Chapter 3: Creating A Repeatable, Reusable, and Reliable Security Data Analysis Workflow and Toolkit

While the majority of the content of this book focuses on how to analyze and visualize “security” data, there is a rhinoceros in the room1 that must be dealt with. If you aspire to become a security data scientist, it is not enough to just take snippets of code, apply them to your own data samples, produce a graph (or two) and declare victory. Just as the comic book hero Captain Marvel2 took on the powers of six mythical figures to save the day, you will need to be infused with the skills and abilities of a librarian, systems administrator/architect, mechanic, programmer, project manager, and forensic pathologist to gain the most personal and organizational value from the concepts and techniques presented in these pages. Unfortunately, it will take a little more work than uttering “*Shazam!*”, but this chapter should help you down the path of acquiring those fundamentals skill with the added bonus of no lightning strikes.

Maintaining an Inventory of Data Sources

It would be difficult to say that no data exists for us to process for the purposes of security-oriented analysis. Even the most nascent security analyst should be able to rattle off a list that would look similar to Table #. But, do *we* really *have* data? Network administrators are usually the owners and caretakers of their device configurations and logs. The same is also true for Windows and UNIX/Linux administrators. The security team may own firewalls, but if you’re a large organization, that may be a completely different team than the one that will perform analytics on the data.  Unless you own the data from generation to deletion, you will be relying on others to either provide it or provide access to it.

Table 3.1 Potential "Security" Data Sources

|  |
| --- |
| Windows Event Logs  Linux/UNIX syslogs  Mainframe Logs  Network Device Logs  Proxy Server Logs  Firewall Logs  Anti-malware Management Event Databases/Alerts  Vulnerability Management Databases  Patch Management Databases  System Configuration Logs  Identity & Access Management/RBAC Records  NetFlow Data  PCAP Data  HR Data Feeds  Application Logs  Web Application Firewall Logs  Business Transactional Data Logs  Database Audit Logs  Asset Management Databases  Physical Security Event Logs  IDS/IDP Alerts  Indicators of Compromise  Help Desk/Non-security Incident Tickets  Risk Assessments  Penetration Testing Results  Application Security Scans  Firewall Port Requests |

Given the plethora and diversity of sources, your first and foremost task is to channel your inner-librarian to create and maintain a comprehensive catalog of these sources, *even if you’re not going to use them for analysis right now*. While some metadata will be unique, there are basic/common elements to record for each component:

**What is the generator of the data?** This could be as broad-based as identifying an organization-wide proxy server farm or as specific as the components of a line of business web application. Identifying the data source generator is important, though, especially as products and applications are retired or upgraded. Having a clear description of the source will make it much easier to make updates or changes when you perform reviews of your catalog.

**What is the actual or potential security purpose of the data source?** The main thrust of this question is to determine how the data can or will be used for security-oriented analyses. While there is a definite school of thought that security practitioners should “log all the things”, this is truly not practical even in the age of “big data” and cheap storage. We’ve used the word “potential” on purpose since there is a huge difference in having a catalog of all data sources and actually using them all. Think if it in terms of an inter-library book loan. The book is in the catalog and you know you can get access to the resource when you need it, even if it isn’t at your local library at the moment. Knowing what and where the data is can save a great deal of time later on, especially if you’re in the middle of an incident.

**Who is the owner/custodian/controller of the data source?** Here, you should be recording the contact information of at least two people entrusted with care and feeding of the data. This is usually the application/service owner and it’s a good idea to go refresh your catalog on at least an annual basis to ensure you have up-to-date contact information. These records will come in handy when you have access or processing issues (and, you will).

**What steps need to be taken to access to the data source?** Most data sources require special permissions to gain access to them and many have special or unique access methods. In some cases, this will involve pulling data via sftp, FTP, http[s], rsync, nfs or CIFS either in real-time (streaming) or at timed-intervals (batch).  You may also need to make direct SQL calls to a myriad of databases or REST3ful and/or SOAP4-based queries to retrieve data from more modern sources or proprietary appliances. Finally, it’s almost a guarantee you’ll be either a Windows event log consumer or syslog consumer and having a solid inventory will help you in building and scaling a good log management environment. Capturing solid details on the access method(s) will also be invaluable when it comes time to debug why data mysteriously stopped flowing into your analytics engines.

**What format(s) are the data elements encoded in?** In a way, your security analytics ingestion “hub” will be a digital Rosetta Stone5, knowing how to read and translate multiple representation formats into ones your engines can process. You can expect to be required to parse comma- and tab-separated (CSV/TSV) records, JavaScript Object Notation (JSON) objects, XML, Common Log Format (CLF) and a myriad of custom log formats. You may also be capturing raw packet capture (PCAP) dumps, SNMP traps or even unstructured text files. Recording this attribute will help you know when you need to add a new translator type to your repertoire.

**What mechanisms are in place to validate the integrity of the data source and transmission processes?** We’ve hinted that your data acquisition setup will be less than perfect and the answer to this question should help provide an early warning system when gremlins decided to creep into your processes. If a source should be generating a mean of “n” events per second and that suddenly drops down a few standard deviations or—even worse—to “0”, you can fairly confidently assume that this is something you should investigate, especially if it’s a more vital data source (say, firewall logs or IDS alerts). Similarly, you may be expecting to consume a data source that has eight fields per event that now has more or fewer fields as a side effect of a vendor “upgrade”. Having a process in place to validate and notify of such integrity issues can help prevent lost time and visibility down the road.

