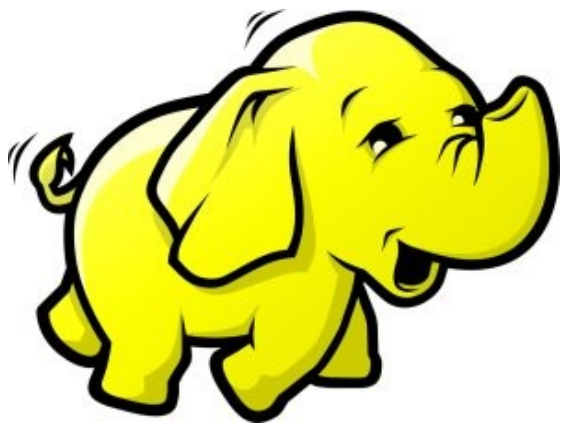
處理海量資料的資訊架構與關鍵技術

Technologies to build IT Stack for Big Data

Jazz Wang

Yao-Tsung Wang

jazz@nchc.org.tw



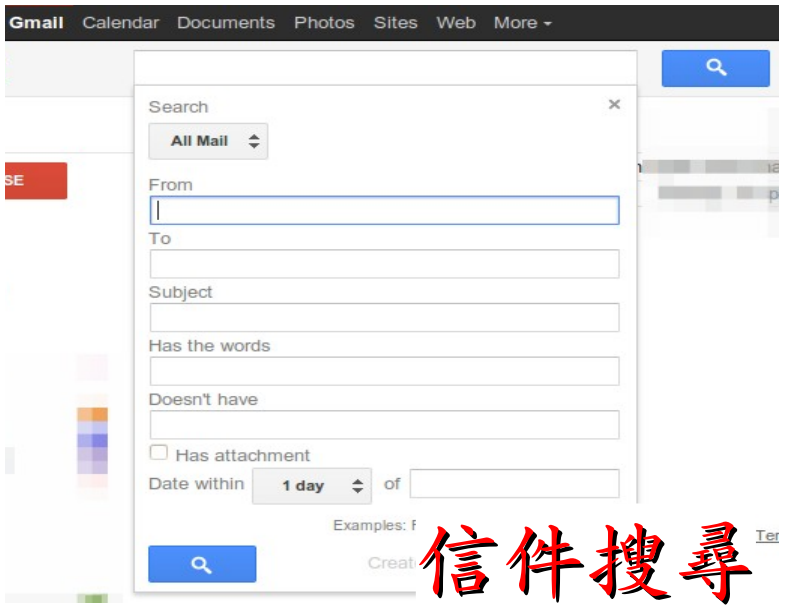
Hot Jobs in Big Data
從海量資料的熱門工作談起

Data Mining	資料探勘
Data Visualization	資料視覺化
Data Analysis	資料分析
Data Manipulation	資料操控
Data Discovery	資料鑑識

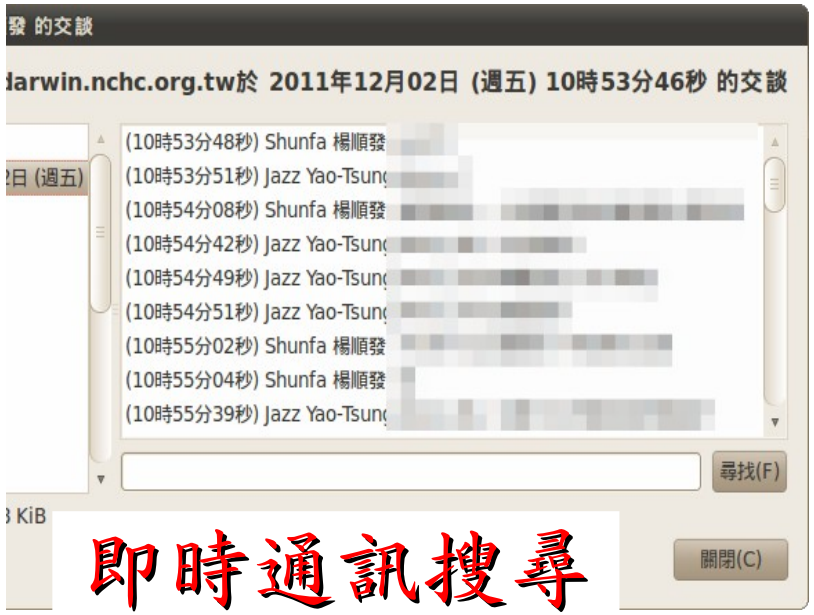
How to Get a Hot Job in Big Data, Dan Tynan, InfoWorld, March 19, 2012
出處：http://www.cio.com/article/print/702388

