Paper SAS670-2017

Data Management for Cybersecurity

Alex Anglin, SAS Institute (Canada) Inc.

ABSTRACT

As an information security or data professional, you have seen and heard about how advanced analytics has impacted nearly every business domain. You recognize the potential of insights derived from advanced analytics to improve the information security of your organization. You want to realize these benefits, but also to understand their pitfalls.

In order to successfully apply advanced analytics to the information security business problem, proper application of data management processes and techniques is of paramount importance. Based on professional services experience in implementing SAS® Cybersecurity, this session teaches you about the data sources used, the activities involved in properly managing this data, and the means to which these processes address information security business problems. You will come to appreciate how using advanced analytics in the information security domain requires more than just the application of tools or modeling techniques.

Using a data management regime for information security concerns can benefit your organization by providing insights in to IT infrastructure, enabling successful data science activities, and providing greater resilience by way of improved information security investigations.

INTRODUCTION

Every day we hear and read about how advanced analytical techniques are being used to solve business problems across practically every conceivable domain. Organizations are constantly reporting that they are experiencing information security compromises. Every industry is a potential target of malicious actors seeking to compromise organizations for a variety of reasons; no industry is immune. Traditional information security systems offering perimeter protection (such as firewalls) or endpoint protection (such as antivirus programs), while still necessary, cannot be relied upon for complete protection. Security systems that rely upon signatures of malicious activities can show known attack vectors, but fail with new and novel attacks. The greatest challenge that organizations face from an information security perspective is that whereas a defender of information systems must ensure perfect security all the time, an attacker needs to be successful only once to achieve their goals. Organizations must adopt new mindsets to confront this threat: They must focus on recognition, resiliency, and response to maintain information security.

Advanced analytics has a strong role to play in supporting these areas, and with it proper data management is a necessity. The current state of recognition in the field means that many incidents are not recognized until weeks or months after the initial compromise. Advanced analytics can shorten this lag time down to hours or minutes. Resiliency requires insight into information systems activities and communications: Analytics can guide remedial actions by showing the breadth and depth of a compromise. This in turn can guide response actions by information security personal so that a thorough response is executed and necessary remedial actions are taken.

In the information security field, there are no panaceas, and advanced analytics is no exception. Despite that, organizations can greatly improve their security posture by adopting data-driven analytical techniques. In doing so, they must be cognizant of the need to manage the data that they will use to recognize that an incident has occurred, and then to respond to it.

UNDERSTANDING THE PROBLEM

Applying analytics to information security is a nascent field because modern technologies are only now able to address several of its challenges. The three Vs of big data (volume, velocity, and variety) apply to the information security field more so than to almost any other: Information security professionals struggle to capture and analyze all the alerts generated by security tools as well as the relevant data sources for

their investigations of incidents. Data is generated in constant streams, supplemented by additional outputs from information technology processes. This data comes from all parts of their organization, as well as from external sources.

Without the application of modern data processing technologies, security professionals are left to their individual understanding and intuition of the situations they encounter. Consistency in analysis across or between teams is not guaranteed. Information security professionals will rely upon the toolkit they have at their disposal without the ability to have a holistic view of the state of their organizations. The tools they use often generate a volume of alerts that is beyond the time availability of security teams to review. They will look to solutions that seek to inform them of the scope and impact of incidents in a way that doesn't overwhelm them with information but that instead provides guidance as to how best to use the limited time and resources available to address and rectify security incidents.

DESIGNING A SOLUTION

Analytics for information security purposes requires a sound design and implementation to be effective. In order to be successful, the system must incorporate a thorough understanding of the data. The system must also describe the environment of the organization and the limitations that are associated with it must be understood. This will enable the team to explain the operation of the system to stakeholders, and in doing so set appropriate expectations for results derived from it. While there are many potential data sources that can be used for security analytics, SAS® Cybersecurity operates with a focused view of several of them, rather than using all available data sources. In the incorporation of various data sources to SAS Cybersecurity, we endeavor to have them be as comprehensive as possible about the topics they describe.

NETFLOW

The core data processing operates on NetFlow data (https://www.cisco.com/c/en/us/products/collateral/ios-nx-os-software/ios-netflow/prod_white_paper0900aecd80406232.html). NetFlow is a protocol developed by Cisco that describes network communications. As opposed to packet level data or host based logs as shown in Figure 1, NetFlow provides metadata about communication within an organization. It does so by providing information about the communications between computers on the network based on their Internet Protocol (IP) addresses. Included is information about the ports used, the length of the communication, amount of data transferred, and other networking information. This can easily provide billions of records per day for processing where unsampled data streams from networking equipment or collectors are provided. Compared to the content of the communications, processing this data holistically across an organization's network is viable with current technologies. Processing the content of the communications, or the packets themselves, is typically limited to specific hosts on a network via tools that identify traffic patterns or that limit the period of the analysis.

