Hewlett Packard Enterprise

Business white paper

A vision for cyber security detection analytics

Table of contents

Introduction

Organizations are in the midst of considering how Big Data can assist in their plans to detect advanced cyber adversaries. Many are starting to build Big Data infrastructure and feed it both structured and unstructured data, but few have determined exactly what they will do with the data after they have collected it. This paper outlines the vision of what to do with all this security data; a vision for detecting advanced adversaries through pairing Big Data and data science.

Detect breaches

The security industry is not catching enough "bad guys". Reports published by the vendor community show that the time between a breach occurring and a breach being detected average at 229 days.[1] In fact, the majority of breaches are reported by external parties and law enforcement agencies based on stolen assets showing up in the underground economy.[2] As a security community, these facts underscore what we all know to be true; that we need to consider new ways to detect our own breaches more quickly.

Why is it so challenging to detect breaches?

Currently available cyber security tools are pretty good at detecting known attack patterns. If an attack matches a signature, talks to a known bad place, uses unencrypted protocols, or happens within the infrastructure that we closely monitor, we can reliably detect it as it occurs (if the technology is set up properly). What we struggle with is detecting unknown attack types, new malicious behaviors, and insider threats. We also struggle with attackers hiding within a bell curve.

For example, in the past many attacks occurred on Fridays at 3 p.m. just before a 3-day weekend. This allowed plenty of time to break in, ransack the place, clean up, and install a backdoor for persistent access. Today things have changed and we now often see attacks on Wednesdays at 10 a.m., because our adversaries understand that we are sensitive to volume and this is peak network traffic based on well-understood network behavior. The adversaries know how to hide in our normal bell curve of network activity.

[1] Mandiant annual threat report, M-Trends, April 2014

[2] 2014 Data Breach Investigations Report, Verizon, 2014

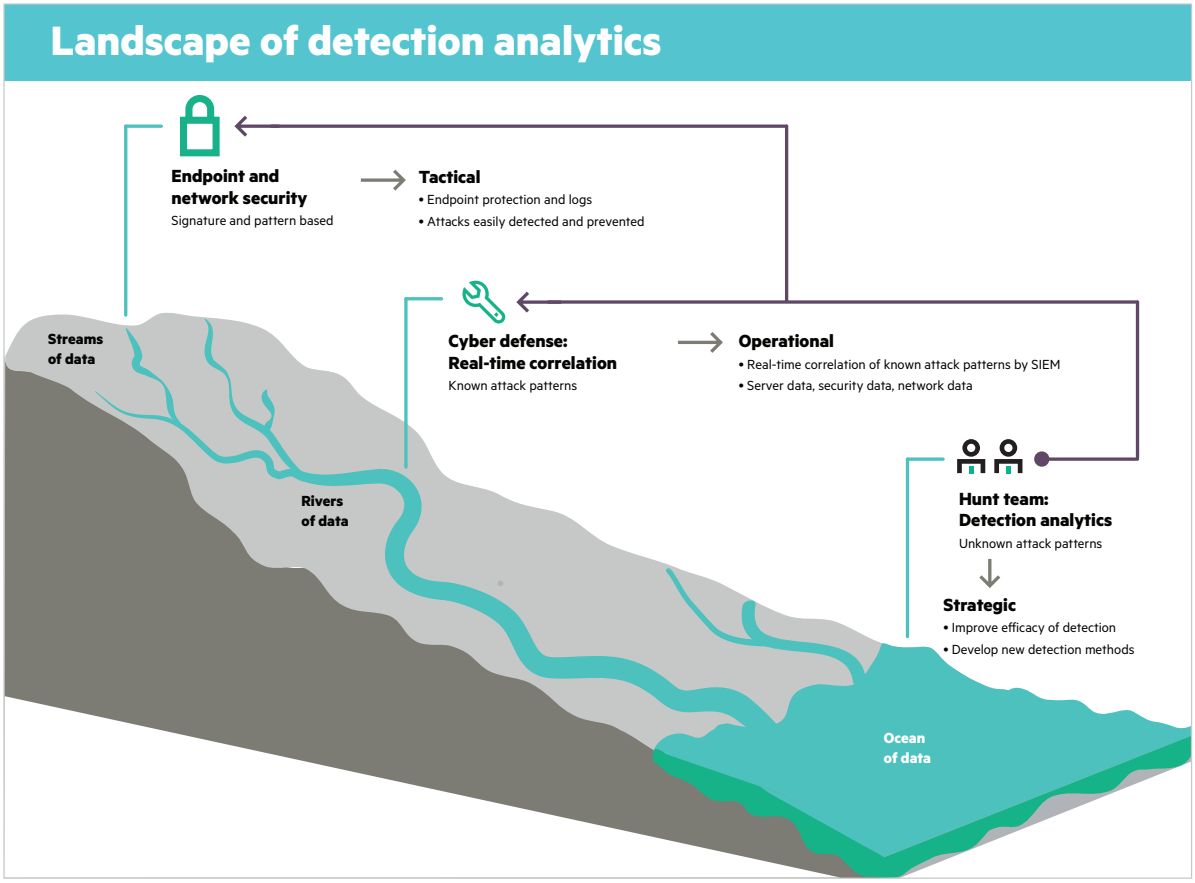

Figure 1. The modern threat landscape and the geography of security detection as analogous to a river delta

