BIGDATA
運算技術成熟下的資料分析與應用
專訪 遠通電收資訊處 陳聲鏗 副總經理

文/王奕超

若論當前最火紅的科技潮字，除了物聯網（Internet of Things， IoT），大概要屬於 Big Data了，從企業到政府，從學術專論到廣告標語，Big Data幾乎無人不談，無人不喊。撇開Big Data到底要譯成「大數據」還是「巨量資料/資訊」，還是「海量資料/資訊」這樣的枝節爭論，我們對於這個萬眾矚目的科技詞彙和焦點，究竟該有什麼正確的認識呢？

舊概念，新效能

「Big Data是一個新的名詞，但並不是一個很新的概念。」說起這個從前年「夯」到今年，並且依舊持續發燒的關鍵詞，成日打滾於龐大的國道交通資料的遠通電收資訊處副總經理陳聲鏗，開門見山強調，Big Data的基本概念，其實就是過去已經存在，並被廣泛運用於管理和決策的資料倉儲（Data Warehouse）[註1]、資料探勘（Data mining）[註2]，只是隨著科技發展讓資料搜集和處理能力的提升，讓這些應用的效益更為驚人。

就資料搜集面來說，各種資訊感測、採集的技術在精準度和效率的提升以及建置成本的下降，讓我們更容易搜集到多樣的資訊；以資料處理能力而言，Hadoop、Spark [註3]這些新興運算框架，則讓過往的資訊處理、分析效能有極大的躍進，傳統上難以處理的非結構化資料，如圖片、圖形、膠捲影片、手寫文檔、聲音檔等各種類型的資料格式得以迅速被轉化為格式相同、關係明確的結構化乃至於數位化的資訊。透過物聯網時代發達的資料搜集能力和強大運算框架，過去不會被搜羅、無法被分析的資訊都將被含納進來，讓可運用、參照的資料種類更為寬廣，而參照面向越寬廣，所做出來的分析也將越精確，有了精確的分析，預測也就越準確，決策也將更正確。

註1資料倉儲是一種將大量資料透過系統性整理的儲存架構，其具有整合性、主題導向性、注重資料隨時間的變化、存入後不會變動等特性，利於決策者快速有效的自大量資料中，分析出有價值的資訊進行運用。

註2資料探勘是指透過統計學方法與電腦運算從大量的資料中搜尋、挖掘出隱藏於其間的關聯性、價值、意義的過程。

註3 Hadoop 是一種開源式叢集運算框架，它能將數以千計的伺服器，整合應用起來像是一台超級電腦。該框架採用HDFS 分散式檔案系統（Hadoop Distributed File System），透過數以千計的節點來存放資料。為了妥善保存檔案，它會先將檔案分割成數小塊，並且把每小塊拷貝成三份，再將這些小塊分散給各工作節點保管。這些工作節點的運作再由一個主節點監管，如果發現有哪個工作節點上的檔案遺失或遭到損壞，就會尋找其他副本進行複製，保持每小塊的檔案在整個系統都有三份的狀態。在資料處理上，傳統作法，是將整個檔案丟進程式軟體中做運算出結果，而面對龐大的資料，Hadoop則是採用分散式計算的技術處理各節點上的資料，並將各節點運算出的結果直接傳送回來歸納整合。這樣多方並進平行處理，可以大大節省龐大資料處理的時間。

至於Spark則是更新竄起的開源式叢集運算框架，採用了記憶體內運算技術（In-memory），在資料尚未寫入硬碟時即在記憶體內分析運算。Spark在記憶體內執行程式的運算速度能做到比Hadoope的運算速度快上100倍，即使是執行程式於硬碟時，Spark也能快上10倍速度。可以用較少的節點數量，達到比Hadoop還高的執行效能。

