非關聯型資料庫（NoSQL）在大數據(Big data)的應用

僑泰中學 龍清榮

前言

網際網路的應用愈來愈廣泛，除了傳統使用電腦外，現在因行動裝置的應用普及與便利，全世界已有 10 億以上人連結網際網路，每日使用移動裝置，存取、建立資料。也因世界經濟的開發成長快速，促進了密集資料科技的使用，更進而帶動資訊量的成長。全世界透過電信網路交換資訊的容量在今年流量將會達到 ZB (10 的 21 次方)，已是我們常使用 TB (10 的 12 次方)的一億倍大。由於數據的增多，產生了一個新的應用領域，大數據（Big data 或 Megadata），或稱巨量資料、海量資料、大資料。

由於資料量規模愈來愈巨大且產生非常快速，例如高速公路 ETC 系統，每日各偵測點所通過的車輛，高達百萬輛，都已到無法透過一般傳統的資料庫可以處理，例如最近伊波拉病毒發生，每個國家都非常注意防範及預防，如何在考慮人口移動、飛機的航班擴展的範圍、天氣溫度對於病毒的影響，利用大數據計算出，短期內如果發生疫情，預測出其擴散的程度，避免疾病擴散也都是大數據的應用一環。

簡介大數據與應用

人類每日產生的內容，包羅萬象，最簡單的如我們每天建立的檔案、電子郵件、手機通訊、網路查詢、社群活動等。各種資料有可能是檔案(影片)或是很簡單的一個通知訊息(line)，如果是檔案，其檔案格式又很多樣化，如圖檔、PPT、音樂等，如果是訊息，有的是資料內容，如姓名、住址、交易明細等，日復一日所產生的資料，而且還附帶大量的相關中繼資料，這些中繼資料更是驚人。

所謂中繼資料是指檔案的相關資訊，如資料是由誰建立檔案、檔案的類型是什麼、檔案內容關鍵值(key)有那些、誰讀取這個檔案等。這些原始內容和中繼資料，共同構成了人類的巨型資料。

大數據是由巨型資料集組成，這些巨型資料集是所有大小資料集的成長而來，已經超出我們能夠分析及處理的範圍，若不借助複雜的自動化技術，勢必無法達到目標。 我們必須仰賴技術來分析和處理這一波龐大的內容與中繼資料潮。

分析人類產生的巨型資料具有極大的潛力，不僅如此，運用中繼資料的力量，也成為商業應用極重用的一環。例如 Google 運用人們上網使用的習慣，分析個人喜好，讓你在旅遊中，以你所在位置，利用 GPS 定位，隨時提供你最佳旅遊資訊，讓你所在區域的店家增加商機，收取店家廣告費用，其廣告收入驚人，由其股價可以得知。

大數據資料的應用已廣泛在我們周遭應用，Facebook、Google、Line、Youtube 等世界型企業都提供很多服務，讓人們每天透過不同設備來存取他，當然許多企業的難題也才正要開始，因為他們如果要提供更好的服務，傳統的資料運用已無法滿足客戶的需求。加上雲端的運用，資料取得更方便，資料的取得又產生另一些中繼資料，資料以倍數自行成長，問題

就更加複雜了。企業為這些服務又另外建立了一個也會逐漸擴大的資料庫，來存放更多產生的內容，而這些資料也同樣需要加以管理及保護，如何架構一個全新的資料庫，應付大數據資料的應用，已是資訊產業的顯學。

資料庫的應用與類別

人類儲存文件，為便利取用與分類，一定會使用資料櫃來分類與儲存，資料庫可視為人類電子化的資料櫃，也就是儲存電子檔案的地方，使用者可以對檔案中的資料執行新增、擷取、更新、刪除等操作。由於任何數據的產生，必須能夠提供足夠的容量來儲存，並能達到快速應用擷取、管理、處理、並整理成為人類所想解讀的資訊。在傳統應用上，資料庫目前都是關聯式資料庫，例如：MySQL、Microsoft Access、Microsoft SQL Server、Oracle、Sybase、dBASE、Clipper、FoxPro，但是受限於其結構，這些都無法應用於大數據領域。

大數據幾乎無法使用目前傳統的資料庫管理系統處理，其特色是必須使用「在數十、數百甚至數千台伺服器上同時平行運行的軟體」。目前應用最廣泛的資料庫架構是非關聯型資料庫（NoSQL），例如：BigTable、Apache Cassandra、MongoDB、CouchDB。

BigTable 發展於 2004 年，現今廣泛應用於 Google 檔案系統（Google File System，GFS）的資料儲存系統，他是一種壓縮的、高效能的、高可擴展性的，用於儲存大規模結構化資料，適用於雲端計算。

Apache Cassandra 是一套 Key-Value 架構的儲存系統。它由 Facebook 開發，用於儲存大数據，集 Google BigTable 的數據模型與 Amazon Dynamo 架構於一身。由於 Cassandra 良好的可擴展性和性能，被 Digg、Twitter、Hulu、Netflix 等知名網站所採用。

MongoDB 是一個高性能，無模式的文檔型資料庫，以檔案導向的資料庫管理系統，它在許多情況下可用於替代傳統的關聯式資料庫，是當前 NoSQL 資料庫中比較熱門的一種。

