



大數據資料處理 – Hadoop MapReduce Programming

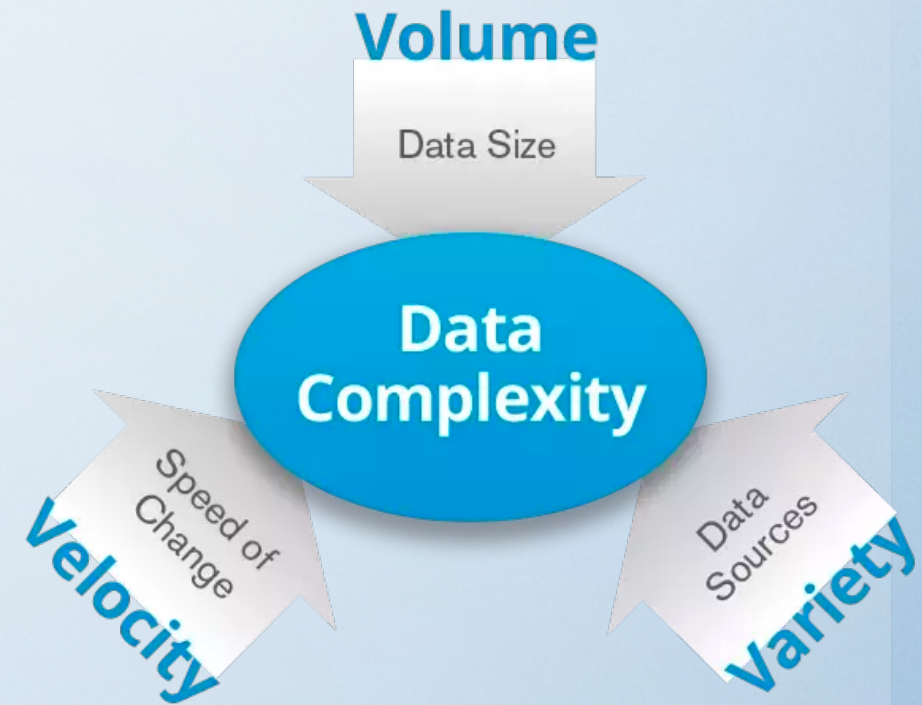
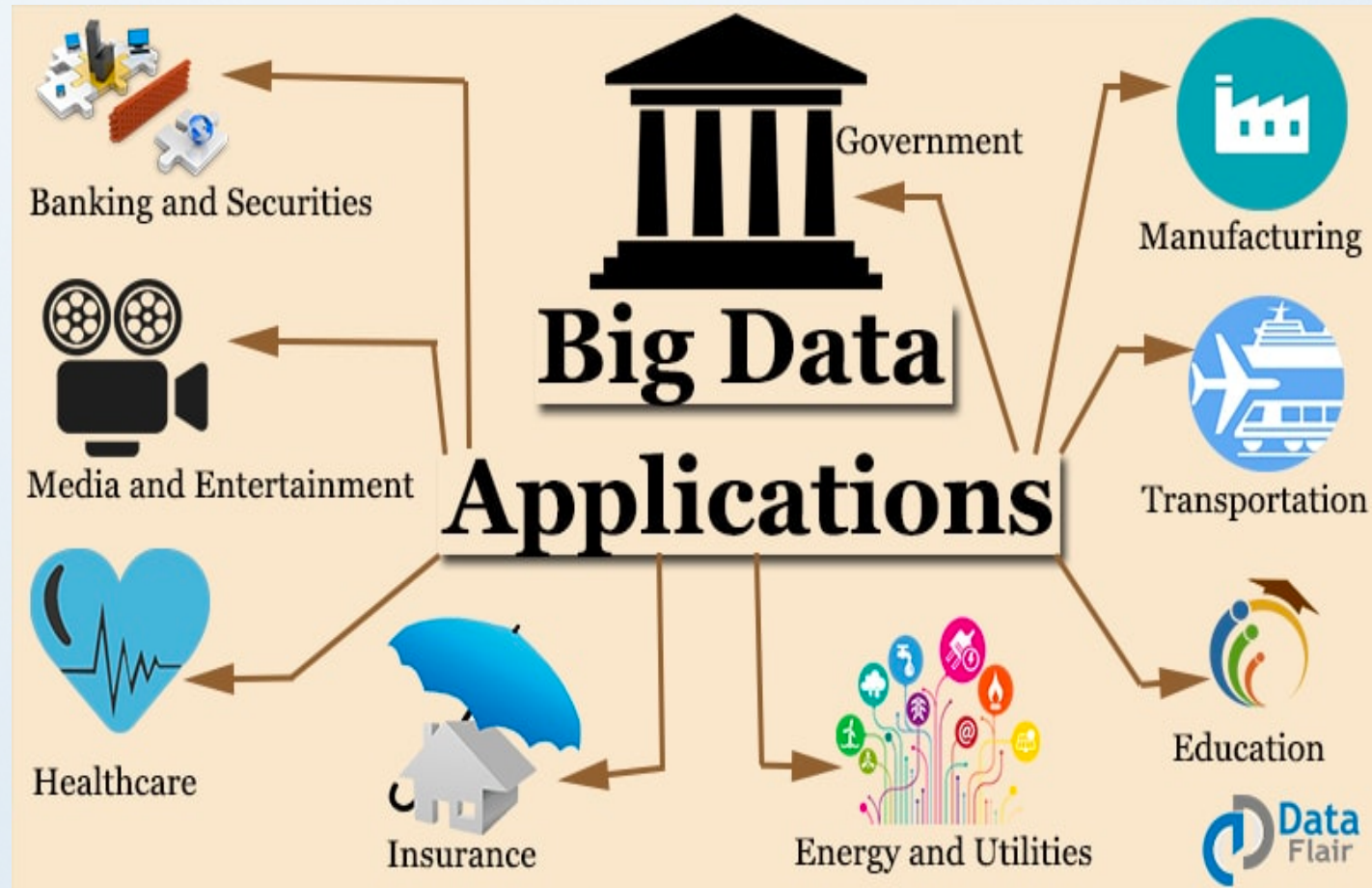
“Why” and “When” we adopt
“Hadoop”?

