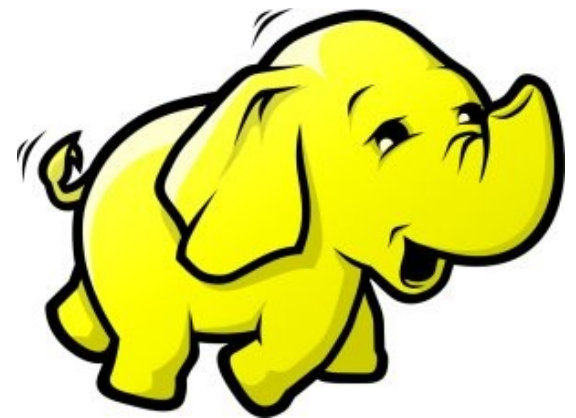




雲端核心技術 Hadoop 與 MapReduce 概念

Introduction to Hadoop, MapReduce and its Ecosystem

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



Course Information 課程資訊



- 講師介紹：
 - 國網中心 王耀聰 副研究員 / 交大電控碩士
 - jazz@nchc.org.tw
- 所有投影片、參考資料與操作步驟均在網路上
 - 由於雲端資訊變動太快，愛護地球，請減少不必要之講義列印。
- 礙於缺乏實機操作環境，故以影片展示與單機操作為主
 - 若有興趣實機操作，請參考國網中心雲端運算課程錄影
 - <http://trac.nchc.org.tw/cloud>
 - <http://www.classcloud.org/media>
 - <http://www.screentoaster.com/user?username=jazzwang>
- 若需要實驗環境，可至國網中心雲端運算實驗叢集申請帳號
 - <http://hadoop.nchc.org.tw>
- Hadoop 相關問題討論：
 - <http://forum.hadoop.tw>



淺談雲端運算趨勢與關鍵技術

The trend of cloud computing and its core technologies

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



Powered by DRBL



什麼是雲端運算啊？

What is Cloud Computing ?



雲端概念

雲端運算不是一項全新技術，而是一項概念。雲端的意義不在技術，而在商業模式的改變。

雲端概念

雲端基礎架構的相關IT建設，如伺服器、網路設施、電源供應器、散熱、儲存裝置等硬體產品，都是台灣科技業的強項。

雲端概念

在硬體設備的優勢基礎上，雲端平台與應用服務，提供台灣ICT產業一個轉型的新契機，台灣的創新能力，不落人後。

<http://www.youtube.com/watch?v=bJLSAcU6O3U>

<http://www.youtube.com/watch?v=VIMtd3nfPqc>

當紅「雲端運算」你瞭解了嗎？
雲端產業 8分鐘就上手

National Definition of Cloud Computing

美國國家標準局 NIST 給雲端運算所下的定義

5 Characteristics

五大基礎特徵

4 Deployment Models

四個佈署模型

3 Service Models

三個服務模式

1. On-demand self-service.

隨需自助服務

2. Broad network access

隨時隨地用任何網路裝置存取

3. Resource pooling

多人共享資源池

4. Rapid elasticity

快速重新佈署靈活度

5. Measured Service

可被監控與量測的服務

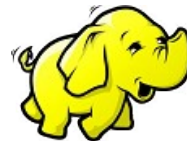
2 perspectives : Services vs Technologies

您想聽的是「雲端服務」還是「雲端技術」？



雲端服務

Microsoft



雲端技術



Cloud computing hype spurs confusion, Gartner says
<http://www.computerworld.com/s/article/print/9115904>

淺談雲端運算 (Cloud Computing)

http://www.cc.ntu.edu.tw/chinese/epaper/0008/20090320_8008.htm

The wisdom of Clouds (Crowds)

雲端序曲：雲端的智慧始終來自於群眾的智慧

2006年8月9日

Google 執行長施密特 (Eric Schmidt) 於 SES'06 會議中首次使用「雲端運算 (Cloud Computing) 」來形容無所不在的網路服務

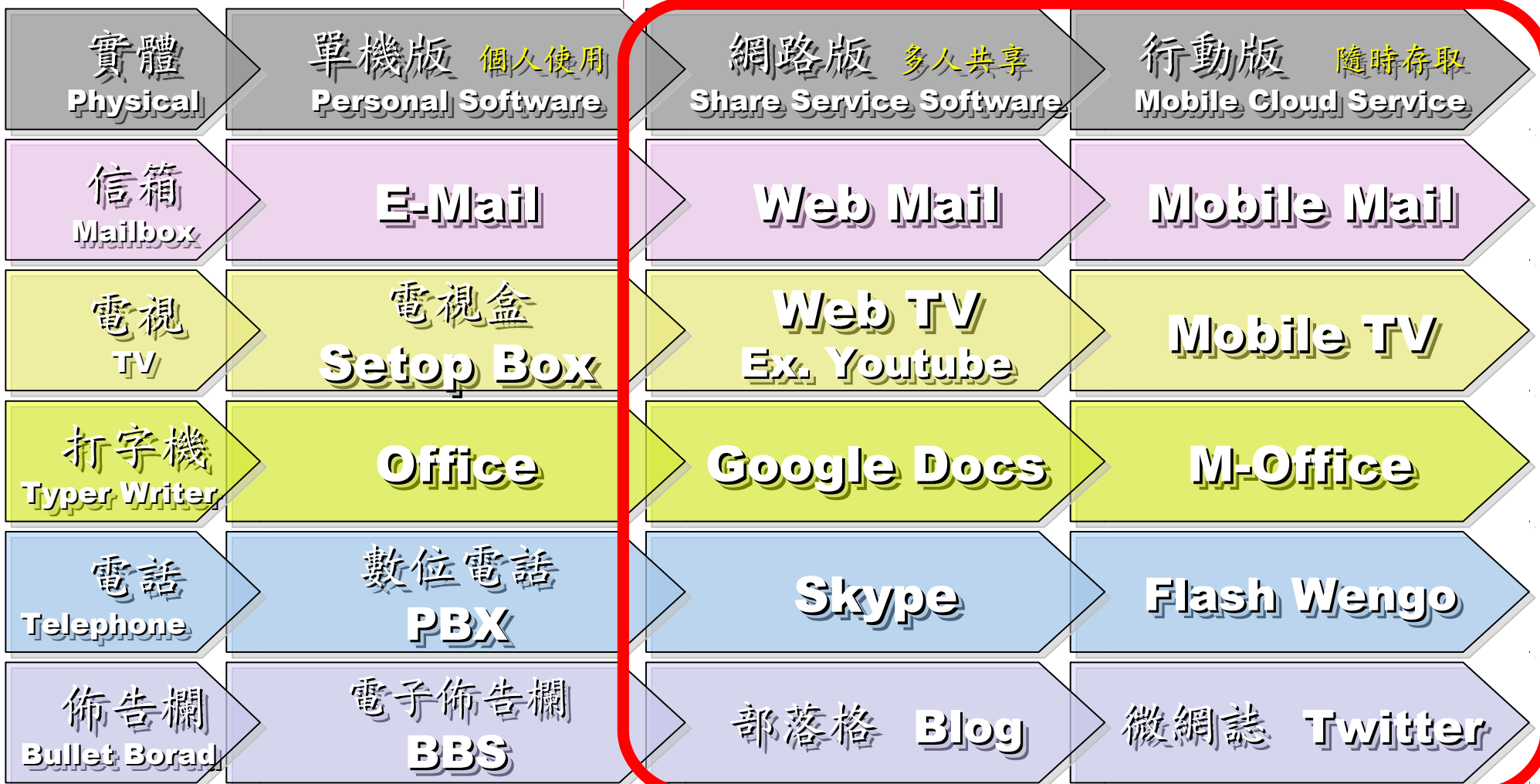
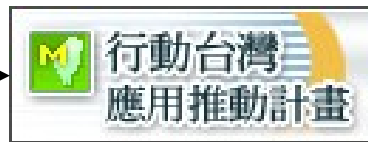
2006年8月24日

