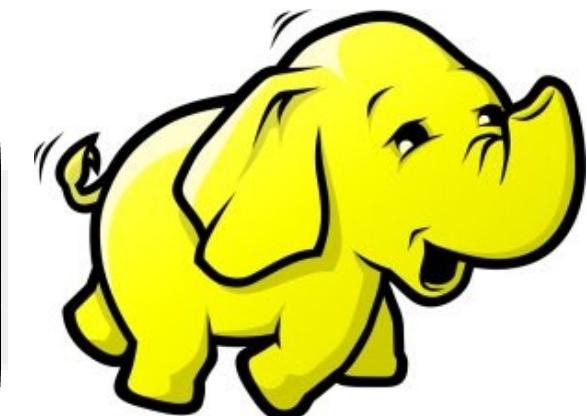




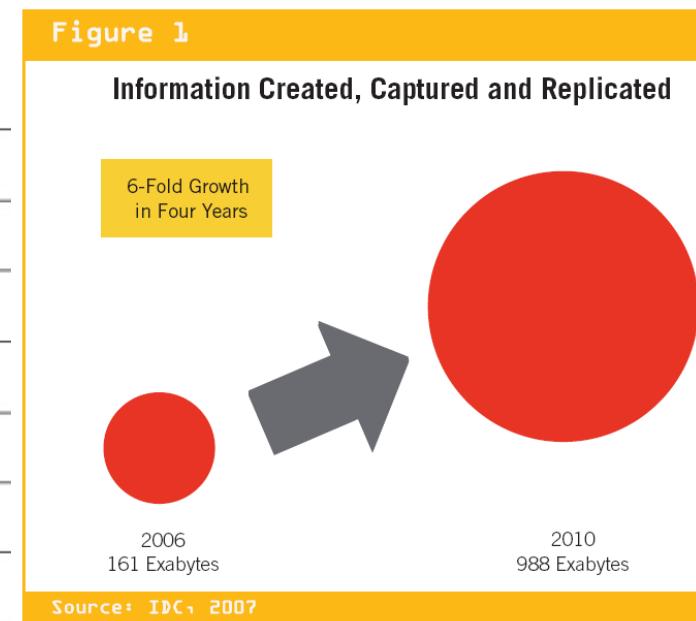
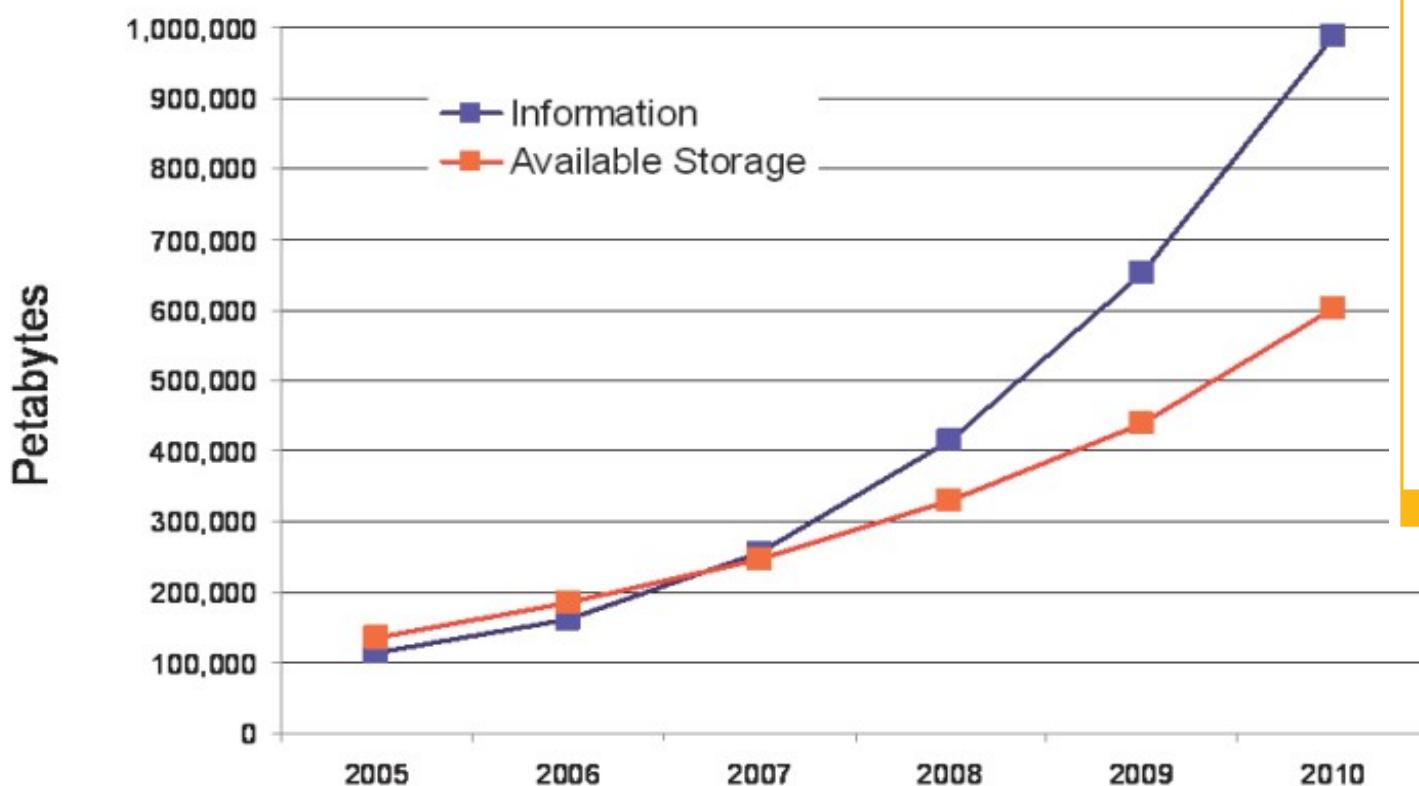
運用 hiCloud 打造 Hadoop 簿集
Build Your Own Hadoop Cluster on hiCloud

