出國報告（出國類別：研究）

巨量資料分析應用之研究

服務機關：臺灣土地銀行
姓名職稱：洪妤如高級辦事員
派赴國家：美國
出國期間：104年10月3日至103年10月23日
報告日期：104年12月24日

摘　　要

考量巨量資料浪潮下的衝擊，隨著資料量急速成長及資料類型漸趨複雜，為因應數位金融環境時代來臨與巨量資料改變金融業生態，本次出國參訪研習，汲取各企業於巨量資料分析之技術發展趨勢，觀摩巨量資料資料分析方法、工具與平台，並評估本行現況後，建議本行規劃導入巨量資料平台、導入探索平台或視覺化分析工具，及成立資料科學家團隊。

目　　錄

壹、 研究目的

巨量資料時代來臨，近年來國內外各領域的企業紛紛加入尋找巨量資料價值的行列，本行的資料應用雖然仍以資料倉儲的結構化資料為主，資料量也未達千兆位元組(PETABYTES)，但隨著資料量急速成長及資料類型漸趨複雜，為因應數位金融環境時代來臨與巨量資料改變金融業生態，確實需開始考量導入巨量資料應用。

本次出國參訪研習，汲取各企業於巨量資料（或稱大數據、Big Data）分析之技術發展趨勢（如巨量資料平台、資料處理技術）、商業創新應用（如資料分析）、資料創新與數據驅動（如資料科學及資料科學家）等資訊，觀摩巨量資料資料分析工具、方法與平台架構、學習各企業應用巨量資料分析提昇企業效率與營收之成功案例，擷取適用於本行現況之架構技術與分析工具，以幫助本行未來規劃建置巨量資料相關應用之參考。

貳、 研習內容

本次赴美從事「巨量資料分析應用之研究」，研習內容為參訪 Pivotal 公司的資料科學研究 LAB，以及參加由 Microsoft、Teradata 公司舉行 Big Data 相關之研討會，議程涵蓋了「商業創新」、「資料創新」、「資料科學與進階分析」、「數據驅動」、「用戶體驗與視覺化應用」、「巨量資料分析工具」等。

一、 Strata+Hadoop World Conference 2015

（一）研討會參加心得

本次參加研討會為 Microsoft 推薦且贊助舉行之 Strata+Hadoop World 2015 Conference，是由 O'REILLY 及 CLOUDERA 公司主辦，自 2011 年起已連續舉行五年，每年舉行 1~3 場不等，舉辦地點跨足美東(紐約)、美西(聖荷西)、歐洲(倫敦)及亞洲(新加坡)，是近年來巨量資料（Big Data）發展的年度大型研討會之一，2014 年參與人數已超過 5000 人，今年的研討會估算更有超過七千人參與，參與的對象多數由資料分析師、資料科學家、業務經理、工程師、高階主管…等組成，藉由參與研討會互相交流，並且從議程安排的主題觀察巨量資料發展趨勢與技術走向。本次研討會於紐約曼哈頓的會議中心 Javis Center（全名 Jacob K.Javits Convention Center）舉行，建築外觀為漸層高度的玻璃帷幕，佔地十分廣闊，走進會展中心時更被採光良好且挑高的寬闊空間震撼！

圖 2-1：Javis Center

（資料來源：自行拍攝）

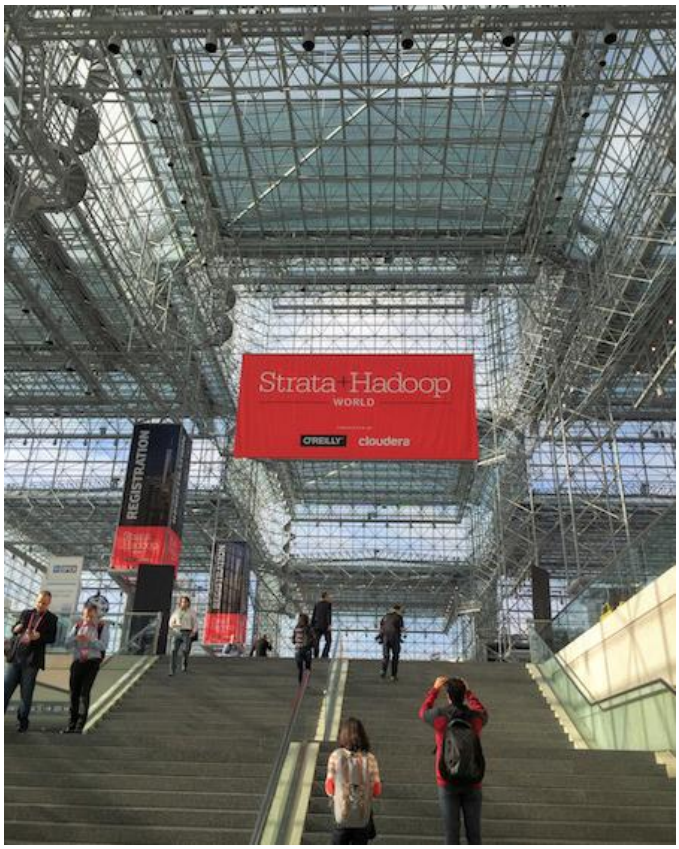

圖 2-2：Javis Center 內部

（資料來源：自行拍攝）

會展中心最北端的演展廳 Javits Center North 做為研討會的 KEYNOTES 大型演說場地，演說者都是相關領域中的佼佼者、專家、大老闆，精彩的演說搭配巨型投影布幕及專業的影音設備，能有幸參與其中聆聽演說是一件非常難能可貴的經驗。

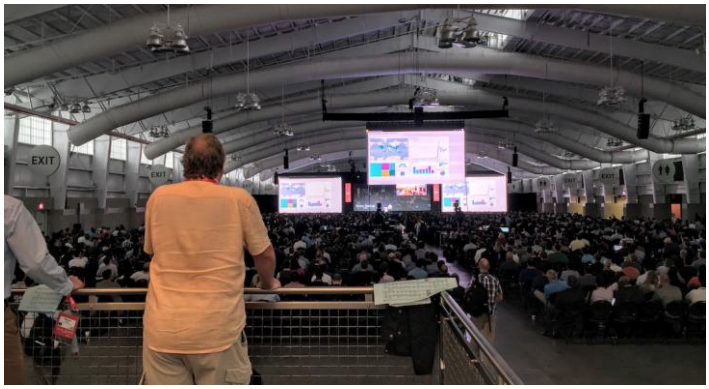

圖 2-3：Javis Center North

（資料來源：自行拍攝）

午膳是在會展中心的 3A 與 3B，採取自助餐式用餐，座位區特地規劃成大圓桌，讓參與者在用餐的同時也有機會和其他參與者交流分享心得、拓展人脈。

圖 2-4：Javis Center 會議廳

（資料來源：自行拍攝）

以下補充本研討會相關主辦與贊助企業簡介：

1. Microsoft

微軟為本行合作廠商，亦於本次美國參訪引薦參與 Strata+Hadoop World 研討會，以下介紹摘錄自為台灣微軟官網。

微軟是全球科技產業的領航者，提供全球領先的軟體、服務、設備和解決方案，自 1975 年成立以來，一直致力於幫助個人和企業用戶，全面發揮科技潛能，實現夢想。四十多年來，微軟一直專注於技術創新與變革，透過卓越的軟體、設備和服務，能夠幫助用戶提昇生活和工作生產力，讓數億的使用者真正受益於科技。微軟公司總部位於美國華盛頓州 Redmond，在全世界超過 190 個國家和地區設有分公司或是分支機構，擁有超過 125,000 名員工。微軟執行長 Satya Nadella 在上任後提出了新願景：「在這個行動優先、雲端至上的世界裡，微軟的核心競爭力，就是成為一個提供生產力與平台的公司，幫助在這個地球上的每一個人到每一個組織，都能貢獻更多、成就更大。」台灣微軟成立於 1989 年，身為政府、學校與企業最值得信賴的夥伴，始終致力於幫助台灣提升創新力、競爭力以及促進經濟繁榮，共同來亮點台灣！台灣微軟公司位於台北市松仁路 7 號 8 樓，並設有中部及南部辦公室。正職員工人數約為 400 人。

