>]
    [-get [-ignoreCrc] [-crc] <src> <localdst>]
    [-getmerge <src> <localdst> [addnl]]
    [-cat <src>]
    [-text <src>]
    [-copyToLocal [-ignoreCrc] [-crc] <src> <localdst>]
    [-moveToLocal [-crc] <src> <localdst>]
    [-mkdir <path>]
    [-setrep [-R] [-w] <rep> <path/file>]
    [-touchz <path>]
    [-test -[ezd] <path>]
    [-stat [format] <path>]
    [-tail [-f] <file>]
    [-chmod [-R] <MODE[,MODE]... | OCTALMODE> PATH...]
    [-chown [-R] [OWNER][:[GROUP]] PATH...]
    [-chgrp [-R] GROUP PATH...]
    [-help [cmd]]
```



# MapReduce 簡介

*Introduction to MapReduce*

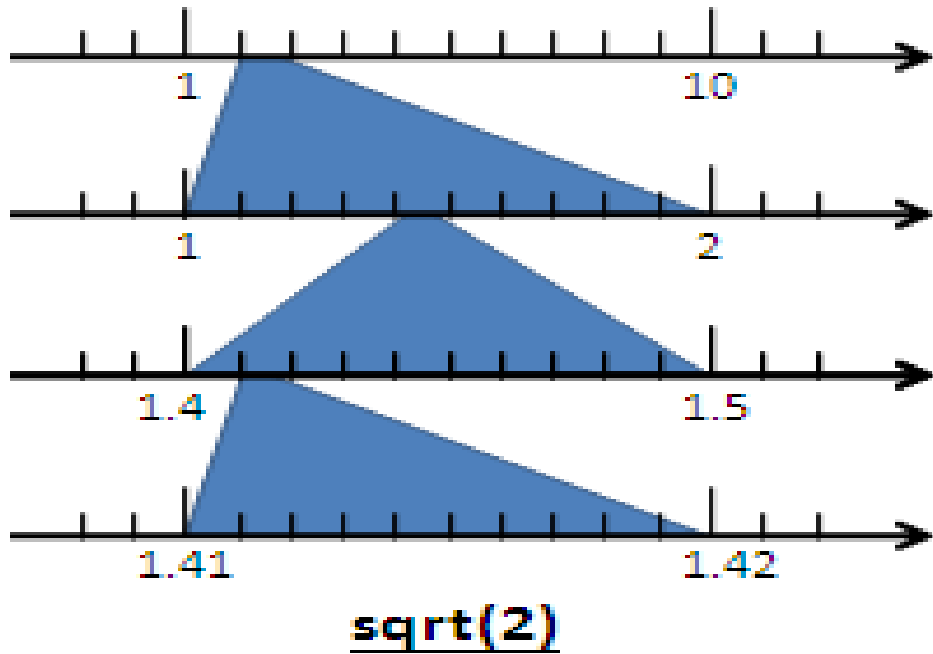
***Jazz Wang***  
***Yao-Tsung Wang***  
***jazz@nchc.org.tw***



# Divide and Conquer Algorithms

## 分而治之演算法

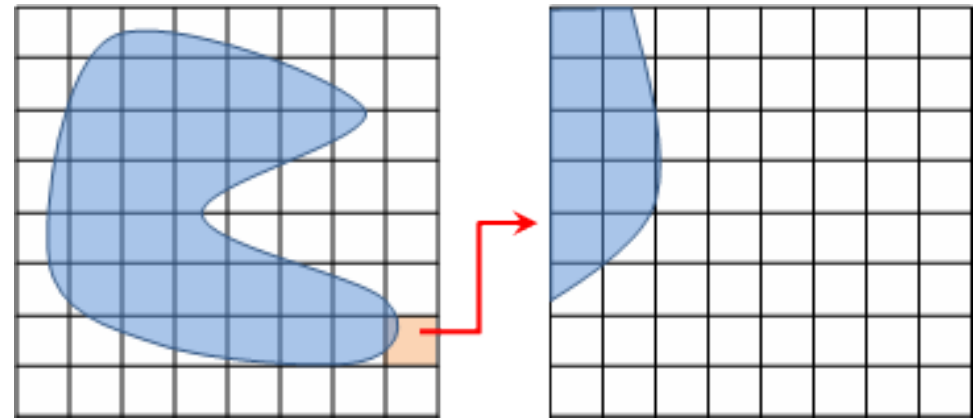
Example 1:



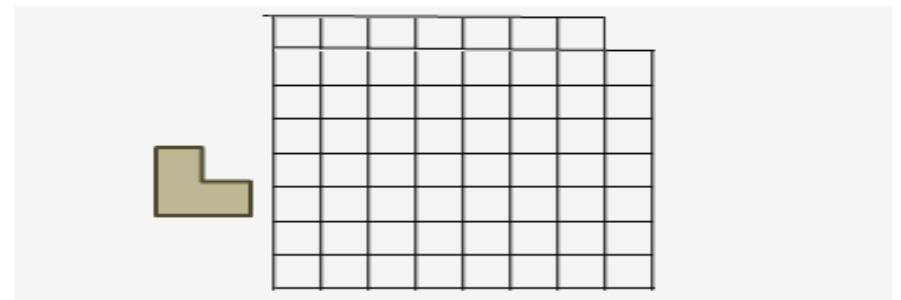
Example 4: The way to climb 5 steps stair within 2 steps each time. 眼前有五階樓梯，每次可踏上一階或踏上兩階，那麼爬完五階共有幾種踏法？

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

Example 2:



Example 3:



