

資料整合雲架構設計

Infrastructure of Data Integration Cloud

林哲暉
逢甲大學
jer@gis.tw

何丁武
經濟部水利署
norritho@msl.wra.gov.tw

摘要

本文提出一個架構於雲端運算基礎上，適用於轄下資訊系統眾多的機關一個資料整合雲服務平台，解決資訊揭露不易與降低資料交換的複雜度。隨著雲端環境的成熟，採用雲端運算技術提升資料交換機制的效率，運用雲端運算的高效能及高可用性特性，以滿足機關內具備各式資料乃至未來巨量資料之即時運算、快速回應之需求，同時使用雲端整合之架構將機關內對內、內對外資料交換服務品質進行提升。

Abstract

This research proposes a software architecture based on distributed computing environment, and suitable for governmental agencies. With the maturation of distributed computing, it is able to satisfy the need and requirement of near-real time big data processing as well as increasing the service level by this software architecture. Besides, this research also refine the workflow and user interface, and enhance the usability of the current architecture as well as developing the authentication and authorization mechanism for mobile devices.

一、前言

雲端運算為近期熱門的資訊技術之一[1, 2]，政府機關已陸續導入雲端運算中「基礎設施即服務」(Infrastructure as a Service, IaaS)的虛擬化作為解決方案之一。為響應電子化政府已陸續將歷史、即時、時序性資料進行電子化儲存或處理，例如業務權責產製資訊、環境監控等格式化或非格式化資料，並建立資料交換機制且行之多年，以經濟部水利署為例，時至今已累積水文、水情、河川與水庫等各式巨量資料，然而受限於原先採用之技術及系統架構，無法滿足相關主管機關及單位即時掌握水利資料之需求。

在地球環境變遷影響下，水所帶來的危險越來越不可輕視；在水對我們危機並存之際，惟有在主管機關及單位能全時快速掌握水資料，才能化危險為安全，化機會為經濟，因此利用「資料交換機制」讓相關主管機關及單位掌握水利資料外，應著重在「即時」與「大量」的資料要求，即便在水災等危險情況亦不因大量的資料交換，導致效能低落，甚或無法進行資料交換，致全國所有資料交換引用系統因此停擺，同時也能迅速回應資料運算並交換資料，即便在瞬間大量取得、歷史查詢及民眾應用的

場合，均能確保毫無延遲地得到所需的資料。

水利署雖已建立「資料交換機制」，然而因為先前的技術限制及採行的系統架構，並無法達到「全時」及「快速」之要求。隨著資訊技術的進步，雲端運算科技已臻成熟，且經國際研究及實務經驗指出，其之應用可讓應用者達到「即時」與「大量」之要求。

二、研究目的

資料交換需求在於提供「全時」的 24 小時服務，應用於環境監控或防災單位則會因大量的時序性資料的產生，引發交換頻率暴增導致效能低落，甚或無法進行資料交換，致重要資料交換因此停擺。「快速」的要求重點在迅速回應資料運算並交換資料，即便在瞬間大量資料、歷史大量資料，及民眾大量存取的場合，均能確保毫無延遲地得到所需的資料。資料整合雲服務平台為經由機關訪談、分析設計與系統開發之線上系統，針對機關需求進行使用者介面親合度提升與流程調整，強化平台之便利性與易用性，並與行動裝置進行整合，解決機關與機關間資料系統性交換問題，並以開放格式供應公開資料(Open Data)以促進資料加值應用。有關資料入庫的設計則透過代理 Agent 服務將資料儲存至雲端資料庫 Hadoop，以利用雲端環境的高效能、大容量及高可用性等效益，並提前具備挖掘資料價值成為資料領域知識，另一方面結合雲端運算發展自動化資料分析能力，期使資料整合雲平台從巨量資料儲存體晉身巨量資料分析平台，協助機關藉由資通訊技術邁向巨量智識時代。

三、雲端資料庫之應用

為提供業務單位將常用的文件檔案格式或 PDF 格式與資料整合雲進行整合，避免使用公有雲檔案服務資安隱憂，將業務產生之各種格式檔案以入庫集中管理方式結合資料整合雲上之使用者驗證、授權及容量管制機制，以私有雲架構來確保資料安全性。

