Digital Forensics

The Art of Cyber Archaeology

According to the Ponemon Institute's 2018 "Cost of a Data Breach" study, based on interviews with 2,200 professionals from 477 companies, the average time to detect a data breach is over 6 months (197 days).

The 2017 Verizon Data Breach Investigation Report tells us that there was a 3:1 ratio of external to internal breach detection methods. In other words, a significant majority of company breaches were **not** detected by internal security tools but were reported by law enforcement, fraud protection services or other third parties outside the company.

Time between breach and detection:

197 days

When a breach is discovered, and the company CEO asks, "How bad is it?", having an effective long-term digital forensics strategy can be the difference between answering "**I don't know**", and "**I don't know yet, but I will be able to find out**".

This white paper helps you understand the potential value of forensic log data, as distinct from day-to-day operational security logs, and the potential challenges involved.

The case for digital forensics

There is an old stereotype of 90s era IT security; the security administrator would be stuck in a room deep within the IT group, they'd do some sort of esoteric magic to turn pizza and cola into firewall rules and virus signature updates. In those days, data flows in and out of the organization were usually rigidly defined, there was a clear delineation between 'inside' and 'outside' the corporate network, and IT risk and corporate risk were not closely aligned.

Breaches detected using internal resources:

1 in 3

Those days are well and truly gone. The CISO is usually a board or executive level position, the corporate risk profile is intimately tied into digital data security, data flow into and out of an organization is often mercurial in nature, and a defense in depth strategy is just one of the things designed to protect assets in a world dominated by BYOD (bring your own device) and cloud services.

The consequences of a breach can be significant - money, reputation, cyber insurance premiums, regulatory repercussions, and so on; but sometimes the fact that a breach happened is less important than discovering the scope of the breach. How did they get in? What did they see? What did they take? Are there any backdoors remaining? Not knowing the answers to these questions is a painful prospect for a CEO at a board meeting. Not having the ability to find out, even with additional time and resources, is a nightmare scenario.

Organizations often have a fiduciary responsibility to divulge the scope of attacks, and the resulting data breaches not just to internal staff and shareholders but also to users of a service. The global nature of services and customers means that an individual organization may be subject to dozens or hundreds of different national regulations relating to the privacy and security of customer information. Having the data to clearly indicate the scope of a breach can be an immensely valuable asset.

The security cat and mouse game

As technology, hardware, operating systems and applications have evolved, so has the number and sophistication of attackers. The time required to turn a zero-day vulnerability into a global security pandemic is now measured in hours or less, and nation-state or well-funded criminal actors, assisted by our willingness to post our professional and personal lives on LinkedIn, Instagram and Facebook, will configure their campaigns to laser-target particular staff, technology or procedures. Even organizations that normally fly under the radar and have little of current perceived 'value' to an attacker may be caught up in the blast-wave of targeted attacks, with their systems compromised with ransomware or Advanced Persistent Threat (APT) packages installed as a beachhead for future incursions against the company.

Ransomeware cost per year:

$75 billion

The requirement for network access is pervasive; systems or devices that were previously air gapped now require an internet connection for full functionality. On the consumer side, watches, phones, televisions and even lights are often connected to the Internet in some fashion. Despite the best of intentions, such advancements tend to seep inside the walls of organizations, leading to a broad potential attack surface incorporating devices that may have minimal functionality and a low threat profile, but poor security update support.

In live attacks, bug bounties, and Capture-The-Flag (CTF) tournaments, we regularly see situations where several low to moderate risk vulnerabilities or devices can be strung together to fully compromise a system or network. This changes the way a security team needs approach risk management to more of a holistic and thorough approach using defense in depth strategies, rather than concentrating just on critical assets or network boundaries. A security team must get it right, all of the time, across all devices to stay safe. An attacker only has to get it right once to compromise a potentially essential asset or beachhead device. The normal business continuity and recovery strategies such as backups, are an effective part of an administrator's toolkit, but to reduce the risk of reinfection, they need to know what happened, when, who, how, where and why. Having relevant and high-quality log data is key for any remediation and forensics investigation.

The resource tug of war

Log data accumulates quickly and relentlessly, and much of it is of minimal value for day-to-day security posture analysis or intrusion signature detection. When a security incident is detected, though, every single event your investigators can get their hands on may as well be gold plated.

Tens of thousands of events per second is gigabytes per day and terabytes per month, equating to petabytes of storage to meet regulatory retention requirements.

Each event collected and retained is a cost to the organization - whether it be displaced network bandwidth, CPU usage or disk storage - so most log analysis solutions are designed to retain a small amount of operational data to meet a security operation center's immediate need. Some software even charges on a per-event basis for collection and storage, which provides a natural disincentive for broad and long-term forensic logging.

Every byte included and every event ingested means additional indexing, scanning and storage overhead. Organizations that have a more strategic approach to log data and value the ability to retain and query historical forensic security information are often faced with the difficult choice between operating large expensive clusters to scale up operational logging infrastructure to cope with event and storage volumes beyond day-to-day logging needs, investing significant time and effort to filter incoming event data, or using a specialist forensic log storage toolchain.

