www.pipelinepub.com
Volume 17, Issue 1

The Third Wave in Cybersecurity

By: Aimei Wei

Cybersecurity advances come in waves, and the waves of the past few years have revolved around data collection and incorporating AI technology as a means of sifting through that data. But as cybersecurity tools proliferate, it's clear that data collection and AI are not enough. Security analysts must monitor a dozen or more different consoles to inspect traffic across their entire attack surface, and they miss complex attacks that consist of seemingly innocuous events reported by several different tools.

This is why we have entered a third wave: correlation. Correlation involves inspecting and responding to data inputs across tools to achieve higher fidelity alerting. While data collection alone resulted in seas of data that were impossible to effectively inspect and AI produced better results within security tools, correlation pulls together data from across disparate tools to more effectively spot complex attacks.

In this article, we'll look at the progress of cybersecurity technology through the first two waves, and see how the correlation wave is now emerging to deliver faster detections and responses to complex attacks.

Wave One: The Rise of Data

Data has been at the heart of cybersecurity since its inception. It began with analysts looking at firewall and server logs and expanded as the number of security tools grew. The main purpose of security information and event management (SIEM) systems was to collect and aggregate logs from different tools and applications for compliance, incident investigation and log management. ArcSight, one of the SIEM tools released around the turn of the century, was a typical example of an SIEM and log management system. The raw packets were collected and stored as-is for

forensics despite the fact that they require lots of storage space and it is very hard to sift through these huge numbers of packets to find any indication of breaches.

Quickly, however, we realized that neither raw logs nor raw packets individually are enough to be effective in detecting breaches, and raw packets are too heavy and have limited usage besides forensics. Information extracted from traffic such as Netflow/IPFix, traditionally used for network visibility and performance monitoring, started to be used for security. SIEMs also started to ingest and store Netflow/IPFix, too. However, due to both technical scalability concerns and cost concerns, SIEMs have never become the mainstream tool for traffic analysis.

As time went by, more data were collected: files, user information, threat intelligence, etc. The goal of collecting more data was valid—get pervasive visibility—but the net challenge, responding to critical attacks, is like finding needles in a haystack, especially via manual searches or rules manually defined by humans. It's both labor intensive and inefficient.

There are two technical challenges facing data-driven security: how to store large volumes of data at scale (allowing for efficient searches and analysis) and how to deal with the variety of data—especially unstructured data—as data can be in any format. Traditional relational databases based on SQL ran into both of these problems. Earlier vendors scrambled to solve these problems with many homegrown solutions. Unfortunately, most of them were not as efficient as what we are using today—NoSQL databases for Big Data lakes.

There is one more challenge facing data-driven security: the software architecture to cost-effectively build a scalable system for enterprise customers. The typical three-tier architecture with front-end business logic and database tiers became a big hurdle. Today's cloud-native architectures, building upon micro-services architecture with containers, provide much more scalable and cost-effective solutions.

Wave Two: The Rise of AI

Once you have lots of data, what do you do with it? As mentioned previously, with a large of volume of data, sifting through it looking for meaningful patterns is tedious and time-consuming. If your IT infrastructure gets hacked, it might take days or weeks to find out. Meanwhile, the damage is already done, or sensitive data has already been stolen.

In this case, too much data becomes a problem. Fortunately, we have seen the rise of machine learning (ML) thanks to advances in machine learning algorithms and computing power. Machines are very good at doing repetitive and tedious work very quickly, efficiently and tirelessly, and do so 24x7. When machines are equipped with intelligence-like learning capabilities, they help humans scale. Lots of researchers and vendors in security started leveraging AI to solve the data overload problem, to help them find those needles, or to see trends that are hidden inside large datasets. For example, endpoint detection and response (EDR) companies are using AI to address endpoint security problems; user and entity behavior analysis (UEBA) companies are using AI to address insider threats; and network traffic analysis (NTA) companies are using AI to find abnormal network traffic patterns.

On the surface, having lots of data becomes less of a problem with AI-driven security, as ML usually requires lots of data to train the model and learn the patterns. On the contrary, not enough data is obviously a problem as the less data, the less accurate and thus the less useful the ML model becomes. However, as time went by, researchers gradually realized that having the *right* data was far more important. Too much data without the right information is just a waste of computing power for ML as well as a waste of storage space. Earlier UEBA vendors with solutions based on logs from SIEM tools learned this hard lesson: the SIEM might have collected lots of logs, but only a few of them contain the right information related to user behaviors. So, although data-driven security builds a great foundation for AI-driven security, in order to build scalable and accurate AI-driven security, the right data is far more important.

Using AI definitely helps alleviate the pains with Big Data, but it has its own challenges. For example, both UEBA and NTA leverage unsupervised ML for behavior analysis. However, an abnormal behavior observed for a user or from network traffic does not necessarily mean a security incident. These tools can thus generate lots of noise, causing alert fatigue. Furthermore, the smart hacks usually go through several stages of the kill chain before they can be caught. How can you recover the trace of a breach and fix the root cause?

Finally, another big challenge facing AI-driven security collectively is cost: the capital cost of the tools themselves, the cost of the infrastructure of compute and storage used by these tools, and the cost of operating so many different tools in their silos with different screens.

AI and ML technologies can facilitate automated decision-making that blocks firewall ports or performs other attack-killing functions, but only if the right data is being fed to them.

Wave Three: The Rise of Correlation

Correlation provides the means to wade through the data to come up with the right data that enables us to spot zero-day attacks and other sophisticated hacks. The wave of correlation is built upon the two previous waves. However, it is all about getting above the data as well as the tools, and it is about wrapping everything together in a single platform. Correlation means consolidating tools so their results can be examined more easily, and correlating results from disparate tools so multi-phase attacks can be spotted and stopped.

Security analysts from ESG, Gartner, Forrester, IDC and Omdia all agree this change in thinking from siloed tools to a consolidated platform is key to helping us see and respond to critical breaches. Specifically, the platform needs to take a holistic approach and look at correlating detections across network, cloud, endpoints and applications: in short, the entire attack surface.

The key objectives of correlations of detections across tools, feeds and environments are to improve detection accuracy, to detect complex attacks by combining weaker signals from multiple tools to spot attacks that might otherwise be ignored, and to improve operational efficiency and productivity. No longer does comprehensive visibility mean finding the right data—rather, it means correlating data to spot the complex attacks.

For example, suppose an employee gets a phishing email with an embedded link. The employee clicks the link and downloads a malware file. The malware accesses a corporate server at 2 AM.

The malware begins sending sensitive data to an external address using DNS tunneling. Each of these activities would show up in a different log, so there's a need to correlate these actions to reveal the attack.

To implement the third wave, many security vendors are building so-called anywhere (x) detection and response, or XDR platforms. Companies like Palo Alto Networks and Trend Micro are buying and consolidating tools from other vendors, and pure-play XDR vendors like Stellar Cyber have built XDR platforms from the ground up. XDR is a cohesive security operations solution with tight integration of many security applications in a single platform with a single pane of glass. It automatically collects and correlates data from multiple tools, improves detections, and provides automated responses. A platform tying together tools and applications innately drives the cost down, in both tools cost and infrastructure cost, while it improves operational efficiency by eliminating manual correlation.

The third wave of cybersecurity is about collecting the right data, consolidating it in a data lake, and correlating the output of a dozen or more security tools into a single console for rapid detection and response. Data is the engine that runs the system: while AI and ML discover the right data and initiate automated responses, correlation spots the most complex attacks to prevent them from ruining the CISO's day.