預測分析運用於資料儲存建置成本之研究

朱金城　　趙原宏
國家高速網路與計算中心
{a00chu00,1303024}@nchc.narl.org.tw

彭煥淇
國立交通大學資訊管理研究所
luckystar7810@gmail.com

摘要

本研究透過資料儲存伺服器實際維運及用量記錄，利用該長時期的觀測數據，選擇適當的預測期間，進行成長量的預測分析，尋求最佳預測模式，做為資料儲存建置週期的決策依據，以本研究最佳預測模型為選取最近三個月內的實際用量，以線性迴歸分析實證，可成功試算未來三個月至一年的使用量預測，其精準值在平均百分比誤差為8.57%。

為達到有效控制建置成本，調整建置策略，由一次購置改為分批採購，試算當儲存容量固定以5%持續成長，依照使用量預測值，進而找尋安全的建置量公式，其結果與 Google 以三個月為週期的預測模型[9]相符，可見本研究提出的預測模式確為有效，具實務應用價值。

關鍵詞：資料儲存建置、線性迴歸分析、資料儲存。

Abstract

It is very common for an IT organization to manage capacity plan. The goal of capacity planning is provide cost effectively growing storage capacity to keep pace with data growth.

For three years, we collected detailed usage information for data stored in data servers spanning dozens of clusters in Taiwan. In this paper we propose a predictive model that predicts data storage growth rate from historical trends. This study utilizes linear regression analysis to analyze the disk usages of data server. The forecasting results can be compared with the actual operated data by the mean absolute percentage error approaches. Our results show three-month out forecasts with the errors of less than 8.57% of the short-term prediction.

The easier way of reducing overall purchasing cost is to make sure we are buying at cheapest prices in every possible way. Finally, in our modeling approach we assume that the data storage growth rate is 5%, Experiments have been conducted to evaluate the accuracy of our prediction model in the real world purchasing situation.

Keywords: capacity planning, linear regression analysis, data storage

1. 緒論

從任何領域幾乎所有的資訊及資料收源，橫跨科學研究、民生和醫療，有不同的資料來源，從消費電子設備，如數位相機機和攝影機，醫療設備，如各種形式的基因圖譜掃描儀，電子顯微鏡和廣泛的感測器。

如今科學家對於儲存空間的使用管理，也越來越重視，而對於研究人員來說，資料可能需要保存的時間也愈來愈長，在生命科學及數位典藏等領域，資料可能需要保持 30 年以上。

雲端運算環境的發展，日益穩固成熟的現代，越來越多研究人員願意放下過往對資安、網路連線品質、隱私權的疑慮，進而將大量本地端的研究資料上傳至雲端儲存、運算，而基礎架構即服務(Infrastructure as a Service，IaaS)的供應商在獲得市場大量成長的同時，也面對了關於儲存資源配置的問題。

隨著資料量的增長，在配置儲存資源時，預估使用者的未來一段時間內的使用量是個相當重要的問題，資料儲存的供應商可以藉由預測未來使用者的使用量來提醒使用者正確的時間點對於自己的需求做出修正，例如追加雲端的儲存空間或著是調整資料上傳的計畫。對於資料儲存的供應商而言可以藉由預估未來使用者的使用量推得自己將會需要多少設備，並且於設備價格較便宜的時段大量購買以降低基礎建設建置之成本。

本論文各章節安排如下。第一節緒論說明本研究之背景及研究動機。第二節則針對儲存建置如何有效控制成本進行說明。第三節提出符合資料儲存的需求預測方法。第四節將說明選擇最佳的預測模式，比較其產生的效益並探討結果。最後一節為本論文所獲得的成果。

2. 資料儲存建置成本

在資訊硬體競爭激烈的時代背景下，因為閒置的儲存空間其資產在各研究單位內的總資產額，佔有極為可觀的比例，如何有效管理，是資訊設施營運人員的永續追求目標，藉由追蹤「完整使用行為及用量記錄」，實行用量分析及預測，做為辨識設備採購的最佳時機，及任何有待改善及可能最佳化的機會，包括資料轉換的空間規劃或進而有效提升

