論述 》管理 · 資訊

大數據在主計資訊的發展與聯想

第五波的資訊浪潮「大數據」襲來，大數據能預知未來嗎？我們該如何使用？我們能掌握大數據嗎？還是大數據在操控著我們？

魏明達（行政院主計總處主計資訊處研究員）

壹、前言

隨著數位時代來臨，行政院毛院長自接任閣揆以來，宣示以資料開放（Open Data）、大數據（Big Data）與群眾外包（Crowd-Sourcing）的「網路溝通與優化施政三支箭」做爲科技內閣施政方針，期使政府能有效運用網路等新興科技，加強政府與民間資訊交流，改善各公部門施政方向，讓人民有感於政府的施政改善。當然行政院主計總處（以下簡稱總處）也在這波趨勢下，在開放資料上成績顯著，亦開始研析大數據的技術與應用。

貳、大數據的特性

大數據具有 3V（大量性 Volume、時效性 Velocity、多樣性 Variety）、或 4V（加眞實性 Veracity 或價值性 Value）特徵，難以用一般的技術來收集、處理與管理，於是產生了新一代的處理邏輯與決策方法。以往資訊電腦處理的資料有限且全查作業之高成本及耗時，遂發展以抽樣資料來推估母體之統計技術，這樣的統計技術存在抽樣誤差（抽樣數量與代表性）及資料誤差（受訪者未眞實回答）而影響結果，決策也可能因而產生偏頗的施政方案。科技的發展促使大數據這類科技運用，能取得大量性與即時性的資料，使資料更接近母體，當樣本趨近於母體時，我們便能得到更爲眞實的完整資訊，那麼決策便能精準地改善施政結果。

大數據應用的成功典範，Google 應該是其中翹楚，Google 是目前蒐集資料的巨擘，從收集的大數據中分析，

並發現新的應用。其中成功的例子之一便是 Google 街車，他不僅僅只是拍攝街景、更收集了許多的路線及座標資料、甚至是路上經過的 Wi-Fi 信號，而今，Google 正利用大數據研發無人駕駛的汽車。

大數據應用的決策模型（圖1）第一步是不斷地蒐集資料、第二步是分析資料找出其中資料的組合邏輯、接著第三步是發現新的應用。並非是傳統的應用模式，先想好需要什麼應用，然後再去蒐集資料。故未來決策模式將逐漸以完整的資料來進行決策（即資料驅動決策，Data driven decision），取代運用有限的資訊來進行決策（即資訊驅動決策，Information driven decision）。

參、總處的資料是大數據？

總處所擁有的資料究竟算一般數據還是大數據？我們把總處的大部份資料概分三類：歲計、會計、統計，而統計又包含了公務統計、抽查、普查資料。若以大數據 3V 的特質來看，歲計、會計不具有多樣性，抽樣調查資料與母體的量相對來說並不算大量；就時效性而言，抽樣資料也並非每天更新，統計所用的數據往往是一個月前的資料，普查資料也是 5 年普查（農林漁牧業普查、工業及服務業普查）或 10 年普查（人口及住宅普查）一次，換句話說，總處的數據資料顆粒若能再細緻、時間再即時些，這些資料數據再輔以大數據 3V 特性，必能使決策資訊更貼近民意。

歲計、會計、統計資料系統獨立運作，但這三種資料意義存在著密切的關聯性，而統計資料更具有非常重要的參考價值與指標特性。一旦這三種資料系統能整合成資料銀行或是總處大數據庫，配合資料科學團隊的分析，那麼讓人興奮的應用將隨之而來（下頁圖2）。

肆、啓動大數據列車

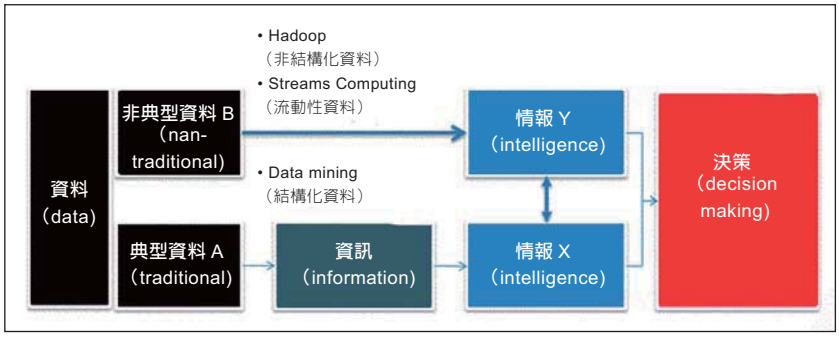
大數據應用成功重要因素，包括資料面、技術面、應用議題及資料科學人才培育等，因此建議以下幾點作法：

