Feature

雲端安全工具－資料科學
雲端世代的可視性、偵測和保護
Data Science as a Tool for Cloud Security—Cloud Generation Visibility, Detection and Protection

作者：Aditya K. Sood, Ph.D.
Is the director of security and cloud threat labs at Elastica, Blue Coat Systems. His research interests are malware automation and analysis, app security, secure software design, and cybercrime. The author of the book *Targeted Cyber Attacks*, he has also authored several articles for IEEE, Elsevier, *CrossTalk*, ISACA®, *Virus Bulletin* and Usenix. Sood has been featured in several media outlets including The Associated Press, Fox News, *The Guardian*, *Business Insider* and the Canadian Broadcasting Corporation. He has also been an active speaker at industry conferences such as Black Hat, DEFCON, Hack In The Box, RSA, Virus Bulletin and OWASP.

Michael Rinehart, Ph.D.
Is a chief scientist at Elastica, Blue Coat Systems, leading the design and development of many of its data science technologies. He has deployed machine learning and data science systems to numerous domains, including Internet security, health care, power electronics, automotives and marketing. Prior to joining Elastica, he led the research and development of a machine learning-based wireless communications jamming technology at BAE Systems.

譯者: 徐立群, 國立成功大學會計學系,教授、電腦稽核協會編譯出版委員會委員

共享、合作和在任一地點連接雲端應用程式是現今雲端主要特色。然而，雲端安全面臨擴展性的挑戰。與雲端同領域的其他產業同時也面臨一樣的問題，同樣的擴展性問題、數據科學技術已經證明非常成功了。舉例像是網路搜尋、高速金融、大量圖像和視頻處理，甚至像大規模的防禦系統。最近，數據科學技術也越來越多地在內部採用計算機和網絡安全應用程式。毫無疑問地，資料科學可被用作一個核心技術以確保、加強雲端應用程式藉由實施演算法來利用大規模資料探勘檢測威脅。

> “安全的基石是可視性”

有了資料科學，從結構化或非結構化數據識別和萃取關鍵訊息像是資料探勘、機器學習、統計、自然語言處理是有可能發生的。萃取資料可被用於分析和洞察目標環境的數據。圖1強調用於資料科學演算法基石的不同技術。

安全的基石是可見的。對於有效的雲端應用安全性，可見性意味著去了解：

- 員工是使用那些雲端應用程式
- 員工採取的策略
- 員工使用應用程序創建與分發的訊息

一旦實現了這種可見性，惡意內部人員和惡意軟件威脅的檢測和資產的保護就不再具有與雲應用程序進行互操作的安全系統，這有助於警報，自動預防和修復策略。資料科學在實現可見度方面扮演著重要角色。一旦達到可見度，就會面臨檢測威脅的挑戰。對於雲應用，面臨的挑戰是檢測異常用戶活動、駭客攻擊或可能暴露或破壞存儲在雲服務中的信息的其他威脅。這需要有意義的可見性級別來捕獲用戶操作及其訪問的資源。例如，是用戶上傳異常大量的加密文件（例如，ransomware）的嗎？用戶是否查看他/她通常不訪問的異常大量的特定信息（例如銷售聯繫人）？固定使用量的閾值（例如，上傳限制）可以正確地識別大多數異常行為，但可能導致昂貴的假陽性警報或大量錯過檢測。

傳統的安全解決方案並非專為雲端應用而設計的；他們為內部部署系統提供的保護並沒能有效地轉化為雲端版。隨著服務供應商繼續簡化這些功能，資料外洩（有意圖地或意外地）的威脅增加，使資料遺失防護（DLP）成為任何雲端安全解決方案的重要特徵。例如，高級本地DLP系統不了解鏈接語義，因此可能無法識別通過電子郵件發送的鏈接與破壞支付卡行業（PCI）[1]關聯性的文件。原因可能很簡單，因為DLP系統無法辨識它應該遵循的鏈接，或者它根本無法讀取文件或解釋網站上的流量。

問題是資料科學是否可以用作一種機制：

- 確保用戶不會意外公開包含合規性問題的文件
- 防止和修復數據曝光
- 檢測並防範構成內部人員的惡意內部人員、攻擊者或惡意軟件

答案是肯定的；數據科學可以解決所有上面列出的問題。本文討論了雲端安全性如何受益於資料科學的擴展能力，以提供一致和廣泛的雲端應用程序可見性，新的和動態的雲端威脅的可解釋性檢測，以及對雲端服務敏感內容的準確檢測。

圖一：資料科學技術

Source: A. Sood and M. Rinehart. Reprinted with permission.