A "river delta" analogy

When considering the future of detection analytics, it is helpful to consider a "river delta" as analogous to the current enterprise security landscape. A river delta consists of streams, rivers, and an ocean (see figure 1). Endpoint and other security devices produce small streams of operational data. This data flows downstream and is aggregated into rivers of enterprise log and security data. Think of these rivers as including business operations context, IT operations logs, and information security events. When these are aggregated and monitored by real-time correlation in a security information and event management (SIEM), they anchor the modern Cyber Defense Center (CDC). This CDC monitors real-time correlated security events to detect indicators of potential attacks in progress.

Given the modern threat landscape, we can now picture how real-time capability requires a correlation, and longer-term analytical capability as a supplement. We need to expand operational post-hoc analytics to the data "ocean". This work in the data ocean is commonly assigned to a newly formed "hunt team". There should be an important operational link between the hunt team and CDC, especially when an unknown attack takes place. Once this attack type is detected, it is then converted into automated (and hopefully) real-time detection, so in the future it is caught in real-time.

In terms of geography, the tactical technologies for breach detection and prevention are in the "streams" of data (e.g., intrusion prevention), operational monitoring capabilities sit across the "rivers" of data (e.g., SIEM), and any strategic data analysis for breaches reside in the oceans of data (e.g., hunt teams). People and process are a critical link between these levels.

All data is not equal

One important counter argument to Big Data is the fact that all data is not equal. There is a Big Data counter movement in the data analytics community that is referred to as the small data movement. The thought process is that you find the application or use case for the data you collect before you collect terabytes of it. The conventional wisdom, "collect it all and figure out what to do with it later," is both false and expensive.

Every bit of data collected must be processed, transmitted, normalized, analyzed (labor intensive), stored, and managed through a complete lifecycle. An example of wasteful data collection can be seen in typical router log collection. Some of the common logs seen from routers include: route up, route down, and route flap. Those three messages are not often (if ever) useful in enterprise security detection and having a terabyte of them will not magically make them useful. There is a strong tendency in enterprise security monitoring to collect the information that is easiest to get in big scale. A better way to approach Big Data and data science is to first find an application for the data and only then collect it on a large scale.

Security analytics is not a product

There are commercial products and open-source tools that organizations can use to perform analytics, but you can't get the full value of enterprise security analytics by installing hardware or software only. Organizations are building systems that ingest hundreds of megabytes of security data per second, but their analysts can only read a few bytes per second. Even with great visualization tools, analysts will only be able to mentally process a few kilobytes of data per second. Tools can help pare down the dataset to a smaller size, but analysts also have to know what questions to ask.

Enterprise security experts need to learn to think with an analytical mindset. They need to be curious, explore the data they collect, find patterns, and follow the trail of an investigation wherever it leads. Looking at individual events or correlated events are not sufficient anymore. While the hunt team is a separate role for cyber defense analyst in larger organizations, smaller organizations will not have this luxury. Existing security analysts should be trained to look at their data with an analytical and curious mindset.

A practical note on retrieval speeds

One critical factor preventing enterprise security teams from optimally exploring data is poor retrieval speeds. The Big Data space now has technology, such as columnar data stores, which can quickly retrieve results, provided the data is stored in an optimized fashion (structured data is the fastest). As you build out plans, we caution to not only consider how fast you can place data in your ocean, but also to consider if you can get the data out of that ocean, and how quickly it needs to be retrieved in order to be useful to your security teams. Our experience shows that optimal exploration environments require results in seconds (not minutes) to capture the creativity and the "Aha" moments of a hunt team.