設施與資產使用率等等。

對於資訊設施建置及維運方面而言，儲存設施的成本若占 IT 設施的很大一部分，就營運者的角度來看，若可以大幅的精簡在這方面的開支，整體的營運資金方面勢必可以得到更大的緩衝。

基於以上這點，若能以精準的方法預測每一週期的使用量，則一方面可減少採購資金的支出，一方面也較可以減少閒置的空轉設備，積極思考如何避免一次性的建置大批儲存空間。避免設備閒置，造成資源的浪費，並且可以有效節省機房能源。

2.1 建置成本

科技發展趨勢的黃金定律摩爾定律（Moore's Law）[1]，以每 18 個月降價一半的幅度下滑。而儲存設備，則以克來德法測 (Kryder's Law)[2] 每 13 個月同一價格的存儲容量會變成兩倍，並預測未來十年不到的 2020 年時， 史上會出現 2.5 英寸磁碟，且單顆將可高達 14TB 儲存空間，而費用上僅僅為 40 美元左右。這都意味硬體若能在滿足未來成長的條件下，以分期分批採購，就可以取得最實惠的方法建置儲存空間。

通常儲存空間的建置是以年度預算編列的方式，進行一次性的採購，但儲存設備如硬碟，磁帶的硬體，不管售價如何下跌，長期持有資料儲存服務，就僅僅簡單地增加額外的存儲容量不是一個可行的長期策略，因為除建置成本外，仍需要大量的費用開支，包括電源和冷卻系統及機房地板空間等。要處理這些日益增加的資料儲存設施，有效地管理且降低長期總擁有成本（Total Cost of Ownership，TCO），需要 IT 營運者和資料擁有者，具備成本效益的策略[3]包括：配合資料的成長速度，以具備成本效益的方式建置、長期的管理資料存取及保留及最小化成本的建置策劃和管理大量的數據。

2.2 維運成本

儲存設備的總購入成本(Total Cost of Acquisition，TCA)一般佔總持有成本的 20%[4]，而總購入成本，隨著維運時間的累積，伴隨設備故障率提昇，系統的穩定度及維修成本的增加，以致總持有成本將比當初一次性建置的購入費用高出 3 至 4 倍，但又要以多少週期，進行設備擴充或汰換舊設備呢？此篇論文[5]指出當維運進行至第三年後，總擁有成本將逐年增加，如圖 2-1，建議以三年為週期，進行儲存設備汰舊及更換，隨著儲存設備新科技的發展，緒如高容量、高密度的儲存裝置的出現，所以如何顧及儲存成長的需求及選擇設備汰換的時機，在適當的時間進行更換或設備擴充是一件很重要的課題。

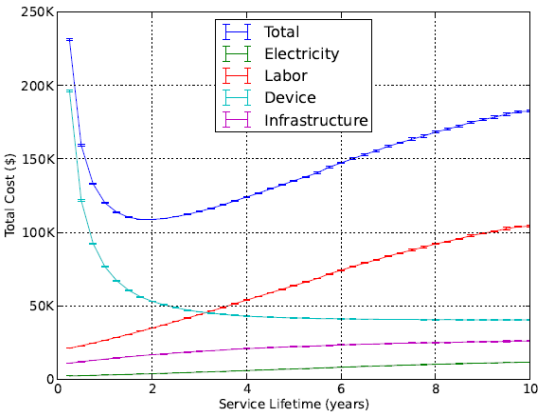

圖 2-1 儲存服務總持有成本

定期的汰換及擴充儲存儲存，才是長治久安的治理之道，所以本文所探討建置的成本，主要是以採取週期性建置儲存設施，其初期所需花費的購置數量做為效益分析，暫不納入維運及基礎設施的固定成本。