Amazon 以 Elastic Compute Cloud 命名其虛擬運算資源服務



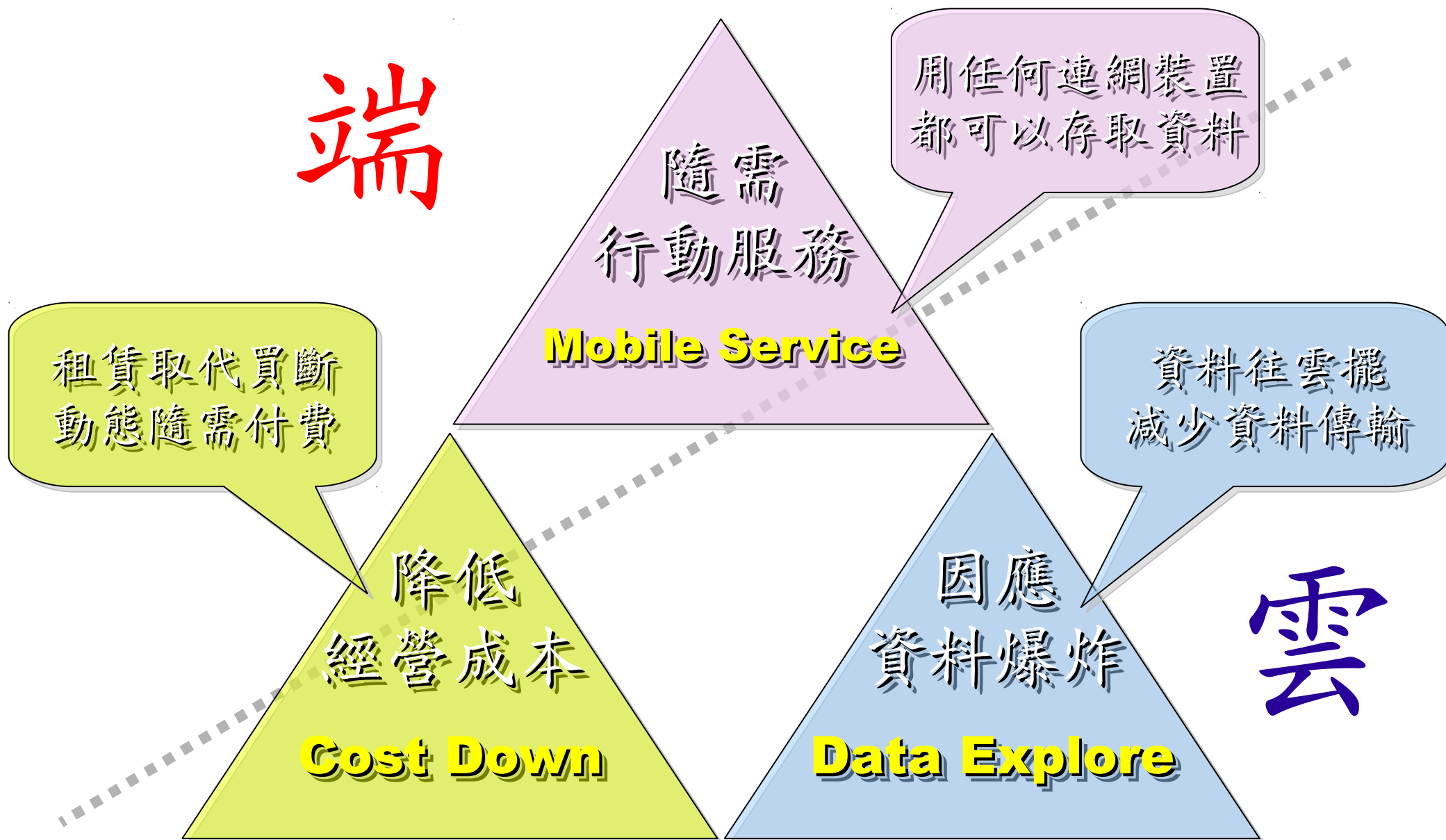
Evolution of Cloud Services

雲端服務只是軟體演化史的必然趨勢

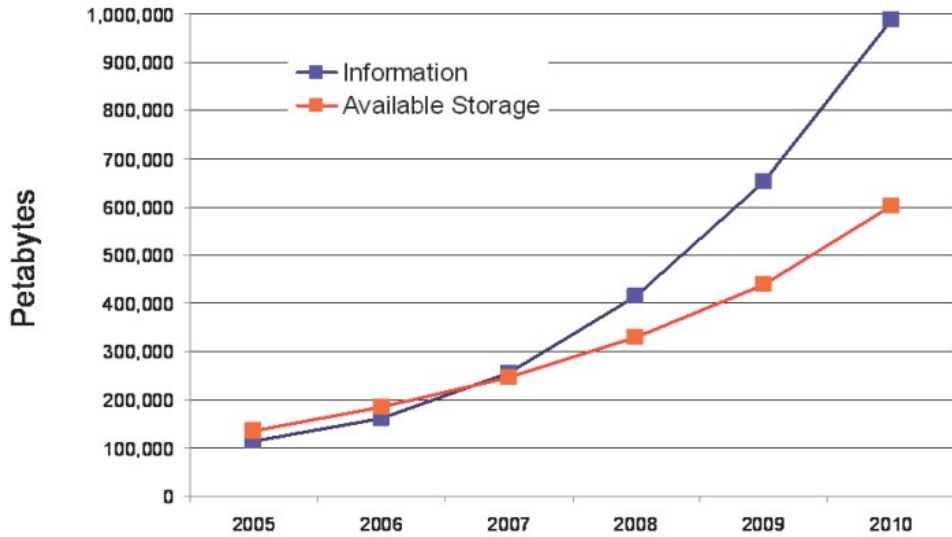


Key Driving Forces of Cloud Computing

雲端運算的關鍵驅動力



Information Versus Available Storage



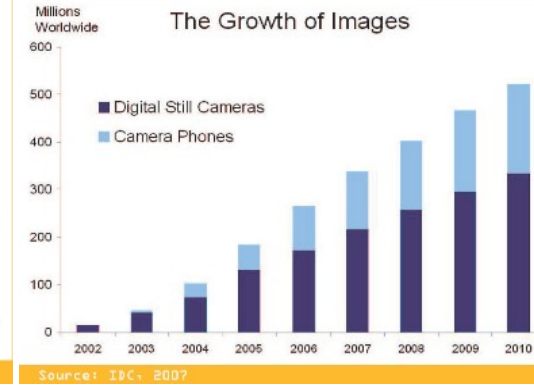
Source: IDC, 2007

2007 Data Explore

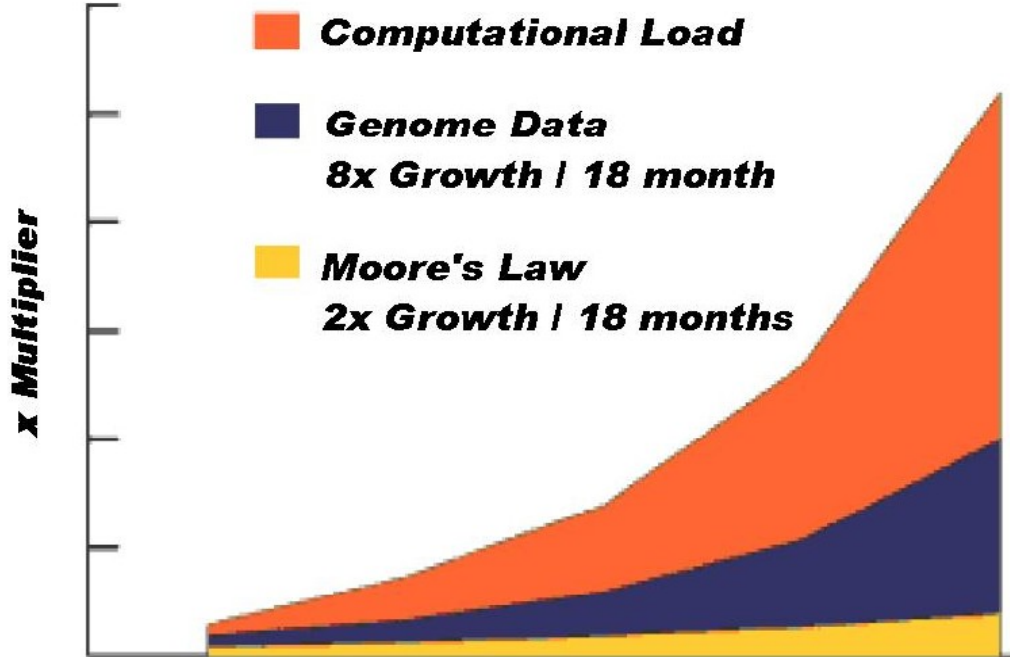
Top 1 : Human Genomics - 7000 PB / Year
Top 2 : Digital Photos - 1000 PB+ / Year
Top 3 : E-mail (no Spam) - 300 PB+ / Year



Source: IDC, 2007



Source: IDC, 2007



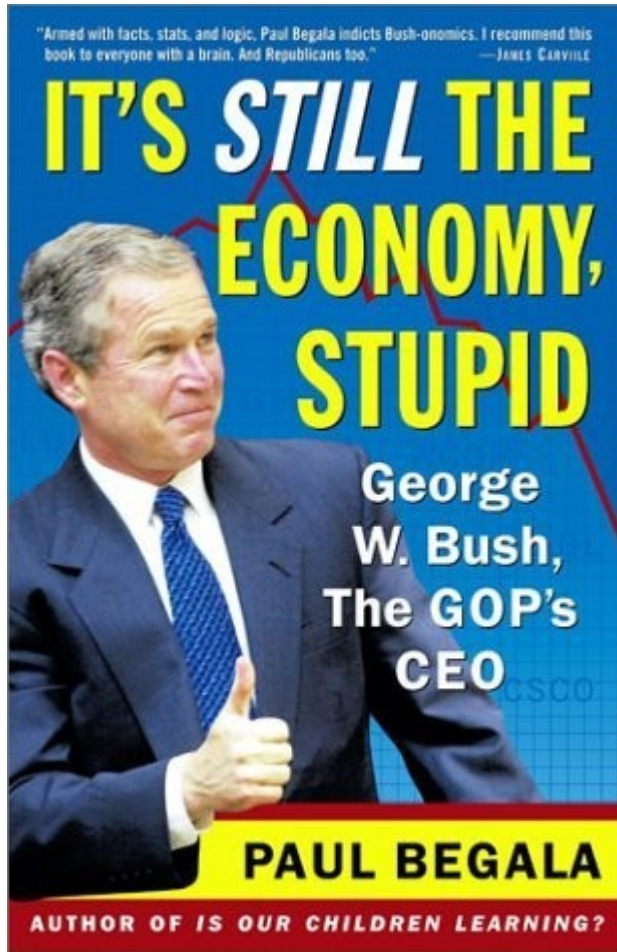
Particle Physics Large Hadron Collider (15PB)	Human Genomics (7000PB) 1GB / person 200PB+ captured 200% CAGR	World Wide Web (~1PB)	Wikipedia (10GB) 100% CAGR
Annual Email Traffic, no spam (300PB+)	Internet Archive (1PB+)	Estimated On-line RAM in Google (8PB)	Personal Digital Photos (1000PB+) 100% CAGR
200 of London's Traffic Cams (8TB/day)	2004 Walmart Transaction DB (500TB)	Typical Oil Company (350TB+)	Merck Bio Research DB (1.5TB/qtr)
UPMC Hospitals Imaging Data (500TB/yr)	MIT Babytalk Speech Experiment (1.4PB)	Terashake Earthquake Model of LA Basin (1PB)	One Day of Instant Messaging in 2002 (750GB)
Total digital data to be created this year 270,000PB (IDC)			

Phillip B. Gibbons, Data-Intensive Computing Symposium

Source: <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf>

Source: http://lib.stanford.edu/files/see_pasig_dic.pdf

IT'S THE DATA, STUPID!



「笨蛋！重點在經濟」

(**"It's the economy, stupid"**)

卡維爾 (**James Carville**) 自創這句標語，促使柯林頓當上美國第 **42** 屆總統。

- **1992** 年

「笨蛋！重點還是在經濟」

(**"It's STILL the economy, stupid"**)

卻讓小布希嘲笑是幼稚的總統。

- **2002** 年

雲端時代，谷歌會說：「笨蛋！重點在資料」

(**"It's the data, stupid"**)

誰掌握了你的資料，就有機會掌握你的荷包
想想看，電腦、手機掉了，您心疼的是甚麼呢？

- **2007** 年

Reference Cloud Architecture

雲端運算的參考架構

應用軟體 **Application**

Social Computing, Enterprise, ISV, ...

程式語言 **Programming**

Web 2.0 介面, Mashups, Workflows, ...

控制管理 **Control**

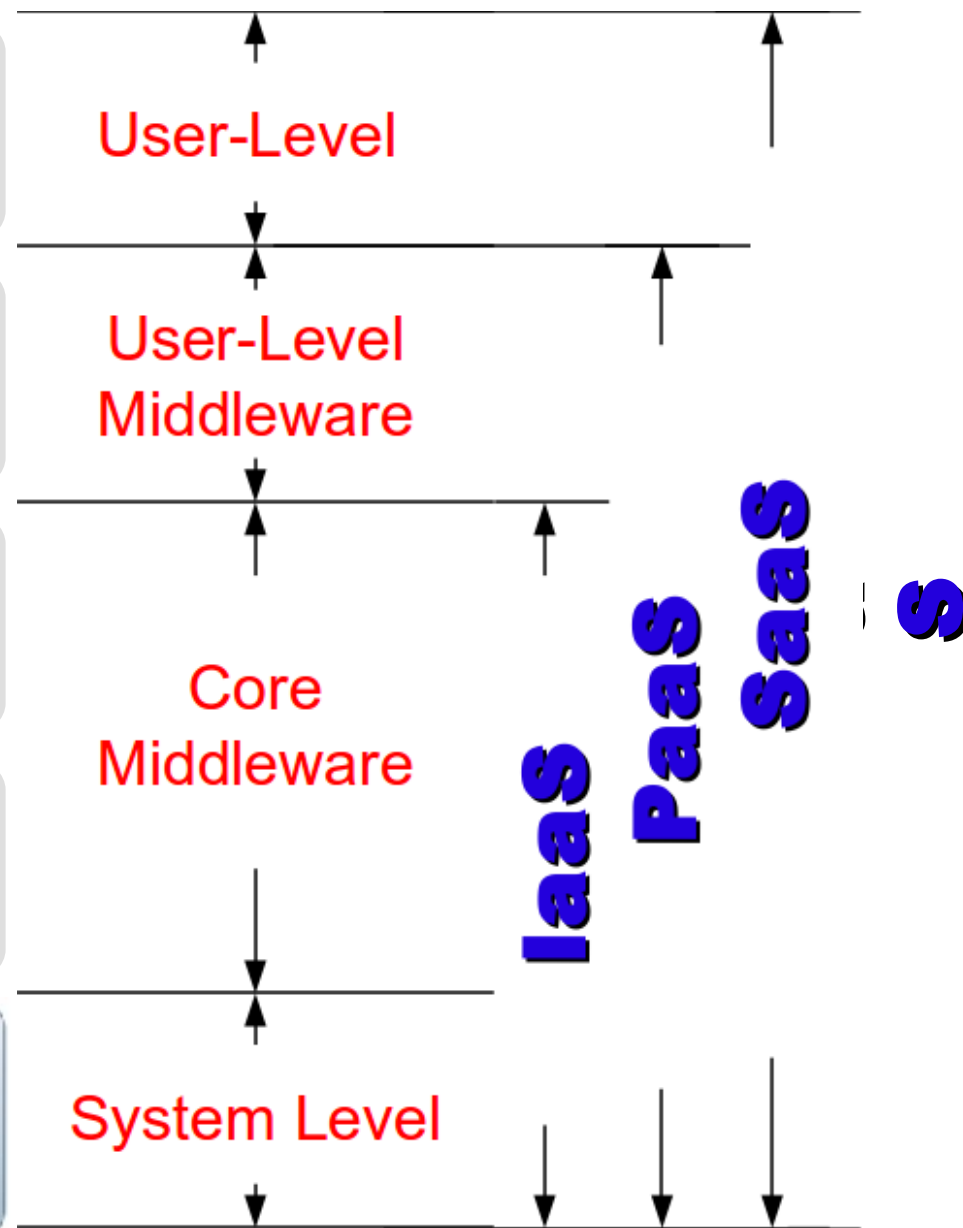
Qos Negotiation, Admission Control, Pricing, SLA Management, Metering...