NetFlow is available in different versions, though the main ones currently implemented are Version 5 and Version 9. The Internet Engineering Task Force (ITEF) has created a standard for transferring flow information: The Internet Protocol Flow Information eXport (IPFIX) standard, which is based on Version 9 of NetFlow. Vendors in addition to Cisco have also implemented equivalents where IPFIX is not available in their networking equipment.

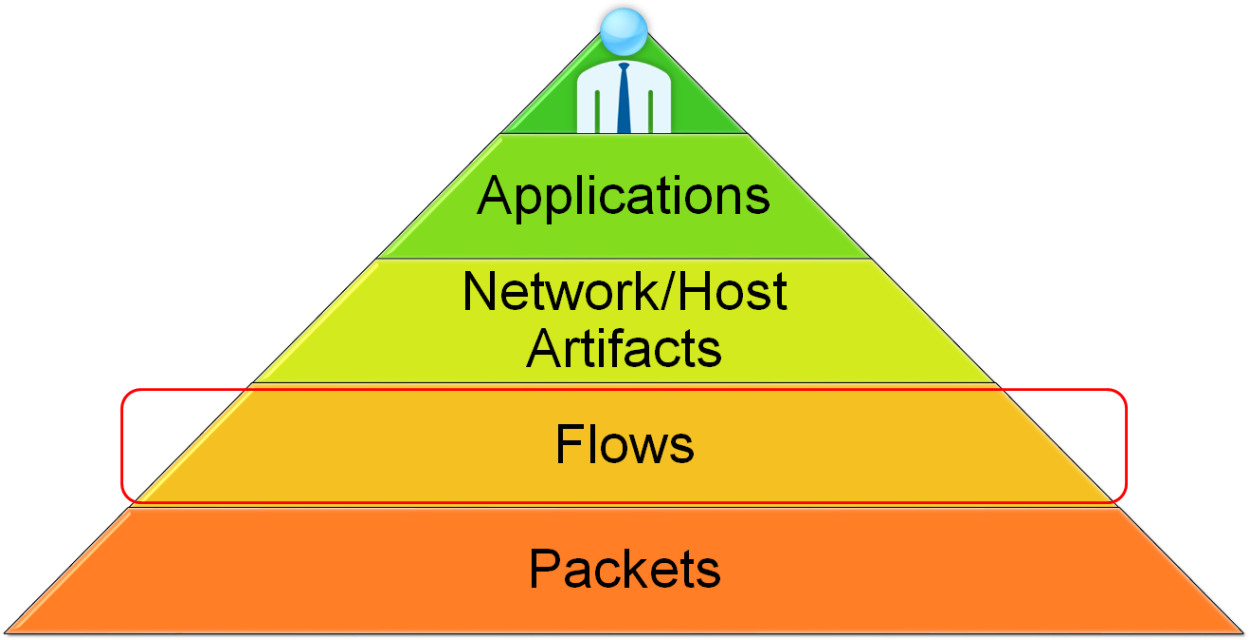

Figure 1. Flow Illustration

AUTHENTICATION EVENT DATA

Reviewing information about communications in a network is of limited use by itself. To gain insight into the communications within an organization, it quickly becomes important to understand the individual users or user accounts who are associated with the network traffic. To do so, you need a data source that provides this association. Corporate directories provide this information as they act as a central repository for user credentials. When a user logs on to their computer or accesses systems that rely upon the directory for authentication, the event is logged and the authentication event, the source of the event, and the user ID that performed this event is tracked. Having the source of the event in the data stream enables real-time joins with NetFlow data, which provides an association between user IDs and IP addresses.