# What is MapReduce ??

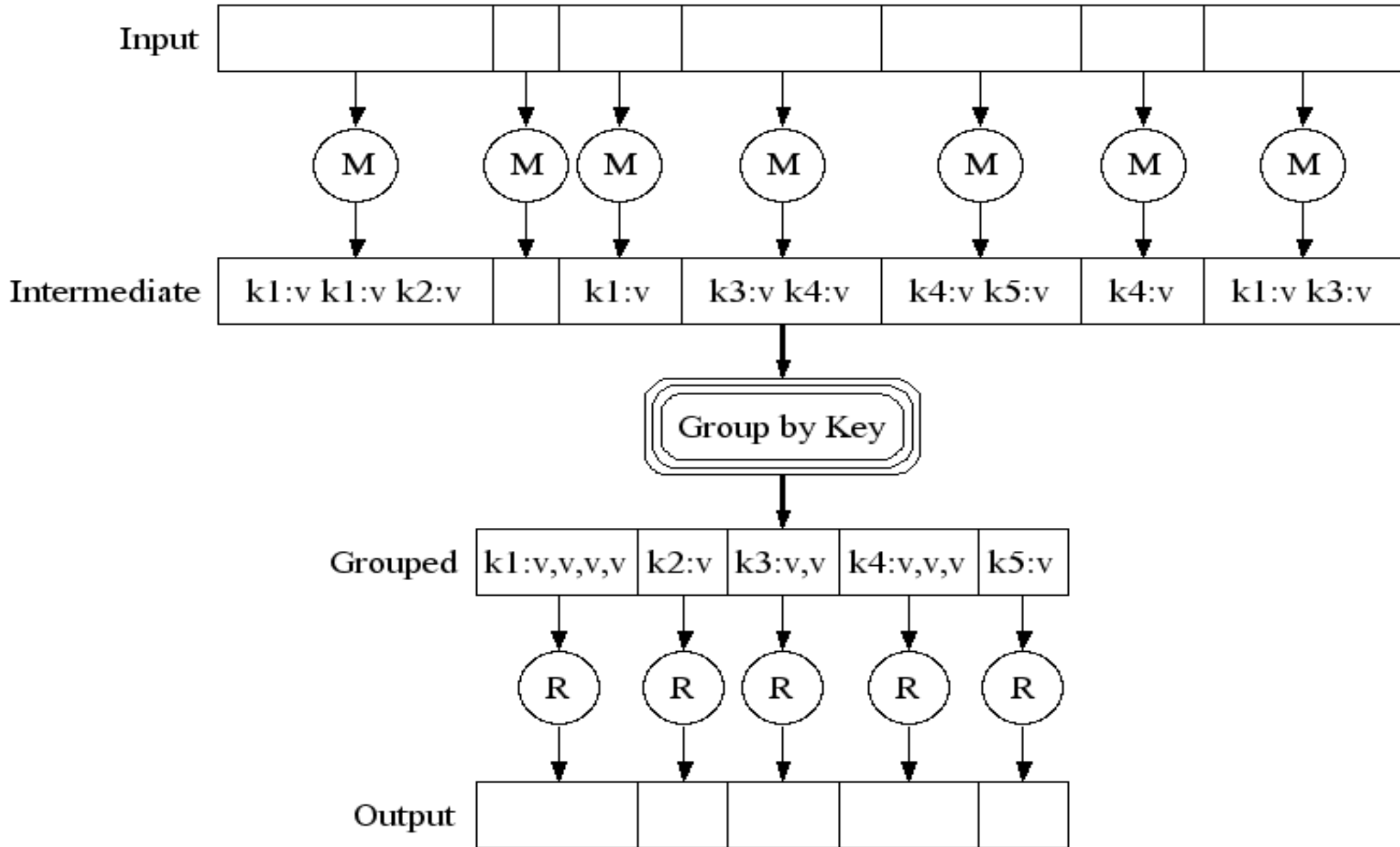
## 什麼是 MapReduce ??

- 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.
