Hadoop＆Spark系列 –
深入淺出Hadoop＋Spark安裝與開發

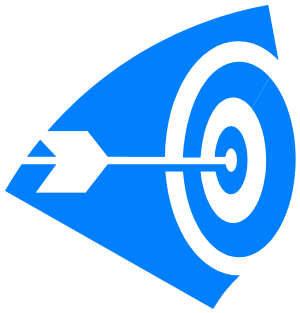
Yung-Chuan Lee

About Me

- 姓名：李泳泉
- 任職：南部某家幸福企業裡，微不幸的攻城獅
- 歷程：
  - 2015年初 - **21世紀最性感的職業：Data Scientist**
    - 被這句話吸引，走向不歸路
  - 2015年中 – **開始學習Hadoop**
  - 2015年底 – **開始學習Spark**
  - 2016年中 – **EHC Hadoop 佈署大賽2016季軍**
  - 2016年底 – **資料年會及群環科技講師**
- EMail：whitelee@gmail.com

課程目標

- 開始Hadoop及Spark學習的大門
  - 了解基本名詞的內容，為深入的學習鋪路
- 架設可以練習及開發的Lab環境
  - 練習方熟能生巧，但總得有個環境吧 ~
    - Hadoop、Spark、Hive ……
- 學習Hadoop&Spark平台程式開發
  - MapReduce
  - RDD操作
  - DataFrame操作
  - Spark MLlib應用



範例程式下載專區

- 本課程的github page: https://github.com/yclee0418/hadoopTeach

Outline

Module 1：大數據簡介

Module 2：Hadoop Ecosystem介紹

Module 3：Hadoop 平台安裝

Module 4：Hadoop 分散式檔案系統（HDFS）

Module 5：Hadoop MapReduce

Module 6：Apache Hive

Module 7：Sqoop與Flume

Module 8：Apache Spark

Module 9：Spark 平台安裝

Module 10：RDD － Resilient distributed dataset

Module 11：Scala 程式開發基礎

Module 12：Spark SQL 及 DataFrame

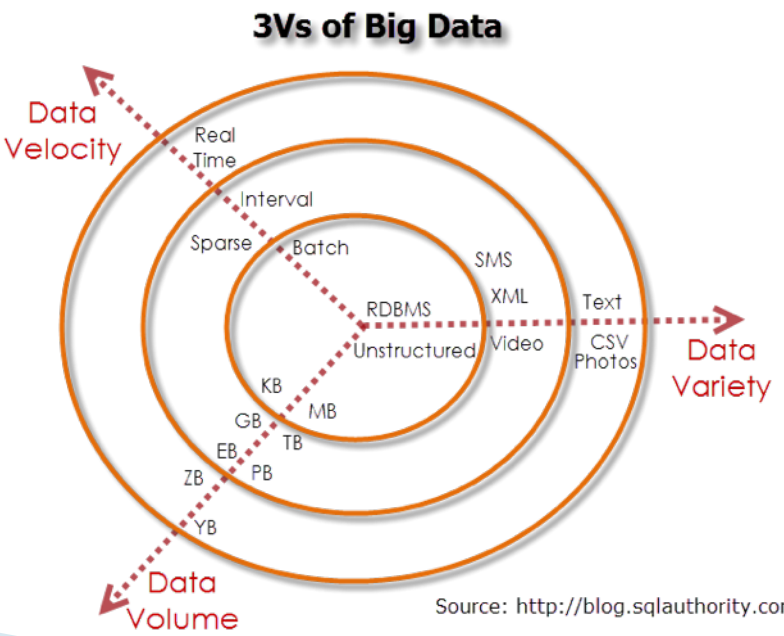
Module 13：Spark 機器學習函式庫(MLlib)

What is Big Data

- Big Data（中文：巨量資料、海量資料、大數據）
- 過去廣泛用於企業內部的資料分析、商業智慧（Business Intelligence，BI）和統計應用之大成
- 資料量急速成長、儲存設備成本下降、軟體技術進化和雲端環境成熟等客觀條件成立
- 資料分析從過去的洞悉歷史進化到預測未來，甚至是破舊立新

