建構輕鬆管理易於擴充的 Hadoop 運算環境

高鈺棟

麟瑞科技股份有限公司

david_kao@ringline.com.tw

摘要

Hadoop 是 Apache 基金會的一個開源專案計劃，最初是使用 Lucene 的子項目 Nutch 做為搜尋引擎的一部分。Hadoop 是以 JAVA 寫成的，可以提供巨量資料的分散式運算環境，而 Hadoop 它的架構是由 Google Lab 開發的 Big Table 和 Google FileSystem(GFS)的概念實做而成。

關鍵字：Big Data、巨量資料、Hadoop、運算叢集、虛擬化

Abstract

Hadoop is an open source project Apache Foundation project, initially using Lucene subproject as part of Nutch search engine. Hadoop is written in JAVA, can provide a huge amount of data distributed computing environment, and Hadoop Its architecture is developed by Google Lab's Big Table and the Google File System (GFS) concept made real to do.

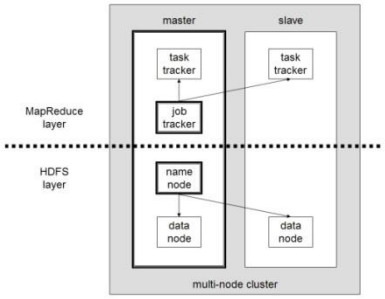
Keyword: big data, hadoop

1. 前言

Hadoop 是 Apache 基金會的一個開源專案計劃，最初是使用 Lucene 的子項目 Nutch 做為搜尋引擎的一部分。Hadoop 是以 JAVA 寫成的，可以提供巨量資料的分散式運算環境，而 Hadoop 它的架構是由 Google Lab 開發的 Big Table 和 Google File System(GFS)的概念實做而成。

2. Hadoop 架構

Hadoop 由許多元素構成，採用 Master/Slave 架構。其最底層元件是 Hadoop Distributed File System (HDFS)，儲存運算叢集中所有 Data Node 上的檔案。HDFS 的上一層是 Map Reduce，是由 Job Trackers 和 Task Trackers 組成（圖一）。



圖一 Hadoop 架構圖

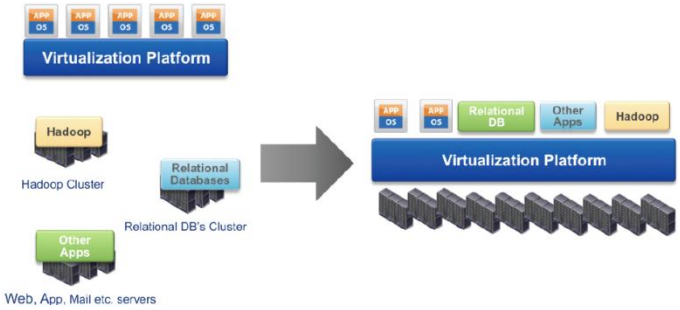
HDFS 就像一個傳統的檔案系統。可以 Create、Delete、Move 或 Rename 檔案等等。但是 HDFS 的架構是基於一組特定的節點建構的，這是由它自身的特點決定的。這些節點包括 Name Node，它在 HDFS 內部提供中繼資料服務。Data Node 為 HDFS 提供儲存。由於只存在一個 Name Node 上，並不具備防止單點故障功能，這是 HDFS 的一個缺點。Name Node 是在 HDFS 中的一台單獨機器執行的軟體。它負責管理檔案系統名稱空間和控制外部存取。Name Node 並不具備防止單點故障功能。

Data Node 是由許多個的節點擔任，一個資料檔會被切割成數個較小的資料區塊，並且儲存在不同的 Data node 上，每一個區塊還會有數份副本存放在不同節點，這樣當其中一個節點損壞時，檔案系統中的資料還能保存無缺。

3. vSphere 高可用度提供 Hadoop 穩定運算平台

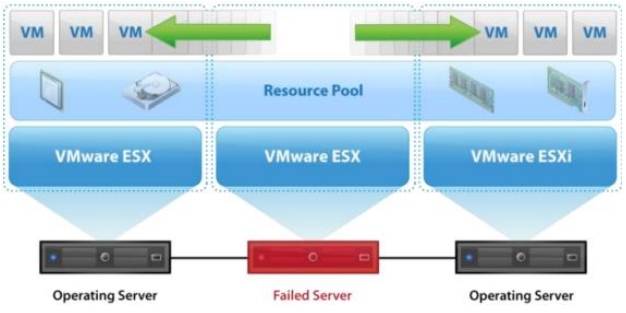
Hadoop 技術做為一個開放源始碼分散式運算平台，近幾年已被大家做為巨量資料運算標準平台，而 Hadoop 需要實體伺服器、儲存設備等專屬硬體，運算能力取決於實體伺服器多寡，當硬體越多其管理及部署、安全性、網路等設計就相對於複雜，部署能力費時又費力，系統調校困難，專屬硬體使用率過低、彼此資源無法共用、專屬硬體投資成本過高，再加上 HDFS 及 Name Node 單點故障風險很高，所以就造成評估規劃巨量數據平台時就會卻步。

