地理資訊系統服務於雲端環境之適用性探討
A study of GIS services in the cloud

彭逸帆[*]　　白璧玲[**]

Yi-Fan Peng　　Pi-Ling Pai

摘　要

近年來由於網際網路技術蓬勃發展，連帶使得地理資訊科學所應用的各項技術與處理能力均較以往更為先進。首先，最顯而易見的改變即是資料品質的大幅改善，像是地理資訊系統常被應用的圖資影像，其品質均較過去有長足的進步；其次，針對資料處理能力而言，隨著硬體計算能力的提升，使得在處理地理資訊相關問題時可以有更強大的處理能量。

不過，畢竟單一軟、硬體的處理能力仍有其上限，若碰到需要處理巨量資料時，僅憑單一系統仍稍嫌不足。所幸近來雲端技術的盛行，有愈來愈多的領域紛紛嘗以雲端技術來解決該領域之問題；地理資訊科學領域當然也不利外。

對於雲端技術的適用性，目前許多專家、學者仍進行相關討論與實驗，現階段在公有雲（Public Cloud）安全性仍有疑慮的情況下，以私有雲（Private Cloud）進行相關測試是目前較為可行的方案。本文即是以中央研究院計算中心 GIS 組為例，嘗試透過 Hadoop 自行建立小規模之私有雲，並將部份 GIS 服務結合私有雲進行說明，並探討地理資訊系統服務在雲端環境下的適用情況。

關鍵字：地理資訊科學、雲端技術、Hadoop

Abstract

The recent accelerating development of Internet technology has allowed the improvement in the ability of dealing with several issues confronted in the field of geographic information science. First and foremost, there has been a vast improvement of data quality. Base maps used for GIS are in higher quality than before. Secondly, computing technology has also been fortified, and therefore more difficulties can be overcome more efficiently.

However, there are still limitations whether in terms of software or hardware. The current status remains insufficient when it comes to dealing with huge amounts of data by using a single system. Because of the rising popularity of cloud technology, more and more researchers are trying to apply this as means of solving issues. And the field of geographic information science is no exception.

*　中央研究院計算中心資訊人員
Research Assistant, Computing Center, Academia Sinica

**　中央研究院計算中心 GIS 組長
GIS Team Leader, , Computing Center, Academia Sinica

Specialists and scholars are still debating over the suitability of applying cloud technology in researches. Due to security reasons, it is better for some experiments to be done in private cloud environment. This article discusses the experience of the GIS team of the Computing Center, Academia Sinica in our attempt in building a small private cloud by incorporating Hadoop technology. Issues on running GIS services on the cloud are also discussed.

Keywords： Geographic Information Science, Cloud technology, Hadoop

前　言

隨著資訊技術的演進，依賴著新穎設備所產製的資料也較以往更為細緻。以地理資訊科學（Geographic Information Science，GIScience）領域為例，相關研究或應用的資料品質都較以往大幅提升許多。舉例來說，在過去因應地理資訊系統建置的需求，大多都會需要使用到基礎圖資（Base Map），而這類資料內容包含向量資料或是影像資料，因為產製設備的提升，現階段可以使用到的影像資料容量已由早期的幾MB到目前可以到GB甚至TB的等級。由此即顯而易見地看出，透過資訊技術的進步可以替相關研究成果帶來的益處。

雖然目前電腦設備及其性能也較過去大幅增進許多，然而對於巨量資料的處理，以現有系統架構仍有其極限，所以當面臨處理巨量資料的問題時，勢必需要引入新的想法及架構，以處理不斷增長的資料內容。

目前資訊科技領域對於處理類似問題時，最常被討論的解決方案即是引入雲端運算（Cloud Computing）的技術，透過雲端強大的計算與儲存能力，用以解決過去所無法處理的問題。此概念不僅適用一般資訊領域的問題，對於地理資訊科學領域也是可以應用。本文即是針對地理資訊科學領域結合雲端技術可能產生的一些議題進行討論與分析。

相關研究

(一) 一般地理資訊系統服務環境架構

在一般資訊系統中，如果要將服務提供多人服務，則需將該資訊系統安裝於網路環境中；而所採用的系統架構也必須從原本的單機環境轉變成多人的作業模式。以目前常見的處理方式，即是使用主從式的架構（Client-Server Architecture），使用者對伺服器提出需求（Request），伺服器處理之後將結果回傳（Response）給使用者。如圖 1 所示。

