雲報專欄：巨量資料分析在雲端運算部署上的發展—王雲

工研院專家/技術專家委員會委員

近年來談到 IT 產業的發展主軸，巨量資料分析及雲端運算毫無疑問的是所有人的首選。每當談到巨量資料的「三個 V」，亦即多元性 (Variety)、數量 (Volume) 和速度 (Velocity) 時，自然會聯想到應該透過雲端技術來部署巨量資料分析的系統。當我們進一步檢視巨量資料分析系統的架構，一個很簡單的層次架構可以包含應用層、分析層、資料處理層以及基礎系統層。同樣的，雲端運算的服務模式也包含了軟體即服務 (SaaS)、平台即服務 (PaaS)、以及基礎架構即服務 (IaaS) 的服務模式。因此巨量資料分析在雲端的發展也就產生出各種不同的模式。本篇短文對全球 IT 業界於巨量資料分析在雲端運算的部署做一個簡單的介紹。

一、亞馬遜的發展

從全球雲端運算產業界的發展至今來看，國際研究暨顧問機構 Gartner 在 2014 年的雲端基礎架構即服務 (Cloud IaaS) Magic Quadrant 的研究報告中列出亞馬遜 (Amazon) 大幅度領先其他業者，為此領域的領頭羊。在此先看一看這些年亞馬遜在其雲端服務 AWS (Amazon Web Services) 中提供了哪些有關巨量資料分析的服務？亞馬遜在 2006 年啟動了雲端服務的業務 AWS 之後，從 2009 年開始提供了以 Hadoop 為基礎的 Amazon Elastic MapReduce (Amazon EMR) 服務、以及後續的數據倉庫及商業智能的 Amazon Redshift (2012) 服務、資料串流的 Amazon Kinesis (2013) 服務、配合巨量資料分析的其他各種服務如 NoSQL 資料庫功能的 Amazon DynamoDB (2012) 服務、雲端儲存及運算的 Amazon S3 (2006) 服務、關聯式資料庫 Amazon RDS (2009) 服務、Amazon EC2 (2006) 服務等等。其中 Amazon EMR 包括了 Apache Hadoop 平台以及常用的 Hadoop 生態系統，譬如：Hive、Pig、HBase、Mahout、Impala 等等存儲及分析的功能。Amazon 2014 年在美國拉斯維加斯舉辦的年度大會中 (re:Invent) 宣布推出了商用級關連資料庫 AWS Aurora，以及推出事件驅動運算服務 AWS Lambda，這也都顯示著亞馬遜持續研發其雲端服務中巨量資料分析平台的策略。作為一個巨量資料分析雲端運算平台，亞馬遜也提供給第三方供應商一個技術及商業的市場平台，在亞馬遜市集 (AWS Marketplace) 中能發現許多第三方的巨量資料分析解決方案提供軟

體即服務及基礎架構即服務的服務模式。這樣的產業生態系統幫助了用戶能更加容易地開發及使用雲端的巨量資料分析應用及系統。

二、微軟的發展

微軟 (Microsoft) 在其近年來的年度財報中都一再的指出雲端計算是微軟最重要的公司戰略，在微軟 2014 年的財報中甚至提到微軟商業雲收入可達到 44 億美元。Gartner 也在 2014 年的雲端基礎架構即服務 Magic Quadrant 的研究報告中把微軟提升為與亞馬遜並列為產業的唯二領先者，並且預估微軟將會持續縮短其差距。談到微軟的巨量資料分析在雲端運算的部署，首先會想到的就是：微軟的 Apache Hadoop 微軟版- Microsoft Azure HDInsight (Oct，2013)了，HDInsight 採用了 Hortonworks Data Platform (HDP)為主體包括了大部分 Hadoop 生態系統譬如 Storm、HBase、Pig、Hive、Sqoop、Oozie、Mahout、Ambari 等等。身為數據庫，數據倉庫及商業智能的領先廠商微軟也將其傳統產品雲端化，其中 SQL Azure 是跟 Windows Azure 早在 2010 年就推出了。在資料存儲服務中除了 SQL Azure 外還提供了 Azure DocumentDB(NoSQL)、Azure Redis(Cache)等等及一系列的分析服務，包括 Azure Machine Learning、Azure Stream Analytics、Azure Event Hubs (2014)等等。同樣地 Azure 也提供給第三方供應商一個技術及商業的市場平台，在 Azure 服務商場 (Azure Marketplace) 中能發現許多第三方的解決方案。