Diamonds in the rough

Security event log collection often comes down to a compromise. Logging is not the core business for anyone other than a few strange outlying organizations like the team that created Snare. On the surface, most IT administrators would struggle to provide a rationale for collecting and storing the entire raw log data feed that operating systems, services, IoT devices and applications can potentially produce. Justifying the human and computing resources required to comprehensively trawl through such logs would be similarly challenging for all but a few high security organizations. However, in situations where an active intruder has been detected or if a historical breach has been identified, access to the widest range of possible data to support an investigation is sometimes crucial. The critical nature of the incident will generally mean that hardware and personnel resources can be easily justified to determine the scope of the problem, and therefore reviewing a wider range of data is both possible and necessary.

There have been situations in the past when non-security log data has been a critical element in identifying a rogue actor's path through an organization, particularly when stakeholders have to be consulted months after the event to determine whether actions on a particular machine might have been legitimate or not. Sometimes, context is king – and retaining a broad and deep range of security and non-security logs can be the difference between success and failure when tracking an incident or providing court-quality data for legal prosecution where chain-of-evidence and verifiability is crucial.

Gartner, in their "Top 10 Strategic Technology Trends for 2020" report, highlighted artificial intelligence (AI) security and machine learning (ML) as key components to understand patterns uncover attacks and automate parts of the cyber security processes. Machine learning algorithms function well in a rich data environment. Collecting and storing a wide, deep and chronologically diverse range of log data can enable a machine learning inference engine to detect indications of compromise when a narrower scope of data may not.

Digital Archaeology

Information technology moves fast. In the year 2000, if you told someone that MySpace, Yahoo, AOL and Nokia would be practically relegated to footnotes in history within 15 years, you would have been laughed out of the room. Time is littered with the detritus of once massive digital giants like Wang, MicroPro, Lotus, Corel, Palm, Commodore, and Blackberry. The rise and fall of these entities are a useful cautionary tale when consideration is given to storing long-term forensic log data in proprietary or hard-to-access formats.

In situations where a significant system or data breach has been encountered, third parties - either consultants or federal government representatives (FBI/ASD/GCHQ/CERT) may be available to assist with the process of incident identification and mitigation. At that point, easy access to forensics log data can significantly reduce the time to identify and time to react for an organization when taking advantage of the extra resources. From raw 'greps' at the command line to AI data corpus uploads, complex data analytics and visualization tools, using common or open logging storage formats significantly increase the range of tools available to conduct quality analysis.

Key Points

1. The time between breach and detection can be significant, often going beyond the time that many organizations retain log data.
2. Collecting, analyzing and storing forensic-quality log data can be expensive using tools that are designed to process a limited log data volume or duration; particularly those that have a per-event or per gigabyte cost model. This will have the unintended impact of companies not storing valuable information that would be needed for any forensics investigation.
3. The availability of personnel (internal or via a third party) and processing resources increases significantly when a breach occurs, but they are of minimal benefit unless you retain forensic quality data for later analysis.
4. Beware proprietary log storage lock-in to your data. Access to, and portability of, the data is a key requirement.

Operating large expensive clusters of servers

Many systems that are designed to analyze and report on event log data are very frontend heavy. Systems are often optimized for query performance on low volume data stores and are architecturally unsuited for forensic storage or mass data query. This generally implies:

- Compression rates are low.
  - In many cases, a single event of 1k in size may result in 1.5k of data, index and metadata, once on disk. Replication increases that number again.
- Ingest rates are very low
  - The requirement to analyze, index and extract intelligent meta-data from each and every incoming event generally means that any individual collection node can process a very small number of events per second.
  - Organizations with large EPS rates will often need many collection nodes for their operational collection rates or will need to find ways of cutting down the volume of source data.
- Query speeds are high, but as volume increases speed decreases, and the performance impacts may not scale linearly with event rates and storage volumes.
- Hardware requirements can be significant when moving beyond baseline functionality.
  - Using the sizing guide available for some log storage solutions, assuming 10,000 events per second, and applying the log retention requirements of HIPAA/SOX, the following approximate hardware requirements are recommended:
    - 8+ ingest servers
    - 3+ petabytes of storage

Investing significant time and effort to filter incoming event data, in the hope of retaining just enough log information to preserve a reasonable long-term picture of the historical security posture

- Filtering data generally shifts processing from the servers to the systems which generate the data and reduces event volume.
- Decreases tool costs, for solutions that charge per-event.

Specialist forensic log storage toolchain

- In parallel with your day-to-day security solution, with significantly extended retention settings, or
- As a filtering proxy server that pushes a limited range of data to your primary analysis engine, while preserving forensic data.

About the authors

The Snare team members are specialists in the area of computer forensic log data and have been at the forefront of technology and processes for over 25 years. We work in high security/high threat environments to claw actionable threat intelligence and forensic tracking out of system, application and device logs. Our software is deployed in a wide range of industry segments including defense, intelligence, healthcare, logistics, finance, and government agencies, and is designed to work well in high volume, resource constrained environments such as deployable assets, air-gapped environments, SCIFs or segregated networks.

Contact us

Toll Free US: 1 (800) 834 1060
Denver Office: 1 (303) 771 2666
Asia Pacific: +61 8 8213 1200
UK/Europe: +44 (797) 090 5011

Head Office

Level 1, 76 Waymouth Street
Adelaide, South Australia 5000, Australia

ABN: 84 151 743 976