一、資料面

（一）建立資料匯集、整合與保存機制

資料驅動決策模型需具有完整且即時的資料，才能產生精確的情報資訊。前述大數據應用模型必須先仰賴大量的資料，新的應用才能接踵而至，因此資料的基礎

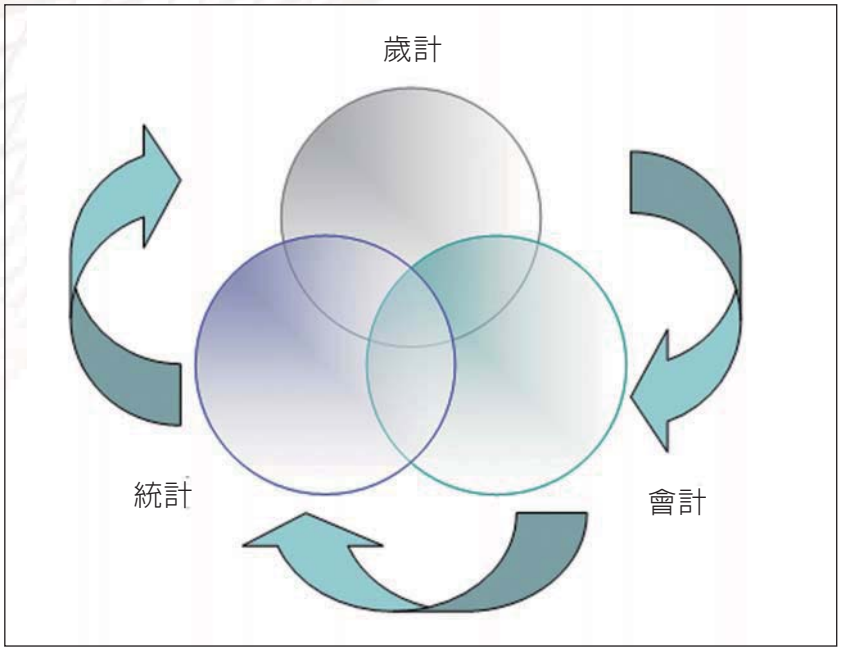
圖 1　大數據從資料到決策的流程



資料來源：科技政策研究與資訊中心－科技產業資訊室整理。

論述 》管理 · 資訊

圖 2　主計三連環是相輔相成，缺一不可



資料來源：作者自行整理。

建設尤爲重要，總處刻正著手規劃的民眾數據查詢平台便是爲了整合跨機關的主計資料、提供資料永久保存、確保資訊安全（含個資保護）、降低技術門檻，且將成爲一個主計資料銀行（data bank），有了這個平台後，總處的大數據高樓才會有平穩的地基。

大數據的應用有賴於數據分析師在各個資料集中找出可能的關係與應用，越是完整的資料集便能帶來有效且全面的應用。目前要取得總處各項資料結構與定義並不方便，沒有統一窗口；在實際應用上更遭遇資料完整、資料格式及機敏資料限制等種種問題，爲解決以上問題，本平台建置期使實際應用時資料的取得將會更爲便利，並加速應用開發的時程。

（二）強化資料時效

有了即時有效的資料才能產出讓人民當下有感的資訊與決策。以往資訊情報落後的情況，縱使政府相關單位得到資訊情報，快速召集相關部會研擬對策，但情報不夠即時，仍無法立即下達有效的決策幫人民解決當下所遭遇的困境。抽樣調查可能會帶來抽樣誤差或是資料誤差，若非母體數據，很難達到精準無誤。而抽樣調查耗時費力，若能借群眾力量來蒐集資料，或透過科技通訊的運用，主動由群眾的智慧裝置提供最新資料，那麼當群眾數量大到一定規模時，數據便會符合現況且趨近於母體，決策便會更爲精準。

（三）結合其他公部門資料

單一特質的資料集合只能得到一種事實的陳列，多種單一特質數據的結合，才能成就大數據中的黃金價值。因此，跨機關的資料整合就極爲重要。例如：人口普查的數據往往是公部門建設或預算分配的重要參考資料之一，戶政單位每月也有掌握人口的設籍資料，值此大數據時代，只要結合其他