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



Data Explosion!! 始於 2007 的「資料大爆炸」時代

Information Versus Available Storage



2007 年，IDC 預估
2010 年會成長六倍！
(相較 2006 年)

Source: IDC, 2007

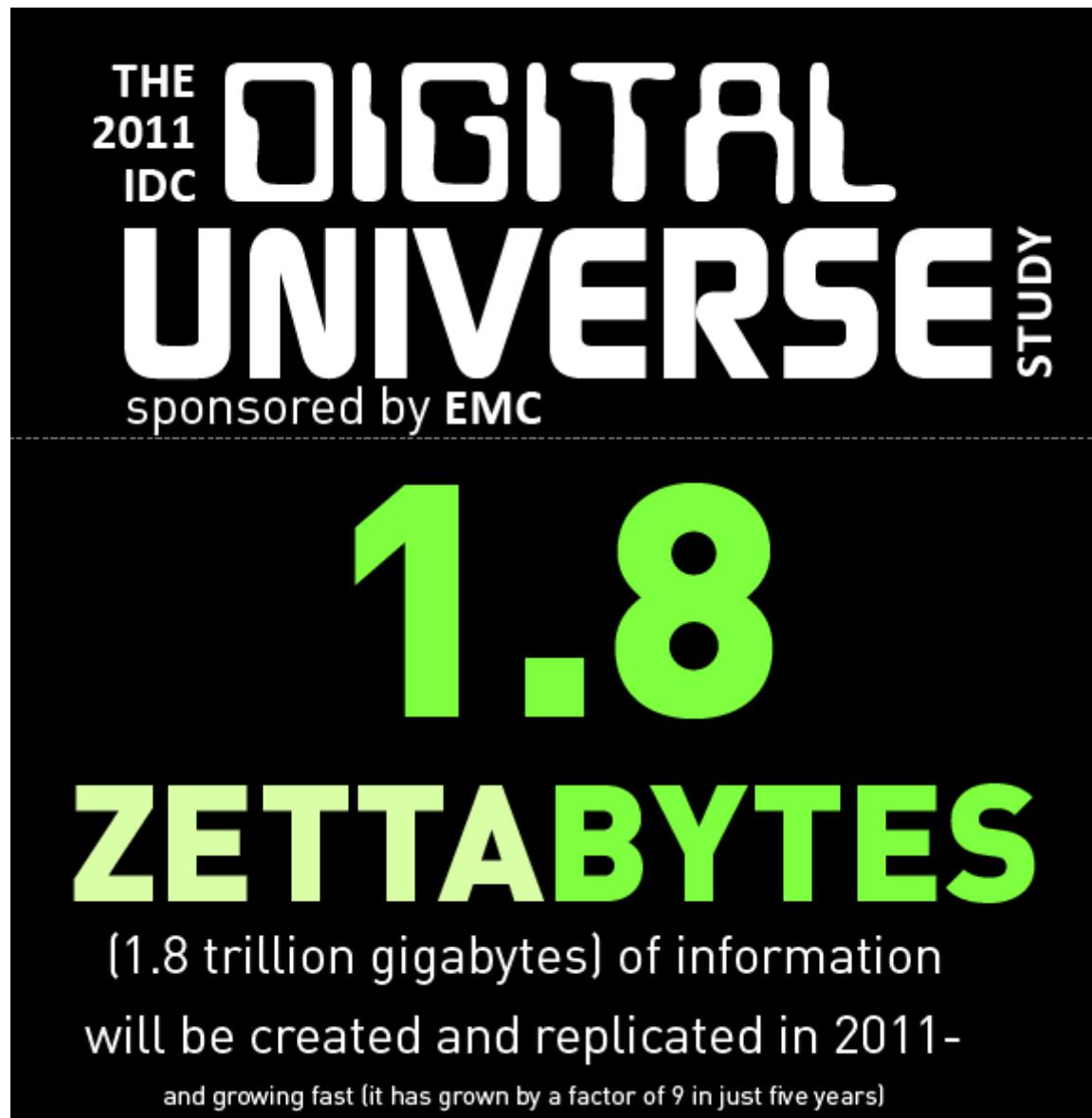
出處：[The Expanding Digital Universe](#),

A Forecast of Worldwide Information Growth Through 2010,
March 2007, An IDC White Paper - sponsored by EMC

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

2006 161 EB
2010 988 EB (預測)

Data expanded 2x each year !! 每年約略兩倍



出處 : Extracting Value from Chaos,
June 2011, An IDC White Paper - sponsored by EMC
<http://www.emc.com/collateral/about/news/idc-emc-digital-universe-2011-infographic.pdf>

追蹤歷年的 IDC 數據：

2006	161	EB
2007	281	EB
2008	487	EB
2009	800	EB (0.8 ZB)
2010	988	EB (預測)
2010	1200	EB (1.2 ZB)
2011	1773	EB (預測)
2011	1800	EB (1.8 ZB)

景氣差而成長趨緩？
或受新技術抑制？

What is Big Data?! 何謂『巨量資料』？

巨量資料泛指資料大小已無法用一般軟體擷取、管理與處理；
單一資料集大小介於數十 TB 至數 PB 的資料。

'Big Data' = few dozen TeraBytes to PetaBytes in single data set.

Definition

[edit]



Big data is a term applied to data sets whose size is beyond the ability of commonly used software tools to capture, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target currently ranging from a few dozen terabytes to many petabytes of data in a single data set.

