中文

成大研發快訊　第三十卷 第五期 - 2016年五月六日
[http://research.ncku.edu.tw/re/articles/c/20160506/4.html]

使用動態資料配置策略改善於異質環境下之Hadoop效能

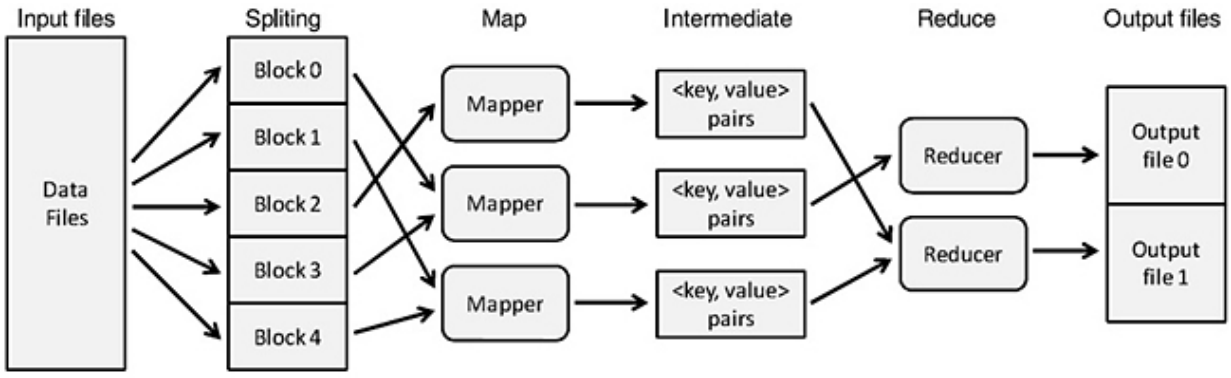
李佳衛[1], 謝光昱[1], 謝孫源[1,2,*], 蕭宏章[1]

[1] 國立成功大學資訊工程學系
[2] 國立成功大學製造資訊與系統研究所

Big Data Research, (special issue on Scalable Computing for Big Data), vol. 1, pp. 14-22, August 2014.