公部門的資料，這份人口資料將會成爲前所未有的重要黃金數據。舉凡交通、建設、經濟、教育、社會福利甚至治安方面都能靠這份黃金數據來加以改善。

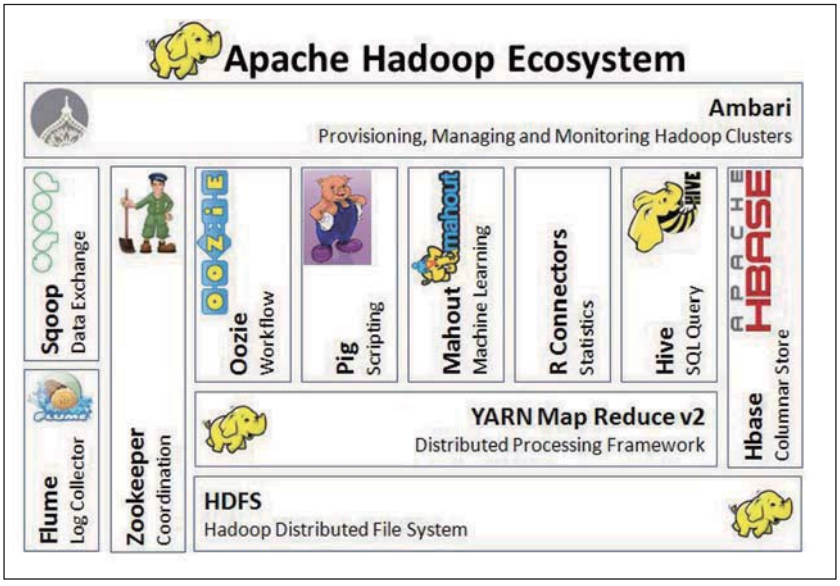
二、技術面

大數據若依傳統的關聯式資料庫儲存並無法確實地運用，需要靠新技術來儲存大量、多樣且快速產生的資料，於是站在 Google 肩膀上的 Apache Hadoop 迅速竄紅，它提供了相對可靠、平價、架構簡單的儲存與運算平台，Hadoop 雲端平台的兩大核心：分散式檔案系統（HDFS）與平行運算架構（Map-Reduce）恰好迎合了大數據儲存與分析運算的兩大主軸（圖 3）。

其他因大數據潮流而發展的技術有：非傳統關聯式資料庫管理系統（NoSQL database）、機器學習（Machine Learning）、資料探勘（Data Mining）、資料視覺化（Data Visualization）、叢集運算引擎（Apache Spark）等等。

然而這些科技技術的應用技術門檻較高，無法普及到業務層級同仁，除藉由教育訓練來瞭解新技術，培養技術專才、加強整合應用能力外，也要透過技術包裝，將大數據應用的技術簡化，業務同仁或主管只要透過幾個簡單的步驟，即能取得決策資料，以達到實際運用的目標。

圖 3　由 Apache Hadoop 所衍生的技術生態系統



資料來源：Apache Hadoop document.

三、應用議題

空有龐大的資料而無應用並無法讓大數據成功，總處計劃透過學研界在大數據之研究創意與能量，對主計資料進行深度分析，以產生對施政更有參考價值之研究成果。底下舉兩個可能的應用例子：

（一）借群眾力量應用於價格調查

價格是民眾非常關心的資料，但要即時收集到消費數據並不容易。如果今天我們提供一個免費的記帳 APP 軟體給民眾，使用的代價是必須傳回無記名的價格資料，那我們便可每天不費吹灰之力蒐集到當日或近日的價格資料，有了這些數據，

我們便可統計出最新的民生物品價格趨勢，並同時在記帳 APP 上顯示出來，也讓使用 APP 的民眾可即刻地瞭解自己購買物品的價格帶。

（二）主計資料結合交通資料

根據總處的人口普查資料與跨縣市工作遷徙人口資料，我們可以概估當連續假日發生時，大部分的外地工作人口將會回到設籍地，再配合交通部 ETC 的歷史數據，便可更準確地估算出各地路段可能的返鄉交通情形。若再能結合觀光局或是民間旅遊地點的歷史入園票數，可以推估出遊及返回工作崗位的各地交通流量，民眾便可透過這些情報改變交通計劃。

當然以上的發想還需透過研究以驗證其可行性，但透過上例，可以瞭解未來業務目標，可以透過更多元多樣的技術及資料來達成。

四、資料科學人才培育

大數據的應用發展非一蹴可幾，也絕對不單只是資訊人的事情，培養大數據相關人才刻不容緩。資料需要時間累積到一定的規模，才具有挖掘的意義，而在大數據中要挖掘出有用的關聯與運用，更需要藉由資料科學家、資料分析師靠專業與經驗找出珍貴的邏輯關聯性；或是靠與資料相關的業務承辦人才能在熟稔的大量資料中找出可能的應用需求。主計資料的應用發展需要集思廣益；需要各機關同仁的共同付出與合作才能不斷地成長茁壯；更需要業務單位的創意、精益求精才能產出讓人民更有感的應用。

未來可廣納各方專家大數據經驗、積極培養開放資料與大數據人員，規劃辦理大數據系列研討訓練，建立主計同仁大數據基本觀念、激發創意，以共同開創主計大數據相關應用與發展。

伍、結語

主計三連環的大數據發展正逐步展開，但目前尚面臨跨部會資料取得困難、處理機敏資料去識別化等挑戰。若有朝一日數位資料分享有充分法源依據，政府各公部門資訊連線發達，取得母體資料輕而易舉，那我們將重新思考，在大數據的相關資料分析參考下，不僅能做到合理地審核預決算及會計資料，更能即時且無抽樣誤差的統計，還能找到更多讓人民有感的數據資訊。

最後以「聖人見微以知萌，見端以知末，故見象箸而怖，知天下不足也。」這句話來結尾，有了大數據以後，我們可以不是聖人也能做到見微知著，如果政府能靠大數據的解析而洞察機先、防患於未然，這不啻是政府的德政，亦是人民的福祉。❖