資料整合雲依使用者分享檔案之需求，提供上傳檔案入庫至雲端環境儲存或進一步分享給單位內其他使用者，透過檔案連接分享以方便使用者彈性、便捷的操作方式進行雲端應用，參考目前業界廣為使用的 Dropbox 與 Google Drive 架構，作為雲端檔案入庫/出庫整體架構之基礎，並建置使用者端(使用者本機存取)及平台端(使用者連結網頁存取)兩種雲端檔案存取機制與介面。

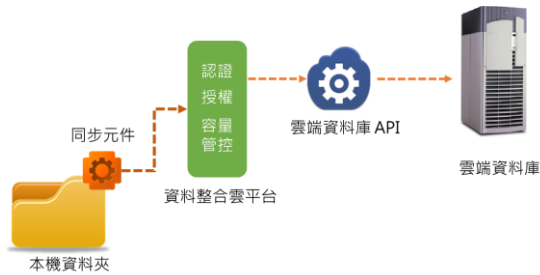


圖 1 雲端檔案分享流程

另一方面則考量資料上稿人員線上登打資料與即時可見的便利性，資料整合雲自動計算彙整特定欄位值，以免去雲端資料上稿人員自行計算彙整並登打欄位值造成的誤植問題。雲端資料庫模組「維護資料表」擴充編修計算彙整欄位功能，使用者執行「新增彙整欄位」開啟「新增彙整欄位運算設定」畫面，使用者自行設定運算公式，並透過系統自動檢核運算式正確無誤後即完成新增「計算彙整」欄位。



圖 2 雲端資料庫欄位計算功能

政府機關公文系統因負責業務範圍較廣致附件資料量逐日增加，對於公文系統進行歷史資料查閱時造成系統瞬間負荷增加。資料整合雲利用「公文附件雲端備份模組」定時讀取歷史公文附件儲存於雲端資料庫，再提供「雲端檔案查詢服務」供公文系統整合歷史公文附件查詢下載，以降低歷史附件儲存與查詢的效能問題。

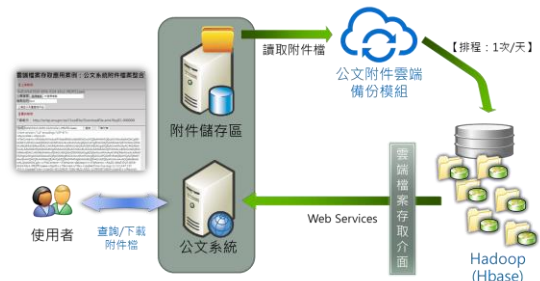


圖 3 公文附件雲端儲存架構

配合行政院電子化政府政策，落實資料典藏目標，俾使政府機關歷年執行計畫成果能永續保存，並且提供未來推動業務時，可知悉歷年的業務辦理狀況，使業務規劃與推動方向能夠一致地往更完善的方向發展。資料整合雲提供「雲端轉換與入庫模

組」將業務計畫成果 PDF 檔案進行線上批次與分散式格式轉換成 PNG 或其他行動式 APP 可讀取之格式，來取代傳統的人工單檔轉換工具耗費大量時間進行轉檔工作。

因各業務計畫結案時，並需將成果報告電子檔歸檔至電子圖書系統，以往使用的檔案格式為 Adobe PDF，而為了使檔案於行動裝置瀏覽時不用整個下載，要再經由轉換程式轉換為 Flash SWF 檔案格式。由於每年執行之業務計畫量相當多，產生的 PDF 檔案量也相當的龐大，傳統的轉換工具需耗費大量時間進行 PDF 轉 SWF 的工作。資料整合雲採用的雲端平台架構，後端使用 MapReduce 框架，特性為分散式處理，將需要轉換之 PDF 電子檔清單整理後，當成 MapReduce 的輸入，由雲端平台將這些待轉換的 PDF 電子檔，分派到雲端平台上的節點，平行分散地進行轉換程序，最後將轉換後的成果歸檔回雲端平台上的資料庫，透過這樣的機制來解決電子圖書系統所面臨的檔案轉換問題。