虛擬化 **Virtualization**

VM, VM management and Deployment

硬體設施 **Hardware**

Infrastructure: Computer, Storage, Network



Open Source to build Private Cloud

建構私有雲端的自由軟體

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

端

平板行動應用

社交溝通協作

多媒體內容

次世代分析

社交分析

情境感知運算

儲存等級記憶體

無所不在的運算

模組化基礎建設

雲端運算

SaaS :
Web 2.0

PaaS :
Big Data

IaaS :
Virtualization

社交網路

評價排行榜

即時搜尋

智慧裝置

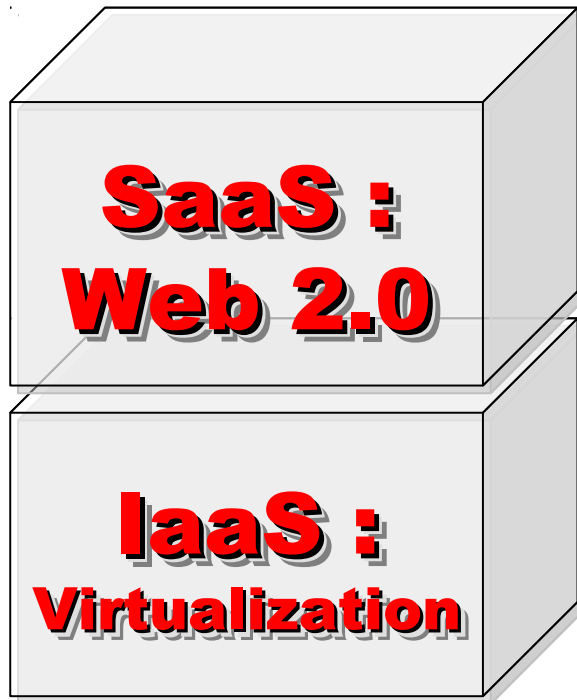
大量資訊分析

雲端運算

雲

Two Type of Cloud Architecture ?

雲端架構的兩大陣營？



想盡辦法誘你用計算跟網路
Computing Intensive



想盡辦法誘你提供資料作分析
Data Intensive

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

Three Core Technologies of Google

Google 的三大關鍵技術

- 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>



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/>

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, Twitter
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 [MalStone 效能評比軟體](#)
- **Open Cloud Consortium provide Open Cloud Testbed and develop MalStone toolkit for benchmark**



National Center for Data Mining
University of Illinois at Chicago



Open Data Group
<http://www.opendatagroup.com/>

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

Total Nodes	4000
Total cores	30000
Data	16PB

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



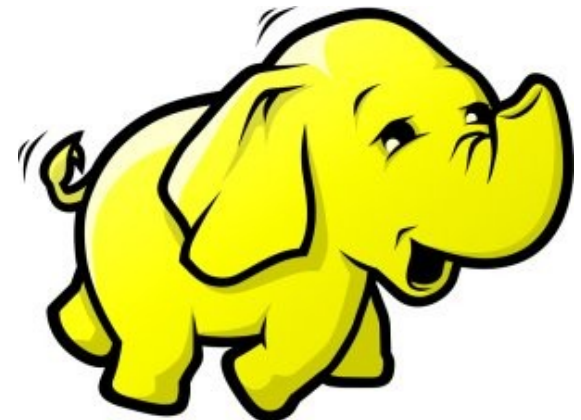
Hadoop 簡介：源起與術語

Introduction to Hadoop : History and Terminology

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



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 in 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>



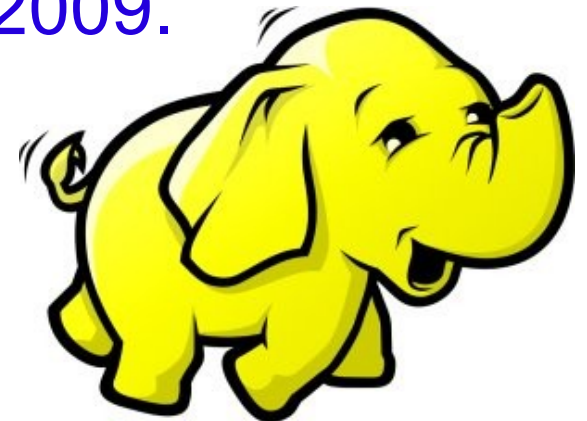
History of Hadoop ... 2004 ~ Now

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



Ticket #HADOOP-1 @ 2006-02-01

Hadoop 這套軟體的起源紀錄 ... 2006年二月一日



The Apache Software Foundation

<http://www.apache.org/>

Log In ▾

Quick Search

Dashboards ▾

Projects ▾

Issues ▾

Agile ▾



Hadoop Common / HADOOP-1

initial import of code from Nutch

Log In

Views ▾

Details

Type:	Task	Status:	Closed
Priority:	Major	Resolution:	Fixed
Affects Version/s:	None	Fix Version/s:	0.1.0
Component/s:	None		
Labels:	None		

People

Assignee:	Doug Cutting
Reporter:	Doug Cutting
Vote (0)	Watch (0)

Description

The initial code for Hadoop will be copied from Nutch.

Dates

Created:	01/Feb/06 02:54
Updated:	03/Aug/06 17:46
Resolved:	04/Feb/06 05:57

Issue Links

This issue is part of:

[NUTCH-193](#) move NDFS and MapReduce to a separate project

Activity

All Comments Work Log History Activity Subversion Commits ▾

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
- **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

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

Total Nodes	4000
Total cores	30000
Data	16PB

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

Comparison between Google and Hadoop

Google 與 *Hadoop* 的比較表

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

Why should we learn Hadoop ?

為何需要學習 **Hadoop ??**

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

simplyhired[®]
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!!