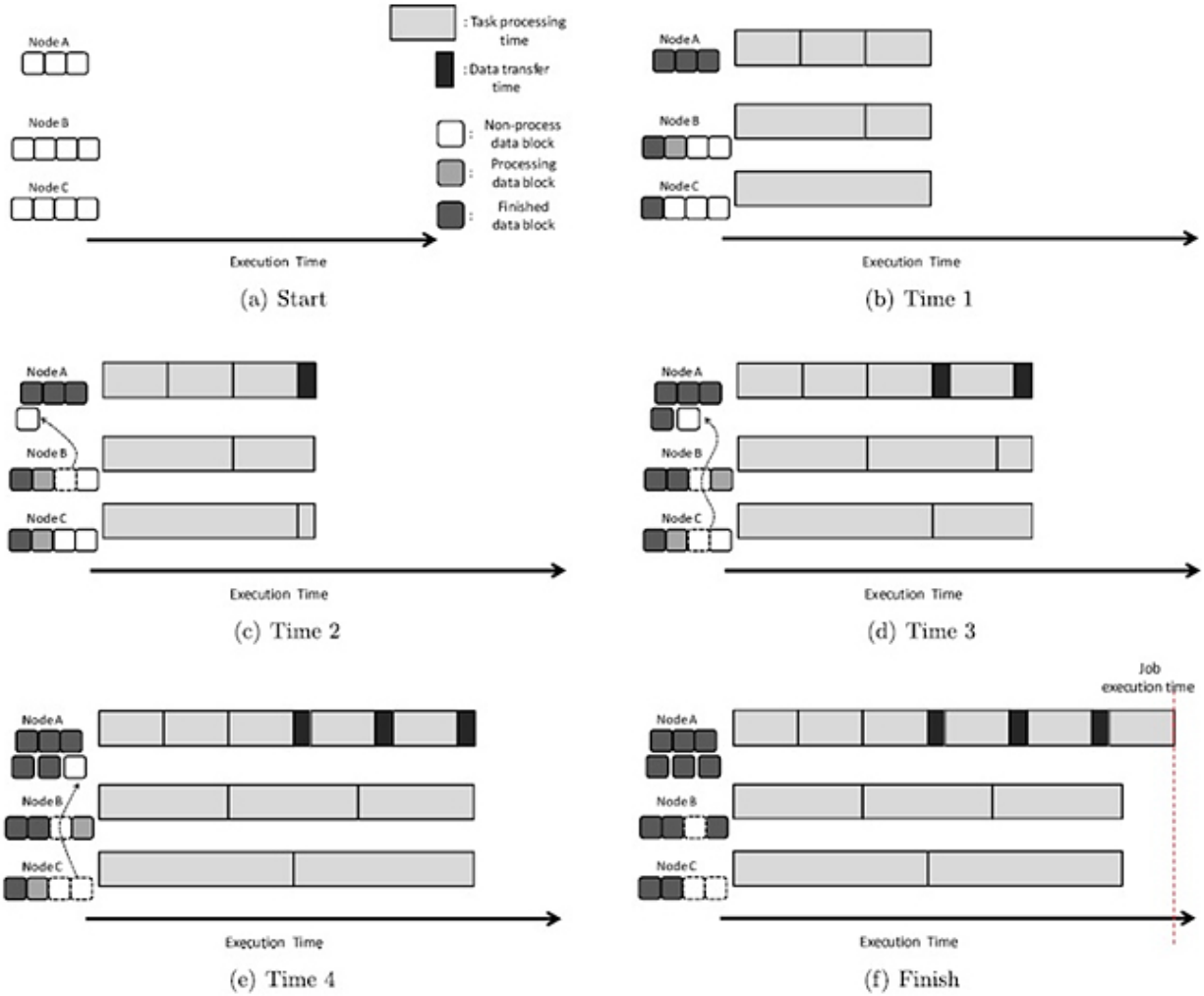
hsiehsy@mail.ncku.edu.tw

在雲端架構裡，MapReduce對於大規模的資料平行應用是一個非常重要的程式設計模組，圖一為MapReduce模組的架構。



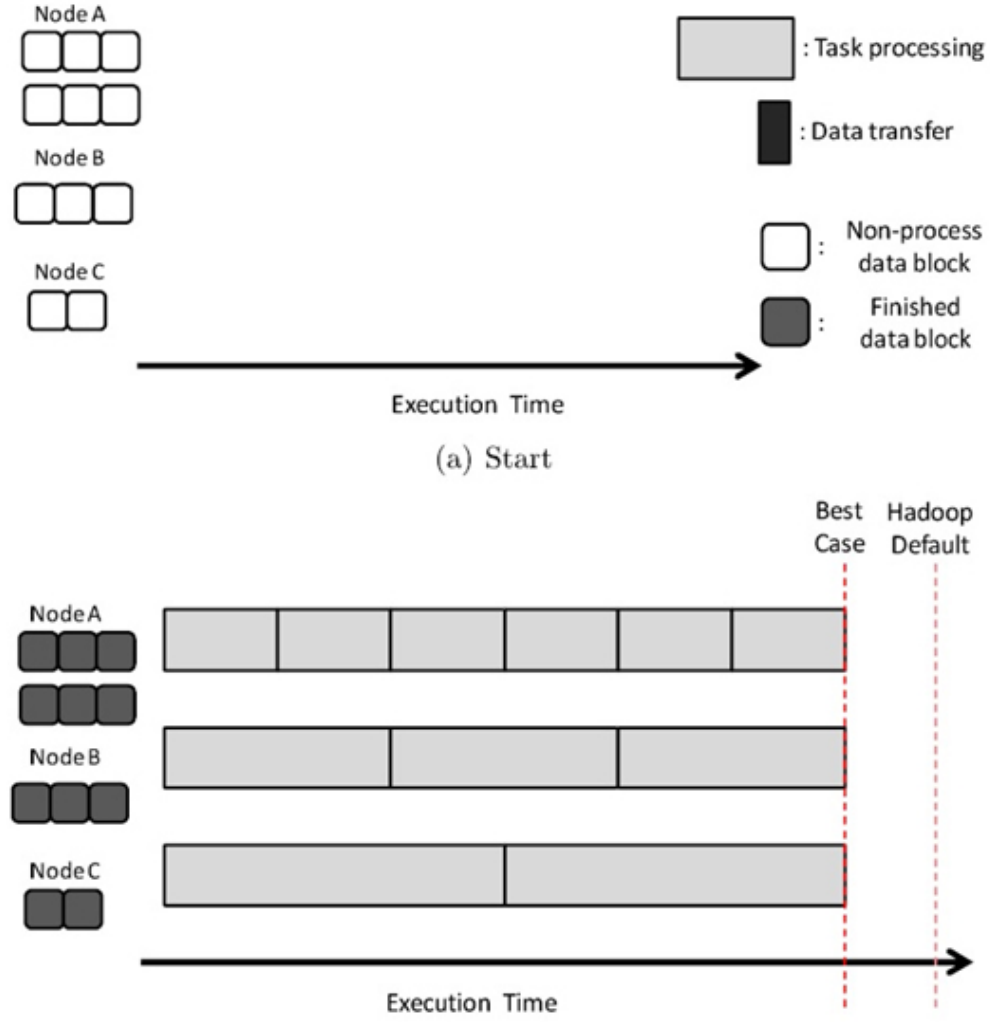
圖一：MapReduce模組的架構。

Hadoop則是一個將MapReduce模型實作出來的平台，他是屬於開放原始碼的軟體，並且Hadoop經常被使用於資料密集的應用上，像是資料探勘以及網路索引。Hadoop在運行時會假設在叢集裡所有的機器節點都擁有相同的計算能力，並且每台節點執行工作所需的資料都是在本機上的，不需要進行資料的傳輸。然而，在一些私人的叢集或是計算中心並不會符合同質性，而在這樣的異質環境底下則可能會增加額外的開銷並且降低MapReduce的效能。圖二為Hadoop預設之資料放置策略