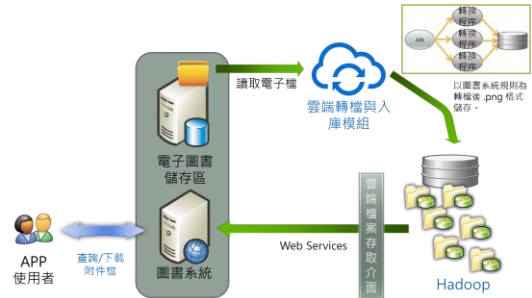


圖 4 雲端轉換與入庫模組架構

四、外部資料交換中心之串接

政府機關彼此間資料交換除了採用傳統 Web Services 方式，環保署另建立了「環境資源資料交換平台」(Central Data Exchange)彙集相關部會與環境相關的資料，並可藉由該平台進行資料發布與接受作業。而資料整合雲為了與該平台進行自動化介接，採用雙向資料交換機制。以常駐背景程式監控環保署 CDX Node 資料夾以即時讀取該平台交換資料以寫入資料整合雲資料庫，之後內部使用時則直接透過 CDX 轉介服務，將前述入庫之 CDX 資料轉供內部各應用系統申請使用。



圖 5 資料轉入流程

水利資料整合雲使用者則可先透過「CDX OutBox 服務管理」功能，選取設定交換至環保署的服務清單，由資料整合雲平台定時自動讀取服務資料並依 CDX 交換格式組成 XML 檔案，放置於 CDX Node 資料夾供環保署讀取以交換至環境資源資料交換平台。



圖 6 資料轉出流程

政府機關另一資料交換平台為「e 政府服務平台」，此平台規範了服務規格。為順利被 e 政府服務平台之服務引用端引用，促進資料流通共享並提升服務交換績效資料整合雲提供符合 e 政府服務平台規範的轉換服務。當服務引用端透過介接服務呼叫服務時，此轉換服務可根據 e 政府服務平台的 XML 訊息交換標準，解析服務請求中的 SOAP 與服務訊息。此服務訊息內部資訊由本平台自訂，故可輕易轉換為對原始服務的呼叫，取得原始服務回傳結果後，轉換服務再將結果轉為 e 政府服務平台格式，最後依原路徑傳遞服務執行結果至 e 政府服務平台的介接服務、服務引用端，完成服務引用流程。

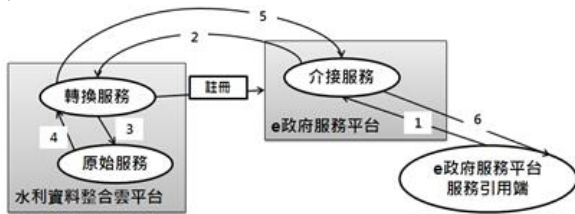


圖 7 e 政府服務平台服務轉換與引用機制

五、資料入庫與 Open Data 機制

服務導向架構最重要的資產就是服務，也唯擁有資料單位將資料透過服務對外做分享，才能發揮資料整合的效果。為了讓這些資料能夠發揮最大的效益，資料整合雲將各資訊系統擬分享與交換的資料匯入雲端資料庫外，並整併平台服務註冊功能及單機版、Web 版服務產生器，從服務註冊、產製、部署整併為單一流程並透過單一介面執行，使用者依流程指示順序執行，即可完成服務自動發佈。



圖 8 服務自動發布

為讓資料整合雲可自動將資料儲存至雲端資料庫，藉由代理 Agent 程式，透過連結資料庫或服務的方式，將關聯式資料庫資料匯轉入庫，鑑於非屬於標準資料入庫欄位對應係由使用者於平台註冊服務時進行設定，而屬於標準資料入庫欄位對應資訊係以 Web Services 方式提供，為使代理 Agent 服務對外的介面統一，提供平台統一控管各 Agent 服務運作現況

代理 Agent 包括設定資料來源模組、資料匯轉模組、接收資訊模組，資料來源模組提供使用者透

過介面設定資料庫或服務來源，以供資料匯轉模組連結資料庫或呼叫服務取得入庫資料之用；資料匯轉模組則定時連結資料庫或呼叫服務取得入庫資料並透過 API 將資料匯入雲端資料庫，匯入資料時若發生資料項不存在情況，需記錄資料匯入異常紀錄並寄送資料匯轉異常通知給平台管理者及資料管理者；接收資訊模組負責接收、解析並儲存資料入庫對應資訊。