<14>Dec 18 19:23:03 DOMAINCONTROLLER.company.com MSWinEventLog 1 Security 11195349 Sun Dec 18 19:23:03 2016 4624 Microsoft-Windows-Security-Auditing N/A N/A Success Audit DOMAINCONTROLLER.company.com Logon An account was successfully logged on. Subject: Security ID: S-1-0-0 Account Name: - Account Domain: - Logon ID: 0x0 Logon Type: 3 Impersonation Level: Impersonation New Logon: Security ID: S-1-5-21-1111111111-111111111-111111111-11111 Account Name: USERID Account Domain: COMPANY Logon ID: 0xFFFFFFFFF Logon GUID: {FFFFFFFF-FFFF-FFFF-FFFF-FFFFFFFFFFFF} Process Information: Process ID: 0x0 Process Name: - Network Information: Workstation Name: Source Network Address: 192.168.101.101 Source Port: 52465 Detailed Authentication Information: Logon Process: Kerberos Authentication Package: Kerberos Transited Services: - Package Name (NTLM only): - Key Length: 0

Figure 2. Example Active Directory Authentication Event

NETWORK

Understanding the nature of an IP address requires context around its use in the corporate network. Networking teams design the layout of IP addresses such that specific ranges or subnets are associated with a geolocation and segment of the network. These segments in turn can be allocated for purposes such as servers in a data center, networks at specific offices, or systems hosted by cloud providers. Gathering this information requires close collaboration with the teams responsible for the design and management of the network. Their subject-matter expertise and knowledge of the systems they use to track the allocation of IP addresses on the network is necessary for gathering and validating this data.

USER AND ORGANIZATIONAL STRUCTURE DATA

An understanding of authentication events provides clear value when correlating individual accounts with network traffic. However, to gain insight into the observed behavior of individuals, it is helpful to have greater contextual analysis of their position and role within the organization. Typically, we expect the network behaviors of information technology staff to differ from that of administrative staff.

Data about an organization is normally the easiest information to gather. All organizations have administrative and technology needs that require good data quality as it pertains to their employees and the structure of the organization. This in turn is reflected in Human Resource Information Systems (HRIS)/Human Resource Management Systems (HRMS) and in corporate directories such as Microsoft Active Directory. HRIS/HRMS systems generally have database back ends or APIs through which data about the organization can be gathered. Corporate directories can be queried using the Lightweight Directory Access Protocol (LDAP).

IT ASSET INFORMATION

Assets describe all the devices connected to the corporate network, and their type and function. This includes devices such as servers and hosted virtual machines, as well as devices such as networking equipment, printers. Data about devices can be gathered from systems such as configuration management databases (CMDB), asset inventories, or Microsoft Active Directory (for Windows servers). However, even when these systems are implemented, it is often difficult to ensure that they remain complete and up-to-date. With IT challenges such as the Bring Your Own Device (BOYD) and Internet of Things (IoT) trends, IT organizations face additional challenges in understanding the devices on their network. With the adoption of public and private cloud platforms, it is also difficult to track servers in those environments: Many can be ephemeral for the purposes of addressing capacity needs or for development purposes.

Teams tasked with gathering this data should quickly seek not only to understand the scope and depth of the data they can gather for this purpose, but also to understand its limitations. Data from other sources described above can be used to identify devices that are not tracked by systems that are used to gather asset information. For example, segments of the network can be allocated for specific uses such as Voice over IP (VoIP) devices. For the purposes of comparative analysis, devices on these segments can be assumed to be used for that purpose. Analysis of the traffic from devices can reveal information about their nature: For example, VoIP devices use protocols such as Session Initiation Protocol (SIP) that can be recognized from the communications described in the NetFlow data. Analysis and utilization of this information is recommended to be a cyclical set of activities rather than a single phase: As the data management team gathers and validates this information with subject matter experts in the IT organization, this information can be fed back into systems that track assets in the organization, thereby enabling a virtuous information cycle that informs the IT organization about the devices on their network.

WEB PROXY DATA

Web proxies are devices or applications that provision external access to Internet resources by acting as an intermediary to communications that use the Hypertext Transfer Protocol (HTTP). As they sit between the communications of users browsing websites and the websites themselves, web proxies can be configured to track various information such as the uniform resource locator (URL) being requested, data being transferred, the user ID of the account making the request, and whether the website being accessed is compliant with acceptable use policies. This information can be used by security analysts to identify issues of malicious websites being visited or incidents of data exfiltration following a compromise.

THREAT INTELLIGENCE

Threat intelligence data describes IP addresses on the Internet that have been observed to have exhibited malicious behavior such as the operation of botnets or hosting phishing websites. They are provided by a variety of organizations that offer this information on a free-to-use basis, on a commercial basis, or on a consortium basis such as those provided by various Information Sharing and Analysis Centers (ISAC) (https://www.nationalisacs.org/). In each case, the data is gathered via observation points

across the Internet that seek to detect and correlate indicators of malicious activities such that the organization consuming the threat intelligence data can in turn recognize security incidents more quickly.