授課教授：王經篤 (Jing-Doo Wang)

Email: jdwang@asia.edu.tw

亞洲大學 資訊工程系 (CSIE, Asia University)

Big Data problem



From: <https://www.quora.com/What-is-big-data-analytics-how-does-Hadoop-provide-it>

From: <http://data-flair.training/blogs/big-data-applications-various-domains/>

處理大數據資料問題的趨勢

The trend to handle big data problem.

PC



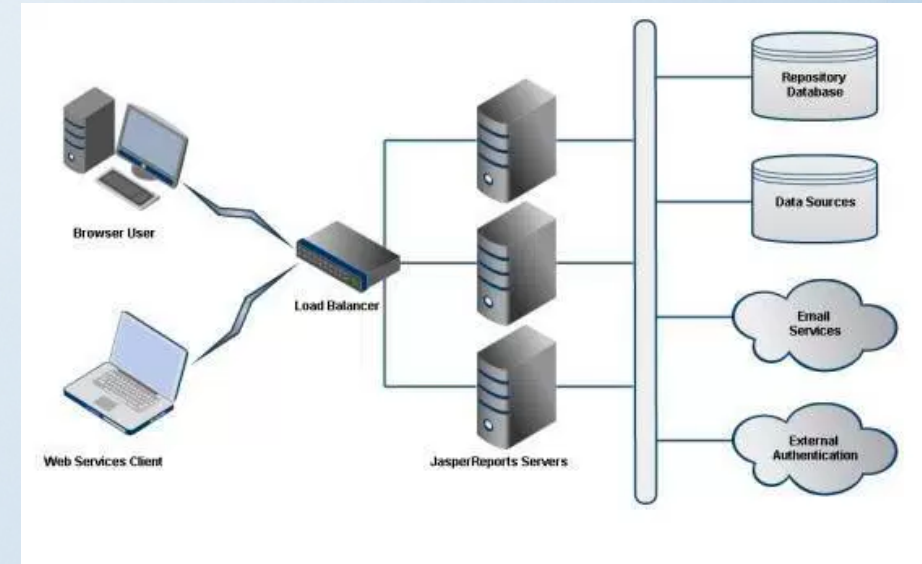
From: <http://www.occomputerrentals.com/rental-products/pc-computer-rentals/>