2. O'REILLY

歐萊禮媒體（O'Reilly Media, Inc.）是以出版電腦資訊書籍聞名於世的美國公司，也是全球著名的創新科技資訊提供者，由提姆·歐萊禮（Tim O'Reilly）創立於 1978 年。該公司既是出版開放原始碼書籍的先驅之一，也常承辦許多開放源始碼社群的研討會議，業界領袖和電腦玩家都是透過 O'Reilly 的書籍、研討會和網站向全球展示最新的電腦科技。出版圖書的選題範圍現在也擴大到數學、心理學、旅遊、日常生活和職業發展等，並創立了 Make 雜誌，從而成為 DIY 革命的主要先鋒。從最暢銷的《Whole Internet User's Guide and Catalog》（被紐約公共圖書館評論為二十世紀最重要的 50 本書之一）到 GNN（全球第一個商業網站：全球網路導航器（Global Network Navigator）），再到 WebSite（第一個桌面 PC 的 Web 伺服器軟體），O'Reilly Meida,Inc.一直處於 Internet 發展的最前端，關注真正重要的技術趨勢——透過放大那些「細微的訊號」來刺激社會對新科技的應用，無論是透過書籍出版、線上服務或者面授課程，每一項 O'Reilly 的產品都反映了公司不可動搖的理念——訊息是激發創新的力量。

3. CLOUDERA

Cloudera（Cloudera, Inc.），於 2008 年正式成立的美國軟體公司，是全球著名的大數據業務處理公司之一。Cloudera 向企業客戶提供 Apache Hadoop（Cloudera Distribution including Apache Hadoop，CDH）的軟體、支援、服務以及培訓，利用 Hadoop 這一開源技術幫助公司搭建他們的大數據系統，Hadoop 可使用一些價格低廉的硬體就完成大量的數據分析，所以非常受大小企業歡迎。在 Hadoop 高層級 Apache 的分散式儲存，以及 open data 分析技術運用方面，Cloudera 在業界都遙遙領先。Cloudera 的創始團隊包括前 Google、Facebook、Yahoo 以及 Oracle 的員工和高級工程師。Cloudera 的客戶中有很多知名公司，如 AOL、哥倫比亞廣播公司、eBay、Expedia、摩根大通、Monsanto、諾基亞、RIM 和迪士尼等。

（二）研討會議程與重點

Strata+Hadoop World 2015 年研討會的討論重點為「善用資料」，認為能掌握如何使用資料的企業才能掌握未來。研討會集結了在策略運用、科學領域及產業中最優秀的人才，在會中提供巨量資料發展的現況發表、探索與交流，主要議程包含：數據驅動型的商業未來（Data-driven Business）、資料科學及進階分析（Data Science & Advanced Analytics）、資料創新（Data Innovations）、用戶體驗設計與視覺化應用、Hadoop 使用案例、Hadoop 核心與技術發展、Spark 技術與應用、開放資料的法律及道德、物聯網與即時應用（IoT & Real-time）、企業組織變革、資料安全。

更進一步觀察研討會議程內容可發現幾個重點：Hadoop 應用已趨向穩定與成熟，有許多分享案例；Spark 在今年是繼 Hadoop 之後新興起的熱門議題。（Hadoop 與 Spark 皆為巨量資料應用的技術）；想導入巨量資料應用，要先認識資料湖泊（Data Lake）的架構概念；巨量資料發展帶動資料科學的興起與資料科學家的重要性，而資料分析更是巨量資料創造價值的具體實現；物聯網與即時應用時代即將來臨，相關主題將是往後幾年的熱門趨勢。上述有關「Hadoop」、「Spark」、「資料湖泊」、「資料科學」、「資料科學家」與「資料分析」議題，在下一章（參、研習心得）中分別有更詳細的內容介紹。

二、 Pivotal LAB 參訪

（一）LAB 參訪心得

本次參訪的 Pivotal LAB 位於加州帕羅奧圖(Palo Alto)，在著名的史丹佛大學附近，LAB 外觀是兩層樓高的建築，內部有寬敞的員工餐廳及裝滿飲料、食物的茶水間、健身房等設施，根據 Pivotal LAB 說法，認為創新的企業必然需要像這樣的工作環境，有助於提升員工的創新能量，實際走訪在 LAB 中，真的有感受到環境舒適、員工有活力的工作情緒氛圍。

圖 2-5：Pivotal Lab 門口與內部

（資料來源：自行拍攝、Pivotal）

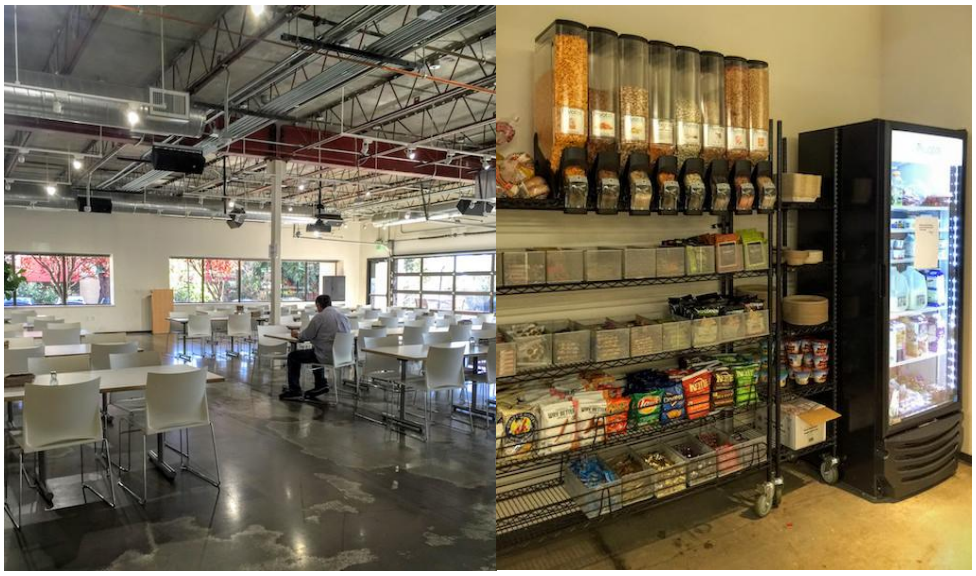

圖 2-6：Pivotal Lab 員工休息室

（資料來源：自行拍攝）

Pivotal 簡介

位於舊金山灣區西南方的 Pivotal Labs 創立於 1989 年，為私有股權的敏捷軟體(agile software)之開發服務與工具供應商，涵蓋網路、行動裝置、巨量資料和雲端服務領域的世界級工程人才，並提供領先業界且在全球超過 24 萬用戶採用的軟體開發工具— Pivotal Tracker 。Pivotal 的客戶包括 Twitter 、 Best Buy 、 Groupon 、 Salesforce.com 、 EMI 、 Urban Dictionary 、 Linden Lab 、 Task Rabbit 以及安妮可希基金會 (The Annie E. Casey Foundation) 等。2012 年 4 月 EMC 公司宣布併購 Pivotal Labs，為 EMC 帶來高度差異化以及最佳的軟體開發方法學，進一步強化了 EMC 原本就相當堅強的產品與服

務陣容，讓企業能儲存、分析、以及運用資料量多到超出傳統 IT 基礎架構所能負擔的「巨量資料」。2012 年初，EMC 推出 Greenplum 統一分析平台 (Unified Analytics Platform；UAP) ，首度提供橫向擴充的基礎架構，用來分析結構化與非結構化的資料。 EMC 並宣布全面供應 Greenplum Chorus，針對資料科學團隊推出類似 Facebook 的社群協作工具，用來合力開發資料集，並確保能快速為企業提供有價值的資訊。在納入 Pivotal Labs 後， EMC 現在能支援 Greenplum Chorus 的資料集，並讓客戶能運用像是 Ruby on Rails 等現代的程式開發環境，快速建立巨量資料應用。如今 Pivotal Labs 加入後，這家敏捷式開發領域的領導者精通於快速疊代(iteration)、測試導向的研發、開放原始碼工具、以及現代開發框架等方面的技術，讓新創公司以及全球化企業客戶都有能力快速建立新應用與服務，善加運用巨量資料帶來的效益。2015 年 10 月，Dell 以 670 億美元買下 EMC，成了 IT 史上最大規模的併購案。Dell 併購 EMC 後，包括 EMC 旗下虛擬化平臺龍頭 VMware、資料分析平臺 Pivotal、融合式架構平臺 VCE、雲端供應商 Virtustream 和資安部門 RSA 團隊也將一併加入 Dell。

（二）簡報重點

本次參訪 Pivotal Lab，由擁有超過 10 年的資深軟體分析師，現為 Head of Data Science，Kaushik Das 接待，針對「Pivotal Data Science」議題進行簡報介紹，議題內容包含 Data Science 的重要性、Data Lake Architecture 及案例分享。節錄重點如下：

隨著 facebook、Twitter、Linkedin 等社群，及雲端及行動應用的興起，人們的生活越來越複雜，選擇也逐漸變多，選項太多甚至成為一件令人痛苦的事。對組織或企業而言也是如此，資料量與資料類型隨著各種科技應用發展，越來越龐大、複雜，而處理複雜資料的基本精神就是要用運更聰明的系統（Smart System），讓人們的生活變得更加簡單。人類本身就是一個 Smart System 的最佳代表，舉例來說，人在準備打擊棒球時，透過眼睛當感應器（Sensors）、人