3. 需求預測方法

3.1 收集歷史資料

國家高速網路與計算中心[6]所建置的儲存系統，包含磁碟及磁帶儲存設施，整合 IBM、Oracle 及 EMC 等不同廠牌的儲存設施，適用於非結構化巨量資料儲存，且具備歸檔及自動資料搬移功能，該儲存系統為穩定之資料儲存空間，可將異地備份資料存放在具備自動搬遷功能之儲存系統，將使用率不高之磁碟儲存設備，強制遷移至磁帶館。

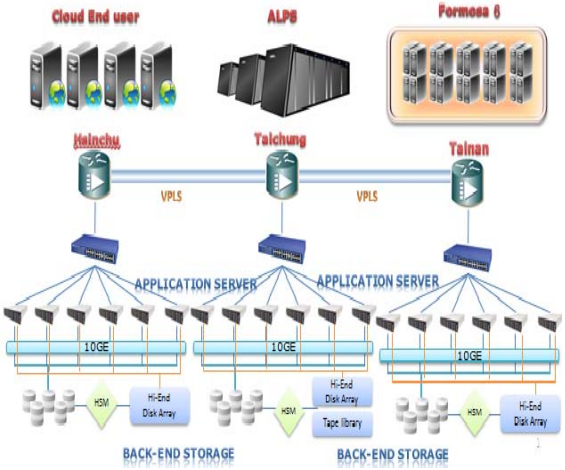

圖 3-1 資料儲存系統架構

整套儲存系統，由磁碟陣列，磁帶館及數台儲存伺服器搭配 IBM TSM[7]及 GPFS[8]軟體所構成，如圖 3-1 所示。該檔案系統可跨越磁碟及磁帶，HPC 數值模擬計算及雲端使用者可以透過該分散式檔案系統，將離線儲存與線上儲存融合的儲存空

間，而使用者的資料最終將會以資料儲存伺服器進行妥善的備份。此次所分析的原始數據，即取自 TSM 資料儲存伺服器。

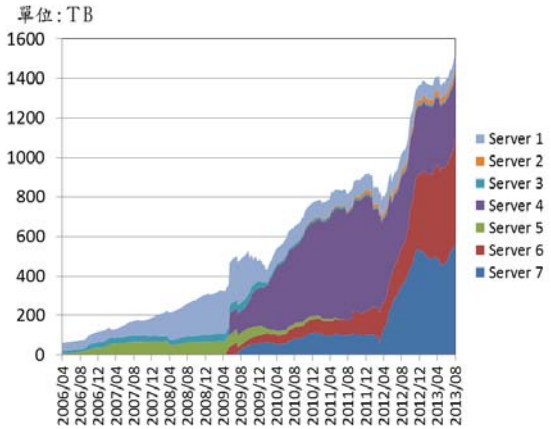

圖 3-2 儲存用量成長圖

如圖 3-2 儲存用量成長圖所示，從 2006 年 4 月至 2013 年 8 月，累積總共七年多原始用量記錄，於儲存維運期間，分別於 2009 年 8 月及 2012 年 5 月，經歷伺服器的更新及儲存媒體格式轉換，當資料轉換時期，可以觀察在不同的伺服器之間，有明顯的資料搬移動作，由於也會額外產生暫存儲存空間，所以，當完成資料轉換後，就會呈現空間負長成的現象，為了彌消在資料搬移過程後資料量巨幅下跌的現象，原始資料經過一次移動平均處理，簡單移動平均(Simple Moving Average)，資料轉換公式如下：

$Sma = (s_1+s_2+\ldots\ldots+s_n)/n$

S_x:第 x 日伺服器中含有的資料量

經過移動平均處理的資料變得相對平滑，可以減少本研究回歸預測時因資料搬移所造成的儲存量離群值。

3.2 預測期間的選擇

本研究旨在利用儲存服務的使用者存取行為之記錄，經過大範圍統計運算(Large-scale statistical computing)後，預估未來一段時間使用者對於資料儲存空間之需求，在得出此需求之後，對設施營運者而言藉由預估未來使用者的使用量推算將會需要多少儲存設備，在增加系統的穩定性之餘，也可以藉由採購計畫降低資源配置的成本。