In a 2001 research report^[14] and related conference presentations, then META Group (now Gartner) analyst, Doug Laney, defined data growth challenges (and opportunities) as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in/out), and variety (range of data types, sources). Gartner continues to use this model for describing big data.^[15]

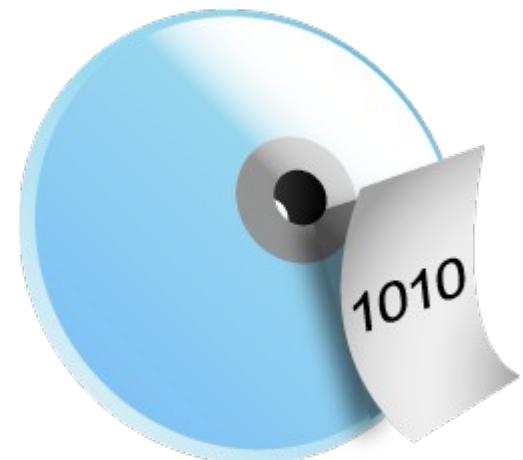
出處：http://en.wikipedia.org/wiki/Big_data



多個檔案，容量 100TB



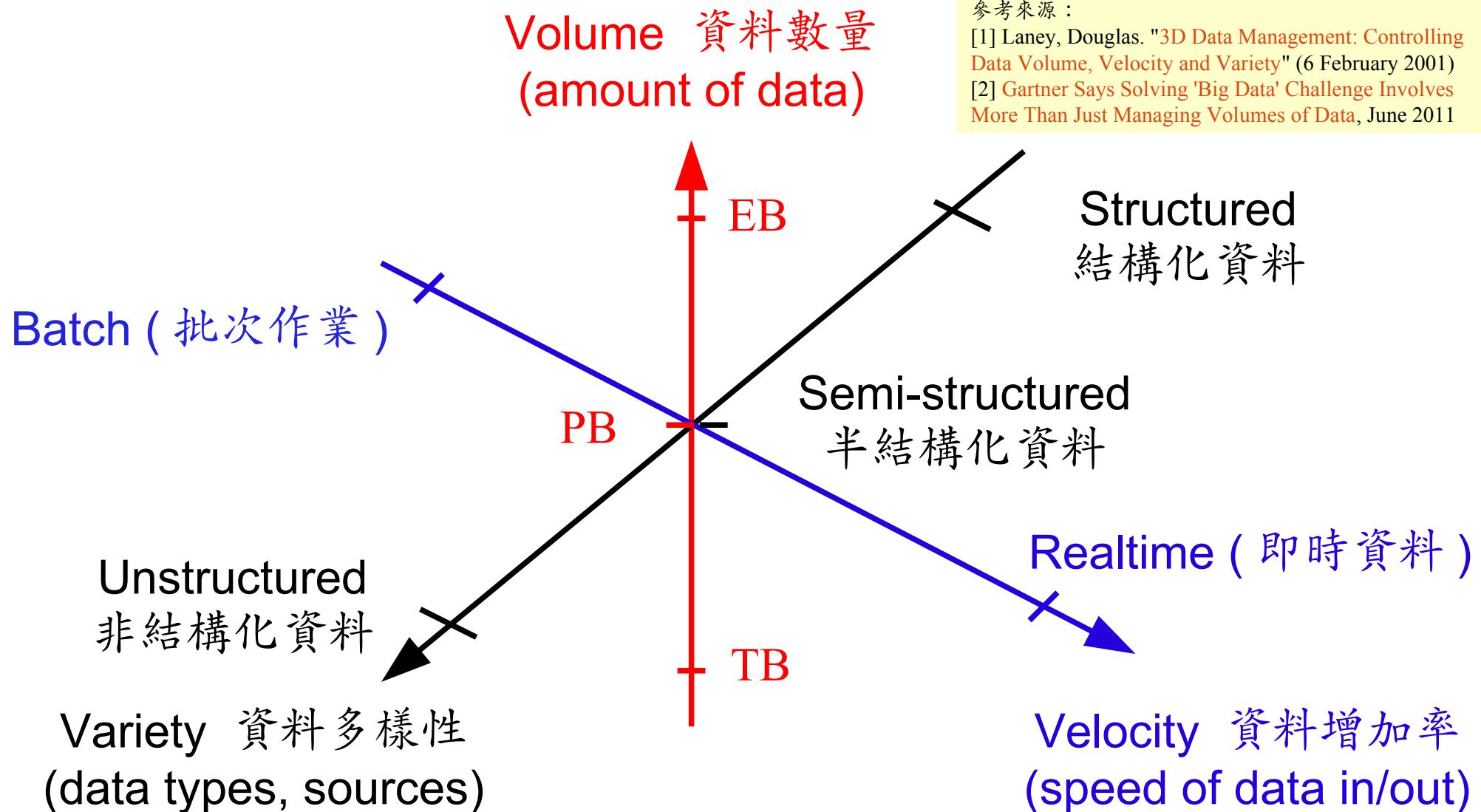
一個資料庫，容量 100TB



一個檔案，容量 100TB

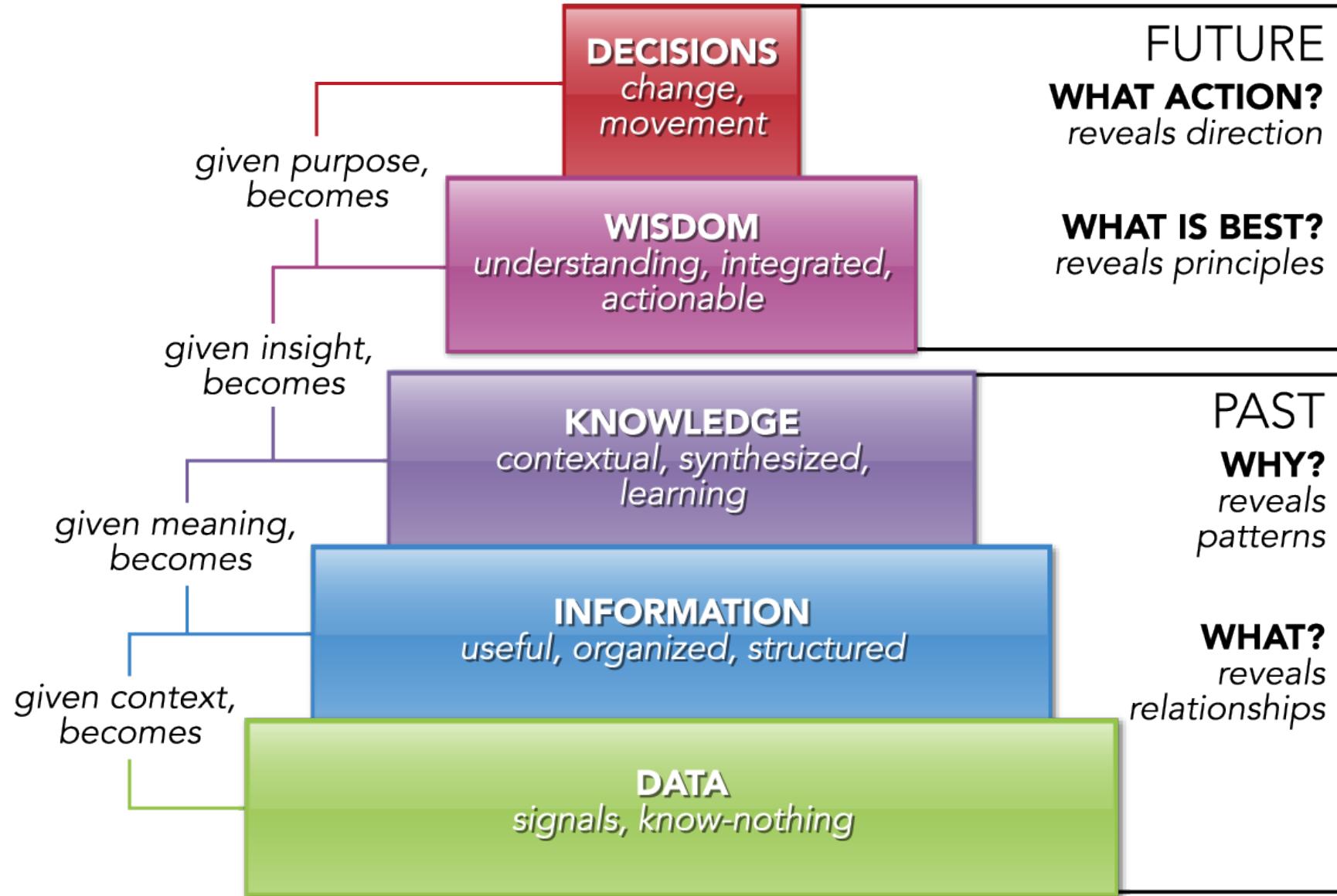
Gartner Big Data Model ? 巨量資料的模型 ?

巨量資料的挑戰在於如何管理「數量」、「增加率」與「多樣性」



Data, Information, Knowledge, Wisdom

知識管理模型：資料、資訊、知識與智慧



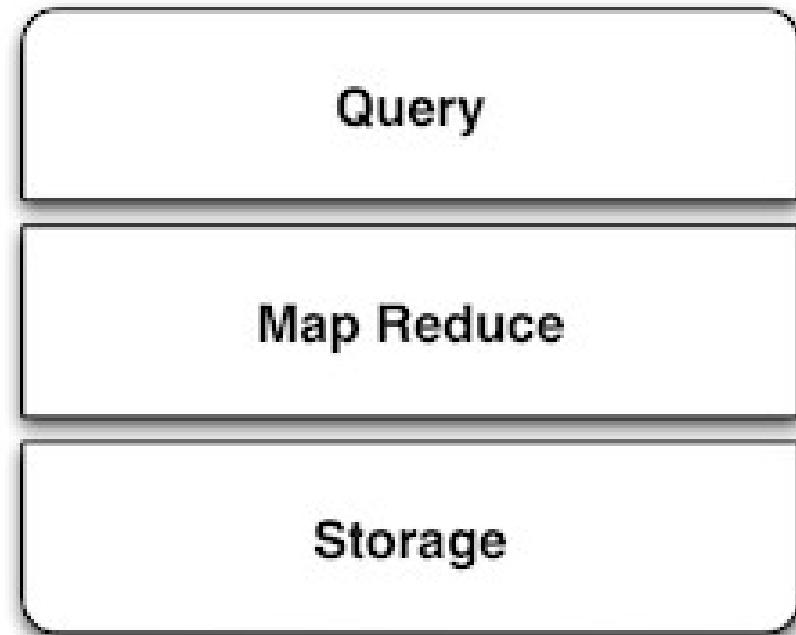
The SMAQ stack for big data

巨量資料處理的資訊架構

做網頁相關的人可能聽過 LAMP



未來處理海量資料的人必需知道
SMAQ (Storage, MapReduce and Query)