腦當計算處理器分析球的軌跡，再告訴身體這個執行器（Actuators）要採取什麼打擊動作來擊中球棒。如圖（2-7），Smart System 就是整合了「感應器」、「計算處理器」及「執行器」的整合系統，在巨量資料的應用中，各種感應器收集到的資料，進入資料湖泊（Data Lake）中，透過資料科學幫助計算處理器做資料分析、建立模型，最後產出各種結果與新應用。有關「資料科學」與「資料湖泊」議題，在下一章（參、研習心得）中分別有更詳細的內容介紹。

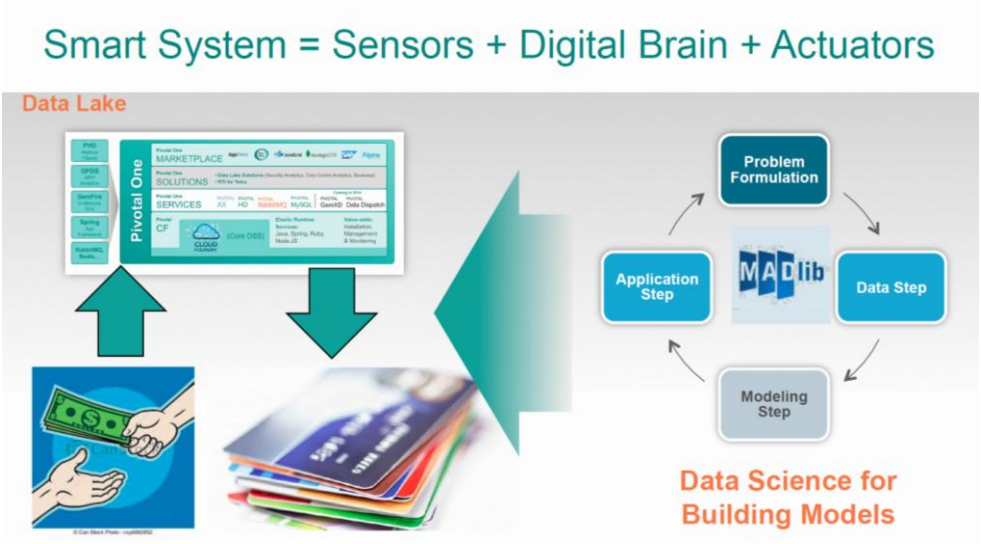

圖 2-7：Smart System 示意圖

（資料來源：Pivotal）

三、Teradata 2015 PARTNERS Conference & Expo

（一）研討會參加心得

本次參加研討會 Teradata 2015 PARTNERS Conference & Expo 為天睿公司主辦，是 Teradata 每年皆邀請全球 PARTNERS 共同參與的盛會，本年度大會於美國加州安那翰 Anaheim Convention Center 舉行，參加人數超過 5000 人，議題以巨量資料（Big Data）為主軸，細分 BUSINESS、MARKETING、TECHNOLOGY 及 VIRTUAL 主題類型，超過 200 個議程，參加的對象多為各領域企業行銷部門、資訊部門的高階主管，在研討會中可和世界各國人員交流、經驗分享。

圖 2-8：Anaheim Convention Center

（資料來源：自行拍攝）

圖 2-9：閉幕演講

（資料來源：自行拍攝）

圖 2-10：Gala Event Extravaganza

（資料來源：自行拍攝）

Teradata 簡介

Teradata(天睿)是全球領先的大數據分析和資料倉儲解決方案公司，專注於整合資料倉儲、大數據分析和業務應用。天睿公司的產品和服務協助各種組織機構，進行資料整合並獲取有利於業務的資訊，從而做出最佳決策，提升競爭優勢。天睿公司成立於 1979 年，前身隸屬於 NCR 公司，自 2007 年 10 月起從 NCR 公司獨立，並由麥可・科勒擔任執行長兼總裁。天睿公司總部設於俄亥俄州邁阿密斯堡。天睿公司資料倉儲產品使用的技術被稱作「無共享（shared nothing）」架構，各個伺服器之間擁有獨立記憶體和處理能力，增加伺服器與節點即可增加儲存的資料量，並由資料庫軟體集中管理各伺服器間的承載負荷量。

（二）研討會議程與重點

Teradata 2015 PARTNERS 研討會以一「Breaking Big」為標語，強調巨量資料的應用現況已不再執著於「大」，而應注重「資料」本身的應用。主題類型概分為「Business」與「Technical」兩項，主要議程包含：巨量資料 vs 商業智慧、數據驅動型的商業未來（Data-driven Business）、感應器分析（Sensor Analytics）、全文分析（Text analytics）、網路安全、資料治理（Data Governance）、資料科學驅動組織（Data Science Driven Organizations）、Hadoop 發展、Teradata Technology、Aster Technology、生態系統架構（Ecosystem Architechture）、巨量資料開源技術（Open Source Big Data Technology）。研討會中有許多企業分享資料應用經驗與成果，擷取部份案例與內容如下：

1. eBay 分享了橋接 Teradata 與 Hadoop 的案例，將 Teradata 資料匯出至 Hadoop 分散式檔案系統（Hadoop Distributed File System，HDFS），以及將 HDFS 資料導入至 Teradata，產出的應用報表是採用 Tableau 軟體，橋接成功的指導原則是需要「容易使用」且「效能要好」。

2. 西班牙電信集團(Telefonica Group)旗下 Vivo 公司（主要據點在南美洲及

歐洲）的 BI 資深經理分享了該公司利用 Teradata 統一資料架構平台（Unified Data Architecture）及 SAS 統計軟體，整合了公司現有的兩座資料倉儲與超過 11 個資料超市系統，萃取資料倉儲中的消費資料、產品所有權、顧客註冊等資訊，執行計算、過濾適合的模型，依照業務規則及客戶使用習慣，替每一個客戶計算好最划算的報價表，大幅縮短行銷團隊的決策時間，將原本需要 30-40 天才能訂出的決策縮短為 1 天。

3. 西南航空（Southwest Airlines）公司使用企業資料倉儲支援各種部門及多樣化的任務，為了因應業務成長，他們認為需要更完整的瞭解客戶。在客戶分析時面臨的問題是無法區隔客戶、找出客戶價值及提供個別客群適合的產品。為了改善此問題，Teradata 提出以客戶為中心的概念性驗證（POC，Proof Of Concept）解決方案，導入 Teradata 旅宿業資料模型（Teradata Travel and Hospitality Industry Data Model），強化對客戶的瞭解與洞察。

CDO（Chief Data Officers）

由 Teradata 資深產業顧問主講的議程，以「數據長」（Chief Data Officers，CDO）這個職位角色的重要性為主題，認為數據時代的來臨改變了組織開展業務的方式，也因此誕生這個職位。CDO 的重要性在於當企業準備導入巨量資料策略，轉型為數據驅動的企業時，幫助企業利用數據分析，得出一個能用事實佐證的企業抉擇，並且能即時擷取出有用的數據。2012 年底，Gartner 調查 CDO 的數量仍低於 50 位，然而 2015 年 9 月，美國市場調查機構（Forrester Research）統計調查千人以上規模的企業，已有將近 48%的企業有任用 CDO 一職。統計各產業任用 CDO 職務的結果顯示，有 16%屬於金融業，是比例最高的產業，其次為電信業 14%、科技業 14%、醫療保健產業 11%、消費服務業 10%、公共產業 9%、媒體產業 9%等。當資料量越大、資料複雜度越高時，資料價值的密度是逐漸降低的，亦即要擷取出有用資訊的困難度是越高的，如圖（2-11），要如何從中創造業務價值，同時又得避免企業轉型應用時的痛苦，是 CDO 的一大挑戰。