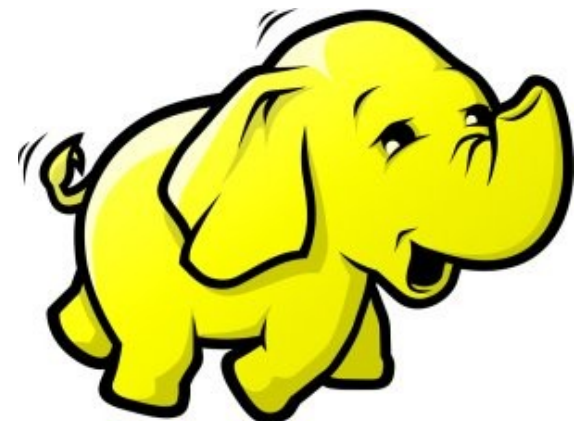
Hadoop 專業術語

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

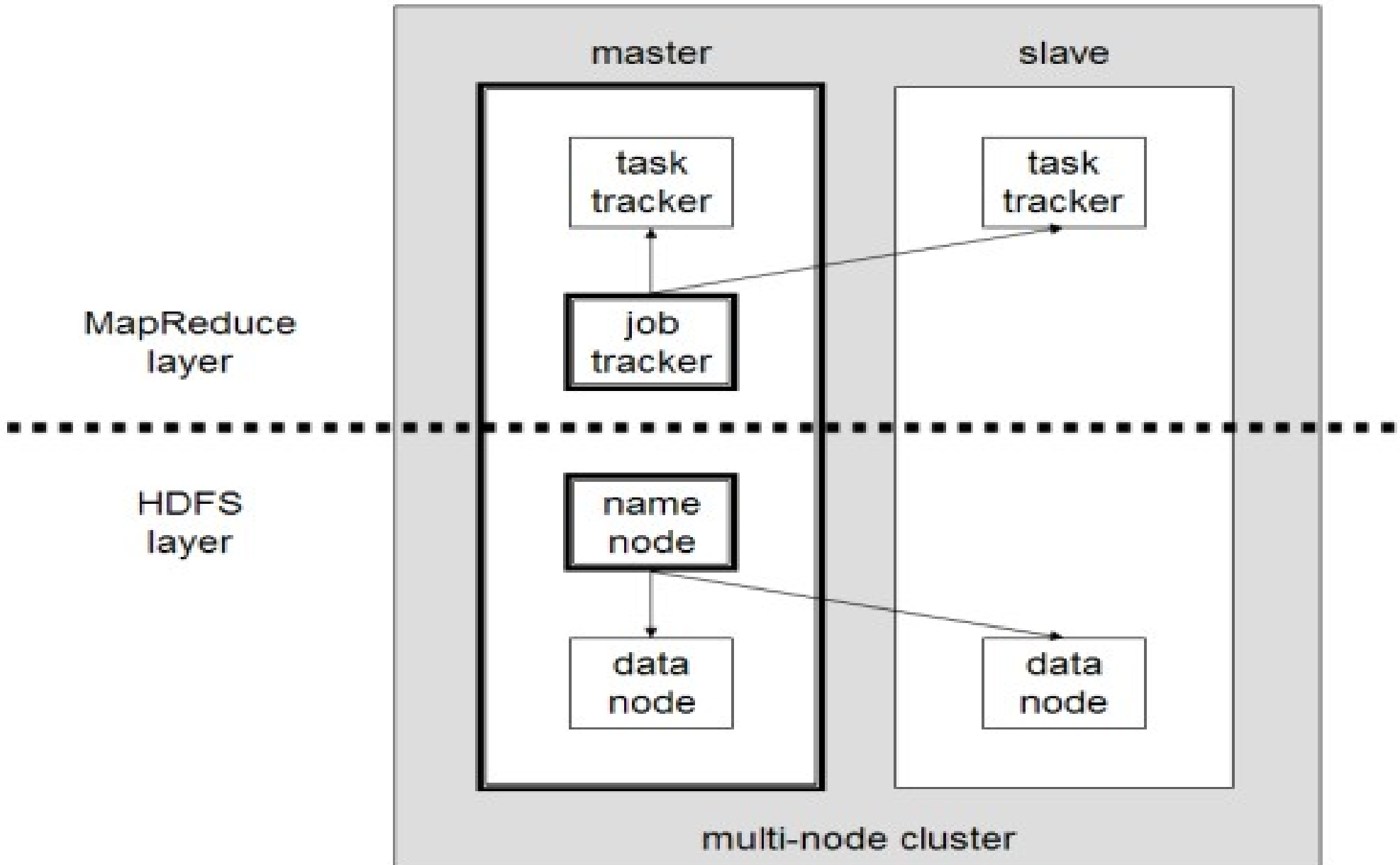
Excute Mapper and
Reducer Tasks

Save Results and
report task status

Multiple Nodes

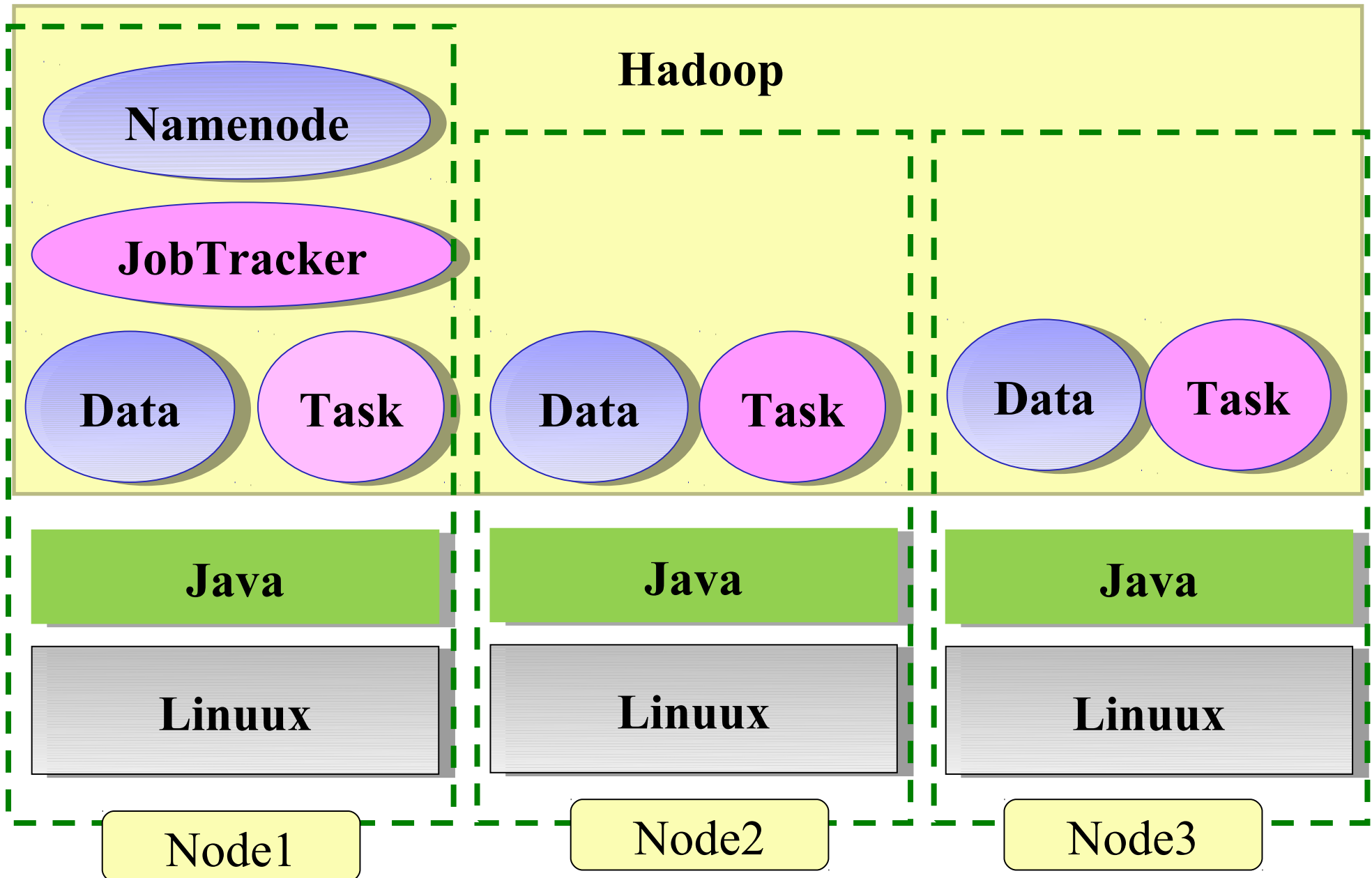
Different Roles of Hadoop Architecture

Hadoop 軟體架構中的不同角色



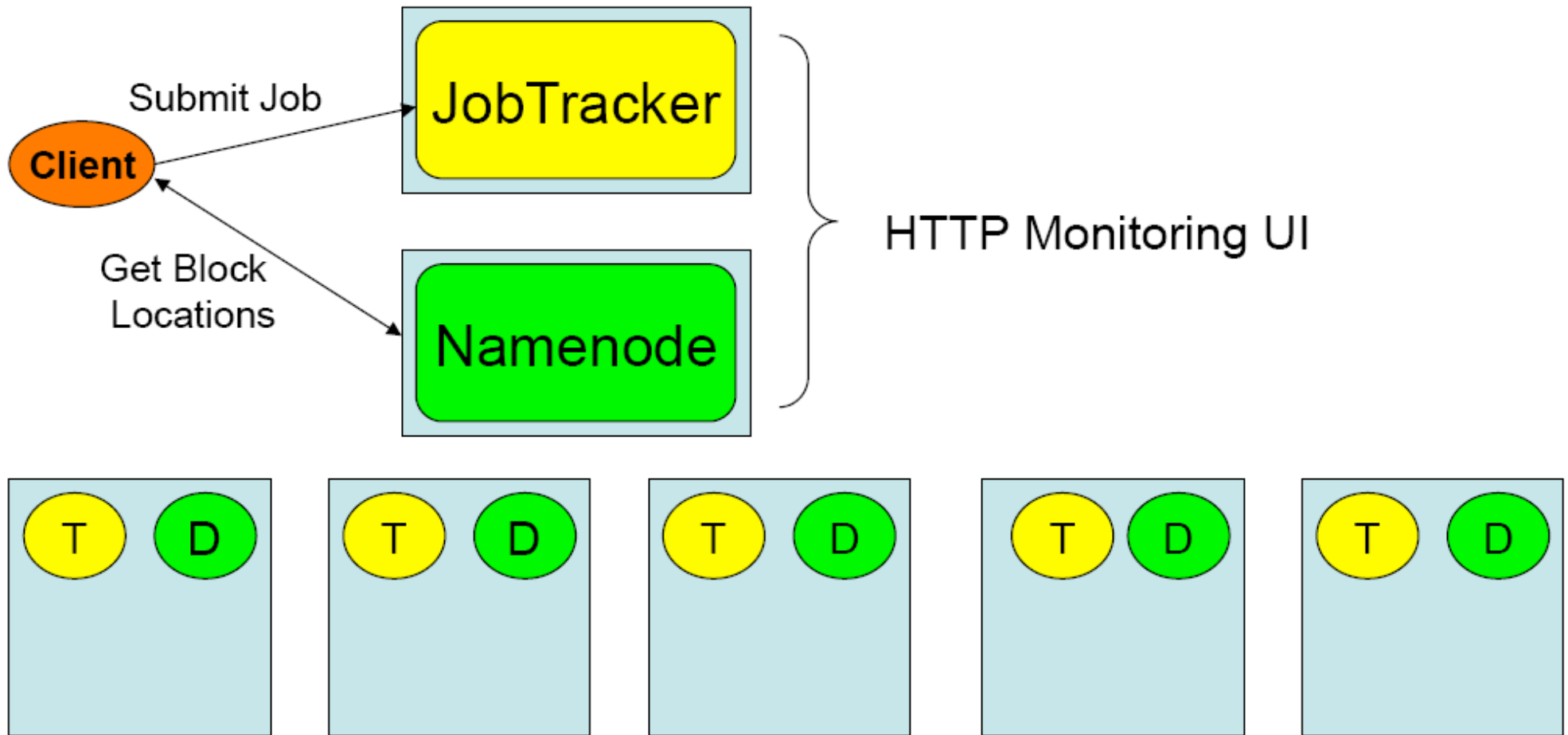
Distributed Operating System of Hadoop

Hadoop 建構成一個分散式作業系統



About Hadoop Client ...

不在雲裡的 *Hadoop Client*



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

建構在大型的個人電腦叢集之上

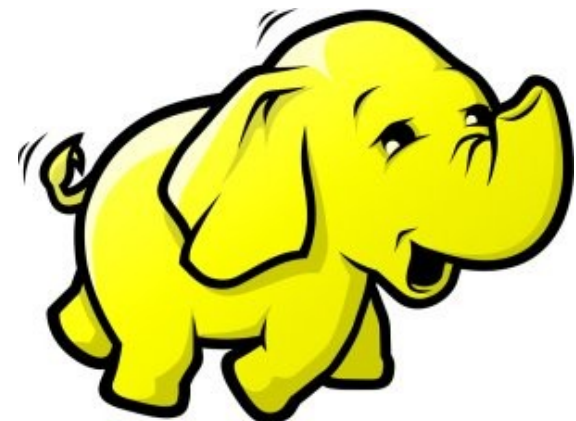
Install on large clusters built of commodity hardware !!



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)

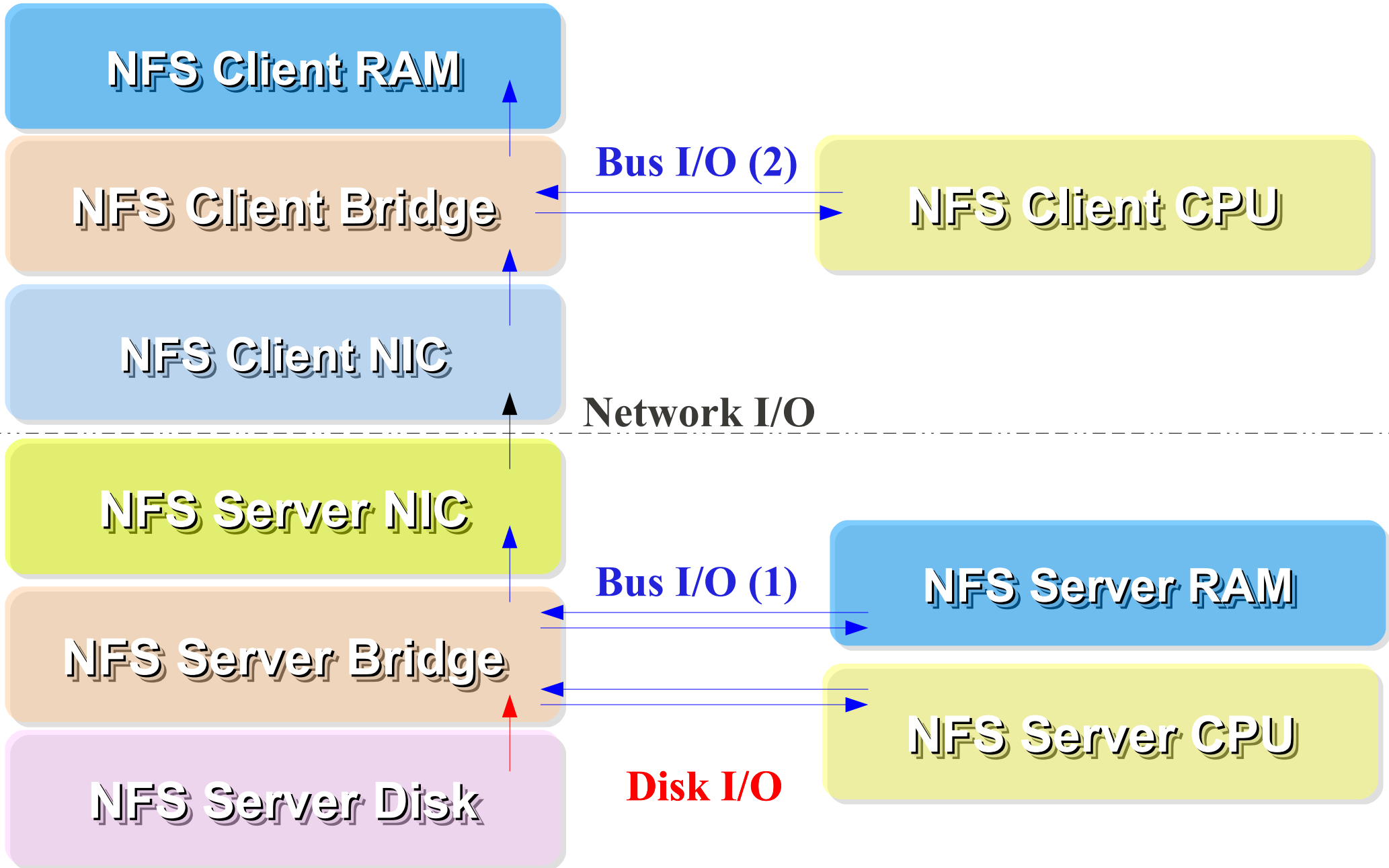
Features of HDFS ...