預測的正確性與預測的長短有著密切的關係，預測所需的時間也不一樣，預測之期間可為短期、中期由長期預測三種，本研究的預測使用對象以研究人員為主，其計算的數值模擬具備時間序列的特性，會隨時間而增加計算題目的複雜程度，也間接增加它資料儲存容量。根據 Google 的用量預測分析[9]指出並不適用超過一年以上的長期預測。因此，為提供一年以內的建置空間規劃，及基於合理採購時程，儲存空間的預測週期，選擇以三個月、六個月及一年以內的中短期預測為最佳，如圖 3-3 所示。

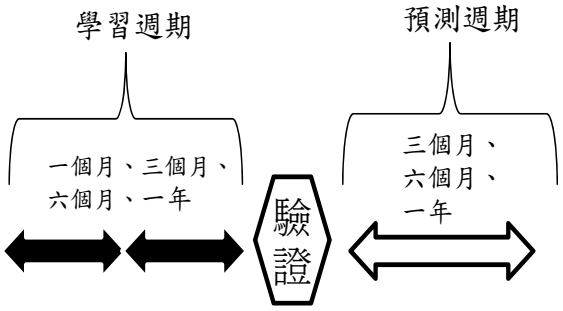

圖 3-3 預測期間

3.3 預算誤差的衡量

所有的預測方法或模型都會有誤差，因此，能讓誤差值保持在一定的水準下，此預測方法或模型就有一定的可信度，而一個優良的衡量指標必須具有正確性、可靠性及容易使用等特睜，以便決策人員可以在最短時間內作出最佳的判斷。

目前用於衡量預測模型之準確度分析方法很多，本研究選擇採用絕對百分比誤差(Mean Absolute Percentage Error；MAPE)來做為模式預測能力的衡量準則。絕對百分比誤差(Mean Absolute Percentage Error；MAPE)：

$$MAPE=\frac{1}{n}\sum_{t=1}^{n}\frac{|F_t-A_t|}{A_t}\times 100\%$$

3.4 線性迴歸預測

為預測未來的儲存空間成長量，考量儲存成長量與時間具備線性關係，適合以迴歸分析(Regression Analysis)進行預測，迴歸分析經常用在解釋和預測兩方面，可經由歷史樣本，計算出迴歸的方程式，再透過迴歸方程式得知每個自變數對依變數的影響力，使用的簡單線性迴歸分析預測法的回歸方程為：

$$\hat{Y}_t = a + bx_t$$

式中，X_t 代表 t 期自變數的值；$\hat{Y}_t$ 代表 t 期因變數的值；a、b 代表線性回歸方程的係數。

a、b 參數由下列公式求得（用$\sum$代表$\sum_{i=1}^{n}$）：

$$\begin{cases} a = \dfrac{\sum Y_i}{n} - b\dfrac{\sum X_i}{n} \\ b = \dfrac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \end{cases}$$

4. 資料儲存預測分析

4.1 選擇最佳預測模式

本研究由歷史的實際用量傳入參數進入學習週期進行資料訓練，再以適合的分析模型，預測出下一週期的預測用量。

每組測試分析，會以一個月、三個月、六個月及一年的實際使用量，以線性迴歸預測方法，算出未來不同週期的預測用量，再以該預測用量與實際用量進行驗證，驗證結果如表 4-1。選其平均絕對百分比誤差最小值之模式為最佳預測模式：

表 4-1 迴歸週期 MAPE 比較

預測模式	預測絕對百分比誤差 MAPE(%)			
	三個月	六個月	一年	平均
一個月	8.15	12.66	25.2	15.34
三個月	8.93	10.69	6.08	8.57
六個月	11.05	12.84	2.27	8.72
一年	8.55	17.93	35.66	20.7