實現可視性

在雲端應用及相關防護措施中實現即時可視性需要解析HTTP流量以確定：

- 訪問伺服器的使用者帳號
- 使用者所執行的動作
- 被存取或修改的資源(例如檔案)

這些資料可用簽章解析HTTP流量後擷取，最後顯示在記錄事件中，例如John Doe用外部電子郵件分享了文件”passwords.txt”。試想如果要解析HTTPS的異動以增進網路流量的可視性。HTTPS流量的解析可以透過佈署一個通透式代理將輸入流量解密，同時能讓HTTPS流量抵達目的地。例如HAProxy[2]，它是一個開源的代理與負載平衡器，可用以連接Tproxy[3]，Tproxy是傳輸控制協定(TCP)的路由代理，能建立完整、自訂的通透式代理，是為解析HTTPS流量的解決方案。[4]

網路安全通常是用靜態簽章的方式實現可視性，但雲端應用頻繁地改變其網路流量模式(通常是以軟體衝刺的速度，也就是每兩周)，使得人工簽章的發展面臨壓力。而且如果確保一個應用發展無虞是項挑戰，那麼同時確保成千上萬的應用，尤其是當它們整合在一起時，將會更加艱難。這迫使生成簽章的方法要能隨著應用的發展而快速地適用，同時能擴展到眾多使用者使用的應用中。簽章通常是人工建立，這是一個費時過程且更加困難了，因為雲端應用中，機器會將重要資訊加密，例如文件名稱。這會造成問題，因為雲端應用會改變它們的流量模式，簽章損壞重建需要相當的花費。還有一項挑戰是在這麼多使用者使用的應用中都需要獨立的簽章。於是安全性的很明顯的：缺少應用是如何被使用的可視性，最終無力辨識雲端威脅。

然而資料科學方法(例如機器學習、資料探勘、文本分析)能正面面對這項挑戰，藉著自動學習簽章，其達到零假陽性率所需的時間僅占人工建立所需的一小部分。如果簽章損毀，資料科

學技術可以操作在回饋迴圈中自動修復簽章，在短時間內回復可視性。這意味著資安團隊可以確信地預期使用者事件跨越大量雲端應用時有著一致與透徹的可視性。

偵測動態威脅

來自惡意內部人員，攻擊者和天真用戶的雲端應用威脅正在迅速增長，雲端應用程式現在正被用於託管和傳送惡意軟件、建立資料洩露的管道、觸發資料破壞行為，暴露關鍵資訊和劫持帳戶。特定的資料科學演算法在提供高品質的威脅偵測處於強大的地位，且其可見度豐富且有意義的，它們被設計來處理大規模資料分析，並從資料中萃取出有意義的訊息，資料科學可以用作檢測雲端安全問題的工具，因為可以在以下多個方面獲得情報：

- 相關性－在特定的安全分析籃中反應出大量資料有助於確定相關性，以了解攻擊的完整情形。此外，當來自多個不同位置的資料彼此有相關性時，可以較細尺度級別來仔細分析攻擊。
- 透視－大數據的探勘意味著大大地描繪出事件的透視圖，當大量的資料集被探勘時，將更容易透視攻擊的全貌，最終可以獲得更多的情報。
- 基準線—當使用與攻擊相關的特定功能來探勘大數據時，它有助於產生可用於測量給定環境中的攻擊強度或放大的基準線。
- 情境－大數據探刊可提供適合的情報，包括在環境中對特定攻擊的情境和情況做警示。

以下是一個簡單的例子：

- 將用戶（A）的行為用資料科學和機器學習建模，以生成基準線。
- 用戶（A）已經兩到三個月都不再從外部透過雲端分享檔案，但最近分享了一個檔案。
- 因為用戶（A）的行為與之前計算所得到的基準線（機率）有偏差，而發出異常的警報。
- 執行附加安全元件，分析出潛在的威脅的異常現象。例如，深度內容檢測（DCI）分析異常情況，以檢測是否通過文檔洩露任何敏感的合規性相關資料，如個人身份信息（PII），PCI或受保護的健康信息（PHI）。
- 計算出一個風險分數，並依據此風險分數來偵測威脅。

