期末主題中，我選擇了第9個Hadoop a Solution for Big Data - Webinar Video | Hadoop Webinar。

這學期因為上了「雲端資料處理與探勘」是hadoop去算wordcounts，課堂上老師也說明了hadoop的原理及運作方式，所以對hadoop比較不陌生。

當影片在介紹hadoop的時候，多半有些聽不懂。但是裡面的圖文表有介紹許多hadoop的相關資料，介紹hadoop在哪裡運作、通常用在甚麼地方。  
下面是影片中具體在介紹的內容。

Hadoop a Solution for Big Data :

Mike Olson is one of the fundamental(十分重要的) brains(參見) behind(背後) the Hadoop development(發展). Yet even he looks at the new type of "Big Data" programming utilized(利用) inside Google. Mike Olson runs an organization(組織) that represents(代表) considerable (相當多的)authority(權威) on the planet's most sultry(悶熱的) programming. He's the CEO of Cloudera, a Silicon Valley (矽谷)startup(啟動) that arrangements(安排) in Hadoop, an open source programming stage focused around(重點圍繞) tech that transformed(轉化成)Google into the most predominant(優越的) drive on(傳動) the web.

Hadoop is relied upon(依靠) to fuel(燃料) an $813 million product advertise by the year 2016. In any case even Olson says it’s as of now old news. Hadoop sprung from two exploration(探勘) papers Google distributed in(分佈在) late 2003 and 2004. One portrayed(描寫) the Google File System, a method for putting away(收拾) enormous(巨大) measures(測量) of data crosswise over(橫向比) a great many extremely(非常) inexpensive(低成本) machine servers, and the other nitty gritty(事實真相) Mapreduce, which pooled(匯集) the preparing power inside each one of(中的每一個) those servers and crunched all(緊縮) that data into something valuable(有價值). After eight years, Hadoop is generally(通常) utilized over the web for data dissection(解剖) and assorted types(分類) of other number-crunching(數字運算) assignments(任務). Anyway Google has proceeded(繼續) onward(前進).

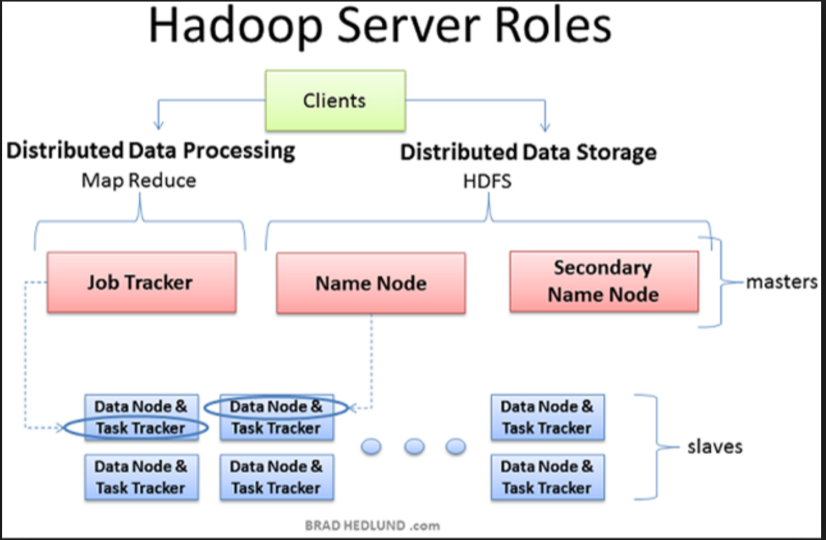
In 2009, the web monster began supplanting(取代) GFS and Mapreduce with new advances(發展), and Mike Olson will let you know that these innovations(創新) are the place the world is going. "On the off chance that you need to comprehend(意識) what the expansive(擴張) scale(規模), elite data(原種數據) preparing foundation (基礎)without(沒有)bounds(範圍) resembles(酷似), my recommendation(建議) would be to peruse(細讀) the Google exploration(探勘) papers that are turning out(發生) at this time," Olson said amid(在..之中) a late board talk close(緊密) by Wired.

On the off chance that you need to realize what the extensive scale, elite data preparing framework without bounds resembles, my recommendation would be to peruse the Google examination papers that are turning out at this moment.

Since the ascent of Hadoop, Google has distributed three especially fascinating papers on the framework that underpins its monstrous web operation. One subtle elements of Caffeine is the product stage that assembles the file for Google web search tool. An alternate show off Pregel, a "diagram database" intended(意) to guide(指引) the connections(連接) between unfathomable(叵測) measures of online data. However the most charming paper is the particular case that depicts an instrument called Dremel.