關鍵在於解決什麼問題

「然而Big Data真正重要的是在於用途，不管是分析、預測跟決策，總會繞著一個目的，也就是要解決什麼問題。」陳聲鏗認為，在談或應用Big Data之前，最重要的是要先確認待解決的問題，任務確立後，才能了解需要哪些資訊、多大的資訊量，要運用什麼樣的方法。

Big Data最常被運用的領域是商業行銷。陳聲鏗指出，有關Big Data最被廣為徵引的例子，就是美國大零售商 Target的商業運用案例。這個案例是這樣的，據說曾經有一名中年男子怒氣沖沖跑到Target的某家分店裡，質問他們為何寄發嬰兒用品優惠券給他的未成年女兒，難道是鼓勵她懷孕嗎？幾日後，該店店長致電給男子，準備向該名男子道歉，誰知道那名男子在電話中竟主動道歉，原來，他女兒確實懷孕了，同時男子也好奇詢問他們是如何能夠判斷出他的女兒懷孕。究其原因，是Target電腦系統透過少女的購物紀錄推斷出少女有很大的機率懷有身孕，對於嬰兒用品可能有其需求，因此才會主動寄發相關優惠資訊。這個案例雖然無從考證其真實性，卻也生動地說明了Big Data在商業行銷上的用處和預測的精準程度。

除了商業行銷，Big Data還有一個很大的用途，就是降低維運成本。陳聲鏗以遠通電收自身在高速公路電子收費（Electronic Toll Collection， ETC）系統的維運為例，為了讓颱風天或各種其他斷電情況下，ETC系統能夠持續運轉服務，其計費門架旁都設有類似不間斷電源（Uninterruptible Power Supply，UPS）系統的SMR(Switch Mode Rectifier，開關整流器)機櫃，然而SMR的電池是有壽命的，原本可能撐4到8小時的電量，到了老化的階段可能只能撐10分鐘、5分鐘，也就是一斷電就無法支撐太久，為了防範ETC在斷電時系統停擺，過往的做法是以兩年為限，定期全部汰換，可是，定期更換的問題是，汰換下來的電池，其實有9成都還有足夠的電量去運作，這無形中也造了一種成本上的浪費。為了讓維運管理更有效益，遠通電收後來特別在機櫃裝設電壓感測器，量測電池放電的電壓變化，透過對於歷年電壓曲線的變化進行觀察，就可以分析出電壓曲線在什麼狀態時，代表電池壽命即將耗盡，因而可以即時進行更換，讓系統的維運更為精準。陳聲鏗強調，現在像這樣的應用其實很多，對於各種設備運作的數據監管其實十分通行，可說是當前Big Data最熱門的應用項目之一。

現階段，遠通電收與高公局合作致力於運用ETC 系統累積下來的Big Data來探索交通問題，建構高速公路的智慧型運輸系統（Intelligent Transportation System，ITS），增進國道交通服務的效率。像前陣子，他們為了研究年假國道的壅塞狀況，除了分析車流資料和天氣資訊，還特別將社交媒體、社群網站的討論訊息也放進處理平台進行交叉比對，結果就發現這些塞車路段周邊原來都有新興景點，例如屏東的琉璃橋、嘉義的高跟鞋教堂和故宮南院、南投的龍鳳瀑布、林口的三井outlet等，因而也初步歸納出年假交通壅塞與觀光景點的因果關係；近期他們則利用高速公路實施計程收費所累積的114億筆資料建置出一套民眾在高速公路的用路習慣的模型，透過即時資料匯入，可望分析出哪些時段與路段的壅塞情況，未來他們打

算將以該資料模型為基礎，開發成一個交通App，為用路人提供30分鐘後的即時交通路況，並為其規劃出最佳交通路徑。陳聲鏗表示，將來如果各縣市的交通資訊也能開放出來，遠通電收將可建置出更強大、關照範圍更廣的交通App。