Applications of Data Mining
資料探勘的應用 - 搜尋引擎

檔案搜尋



信件搜尋

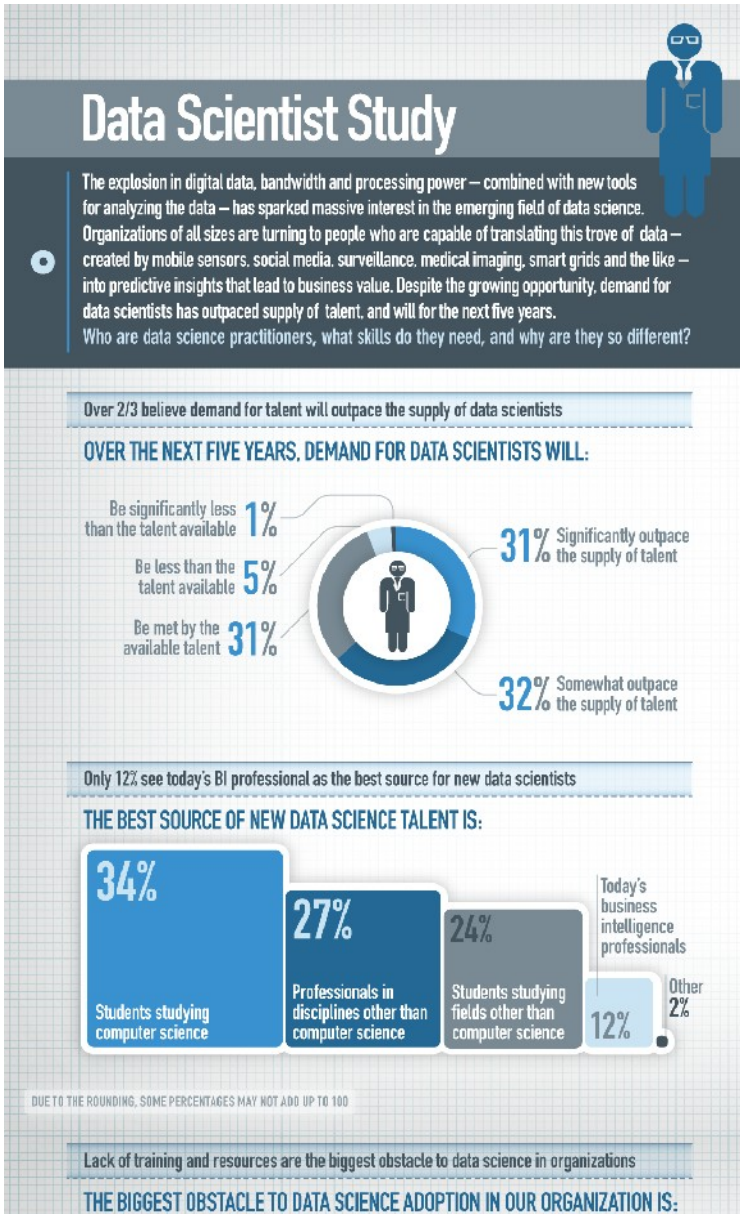


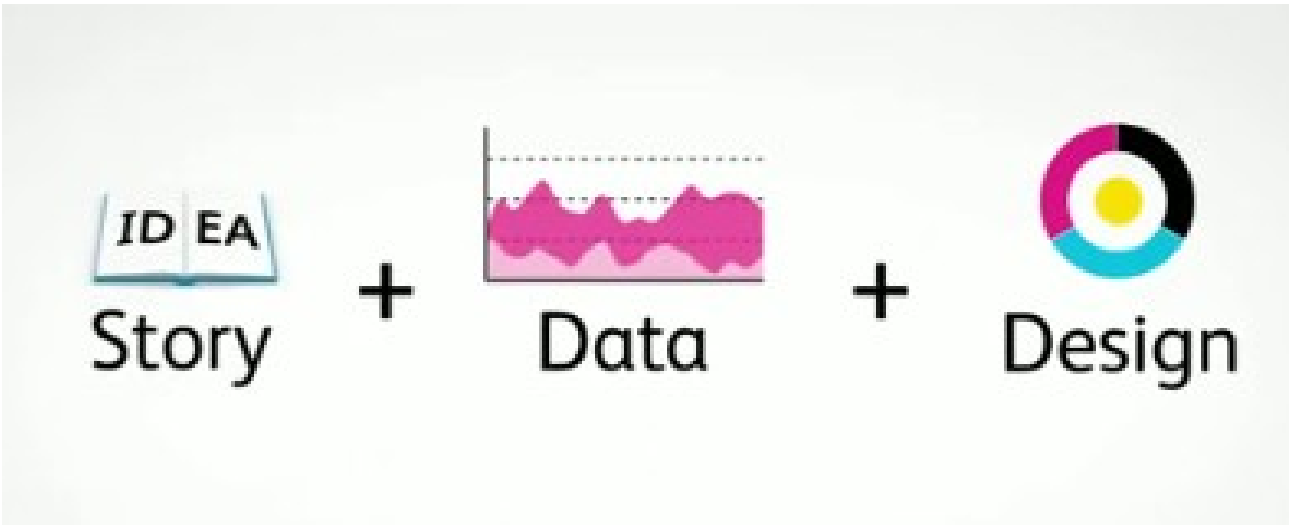
即時通訊搜尋

資料庫搜尋

「網頁搜尋」

Applications of Data Visualization
資料視覺化的應用 -Infographics





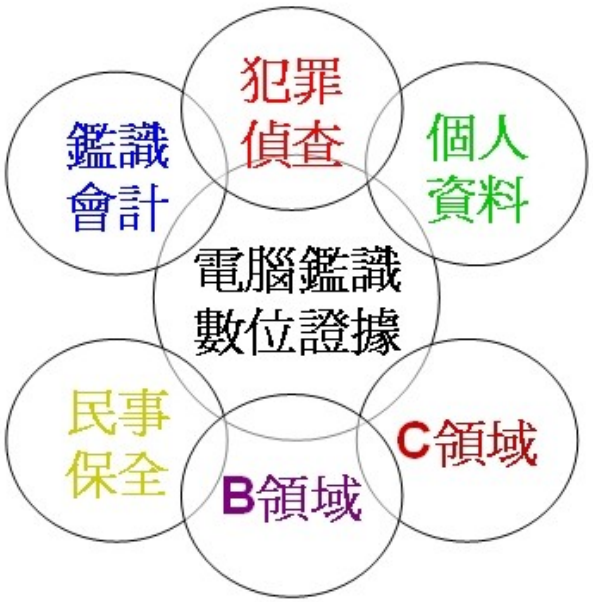
參考來源：未來「夯」職業：資料科學家 http://www.bnext.com.tw/print/article/id/21740
淺談超吸睛的資訊圖表 http://www.inside.com.tw/2011/04/13/infographics

Applications of Data Analysis
資料分析的應用 - 商業智慧 (BI)

圖片來源： http://bizcubic.com/wp-content/uploads/2010/01/business-intelligence-dashboard.jpg