CouchDB 的全名是 Cluster Of Unreliable Commodity Hardware Database，由 Erlang 開發的文檔型資料庫系統，是 NoSQL 資料庫中的重要代表，提供具有高度可伸缩性，提供高可用性和高可靠性，即使運作在容易出現的故障硬體上也是如此。

關聯型資料庫（SQL）與非關聯型資料庫（NoSQL）的差異

No SQL 全稱是 Not Only SQL，是一種不同於關聯型資料庫管理系統設計方式，簡單比喻關聯型資料庫如同一個制式書櫃，分每層及每格，能放的書大小是有規定的，利用 table 表示，也就是資料的樣態如同書本般，是有一定大小規範的，那就很容易處理，但是資料如果是包裹，如同檔案，有大有小，如何放置在書架上，也就是說關聯型資料庫是不合適處理。

傳統應用資料都是結構化，在應用上就很便利，但是在大數據應用中，資料格式多元，有大有小，所以結構化的資料庫在擴展及應用上就受限，典型的現代關連式資料庫在資料密集型應用上都有效能欠佳的問題，包括索引大量的文件，對於高流量的網站無法滿足其效能，對於串流媒體格式無法提供存取。所以在非結構化資料的領域中，代而起之的是 NoSQL，這種技術允許使用者以不同於關聯式資料庫概念，用行列存儲的格式來抓取信息提供服務。

關聯式資料庫因為具有支援非常複雜的資料表結構，關聯式資料庫每次的 Transaction 所花費的計算是相當大的，如果每秒都有上千次，甚至上萬次對關聯式資料庫進行讀寫，就會讓整個資料庫負荷相當沈重。一般社交網站每天可以產生數以千萬計的用戶動態訊息，每月達到以億為單位，對關聯式資料庫來說，查詢一個以億為單位的資料表，是極沒有效率且無法忍受的，而且關聯式資料庫是相當難以進行橫向擴充的，往往必須進行停機來進行資料庫的擴充及轉移，無法動態進行資料庫新增節點和負載平衡的工作。

NoSQL 資料庫並非是近年來的產物，但在上述缺點下，NoSQL 又再度被探討。NoSQL 資料庫其實是多種非使用 SQL 語法查詢的資料庫總稱，大致可以分為下列幾類。

Key/Value 型：存放 Key/Value 成對的簡單構造。如 Amazon.com 的 Amazon Dynamo 採用的資料模式。

列指向的表形式型：具有能夠處理列方向的構造。如 Google 的 BigTable。

文件指向型：轉換成 XML、JSON 等文件形式保存方式。如 10gen 公司的 MongoDB，Apache 的 CouchDB 等。

目前最有名的 NoSQL 的大型商業應用就是 Google 自主開發的 BigTable 和 Amazon 的 Dynamo，而在開放原始碼計畫上則有 HBase 和 Apache 的 Cassandra。

大數據(Big data)的應用

大數據的應用範例包括了大科學、RFID、感測裝置網路、天文學、大氣學、基因組學、生物學、大社會資料分析、網際網路檔案處理、製作網際網路搜尋引擎索引、通訊記錄明細、軍事偵查、社群網路、通勤時間預測、醫療記錄、照片圖像和影像封存、大規模的電子商務等。

當 Google 提供搜索引擎，可以在數以兆計的網路世界中，提供你即時資料，當亞馬遜靠大數據發展出無人能敵的電子商務，迫使美國零售業龍頭沃爾瑪跟進改進資料分析，當大陸阿里巴巴一日的商務交易量，可以大到一家公司一年的交易量，這些大數據帶來的熱潮，應用方興未艾，臺灣健保制度讓人民受惠，所有醫療資訊都在其中，如果使用大數據分析，將可以知道這幾年臺灣人口健康狀況，對於國家發展助益良多。

感覺上，大數據如果要應用，應該很難，其實大數據的應用並不難，而且隨手可得，這幾年民間有許多應用都與大數據應用有關，例如公司為了要辦戶外活動，如果要避開下雨，這時可用 Big Data 來給了他最佳判斷，使用「KNY 台灣天氣資訊」App，這個 App 收集了過去幾十年來的台灣氣象資料，數十年的資料，經過海量資料分析後顯示，讓使用者可以判斷當天天氣狀況，是否適宜辦活動。

結論

由於大數據(Big data)的應用是趨勢，在資料無結構性的前題下，傳統關聯式資料庫在擴展不易、效能搜尋不足、無法在不停機下擴增設備等限制下，已無法滿足需求，在大數據的運用上，每家公司對於資料定義取決於他對於持有資料的完整度，以及其平常用來處理分析資料的軟體之能力。並不是一定要資料達到 TB、PB 以上才是要使用大數據，對於某些公司來說，可能對於數百 GB 的資料集，就已是大數據的運用，但對於其他公司來說，資料可能

需要達到數十或數百兆位元組才會對他們造成困擾，所以對於資料庫的運用要求也不同，但是不管如何，NoSQL 資料庫在大數據的應用已是主流，應用面及成熟度也更臻於完美。