重要的是領域知識

由於相關技術工具的普遍化而取得容易，目前一般企業或單位要進行Big Data的基本部屬，無論是感測裝置或後端運算平台並不困難。「除非像是遠通電收這種資料高度機密，不能出Data center（資料中心）的，運算分析要自己處理的事業，否則直接向電腦大廠購買雲端解決方案，其實就可以實現基本的效果。」陳聲鏗倒是認為，Big Data真正專業、有門檻的部分，其實是所要應用領域的領域知識，這部分須借重各領域的專家學人，就像他們從事交通領域的應用，須借重交通管理的專家。雖然，沒有領域知識仍可藉由演算法的方式，借助運算平台進行分析，並透過機器學習（Machine Learning）能力去驗證、修正，逐漸摸索出結論，然而如果擁有領域知識輔助思考，對於該領域的資料連結和應用會有比別人更豐富的想像，因此他強烈建議在實踐Big Data應用時，應讓資訊專家和應用領域專家相互搭配、結合，才可能有事半功倍的成效。

應用目的是資料存留的最終依據

「隨著資料量越來越大，資料的取捨就成了目前資訊人員必須面對的掙扎。」陳聲鏗指出，雖然Big Data時代，的確能讓許多過往的「垃圾」翻身成「黃金」，但是很多東西，起碼在當前的科技發展、眼下的使用目的裡，就只是個無用的「垃圾」，面對儲存空間有限的現實，勢必得做出存留的決斷。面對這樣的抉擇，陳聲鏗建議，一切還是回歸到「要解決什麼問題」這個目的去思考，保存當下及中短期派得上用場的資料，放棄用不著的部分。

Big Data的個資應用

不過，在Big Data趨勢下，資料的來源、搜集方式、應用日趨千變萬化，對社會大眾來說，最大的疑慮恐怕是個資暴露的問題。對此，陳聲鏗表示，將資料「去識別化」，並且賦予不可回溯的設計，是資料應用於公共服務上的一種有效保護方式，例如遠通電收在自己的資料平台上，就設計讓能對應到特定車輛車籍資料與行駛路徑的eTag代碼進入後，會轉換成一組無法對應且不可回推關聯性的數據，確保有心人士無法利用這些數據連結出任何個資。然而，這樣的保護機制卻不適用於商業行銷，這個當前Big Data最被廣泛應用的領域。

「商業行銷的數據一定要可以對應，因為他是要去分析顧客行為，了解每個客戶的愛好和需要，針對這樣的偏好、需求去行銷，所以他一定要能夠標定出這些資訊是屬於誰的，分析出來才有意義。」陳聲鏗提醒，在這方面的應用上，是不可能「去識別化」的，他進一步強調，保障個資並不是要把所有資料「去識別化」，因為在許多用途上，「識別化」是必須的，重要的是如何確保個資的取得經過授權，並且僅用於授權時講定的目的。

對於時下許多個資授權徵詢，尤其是在APP和社群網站，往往藏在冗長的聲明、注意事項或「我同意」裡，陳聲鏗一方面提醒社會大眾應盡量耐著性子把它們讀完，以搞清楚自己要讓渡出甚麼個資；另一方面他也建議廠商應盡量讓個資授權徵詢以更簡單明瞭的方式呈現，把需要的個資、限定應用範圍講清楚，必要時提供優惠作為交換，最重要的是，還必須提供當事人安全退出的機制。

陳聲鏗強調，開誠佈公而不是矇騙過關的態度，才能讓Big Data的商業應用擁有比較健康的環境。正如知名數據科學家、「Big Data之父」維克托· 邁爾-舍恩伯格（Viktor Mayer-Schönberger）教授所強調的：「沒有信任也就沒有大數據。」當大眾對於資料應用不能信任，Big Data的功效就無法徹底發揮；當大眾對於資料應用的擔憂持續擴大，Big Data的發展必將面臨重重阻礙。或許，當應用與保護取得平衡的一天，屬於Big Data的時代才會真正的到來。