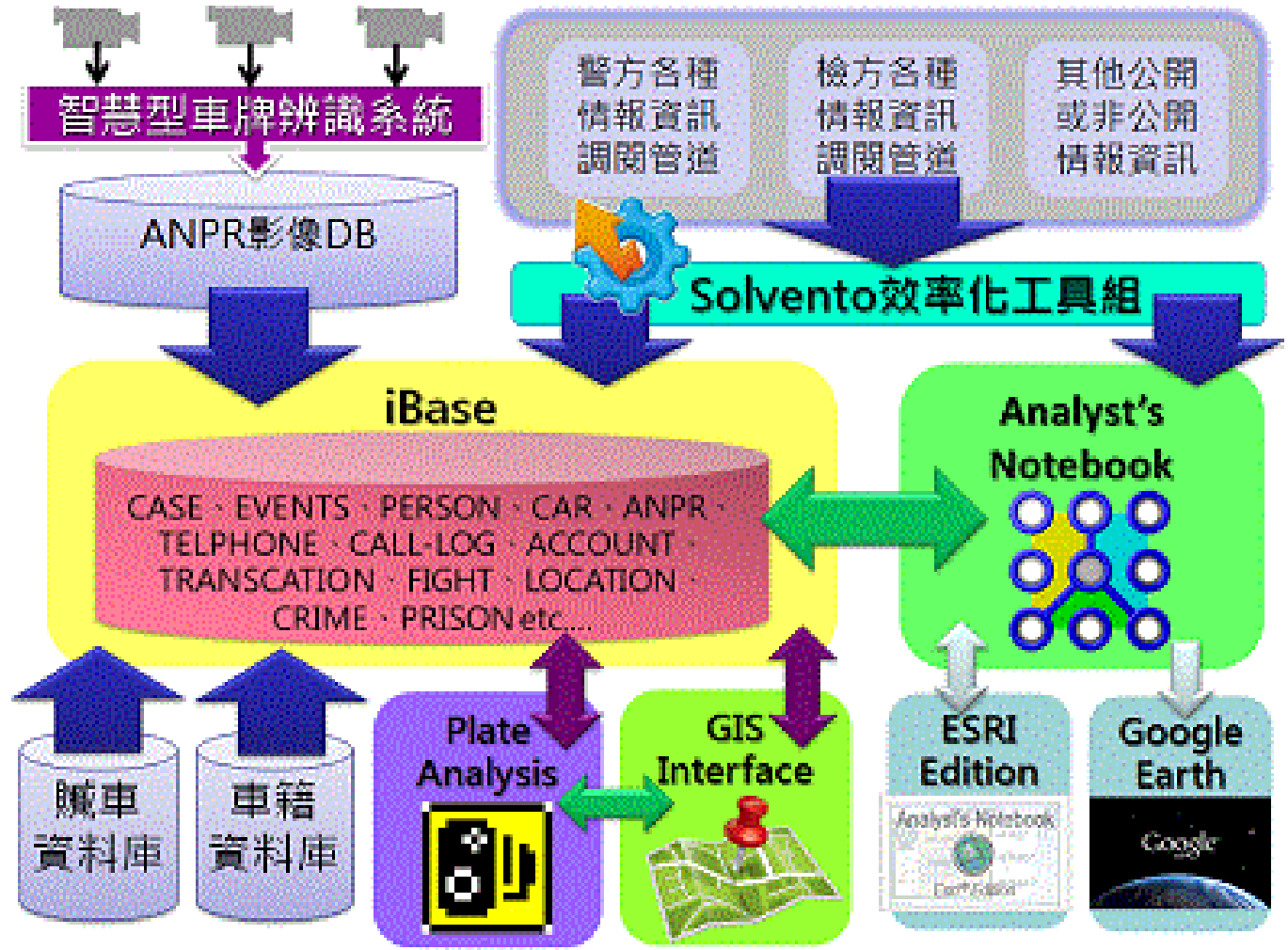
Applications of Data Discovery
數位鑑識 - 資訊與法律的結合