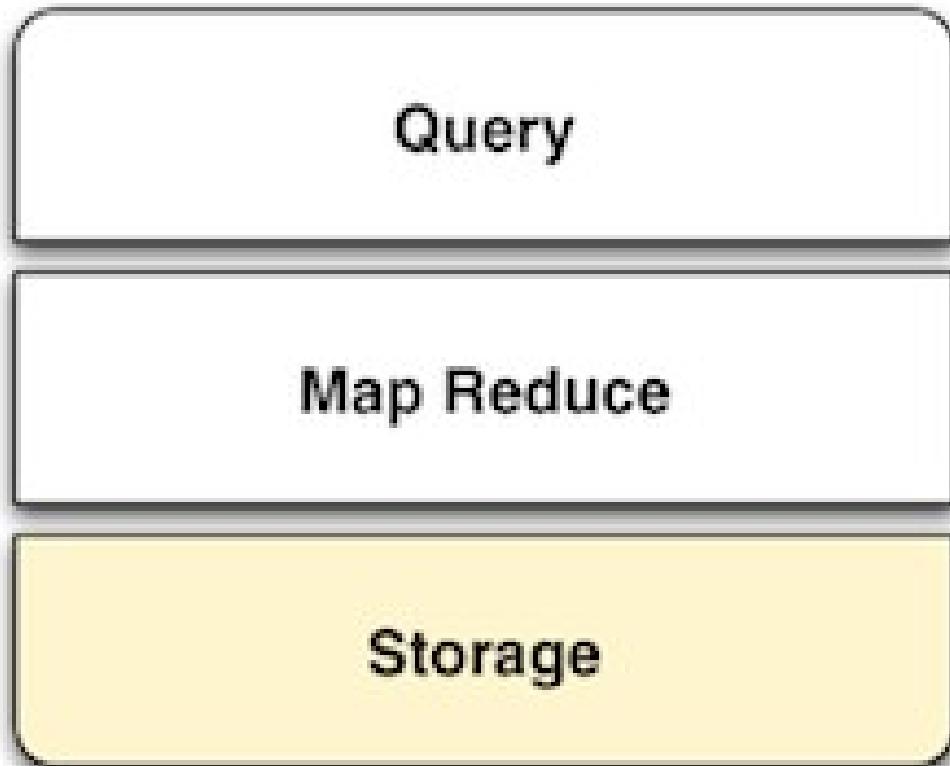
參考來源：The SMAQ stack for big data , Edd Dumbill , 22 September 2010 ,

<http://radar.oreilly.com/2010/09/the-smaq-stack-for-big-data.html>

圖片來源：<http://smashingweb.ge6.org/wp-content/uploads/2011/10/apache-php-mysql-ubuntu.png> 7

The SMAQ stack for big data

巨量資料處理的資訊架構

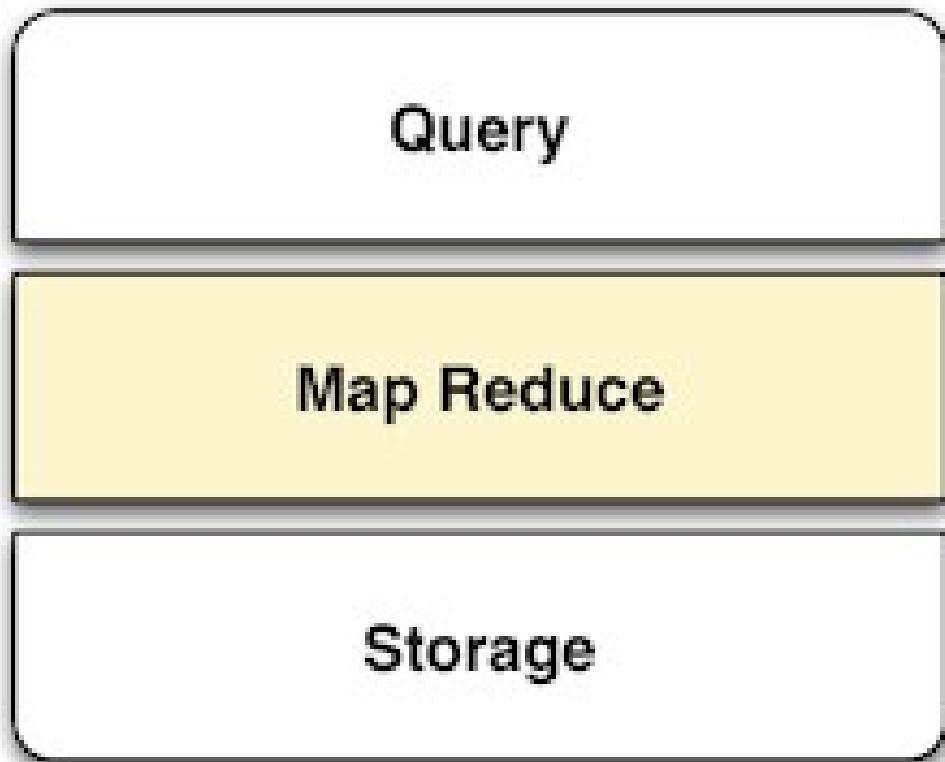


用來儲存分散、沒有關聯
的非結構化資料



The SMAQ stack for big data

巨量資料處理的資訊架構



運用批次處理的方式，將運算工作平均分散到許多的伺服器做運算。

Key features

- Distributes computation over many servers
- Batch processing model

The SMAQ stack for big data

巨量資料處理的資訊架構

Query

Map Reduce

Storage

Key features

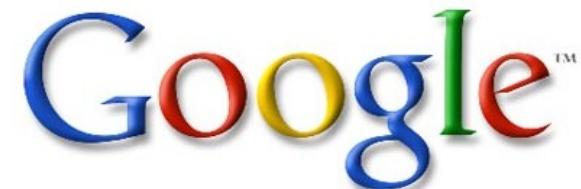
- Efficient way of defining computation
- Platform for user friendly analytical systems