資料科學演算法還可以有意義地整合多個資料來源，以便向組織提供用戶對風險的評估更全面的描繪，隨著輸入訊號的（用戶，應用程序，動作，位置和設備）數量增加，這些演算法會自動橫向擴大。有意義的透視使用者動作的紀錄是為了有意義的威脅偵測，例如，有一個警示是：「John Doe瀏覽銷售聯絡人資料的數量異常的高」，如果John Doe並沒有在銷售人員名單中，對於資訊安全團隊而言這可能是一個很重要的警示。

> “雲端中的“雜訊”的潛力遠遠高於內部部署系統，而這種增加可能會增加昂貴的偽陽性警報。”

資料科學演算法減輕了資訊安全團隊制定出低偽陽性率的異常行為檢測方針的負擔，這是因為他們能夠擴展到開發跨應用程序、行為甚至訊息類別（例如，檔案、資料夾、文件、部落格）的高保真度用戶級行為模型。

建立雲端世代數據遺失防護的解決方案

傳統的安全機制中，主要是透過數據遺失防護(data loss prevention,DLP)系統掃描傳送中的電子郵件以及儲存在伺服器的檔案[5]，這樣的系統可以按照正規表示法、關鍵詞及文件副檔名有效地辨識敏感資訊，有許多傳統的DLP解決方案由Symantec[6]、Fortinet[7]、McAfee[8]、 Checkpoint[9]、Websense[10]、EMC[11]和TrendMicro[12]等公司提供，

使用標準技術來處理數據洩漏。儲存在雲端中的數據與儲存在內部伺服器的數據是不同的，因為員工可以使用雲端從事更廣泛的活動，舉例來說，文件共享服務可以包含大量的短訊息片段(來自網際網路的密碼或本文)；檔案如電子郵件、收據及網路瀏覽紀錄檔等；媒體檔案；尚未標記的敏感文件草稿；以及員工表格和客戶發票等正式文件。

雲端中"干擾"的潛力遠高於內部系統，這樣的干擾會增加昂貴的誤報，數據科學技術可以再評估文檔時利用增加的資訊來解決這個問題，舉例來說，比起存在於瀏覽紀錄檔或者電子郵件原始本文的九位數字，個人健康表格的九位數更可能構成PII，也通過使用上下文，數據科學演算法保持較高的靈敏度、較低的誤報率。

資料科學更進一步擴大了DLP系統可識別的敏感文件範圍，並減少管理工作，例如，數據科學可以使用文檔結構和自然語言程序來檢測為標記的設計文檔及財務文檔，使用數據科學技術來提供更廣泛有效的原始碼檢查，而不依賴於降低總體靈敏度的高度特定關鍵字組合。

最後，儲存在雲端中的數據內容龐大及範圍廣泛，對DLP系統是一大挑戰。在雲端以前，許多使用者檔案儲存於本地，而更重要的公司檔案則是被共享或歸檔，但是雲端的便利性導致員工用來儲存本地存儲的許多文件類型，包括電子郵件、收據、密碼和證書文件、下載的文件和事件日誌，大量的"干擾"導致潛在的誤報來源更多。為了對資訊安全團隊有價值，雲端DLP必須保持並提高檢測敏感內容的能力，而不增加誤報率。[13]

應用自動預防和補救政策

數據科學提高可見性及準確性的好處為資訊安全團隊提供了新機會，去定義自動策略來保護雲端應用內容。及時可見性可用於阻止某些雲端應用程序的操作，當與進階威脅檢測結合使用時，可以自動限制風險高的使用者帳戶，直到資訊安全團隊清理。最後，快速修復也有可能，如果使用者要分享敏感檔案，系統可以自動取消分享。除策略外，細粒狀事件日誌紀錄還為資訊安全團隊提供了增加根本原因分析的潛力，從而有助於發現新的或更廣泛的網路威脅。