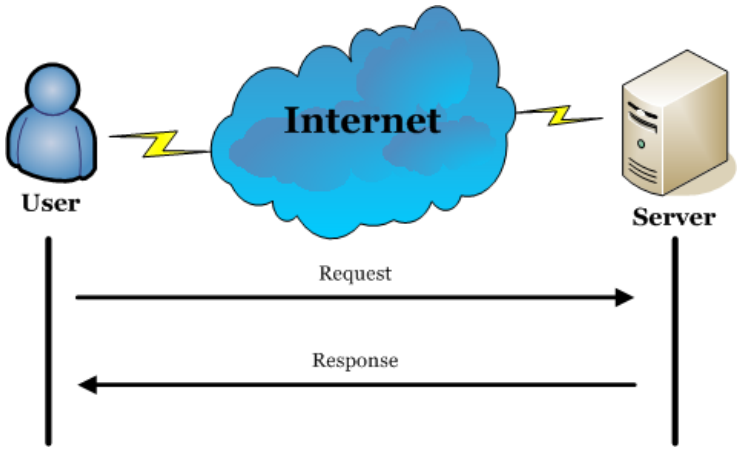

圖 1　主從式架構示意圖

採用此種架構的好處是，整個資訊系統是於網路環境運行，使用者僅需透過網路，即可存取資訊系統內的相關內容進行工作的處理；而管理人員只需維持好伺服器的運作正常即可。

在一般地理資訊系統運作的模式上，當資料提供者將資料處理好後，也需將相關成果放置在伺服器上，才能夠提供使用者存取使用。以現行的系統架構大多也是採主從式架構進行系統的建置；而為了讓伺服器能發揮更大的效用，一般在單一伺服器中往往會安裝一個或一個以上的服務，好讓伺服器盡可能發揮機器的效能。當使用者有特定地理資訊服務的需求時，僅需透過網路服務的方式提出需求，伺服器即可針對使用者所提出的需求進行資訊的計算或處理，最後將執行的結果回傳，或是透過電子圖台顯示的方式將結果呈現，如圖 2 所示。在地理資訊系統中有多項圖資內容的服務供使用者使用，當使用者針對地理資訊系統提出檢索的功能時，伺服器經過處理之後將計算完畢的結果回傳給使用者。

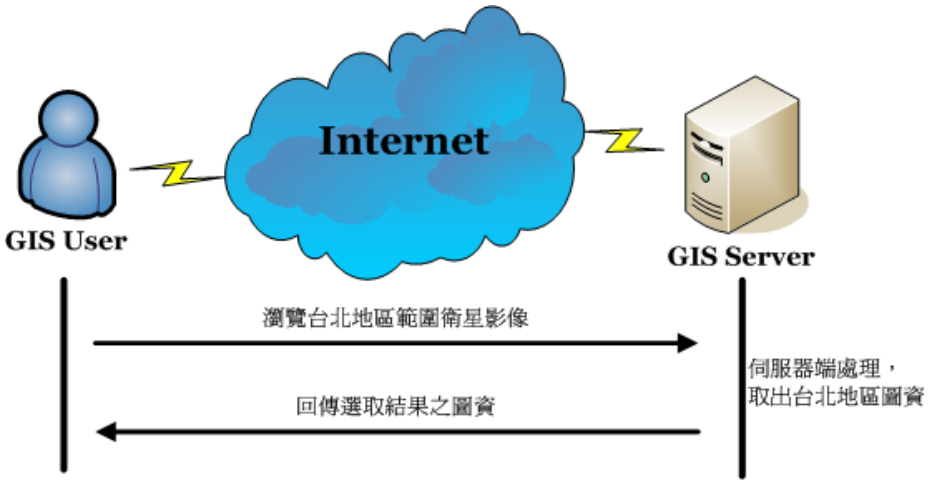

圖 2　地理資訊系統資料處理示意圖

(二) 已知存在問題

雖然上述資訊系統架構具有系統資源分配的彈性，也就是當使用者針對地理資訊系統提出服務需求時，系統才會對相關要求進行處理；若沒有任何需求請求時，系統即處於閒置的狀態。

然而如同前面提到的問題，因為資訊技術的提升，目前地理資訊系統所需處理的資料量往往需要耗費更多的系統資源；若於資料處理的期間有新的使用者需求提出，即會影響整體的處理效能。