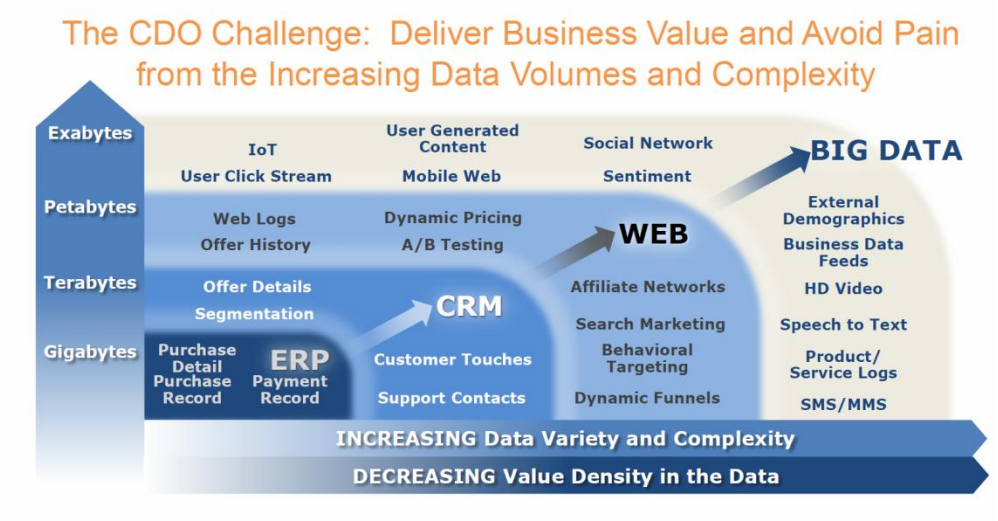

圖 2-11：The CDO Challenge

（資料來源：Teradata）

參、 研習心得

一、巨量資料定義

巨量資料，或稱 Big Data、海量資料、大數據，指的是在 3V(Volume、Variety、Velocity)方面難以管理的大量資料，以及為了儲存、處理與分析這些資料的技術；此外，更包括分析這些資料並能夠從中萃取出有用資訊或洞見的人才與組織之全盤概念。所謂為了儲存、處理與分析資料的技術，指的是大規模資料分散式處理架構的 Hadoop、擴充性優異的 NoSQL 資料庫，以及機器學習與統計分析等。而分析資料、並從中萃取出有用資訊或洞見的人才與組織，指的則是目前炙手可熱的「資料科學家」（Data Scientist）與能夠有效運用巨量資料的組織型態等。

那麼企業需要管理的資料有哪些分類呢？一般而言，企業的資料類型包含：(1)結構化資料，指放在 RDB (SQL DB) 或 資料倉儲中的資料；(2)半結構化資料：如各種 Log 檔、XML 檔、CSV 檔，一般可統稱為 Machine Data；(3)非結構化資料：如圖檔、語音檔、影像檔、PDF 檔、Office 檔案、電子郵件、網頁等資料。這些不同結構的資料，一般在企業中會搭配各種應用系統來使用，如 ERP、CRM、SCM、Reporting 之於結構化資料；Logging 日誌管理系統之於半結構化資料；E-Mail、Web 之於非結構化資料。有一個調查顯示，結構化資料，只佔企業內全部資料的 15%，其餘的 85%，則為半結構化與非結構化資料。

二、巨量資料平台

巨量資料的處理採用平行架構或分散式架構來加強系統的擴充性已成為必然，目前市場上有兩大主流方向：一是 MPP 資料庫 (即 Massively Parallel Processing，大規模平行資料庫)為首的平行關聯式資料庫陣營，可以處理與分析結構化資料，如 RDB、EDW；另一個是以 Hadoop / MapReduce 為首的分散式 NoSQL 陣營，可以處理非結構化與半結構化資料，藉由辨識、解析非結構

化資料，或是彙整、清洗、排序、關聯、轉換半結構化資料，最終產出高品質的半結構化或結構化資料，以提供後續分析或其他應用；除此之外，近年來新興發展中，企圖取代 Hadoop 的 In-memory 分散式運算技術—SPARK 開源軟體也值得注意，以下分別簡介：

（一） MPP 架構

在目前大多數資量倉儲應用中，大部分使用大規模平行處理(Massively Parallel Processing，MPP)的架構，通常被稱為 MPP 資料庫。MPP 資料庫又稱為無共用(Shared-nothing)資料庫。MPP 架構可以有效加強查詢效率及平台的可擴充性，在 MPP 電腦系統架構裡，每一個含 CPU 與本地記憶體(Local memory)的子系統皆為各自獨立(Shared-nothing)，整個系統也是由單一作業系統管理。CPU 與 CPU 之間只能靠訊息傳遞(Message Passing)的方式來溝通，電腦系統可以連結的 CPU 個數相當多，其優點是整個系統擴充性非常高。

本行選擇以使用 SQL Language query 進行平行運算處理的 MPP 架構做為長期發展的起點，於 103 年將資料倉儲導入 MPP 架構。

（二） Hadoop 技術

Hadoop 是由 Google 發表的 Map Reduce 及 Google File System 等文章提出的概念，由 Apache 軟體基金會實作而 成，所以跟 Google 內部使用的雲端運算架構相似。Hadoop 是一個分散式運算的標準，它可以將很大量的資料，由主電腦分解成小部分的資料，分別傳給各個節點電腦進行運算，而節點電腦再將運算完的結果傳回主電腦進行結合，此架構多用於新形態、半結構、非結構性的資料應用。Hadoop 軟體平台核心包含兩大部分：分散式檔案系統 HDFS 與平行運算框架 Map/Reduce。Hadoop 軟體平台以業界公認的 Apache Hadoop 開源版本為標準，不同廠商可以取其為核心，添加更多的開源軟體模組，或是自行開發各種的管理功能，成為自己的發行版本 (Hadoop Distribution)，如 Cloudera

的 CDH、Hortonworks 的 HDP。而相關硬體，如天睿公司 Teradata Appliance for Hadoop、台灣意圖公司的 Etu Appliance，皆是整合 Hadoop 軟體平台與硬體，方便快速部署與管理的一體機。

Hadoop 使用的 NoSQL（Not Only SQL 的簡稱）資料庫是相較於傳統 SQL 關聯式資料庫的一種資料庫類別的通稱，開源專案(Open Source Project)如 HBase、MongoDB、Cassandra、Neo4j、CouchDB 等皆屬 NoSQL 資料庫。這類產品大部分都可以在 Hadoop 平台上實行，以應付大規模線上同時查詢的需求。既然是資料庫，還是一樣有結構化的 Table Schema，但差異是兩兩 Table 之間，不做關聯；也沒有多次存取的資料交易特性；且存在其中的，可以是非結構化資料。

有趣的是，Hadoop 陣營與非 Hadoop 陣營，並非是兩個平行不相往來的世界，隨著 Hadoop 漸獲企業的重視，成為巨量資料的主流技術平台，有越來越多的非 Hadoop 陣營產品，廠商也自己或透過第三方，搭起了通往 Hadoop 的橋樑，可以存取放在 Hadoop 陣營中的原始資料或處理過後的資料。最常見的橋樑為連接 Hive（建構於 Hadoop 之上的資料倉儲框架）、HDFS (分散式檔案系統)、或 HBase (一種 NoSQL 資料庫) ，尤其是前兩者。

（三） SPARK 技術

Spark 概念為 In-memory 分散式運算開源軟體平台，這個眾所矚目的新世代運算技術架構最初是由加州大學柏克萊分校 AMPLab 專案的研究團隊，於 2010 年以 Scala 語言開發完成。2013 年加入 Apache 基金會後，於 2014 年成為 Apache TLP (Top-Level Project)，算是一個非常年輕，成長很快的開源專案。Spark 為一個用於擴充資料處理的叢集運算引擎，不需要 Hadoop 即可執行於一般商用硬體環境，並具有容錯能力，後來更發展成為一個開源巨量資料分析工具，可用來彌補 Hadoop 即時分析的不足。Spark 因為使用了記憶體內運算（In-Memory Computing）技術，能在資料尚未寫入硬碟時，就在記憶體內分析

運算，甚至透過 Spark 的串流處理套件（Spark Streaming），也能即時處理串流資料。在進行資料處理上，Spark 也採用了映射（Map）和化簡（Reduce）兩個步驟，可以支援各種資料庫可用性群組（Database Availability Group，DAG）任務和快速資料分享的操作。隨著巨量資料技術不斷演進，不少企業如電子商務、零售業及半導體製造等企業，開始廣泛運用巨量資料分析工具，如 Hadoop 等，甚至，更逐漸從傳統批次或離線分析，轉而走向了採用像是 Spark 等接近即時（near real time）分析的巨量資料分析，做為更進階分析的加值應用。

根據 Apache Spark 官方的說明，Spark 在記憶體內執行程式的運算速度，可以做到比 Hadoop MapReduce 的運算速度，還快上 100 倍，即便是執行於硬碟時，Spark 也有達到 10 倍速度，今年在 Sort Benchmark Competition（資料排序基準競賽），僅以 23 分鐘就完成排序多達 100 TB 的資 料量，打破了此前由另一個巨量資料分析工具 Hadoop 保有 70 分鐘的世界記錄， 過程中只花了原本不到三分之一的時間，僅靠著部署 207 臺 Amazon EC2 i2.8xlarge 的虛擬機器，及配置了 6,624 顆虛擬核心處理器，平均每分鐘可以排序處理 4.27 個 TB 的資料量。