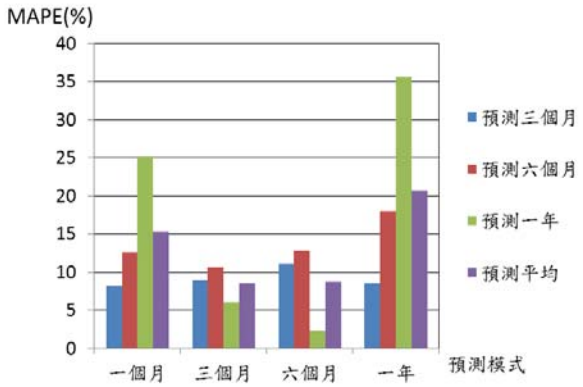

圖 4-1 迴歸週期與測試組合比較

根據表 4-1 及圖 4-1 的分析結果顯示，最佳預測模式為三個月，絕對百分比誤差平均值為 8.57%，其次為六個月，以一個月及一年的預測值較差。最佳預測模式，表現在未來三個月、六個月及一年皆有不錯的預測效果，也最符合實際用量的表現。

4.2 建置量預測

為達到有效控制建置成本，調整建置策略，在驗證出最佳的預測模擬後，依照使用量預測值，進而找尋安全的建置量公式，計算出下一週期的儲存設施採購量。

挑選三個月、一年的這兩組預測模式，選擇一年的預測模式主要是做為對照樣本，比較以年度一次採購及每季或每半年分批採購的可能成本差異。

由一次購置改為分批採購，以採取減少購置儲存設備的成本支出及減低閒置可能，場策略及不同的資金分配策略，進行歷史資料回溯測試

為符合建置量大於未來需求量，我們將採取安全增量逐量遞增估算，將實際的儲存使用量 S 乘上一個安全增量參數，得到一筆新的儲存使用量 S'，使用 S'的資料進行回觀預測的分析，並在分析後進行回溯測試(Back test)，回朔測試中以採購周期(一季、一年)為單位，於第 N-1 期預測第 N 期之儲存設備建置量，並於第 N 期驗證建置之資料儲存設備是否足夠使用，若是不足則調整安全增量參數直到該增量符合所有歷史用量，試算結果，如表 4-2。

表 4-2 安全採購量推算

預測模式	預估安全增量（%）			
	三個月	六個月	一年	平均
三個月	15	50	40	35
一年	40	50	55	48

4.3 效益分析

參考美國聖地牙哥高速電腦中心(San Diego Supercomputer Center，SDSC)於 1997-2009 年八年期間，該電腦中心儲存空間以 15.2 個月成長兩倍的速度為為例[10]，並設訂假設條件為：累積實際使用量以 2000TB 為起始值，儲存容量固定以 5%持續成長。

針對上述的使用情境，在基於不同預測模式、安全存量及採購方式的條件下，試算其一年的儲存建置成本，挑選四組進行試算，如下表 4-3：

表 4-3 模擬條件分組

	預測模式	安全存量	採購方式
實驗組合#1	三個月	15%	每季
實驗組合#2	三個月	50%	每半年
實驗組合#3	三個月	40%	一年
實驗組合#4	一年	55%	一年

