資料爆發與儲存系統－陳宗治/廣達電腦協理，技術專家委員會委員

因為資料爆發所引發的需求，使得使用者在要求低價的儲存方式的同時，也促使儲存系統被大眾化(commoditized)，而資料的主要價值在於該資料所要表達的涵義(the insight of data)。對於資料儲存設備這一類的產品而言，其所提供的產品差異性及附加價值在於洞悉資料價值的資料服務(data service)。海量資料分析(big data analytics) 已成為儲存系統公司增加產品價值的主戰場，例如 EMC 收購了 Greenplum，NetApp 與 Hortonworks 合作皆顯示了上述的趨勢。然而因為儲存的特性，儲存系統並不會分析和了解其儲存資料內容的涵義，因此傳統的儲存方式只能有限制地達成存取最佳化(access optimization)，例如經常存取資料的快取(caching)即為其中一個例子。儲存基礎建設(storage infrastructure)的設計已成為資料結構及佈局(data architecture and layout)上整體設計的一部分，藉由資料服務的提供，儲存設備可以對資料數據有充分地了解並藉此達到存取最佳化。

未來的運算將會以資料為中心。只要資料被安全地保存著，伺服器可以故障，只要有備用的伺服器即能接手執行故障的伺服器所未完成的工作。然而資料

是放在儲存系統上，所以儲存系統是絕對不能故障的，否則之前的工作狀態都將付之一炬。多年來，傳統企業等級儲存系統(enterprise storage)的供應商持續在市場上以非常昂貴的價格提供可靠的(reliable)、可容錯的(fault-tolerance)，以及高可用性(high-availability)的儲存系統，但是這樣的價格對於絕大部分以資料導向(data-driven)為主的應用來說，是無法負擔的。這些企業等級的儲存系統提供了許多儲存功能，包括資料冗餘性(data redundancy)、資料可用性(data availability)、可容錯性(fault tolerance)、可擴展性(scalability)、異地備援功能(remote replication)、快照功能(snapshot)、備份功能(backup)、資源隨需分配(thin-provisioning) 、資料去重複化(de-duplication)、壓縮(compression)、高可用性(high availability)、災難重建(disaster recovery)、服務性(serviceability)、虛擬儲存(virtual storage)，持續資料保護(CDP)和資訊生命週期管理(ILM)等。

將這些產品功能大眾化並不只是明顯地將價格降低，同時還要讓一般大眾可以確實地了解和方便地使用這些功能。對使用者而言，使用者理想中的儲存系統是在有需求時能提供無限的使用空間，並以實際使用的儲存量作為計價方式。為了達到這種隨取儲存(on-demand)的要求，儲存系統不能像現今市面上大部分的儲存產品是安裝在特定硬體上之嵌入式儲存設備(embedded storage

appliance)，而是必須能從資料中心的資源池中動態地配置(dynamically allocate)運算能力、網路頻寬及儲存容量，以符合隨取要求下智慧地提供(intelligently provision)虛擬儲存的空間。為了將現今以資料導向為主的應用程式所產生的資料儲存下來，儲存系統必須在非常低價的前提下，仍能提供這些企業等級儲存系統的功能，以進一步改善資料中心的儲存建設之可靠性(reliability)、可用性(availability)和服務性(serviceability)。

網路上提供愈來愈多的服務，而使用者利用這些服務在線上與他人互動、合作和分享資訊，任何人都可以在任何時刻、任何地點，使用任何的連線裝置來產生和使用資料。傳統的儲存系統只能讓使用者存取資料，無法藉由觀察使用者的資料內容進而達成整理及管理資料的目的，以便日後存取。隨著海量資料(big data)的來臨，提供建構於儲存系統上的資料分析服務(data analytics service)能夠幫使用者整理及管理資料，並從資料中擷取價值。對使用者而言，理想中的儲存系統是在有需求時能提供無限且虛擬的隨取儲存空間以儲存大量資料，並能以可負擔的價格快速取得所需的資料片段。

使用者可以從大量資料所獲得的價值是少量資料所無法提供的，海量資料的分析和管理已經成為商業智慧的下一個重大指標，現今許多公司正處於前所未有

的資料爆發，而這些資料大部分是未結構化的資料(unstructured data)，其中的資料涵括潛在的商業價值(the insight of business value)，這類典型未結構化的資料只能經由人為解讀，卻很難被機器所處理。藉由分析資料了解其涵義後，機器方能建構資料的涵義與關聯性的詮釋資料(metadata)，進一步有效率地擷取資料的價值。藉由分析未結構化的資料，並以分析結果去建構這些資料的涵義與關聯性的結構化資料(structured data)對資料探勘(data mining)至為重要。這一切都只是為了將人和相關的資料(relevant data)連結在一起，為了達到這個目的，資料搜尋只是開始的第一步，卻不是最佳的解決方法。就像未來的網頁應該轉化為概要網頁(schematic web)，以便更容易且有效地讓資料在應用程式間分享和重新應用。

計算機科學(Computer science)被用來處理運算資源的管理，而資料已成為資料中心裡最重要的資源，資料管理的方式亦有別於以往管理運算資源的技術。資料科學(Data science) 將為這個世代的「計算機科學」，以現今的資料成長速度來看，資料爆發已成為資料中心管理者及企業主所頭痛的議題。因此未來是屬於可以有效蒐集資料及從大量資料中擷取出商業價值的公司。