（四） 資料湖泊架構

過去企業從導入資料庫、資料倉儲、發展商業智慧，以面對不斷增長的資料以及資料的應用問題，隨著時間與科技發展，資料量持續增加、資料類型漸趨複雜，傳統架構將可能不敷使用，也促使大家思考新架構的產生，2011 年富比士雜誌在「Big Data Requires a Big, New Architecture」一文中提出了 Data Lake（資料湖泊）這個新架構。如今大多數企業已將巨量資料交付給採用 Hadoop 這類技術的資料湖泊，撰寫客製化程式碼並從資料中擷取出有價值的資訊。下表（3-1）整理了資料庫、資料倉儲、資料湖泊的不同特點：

功能	資料庫(Data Base)	資料倉儲(Data Warehouse)	資料湖泊(Data Lake)
設計目的	作為應用服務的資料儲存用。	作為企業決策分析資料儲存用。	作為企業的資料庫整合資料的儲存平台。
儲存內容	儲存於各應用服務的精細、短期間的資料，資料會持續變動。	儲存長期、歷史性且經過整理的企業決策分析資料。	儲存多樣性、高不確定的資料，作為企業資料分析素材。
資料擷取	具有資料新增、修改、刪除、查詢的功能，內容經常變動。	主要提供資料查詢與分析功能，內容較少變動。	主要提供資料快速 讀寫內容變動性大，且較為龐雜。
分析能力	提供短期資料查詢。	提供多構面的分析，及長期的資料分析。	提供資料的萃取、轉換讀取，以及資料分析的工作。

表 3-1：資料庫、資料倉儲、資料湖泊比較表

（資料來源：[9]）

資料湖泊有別於資料倉儲，資料倉儲的資料通常是品質較高，且是被預先處理過的資料。而資料湖泊架構的設計是可擷取大量的、各種類型的資料，作為資料素材(Data Material)的儲存管理，以利分析應用。正因為資料湖泊架構範圍更廣、在資料分析上擁有更多彈性，很適合做為企業導入巨量資料應用的架構藍圖，此行到美國參加的研討會上也發現有許多以資料湖泊為主題的分享與討論議程，探討此架構的概念與分享導入巨量資料的歷程。圖（3-1）為 Pivotal 提供的完整資料湖泊架構示意圖，由圖可看出資料湖泊資料包含了 MPP 資料庫、In-Memory 資料庫及 HDFS 分散式儲存資料庫，各種來源、各類型的資料，在資料湖泊經由不同方式的儲存、處理、淨化、管理後，有彈性的產出各種分析資料。

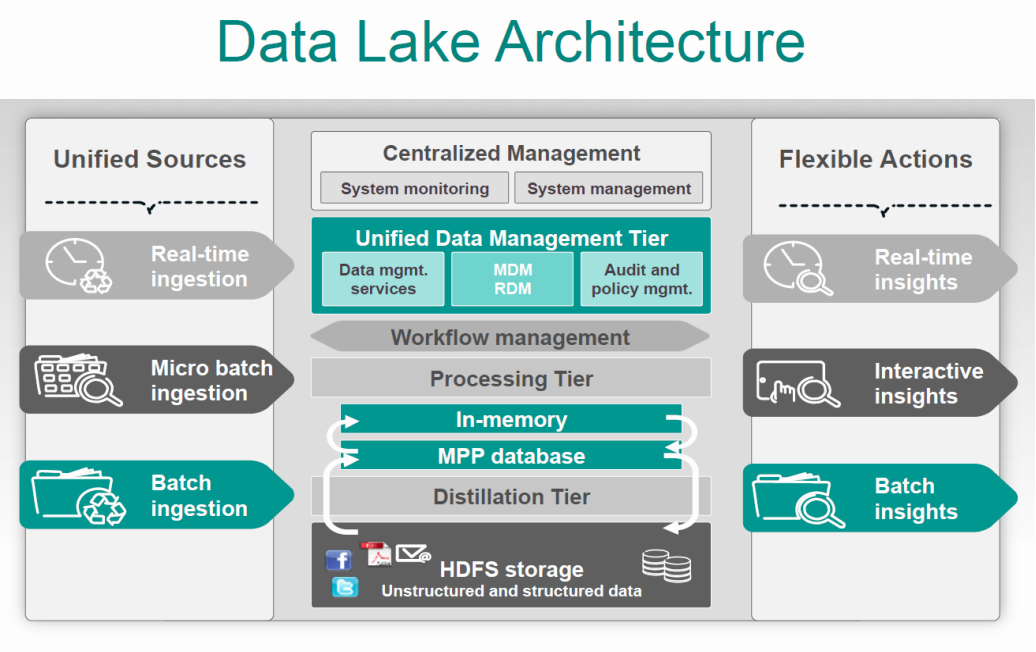

圖 3-1：資料湖泊架構

（資料來源：Pivotal Labs）

更進一步的解釋資料湖泊的發展過程，根據 2014 年 1 月富比士雜誌的「The Data Lake Dream」一文，提到資料湖泊的發展成熟度可分為四個階段，幫助企業瞭解發展資料湖泊發展到極致會達到的狀態，並對比目前的現況屬於哪個階段：

階段一「尚未導入 Hadoop」：此階段是目前企業普遍的資料架構，紅色圓圈代表應用系統(資料庫)，綠色方形代表資料倉儲，各個應用系統(資料庫)相互獨立，部分應用系統間有做資料交換，部分應用系統資料傳遞到資料倉儲進行分析與報表產出。如圖（3-2）。

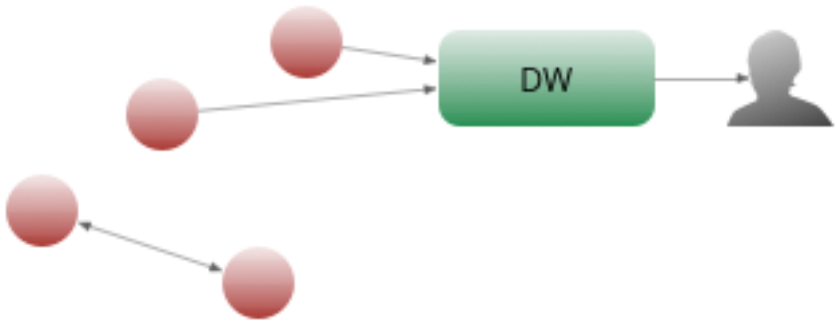

圖 3-2：資料湖泊成熟度階段一

（資料來源：[8]）

階段二「導入 Hadoop」：藍色橢圓為 Hadoop、藍綠色方形為資料分析器。企業將部份應用系統的資料儲存到 Hadoop，執行批次 Map/Reduce 運算，進行 ETL(資料擷取、轉換、載入)後的資料傳遞至資料倉儲或資料分析器，Hadoop 處理好的資料再回饋給應用系統。如圖（3-3）。

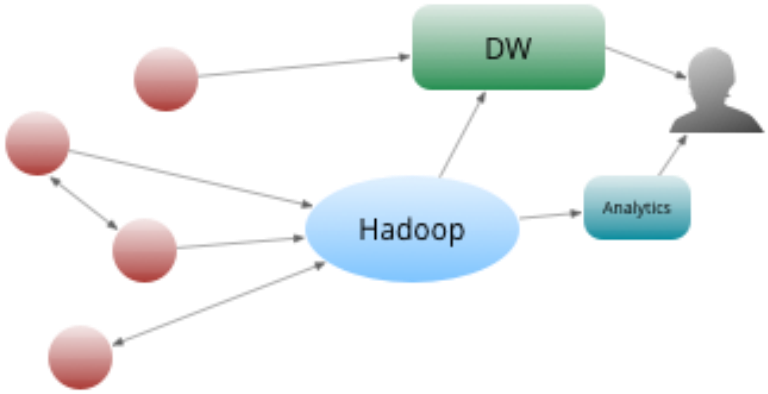

圖 3-3：資料湖泊成熟度階段二

（資料來源：[8]）

階段三「成長中的資料湖泊」：企業將新建置的應用系統資料都儲存到 Hadoop，各應用系統透過 Hadoop 互相取用資料，將資料分析工具（例如 Impala, Greenplum, Spark）佈署建置在 Hadoop 平台上，與 Hadoop 交互使用資料，至此 Hadoop 成為所有資料儲存的集散地，資料治理(Data Governance)與中介資料變得非常重要，資料倉儲則退居於後，僅供例外或特殊需求時才使用，外部的資料來源(例如 Open Data)也儲存至 Hadoop 進行整合。如圖（3-4）。

