美國醫療保健領域對新興資料儲存系統理論「資料湖泊」(Data Lake)的應用

在現今資訊流通快速蓬勃發展的時代，巨量資料(Big Data)帶來效率與生產力等龐大效益已無庸置疑。相較於將資料以「資料倉儲」(Data Warehouse)模式儲存，「資料湖泊」(Data Lake)被廣泛視為巨量資料快速演進的下一步。

美國的醫療保健領域為因應巨量資料發展並提升醫療保健系統的透明度與有責性，美國醫療保險與補助中心(Centers for Medicare & Medicaid Services, CMS)於2013年底建立CMS虛擬研究資料中心(Virtual Research Data Center, VRDC)，讓研究員能夠以安全有效率的方式取得並分析CMS的龐大醫療保健資料。此種資料倉儲模式會對進入的資料預先分類，並整合為特定形式以指導後續分析的方式。缺點在於為讓資料更易於分享，會進行「資料清理」(data cleaning)以檢測及刪除不正確資訊並將其轉換成機器可讀取格式，各資料版本會被強制整合為特別形式，但資料清理和轉換的過程會導致明顯的數據流失，對研究產生不利的限制。有鑑於此，為更有效益的應用巨量資料，Pentaho首席技術官James Dixon提出新的資料儲存理論—資料湖泊(Data Lake)，此概念於2011年7月21日首先被討論於美國《富士比》雜誌中，目前在英美國家公部門和民間企業間已被熱烈討論。

與Data Warehouse最大不同在於Data Lake可包含「未被清理的資料」(unclean data)，保持其最原始的形式。故使用者可取得最原始模式的資料，減少資源上處理數據的必要，讓來自全國各政府機關的資料來源更易於結合。Data Lake主要有四點特性：1.以低成本保存巨量資料(Size and low cost)2.維持資料高度真實性(Fidelity)3.資料易取得(Ease of accessibility)4.資料分析富彈性(Flexible)。儲存超過百萬筆病患資料的加州大學歐文分校醫療中心(UC Irvine Medical Center)即以Hadoop架構為技術建立了一個Data Lake，該中心能以最原始的形式儲存各種不同的紀錄數據直到日後需要被分析之時，可協助維持資料的來源與真實性，並得以不同形式的醫療數據進行分析項目，例如患者再住院可能性的預測分析。

但相對的Data Lake在安全性和檢視權限上也有一定的風險，尤其是醫療保健領域，因為這意味著病患的資料在個資生命週期裡隨時可被取得，因此資訊的取得應被嚴密控制以維持各層級的安全與保障，在建立安全的Data Lake之前，必須審慎考慮誰有資訊檢視權限以及透過什麼媒介取得Data Lake中的資料等問題。

相關連結

- Dan Woods, Big Data Requires a Big, New Architecture, Forbes, July,21,2011
- Rachel Haines, Is the “Data Lake” the Best Architecture to Support Big Data? InFocus- EMC, January 14, 2014
- Edd Dumbill, The Data Lake Dream, Forbes, January 14,2014
- From The Cloud To The Lake: The Future Of Data In Healthcare
- In the Future We Will Store Data Not in a Cloud But in a Lake
- What big data could do for health care
- Finance Committee chairman seeks input on health data transparency

相關附件

- Brian Stein & Alan Morrison, The Enterprise Data Lake:Better Integration and Deeper Analytics, PwC Technology Forecast, Issue 1(2014) [pdf]

許芳瑜
組長 編譯整理

上稿時間： 2014年08月

資料來源：
Brian Stein & Alan Morrison, The Enterprise Data Lake:Better Integration and Deeper Analytics, PwC Technology Forecast, Issue 1(2014), available at http://www.pwc.com/en_US/us/technology-forecast/2014/issue1/assets/pdf/pwc-technology-forecast-data-lakes.pdf (last visited August 11, 2014)
Finance Committee chairman seeks input on health data transparency http://thehill.com/policy/healthcare/209196-finance-chairman-seeks-input-on-health-data-transparency#ixzz3A4O3ysjV (last visited August 11, 2014)
What big data could do for health care http://www.washingtonpost.com/blogs/wonkblog/wp/2014/07/09/what-big-data-could-do-for-health-care/ (last visited August 11, 2014)
In the Future We Will Store Data Not in a Cloud But in a Lake http://www.brookings.edu/blogs/techtank/posts/2014/07/28-big-data-lakes(last visited August 11, 2014)
From The Cloud To The Lake: The Future Of Data In Healthcare http://www.bsminfo.com/doc/from-the-cloud-to-the-lake-the-future-of-data-in-healthcare-0001(last visited August 11, 2014)

延伸閱讀：
Edd Dumbill, The Data Lake Dream, Forbes, January 14,2014 http://www.forbes.com/sites/edddumbill/2014/01/14/the-data-lake-dream/(last visited August 13, 2014)
Rachel Haines, Is the “Data Lake” the Best Architecture to Support Big Data? InFocus- EMC, January 14, 2014 https://infocus.emc.com/rachel_haines/is-the-data-lake-the-best-architecture-to-support-big-data/(last visited August 13, 2014)
Dan Woods, Big Data Requires a Big, New Architecture, Forbes, July,21,2011 http://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/(last visited August 13, 2014)

文章標籤

推薦文章