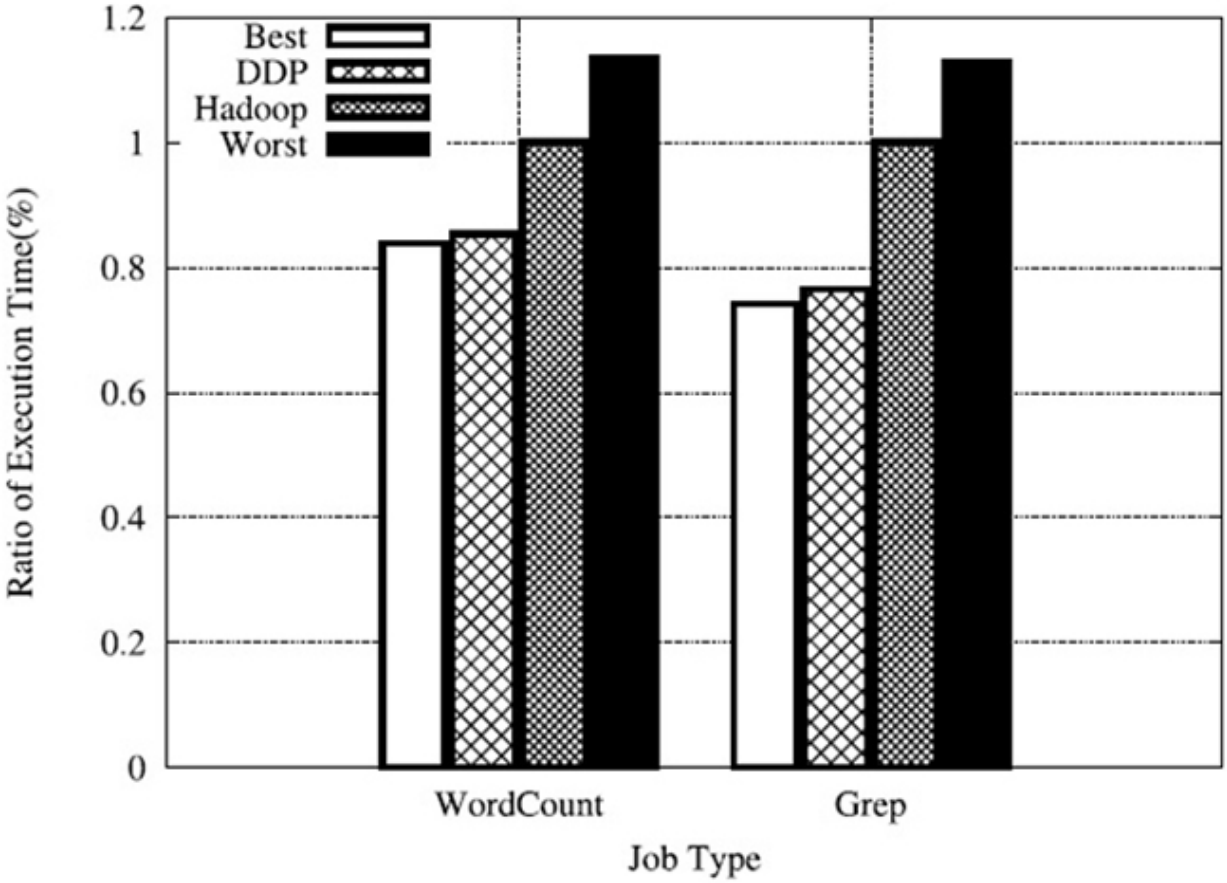
圖二：Hadoop預設之資料放置策略

本論文設計了一個資料放置的演算法，用來解決節點會有工作量不平衡的問題。我們所提出的方法可以動態的調整以平衡在每台節點上資料的儲存，而調整的方式則是根據在異質環境的Hadoop叢集裡每台節點各自的運算能力來調配，這樣可以減少時間花在資料傳輸上，來達成改善Hadoop的效能。圖三為最佳的資料放置策略。



圖三：最佳的資料放置策略

在實驗的結果顯示出，使用本論文發展的動態資料放置演算法在異質環境底下可以降低執行時間並且提升Hadoop的效能。圖四為實驗的比較結果。



圖四：實驗比較結果