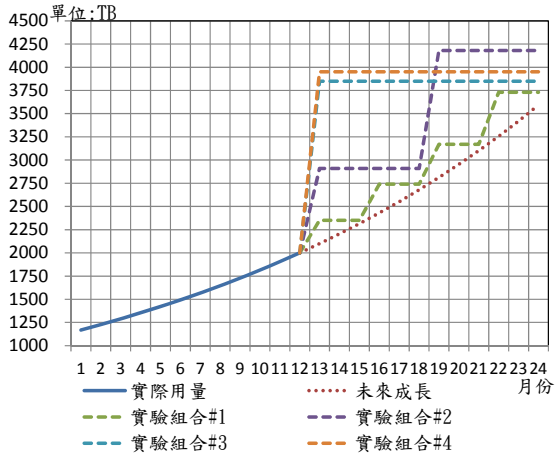

圖 4-2 採購週期曲線

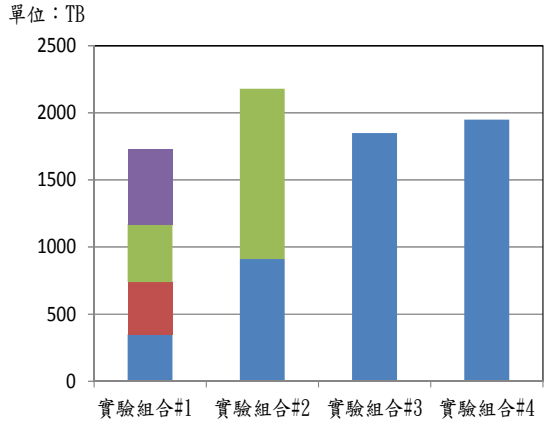

圖 4-3 迴歸週期與測試組合比較

根據圖 4-2 及圖 4-3 試算的結果，發現本方法所預測的採購成本，明顯優於聖地牙哥高速電腦中心以多年期一次性的採購方式。以實驗組合#1 的結果為最佳採購模式，即採取每季分批採購，總購得容量為 1730TB，與以一年採購一次的實驗組合#4 比較，其採購數量為 1950TB，兩者差距達 11%，表示在符合未來一年度的需求下，採取實驗組合#1 的採購方式可以避免閒置的儲存空間高達 220TB，在本節不僅驗證精確估採購容量，亦可合理選擇採購週期及最佳時間點。

5. 結論

利用資料儲存長時間用量分析，估算及找尋最佳資料儲存的建置方法，本研究共分析七部儲存伺服器實際用量資料，以數據分析實證，以本研究模型選取最近三個月內的實際用量，以線性迴歸方式計算，平均可成功預測出誤差率 8.57%以下的未來三個月、六個月至一年的使用量預測，並以該預測使用量，以三個月為單位，推算符合未來需求的儲存空間建置量，將該最佳模式套用在美國聖地牙哥高速電腦中心(SDSC)的儲存空間成長曲線時， 若一年以每季分批增購 15%，即可滿足一年的使用量，與一年採一次的採購方式相比較，可以避免閒置的儲存空間高達 11%。且試結果與 Google 以三個月為週期的預測模型[9]，其結果相符，可見本研究提出的預測模式確為有效，具實務應用價值。

參考文獻

[1] Moore's Law, http://en.wikipedia.org/wiki/Moore%27s_law

[2] Kryder's Law, http://en.wikipedia.org/wiki/Mark_Kryder

[3] Peter McGonigal, James Hill, "Dramatically Lowering Storage Costs", SGI White Paper, 2012.

[4] David R. Merrill,"storage economics", Hitachi Data System White Paper, 2011.

[5] Rosenthal, D.S.H., Rosenthal, D.C., Miller, E.L., Adams, I.F., Storer, M.W. & Zadok, E. (2012). "The economics of long-term digital storage". Paper presented at The Memory of the World in the Digital Age Conference， Vancouver， BC. Retrieved from http://www.lockss.org/locksswp/wp-content/uploads/2012/09/unesco2012.pdf

[6] National Center For High-Performance Computing, http://www.nchc.org.tw

[7] TSM,Tivoli Storage Manager, http://www-01.ibm.com/software/tivoli/products/storage-mgr/

[8] GPFS, General Parallel File System, http://www-03.ibm.com/systems/software/gpfs/

[9] M. Stokely, A. Mehrabian, C. Albrecht, F. Labelle, A. Merchant, "Projecting disk usage based on historical trends in a cloud environment", Workshop on Scientific Cloud Computing, 2012.

[10] Dr. Francine Berman，"A Vision for a New Era in Computational science" ， , http://www.nsf.gov/eng/cbet/workshops/combustion/f07MarSanDiego_HPC_Berman.ppt, 2007.