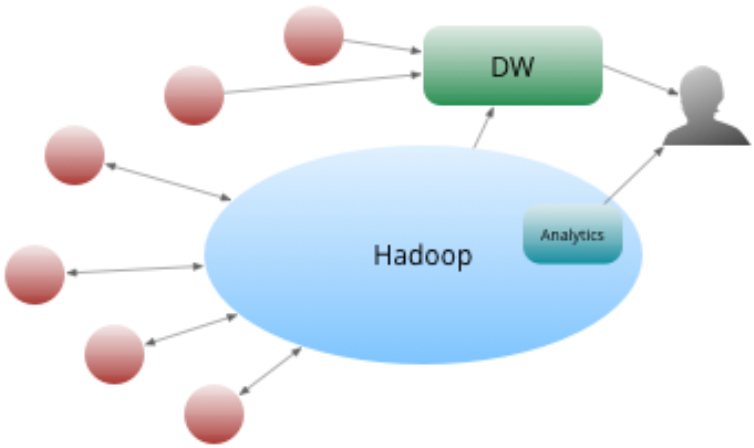

圖 3-4：資料湖泊成熟度階段三

（資料來源：[8]）

階段四「資料湖泊與應用雲端平台」：新的應用系統建置在 Hadoop 的應用雲端平台上，Hadoop 已成熟為彈性的分散式運算平台，提供營運需求與分析功能，資料可用性提昇，應用系統的部署時間則因此縮短，此階段須強化資料湖泊的安全性與資料治理的功能，部份應用系統因其特殊需求及歷史因素仍維持獨立運行。目前僅少數公司（如 Google、Amazon、Alibaba、Facebook）因其資料規模過大，需要高效率的資料處理架構而發展至這個階段。如圖（3-5）。

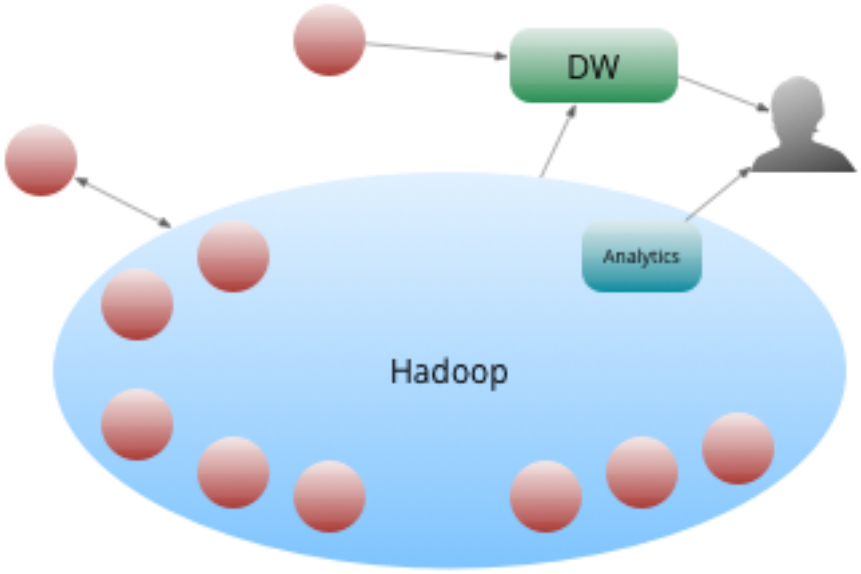

圖 3-5：資料湖泊成熟度階段四

（資料來源：[8]）

發展成熟的資料湖泊其內部的處理過程可參考圖（3-6），步驟 1-3 是資料蒐集階段，將批次交易資料、異動的文件、檔案、即時的串流資料傳遞到資料湖泊；步驟 4-6 為資料的處理，可能包含計算、過濾、字串處理、遮罩敏感資料、存取客戶具參考價值的資料、提煉與清洗資料，亦即 ETL 的過程；步驟 7 是將部份處理後的結構化資料回饋儲存至資料倉儲；步驟 8-10 是資料探索、分析、治理的階段可能用到統計、資料探勘、機器學習、建模、演算法等分析方法與視覺化分析工具；最後步驟 11 則是資料的應用與呈現，包含資料視覺化、報表、數位儀表板、說故事等。

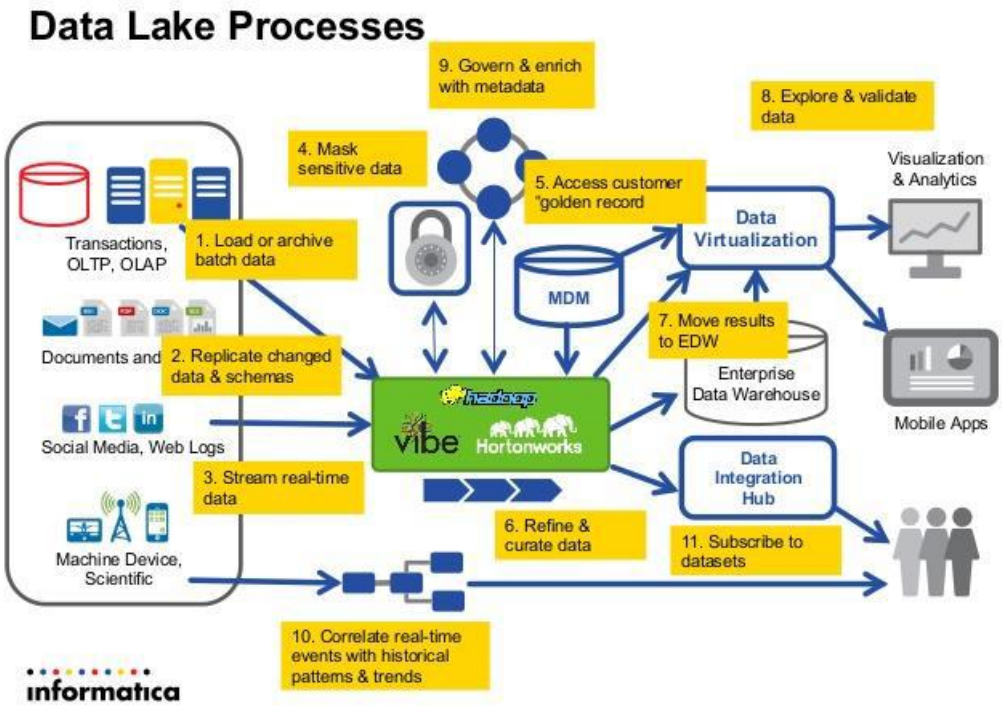

圖 3-6：資料湖泊處理過程

（資料來源：Informatica）

三、巨量資料分析

資料不僅隨處可見，且正不斷的急速成長，對大型企業而言，資料量可達到千兆位元組(PETABYTES)或更多，而中小企業的資料量也可能達到十或百兆位元組(TERABYTES)。引用翟本喬的說法，「想要做 Big Data 的人請先認清楚: 什麼是 data? 存得起來的，就是 storage；看得到的，才是 data；看得懂的，叫做 information；用得出來的，才能稱為 intelligence。」

（一） 資料科學

巨量資料分析處理的發展形成了一門新興的科學—資料科學（Data Science）以及其中的角色—資料科學家（Data Scientist）。因為巨量資料要做到的不僅是儲存和管理資料，而要能做預測式的分析（例如如果這樣做，會發生什麼？）資料科學是根據統計學的論證，從資料中獲得經驗和日後的方向。但資料科學並非簡單的統計學，僅需要現有的傳統基礎架構與軟體平台，需要新的應用、新的平台和新的資料觀。

「資料科學」的定義為，在分散式運算環境中，將大量且多種結構的資料，使用統計及機器學習技術，指出相關性及因果關係、能分類並預測事件、辨別出行為模式及異常事件、推論概率、興趣及情緒；「資料科學家」雖尚未有明確的定義，但大致上指的定義是:「資料科學家指的是，能運用統計分析或機器學習、分散式處理等技術，由大量資料中萃取出在商業上有意義的洞見，然後以簡單易懂的方式傳達給決策者的人才；或是用資料創造出全新服務的人才」。也有人打趣的說，資料科學家是擁有比軟體工程師好的統計能力且擁有比統計學家好的軟體工程能力的綜合體。自從哈佛商業評論在 2012 年 10 月刊載《Data Scientist: The Sexiest Job of the 21st Century》一文後，資料科學家吸引許多的關注與討論，包含：媒體到處尋找資料科學家的故事；企業審視要內訓還是外求資料科學家，以便創造更多的商業價值；學術單位探討要如何設計課程，以便培養更多的資料科學家。優異的資料科學家就像獨角獸一樣珍貴難尋，而且可不是只有科技公司在搶人，傳統金融界、零售商、廣告、教育，幾乎所有產業都需要資料科學家從大量數據中萃取精華。

