Data Science as a Tool for Cloud Security

Cloud Generation Visibility, Detection and Protection

Sharing, collaboration and anywhere access are the prominent features of modern cloud applications. However, cloud security faces scalability challenges. In industries other than the cloud that are facing this same scalability problem, data science techniques have proven highly successful. Examples include web search, high-speed finance, high-volume image and video processing, and even large-scale defense systems. Recently, data science techniques have also been increasingly adopted in on-premises computer and network security applications. There is no doubt that data science can be used as a core technology to secure and strengthen cloud applications by implementing algorithms that can detect threats through large-scale data mining.

> **"The cornerstone of security is visibility."**

Using data science, it is possible to identify and extract critical information from a variety of structured or unstructured data by using techniques such as data mining, machine learning, statistics and natural language processing. The extracted information can be used to perform analytics and to gain insights into the target environment from which data are fetched. **Figure 1** highlights the different techniques that are used as building blocks of data science algorithms.

The cornerstone of security is visibility. For effective cloud application security, visibility means understanding:

- What cloud applications are used by employees
- What actions employees take
- What information employees create and distribute using the apps

Once this visibility is achieved, detection of malicious insiders and malware threats and protection of assets follows from having security systems that interoperate with cloud applications, which facilitates alerting, automatic prevention and remediation policies. Data science plays a significant role in attaining that visibility. Once visibility is achieved, there is the challenge of detecting threats. For cloud applications, the challenge is detecting abnormal user activities, hacking attempts or other threats that could potentially expose or destroy information stored on a cloud service. This necessitates a meaningful level of visibility that captures both user actions and the resources they access. For instance, is a user account being used to upload an abnormally large number of encrypted files (e.g., ransomware)? Is a user viewing an abnormally large amount of specific information (e.g., sales contacts) that he/she typically does not access? Fixed usage thresholds (e.g., upload limits) can correctly identify most aberrant behavior but are likely to result in costly false-positive alerts or a significant number of missed detections.

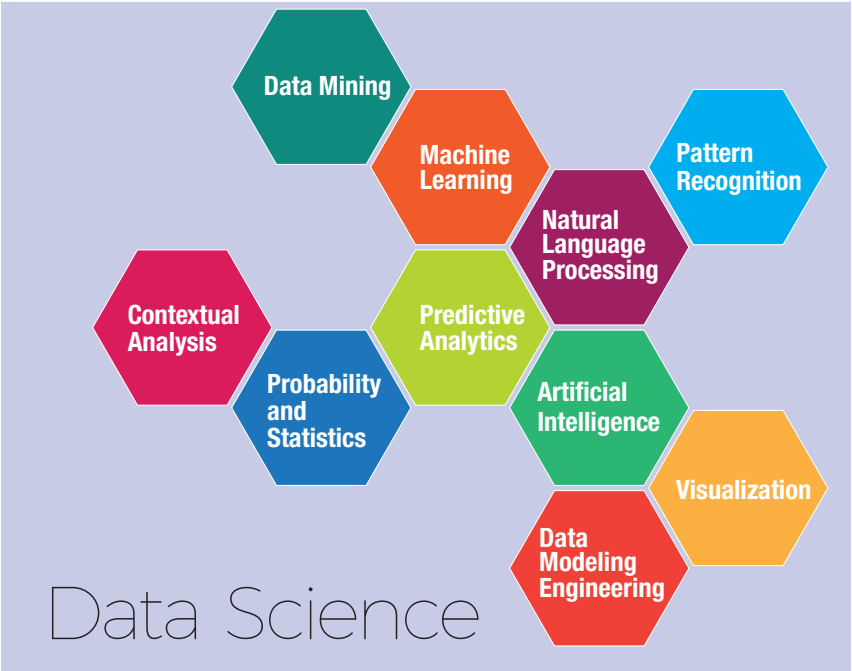
Traditional security solutions are not designed specifically for cloud applications; the protection they afford to on-premises systems does not effectively translate to the cloud. As service providers continue to simplify these features, the threat of data exfiltration (intentional or accidental) increases, making data loss prevention (DLP) an essential feature of any cloud security solution. For example, an advanced on-premises DLP system does not understand link semantics, so it may not

Do you have something to say about this article?
Visit the *Journal* pages of the ISACA web site *(www.isaca.org/journal)*,find the article and click on the Comments link to share your thoughts.

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.