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

Parallel Computing using NFS storage

使用 **NFS** 進行平行運算



Parallel Computing using HDFS

使用 **HDFS** 進行平行運算

TaskTracker RAM

TaskTracker Bridge

Disk I/O x N Node

DataNode Local Disk

Bus I/O (2)

TaskTracker CPU

Network I/O

TaskTracker NIC

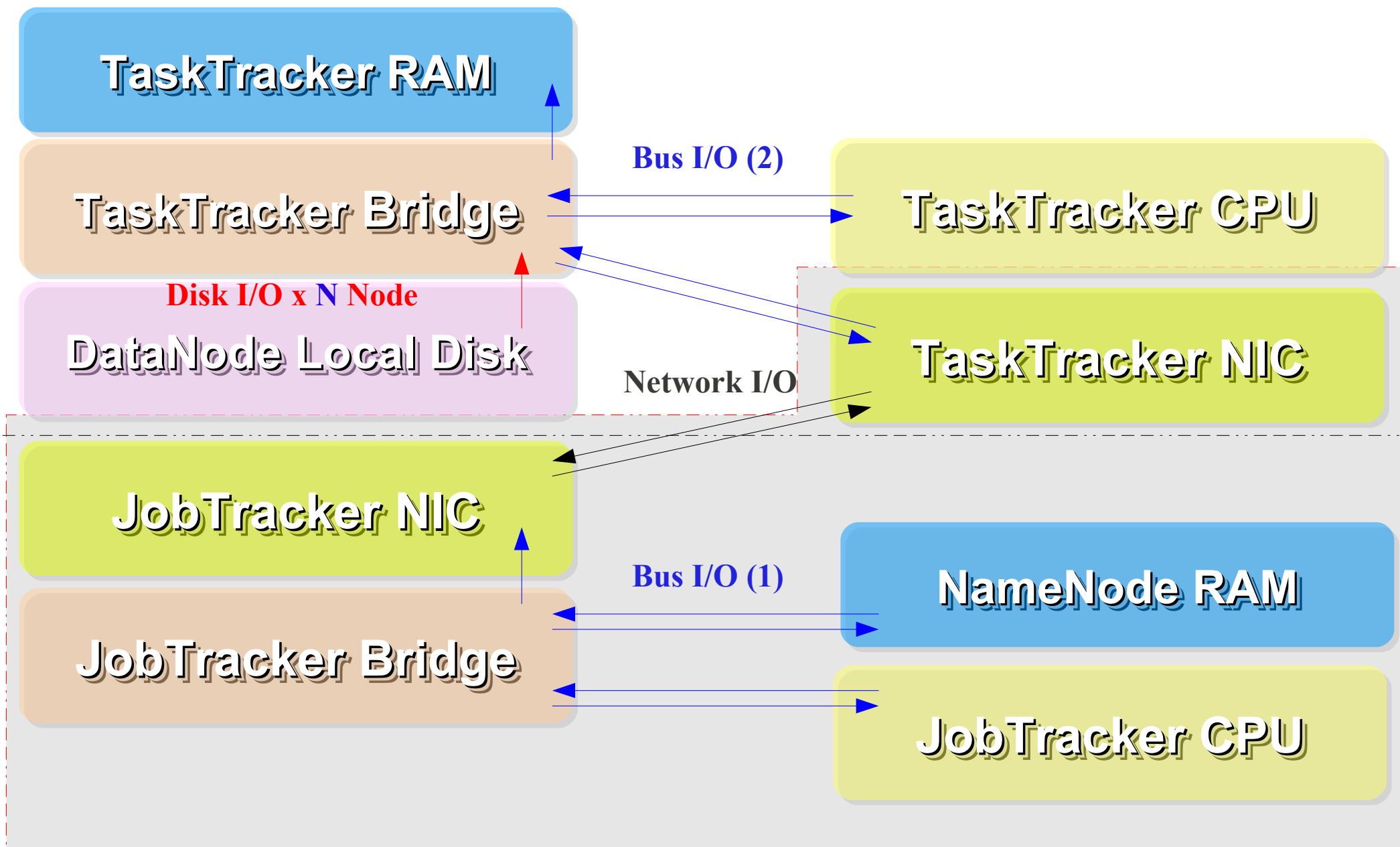
JobTracker NIC

Bus I/O (1)

NameNode RAM

JobTracker Bridge

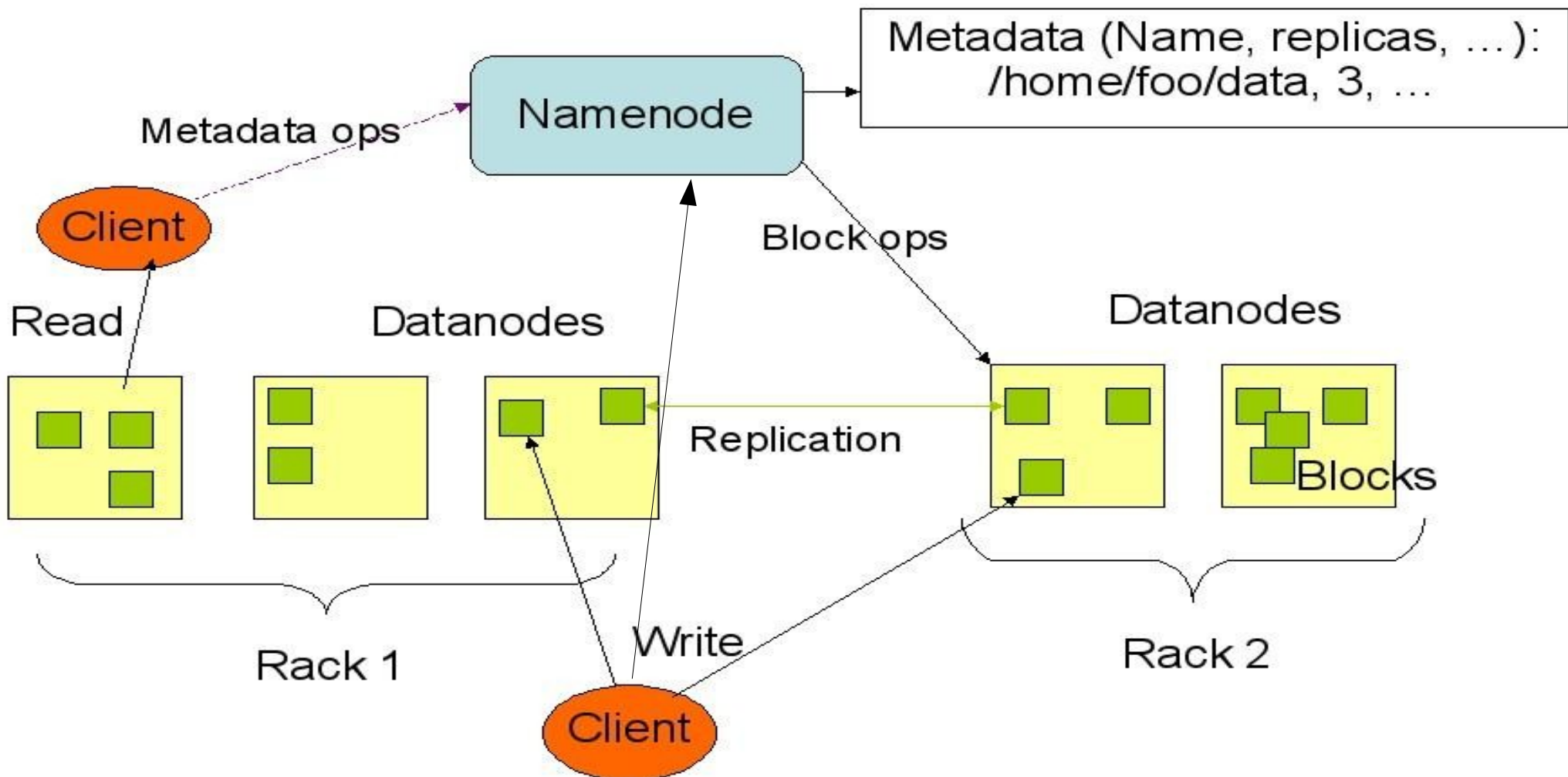
JobTracker CPU



How HDFS manage data ...

HDFS 如何管理資料 ...

HDFS Architecture



How does HDFS work ...

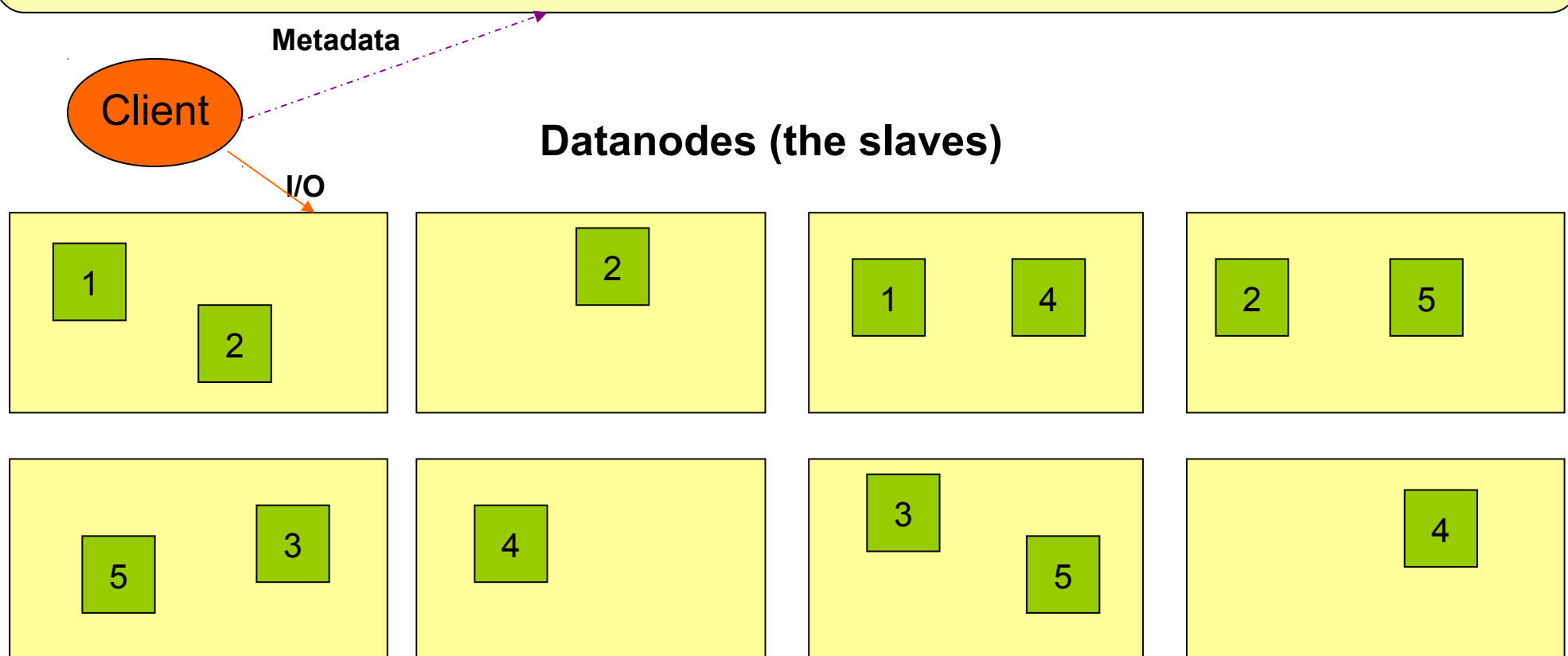
HDFS 如何運作 ...