電腦鑑識&會計鑑識

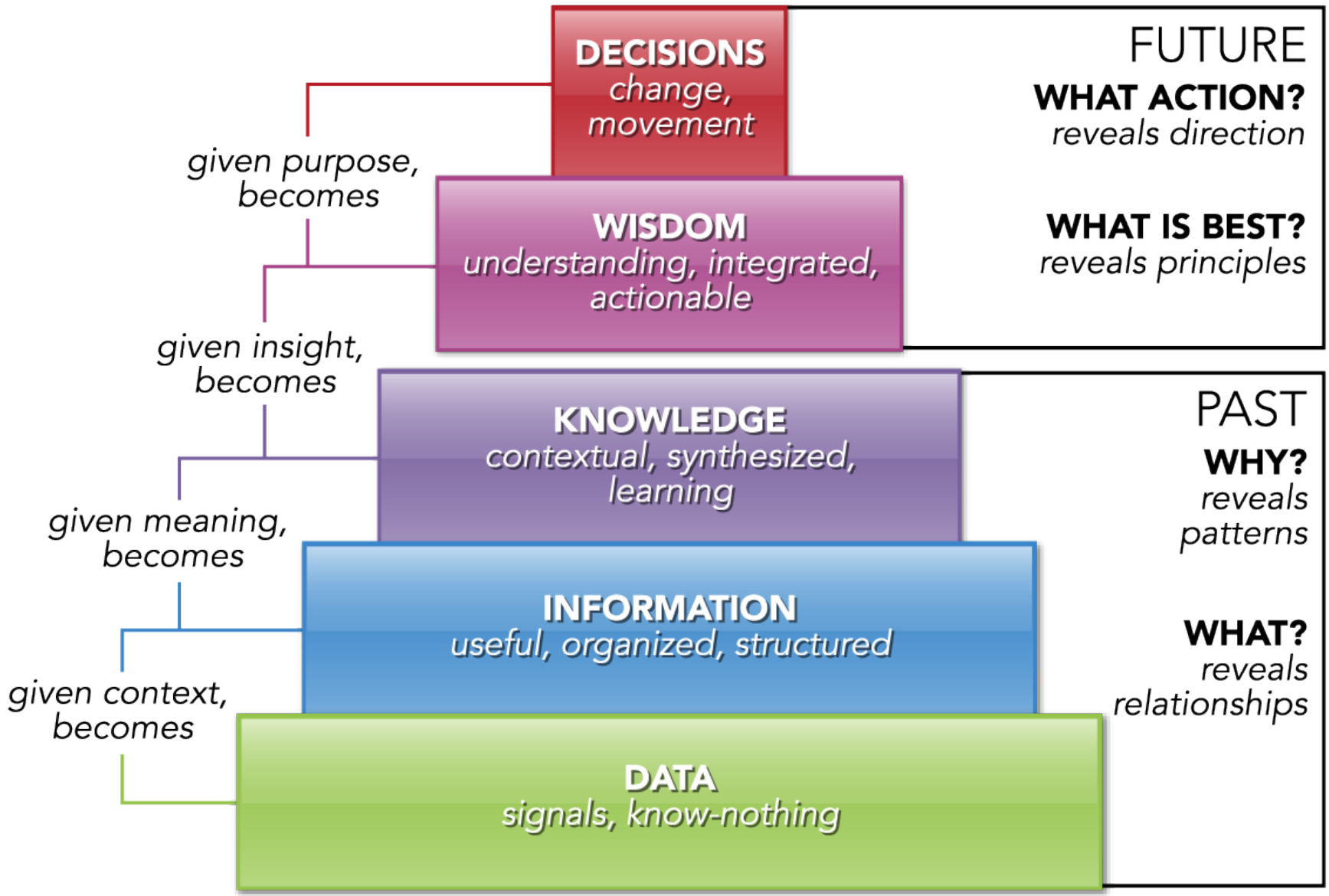


http://blog.udn.com/kf0630/6018593

http://www.solventosoft.com/upload/ANPR_02s.gif



Data, Information, Knowledge, Wisdom
知識管理模型：資料、資訊、知識與智慧

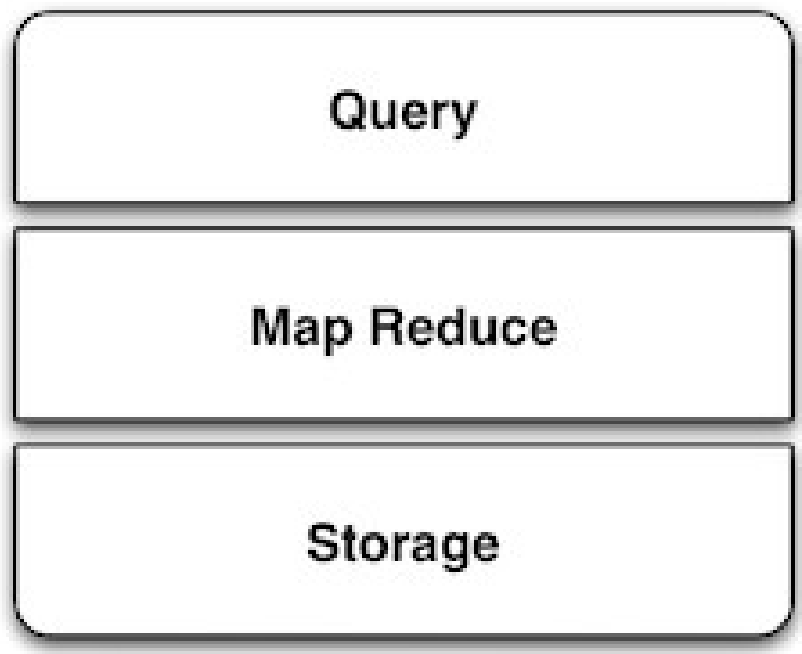


http://www.pursuantgroup.com/blog/tag/dikw-model/

The SMAQ stack for big data
海量資料處理的資訊架構

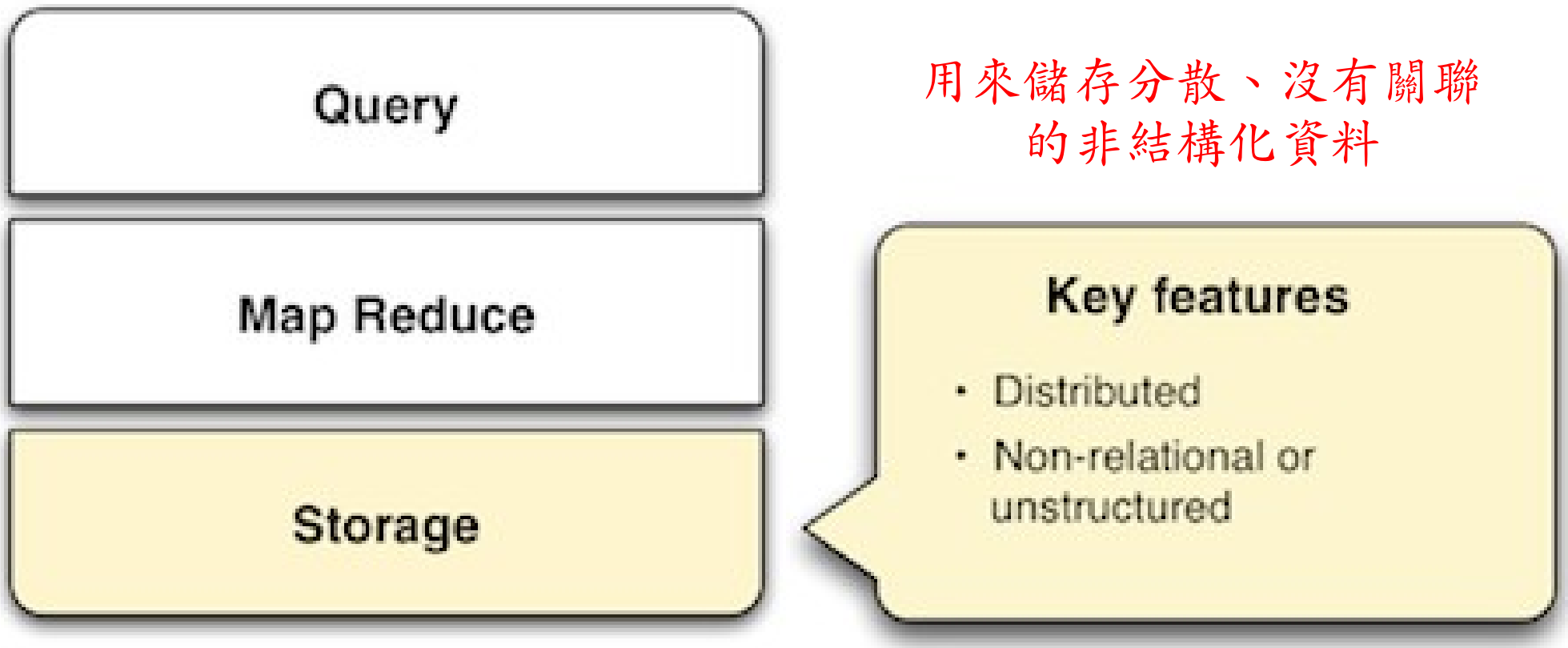
做網頁相關的人可能聽過 LAMP

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

The SMAQ stack for big data
海量資料處理的資訊架構