Hadoop 採用 VMware 雲端運算平台，透過 VMware 領先業界虛擬化技術，將原先 Hadoop 所需實體伺服器透過虛擬化技術，簡化 Hadoop 所需硬體管理及增加平台高可用度及安全性具備快速佈署、高可用性、絕佳系統延展性、彈性資源分配及利用（圖二）。



圖二　簡化 Hadoop

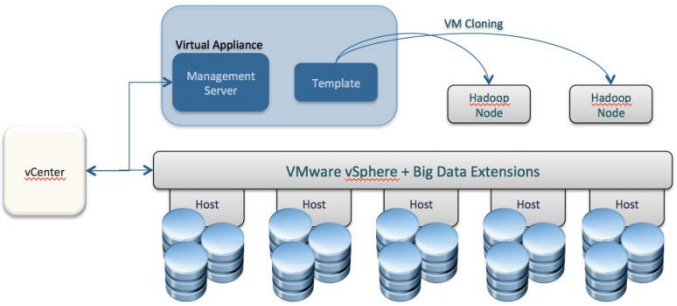
雖然 Hadoop 是透過複製方式提供 Hadoop 運算平台高可用性，包含 Datanode、namenode、job tracker、Pig、Hive、HBase 等元件，VMware vSphere High Availability 能針對 Hadoop 虛擬機提供一致性、自動化保護，建構第一道防線，VMware vSphere High Availability 會監控 vSphere 主機和虛擬機，當偵測到伺服器中斷服務，能將運算叢集故障伺服器所有虛擬機切換至其它未中斷服務 vSphere 主機，並且將虛擬機重新啟動，過程中不需管理人員任何介入，當偵測到虛擬機作業系統中斷時，會自動重新啟動虛擬機，縮短虛擬機停機時間，大幅改善原先運算叢集高可用性可靠度，簡化及加強運算平台保護，另外 VMware vSphere Fault Toerance (FT)提供連續高可用性及 vMotion 功能，將運算平台單點故障及停機時間縮至最短。



圖三 VMware vSphere High Availability 可以解決 Hadoop 高可用性難題

4. 快速佈署 Hadoop 運算節點，增加運算能力

VMware 各種工具，如複製(Clone)、創建範本(Template)和資源分配等功能都可以增快運算叢集佈署速度，而 VMware 近幾年深耕開放源始碼專案及 Apache Hadoop 社群共同合作開發，讓 Hadoop 主要元件有感知虛擬化支援能力，比起 Apache Hadoop 平台擁有更彈性擴充及提高 Hadoop 在 VMware 虛擬平台效能，VMware 於 2012 年 6 月啟動 Serengeti 開源項目計劃，透過此計劃開發 Hadoop 部署管理工具，以讓 Apache Hadoop 管理者可以在短時間內輕鬆部署一個 Hadoop 運算叢集及增加 Apache Hadoop 高可用度能力，日前公佈 VMware vSphere 5.5 版本中提出 Serengeti 計劃針對 Hadoop 平台提供 Big Data Extension 圖型介面操作軟體，使 Hadoop 擁有更彈性擴展能力及資源分配管理，以及增加運算叢集平台高可用度，此軟體被設計為 VMware vCenter 操作介面一個 Plugin，此後 Apache Hadoop 管理者可以透過 VMware vCenter 圖形化管理介面簡單、輕鬆管理 Hadoop 運算平台，讓平台管理者管理 Hadoop 平台不再是難事（圖四）。

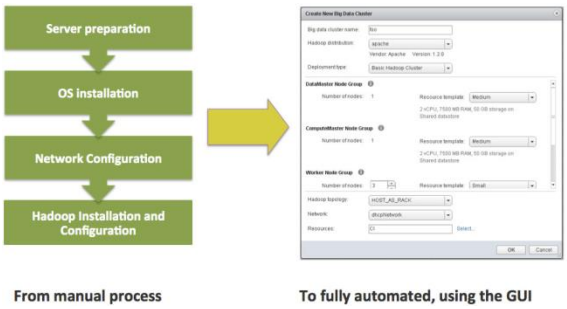


圖四 Big Data Extension 架構圖

5. Big Data Extension 優點

- 輕鬆且快速部署、管理、擴展 Hadoop HDFS、MapReduce、HBase、Pig、Hive、Hive Server
- 彈性擴充運算叢集所需資源，且自動化平衡運算資源、擴展平台以滿足性能 SLA
- 透過虛擬機獨立的數據和節點計算，建構一個真正的多租戶環境透過虛擬化技術，控制資料存取及管理資源

- VMware vSphere 高可用性技術與 Hadoop 高度整合，提供一鍵式為 Hadoop Name Node 及 Job Tracker、Hive、Pig、HBase 輕鬆建置各項元件高可用性架構，增加運算叢集高可用性，避免許多單點故障問題及風險
- 支援 Apache Hadoop、Cloudera、Greenplum、Hortonworks、MapR（圖五）。