舉例來說，當有一使用者對地理資訊系統進行特定區域的環域計算（Buffer）時，因為此項工作需要先將相關資料載入至記憶體後，才能進行後續的分析工作。在有限的資源下，若因為執行一位使用者的 Buffer 工作就會影響到另一位使用者可以運用的資源。若將此問題繼續放大，當有一萬名使用者同時對單一地理資訊系統提出要求時，可以想見單一地理資訊系統是無法抵擋如此龐大的計算量。

因此，如何處理或解決此類問題，已是目前多數地理資訊服務提供者所面臨的重大挑戰之一。

(三) 分散式架構

為解決此類單一系統可能發生的問題，在早期就有學者提出分散式的架構（Distributed Architecture），藉以改善單純主從式架構的問題。所謂分散式架構相對於主從式架構，其概念是在於將原本集中於單一伺服器的架構改為將工作分散到多台伺服器上，如圖 3 所示。

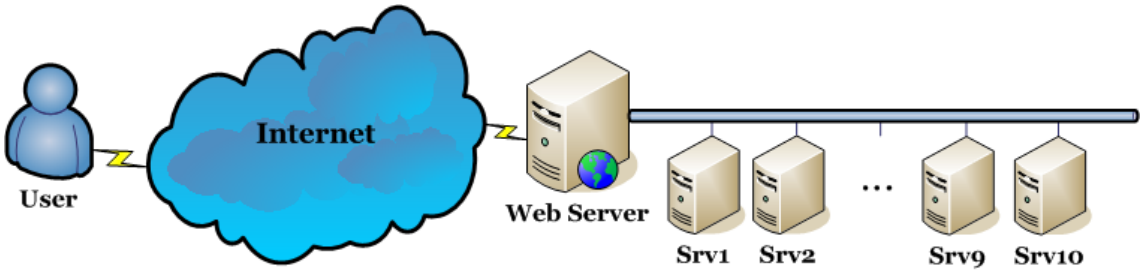

圖 3　分散式架構示意圖

舉例來說，原本同一台主機可能要同時提供十種不同類型的圖層資料，由於每一圖層的資料內容都相當複雜且龐大，如果單一伺服器主機要處理多個圖層資料的計算是非常容易就造成系統效能的降低，但如果採用分散式的架構，即可將原本一台機器十項服務的架構改為十台機器，每一台機器負責一項圖層的資料，在不考慮成本的情況下，即可增加整體的服務能力。

雲端環境簡介

由於雲端運算的概念已廣為大家所接受，所以許多資訊大廠均投入相當的能量進行相關雲端運算環境的建置與應用。針對雲端運算的產業約略可以分成如圖4所示的幾大類別 (Hrushikesh Zadgaonkar，2011) 。