Namenode (the master)

Path and Filename – **Replication** , **blocks**

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

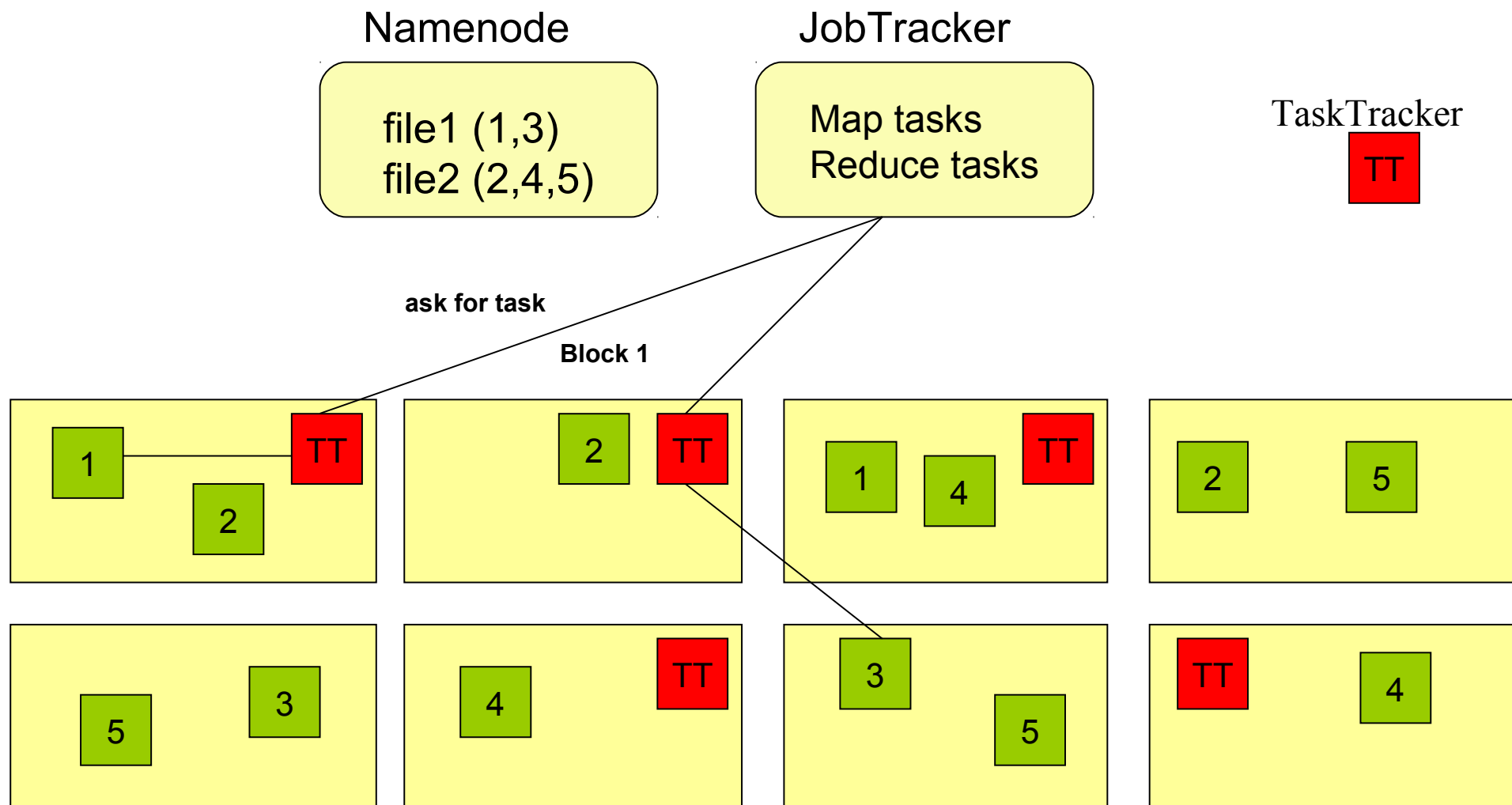
name:/users/bobYahoo/someData.gzip, copies:3, blocks:{2,4,5}



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 ...