"If you had let me know heretofore me what Dremel cases to do, I wouldn't have trusted you could manufacture(營作) it," says Armando Fox, an educator(教育家) of software engineering at the University of California, Berkeley who has some expertise in(在專業) these sorts(各式各樣) of data-focus measured(測量) programming stages.

Hadoop運作原理:

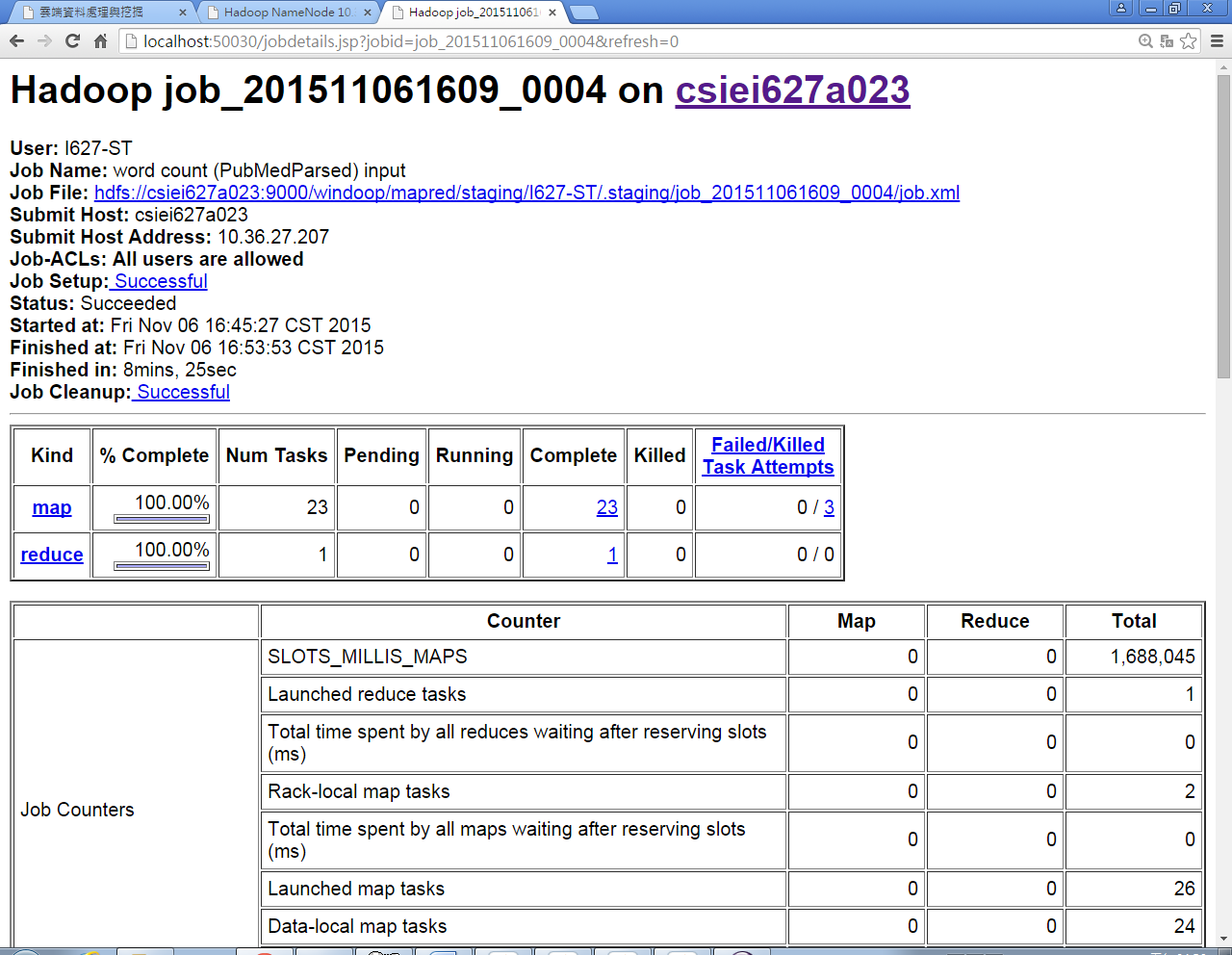


**透過分散式架構的HDFS檔案系統、搭配可分散運算的MapReduce程式演算方法，可以將多臺一般商用等級的伺服器組合成分散式的運算和儲存叢集，來提供巨量資料的儲存和處理能力。**

對於hadoop的使用，他的方便是可以叢集運算。用1台主機當master，其他主機當slaver，修改一些site的內容就可以同時用多台主機去算。Localhost會顯示那些主機運算失敗。在相同的背景環境下↓