Server



From: <http://www.dell.com/en-us/work/shop/cty/pdp/spd/poweredge-t430>

Cluster

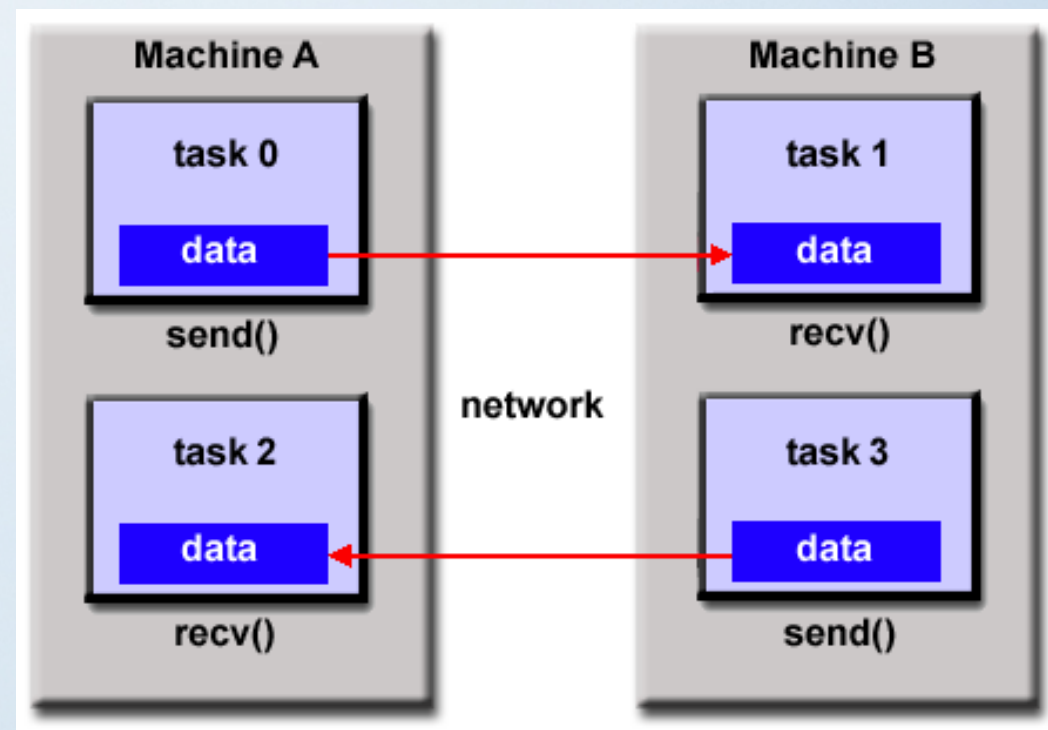


From: <http://www.surfcloudtech.com/cluster-computing-architecture/>

平行計算模組

Parallel Programming Models

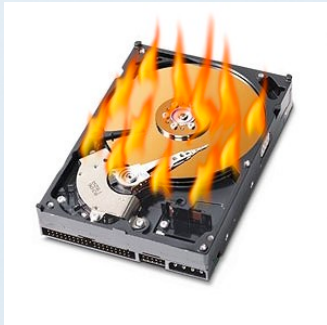
- 分散式計算 (Distributed Memory)
 - 訊息傳遞介面 Message Passing Interface (MPI)
 - [Open MPI](#)
- 困難：
 - 軟硬體價格昂貴
 - 程式設計者須自行控制
 - 資料如何存取？
 - 資料如何傳遞？
 - 無容錯？



叢集電腦的問題

The problem of using cluster computing

- 平行演算法 (Parallel Algorithm)
- 擴充 (Scale out)
- 工作分配 (Job Partition)
 - 負載平衡 (Load Balance)
- 錯誤容忍 (Fault Tolerance)
 - computing node crash?