Using the Structured Threat Information Expression (STIX) standard (https://oasis-open.github.io/cti-documentation/), threat intelligence feeds can describe a variety of topics related to security incidents including threat actors and the tactics, techniques, and procedures (TTPs) of attackers. Despite this rich description of security-related topics, most threat intelligence feeds typically focus on file hashes of known malware and IP addresses associated with misuse such as malware hosting.

ACTING

OBJECTIVE—IDENTIFICATION OF INCIDENTS

Together, the data sources in the previous section describe behavior and context of an organization's network. Through the integration of those sources, information can be provided to security analysts that assists with guiding their day-to-day activities toward ensuring the information security of their organization. The key objective is to quickly identify incidents to ensure that they are addressed as fast as possible to mitigate their adverse effects. By providing this information as well as the tools to analyze it, security analysts can gain greater insight into issues they are investigating. They can also identify of incidents that previously would have remained unrecognized.

DEFINITION OF NORMAL

Establishing an analytics system for information security is not a simple matter of build and deploy. While continuous refinement is advised to address the ever-changing landscape of IT infrastructure, an initial baselining phase is also required. This phase goes beyond typical testing activities such as unit testing, integration testing, and system testing. Instead, it involves the operation of the system not only for the sake of identifying security incidents (though that can still be addressed through findings from the system), but also for validating the results of the analytics. At this phase, the team should be concerned with aspects of the analytics that lead to false positives such as incorrect categorizations of individuals or devices. The team should also review the completeness and accuracy of the results to ensure that the scope of the information provided to security analysts is complete based on the data processed by the system, and that users and devices are being correctly reflected in the results.

ALERTS

The activities of information security analysts can benefit from receiving all this information, but just providing it in a raw form doesn't realize the full potential of advanced analytics. To do so, the data described earlier must be passed through a model that provides insight for users that they would not necessarily be able to ascertain themselves and in doing so generate alerts that are actionable or investigation-worthy. SAS Cybersecurity does this by using the deviations of a combination of metrics derived from the data consumed by the system. With a composite metric, security analysts can focus on aspects of the risk alerts to understand their composition and in turn to make decisions as to what actions to take in response to the information provided.

INVESTIGATIONS AND INCIDENT RESPONSE

Analytics by itself cannot ascertain whether information provided by the system is a result of malicious acts. To do so, security analysts need to conduct investigations to understand the nature and impact of an incident. In doing so, they rely upon information provided by a variety of tools, depending on the nature of the investigation. Security tools of all types generate alerts for security analysts to review and act upon. The analysts are deluged by alerts and receive many more than they can act on, much less come to definitive conclusions about as to the nature and scope of a potential incident. In this context, the key value of adopting analytics for information security becomes to direct investigations toward incidents that are of the highest value in terms of addressing the adverse impact to the organization. To ensure consistency across investigations, organizations benefit from common guidelines for the operations of information security staff. In conducting activities that involve multiple parties, this can aid the reliability of investigations. Despite that, security teams need to be continually reviewing, refining, and improving the

guidelines. Attackers are continuously updating their processes and tools, and it is incumbent on defenders to do so as well.

CONCLUSION

Even though information security remains a pressing concern for all organizations, solving the security problem writ large remains an elusive goal. Technological considerations mean that many systems and components are built with imperfect security. Psychological considerations mean that attackers can compromise organizations via techniques such as phishing attacks, where malicious emails that appear legitimate are sent to targets that in turn lead to system compromise. Economic considerations mean that it is less costly for an attacker to attempt to compromise an organization than it is for the organization to defend its network. Furthermore, criminal investigations into attacks do not often succeed due to the challenge of the complexity of the investigations and jurisdictional considerations. In the end, attackers maintain the upper hand.

Organizations need to do all they can to protect their networks, but they do not have unlimited resources to do so. Understanding how best to focus their investigative and remediation activities in the event of a security compromise is a key benefit that advanced analytics offers the security domain. Yet, as with other domains, poorly managed data can quickly hinder these capabilities. Understanding and properly implementing data management processes for security analytics systems is a key area in which organizations have an advantage in their defense: By holistically understanding their network, information technology assets, and organization, they are in a position to recognize and respond to security incidents and in turn greatly mitigate information security risk.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Alex Anglin
SAS Institute (Canada) Inc
alex.anglin@sas.com