將算完的結構化資料儲存
到可供查詢的資料庫系統

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/>

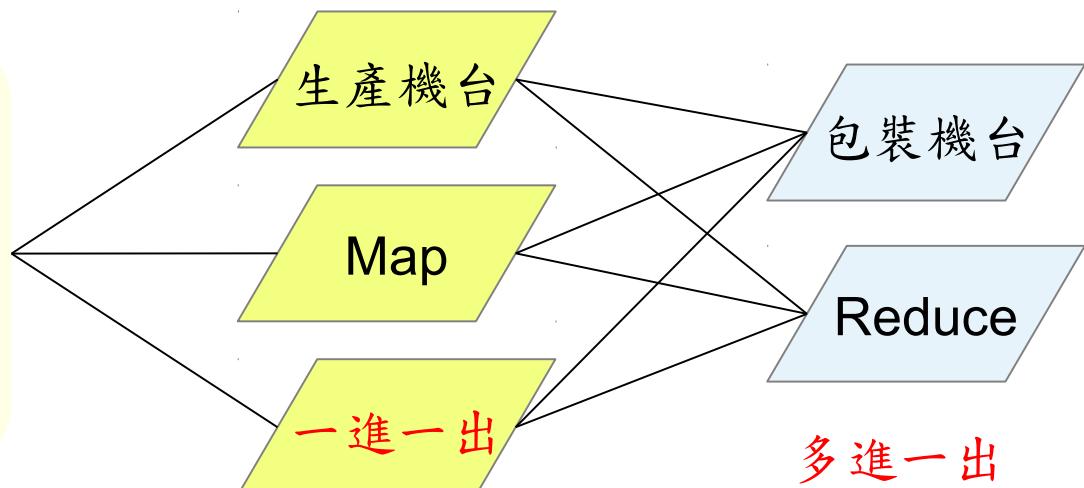
Hadoop 簡介

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** 工工作站該做哪些事情。

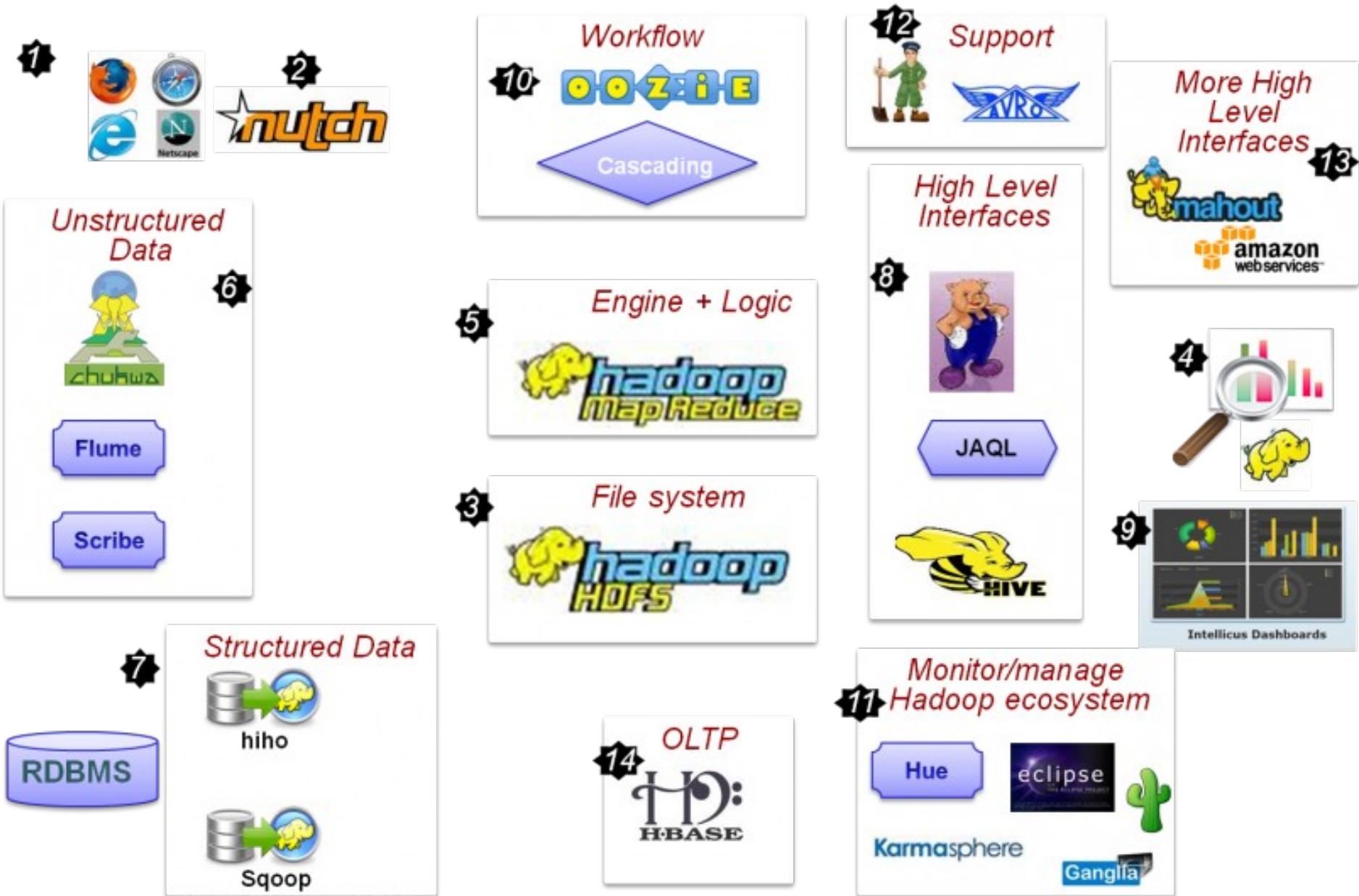
就像工廠的倉庫
存放生產原料跟待售貨物

HDFS 存放
待處理的**非結構化**資料
與處理後的**結構化**資料



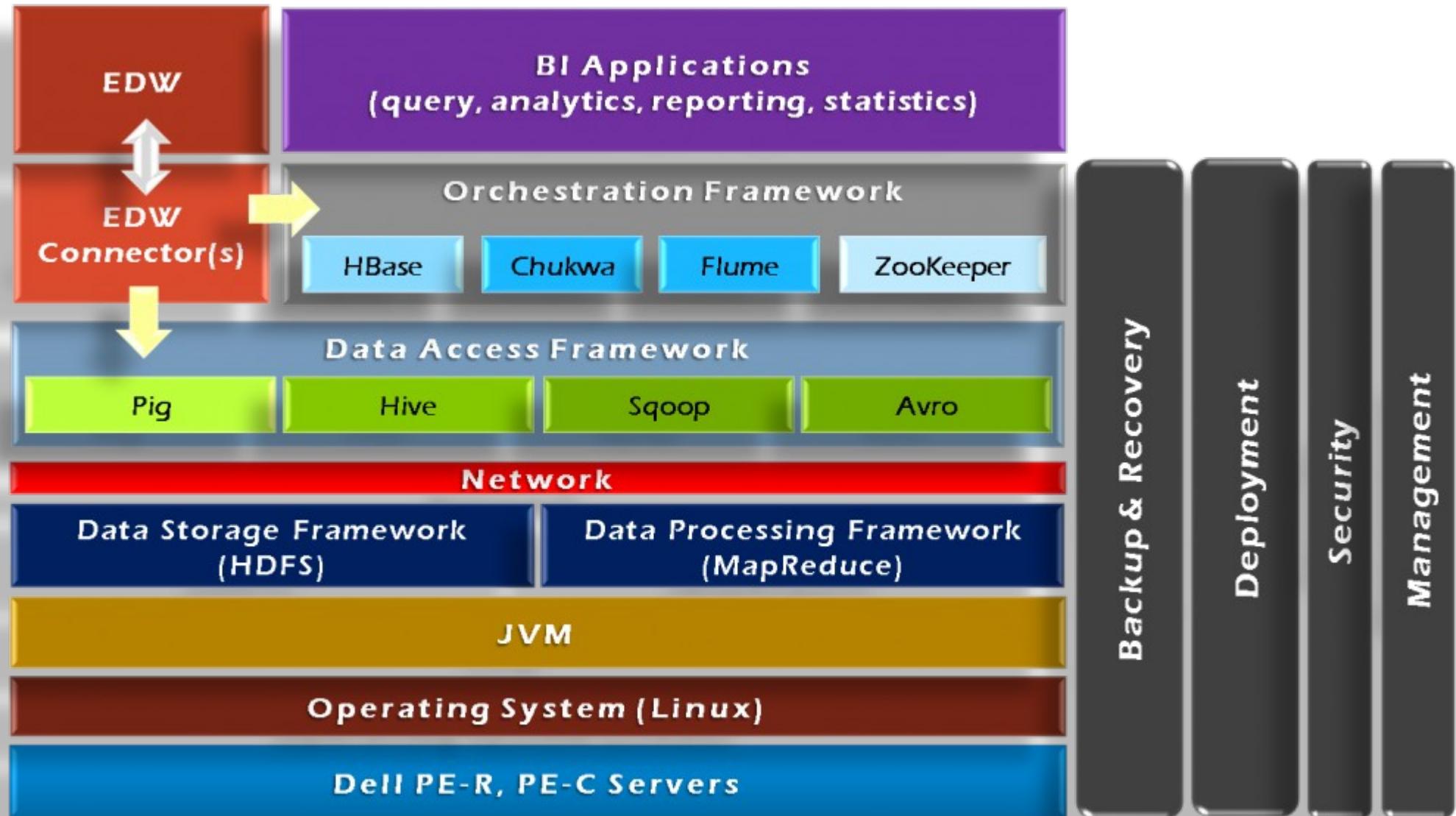
Why we choice Hadoop? Good Ecosystem!

豐富的生態系建構出處理海量資料的工具庫



BI and EDW build on Hadoop Ecosystem

運用 Hadoop 生態系搭建資料倉儲與商業智慧分析



Microsoft love Hadoop, too 微軟幫 Azure 還有 SQL Server 都接上 Hadoop

SQL Server | All Microsoft Sites United States | Change Search Microsoft bing Web

 Microsoft® SQL Server®

About SQL Server Solutions & Technologies Editions Get SQL Server Learning Center Partners

Business Intelligence

 Share this page

Big Data Analytics

Strata Big Data Conference 2012 and Power View Contest

Strata Big Data Conference 2012...



0:00 / 2:35

YouTube

Big Data Solution

Unlock business insights from all your structured and unstructured data, including large volumes of data not previously activated, with Microsoft's Big Data solution. Microsoft's end-to-end roadmap for Big Data embraces Apache Hadoop™ by distributing enterprise class Hadoop based solutions on both Windows Server and Windows Azure. Our solution is also integrated into the Microsoft BI tools such as SQL Server Analysis Services, Reporting Services and even PowerPivot and Excel. This enables you to do BI on all your data, including those in Hadoop.

Key Benefits

- Broader access of Hadoop to end users, IT professionals and Developers, through easy installation and configuration and simplified programming with JavaScript.
- Enterprise ready Hadoop distribution with greater security, performance, ease of management and options for Hybrid IT usage.

參考來源：Big Data Solution | Microsoft SQL Server 2008 R2

<http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/big-data-solution.aspx>

Oracle love Hadoop, too Oracle 也接上 Hadoop

The screenshot shows the CNET News homepage. The top navigation bar includes links for Home, Reviews, News, Download, CNET TV, How To, and Mail. A banner for HP Officejet Pro printers is visible, with the text "MAKE YOUR BUSINESS SHINE FOR LESS" and a "LEARN MORE" button. The main content area shows a news article titled "Cloudera teams up to connect Oracle and Hadoop".

CNET > News > Software, Interrupted

Cloudera teams up to connect Oracle and Hadoop

Cloudera and Quest software are partnering to provide connectivity between Oracle and Hadoop.



by [Dave Rosenberg](#) | June 21, 2010 5:30 AM PDT

[Follow](#)

This week [Cloudera](#), a provider of software and services for the Apache Hadoop project, is set to announce a partnership with [Quest Software](#) to develop, support, and distribute an Oracle connector for Hadoop.