結論

部署企業認可的應用程序在本地的情況下，端口和應用程序阻擋的組合已經成功地減輕了各種網絡攻擊，但隨著企業開始使用雲端，這些機制卻變得不理想，現在需要主動保護企業認可的雲端應用程序，仔細檢測和阻止惡意行為，同時促進生產力。數據科學是一種工具，有助於將現有的專家級安全常規和技術擴展到雲端應用的規模和速度。具體來說，它可以改善雲端應用使用者行為的可見性、對潛在威脅的可解釋性檢測以及敏感內容更深入和更廣泛的檢測。這些優點可減少信息安全團隊的負擔，減少誤報，而不會犧牲對威脅的敏感性，並進一步促進自動預防和修復策略的準確使用。

END NOTES

1 SANS Institute, Data Loss Prevention, USA, 2008, *www.sans.org/reading-room/whitepapers/dlp/data-loss-prevention-32883*

2 HAProxy, *www.haproxy.org*

3 GitHub, *github.com/benoitc/tproxy*

4 Turnbull, M.; "Configure HAProxy With TPROXY Kernel For Full Transparent Proxy," loadbalancer.org, 11 February 2009, *www.loadbalancer.org/blog/configure-haproxy-withtproxy-kernel-for-full-transparent-proxy*

5 Elastica, The 7 Deadly Sins of Traditional Cloud Data Loss Prevention (DLP) in the *New World of Shadow IT*, 2014, *https://www.elastica.net/ebook-7sins-dlp/*

6 Symantec, "Data Loss Prevention," 2015,*www.symantec.com/products/informationprotection/data-loss-prevention*

7 Fortinet, "Data Leak Prevention (DLP)," *Inside FortiOS*, 2013,

http://docs.fortinet.com/uploaded/files/1118/inside-fortios-dlp-50.pdf

8 McAfee, "McAfee Total Protection for Data Loss Prevention," *www.mcafee.com/us/products/total-protection-for-data-loss-prevention.aspx*

9 Check Point, "Data Loss Prevention Software Blade," *www.checkpoint.com/products/dlpsoftware-blade*

10 Websense, "Websense Data Security Suite," 2013, *www.websense.com/assets/datasheets/datasheet-data-security-suite-en.pdf*

11 RSA, "Data Loss Prevention Suite," *www.emc2.*

bz/support/rsa/eops/dlp.htm

12 Trend Micro, "Integrated Data Loss Prevention

(DLP)," *www.trendmicro.com/us/enterprise/dataprotection/data-loss-prevention*

13 Elastica, "Cloud Data Loss Prevention (Cloud DLP)," *www.elastica.net/data-loss-prevention*

Quality Statement:
This Work is translated into Chinese Traditional from the English language version of Volume 4, 2016 of the ISACA Journal articles by the Taiwan Chapter of the Information Systems Audit and Control Association (ISACA) with the permission of the ISACA. The Taiwan Chapter assumes sole responsibility for the accuracy and faithfulness of the translation.

品質聲明：
ISACA臺灣分會在ISACA總會的授權之下，摘錄ISACA Journal 2016,Volume 4中的文章進行翻譯。譯文的準確度及與原文的差異性則由臺灣分會獨立負責。

Disclaimer:
The ISACA Journal is published by ISACA. Membership in the association, a voluntary organization serving IT governance professionals, entitles one to receive an annual subscription to the ISACA Journal.

Opinions expressed in the ISACA Journal represent the views of the authors and advertisers. They may differ from policies and official statements of ISACA and/or the IT Governance Institute and their committees, and from opinions endorsed by authors' employers, or the editors of this Journal. ISACA Journal does not attest to the originality of authors' content.

Instructors are permitted to photocopy isolated articles for noncommercial classroom use without fee. For other copying, reprint or republication, permission must be obtained in writing from the association. Where necessary, permission is granted by the copyright owners for those registered with the Copyright Clearance Center (CCC), 27 Congress St., Salem, MA 01970, to photocopy articles owned by ISACA, for a flat fee of US $2.50 per article plus 25¢ per page. Send payment to the CCC stating the ISSN (1526-7407), date, volume, and first and last page number of each article. Copying for other than personal use or internal reference, or of articles or columns not owned by the association without express permission of the association or the copyright owner is expressly prohibited.

免責聲明：
ISACA Journal係由ISACA出版。ISACA 為一服務資訊科技專業人士的自願性組織，其會員則有權獲得每

年出版的ISACA Journal。