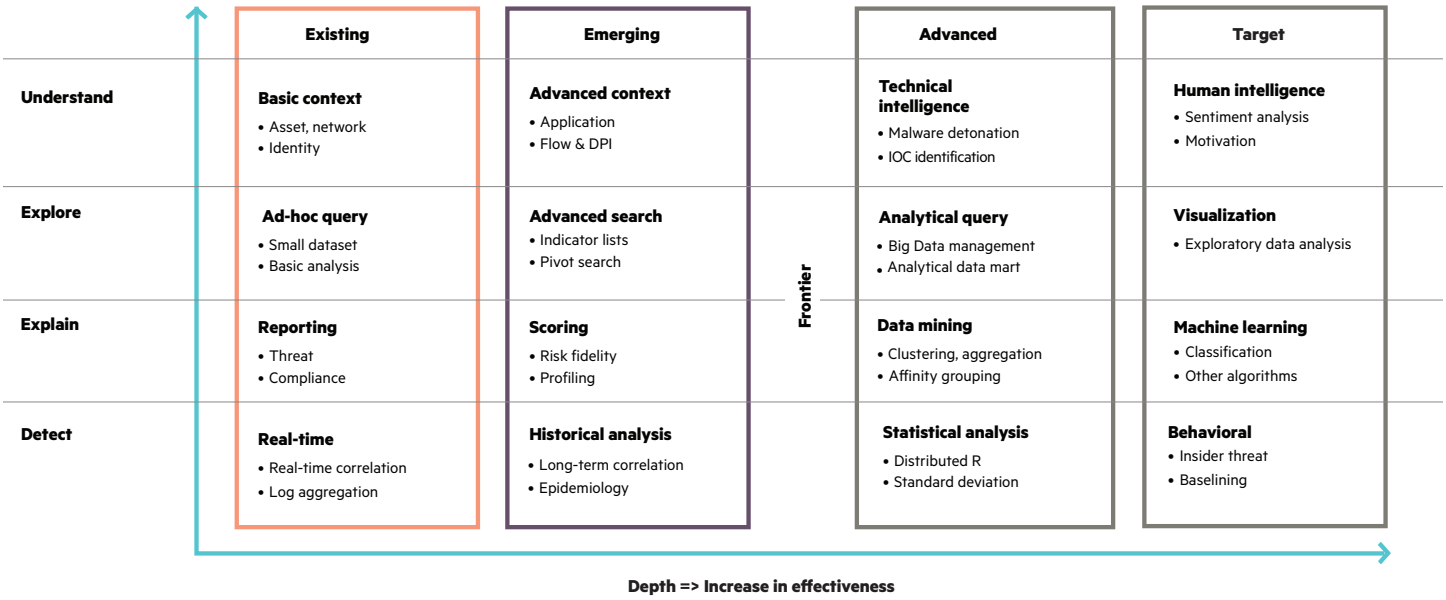

Figure 2. A vision for detection analytics

A vision for the future

To apply data sciences for efficient security detection in the future, you need mature capabilities as you work to detect, explain, explore, and understand security events in your environment. Figure 2 provides a vision, mapping, and strategy for your current and future detection analytics needs.

Existing detection is the base detection technology stack we commonly use today.

- **Detect:** Current detection methods include real-time correlation and log aggregation; this is the single pane of glass and the ability to highlight the events of interest.
- **Explain:** Explaining often results in reporting, this includes threat, vulnerability, and compliance reports. These are topics that need careful construction and effective communication to leadership.
- **Explore:** Exploration requires the ability to query data in an ad-hoc manner to produce small datasets on which you can conduct basic formal analysis.
- **Understand:** Context is an important component of detection analysis because it can be tricky to decide whether unusual activity is truly a breach or if it is simply a benign event. Having context is a critical component in that decision. To gather context, we can collect information about assets, networks, and identities that we are protecting. This context will be critical in our decision on whether an event is indeed malicious and should be escalated as a breach or if, for example, the event is simply a compliance scan.

Emerging detection—an analytical capability—is available either as a piecemeal component of enterprise security programs or as individual point solutions, but not yet as a coherent technology stack.

- **Detect:** Consider historical analysis and imagine the ability to conduct long-term correlation as an immediate follow on to real-time correlation. You could escalate an event of interest and then immediately search for every historical instance of that correlation event. This kind of cyber epidemiology makes policy decisions about your enterprise security possible, which is based on an analysis of "patient zero" rather than just cleaning up an infected host.
- **Explain:** Emerging scoring requires the ability to profile IP addresses, users, systems, servers, and other elements of the risk picture in our enterprises at a much higher level of fidelity. There is a real danger in what is called "feel good" data science here, that is mixing quantitative elements with qualitative descriptors and this can water down the value of the measurement.

- **Explore:** Advanced searching assists with exploration. The ability to pivot through security data is particularly useful; for example, if you explore a user's behavior and see the user interact with a particular server, you can then pivot and see what else that server has done subsequently. This also affords the ability to conduct effective searches of the enterprise security data looking for known lists of indicators of compromise (IOCs).
- **Understand:** True understanding of a breach will often require advanced context. This context would include things like internal application error logs. One example is, collecting application logs from the HPE Fortify runtime agent so they can report into an HPE ArcSight SIEM. Another key to understanding a breach is network flow and deep packet inspection data. It allows deeper root-cause analysis in order to truly understand if an ongoing attack deserves to be escalated.

Advanced detection capabilities are the current frontier of development, both in the vendor community and in advanced enterprise security programs.