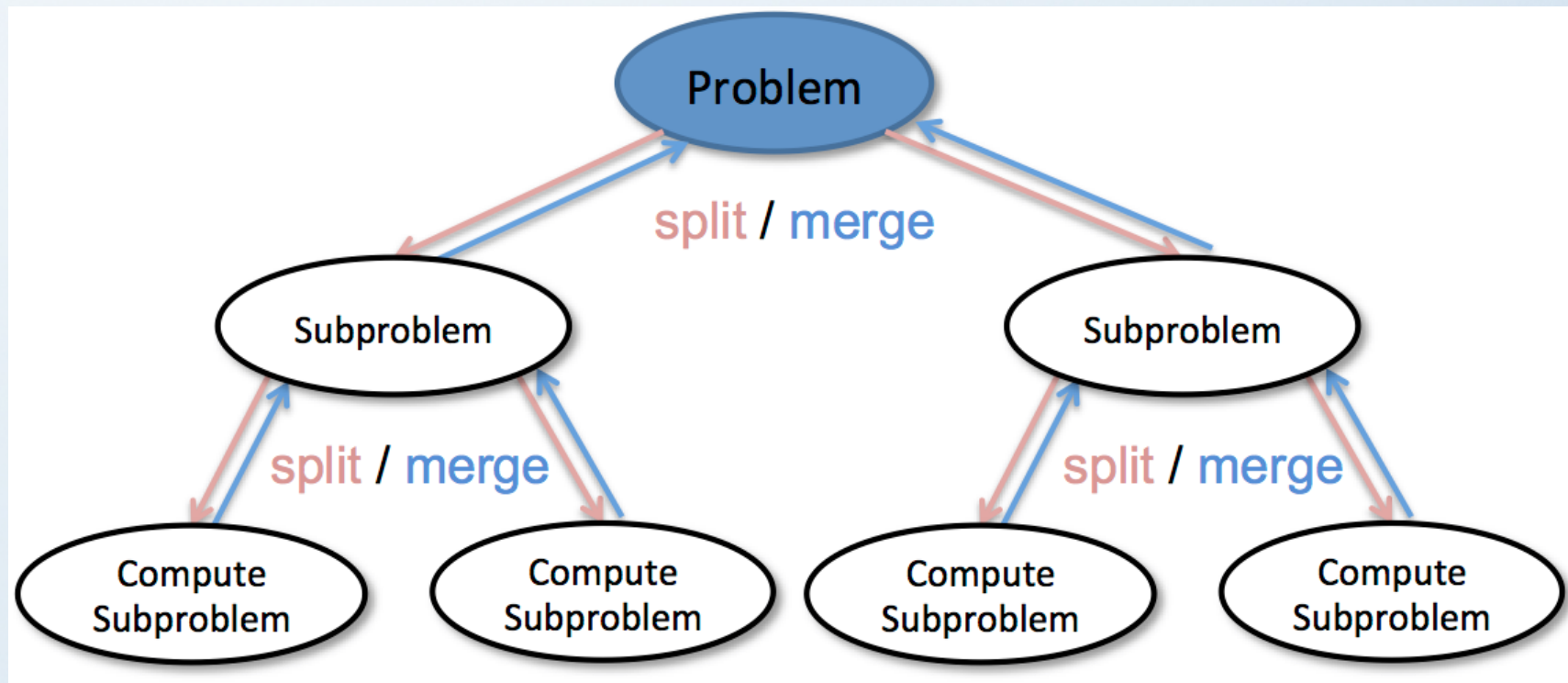


From: <http://www.surfccloudtech.com/cluster-computing-architecture/>

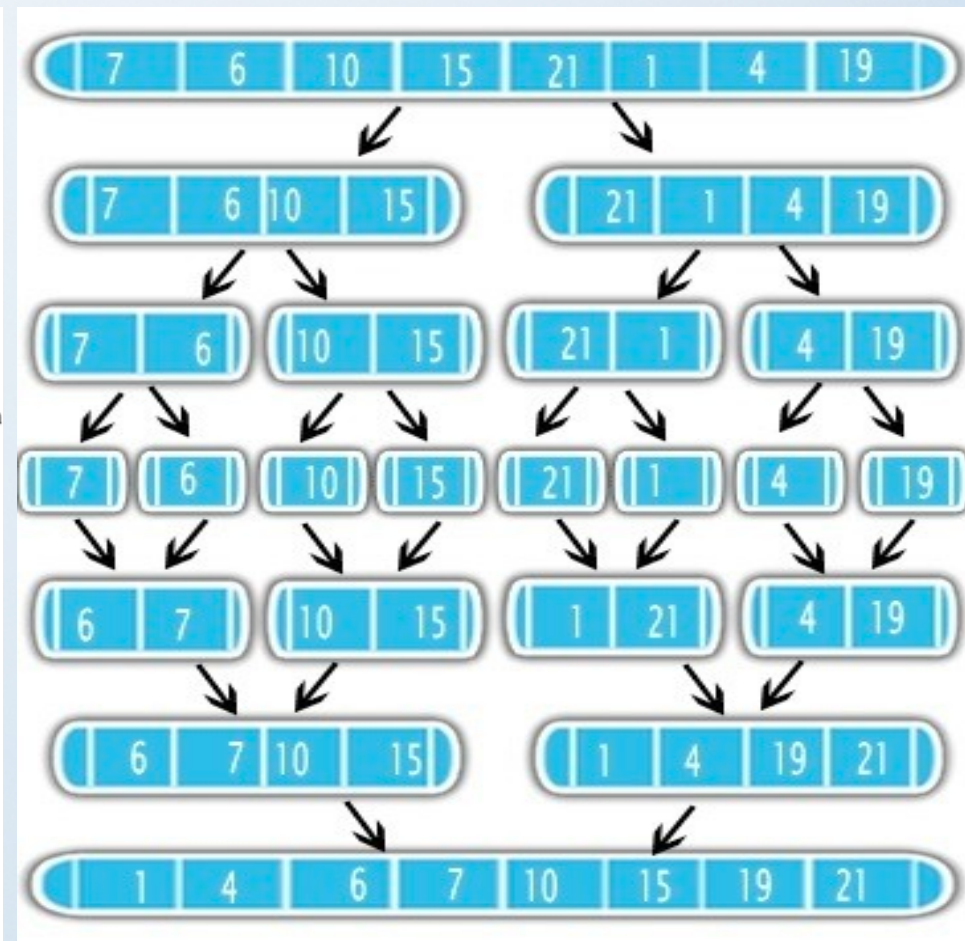
不是所有的問題都可以直接用 Hadoop 解決

Not all problems can be solved **directly** via Hadoop

- 分割和克服 演算法 (Divide-and-Conquer Algorithm)