參考來源 : Cloudera teams up to connect Oracle and Hadoop

http://news.cnet.com/8301-13846_3-20008242-62.html

Hinet Application of Big Data

中華電信已經在做的海量資料應用



中華電信：分析駭客行為，拓展對外新服務

撰文者：趙郁竹

發表日期：2012-03-06



[214期雜誌精選]

全台最大的中華電信提供行動電話、市話、寬頻固網、MOD……，各種業務服務，加起來的用戶數就有3000萬，比全台灣人口還多，光是單月帳務數量就高達100億筆資料。除了電信、寬頻服務，還有日益增加的數位服務、行動增值服務，從服務內容到客戶端，累積出的資料相當驚人。

「資料量越來越大，日常分析工作需要很多時間，但新的運算技術有效解決了這個問題，」中華電信資訊處處長陳明仕說。2010年開始，因為中華電信本身的資料運算需求，採用分散式運算架構Hadoop技術，打造出大資料運算平台，不但解決了自身的資料問題，還能對外提供資料運算應用。

以MOD為例，一天有幾千萬筆資料，如何找出使用者在什麼時段做了什麼事？廣告效益又如何？「用傳統的方法，需要400分鐘才能分析完；用Hadoop大資料平台，13分鐘就能解決，節省非常多時間，」他說。

追蹤再拆解

大資料運算技術除了節省時間，還能防止駭客入侵。「駭客的攻擊行為都有模式可循，」陳明仕解釋，就像球賽一樣，了解進攻模式就能防守。用戶的資料保護是第一要務，因此透過行為模式分析，能有效保護企業資訊安全，也保障客戶的個資安全。

參考來源：中華電信：分析駭客行為，拓展對外新服務，發表日期：2012-03-06

<http://www.bnnext.com.tw/print/article/id/22333>

Hinet Application of Big Data

中華電信已經在做的海量資料應用

IT ithome.com.tw

中華電信用Hadoop技術分析通話明細

 READ LATER

面對資料快速成長以及非結構性資料的增加，中華電信資訊處第四科科長楊秀一表示，中華電信近來利用Hadoop雲端運算技術自行開發了一個專門用來分析非結構化資料的巨量資料（Big Data）運算平臺，嘗試在資料進到資料倉儲系統之前，先進行資料的分析與處理以減少資料倉儲的資料量。

近年來行動語音市場趨於飽和，為了掌握用戶特性進行客製化行銷，一份資料要進行分析，就會被多次複製，因此即使用戶增加趨緩，但中華電信擁有的資料量仍快速暴增。

中華電信用來分析的資料模型最早於10多年前已有雛形，但當初主要用於行動語音分析。一直到2009年，他們完整導入Teradata的電信業邏輯資料模型cLDM 9.0版，整合更多電信服務的用戶資料。楊秀一表示，當初導入該模型的目的主要是為了整合行動語音、固網、數據的資料，進行以人為中心的分析模式。在導入之前，中華電信的資料模型是以設備為中心，因為不同設備的記錄資料儲存在不同的資料庫，無法進行整合性的分析。

參考來源：中華電信用 Hadoop 技術分析通話明細，發表日期：2011-06-12
<http://www.ithome.com.tw/itadm/article.php?c=68023>

雲端產業供應鏈

Supply Chain of Cloud Computing

應用軟體
供應商



端

行動裝置
共通平台

HTC
quietly brilliant

apple

H

各類裝置
存取服務

軟體服務
供應商

Google

amazon
web services™



雲

資料中心
機房維運

IBM

acer

資料中心
提供服務

基本硬體
建設組件

H

FOXCONN®
鴻海科技集團

英業達集團
Inventec

廣達電腦
Quanta Computer

雲端產業供應鏈

Supply Chain of Cloud Computing

應用軟體
供應商



端

行動裝置
共通平台

htc
quietly brilliant



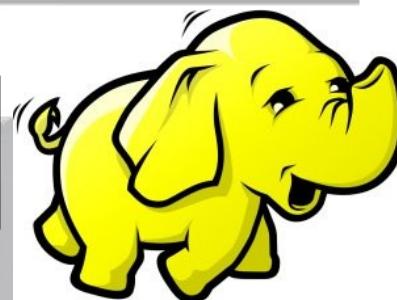
各類裝置
存取服務

軟體服務
供應商

Google

amazon
webservices™

大象擺這裡



資料中心
機房維運



acer®

資料中心
提供服務

基本硬體
建設組件



FOXCONN®
鴻海科技集團

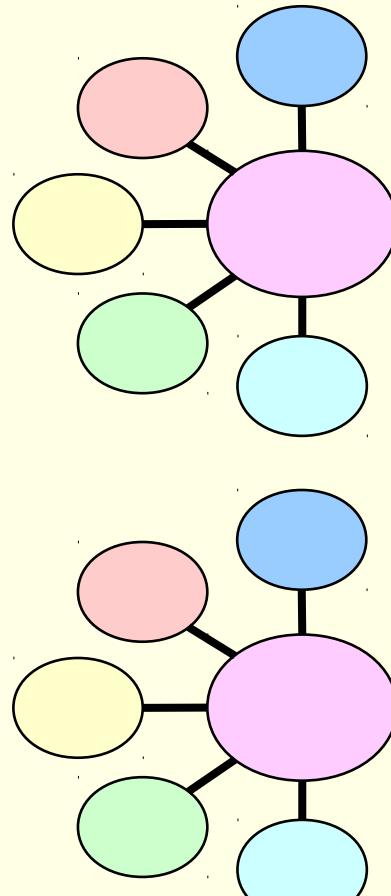
英業達集團
Inventec

廣達電腦
Quanta Computer

巨量資料的奇幻漂流 Life of Big Data

Internet of Things 物聯網

Sensor Network



Smart Grid

雲 資料中心 提供服務



Public Data Hub
Data as a Service

開放資料
Open Data



Web 2.0

Query

Map Reduce

Storage

Big Data

Cloud Computing

雲端運算

端

各類裝置
存取服務



Mobile Computing

Internet of Things 物聯網

