國立交通大學
National Chiao Tung University

出國報告（出國類別：學研訪問）

參訪美國 IBM T.J Watson Research Center

服務機關：交通大學電機工程系
姓名職稱：王蒞君教授
派赴國家：美國 IBM T.J Watson Research Center
出國期間：2014/9/5－2014/9/6
報告日期：2014/10/2

摘要

美國 IBM T.J. Watson Research Center 有意與王蒞君教授合作在交通大學成立江河運算平臺研究中心，王蒞君教授藉由這次到美國 IBM T.J. Watson Research Center 的參訪進行研究中心成立事宜的討論以及聽取研究中心學者的報告以了解目前在江河運算在產業中的應用。

目次

一、目的

海量資料運算為雲端運算發展的趨勢，但是海量資料的運算面臨一個重大的問題：即時性，Apache 所提出的 Hadoop 系統因為檔案存取的時間過久，導致海量資料的運算與分析無法即時處理，但是現今有許多應用需要即時運算出結果，因此 IBM 提出 InfoSphere Streams 平臺，稱為江河運算，利用資料流的概念處理資料，資料不被儲存，資料流過即運算，可即時針對海量資料分析出結果。

有鑑於即時資料分析日趨重要，美國 IBM T.J. Watson Research Center 的張書平博士與本人於今年 7 月在交大舉辦 InfoSphere Streams 教學的 workshop，培訓了交大 24 位的學生，透過培訓將 IBM 的即時資料分析平臺深植學界，除了 7 月的培訓外，張博士更進一步提出希望與本人合作在交通大學成立江河運算研究中心，透過研究中心建立產學合作的橋樑。

因此本人此次前往美國 IBM T.J. Watson Research Center 除了與張博士和其高階主管討論研究中心成立的事宜之外，更透過參訪進一步了解此即時資料分析平臺的操作使用以及目前美國 IBM 是如何將此即時資料分析平臺與業界進行合作，本人還進一步與美國的學者取經，希望能夠在下學期在交通大學開課，在交大培育出更多資料分析的人才。

二、過程

參訪過程如下表，9/5 早上先由 Dr. Daby Sow 介紹江河運算平臺以及江河運算如何實現在醫療照護上面。在江河運算中，資料不儲存在資料庫中，資料流過運算節點時就直接立即進行運算，Streams 平臺中有許多的節點被邊所連接，在

平臺內的運算節點稱為 operator 或是 adapter，operator 或 adapter 會在資料經過時對資料進行運算如圖一，圖二為在 Streams 平臺上的簡單範例，functor 會將進來的資料轉成可被運算的格式並將資料送往 operator，此範例中的 operator 為 split operator，split operator 會將資料分流放入資料庫或是檔案系統，從範例中可以看出在 InfoSphere Streams 平臺的資料在流動時就已經被計算得出結果並記錄資料的特色模型或是被判斷要前往什麼地方，不像傳統的資料會先被存入資料庫，等到運算需求進入後再把資料從資料庫拿出來計算，此種資料不儲存的運算模式在海量資料的運算上可以對資料進行即時分析，即時得出結果。

此種即時資料運算很適合用於需要即時針對病人狀況判斷的醫療照護，因此 IBM 從 2009 就開始將 Infosphere Streams 平臺運用於重症新生兒的監測上，透過分析新生兒的血氧濃度、腦氧濃度等指數歸納出某種突發狀況的前期症兆，當有其他新生兒有類似狀況時可即時通知醫生護士進行搶救，如此可大幅減輕重症病房中醫生和護士的負擔，2009 年的試驗相當成功，因此 IBM 開始陸續與哥倫比亞大學的醫學中心以及 UCLA 的健康中心等單位進行更進一步的研究與合作，如圖三。

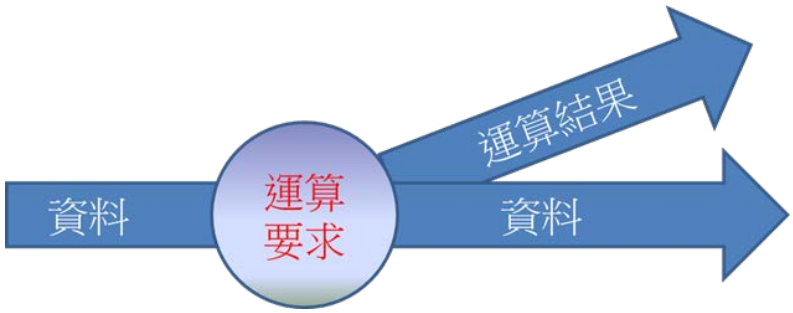
午餐結束之後，本人拜訪 Dr. Deepack Turaga，Dr. Deepack Turaga 曾在哥倫比亞大學教授江河運算課程，王蒞君教授透過此次與 Dr. Deepack Turaga 的取經，預計於下學期在交通大學開設相關課程以培育更多的即時資料分析人才。Dr. Deepack Turaga 的教材一開始會先介紹目前針對大資料運算的平臺，例如：Hadoop，但是 Hadoop 在運算時容易遭遇到硬碟 I/O 的讀取速度而成為運算的瓶頸，因此 Hadoop 並不適用即時資料的分析運算，在即時運算方面，紛紛有 Storm 以及 Spart Streaming 等平臺的提出，但是此些平臺上針對運算效能最佳化、資源擴充性以及資料容錯復原機制的設計都還不盡完善，IBM InfoSphere Streams 就是改進以往平臺的缺點並加入運算資源最佳化、資源擴充、容錯復原等功能，使得不儲存資料的即時資料分析變得更加穩定。本課程即時針對 InfoSphere Streams 上的 middleware 以及各種最佳化容錯機制的設計加以介紹，除了講述理論外，