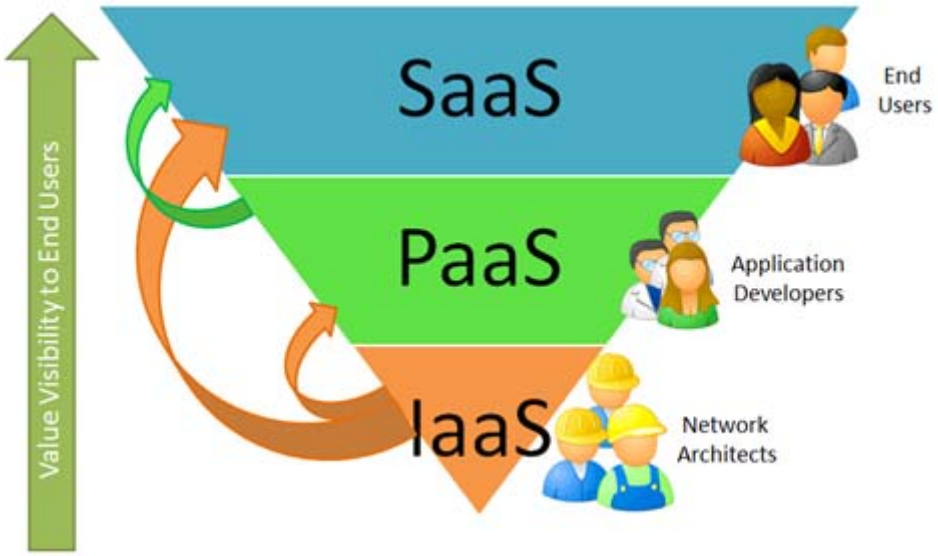

圖 4　雲端產業類別及其適用使用群示意圖

針對幾家發展雲端環境的資訊廠商而言，他們依據自己所定義出的概念建置並提供相對應的雲端環境服務，大略可以分成如表1所列幾項 (王耀聰、陳威宇，2008)。

表 1　主要雲端環境服務提供者之服務比較表

	Amazon EC2	**Google App Engine**	**Microsoft Azure**	**Yahoo Hadoop**
Architecture	IaaS/PaaS	PaaS	PaaS	Software
Service Type	Compute/Storage	Web application	Web and non-web	Software
Manage Technology	OS on Xen hypervisor	Application container	OS through Fabric controller	Map/Reduce Architecture
User Interface	EC2 Command-line tools	Web-based Administration console	Windows Azure Portal	Command line and Web

對於部份商業軟體所應用的技術，由於其細節為該公司之商業機密，一般使用者也無法取得細部內容，底下針對使用開放原始碼技術之雲端環境與資料庫作概略性的介紹。

(一) Hadoop 簡介

Hadoop 原是 Apache 底下 Lucene 裡的一個由 Dong Cutting 所發起的專案；而 Lucene 是一套用於全文檢索和搜尋的開放原始碼程式，後來因為網際網路的蓬勃發展連帶使得 Lucene 大受歡迎，有愈來愈多感興趣的人投入研究；而 Hadoop 則是建構 Lucene 底層環境的程式，發展到後來，將這種底層技術應用到其他領域的人愈來愈多，因此到後期，Hadoop 也晉身至 Apache 最頂層的專案。關於 Hadoop 的一些重要發展歷程，整理如表 2 所示 (Tom White，2009) 。

表 2　Hadoop 重要發展史

2003.02	Google 撰寫出第一套 MapReduce 函式庫
2003.10	Google 的 GFS 論文發表
2004.12	Google 的 MapReduce 論文發表
2005.07	Doug Cutting 在 Nutch 中實作了新的 MapReduce 技術
2006.02	Hadoop 的程式從 Nutch 計畫中提升到 Lucene 的次計畫
2006.11	Google 的 Bittable 論文發表
2007.02	Mike Cafarella 撰寫出第一套 HBase
2007.04	雅虎（Yahoo）執行 Hadoop 的 Node 數超過 1000 個
2008.01	Hadoop 變成 Apache 中的頂層計畫（Top Level Project）

就實際應用的層面而言，目前在 Yahoo 有超過兩萬台電腦（約 100,000CPU）在負責他們公司的雲端平台， Google 則是應用這項技術解決網路上大規模計算的問題，而 Amazon 則是透過 Hadoop 的平台用以建構出他們大量商品的索引資料；防毒軟體趨勢公司則是將此一平台應用於防毒、惡意攻擊、...等危害資訊安全的應用上。

針對 Hadoop 的基本架構（Hadoop Distributed File System）如圖 5 所示。

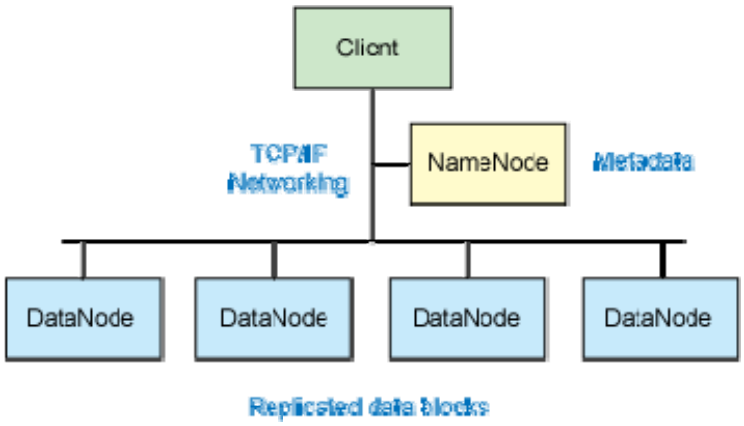

圖 5　Hadoop 分散式檔案系統

當使用者有一項工作要放在雲端環境上執行時，使用者透過 NameNode 將資料存放於多個 DataNode 中，而實際資料存放的方式與位置是由 NameNode 負責處理。為了資料安全與降低負載，存放資料的方式也並非僅單一份放至於單一 DataNode，而是將資料分散放置在不同的 DataNode 中。而不論是存放的資料或是計算的工作，都會被切割成許多小單元，進而分散到各伺服器進行處理。