**What are the record retention policies for the data source?** If you’re not in a large or highly regulated organization, issues such as data retention tend to come up only when storage space becomes scarce. However, in many companies there are strict policies on how long you must maintain access to certain types of data. You should check with both the data source owner and your legal/compliance department to determine what your responsibilities are as a consumer and processor of the data. Many times, a data source owner will believe they have transferred responsibility to the security department without explicitly stating so up front. Getting this confirmation can stave off future headaches and potential legal issues.

**Where, physically and logically, is the data?** Having an understanding of the geographical and network locations of the data source will be very helpful when it comes to actually implementing the data intake procedures. If you have identified that a source is in a Colorado data center but your collection and analytics engine is in the Detroit data center, you know that there are potential availability, capacity and latency issues you may need to deal with when building out and maintaining your intake and analytics engine. Ideally, this information will be obtained from your existing asset management system, since your security analytics inventory should—ideally—be a reference source, not the authoritative source.

**What is the expected volume (*how much*) and velocity (*rate of transmission/expected consumption*) of the data?** You may say you want real-time access to your organization’s firewall logs but that access will do little good if your processing engines are not capable of digesting that fire hose of information. Thankfully, it’s fairly straightforward to measure and estimate the volume/velocity of the data flow and design an appropriate intake configuration. Skipping this step may end up with you only retrieving only one out of every four events and missing potentially critical data elements or worse, launching your own denial of service (DoS) attack on your logging and processing servers.

Building Your Data Intake And Analytics Engine

Now that you have an idea of the data sources that are available and at least an initial estimate at which ones you’ll be using, you’ll need to design the collection and processing systems. If you think of the data as fuel, your goal here is to design the most efficient way to get that fuel into your tank for storage and engine for processing. This will not be a one-time event as you will always be incorporating and potentially retiring data sources and developing new ways of crunching the data. Figure 3-1 provides a conceptual overview of this data flow.

Figure 3.1: High Level Overview Of Data Intake/Processing Flow [f03##.eps]

Keeping the intake process as straightforward as possible should be your primary goal as that will make it much easier to diagnose and fix issues as they crop up and will also simplify what mechanisms you need to put into place to get at this data for analysis.

First, you must decide what your centralized system, network and application event logging strategy will be. The concept of a homogenous data center environment is but a faded memory—if it ever truly existed—since even the most Windows-centric organization cannot help but have its share of routers, switches, firewalls and appliances, not to mention a mainframe or two. Despite the continued prevalence of syslog-based logging across an ever increasingly wide spectrum of systems, applications and devices, Windows systems still hold hard and fast to using Windows event logs as their preferred log language (much like how the French fight the linguistic hegemony of English). Given this data center dichotomy, you will need to make a choice as to whether you’re going to use a solution like Microsoft System Center Operations Manager (SCOM) to handle all events—including, using their syslog adapter for everything besides Windows components—or whether you will be using a logging translator in your Windows environment to get events out into a more common tongue.

Larger organizations with equally as large budgets may benefit from a proprietary log management solution such as HP ArcSight Logger or Splunk, but it’s possible to architect a robust collection hub with something as traditional as syslog-ng6 (since you can even forward Windows events to it using a tool like Snare7) and more modern open source tools such as logstash8 and Graylog29. Figure 3-2 provides a more detailed picture of this part of the intake hub.

Figure 3.2: Expanded Event Logging View Of Intake Hub [f03##.eps]

Keep in mind these three core, guiding principles for your intake hub:

* It must have the ability to consume events from current and future sources. This means it must be scalable and flexible.
* It must have the ability to store events for later retrieval/processing. This means the storage must also be scalable.
* It must have the ability to stream events, possibly translating them into a more usable format, into your analytics engines.

The previous section introduced the concepts of data *velocity* and *volume*. You will need to design your intake servers (the “hub”) well enough to handle the entire incoming flow of data from each source. You may be used to the term *velocity* being referred to as the number of *events (or messages)-per-second* and *volume* as the aggregate *event record (or message) size*. Here is where your systems administrator/architect powers will come in handy. NOTE: You may also end up needing to exercise your budget bending powers as well, since this may be a more expensive task than you might think at first.

Notes

1. Joseph F. McDonald, “Russell, Wittgenstein, and the Problem of the Rhinoceros”, *The Southern Journal of Philosophy* (Volume 31, Issue 4, Winter 1993; pp.409-424)

2. http://en.wikipedia.org/wiki/Captain\_Marvel\_(DC\_Comics)

3. http://www.ibm.com/developerworks/webservices/library/ws-restful/

4. <http://www.ibm.com/developerworks/webservices/tutorials/ws-understand-web-services1/section2.html>

5. <http://en.wikipedia.org/wiki/Rosetta_Stone>

6. <http://www.balabit.com/network-security/syslog-ng/opensource-logging-system/>

7. <http://www.intersectalliance.com/projects/SnareWindows/>

8. <http://logstash.net/>

9. <http://graylog2.org/>

Recommended Reading

*Mastering Unix Shell Scripting: Bash, Bourne, and Korn Shell Scripting for Programmers, System Administrators, and UNIX Gurus, 2nd Edition* by Randal K. Michael (Wiley, ISBN: 978-0-470-18301-4)