Big Data的3V觀點

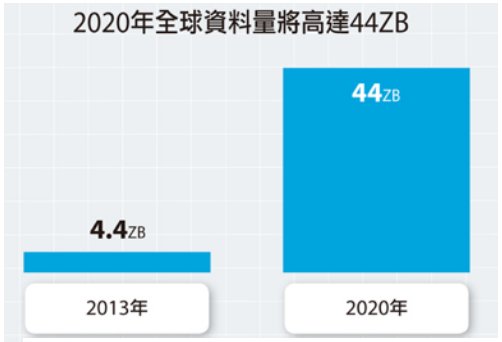
- 由Doug Laney(META Group)於2001提出
- 資料增長的挑戰和機遇的三個方向，合稱「3V」：
  - 量（Volume，資料大小）
  - 速（Velocity，資料輸入輸出的速度）
  - 多變（Variety，多樣性）

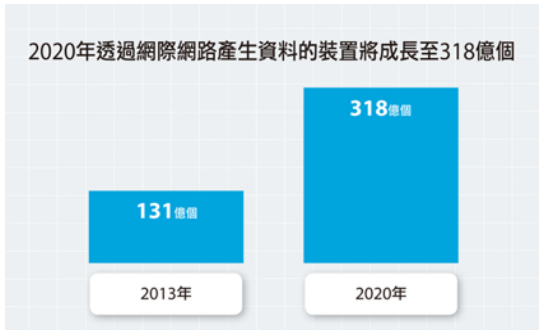


Source: http://blog.sqlauthority.com

大數據時代 – 資料帶來的挑戰

- 資料產生的速度、數量及多樣性超過單一主機所能負荷





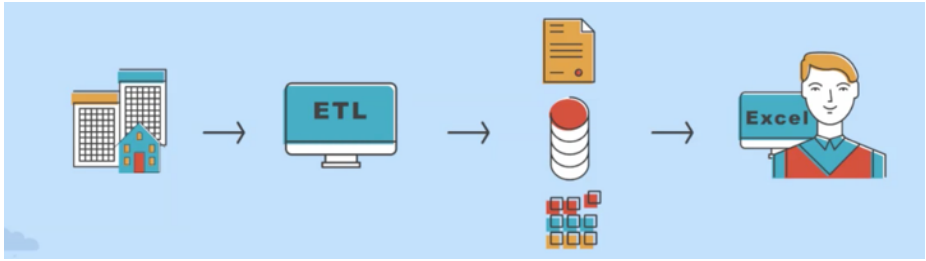
ref: http://www.ithome.com.tw/article/87190

- 需要簡單且高度平行化的資料處理架構協助去蕪存菁(Veracity)，進而產生價值(Value)

Decimal		
Value	Metric	
1000	kB	kilobyte
1000^2	MB	megabyte
1000^3	GB	gigabyte
1000^4	TB	terabyte
1000^5	PB	petabyte
1000^6	EB	exabyte
1000^7	ZB	**zettabyte**
1000^8	YB	yottabyte

傳統資料處理的挑戰

- 由單一的主機(超級電腦)集中化處理
  - 建置成本高、不易擴充
    - 運算速度追不上資料產生速度
    - 儲存空間追不上資料成長量
- 客觀因素改變使分散式架構漸成主流
  - 使用標準化的量產硬體(Commodity Hardware)
  - MOORE's LAW(18個月會將晶片的效能提高一倍)
  - 數據分析需求的普遍化
  - 越來越便宜的儲存設備



分散式資料處理的挑戰

- 分散式應用開發複雜度極高
  - 工作分派、同步控制、……
- 如何有效利用有限的頻寬
  - 網路
  - Storage
- 節點失效的處理
- 資料讀取 / 處理的瓶頸