(二) HBase 簡介

一般系統架構在有了基礎環境之後，對於處理龐大的資料內容往往都會以資料庫的形式進行管理。換言之，即是希望透過資料庫可以有效地查詢、檢索、重複使用龐大的資料內容。

然而，對於雲端環境而言，其系統運作方式與一般在同一伺服器內部運作有非常大的差異。以傳統資料庫系統為例，在單一環境底下，資料庫必須符合四項原則（如表 3 所示），這些原則如放在雲端環境中卻不見得都可以達到。

表 3　資料庫系統四項原則

縮寫	英文名稱	中文名稱	解釋
A	Atomicity	原子性	一整個交易中的所有操作，要麼全部完成，要麼全部不完成，不可能停滯在中間某個環節。交易在執行過程中發生錯誤，會被反轉（Rollback）到交易開始前的狀態，就像這個交易從來沒有執行過一樣。
C	Consistency	一致性	在交易開始之前和交易結束以後，資料庫的完整性限制沒有被破壞。
I	Isolation	隔離性	兩個交易的執行是互不干擾的，一個交易不可能看到其他交易運行時，中間某一時刻的數據。
D	Durability	持久性	在交易完成以後，該交易對資料庫所作的更改便持久地保存在資料庫之中，並不會被反轉。

如同表 3 所示，在單一系統中，如果要達成該四項原則，僅需控制好單一伺服器運作的情況即可符合該四項原則；然而，在雲端環境中，由於資料的傳遞不單只是在單一主機內的匯流排（Bus）內傳送，而是可能在內部網路（Intranet）甚至是網際網路（Internet）上傳遞，所以要達成資料庫的四項原則是有相當的難度。

因此，若要在雲端環境中使用資料庫的功能，資料庫的四大原則勢必需要做些修正，才能夠引入使用。

HBase – Hadoop Database 是一 Column oriented 型態的資料庫，由於該資料庫式架構在 HDFS 之上，所以可以透過該架構將資料內容分散儲存在雲端環境中，因此是一個具有高可靠性的分散式儲存系統。其架構如圖 6 所示 (Lars George，2010) 。

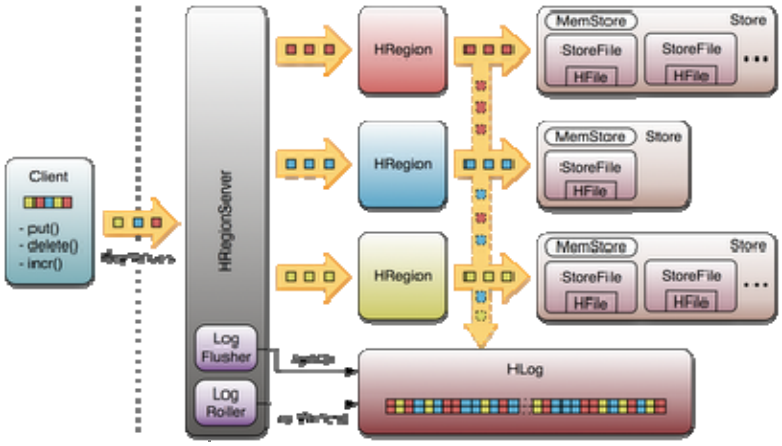

圖 6　HBase 架構圖

地理資訊系統引入雲端計算概念

如同前言所提及，由於地理資訊系統的資料與計算量均呈現大幅度成長，因此現在有越來越多的研究均傾向尋找較大的計算資源與環境以解決原本無法解決的問題，其中，雲端運算的環境即是一個可能解決的方案。

然而，對於地理資訊科學相關的問題是否適合直接以雲端環境原生的架構來解決，這個部份仍是需要思考與探索的地方。底下將針對雲端運算適合的問題處理模式進行分析，接著以地理資訊系統的角度探討與雲端運算結合可能產生的一些議題。