這是用1台電腦算2年份的資料(約2G): (花了20分33秒)

  
這是用4台電腦算1年的資料↓

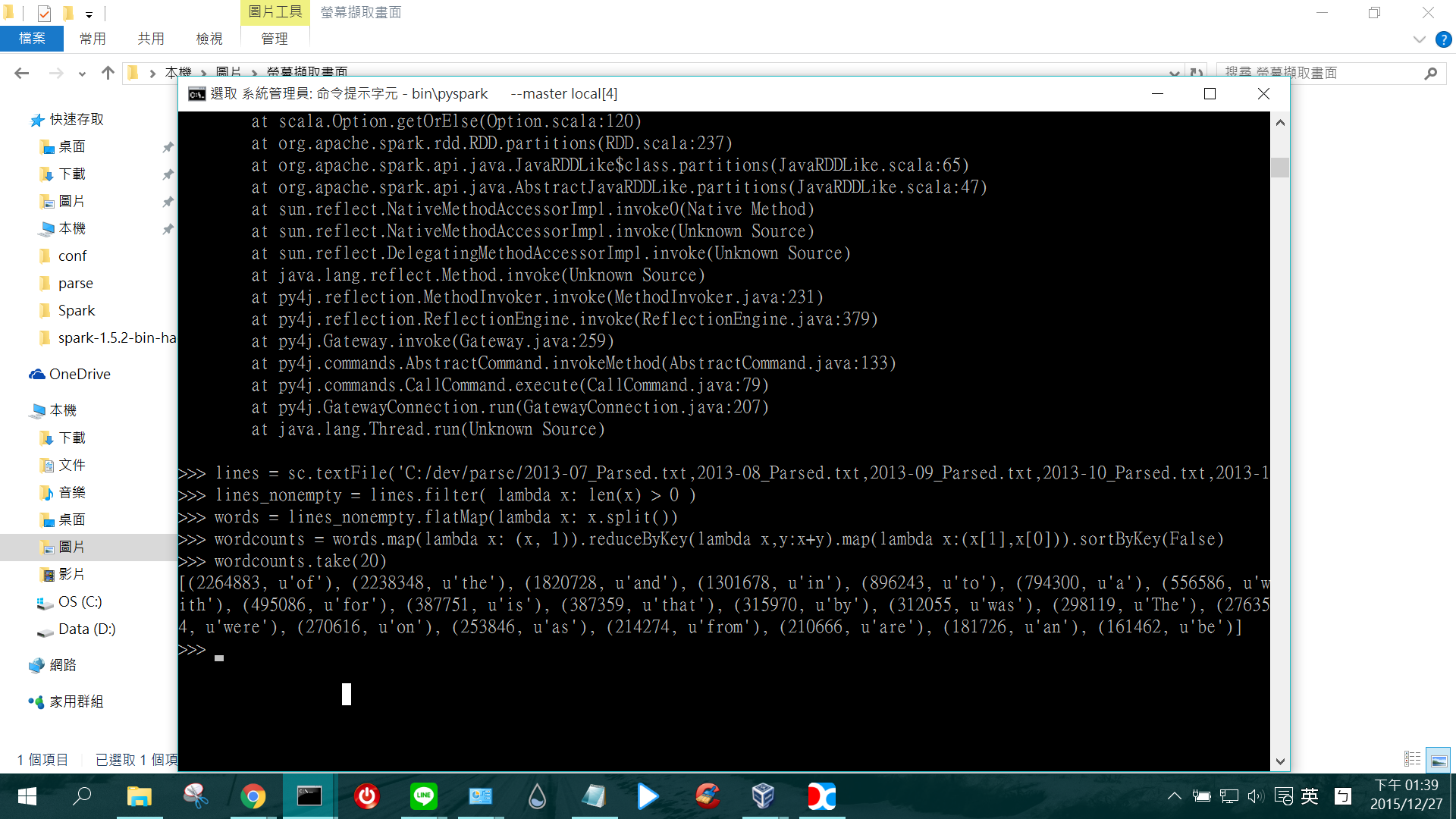
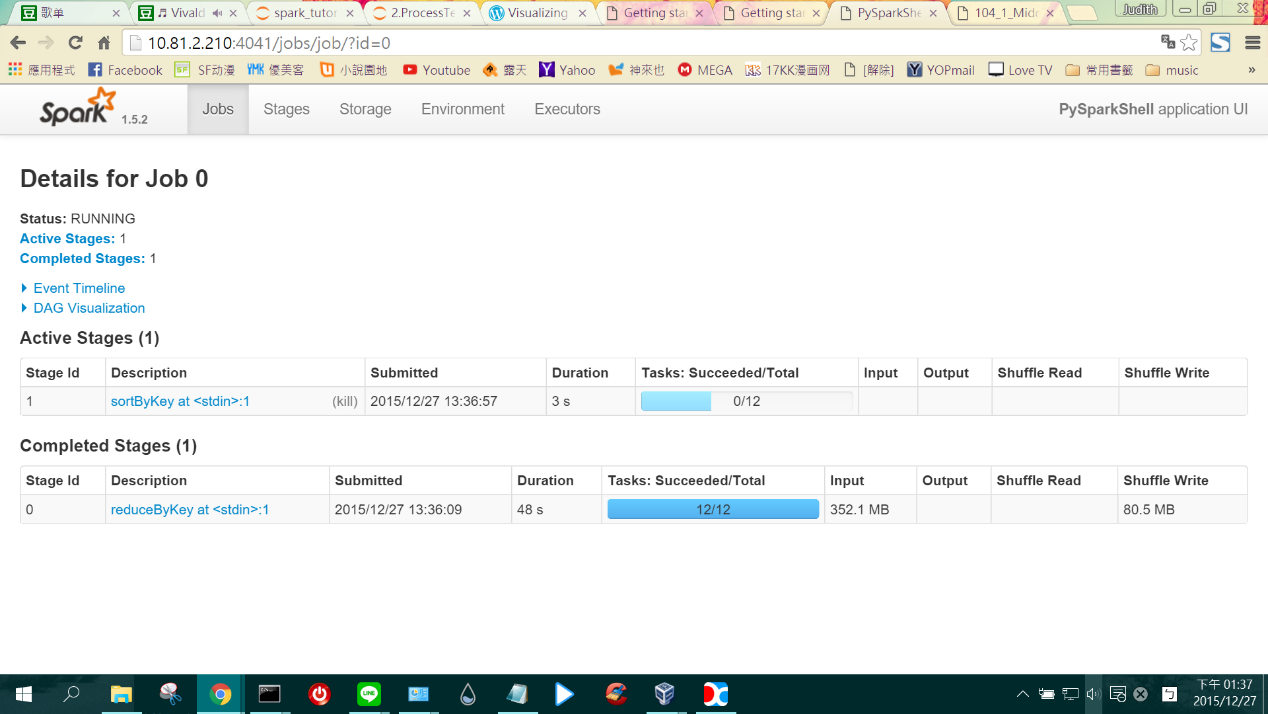
****