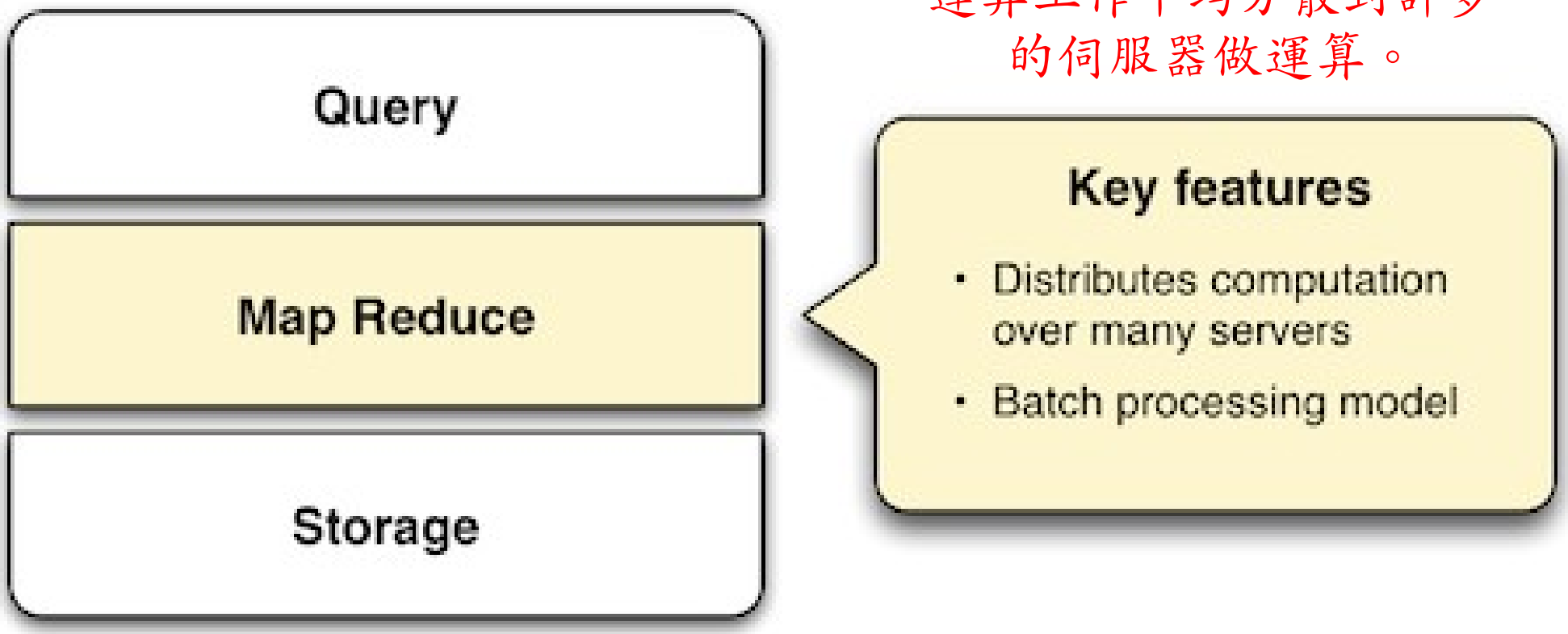


用來儲存分散、沒有關聯的非結構化資料

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

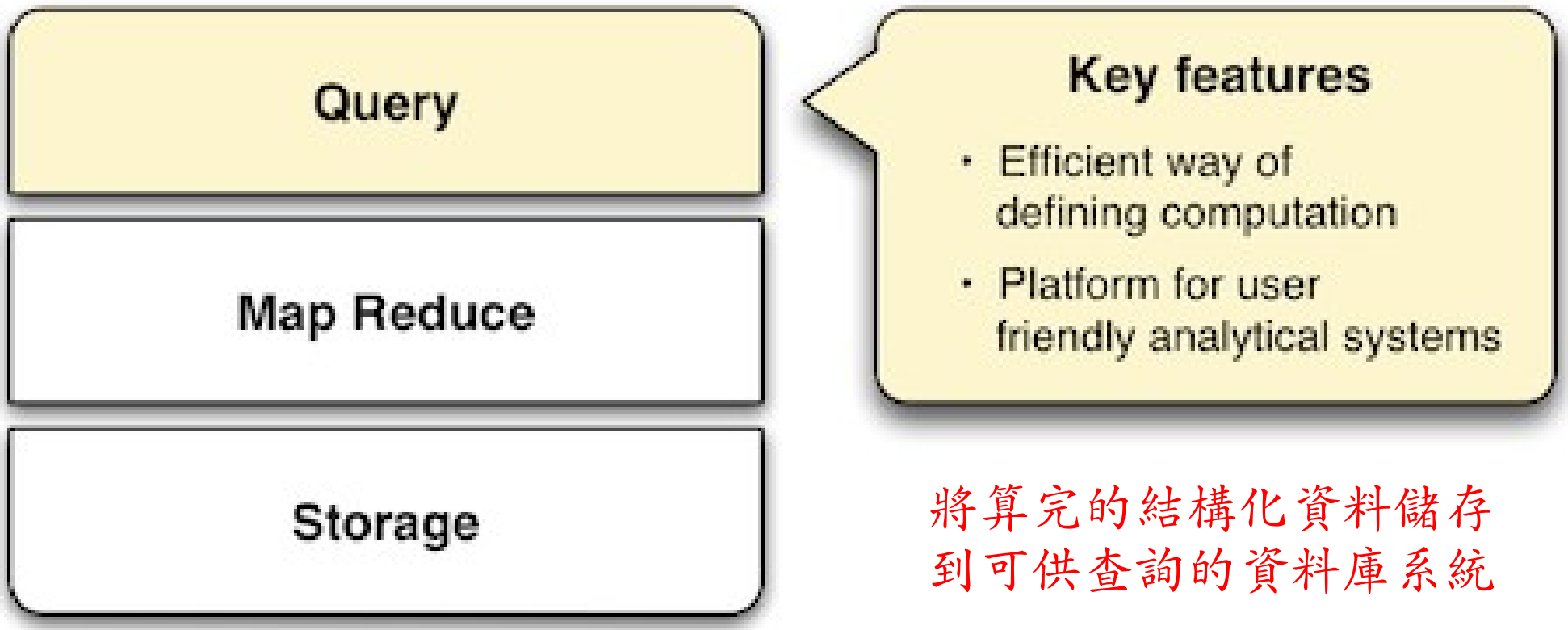
The SMAQ stack for big data
海量資料處理的資訊架構

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。



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

The SMAQ stack for big data
海量資料處理的資訊架構



將算完的結構化資料儲存到可供查詢的資料庫系統

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

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 : Simplifed 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