(一) 雲端運算之適用性

每一種解決問題的環境均有適合其計算的資料屬性，如果想要處理的問題屬性能夠與解決問題所提供的環境一致的話，將可以讓問題更快速地被解決。以雲端環境所處理的資料而言，由於資料均會被切分成多分，才能夠分散到不同的節點進行處理，所以對於欲處理的資料而言，建議使用資料關聯性較低的資料比較適合在該環境中處理。舉例來說，對於趨勢科技公司應用雲端環境進行的一些工作 (Ray Liao, Jerry J Wu，2009) ，他們透過雲端環境存放一些記錄檔，而記錄檔案剛好是以每筆為一單位的儲存方式進行處理，所以每行記錄之間可說是獨立的事件，彼此沒有關係，即便將記錄檔以行為單位進行資料拆解、處理也不會影響最後整體資料的結果。

若於地理資訊系統與雲端運算結合的實例而言，目前逢甲大學地理資訊系統研究中心所發展的車隊管理系統 (辜文元，2010) 也是針對結合過程中適用雲端技術條件的部份所進行的開發。車隊管理系統經過長年的運行，累積非常龐大的資料量，這類資料由於資料量過於龐大，往往已經不適合儲存於一般傳統資料庫系

統中。不過，若需要針對歷史資料進行分析處理，又必須將過去累積的資料一併顯示，因此該單位結合雲端技術嘗試解決過去所不容易解決的問題。在該研究中，逢甲大學研究團隊利用了雲端運算的特點去解決龐大資料量的問題。分析該問題，其實最大的困難點並非是在於地理資訊系統顯示的部份，真正的問題點是在資料存放與搜尋的問題上。所以透過該技術的協助，即可將原本在地理資訊系統領域不易解決的問題，透過結合雲端計算的方式將問題排除。

(二) 雲端運算之應用實例

底下針對兩個結合雲端概念所建置出的系統服務進行介紹。在GIS商業軟體中，ESRI公司一直在GIS產業佔有相當重要的地位，因此該公司對於GIS與雲端結合的研究也不遺餘力的進行，期望在雲端上也可以找到GIS的解決方案。

近一兩年，ESRI公司與Amazon公司合作，將其服務直接安裝於Amazon所提供的Amazon Elastic Compute Cloud (Amazon EC2)上，使用者僅需透過租用Amazon的服務即可使用最新版本的ArcGIS Server，如圖7所示 (Sterling Quinn，2011) ，即是ArcGIS Server於Amazon EC2執行的架構圖。