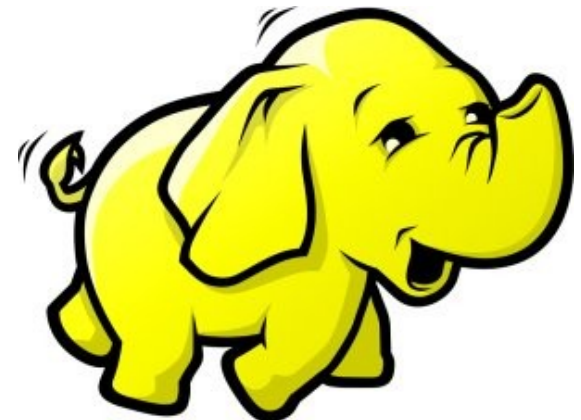
```
jazz@hadoop:~$ hadoop fs
Usage: java FsShell
    [-ls <path>]
    [-lsr <path>]
    [-du <path>]
    [-dus <path>]
    [-count[-q] <path>]
    [-mv <src> <dst>]
    [-cp <src> <dst>]
    [-rm <path>]
    [-rmr <path>]
    [-expunge]
    [-put <localsrc> ... <dst>]
    [-copyFromLocal <localsrc> ... <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

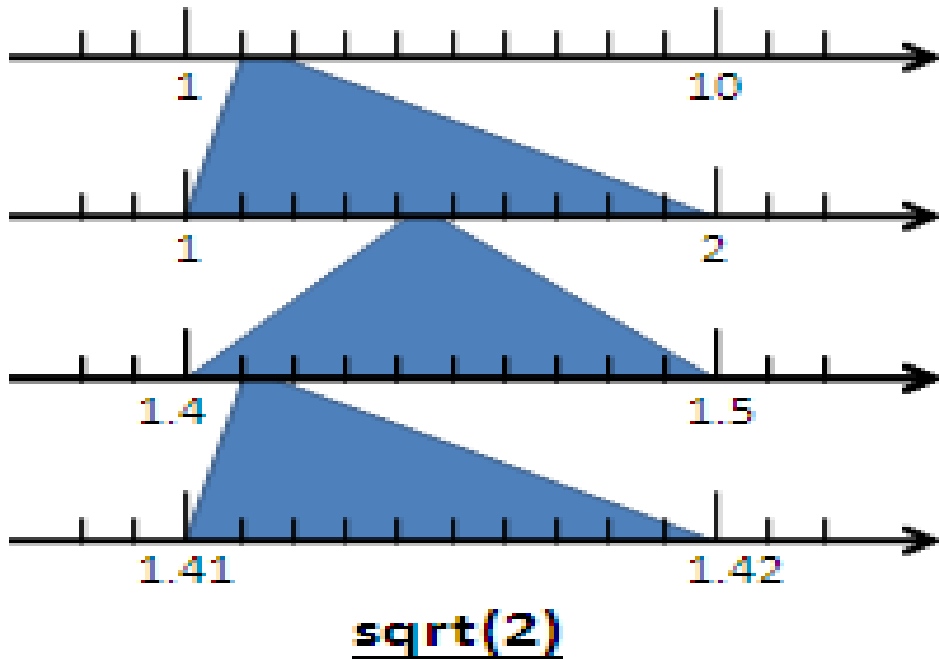
Jazz Wang
Yao-Tsung Wang
jazz@nchc.org.tw



Divide and Conquer Algorithms

分而治之演算法

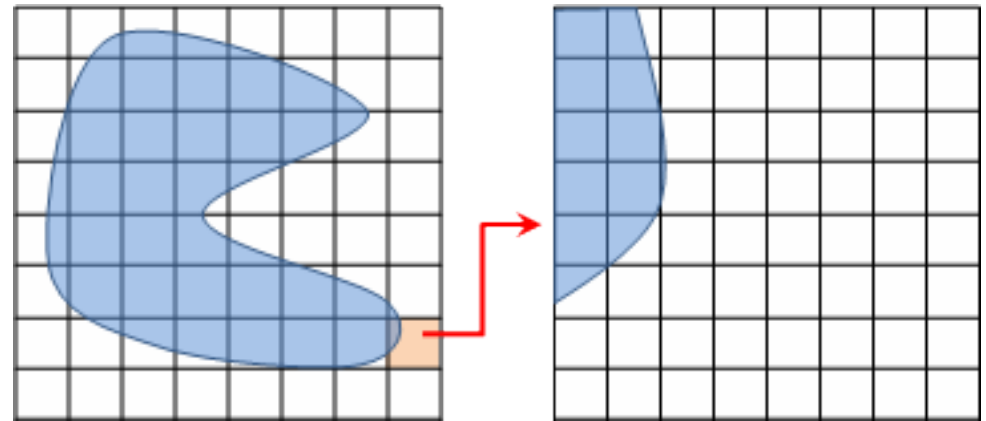
Example 1:



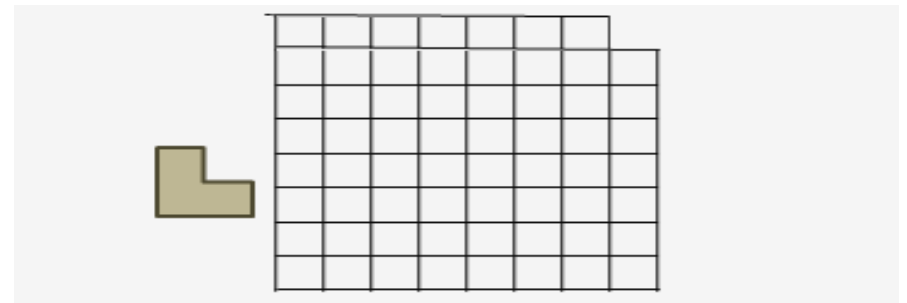
Example 4: The way to climb 5 steps stair within 2 steps each time. 眼前有五階樓梯，每次可踏上一階或踏上兩階，那麼爬完五階共有幾種踏法？

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

Example 2:



Example 3:



What is MapReduce ??

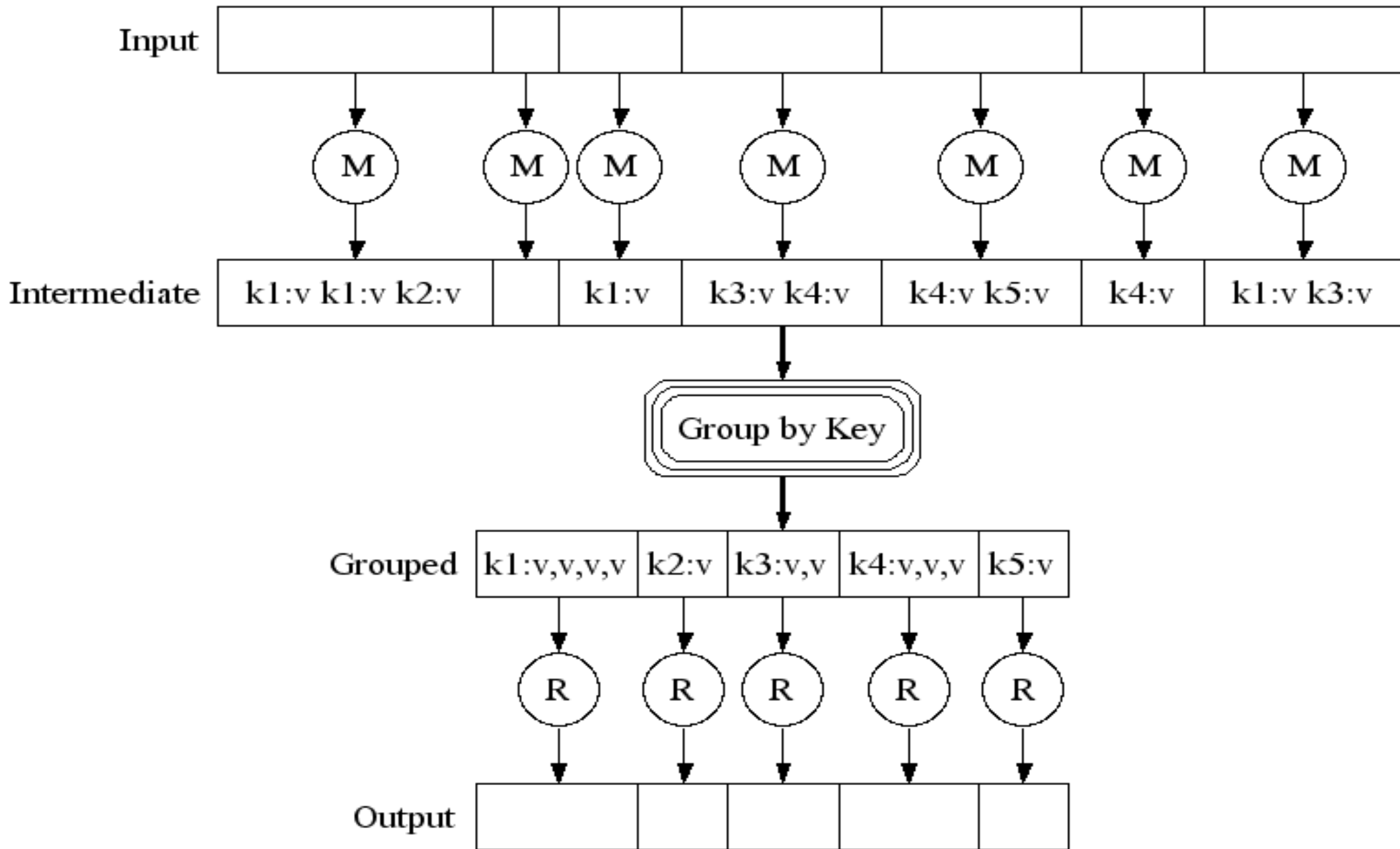
什麼是 *MapReduce* ??

- 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(k3, v3)**

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

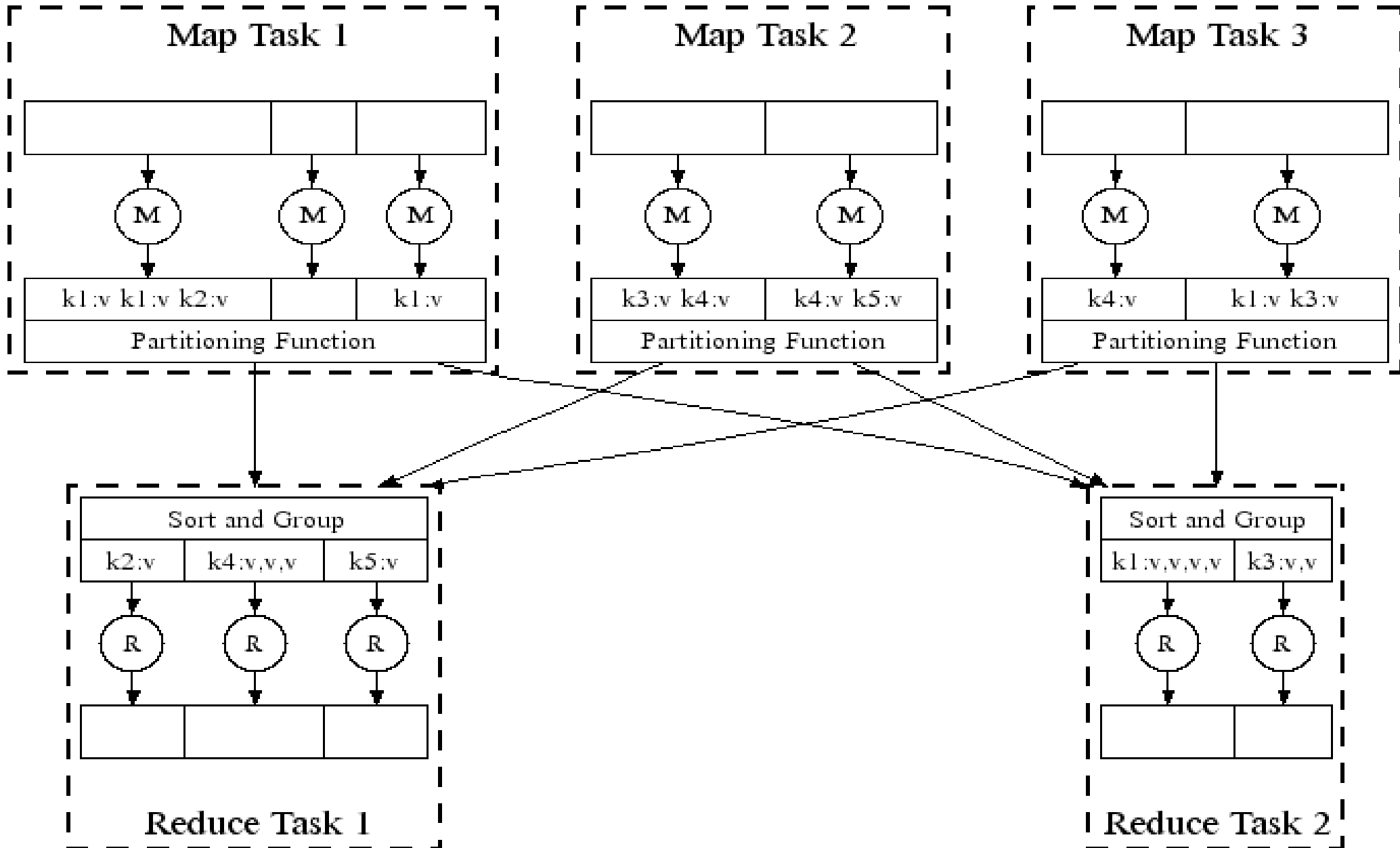
Google's MapReduce Diagram

Google 的 MapReduce 圖解



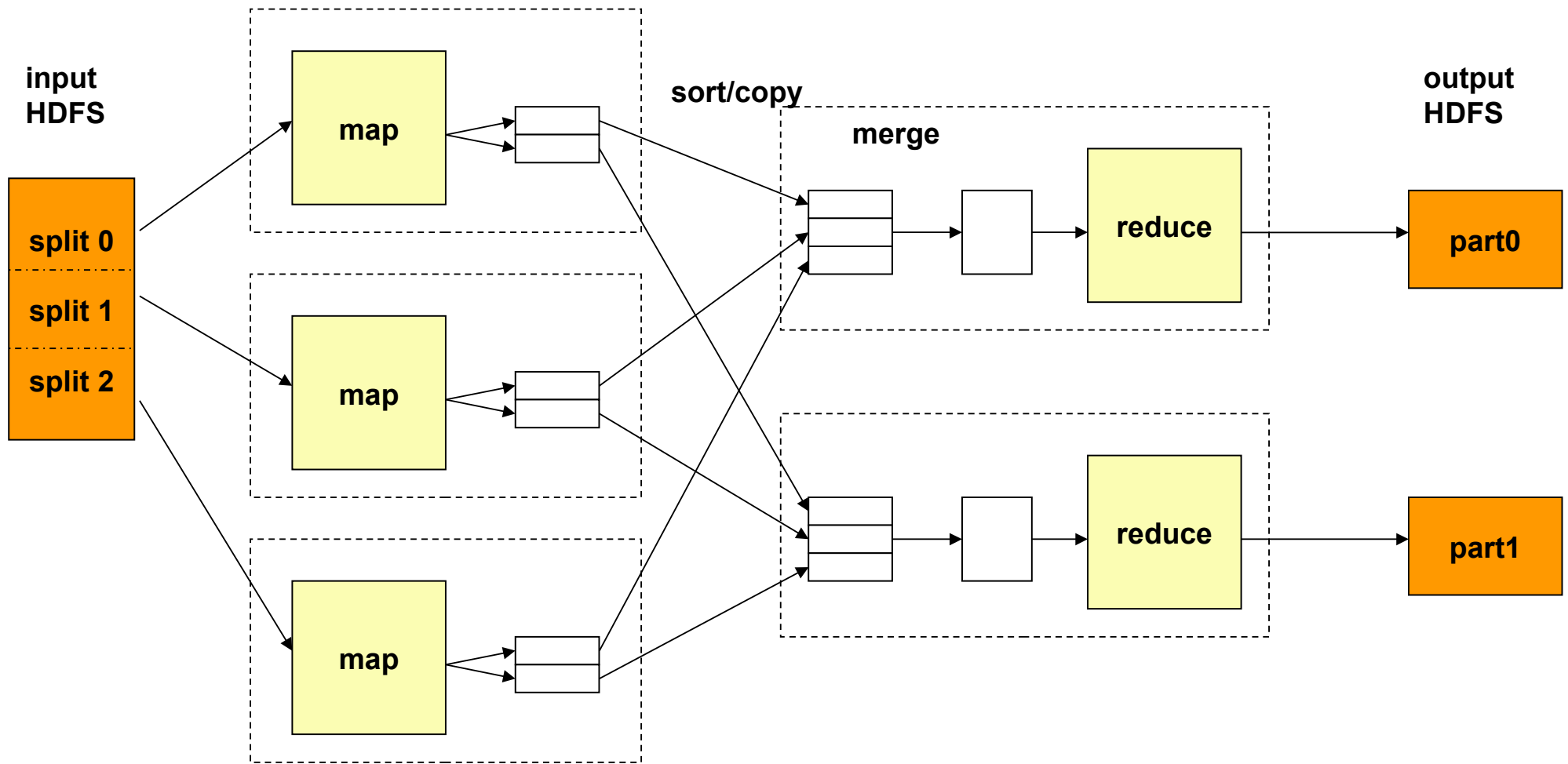
Google's MapReduce in Parallel

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

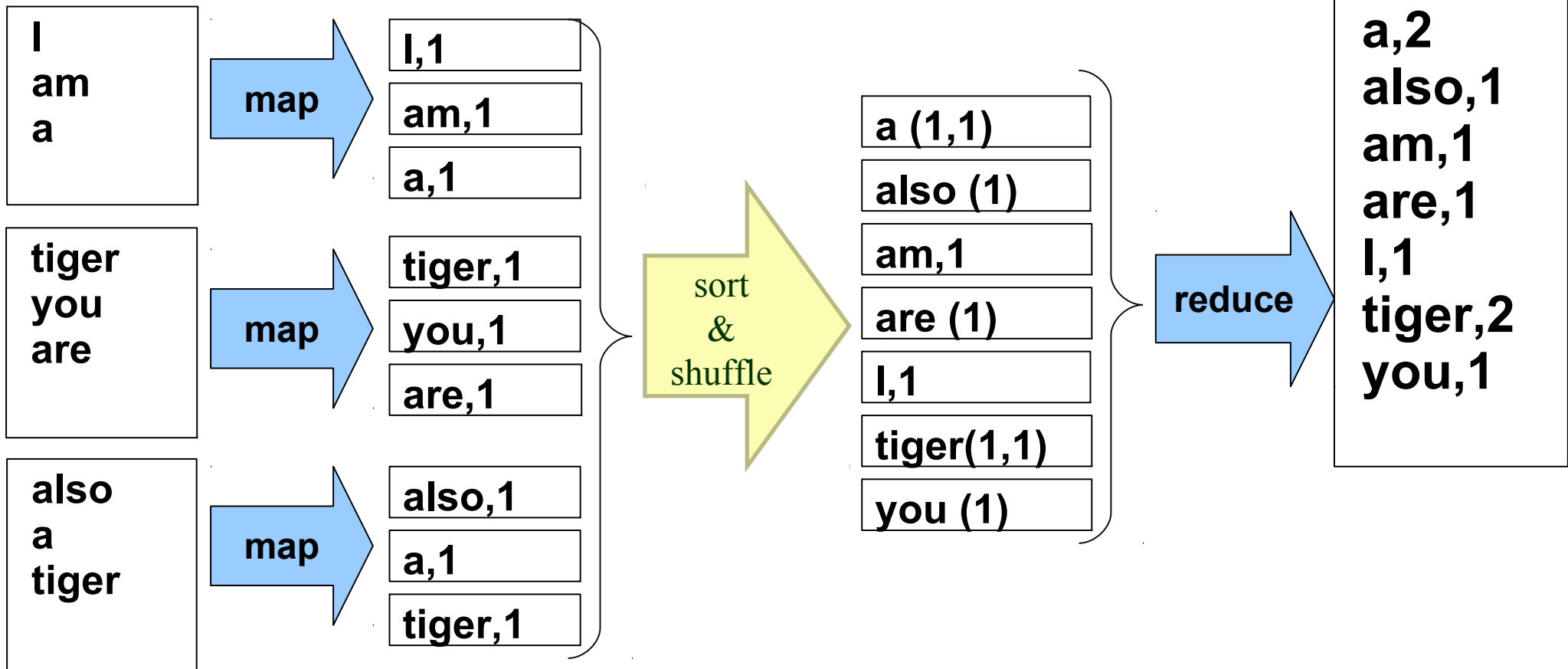
JobTracker 派遣 TaskTracker 作 reduce

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



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/>

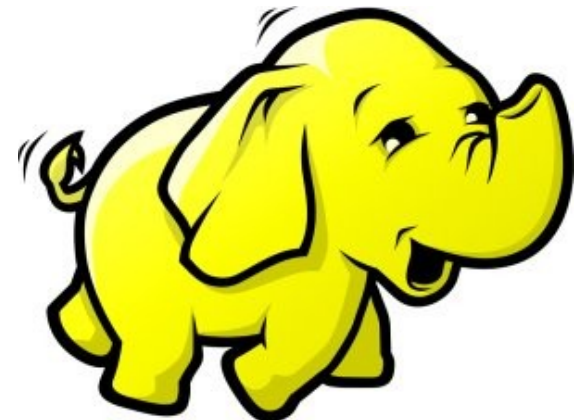
• <http://wiki.apache.org/hadoop/PoweredBy>



Hadoop 相關計畫

Hadoop Ecosystem

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





Hadoop 只支援用 **Java** 開發嘛？
Is Hadoop only support Java ?

總不能全部都重新設計吧？如何與舊系統相容？

Can Hadoop work with existing software ?



可以跟資料庫結合嘛？

Can Hadoop work with Databases ?

開發者們有聽到大家的需求

Yes, we hear the feedback of developers ...



Is Hadoop only support Java ?

- Although the Hadoop framework is implemented in Java[™], **Map/Reduce applications need not be written in Java.**
- **Hadoop Streaming** is a utility which allows users to **create and run jobs with any executables (e.g. shell utilities)** as the mapper and/or the reducer.
- **Hadoop Pipes** is a SWIG-compatible **C++ API** to implement Map/Reduce applications (non JNI[™] based).

Hadoop Pipes (C++, Python)

- Hadoop Pipes allows **C++** code to use Hadoop DFS and map/reduce.
- The C++ interface is "swigable" so that interfaces can be generated for **python** and other scripting languages.
- For more detail, check the API Document of org.apache.hadoop.mapred.pipes
- You can also find example code at hadoop-*/src/examples/pipes
- About the pipes C++ WordCount example code: <http://wiki.apache.org/hadoop/C++WordCount>

Hadoop Streaming

- Hadoop Streaming is a utility which allows users to create and run Map-Reduce jobs **with any executables (e.g. Unix shell utilities)** as the mapper and/or the reducer.
- It's useful when you need to run **existing program** written in shell script, perl script or even PHP.
- Note: both the **mapper** and the **reducer** are **executables** that read the input from **STDIN** (line by line) and emit the output to **STDOUT**.
- For more detail, check the official document of **Hadoop Streaming**

Running Hadoop Streaming