資料科學家需具備的特質

1. 溝通能力：

即使能由巨量資料中發現有用的洞見，如果無法將其應用在實務上，該洞見的價值等於是減半。因此資料科學家必須擁有的特質是，能將資料分析結果轉化為「故事」，有效地將商業價值傳達給對資料分析不具備專業知識的業務單位或管理階層。

2. 創業家精神：

想創造出前所未有的、以資料為核心的全新服務，這樣的創業家精神也是資料科學家必須擁有的重要特質。Google 或亞馬遜、Facebook 等由資料中創造出新服務的企業，都是從龐大的資料辛苦摸索而提出創新服務才獲得成功。

3. 好奇心：

成功的資料科學家們有個共通點，就是不只對龐大的資料背後隱含的秘密擁有強烈的好奇心，對於藝術、技術、醫療、自然科學等各種領域，甚至對所有的事物都具有旺盛的好奇心。有時候透過將完全不同領域的資料結合在一起分析，能獲得以往從來未能得到的、深具價直的洞見。美國企業的資料科學家背景相當多采多姿，有實驗物理學家、計算機化學家、海洋學家，甚至神經外科醫生等。

4. 善於懷疑：

對資料科學家來說，能夠擁有批判性思考、採取批判的眼光來檢查自己的工作，而非採用片面的求同方式，是很重要的特點。

資料科學家需具備的技能

1. 電腦科學：

資料科學家需具備充分的電腦科學背景知識，一般而言是指程式設計、軟體工程、機器學習能力。

2. 數學、統計、資料探勘等：

除了數學和統計的素養外，資料科學家還必須有能力操作 SPSS 或 SAS 等主要的統計分析軟體，其中「R」是最近相當受注目的開放原始碼統計分析用程式語言及執行環境，R 的長處在於擁有豐富的統計分析用套件，且能以簡單的指令執行、具備能把結果視覺化的圖表製作功能。

3. 資料視覺化：

傳達方式的好壞會對一個訊息的品質造成很大的影響。把資料與設計結合，將難以一眼看清的資訊用簡單易懂的方式設計為資訊圖表（Infographics），或是運用外部 API 將大量資料的分析結果與圖表或地圖、數位儀表板（Dashboard）等其他服務結合，即是視覺化，對於資料科家而言是非常重要的複合能力之一。

（二） 視覺化分析

資料視覺化顧名思義就是將資料用視覺化的方式展現出來。人類對圖形的理解能力非常獨到，常常能從圖形當中發現資料的一些規律，而這些規律用正常的方法（例如表格）是很難發現的。在巨量資料時代，資料量變得非常大且繁瑣，要想發現資料中包含的資訊或價值，視覺化是有效的途徑之一。舉例來說在 1854 年處理霍亂過程中，內科醫生 John Snow 在倫敦地圖上標誌出所有病例的發生地點。在將霍亂病例視覺化之後，發現原來病例集結地區的水井被污染了，因此附近居民都染病，而在乾淨水源附近得病的人數就比較少。資料視覺化的目的有二，一是資料分析，二是用資料說故事。資料視覺化根據資料的特性，例如時間資訊和空間資訊等，找到適合的視覺化方式如圖表或地圖等將資料直觀展現出來，幫助人們瞭解資料，找出包含在巨量資料中的規律或寶貴資訊。時間性的資料可以分為連續類型和離散類型資料，連續類型資料是指隨時都在變化的資料，例如大氣溫度。而離散類型資料是間隔較長的變化資料，例如就業指數；空間性的資料視覺化時可依據地理特性分為點和面的視覺化，點的資料視覺化像是結合地圖的分佈圖，面的資料視覺化例如依行政區域劃分的色塊表示圖。視覺化分析就是將分析結果以視覺化方式呈現，告訴決策者這些資料的意義，讓決策者能更精準對這些資料做出判斷。

巨量資料的視覺化工具，以 eBay 公司為例，網路交易平台上有 1.08 億活躍使用者，2011 年售出了 600 億的貨物，產生了 52PB 的資料，包含使用者行為、網路交易、客戶貨運等資訊，eBay 使用 Tableau 公司的視覺化工具將 eBay.com 網站搜索的相關性和品質視覺化、監控最新的客戶回饋、測量客戶情緒、實現資料倉儲系統的業務報告。由於視覺化工具更貼近業務單位使用者（Business User），近年來已發展出多項能介接 Hadoop 陣營資料的產品，例如 QlikView、SAP BusinessObjects、DataWatch、Tableau、Zoomdata 等。

肆、 本行現況及建議事項

一、本行現況

（一） 資料倉儲

本行於 103 年將資料倉儲系統轉移至採用大量資料平行處理（Massive Parallel Processing，MPP）技術架構之平行處理資料倉儲系統，提升資料處理作業（Extract-Transform-Load，ETL）效率與錯誤回報能力，快速將資料匯入資料倉儲系統，系統架構採用 Teradata 資料平行處理伺服器，硬體效能大幅提升，且支援平行擴充，具備日後平行擴充機制，在容量或效能滿載後，可增購相關設備。

（二） 巨量資料浪潮下的衝擊

1. 資料倉儲效能雖已提昇，但資料倉儲的資料處理（ETL）應用僅限於處理結構化的資料，做商業分析、報表系統、商業智慧等。然而巨量資料的應用，必須尋求新的平台架構，能處理例如消費者的行為資料、網頁內容、社群媒體的內容資料、日誌、GPS 點位、XML 檔案、圖檔、影音檔等半結構化或是非結構化資料，並且將這些資料結構化儲存後再提供分析應用。

2. 以業務規劃與應用層面來看，業務單位使用者（Business User）無法以 Excel 或圖形化的 BI 工具直接拉取資料倉儲中的資料，需要倚賴資訊單位介入，從資料源取得資料、建立資料模式(Data Schema)，進入資料倉儲中再取出應用，在這些過程中不僅需要增加人力配置與成本，更重要的是從資料準備到分析、洞察資料價值的回應時間很長，無法即時反應急迫的需求，若能設法將資料分析探索的自由度還給業務單位使用者（Business User），將能大幅提昇本行巨量資料應用的效率與競爭力。

3. 金管會為配合行政院網路溝通與優化施政政策，積極推動金融資料開放，促成政府與民間協同合作創新，於 104 年 2 月 3 日時宣布啟動「大數據（Big Data）應用計畫」與「資料開放（Open Data）計畫」。巨量資料應用正快速改變傳統金融業務模式，巨量資料可透過加值及剖析應用，為投資人、金融業及臺灣企業 創造更大價值。而推動金融巨量資料分析應用、鼓勵創新網路金融服務及普及行動支付及第三方支付應用亦是金管會做為推動數位化金融環境 3.0 的三項重要策略。本行為因應數位金融環境時代來臨與巨量資料改變金融業生態，確實需導入巨量資料平台以增加競爭力。

二、 建議

（一） 導入巨量資料平台

提到巨量資料平台，Hadoop 已是國內外處理巨量資料發展穩定的主流技術，各大資料庫軟體管理供應商相繼發展此技術或平台， IBM、Informatica 自行發展、EMC 與 MapR 合作、Microsoft 與 Hortonworks 合作、Teradata 與 MapR 結盟、Oracle 與 Cloudera 合作，各自發展推出巨量資料處理設備或平台。在此發展之下，以 Hadoop 為核心平台的生態系統 (Hadoop Ecosystem)儼然成型，提供不同的軟體技術堆疊 (Software Stack)產品或服務，包含資料庫軟體、資料整合或資料品質平台、資料科學或進階分析工具、商業智慧工具、視覺化分析工具、交易分析應用工具、技術支援與教育服務等供企業採用。

綜上所述，並評估本行資料倉儲現況，建議以資料湖泊發展的概念逐步導入巨量資料平台，本行現行的狀態為資料湖泊成熟度階段一：各個應用系統(資料庫)相互獨立，部分應用系統間有做資料交換，部分應用系統資料傳遞到資料倉儲進行分析與報表產出。往階段二發展的下一步就是導入 Hadoop 巨量資料平台。環繞在資料倉儲周圍、現存的資料分析、報表、商業智慧（BI）等應用可評估是否移植到 Hadoop 巨量資料平台上，新的應用系統及資料來源逐步開始儲存到 Hadoop 巨量資料平台，導入 Hadoop 巨量資料平台後的架構可參考圖（4-1）。