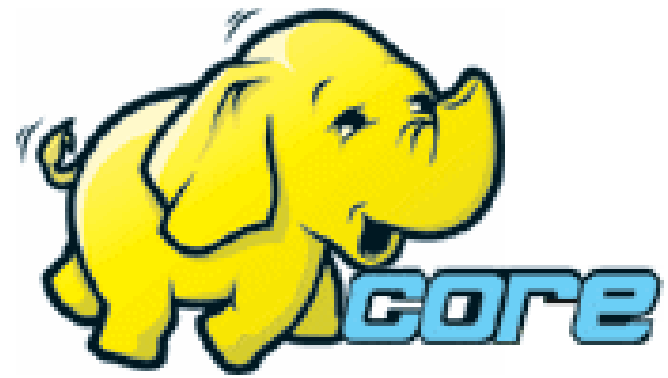
更多不同語言的 MapReduce API 實作：
http://trac.nchc.org.tw/grid/intertrac/wiki%3Ajazz/09-04-14%23MapReduce

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** 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**

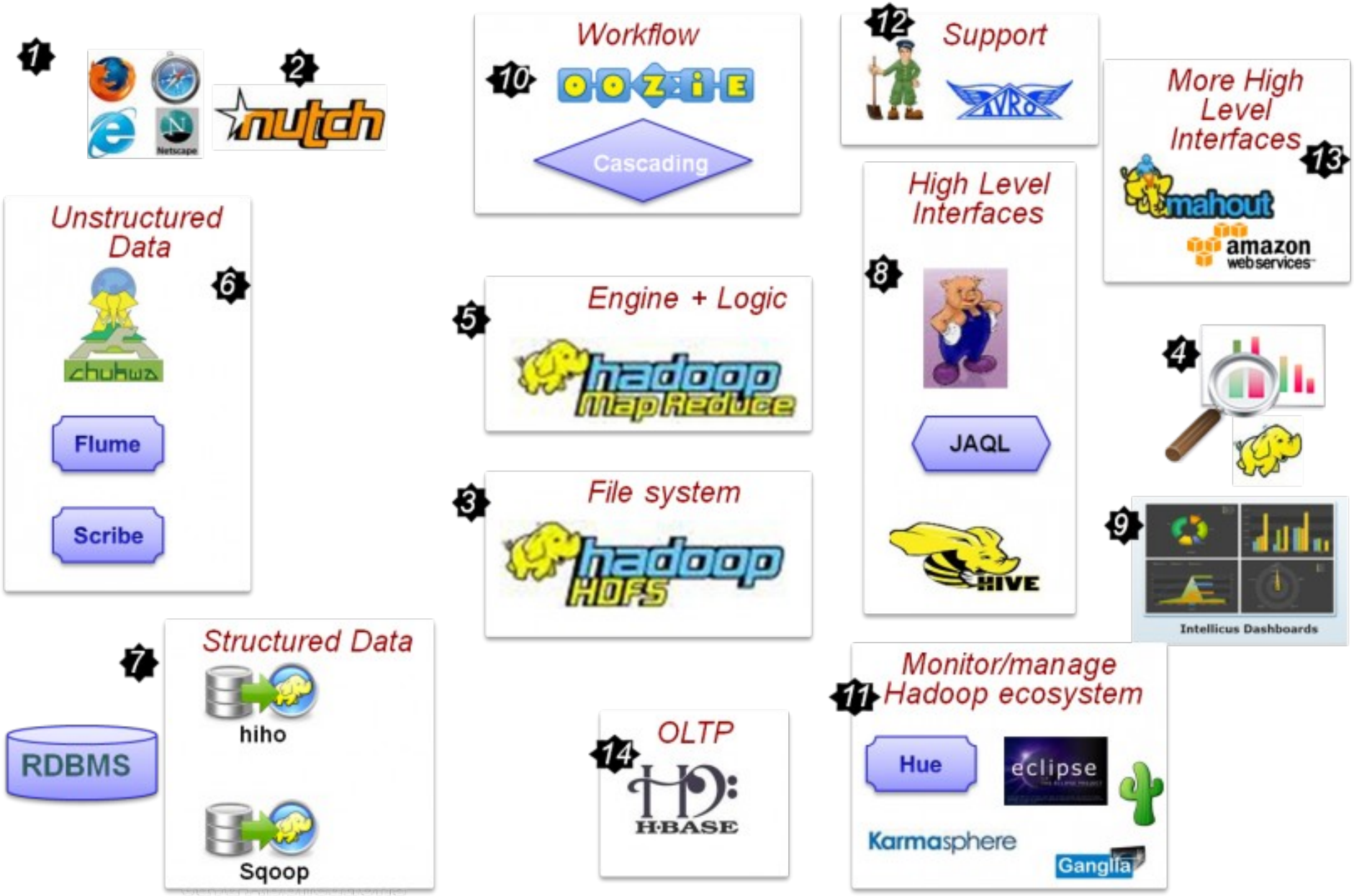
Sector-Sphere

National Center for Data Mining
University of Illinois at Chicago
UIC

open data
Open Data Group
http://www.opendatagroup.com/

Why we choice Hadoop? Good Ecosystem!