- 啟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(...):  $N \rightarrow N$ 
    - Ex. [ 1,2,3,4 ] – (**\*2**) -> [ 2,4,6,8 ]
  - Reduce(...):  $N \rightarrow 1$ 
    - [ 1,2,3,4 ] - (**sum**) -> 10
- **Logical view of MapReduce**
  - **Map(k1,v1) -> list(k2,v2)**
  - **Reduce(k2, list (v2)) -> list(v3)**

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

# Google's MapReduce Diagram

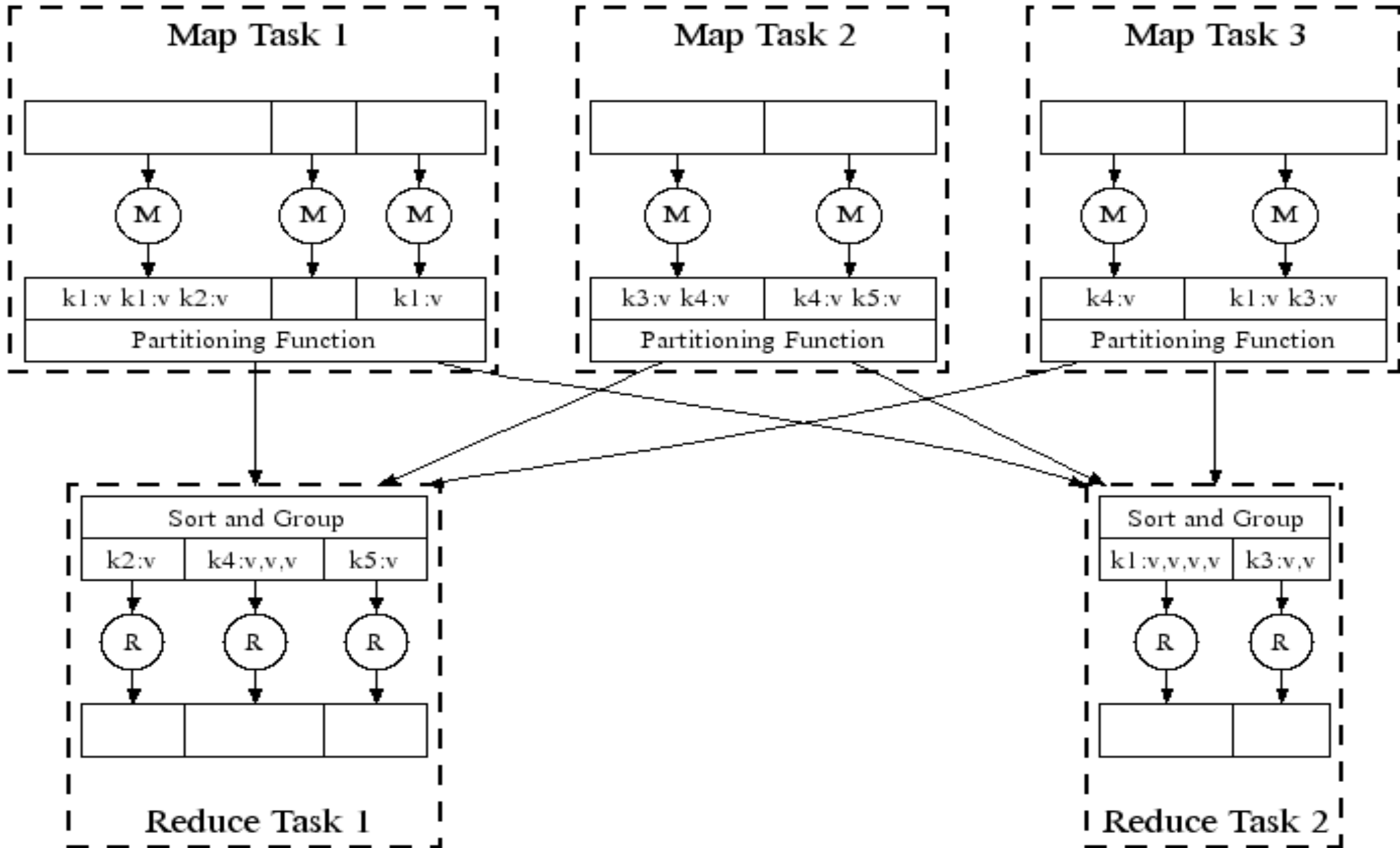
## Google 的 MapReduce 圖解





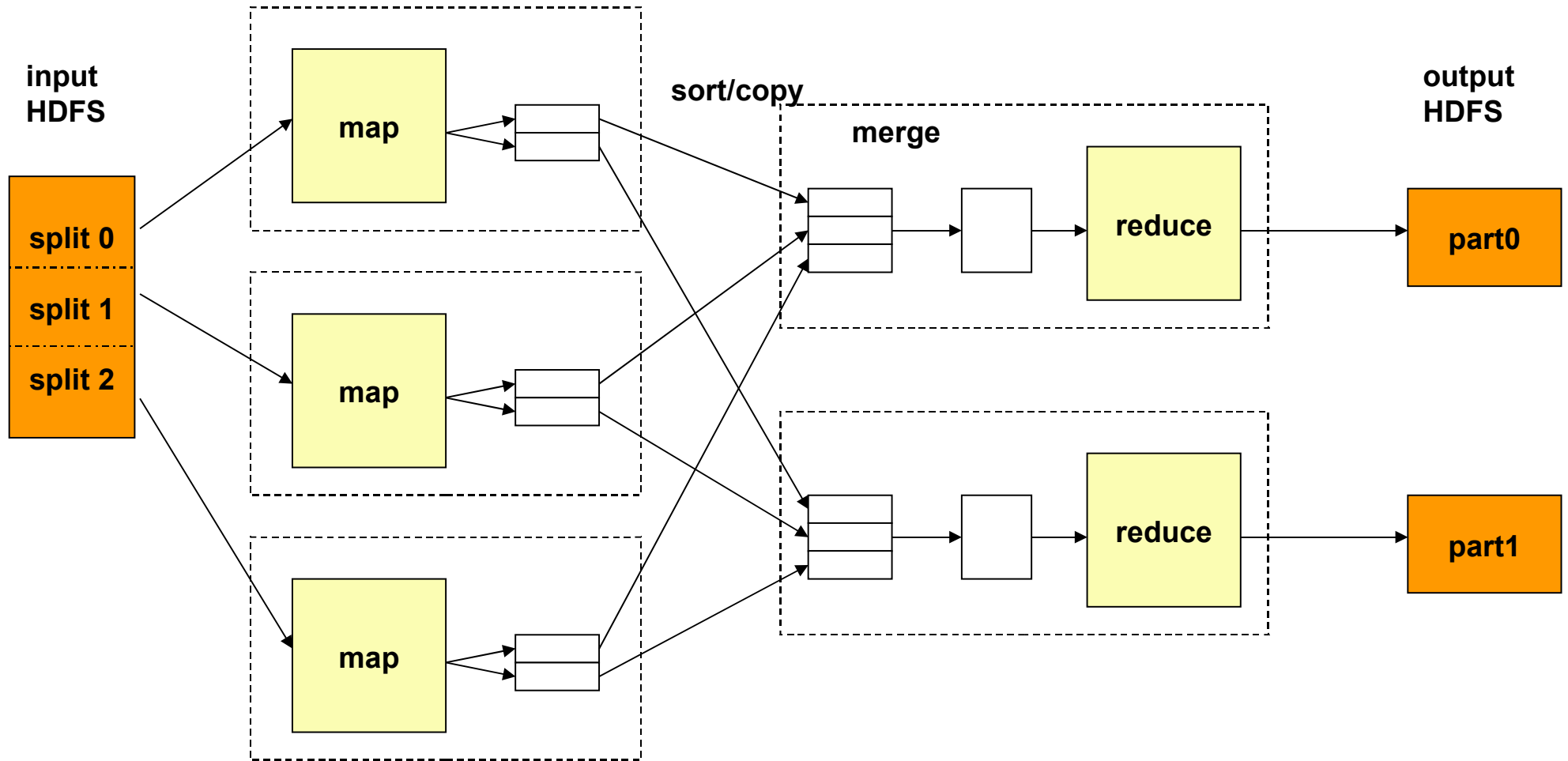