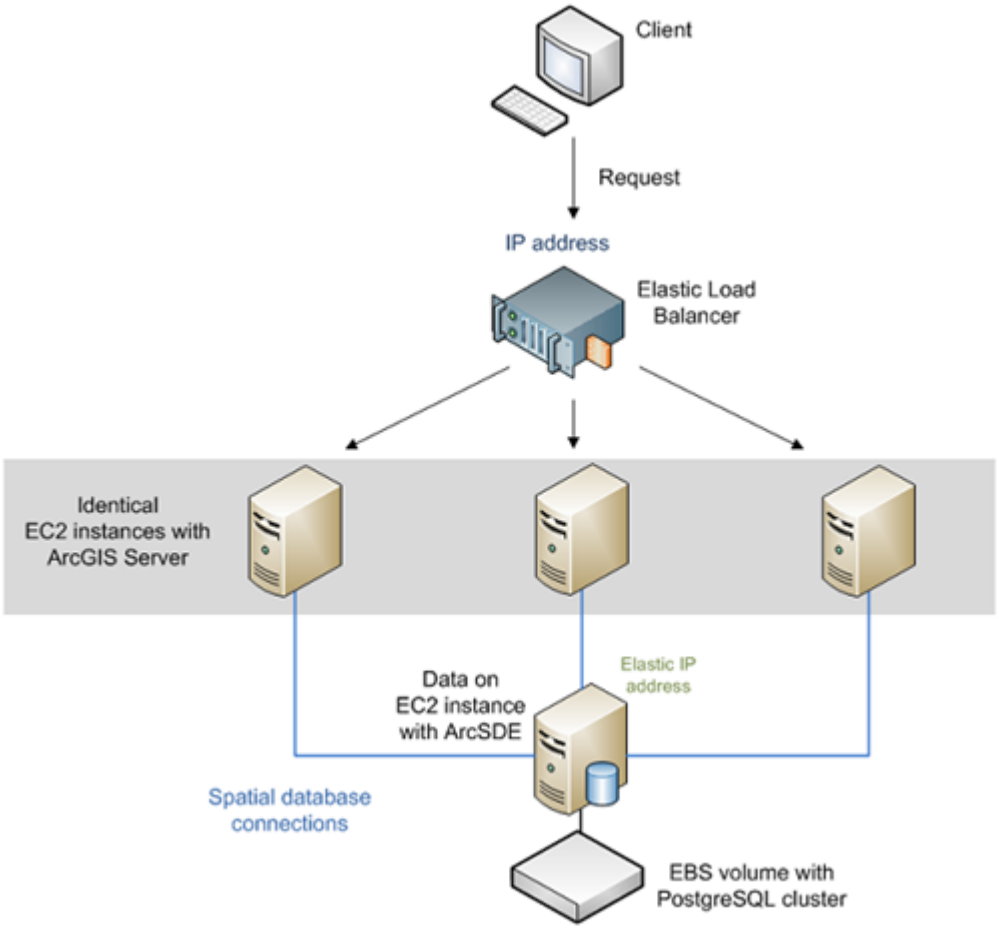

圖7　ArcGIS Server於Amazon EC2上之架構圖

當在ArcGIS Server上資料成長到現有機器無法處理的情況時，使用者只須增加租用的伺服器，讓新增的主機去分擔ArcGIS Server中資料處理的工作，即可輕易解決負載因使用人數過多所產生的問題。