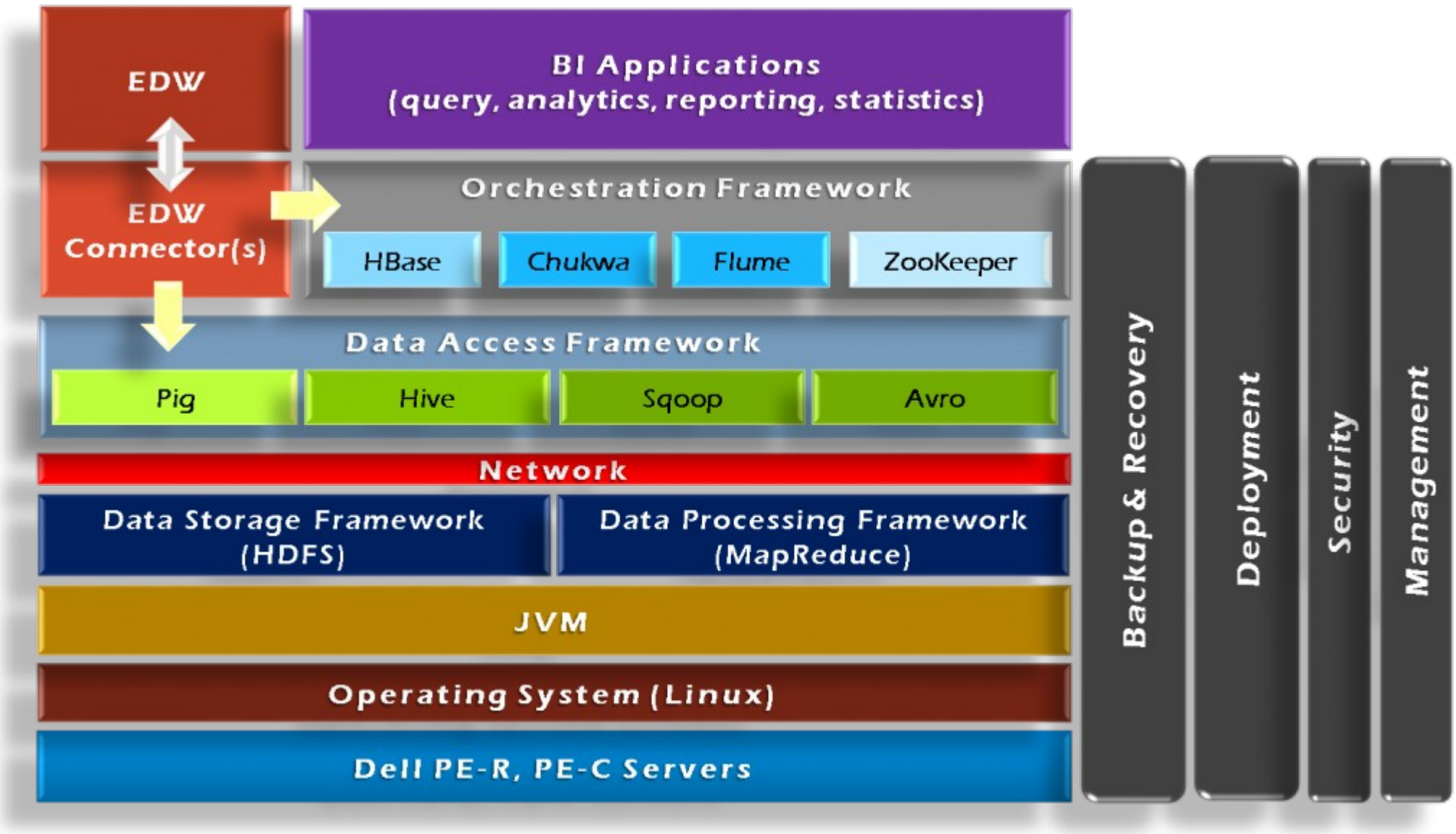
豐富的生態系建構出處理海量資料的工具庫



http://rationalintelligence.com/wp_log/?p=104

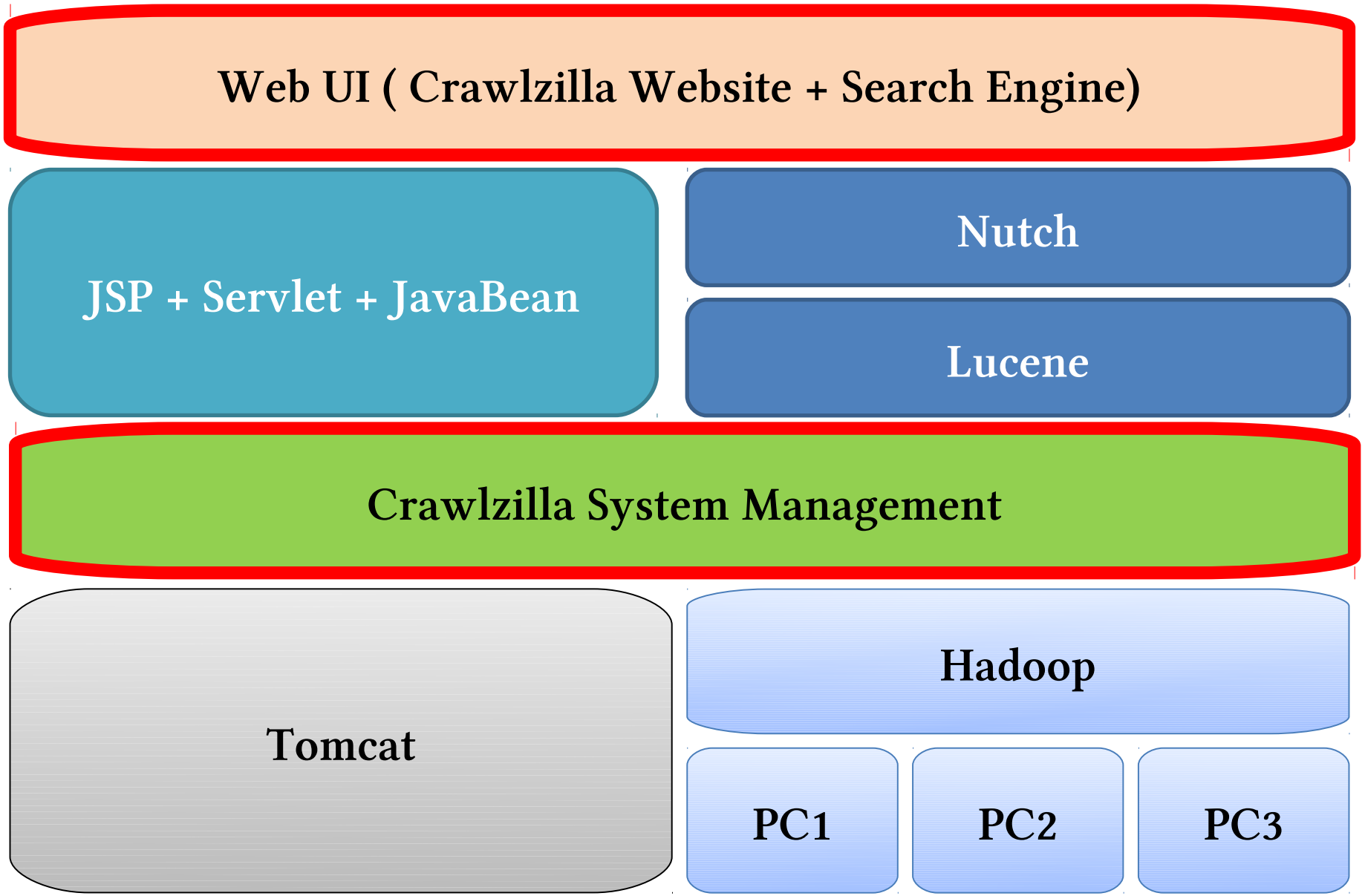
BI and EDW build on Hadoop Ecosystem

運用 Hadoop 生態系搭建資料倉儲與商業智慧分析

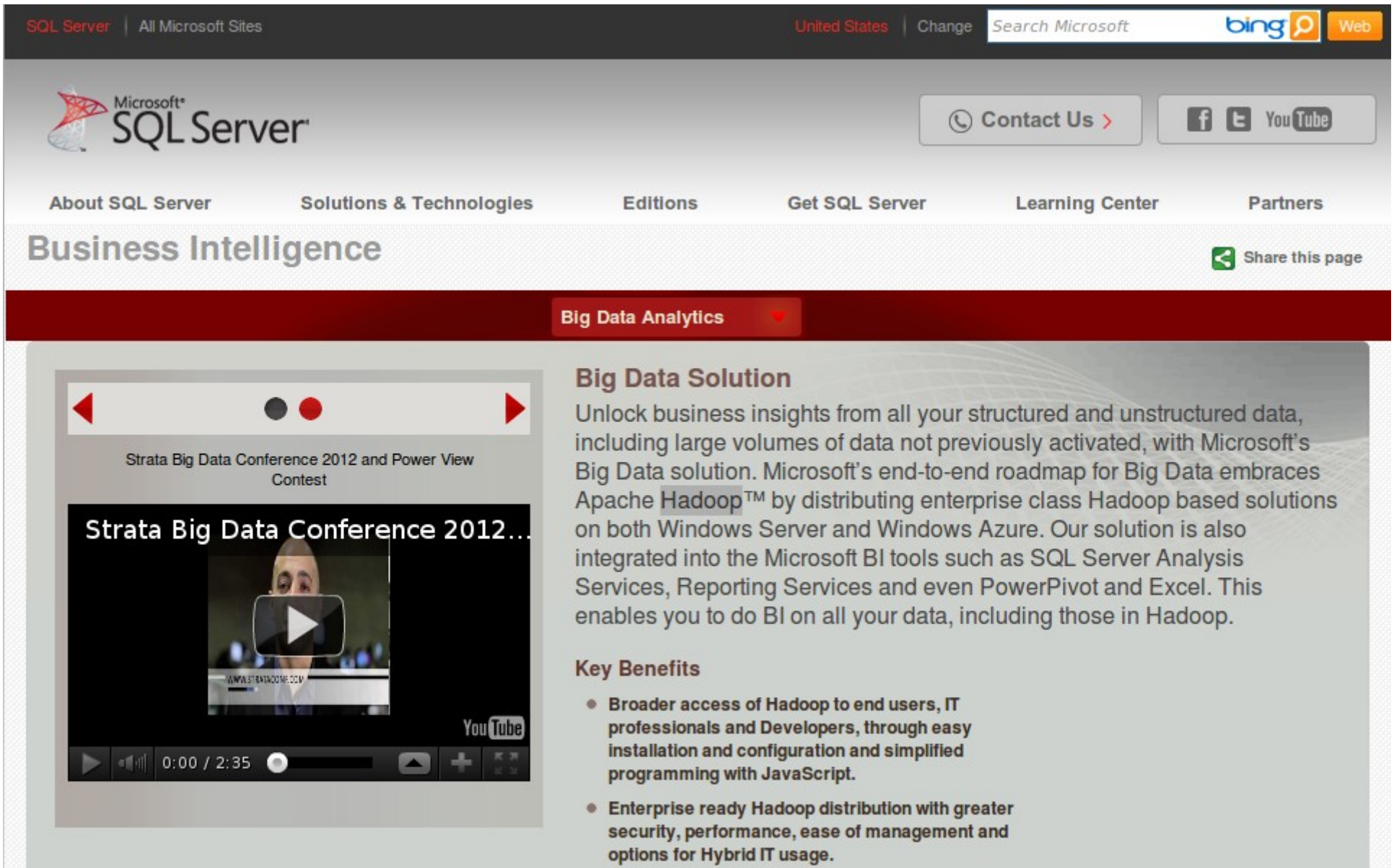


http://rationalintelligence.com/wp_log/?p=104

Build your own search engine, too
您也能用 Hadoop 搭建自己的搜尋引擎



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



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

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

Business Next 數位時代

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、行動加值服務，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：分析駭客行爲，拓展對外新服務，發表日期： 2012-03-06
http://www.bnext.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