# Google's MapReduce in Parallel

## 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

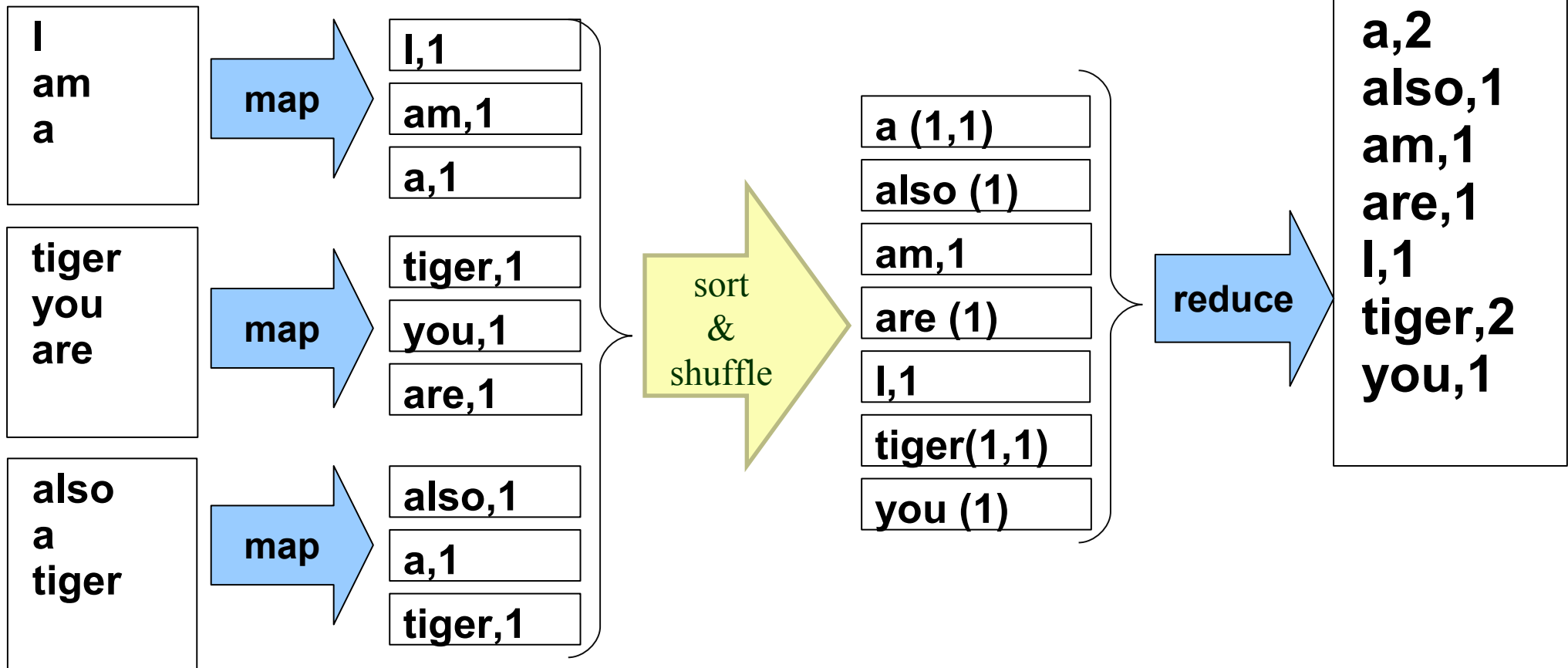
JobTracker 派遣 TaskTracker 作 reduce

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



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)

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

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

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
```

map

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

map

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

sort /  
merge

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

reduce

```
(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>

# What we learn today ?

## WHAT

**Hadoop 是運算海量資料的軟體平台 !!**

hadoop is a software platform to process vast amount of data!!

## WHO

始祖是 Doug Cutting，Apache 社群支持，Yahoo 贊助

From Doug Cutting to Apache Community, Yahoo and more !

## WHEN

**Hadoop 是 2004 年從 Nutch 分裂出來的專案 !!**

Hadoop became separate project since year 2004 !!

## WHY

**資料大爆炸、資料探勘、找工作**

***Data Explore, Data Mining, Jobs !!***

## HOW

**採用自由軟體也能打造私有雲端**

Install on large clusters built of commodity hardware !!