Kerberos 應用於雲端運算之研究

萬明峰*　孫振東**　蔡昌隆**
*中國文化大學數位機電研究所碩士班　**中國文化大學資訊工程學系

摘要

近年來雲端運算已成為趨勢，Hadoop 是以 Google 的 MapReduce 論文實作之相似的系統，使得人們可以簡單地利用來做大量運算的分散式系統。但在使用方便的同時，資訊安全也越來越重要，本論文主要是在 Hadoop 基礎上討論與設計具增強使用者身份鑑別(認證)的系統。

本論文之實驗系統架構使用 Linux 作業系統之平台，再以 Linux KVM(Kernel-based Virtual Machine)模擬多台電腦之雲端環境，並在此系統上使用 Hadoop 之雲端運算軟體，再藉由 Kerberos 鑑別(認證)機制模擬與實作雲端平台之使用者身分驗證系統。

關鍵詞：雲端運算、資訊安全、Kerberos、Cloud Computing、Linux KVM

1. 前言

在這網路發達的時代，各種裝置的連線以及大量的分散式運算已經變的容易，無論是使用電腦、手機、或平板電腦，都可由簡易的 Client 端連上遠處的雲端服務，例如去 Google 搜集資料，上 Facebook 分享資訊，還有各式各樣的端服務，現在網路和雲端已經是人們生活中不可或缺的一部份了。

雖然雲端運算及雲端服務非常方便，但也有安全性的風險。例如，存放在雲端的資料是否正確，資料是否有安全的保護以避免被破壞，隱私的資料是否有授權才能夠存取，此類的資訊安全問題。雲端平台必須要有安全的環境，使用的客戶及企業才能夠安心的使用。

2. 文獻探討

2.1 雲端技術探討

雲端運算是一種可做大量運算，快速配置且有彈性，只需透過網路即可使用之系統，使用者只要使用服務本身，而可以不必去了解或管理其基礎設施，透過大量伺服器，以達到高效能的資料儲存或運算。生活中就有常見的雲端運算，例如：搜尋引擎，網路 APP，網路硬碟等等。

2.2 雲端服務模式

雲端服務以服務內容又可分為下列幾種模式[1]：

(1) 軟體即服務 Software as a service (SaaS)：提供給使用者完整的應用程式服務，但使用者不須自己開發程式，且不必去理會軟體底層的硬體設備、作業系統、以及網路等架構。比較常見的模式是，提供給客戶一組帳號，以登入使用。例如：Google Gmail、Google Docs。

(2) 平台即服務 Platform as a service (PaaS)：此種類提供給使用者一個平台，如程式開發界面、資料庫、資料儲存空間。使用者可以完全控制程式的運作環境，但不控制硬體、作業系統及網路架構。例如：Windows Azure、Google App Engine。

(3) 基礎架構即服務 Infrastructure as a service (IaaS)：提供了虛擬化主機的服務，使用者可以控制整個作業系統、儲存空間、網路元件，但不包含雲端架構(底層的 host 主機)。通常為配發給使用者後，使用者可以透過網路直接存取。例如：Amazon EC2。

2.3 雲端佈署模型

雲端服務又分為 4 種佈署模型，使用雲端服時以使用者需求來做選擇：

(1) 公有雲（Public Cloud）：公有雲服務可透過網路及第三方服務供應者，開放給客戶使用，公有雲並不表示使用者資料可供任何人檢視，公有雲供應者通常會對使用者實施使用存取控制機制。

(2) 私有雲（Private Cloud）： 私有雲具備許多公有雲環境的優點，例如彈性、適合提供服務，這兩者的差別在於，私有雲的資料與程式皆在組織內，而且不像公有雲會受到外界的影響，也比較沒有安全性的問題；由於私有雲是自己建置，所以可以控制所有的資源，包括安全性和人員的存取。

(3) 社群雲（Community Cloud）： 社群雲是由組織與組織之間所建立的一種雲端架構，是為了安全的交換資料。社群成員可以共同使用雲端資料及應用程式。

(4) 混合雲（Hybrid Cloud）：混合雲結合了公有雲及私有雲，通常在這種模型中，都會將企業的機密資料放在私有雲中，而較普通的非機敏資料則是放到公有雲，此方法可以降低把企業資料存放在公有雲的安全問題。

2.4 Hadoop 雲端軟體

Hadoop 是 Apache 軟體基金會 (Apache Software Foundation) 底下的開放原始碼計劃 (Open source project)，最初是做為 Nutch 這個開放原始碼的搜尋引擎的一部份。