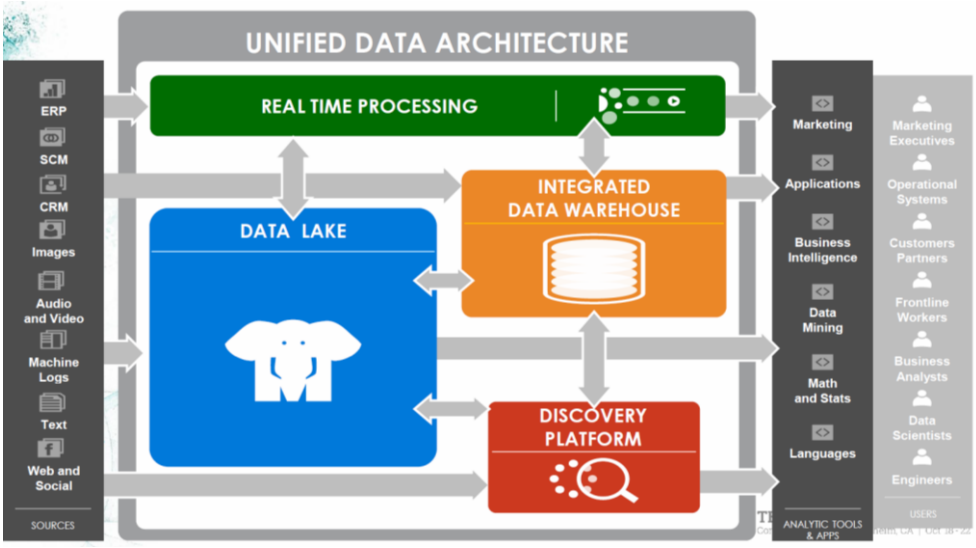

圖 4-1：巨量資料平台架構

（資料來源：Teradata）

（二） 導入探索平台或視覺化分析工具

傳統的商業智慧工具發展，受限於需花費較長時間與較高成本，建置、培訓擁有程式語言技能的分析人員。因此業務單位、研究單位及資訊部門，總要花費大量時間透過資料倉儲、資料超市或報表系統製作報表，完成的報表也常無法即時呈現所需意義。進一步說明，資料倉儲會限制分析人員執行繁重的分析或降低資料探索的靈活性，以本行現況舉例，資料是由資訊處 IT 人員和資料庫管理員來管理和控制的，而業務單位使用者，也就是資料的分析人員必須依賴資訊處 IT 人員來存取和更改資料模式(Data Schema)，需要更長的時間才能獲得資料。

隨著資料量規模擴大、資料源越來越多，建議本行在導入巨量資料平台時，同時考慮導入容易使用的資料探索平台或是視覺化分析工具。所謂探索平台，是能透過簡單易用的 SQL 語法就能分析資料與挖掘資料價值；而視覺化分析工具，能夠輕易的視覺化資料，依靠使用者熟悉的視覺處理切換視覺圖形，不再受限於須編寫程式才能分析數據，藉此看出趨勢和異常，協助人們了解資料，將資料轉化為資訊後，擷取出價值做為決策使用。舉例來說，透過簡單的滑鼠拖放動作，隨之顯示圖表，將複雜的資料快速簡化為易於分析的視覺化結果，節省製作報表及解讀資料的時間。更有研究指出，業務單位使用者（Business User）比

資訊部門 IT 人員更能挖掘巨量資料的價值，許多資料應用帶來的效益，幾乎都是由業務單位使用者探索分析而成。

市場上已出現巨量資料分析專用的探索平台或視覺化分析工具例如 QlikView、SAP BusinessObjects、DataWatch、Tableau、Zoomdata。廣泛應用在各專業領域，例如投資分析、貸款違約分析、貸款拖欠分析、信用卡分析、產業趨勢分析、市場區隔、價格分析、促銷效果分析、價格彈性分析、製程分析、供應鏈分析、預測模型、營收比較、財務分析、人力資源、銷售報告、數位行銷智慧、網站流量、社交網路和關係分析、詐欺偵測等。

（三） 成立資料科學家團隊

企業在導入巨量資料應用時，有以下關鍵角色：

1. 資料處理者：

接觸巨量資料的第一線人員，需具備程式技能、資料庫管理觀念以及資料探勘等資料處理相關技巧，以及 Hadoop、NoSQL 等相關知識，比方多結構化資料的 ETL 處理技術、或是於巨量資料平台上的程式撰寫（Map/Reduce、Pig、Hive、HBase 等），通常是由資工系畢業的軟體工程師擔任，然而對資料處理者而言，不一定要會什麼新潮的程式語言，只要在既有的 Java、Python、Ruby 等語言上有良好基礎，再廣泛學習資料處理的相關知識，就具備了作為資料處理者的基本能力。

2. 資料專家：

瞭解資料的業務專業人員，具統計學基礎，能夠提出明確的商業目標，或是引導商業目標的討論，也能針對最終的結果，提出行動計畫的人，可能是金融分析師、市場研究分析師、或營運經理、業務經理。

3. 資料科學家：

負責提出資料分析的演算法，或是能夠針對結果資料進行判讀，運用統計分析或機器學習、分散式處理、視覺化工具等技術，由大量資料中萃取出在商業上

有意義的洞見，然後以簡單易懂的方式傳達給決策者；或是用資料創造出全新服務。此外，如前述第參章、第三節提及關於資料科學家需具備的四項特質：溝通能力、創業家精神、好奇心、善於懷疑；及需具備的三項技能：電腦科學、數學統計與資料探勘、資料視覺化能力。

以本行而論，「資料處理者」角色為資訊處 IT 人員；「資料專家」角色可能會是各業務單位人員或主管，那麼「資料科學家」這個角色呢？在巨量資料的應用下，資料科學家這個舉足輕重的角色，卻也是本行目前所缺乏的職務。然而，資料科學家的專長是跨足業務及 IT 領域的綜合體，與其期待養成一個同時擅長統計、數學、資訊工程、機器學習、資料探勘、資料視覺化、善於溝通的資料科學家人才，更務實的作法，是成立一個「資料科學家團隊」。這個團隊包含的人員可能有熟悉資料統計模型的數學、統計、物理系背景人員，能將特定形式的資料，包括文字、網頁記錄、語意辨識和聲音檔案等各種資料，依據待解決的問題，套入不同統計模型計算，並產生出有意義的報表；或是具備分析技能，能處理原始資料、非結構化資料、熟悉大規模複雜分析技術的資料分析人員；還有具備業務知識、視覺化能力且善於溝通的業務專業人員，具備「說故事」的能力，能將分析結果以視覺化方式呈現，告訴決策者這些資料的意義，讓決策者能更精準對這些資料做出判斷。

建議本行根據未來巨量資料發展的方向，及前述有關資料科學家的幾項必備技能及特質，成立資料科學家團隊，除招攬熟悉資料統計模型的數學、統計、物理系背景人員之外，可從業務單位挑選擁有資料科學家特質的合適人員，加強視覺化分析能力、基礎程式語言、統計分析軟體課程訓練（例如 R）；或是從資訊單位挑選適合人員進行統計學及業務專業訓練。

參考文獻

1. https://www.microsoft.com
2. http://pivotal.io/labs
3. http://www.emc.com
4. http://www.teradata.com
5. http://www.teradata-partners.com
6. 蔣居裕，**Fred 豢養的雲中象「2015 年台灣 Big Data 市場五大趨勢預測」**、，http://fredbigdata.blogspot.com/2015/11/2015-big-data.html 、
7. 廖于嬋，DIGITIMES 中文網「EMC 併購 Pivotal Labs」，http://www.digitimes.com.tw/
8. Edd Dumbill，「The Data Lake Dream」，http://www.forbes.com/sites/edddumbill/2014/01/14/the-data-lake-dream/，2014 年 1 月。
9. Wu Jerry，**資料科學實驗室**，「什麼是大數據的新架構『資料湖泊』?」，http://dataology.blogspot.com/2014/10/blog-post.html，2014 年 10 月。
10. 李禹道，「巨量資料處理架構之研究-以資料倉儲系統為例」，台灣土地銀行民國 102 年度研究發展報告，2013 年。
11. 曾宗賢，「金融業客戶資料整合系統應用之研究」，台灣土地銀行出國研究報告，2014 年。
12. 城田真琴 著，鐘慧真、梁世英譯，「Big Data 大數據的獲利模式：圖解・案例・策略・實戰」，2013 年 8 月。
13. EMC 研究院、Vmware 研發團隊 周寶曜、劉偉、范承工編著，「巨量資料的下一步 Big Data 新戰略、技術及大型網站應用實錄」，2014 年 8 月。
14. iThome 電腦報「Spark 擊敗 Hadoop 刷新資料排序世界記錄」，http://www.ithome.com.tw/news/92449

15. 翟本喬，「IoT 和 Big Data 商機的迷思」，2013 年 4 月 https://www.facebook.com/notes/ben-jai/
16. Editor Wye，「攻克大數據，21 世紀最性感工作——『資料科學家』的八種技能」，2015 年 3 月，http://www.inside.com.tw/author/editor_wye
17. 金融監督管理委員會，行政院第 3456 次院會「打造數位化金融環境 3.0」，2015 年 7 月