- **Detect:** Advanced detection includes advanced statistical analysis, the ability to train models and detect outliers across any of your enterprise security data feeds. Even an insider has to deviate on at least one measurable parameter in order to conduct malicious activity.
- **Explain:** Here we find one of the greatest assets we have available to us as a community to date; and that is data mining. Here are the main disciplines:
  - Clustering, at first glance, does not appear to be that valuable for security. However, when you cluster security data you almost inevitably identify large numbers of false positives. When you clean up these false positives on your security devices, you effectively make the deep muddy river of data that we monitor clearer and shallower.
  - Classification is used to classify data into types for comparative analysis and cross correlation.
  - Correlation has been around a long time and provides the analytical engine for modern enterprise SIEM deployments.
  - Aggregation also seems less powerful than it actually is. Imagine an aggregate profile of a server, a user, or an IP address based on long-term historical behavior information; this type of aggregation can easily profile attackers and then allow you to identify similar profiles with the addition of a scoring algorithm.
  - Affinity grouping is a very interesting capability. An affinity group is best articulated in an anecdotal (and possibly folklore) retail use case called the "beer and diapers analogy". When retail companies analyze what you buy and demographically categorize you, interesting insights are found. One such insight is, when men buy diapers they also buy beer. Thus keeping beer en route to diapers increases the sale of beer.

    In a security context, malicious command-and-control infrastructure within your environment has a higher affinity for itself and its peer nodes than for the normal infrastructure surrounding it. This allows us a significant advance in detection capability. This type of lesson is called a "domain transfer" and there are a number of effective lessons learned from other domains in data science (such as marketing) that can be applied to information security.
- **Explore:** Analytical query takes your Big Data and turns it into small data, with which you can actually do something. Big Data management takes large amounts of data and turns it into analytically tractable small amounts of data that can be retrieved in a reasonable timeframe. These are often called queries or data marts and these are where you apply your advanced analytical capabilities.
- **Understand:** Technical intelligence in the "understand" row is the concept of producing your own indicators of compromise. Currently, the majority of the industry purchases this intelligence. It is generic intelligence aggregated from open sources and occasionally augmented by honeypot collection projects. The ability to detect new malware in your environment, detonate it, and produce indicators to be shared across your enterprise begins to simulate an immune response and that is clearly a desired direction for the security industry.

Target detection capabilities, the ultimate end game for the technology stack, is behavioral detection of advanced and insider threats, and there's a deeper focus on baselining normal as well as abnormal behavioral profiles.

- **Detect:** Baselining, as much as possible, in your environment will be helpful in advanced detection—and the smarter we can get at baselining, the better we will be at identifying abnormal and noteworthy activity. A behavioral white list, where you deny all traffic and then only allow explicit traffic, is still a ways off for most enterprises but is a long-term target.
- **Explain:** As data matures, we want to increase our depth from data mining and move into machine-learning algorithms. There may not be a clear differentiation but one big impact of machine learning algorithm is automated classification of data types—on which further analysis can be applied. Often, these types define a norm and we can then use this in our dataset for analysis.
- **Explore:** We need true exploratory visualization for security data. While there are visualization tools that are being effectively applied to information security problems, we do not have a dedicated security visualization tool. The detection capability that is made possible by visualizing large amounts of data and being able to cascade disparate visualizations through a selection processes, allows you to perform root-cause analysis and remove data that is not of interest. Then you can dig in visually to the remaining data and it will be richer in malicious activity.
- **Understand:** Human intelligence advances us beyond technical intelligence. The idea is that we could understand and gather human sentiment and motivation indicators based on observed activity (e.g., Twitter scraping or IRC monitoring). While this would be a boon to our work, it is an unstructured data problem and quite complex to address effectively. Imagine the ability to put human sentiment under the same pane of glass as enterprise log data. This may be one privacy bridge too far, but it describes a powerful capability for detecting pending campaign-based attacks.

Conclusion

As a cyber security industry, we need to be catching more attacks, and catching them earlier. In order to do that, we need a shift in how we look at our data. Just collecting large amounts of Big Data is not useful by itself. We need a guiding vision and plan in order to build systems that will grow with our needs as we work to get to the target state, which is reliable behavioral detection of advanced attacks and insider threats.

Learn more at
hpe.com/software/SIOC

About HPE Enterprise Security
Hewlett Packard Enterprise is a leading provider of security and compliance solutions for the modern enterprise that wants to mitigate risk in its hybrid environment and defend against advanced threats. Based on market-leading products from HPE ArcSight, HPE Fortify, HPE Atalla, and HPE TippingPoint, the HPE Security Intelligence Platform uniquely delivers the advanced correlation, application protection, and network defenses to protect today's hybrid IT infrastructure from sophisticated cyber threats.

Sign up for updates

Hewlett Packard Enterprise

4AA5-6876ENW, July 2016, Rev. 2