Hadoop 是以 java 寫成，可以提供大量資料的分散式運算環境，可於不同平台上執行。且 Hadoop 的架構是由 Google 發表的 Big Table 及 Google File System 等文章[2][3][4]提出的概念實做而成，所以跟 Google 內部使用的雲端運算架構相似。

2.4.1 Hadoop 叢集

其主要功能可分為兩個部份：HDFS(Hadoop Distributed File System)及 MapReduce，其中，HDFS 負責資料儲存的工作，而 MapReduce 則是負責運算的工作。HDFS 檔案系統大大的簡化了分散式資料儲存的困難度，使得大量資料處理變的容易，但其缺點是處理少量及小型檔案的效率差。

運算叢集的伺服器角色分工可分為 Master 及 Slave，各伺服器執行不同的任務，分為 JobTracker、NameNode、TaskTracker 和 DataNode。JobTracker 負責指揮各機器內的 TaskTracker 程式運作及回傳結果，TaskTracker 程式即是負責執行 Map 和 Reduce 工作；而 NameNode 負責管理 HDFS 的名稱空間、副本管理和資料存取動作，執行的服務如表 1 所示。

表 1　運算叢集伺服器角色所執行之服務程式

Master	Slave
JobTracker	TaskTracker
NameNode	DataNode

有關 Hadoop 叢集之運作及架構概如圖 1 所表示，其中 Master Server 負責管理其他機器的運作。

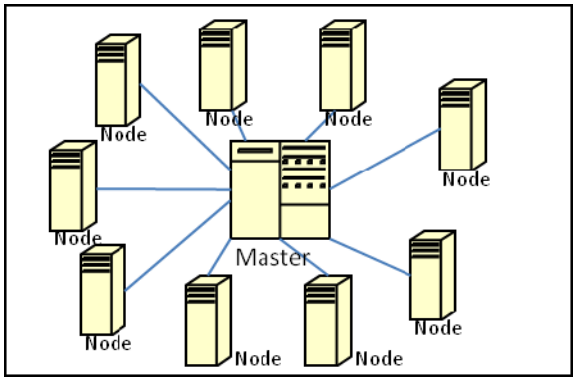

圖 1 Map Reduce 流程

2.4.2 MapReduce 流程

Hadoop 雲端系統輸入需處理的資料後，以 Master Server 來控制其他 Node(Slave Server)作運算，MapReduce 分為以下步驟：

(1) Job 排程

(2) 輸入(Input Data)

(3) Map 階段

(4) Reduce 階段

(5) 輸出結果

承上步驟，檔案輸入後由程式切割成許多小分塊，分別輸入各伺服器作平行處理，如下圖 2 所示。

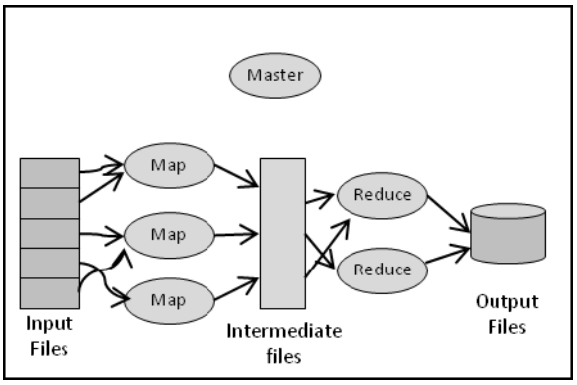

圖 2 Map Reduce 流程

2.5 Kerberos

Kerberos 是由麻省理工學院所研發的一套身份驗證系統，目前最新版本為 Kerberos5。

Kerberos 是一種電腦網路授權協議，可用於不安全的網路環境中，以安全的方式進行身份驗證。軟體設計上採用 Client/Server 結構，並且能夠進行相互鑑別(認證)，也就是說客戶端和伺服器端均可對對方進行身份鑑別(認證)，並且採用對稱式密鑰機制進行管理系統，可以用於防止竊聽、防止 replay 攻擊等。

鑑別(認證)方法是以第三者的角度來提供鑑別(認證)，其構成方式分為 3 個角色：密鑰分發中心(KDC, Key Distribution Center)，使用者端，與伺服器端。KDC 中又包含了鑑別(認證)伺服器(AS, Authentication Server)和票證發行伺服器(TGS, Ticket Granting Server)，且內部包含了一個資料庫儲存了每一位使用者的鑑別(認證)金鑰。