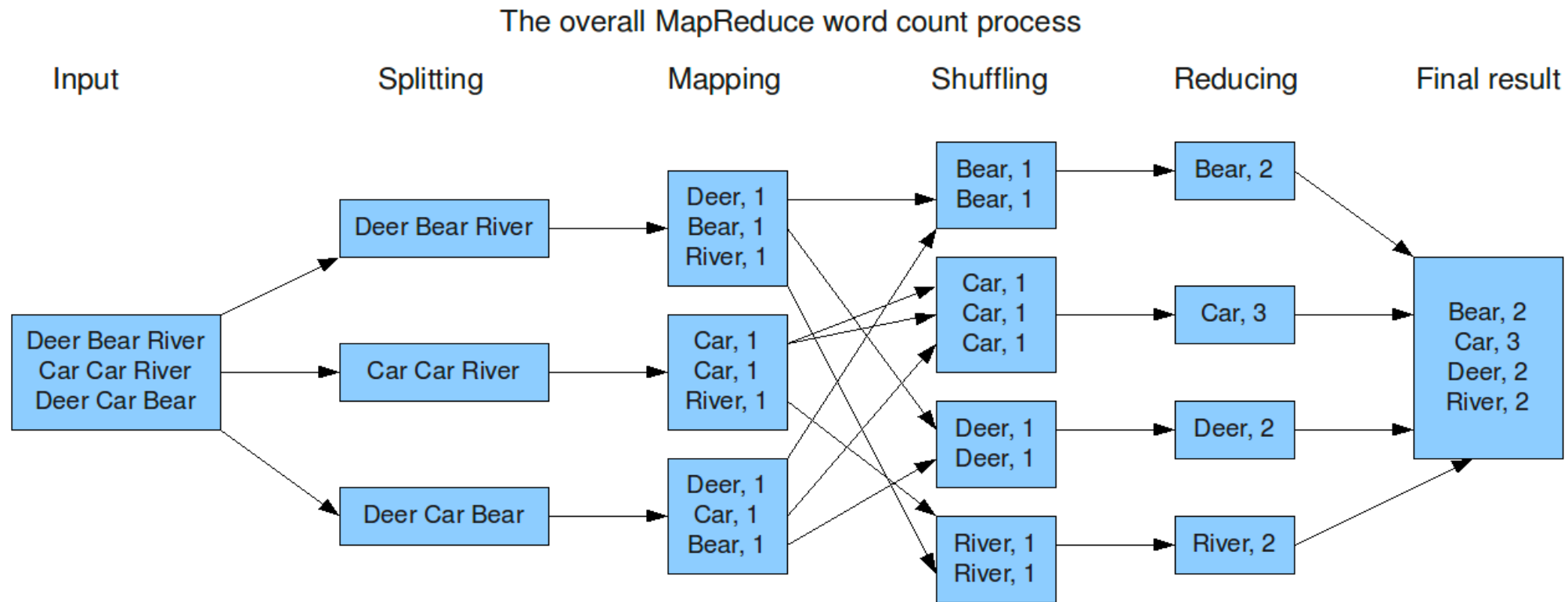


範例：合併排序法 (Merge Sort)