圖五 透過 Big Data Extension 視窗介面，部署 Hadoop 運算資源於彈手之間

6. 更有效利用資源

再中斷運算節點， Big Data Extension 資源擴充技術為Hadoop添加動態資源特性，

叢集動態共享資源，依據資源需求動態的擴展，提高資源使用率。

7. Hadoop on VMware 規劃重點

7.1. 伺服器

每台實體伺服器處理器建議最少要配置兩顆實體處理器，每顆實體處理器核心不可以少於 4 核心，並且開啟 Hyper Treading (HT)功能。記憶體部份每核心最少要配置 4GB 以上記憶體，且需預留 6%。網路卡部份建議至少需 4 埠 10GbE 網路卡，兩埠做為運算叢集傳輸用網路，兩埠做為 vSphere 管理用網卡(vmKernel、vMotion、FT、HA、Management)使用。

伺服器上硬碟配置建議使用小容量多顆硬碟配置，Hadoop 運算需要大量 IOPS 需求，大量 IOPS 需要由多顆硬碟所提供，而 SAS 或 SATA 硬碟單顆 IOPS 並不會因為容量變大而跟著成長，所以當空間需求假設為 3TB，300GB SAS 硬碟需要 10 顆(IOPS 為 1750)，600GB 只要 5 顆(IOPS 為 875)，那麼 300GB 配置效能會遠優於 600GB 配置。

虛擬化系統配置

Hadoop 系統架構設計，採用水平擴充(Scale-out)方式，當運算叢集出現效能問題時，管理者只要在其運算叢集添加更多硬體設備，就可以提高整體運算能力，不再像傳統 IT 基礎架構採用垂直式擴充(Scale-UP)方式，但是當 Hadoop 運算叢集虛擬化後，就可以採用 VMware vSphere Hot-Add 技術線上動態調整運算節點處理器、記憶體、儲存裝置，為運算節點自動添加運算資源，提供運算叢集優越垂直擴充能力，再透過 Big Data Extension 軟體技術與 vSphere DRS、DPM 功能整合，從虛擬化底層感知運算節點效能自動平衡運算節點且分散到各個虛擬主機，當運算能力再搭配提供優質良好垂直及水平擴充方式。這樣子的擴充方式可以帶來極大的系統可擴充性，但是會產生另外一個問題，當不再需要強大運算能力時，系統將會自動地將閒置運算節點下線，會將正在運算的任務中斷，已經完成的任務結果也會遺失，運算節點被強制性中斷，將會造成大量任務要重新執行，整個運算能力將會大大影響到，因此 Big Data Extension 運算資源擴充技術會對所有即將中斷運算節點的運算資源設為零，不再接受新的任務，且等待運算節點上的任務都執行完畢，

可以為 Hadoop 或多個 Hadoop 運算

運算叢集(Data Node)與管理用節點(Name Node)。Job Tracker 虛擬機需配置在一個 NUMA 節點上，並且讀取本地記憶體以取得更低的延遲時間。

Data Node 虛擬機分配最少的伺服器上儲存裝置，最好低於三個以下，Hadoop 對於 I/O 效能非常要求，建議每個儲存裝置最好避免設定 RAID，因為設定 RAID 之後會有所謂的 RAID 效能損耗會影響到效能，每一個實體儲存裝置建議建立一個 Data Store，以取得最好的效能。

7.2. 系統配置

Big Data Extension 將會自動設定 Hadoop 作業系統參數以及優化各項參數，如對運算效能十分要求的話，建議 Hadoop 作業系統換成 CentOS6.x 版本，因為 Linux 6.X 的 Transparent HugePage (THP)以及 Extender Page Tables (EPT)，在虛擬化環境會為 Hadoop 帶來不錯的效能。

7.3. 網路配置

Hadoop 運算叢集網路及管理用網路(如 vmKernel、vMotion、FT、HA、Management)建議實體分離，在 vSphere

設定為不同 vSwitch，實體網路交換器建議最好是能夠是分開為不同交換器。

7.4. 橫向擴充建議

當 Hadoop 運算叢集處理器經常維持在 80%以上，建議擴充新的運算節點，另外每個儲存節點空間建議不要超過 24TB，因為內部是採用複製方式提供運算叢集高可用性，所以當某個節點出現問題時，其資料複製會因為資料過大造成網路擁塞，進而影響到整體運算叢集效能。

8. 結語

VMware 參與 Hadoop 開放源始碼組織多年，藉由 Hadoop 底層與 VMware 優異虛擬化技術及效能，為 Hadoop 運算平台帶來更容易管理、更快速部署、更好的高可用性、資源更有效運用，彈性的運算平台，並且透過優秀的開發團隊在 VMware vSphere 針對 Hadoop 平台進行效能優化調校及軟體開發，擺脫許多人認為在巨量運算需求不適合使用虛擬化技術的迷思，VMware 在大中華區有一個專精於巨量運算研發團隊，可以提供更專業的技術咨詢。

參考文件

VMware：http://www.vmware.com
Hadoop 官方網站：http://hadoop.apache.org/
Hadoop Taiwan User Group：
http://www.hadoop.tw/

（作者現任職於麟瑞科技）