AS 作為使用者身份之認證系統，當使用者發送請求後，如使用者 ID 有效，AS 會發送出一個票證(TGT)和一組 Session Key，而 TGS 則是發行使用者與應用伺服器間之票證。

3. 研究方法與架構

3.1 系統規格

本研究使用了一台電腦來模擬 Hadoop 雲端系統叢集，其中 CPU 必須選擇支援 VT-x 或 AMD-V，才可使用較快的虛擬化技術，否則電腦會以較慢的 QEMU 軟體模擬硬體架構。

電腦之軟、硬體之架構如下表 2：

表 2　伺服器之軟硬體設備

硬體	
CPU	Intel® Core™ i7-3770 3.4GHz
RAM	8GB
軟體	
作業系統	Ubuntu Server 12.04.3 LTS
虛擬機器	Linux KVM 套件
雲端軟體	Hadoop 1.1.1
存取控制	Kerberos V5 krb5-1.11.1

3.2 虛擬系統架構

Hadoop 雲端平台之叢集架構建立在 Linux KVM 虛擬電腦之上，作業系統皆使用 Ubuntu Server 12.04.3 LTS。本研究使用了 4 台虛擬機器執行 Hadoop，分別為 hd1, hd2, hd3 及 hd4，其中、hd1 為 Master，其餘皆為 Slave，而 Kerberos Server 則使用了 CentOS 當作平台。如圖 3 所示：

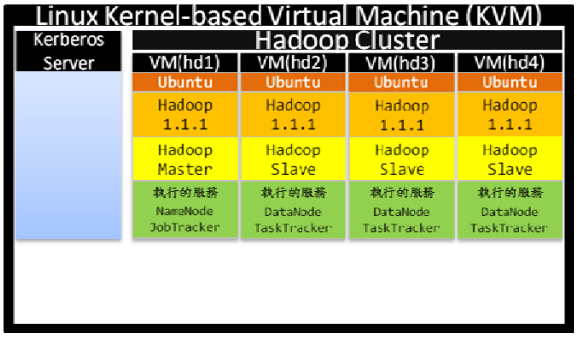

圖 3 實體電腦內的虛擬電腦之架構

由於 Hadoop 可於不同平台上執行，例如：一個公司內有負載較輕的電腦，可使用虛擬化技術增加作業系統來執行 Hadoop，並且跟實體電腦作連結，形成一個大叢集，所以可用虛擬機器來模擬實驗。

3.3 驗證使用者之流程

當使用者欲登入 Hadoop 雲端系統時，必須先經過 Kerberos 鑑別(認證)，帳戶所登入的伺服器為 hd1，亦即 Hadoop Master Server，將經過下列處理步驟，其中，為利描述將以縮寫代表如下：

C：Client，即使用者端。

SS：Service Server，要登入的伺服器(hd1)

