出國報告（出國類別：進修）

參加 EMC 學院聯盟相關訓練課程

服務機關：國立臺北科技大學
姓名職稱：郭忠義 副教授
派赴國家：中國大陸
出國期間：103.07.13~103.07.19
報告日期：103.10.10

摘要

一百年七月一日，臺北科技大學校長接見 EMC 公司所安排大中華地區企業公民政府暨高校合作劉念寧總監研商教育合作事宜，會中決議由台北科技大學資訊工程系推派教師參與相關訓練課程。職授命分別於一百年七月二十五日、一百零一年七月二十九日赴中國大陸上海餐與訓練課程，並於一百年七月取得資訊儲存相關證照後，自一百學年度起每學年均開設一門雲端運算與巨量資料(研究所與四資四合開)相關課程，總修課人數已達 153 位學生，績效良好。今年一百零三年 EMC 學院聯盟訂於七月十四日至七月十八日假中國大陸上海辦理本年度教師培訓課程。此訓練課程以計算機科學、軟體工程、或資訊管理相關學院的教師為主，每一所院校安排符合條件的教師參加，以確保教師進修訓練後，能將資料科學的知識教育大學生。此次進修之後，職擬於一百零三學年度第一學期的物件導向程式設計課程，介紹目前業界與學界對於巨量統計分析運用最廣的程式語言 R，讓同學及早熟悉 R 統計分析程式語言的基本概念與其運用。並預計於一百零三學年度第二學期規劃開設資料科學與巨量資料運算課程。

目次

一、目的

EMC 學院聯盟作為全球公民義務和社會投資的一部分，規劃目的在與世界各地大學合作，因應近年來資訊量不斷增加、資訊複雜度不斷提高而造成的專業技術教育與企業實務斷層，與全球合作聯盟大學提供具有實務特色的暑期密集訓練課程「資料科學與巨量資料分析」。該計畫的目的在於推廣巨量資料科學基礎知識，教育巨量資料分析與建構的概念，銜接企業實務巨量資料技術與大學資訊教育落差。

本人授命分別於一百年七月二十五日、一百零一年七月二十九日赴中國大陸上海餐與訓練課程，並於一百年七月取得資訊儲存相關證照後，自一百學年度起每學年均開設一門雲端運算與巨量資料(研究所與四資四合開)相關課程，總修課人數已達 153 位學生，績效良好。今年一百零三年 EMC 學院聯盟訂於七月十四日至七月十八日假中國大陸上海辦理本年度教師培訓課程。本年度訓練課程為「資料科學與巨量資料分析」，因應網路時代的快速發展，企業需要處理的資料量越來越龐大，商業模式也要能運用這些龐大資料，分析出對商業營運模式有幫助的知識。

參與此訓練課程，可以快速獲得企業需要那些巨量分析與雲端運算的技術應用，也能及時獲得商業營運模式與巨量資料分析的關聯性。受完訓練後，可以將這些寶貴的知識與實務經驗，透過開設雲端技術與巨量資料分析的相關課程，訓練修課同學大學教學與往後實務技術運用的落差，強化同學投入雲端巨量資訊運算與應用職場的競爭力。擬於一百零三學年度第一學期的物件導向程式設計課程，介紹目前業界與學界對於巨量統計分析運用最廣的程式語言 R，讓同學及早熟悉 R 統計分析程式語言的基本概念與其運用。並預計於一百零三學年度第二學期規劃開設資料科學與巨量資料運算課程。

二、過程

1. 參與進修訓練課程：

EMC 學院聯盟為了推廣資料科學的基本技術，啟動資料科學學院聯盟計畫，於 2014 在中國大陸上海研發中心舉辦為期五天「資料科學與巨量資料分析(Data Science and Big Data Analytics)」課程，由 EMC 公司資深講師主持講解訓練。會議中還有來自中國大陸的各大學老師參與訓練。五天的訓練課程分別說明如下：

(1) Introduction to Big Data Analytics

(2) Data Analytics Lifecycle + Lab

(3) Review of Basic Data Analytics Methods Using R + Labs

(4) Advanced Analytics - Theory & Methods + Labs

(5) Advanced Analytics - Technology & Tools + Labs

(6) The Endgame, or Putting it All Together + Final Lab

圖一：訓練課程講師合影

圖一是在 EMC 公司進修訓練課程的地方與課程講師合影，圖右為本次訓練課程講師，圖左為參與訓練課程的老師，圖中右為去年與前年訓練課程講師，今年為訓練課程聯絡人與 EMC 負責學院聯盟教育訓練計畫主持人。

2. 訓練課程實驗環境：

課程實驗環境使用四台 ESX 伺服器，每個伺服器載入支援 4 個學生後臺虛擬機器。伺服器規格為 2 顆六核 x86_64 Intel Westmere X5675 CPU，記憶體 96GB RAM DDR3，本地硬碟 4 x 450GB HDD SAS 每分鐘 15K 轉速，磁片故障保護為內建 RAID 控制卡，支援硬體 RAID-1 級別，連結 IP 網路，使虛擬機器透過前端工具存取到後端。伺服器需有超執行緒能力 Intel VT enabled。第 5 台伺服器執行 windows Server 2008 僅管理虛擬機器群組。執行 vCenter server software，使用一台伺服器作為其他虛擬機器的管理端。