而另一個最近也相當受到注意的服務則是由一家英國軟體公司GIS Cloud Ltd所開發的軟體，其操作界面如圖8所示。

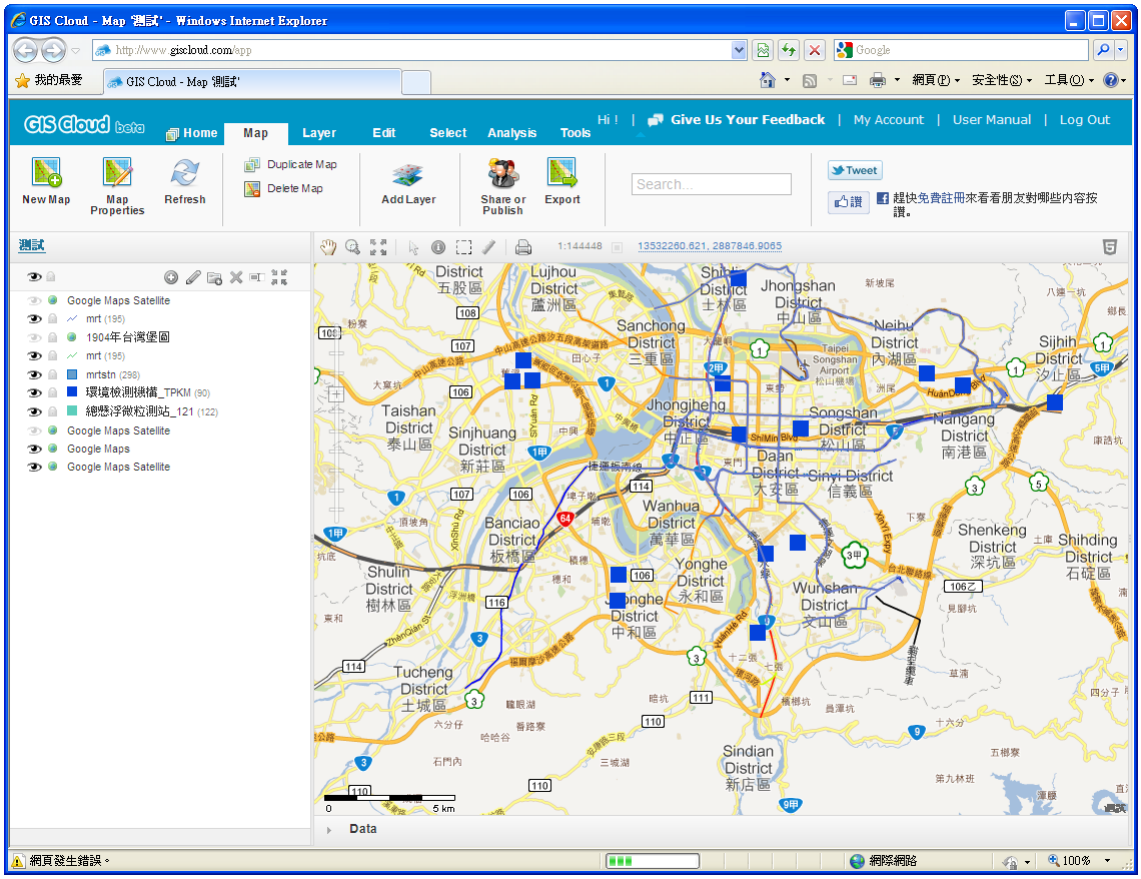

圖8　GISCloud操作界面

該軟體完全以Web Service的方式提供服務，也就是使用者可以將資料完全存放於雲端環境中。此種方式也是類似以SaaS的型式提供服務。在該服務中，使用者可以透過瀏覽器直接在網頁上使用基本的GIS操作功能，如放大、縮小、平移、圖層套疊、...等，甚至對於空間資料庫的連結、向量圖層的進階分析功能（譬如環域分析）也支援。分析此家公司所提供的服務，個人認為更為貼近SaaS的概念，因為對於GIS的使用者而言，完全無須顧慮後端系統資源的分配，GIS專業人員僅需將精神專注於專業領域的工作上即可。

(三) 雲端運算技術於中央研究院計算中心 GIS 組研究情況

目前中央研究院計算中心GIS組延續去年的研究工作「地理圖資共享平台架構之建置與應用」 (彭逸帆、白璧玲，2010) 嘗試將雲端技術整合進共享平台中。因為是先期的研究測試，在各種不同的雲端技術中，期望我們所應用的技術是較為開放的類型。在評估之後，我們選擇以Hadoop作為雲端環境建置的技術，選擇此項技術的考量有底下兩點：

第一，此項技術已經被廣泛應用在商業環境中，因此其穩定性是可以被信賴，同時其技術為開放原始碼，不僅可以避免關鍵技術為特定廠商所持有而不利技術的演進，透過社群的參與更能加速該項技術的更新。

第二，由於是先期的測試工作，為避免因測試而造成資料遺失、外洩的問題

產生，期望透過掌握此項技術可以先行建立僅供內部使用的環境，即希望可以建立一私有雲，待在私有雲測試穩定後再將相關成果放置於公有雲上。

目前中央研究院計算中心GIS組已建立一實驗性質之私有雲，並嘗試將部分資料轉存至該環境中以進行更進一步的測試，期望透過雲端技術的協助可以改善地理圖資共享平台的應用範疇。

結論

對於「雲端運算」的概念，在許多報章雜誌或是相關媒體的介紹下已讓一般民眾有更深一層的瞭解，也體認到此技術將是未來資訊發展相當重要的一環。

在各研究領域均前仆後繼的引入雲端運算技術下，地理資訊科學領域也不例外。然而如同每項應用科學一樣，在應用新的技術之前，必須先分析其特性與適用的範疇，才能有效發揮其能力。

對於地理資訊科學領域而言，部份問題的處理也許因為具有連續性的因素而無法直接以雲端運算的方式解決，不過地理資訊科學領域的研究學者應當可以將研究議題拆解或轉化成符合雲端運算的條件，以引入雲端計算的能量，進而快速解決問題。如此不僅可以透過該技術解決地理資訊領域的問題，甚至可以透過該技術的引入發掘過去所未想見的研究議題與方向。

參考文獻

王耀聰、陳威宇 (2008) ，雲端運算簡介，http://goo.gl/JSQEU

彭逸帆、白璧玲 (2010) ，地理圖資共享平台架構之建置與應用，2010 年亞洲地理資訊系統國際研討會暨台灣地理資訊學會年會、兩岸四地 GIS 與應用遙感研討會

辜文元 (2010) ，Hadoop 於 GIS 上之應用，第二屆台灣 Hadoop 使用者社群會議

Hrushikesh Zadgaonkar (2011), Cloud Computing Concepts and Migration Strategies of an Application to Cloud. http://goo.gl/jlQBN

Lars George (2010), Hadoop on EC2 - A Primer, http://goo.gl/A53PO

Ray Liao, Jerry J Wu (2010), Processing World Wide Domain Information, Hadoop Taiwan User Group Meeting 2010

Sterling Quinn (2011), Deploy a high-capacity ArcGIS Server Web app on Amazon EC2: A case study, http://goo.gl/OS4yi

Tom White (2009), *Hadoop: The Definitive Guide*, O'Reilly