圖 9 Agent 服務架構圖

資料整合雲平台上之資料除提供內部資訊系統引用外，針對需要對外公開之 Open Data 資料服務，亦以服務方式提供資料清單，並轉包為 RSS、CSV、HTML、XML、JSON 等格式之，以服務、檔案連結形式提供外界存取利用。



圖 10 Agent 服務架構圖

資料整合雲具備「公開服務資料下載清單」服務，內部網站可先呼叫「公開服務資料下載清單」服務的「公開資料清單」介面以取得符合標準之 Open Data 資料服務清單，當使用者點選下載特定資料項時，再呼叫「公開服務資料下載清單」服務的「檔案下載介面」並傳入下載的資料項代碼及資料格式，即可取得該資料檔案提供使用者下載前述 Open Data 資料檔案。

六、結合行動載具

為讓機關內使用者以行動裝置登入資料整合雲平台，並兼顧保留完整使用者使用內部應用系統之網站或 APP 登入紀錄，提升使用者執行 APP 毋須重覆執行登入之便利性，資料整合雲平台採用二階段式身分驗證功能，使用者一經完成行動裝置驗證後，即由 APP 保存使用者驗證資訊，後續當使用者開啟 APP 時，即由 APP 自動傳送身分驗證資訊及行動裝置識別碼至平台執行單一登入。

行動裝置驗證依使用對象不同提供 Windows Active Directory 驗證及憑證驗證，前者適用於單位內使用者驗證行動裝置之用，使用者於 APP 輸入網域帳號及密碼後登入，由 APP 向資料整合雲平台傳送帳號及行動裝置驗證碼進行驗證，如身分驗證無誤，則完成身分驗證及行動裝置驗證，後者適用於外單位使用者登入 APP 用，使用者需先至平台請求發送行動裝置驗證碼，再於驗證碼有效期間至 APP 輸入該驗證碼完成憑證驗證。

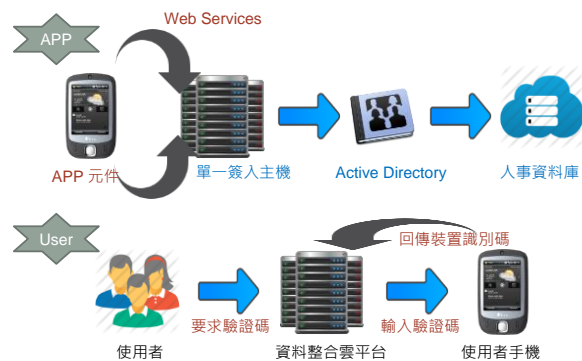


圖 11 行動載具身分驗證流程圖

七、結論與未來方向

資料整合雲以服務導向架構(Service-Oriented Architecture, SOA)為基礎，整併單位內各資訊系統資料與結合資料標準，並利用可擴充性之雲端特性，兼具資料多元性、平台應用性及外部資料交換便利性，妥善整合不同來源與各種格式資料。同時也肩負整個機關資料進、出口重要樞紐，亦提供行動化憑證與單一簽入機制供內、外部使用者進入各資訊系統驗證之用。



圖 12 資料整合雲平台架構圖

政府機關因業務需求建置各類型資訊系統，所擁有之資料藉由雲端水平擴充之特性，將可快速、分散、大量地被儲存至資料整合雲平台，為維持既有雲端環境的高效能、大容量及高可用性等效益，後續應擴充 Hadoop 叢集節點並建立營運監控機制，在運算資源不足時可以藉由 API 來自動建立新的雲端資源。而這些巨量資料具備挖掘價值成為專業領域智識，應於優先採用雲端高效能分析的優勢導入商業分析(Business Intelligence)或大資料(Big Data)的分析模式，從而應用於決策支援或改善既有預測模型等，使資料整合雲平台從巨量資料儲存體晉身巨量資料分析平台。

參考文獻

- [1] B. Sosinsky, Cloud Computing Bible, Wiley Publishing, Inc., 2011.
- [2] Z. Mahmood and R. Hill, Cloud Computing for Enterprise Architectures, Springer-Verlag London Limited, 2011.
- [3] 逢甲大學,水利資料整合雲建置計畫成果報告書, 2013.