KDC：Kerberos Server 內含 AS, TGS

TGT：Ticket Granting Ticket

→：送出請求，回應

下圖 4 為 Kerberos，Client 與 Service 端的運作之流程。

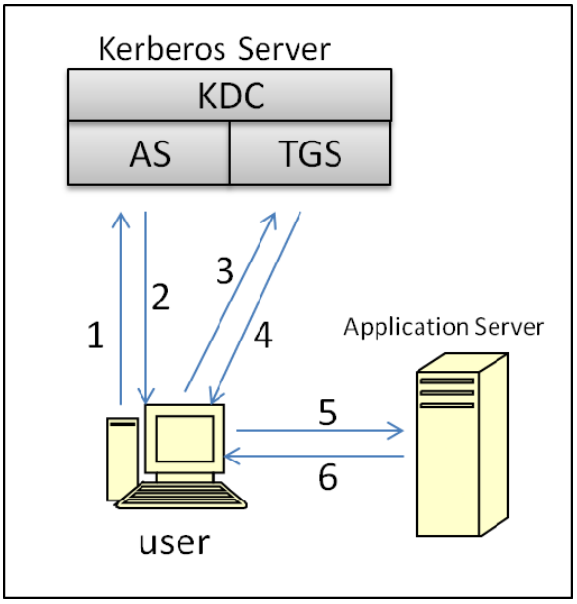

圖 4 使用者登入 Hadoop 之鑑別(認證)流程

其中步驟可簡略分為 6 階段：

(1) C→AS：使用者想跟伺服器通訊時，先對 AS 發送自己身份以請求通行證(TGT)。

(2) AS→C：確認使用者後，AS 將 TGT 及 C 與 KDC 之間的 Session Key 送到 Client 端。

(3) C→TGS：Client 端將之前獲得的 TGT 和請求的服務名發送給 TGS

(4) TGS→C:TGS 送給 Client 端 Ticket 和 C 與 SS 之間的 Session Key。

(5) C→SS：Client 端將剛收到的 Ticket 發送至 Server 端要求服務。

(6) SS→C：開始提供服務。

3.4 系統運作流程

若經過存取控制伺服器驗證成功，使用者即可登入 Hadoop 系統執行雲端運算之程式。

Hadoop 叢集中的伺服器之間也必須經過 Kerberos 伺服器驗證，以防止非法使用者偽裝成合法的使用者去隨意存取系統。

圖 5 表示結合了 Krberos 的 Hadoop 系統運作之流程。

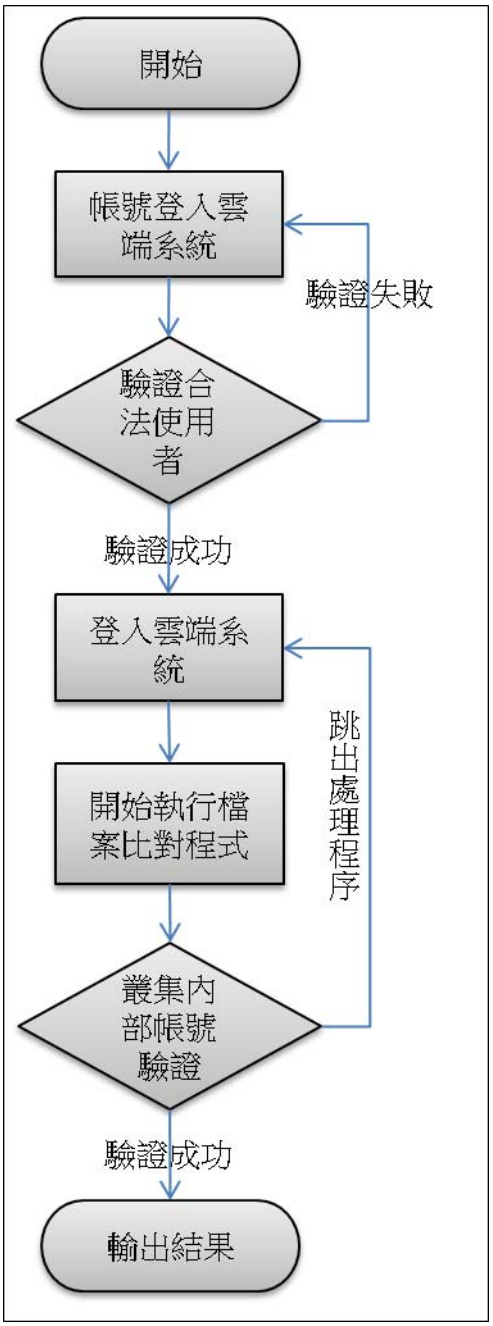

圖 5 執行驗證之流程

4. 結論

雲端平台需具足夠之安全性，方能讓客戶或企業安心的使用。本研究在 Hadoop 雲端系統中加入了 Kerberos 之存取控制，可初步滿足此要求。

本系統所獲之優點如下：

(1) 由於 Hadoop 叢集為大量電腦所組成，導致帳號的更改密碼及帳號停用有一定難度，使用 Kerberos 驗證時可直接於 KDC 資料庫中修改或刪除。

(2) 使用 Kerberos 驗證可防止未經過授權之使用

者隨意存取資料。

(3) 使用 Kerberos 驗證可防止惡意攻擊偽裝成其他帳號執行 Hadoop 服務。

(4) 用此方式不會將密碼傳送到伺服器，可減少密碼被竊取的危險性。

參考文獻

[1] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," 2009.

[2] S. Ghemawat, H. Gobioff, and S.-T Leung, “The Google File System,”19th ACM Symposium on Operating Systems Principles, Lake George, NY, October, 2003.

[3] J. Dean and S. Ghemawat, “MapReduce: Simplied Data Processing on Large Clusters,” OSDI'04: Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, December, 2004.

[4] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, R. E. Gruber, “Bigtable: A Distributed Storage System for Structured Data,” OSDI'06: Seventh Symposium on Operating System Design and Implementation, Seattle, WA, November, 2006.