Figure 1: Data Science Techniques



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

recognize that a link sent over email is associated with a file that breaks Payment Card Industry (PCI) compliance.[1] The cause can be as simple as the DLP system not recognizing that it should follow the link or that it simply cannot access the document or interpret traffic from the site.

The question is whether data science can be used as a mechanism to:

- Ensure a user does not accidently expose a file containing compliance concerns
- Prevent and remediate data exposures
- Detect and protect against a malicious insider, attacker or malware posing as an insider

The answer is yes; data science can address all of the listed concerns. This article discusses how cloud security benefits from data science's ability to scale to provide consistent and broad visibility into cloud application usage, interpretable detection of new and dynamic cloud threats, and accurate detection of sensitive content on a cloud service.

Achieving Visibility

Real-time visibility into cloud applications and related protection requires parsing HTTP traffic to determine:

- The user account accessing the service
- The actions carried out by the user
- The resources (e.g., files) accessed or modified

This information can be extracted using signatures to parse HTTP traffic, resulting in a logged event such as "John Doe shared the document 'passwords.txt' with an external email address." Consider the need to parse HTTPS transactions to gain visibility into network traffic. The HTTPS traffic can be parsed by deploying a transparent proxy that decrypts incoming traffic and simultaneously allows the HTTPS traffic to reach its destination. For example, HAProxy,[2] an open-source proxy and load balancer, can be used in conjunction with Tproxy,[3] a Transmission Control Protocol (TCP) routing proxy to build a full, custom, transparent proxy solution for decrypting HTTPS traffic.[4]

Visibility in traditional network security is typically achieved using static signatures. However, a cloud application changes its network traffic patterns frequently (often at the rate of software sprints, i.e., every couple of weeks), which strains manual signature development. And if securing one application as it evolves is a challenge, securing hundreds or thousands simultaneously, especially as they emerge, is much more difficult. This necessitates approaches to signature generation that adapt as quickly as applications evolve, while simultaneously scaling to the wide breadth of applications available to users. Signatures are typically built by hand—a time-consuming process that is made even more difficult by cloud applications that machine encode critical information such as file names. This is problematic because as cloud applications change their traffic patterns, signatures break and it is costly to rebuild them. Adding to that challenge is the

sheer number of applications available to users requiring individualized signatures. The consequences for security are clear: frequent lack of visibility into how applications are used and, consequently, an inability to identify threats in cloud traffic.

Data science methods (e.g., machine learning, data mining, contextual analysis), however, can scale to meet this challenge by automatically learning signatures that achieve a zero false-positive rate in a fraction of the time required for manual construction. As signatures break, data science techniques can operate within a feedback loop to automatically repair signatures, restoring visibility in a short time. This means that information security teams can confidently expect consistent and deep visibility into user events across a large number of cloud applications.

Detecting Dynamic Threats

Threats to cloud applications from malicious insiders, attackers and naive users are increasing at a rapid pace. Cloud applications are now being used to host and deliver malware, establish communication channels for data exfiltration, trigger acts of data destruction, expose critical information and hijack accounts. Specific data science algorithms are in a strong position to provide high-quality threat detection when visibility is both rich and meaningful. They are designed to handle large-scale data analysis and thereby extract meaningful information out of the data. Data science can be used as a tool to detect security issues residing in the cloud because intelligence can be gained on multiple fronts as follows:

- **Correlation**—Mapping large sets of data under specific security analytics buckets helps to determine correlation to understand the complete posture of an attack. In addition, when data from multiple locations are correlated, attacks can be dissected at a granular level.
- **Visibility**—Mining of big data means big picture visibility. When large data sets are mined, it becomes easier to obtain visibility into the attacks, which ultimately results in gaining more intelligence.
- **Baseline**—When big data are mined using specific features related to an attack, it helps to generate baselines that can be used to measure the intensity or amplification of attack in a given environment.
- **Context**—Mining big data may provide more adaptive intelligence, including contextual awareness and situational awareness of a specific attack in the environment.

A simple example is as follows:

- Behavior of user (A) is modeled using data science and machine learning to generate baselines.
- User (A) had not shared any file externally through the cloud for the last two to three months, but recently shared a file.
- The behavior of user (A) raises an anomaly alert with deviation ratio from the generated baseline (probability) computed earlier.
- Additional security components are executed to analyze the generated anomaly for potential threats. For example, deep content inspection (DCI) dissects the anomaly to detect if any sensitive compliance-related data, such as personally identifiable information (PII), PCI or protected health information (PHI), is leaked through the document.