後來發現資料量如果越大，電腦硬體如果沒有夠多的話，時間也會相對的久。

到後來更發現了spark，它的好處是比hadoop運算更快、低成本。

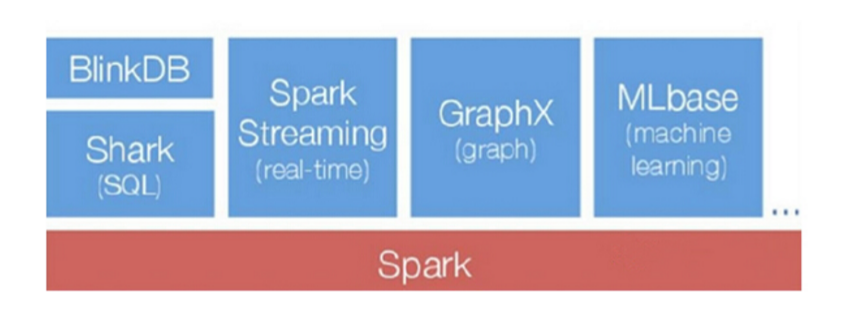
曾經試過入352MB左右的檔案用spark去算，只需要48秒就能算出來。

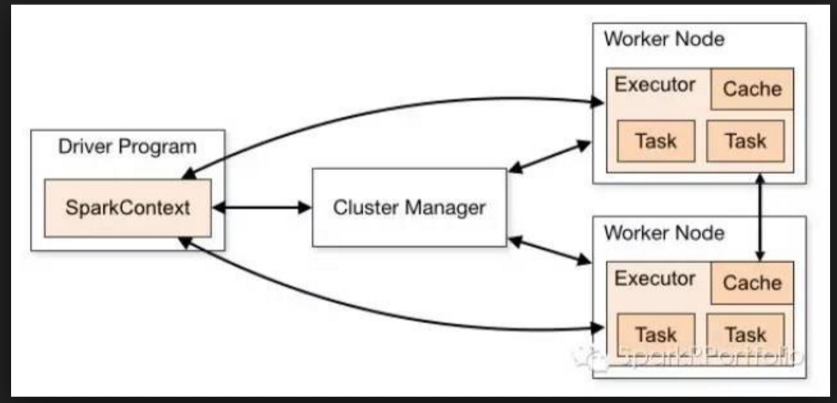
相對地用hadoop去算352MB左右的檔案卻要等5分鐘以上才能跑出結果。



Spark是現在新的技術，也被廣泛使用。可能再過1.2年spark就會像hadoop一樣了，落地深根。

Spark的運作原理:





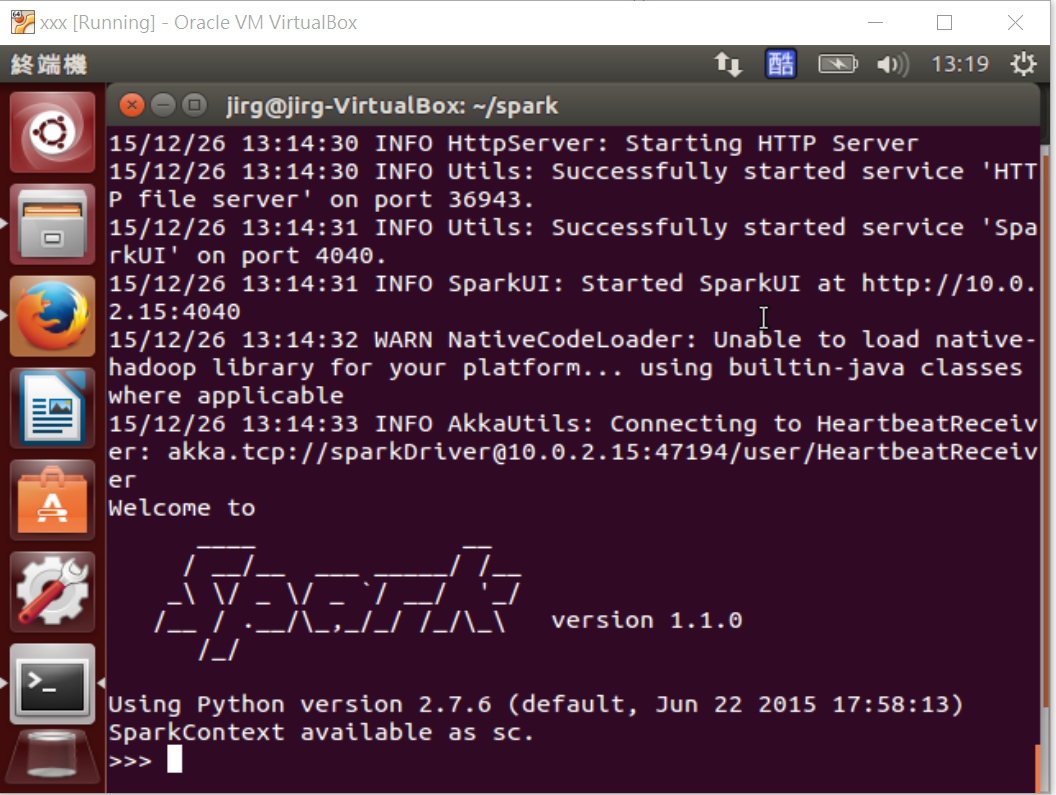
Spark是UC Berkeley AMP lab所開源的類Hadoop MapReduce的通用的併行計算框架，Spark基於map reduce算法實現的分布式計算，擁有Hadoop MapReduce所具有的優點；但不同於MapReduce的是Job中間输出结果可以保存在内存中，從而不再需要讀寫HDFS，因此Spark能更好地適用於數據挖掘等需要選代的map reduce的算法。

這裡有相關的文章: <https://taiwansparkusergroup.gitbooks.io/spark-programming-guide-zh-tw/content/spark-streaming/index.html>

http://student-lp.iteye.com/blog/2158969

Spark是滿有趣的新工具，老師也可以去看看。

這是用VM去裝spark↓



這學期上了潘老師的課，加深了對雲端系統後台運作的服務。雖然不懂的成分居多，但是再多讀幾次終於有稍微了解了。因為生活中和這些息息相關。

老師推薦的主題，10個沒有全部看完，我只看了5個有興趣的，其他的會有空再把它看完，挑了其中一個感觸比較深的來寫心得。

最後謝謝，潘老師！