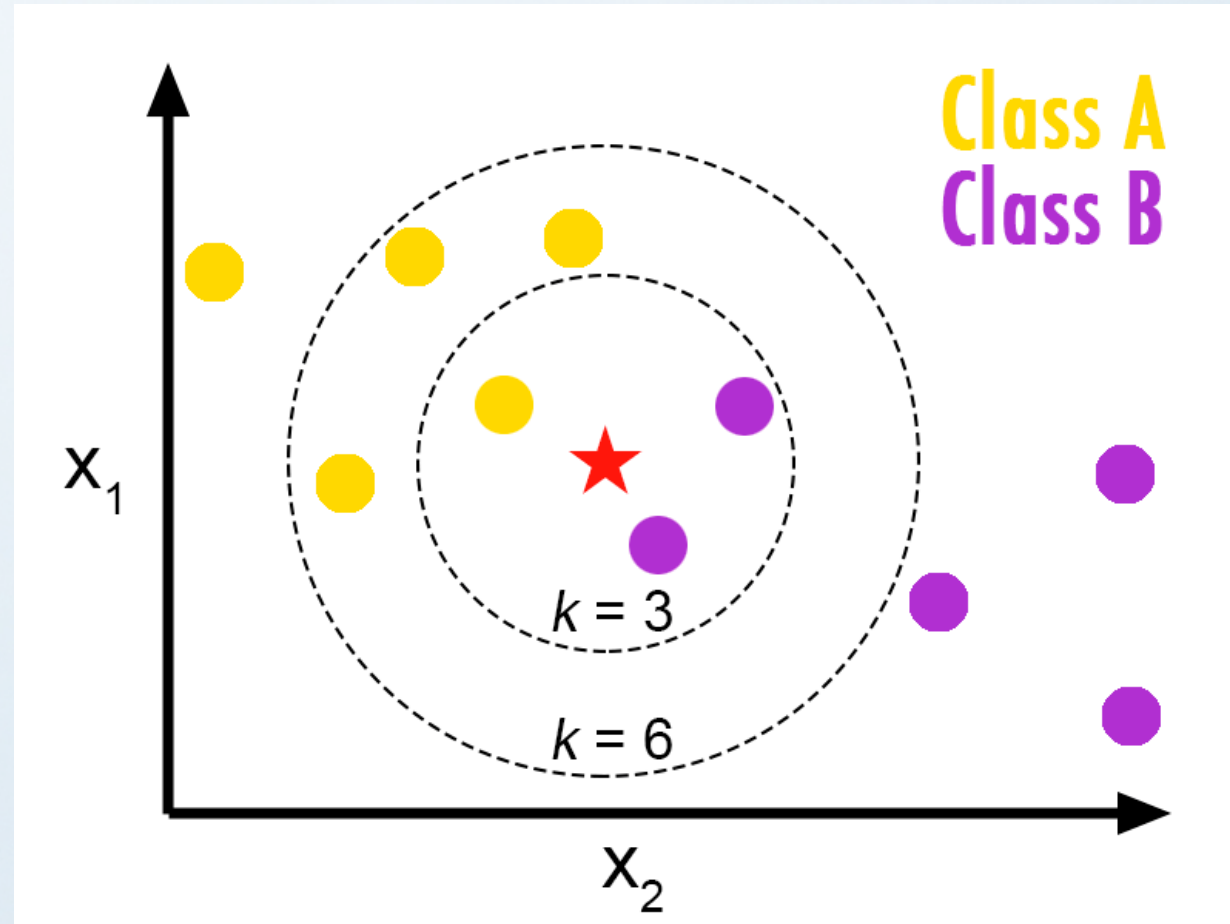


WordCount

The Most Famous Hadoop Example



The k -NN algorithm is among the simplest of all machine learning algorithms



要使用 Hadoop MapReduce來解決複雜的問題 =>不容易！？

2014 Fifth International Conference on Intelligent Systems, Modelling and Simulation

The k-Nearest Neighbor Algorithm Using MapReduce Paradigm

Prajesh P Anchalia

Department of Computer Science and Engineering
RV College of Engineering
Bangalore, India.

Kaushik Roy

Department of Computer Science and Engineering
RV College of Engineering
Bangalore, India.

From: <http://ijssst.info/Vol-15/No-3/data/3857a513.pdf>

PANDA: Extreme Scale Parallel K-Nearest Neighbor on Distributed ... (MMA Patwary 2016)

ieeexplore.ieee.org/document/7516046/

Scheduled System Maintenance

On Tuesday, September 26, IEEE Xplore will undergo scheduled maintenance from 1:00–4:00 PM ET. During this time, there may be intermittent impact on performance. We apologize for any inconvenience.

IEEE.org | IEEE Xplore Digital Library | IEEE-SA | IEEE Spectrum | More Sites

Cart(0) | Create Account | Personal Sign In

IEEE Xplore[®]
Digital Library



Access provided by:
Asia University
» Sign Out



Browse ▾

My Settings ▾

Get Help ▾

All ▾

Enter keywords or short phrases (searches metadata only by default)

Q

Advanced Search | Other Search Options ▾

Browse Conferences > Parallel and Distributed Proc... ?

PANDA: Extreme Scale Parallel K–Nearest Neighbor on Distributed Architectures

View Document

1
Paper
Citation

175
Full
Text Views

Related Articles

Joint trajectory generation for redundant robots

Reconfiguration of resources in middleware

View All

11

使用 Hadoop MapReduce 一定會提升計算速度？

- 多執行緒 (Multi-Thread)
- 圖形處理器 (Graphics Processing Unit, GPU)
- 全部資料量？
- 單一檔案大小？



From: <https://c.8891.com.tw/maserati/grancabrio/Summary.html>



From: <http://blog.udn.com/dk1858camry/1514904>

“Why” and “When” we use “Hadoop MapReduce” ?

- Why
 - 分散式計算環境 (Distributed Computing Environment.)
 - 水平擴充(Horizontal Scalability)
 - General PCs (cheap)
 - 容錯 (Fault Tolerance)
- When
 - 單台電腦無法解決的問題。
 - 問題能夠用『分割和克服』方法解決 (Divide and Conquer)
 - 不是所有的問題都適合 (須找到方法?)
 - 批次模式 (Batch Mode)
 - 不合適於 交談模式 (Interactive Mode) (Spark)