為了操作實驗，伺服器端的軟體環境包括 VMware 軟體－在 VMware 網站可以自由下載，並擁有 60 天的試用期。管理端上需要 vSphere 軟體以及與 vSphere 相容的

windows 作業系統。學生實驗客戶端的機器是 windows 的作業系統，或是 Ubuntu Linux 作業系統，而 Firefox、pgAdmin III 以及 terminal 在 Ubuntu 下為預設程式。客戶端學生的前端機器上， Safari Rstudio 客戶端可使用 Firefox 或 Chrome。透過 PuTTy 的 ssh 協定存取後端虛擬機器。pgAdmin III 視覺化圖形介面存取後端的 Greenplum 資料庫。

圖二：實驗課程使用的工具軟體

在 VMware 環境下部署後臺虛擬模版(ovf)，該範本部署在 vSphere 環境，伺服器規格為 2 顆 vCPU，記憶體 8 GB ，虛擬硬碟 4x 16 GB。使用 snapshot 快照功能，還需要 50%的額外硬碟空間，即每台虛機需要 96GB 存儲。圖二顯示整個實驗操作索賄用到的工具軟體，包括雲端運算平台、雲端平台虛擬機器軟體，以及巨量資料分析軟體等。

圖三：EMC 研發中心實驗室與機房

3. 參觀 EMC 研發中心研發運作：

EMC 公司在中國大陸上海的大中華研發中心，是 EMC 公司重要的資料科學軟體解決方案研發中心，裡面設有多個資料科學與儲存管理研究實驗室。本次主要參觀研發中心的環境、日常研發運作模式，以及其軟硬體配置，了解國際公司實際研發運作模式。圖三為 EMC 公司研發實驗室與機房。

三、心得及建議

1. 進修學習心得

「資料科學與巨量資料分析」課程涉及的範圍比較宏觀，主題是資料分析導向，並不局限在分析工具的使用和技術細節。著重對資料分析的脈絡進行整理，而非對操作細節的描述。實驗部分則涉及操作細節的說明。課程中對於雲端運算技術 MapReduce 和 Hadoop 屬於概念性介紹，提及定義、差別，及 Hadoop 的應用，例如 pig、Hive、Hbase 等。此課程屬於實務應用性，若把「巨量資料」分成「巨量」和「資料」，對於「資料」分析的部分，使用的工具是 R；而對於「巨量」的部分，使用的工具是 Greenplum 的資料庫，因此使用 SQL 語法。

R 語言是目前雲端運算企業界與學術界巨量資料分析使用的最多的分析語言，瞭解其技術內容對於理論研究與實務應用均非常重要。R 語言在網路上可以找到的相關學習資料越來越多，R 實驗環境 RStudio 的 R-document 更有許多完整的例子可以協助同學瞭解。在巨量分析操作指令部分，使用許多 Shell script，因此參加訓練老師需要對 Unix/Linux 系統有所瞭解。課程中沒有針對程式設計的技術細節有過多講解，如果瞭解 Linux script 的使用，對課程的理解很有幫助。

由於課程主題是資料分析，所以課程內容有一定比例在於統計分析方法，包括假設檢驗、t-test、p-value。另外對於機器學習技術包括群聚分析，分類，回歸分析等資料分析方法。參與此次課程，若沒有事前準備或之前有涉略相關領域知識，會對課程內容的理解與吸收造成一定影響。整個課程不涉及艱澀的技術理論，適合讓同學上完課後，很快的運用於即將出社會後的企業應用，以增強其雲端運算實務競爭力。

2. 建議

此次進修訓練可以瞭解全球雲端運算與巨量資料分析在業界最新的應用技術，對於回到大學開設相關課程有很大的幫助。過去兩年所開設的相關雲端運算的課程，修課同學均可獲得最新業界的實務經驗與實驗操作。建議政府部門或學校單位多給予相關的補助經費，讓臺灣的資訊教育能夠與世界產業最新科技應用接軌，

讓學生可以獲致最大的收穫以提昇臺灣整體競爭力。

四、參考資料

[1] Alain F. Zuur, Elena N. Leno, and Erik H.W.G. Meesters. A beginner guide to R.

[2] Brian S. Everitt and Torsten Hothorn. A Handbook of Statistical Analyses Using R.

[3] Guanghui Xu, Feng Xu, and Hongxu Ma. Deploying and researching Hadoop in virtual machines. 2012 IEEE International Conference on Automation and Logistics (ICAL), pp. 395 - 399, 2012.

[4] Y. Demchenko, C. de Laat, and P.Defining Membrey. Architecture Components of the Big Data Ecosystem. 2014 International Conference on Collaboration Technologies and Systems, pp. 104-112, 2014.

[5] P. Chandarana, and M. Vijayalakshmi. Big Data analytics frameworks. 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications, pp. 430-434, 2014.

[6] M. N. O. Sadiku, S. M. Musa, and O. D. Momoh. Cloud Computing: Opportunities and Challenges. IEEE Potentials, 33 (1), pp. 34-36, 2014.