三、谷歌的發展

談到巨量資料分析及雲端運算不能不談到谷歌 (Google)，在許多分析報告中談到 IT 的第三平台 (3rd platform) 時會將蘋果公司列為移動的代表、亞馬遜列為雲端的代表、臉書列為社交的代表、谷歌列為巨量資料的代表。Hadoop 的原始組件 - HDFS 及 MapReduce 均源於谷歌，而谷歌巨量資料分析也都是應用在雲端運算上。谷歌研發了許多雲端運算及巨量資料的技術支持其龐大的互聯網線上應用。在亞馬遜 2006 年啟動了雲端服務的業務 AWS 之後，谷歌在 2008 年推出了平台即服務的 Google App Engine (GAE) 提供能自動縮放的網上應用，2012 年推出了基礎架構即服務的 Google Compute Engine (GCE)。之間，谷歌在 2010 年推出了支持巨量資料分析的 BigQuery 及 Prediction API，BigQuery 提供了對巨量資料以類似 SQL 的極快速查詢功能，而 Prediction API 則提供了機器學習演算法對

數據進行分析及創建預測模型。與亞馬遜的 AWS 及微軟的 Azure 相同地谷歌的雲端平台 (Google Cloud Platform) 也能部署第三方的解決方案，譬如：Apache Cassandra、Hadoop、MongoDB、RabbitMQ、Redis 等等。

四、IBM 的發展

身為 IT 產業界百年老店的 IBM，近年來也一直把雲端運算及資料分析與優化訂為公司成長的主要戰略。在 2011 年 IBM 即推出了 IBM BigInsights (IBM Distribution Hadoop) 能內部部署也可以雲端部署，IBM 具有許多傳統軟體產品及解決方案，並且也都先後被部署到雲端的環境上。在 2013 年 IBM 併購了 SoftLayer 之後，成為 IBM 新的雲端平台，許多 IBM 及第三方廠商以各種不同的服務模式 (IaaS、PaaS、SaaS) 部署了種種產品及解決方案。在 IBM 雲端商場中可以瀏覽到上百項產品及服務，關於巨量資料分析的有：BigInsights (IBM Hadoop)、IBM Streams (數據流)、dashDB (數據倉庫和分析)、VoltDB (內存 NewSQL 數據庫)、Time Series DB (時間序列數據庫)、Geospatial Analytics (地理空間分析)、IBM Watson Analytics (問答分析) 等等。

五、Pivotal Software 的發展

在以上四家傳統企業以及網絡新興企業之外，另外有一家在 2013 年四月成立的公司名叫 Pivotal Software，是由 EMC 及 VMware 最近收購的一些公司合併而成，其中包括了：Greenplum (數據庫、商業智能)、Cloud Foundry (平台即服務)、SpringSource (Spring 框架)、GemStone (分佈式內存) 等等，產品目標是針對雲端運算 (PaaS) 及巨量資料分析，尤其是面對企業界的用戶以及私有雲建置。通用電氣 (GE) 在 Pivotal 2013 年剛成立的時候即做了策略性超過一億美元的投資取得了 Pivotal 10% 的股份，而且通用電氣在其近年來大力推動的工業互聯網 (Industrial Internet) 中也宣布採用了的 Pivotal 的大數據軟件套件 (Big Data Suite)。Pivotal 的大數據軟件套件包含了 Hadoop (Pivotal HD 及 HAWQ)、關聯式資料庫 (Pivotal Greenplum Database)、多種分佈式內存系統(GemFire、GemFire XD) 能提供從批量、交互式到實時的各種處理方式。在資料分析方面 Pivotal 大力借助及推動 MADlib 把各種統計、預測、機器學習計算推送到資料底層運算。

六、結論

綜合來看，現階段巨量資料分析引入了 Hadoop 及相關技術來處理大規模非結構化資料，配合不斷改進的傳統數據庫、數據倉庫、商業智能、分析預測、機器學習等等相關技術，更加入分佈式內存、數據流等等相關技術來處理實時性的需求。而以上的科技更需要能在不斷精進的雲端科技上部署實施。要能即時掌握這些先進 IT 科技應用在各個垂直產業界中，將是我們 IT 工作者面臨的高度挑戰與令人興奮的機會。