本課程在學期末的時候會引領學生製作一個專題，有些學生利用江河運算平臺對股票進行即時分析，希望能夠透過此平臺大撈一筆，也有學生利用此平臺對新聞資料進行分析，透過分析讀者的意見和瀏覽次數可以歸納出哪些是好新聞哪些是不好的新聞以推薦給其他讀者，學生們都相當有創意，著實能夠將 InfoSphere Streams 平臺運用到生活中。相信王蒞君教授下學期在交大的課程也能帶領學生做出有創意發想的即時資料分析系統，圖四為王蒞君教授與 Dr. Deepack Turaga 的合照。

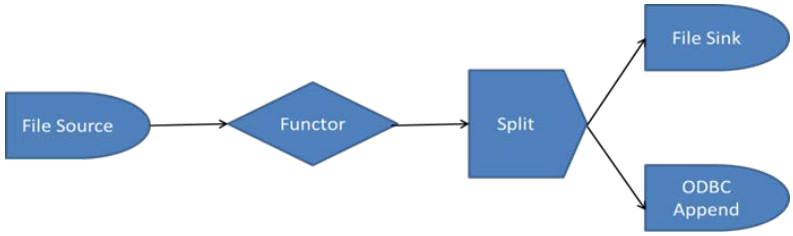
9/6 的行程則是由 IBM THINK Lab 進行 Watson 電腦的展示，如圖五，Watson 電腦是能夠使用自然語言來回答問題的人工智慧系統，由 IBM 公司的首席研究員 David Ferrucci 所領導的 DeepQA 計劃小組開發並以該公司創始人 Thomas·J·Watson 的名字命名。2011 年，華生參加綜藝節目危險邊緣來測試它的能力，這是該節目有史以來第一次人與機器對決。2 月 14 日至 16 日廣播的 3 集節目中，Watson 電腦在前兩輪中與對手打平，而在最後一集裡，Watson 電腦打敗了最高獎金得主布拉德·魯特爾和連勝紀錄保持者肯·詹寧斯。Watson 電腦贏得了第一筆獎金 100 萬美元，而肯·詹寧斯和布拉德·魯特爾分別只有 30 萬和 20 萬。賽後，詹寧斯和魯特表示將一半獎金用於慈善事業，IBM 公司也將 Watson 電腦的獎金分給了兩家慈善機構。Watson 電腦在比賽節目中按下信號燈的速度始終比人類選手要快，但在個別問題上反映困難，尤其是只包含很少提示的問題。對於每一個問題，Watson 電腦會在螢幕上顯示 3 個最有可能的答案。Watson 電腦 4TB 磁碟內，包含 200 萬頁結構化和非結構化的信息，包括維基百科的全文，Watson 電腦在比賽中華生沒有連結到網際網路，圖六即為 Watson 電腦。

Program

Date	Time	Agenda	Speaker
9/5	10:00 – 12:00	InfoSphere Streams on Healthcare	Dr. Daby Sow
	12:00 – 13:30	**Lunch**	
	13:30 – 15:00	How to teach InfoSphere Streams	Dr. Deepack Turaga
9/6	13:00 – 15:00	Watson Demo	IBM THINK Lab
	15:00 – 18:00	Discuss the research center and Dinner	



圖一、江河運算運算模式。

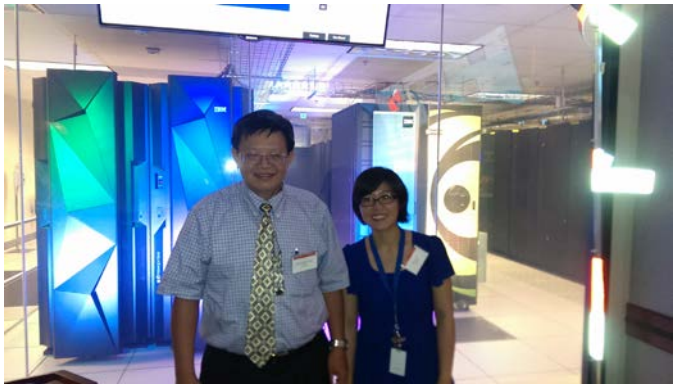


圖二、江河運算範例。

圖三、IBM 在醫療照護上的合作對象。

圖四、王蒞君教授與 Dr. Deepack Turaga 的合照。

圖五、IBM THINK Lab 展示。



圖六、IBM Watson 電腦。

圖七、與 IBM T.J. Watson Research Center 牌子合照。

圖八、王蒞君教授與張書平博士合照。

三、 心得及建議

本次行程收獲相當豐碩，交通大學與 IBM 合作的即時資料運算平臺研究中心預計於今年年底正式成立，研究中心成立後，交通大學的學生將可免費使用 InfoSphere Streams 平臺進行學習與開發，同時本人也將於下半年度在交通大學開設相關課程，積極培養資料分析的人才，為臺灣的科技產業注入新的活力，研究中心成立後也將成為 IBM 與臺灣業界接軌的橋樑，將可帶來更多合作的機會。