- A risk score is calculated and the threat is detected accordingly.

Data science algorithms can also meaningfully integrate multiple information sources to provide a more complete picture of a user's estimated risk to an organization. Such algorithms automatically scale horizontally as the number of input signals (users, applications, actions, locations and devices) increases. Meaningful visibility that logs user actions allows for meaningful threat detection. For instance, an alert such as "John Doe viewed an abnormally high number of contacts in Salesforce" may be very important to the information security team if they discover that John Doe is not in sales.

"The potential for "noise" in the cloud is far higher than for on-premises systems, and such noise increases the rate of costly false-positive alerts."

Data science algorithms reduce the burden on the information security team to develop policies that can detect aberrant behavior while achieving low false-positive rates. This is because they are able to scale to develop user-level behavioral models across applications, actions and even information categories (e.g., files, folders, documents, blogs) with high fidelity.

Building Cloud Generation Data Loss Prevention Solutions

In traditional security, data exfiltration is addressed by data loss prevention (DLP) systems that scan in-flight emails and files stored on servers.[5] Such systems can effectively rely on regular expressions, key words and file extensions to identify sensitive information. There are a number of traditional DLP solutions provided by companies such as Symantec,[6] Fortinet,[7] McAfee,[8] Checkpoint,[9] Websense,[10] EMC[11] and TrendMicro[12] that use standard techniques to handle data leakage. The data stored in the cloud, however, are different than data stored in on-premises servers because employees use the cloud for a much wider variety of activities. For example, a file-sharing service can contain a vast amount of short information snippets (passwords or text from the Internet); archives such as emails, receipts and network logs; media files; drafts of sensitive documents that have not been tagged; and official documents such as employee forms and customer invoices.

The potential for "noise" in the cloud is far higher than for on-premises systems, and such noise increases the rate of costly false-positive alerts. Data science techniques can address this challenge by leveraging increased information from documents when evaluating them. For example, finding a nine-digit number in a health form is more likely to constitute PII than, say, a nine-digit number contained in a network log or in the raw text of an email. By using context, data science algorithms maintain high sensitivity with lower false-positive rates.

Data science further broadens the range of sensitive documents identifiable by a DLP system, and it does so while reducing administration efforts. For example, data science can detect untagged design and financial documents using document structure and natural language processing. It uses data science techniques to offer broader and more effective detection of source code without relying on highly specific key-word combinations that reduce overall sensitivity.

Finally, there is the challenge posed to DLP systems by the vast size and range of content stored in the cloud. Prior to the cloud, many user files resided locally, while more important company files were shared or archived. However, the convenience of the cloud results in employees using it to store many file types that were once stored locally, including emails,

receipts, password and certificate files, downloaded files, and event logs. The sheer volume of "noise" results in a far greater source of potential false positives. To be of value to an information security team, cloud DLP must maintain and improve its ability to detect sensitive content without increasing the false-positive rate.[13]

Applying Automatic Prevention and Remediation Policies

Data science's benefits of improved visibility and improved accuracy provide new opportunities for information security teams to define automatic policies to protect the contents of their cloud applications. Real-time visibility can be used to block certain cloud applications' actions. When combined with advanced threat detection, at-risk user accounts can be automatically restricted until cleared by the information security team. Finally, rapid remediation can take place as well—if a user were to share a sensitive file, the system can automatically unshare it. Aside from policies, granular event logging provides the information security team with increased potential for root-cause analysis, which can help uncover new or broader threats to the network.

Conclusion

A combination of port and application blocking has been successful in mitigating a variety of network attacks in cases where enterprise-sanctioned applications are deployed on-premises. But as enterprises move to the cloud, these mechanisms become less effective. There is now a need to proactively protect enterprise-sanctioned cloud applications at a level of granularity that detects and blocks malicious actions while facilitating productivity. Data science is a tool that helps scale current expert-driven security practices and technologies to the size and speed of cloud applications. Specifically, it leads to improved visibility into user actions on cloud applications, interpretable detection of potential threats, and both deeper and broader detection of sensitive content. These benefits reduce the burden on information security teams by reducing false-positive alerts without sacrificing sensitivity to threats, and they further facilitate confident usage of automatic prevention and remediation policies.

Endnotes

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-with-tproxy-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/information-protection/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/dlp-software-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/data-protection/data-loss-prevention*
13 Elastica, "Cloud Data Loss Prevention (Cloud DLP)," *www.elastica.net/data-loss-prevention*