```
jazz@hadoop:~$ hadoop jar hadoop-streaming.jar -help
10/08/11 00:20:00 ERROR streaming.StreamJob: Missing required option -input
Usage: $HADOOP_HOME/bin/hadoop [--config dir] jar \
      $HADOOP_HOME/hadoop-streaming.jar [options]
```

Options:

```
-input      <path>          DFS input file(s) for the Map step
-output     <path>          DFS output directory for the Reduce step
-mapper     <cmd|JavaClassName>   The streaming command to run
-combiner   <JavaClassName>   Combiner has to be a Java class
-reducer    <cmd|JavaClassName>   The streaming command to run
-file       <file>           File/dir to be shipped in the Job jar file
-dfs        <h:p>|local   Optional. Override DFS configuration
-jt         <h:p>|local   Optional. Override JobTracker configuration
-additionalconfspec specfile  Optional.
-inputformat TextInputFormat (default) |SequenceFileAsTextInputFormat |
JavaClassName Optional.
-outputformat TextOutputFormat (default) |JavaClassName  Optional.
```

... More ...

Hadoop Streaming with shell commands (1)

```
hadoop:~$ hadoop fs -rmr input output
```

```
hadoop:~$ hadoop fs -put /etc/hadoop/conf input
```

```
hadoop:~$ hadoop jar hadoop-streaming.jar -input  
input -output output -mapper /bin/cat -reducer  
/usr/bin/wc
```


Hadoop Streaming with shell commands (2)

```
hadoop:~$ echo "sed -e \"s/ /\n/g\" | grep ." >  
streamingMapper.sh
```

```
hadoop:~$ echo "uniq -c | awk '{print \  
$2 \"\t\" \"$1}'" > streamingReducer.sh
```

```
hadoop:~$ chmod a+x streamingMapper.sh
```

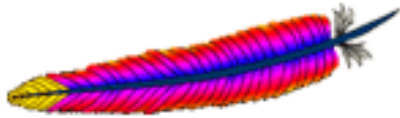
```
hadoop:~$ chmod a+x streamingReducer.sh
```

```
hadoop:~$ hadoop fs -put /etc/hadoop/conf input
```

```
hadoop:~$ hadoop jar hadoop-streaming.jar -input  
input -output output -mapper streamingMapper.sh  
-reducer streamingReducer.sh -file  
streamingMapper.sh -file streamingReducer.sh
```

There are several Hadoop subprojects

Apache > Hadoop >



Top

Common

Chukwa

HBase

HDFS

Hive

MapReduce

Pig

ZooKeeper

▼ About

▫ Welcome

▫ Who We Are?

▫ Mailing Lists

Welcome to Apache Hadoop!

- **Hadoop Common:** The common utilities that support the other Hadoop subprojects.
- **HDFS:** A distributed file system that provides high throughput access to application data.
- **MapReduce:** A software framework for distributed processing of large data sets on compute clusters.

Other Hadoop related projects

- **Chukwa**: A data collection system for managing large distributed systems.
- **HBase**: A scalable, distributed database that supports structured data storage for large tables.
- **Hive**: A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Pig**: A high-level data-flow language and execution framework for parallel computation.
- **ZooKeeper**: A high-performance coordination service for distributed applications.

Hadoop Ecosystem

Pig	Chukwa	Hive	HBase
MapReduce		HDFS	ZooKeeper
Hadoop Core (Hadoop Common)		Avro	

Source: *Hadoop: The Definitive Guide*

Avro

- Avro is a **data serialization system**.
- It provides:
 - *Rich data structures.*
 - *A compact, fast, binary data format.*
 - *A container file, to store persistent data.*
 - *Remote procedure call (RPC).*
 - *Simple integration with dynamic languages.*
- Code generation is not required to read or write data files nor to use or implement RPC protocols. Code generation as an optional optimization, only worth implementing for statically typed languages.
- For more detail, please check the official document:
<http://avro.apache.org/docs/current/>



Zoo Keeper



- <http://hadoop.apache.org/zookeeper/>
- ZooKeeper is a **centralized service** for **maintaining configuration** information, naming, **providing distributed synchronization**, and providing group services. All of these kinds of services are used in some form or another by distributed applications.
- *Each time they are implemented there is a lot of work that goes into fixing the bugs and **race conditions** that are inevitable. Because of the difficulty of implementing these kinds of services, applications initially usually skimp on them, which make them brittle in the presence of change and difficult to manage. Even when done correctly, different implementations of these services lead to management complexity when the applications are deployed.*

Pig

- <http://hadoop.apache.org/pig/>
- Pig is a platform for **analyzing large data sets** that consists of a **high-level language** for expressing data analysis programs, coupled with infrastructure for evaluating these programs.
- Pig's infrastructure layer consists of a **compiler** that produces sequences of **Map-Reduce programs**
- Pig's language layer currently consists of a textual language called **Pig Latin**, which has the following key properties:
 - **Ease of programming**
 - **Optimization opportunities**
 - **Extensibility**



Hive

- <http://hadoop.apache.org/hive/>
- Hive is a **data warehouse** infrastructure built on top of Hadoop that provides tools to enable easy **data summarization**, **adhoc querying** and analysis of large datasets data stored in Hadoop files.
- **Hive QL** is based on SQL and enables users familiar with SQL to query this data.



Chukwa

- <http://hadoop.apache.org/chukwa/>
- Chukwa is an open source **data collection system** for monitoring large distributed systems.
- built on top of HDFS and Map/Reduce framework
- includes a flexible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.



Mahout

- <http://mahout.apache.org/>
- Mahout is a scalable **machine learning libraries**.
- implemented on top of Apache Hadoop using the map/reduce paradigm.
- Mahout currently has
 - Collaborative Filtering
 - User and Item based recommenders
 - **K-Means, Fuzzy K-Means clustering**
 - Mean Shift clustering
 - More ...





Hadoop 與 HBase 簡易安裝 (單機模式)

Hadoop4Win : an Easy Way to install Hadoop and HBase on Windows

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



Powered by DRBL

 Search

[Login](#) | [Preferences](#) | [Help/Guide](#) | [About Trac](#) | [Forgot your password?](#)

	Wiki	Timeline	Roadmap	Browse Source	View Tickets	Search
--	------	----------	---------	---------------	--------------	--------

[Start Page](#) | [Index](#) | [History](#) | [Last Change](#)

hadoop4win

-- Hadoop for Windows using Cygwin

軟體簡介

hadoop4win, 顧名思義為『Hadoop for Windows』, 主要是提供 Windows 平台上簡易安裝 Hadoop 的批次安裝檔。此批次安裝檔內容, 主要參考自國網中心企鵝龍與再生龍團隊成員孫振凱先生之 [drbl-winroll](#) 作品, 抽取安裝部分程式改寫成 hadoop4win 所需的步驟。

hadoop4win 目前包含五大軟體組成:

- [Cygwin](#) - 提供精簡版, 類似 Linux 的環境
- [JDK 1.6.0 update 18](#) - 運行 Hadoop 必須的 JRE(Java Runtime Environment) 與編譯程式所需之 javac 編譯器
- [Hadoop 0.20.2](#) - 包含 Hadoop 0.20.2 原始程式與中英文說明文件檔
- [HBase 0.20.6](#) - 包含 HBase 0.20.6 原始程式碼
- [Ant 1.8.2](#) - 包括 Apache Ant 1.8.2 執行檔

硬體需求

- 已知最低 512 MB 記憶體需求, 建議至少 1024 MB。
- 安裝相關軟體至少需要 500 MB 以上硬碟空間。

軟體需求

- Windows 2000, Windows XP
- 目前已知 **Windows 7 無法正常執行**

hadoop4win
軟體簡介
硬體需求
軟體需求
檔案下載
源碼下載
改版紀錄
臭蟲回報
安裝方法 (1) 安裝檔
安裝方法 (2) 批次檔
反安裝方法 (1) 安裝檔
反安裝方法 (2) 批次檔
測試方法
測試 Hadoop 的步驟
測試 HBase 的步驟
測試 WordCount 編譯
關閉視窗
電腦重開
已知問題



Questions?

Slides - <http://trac.nchc.org.tw/cloud>

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



Powered by DRBL