WHITE PAPER

Enterprise Logging: A Best Practices Approach

GUIDEPOINT
SECURITY

At some point, whether first getting started or in existence for a long time, each organization must ask:

What reasons should we start logging?

What is the goal we are trying to meet?

Are we logging only because it's a "best practice" or "everyone is doing it?" What will the logs be used for?

Is it for security only? Security and compliance? Or other factors?

If for compliance, which regulation? SOX, PCI, HIPAA, something else?

Logs are the lifeblood of security and compliance for any organization.

Blue Team members, the defenders of the enterprise network, live or die by the quality of the logs they can collect or can view in their analytic tools. Auditors use logs to determine how compliant an organization is to mandates. Logs are often the basis of alerts and investigations and analysis and reports for various interested groups such as Security, IT, Insider Threat, Executive Management, Engineering, and even sales. All of this depends on the quality of the logs generated and the dependability of the logs collected.

This challenge was difficult 20 years ago, but now can be downright overwhelming with the number of devices and applications capable of generating logs.

The objective is to figure out how to take all the collected data and seemingly chaotic, disparate details and transform them into something useful to meet the organization's specific needs. Knowing why logging is done will drive decisions on what to log and how to log. This will go a long way toward taking all that generated information and putting it to the best use possible.

What is your logging goal?

You need to ask this because the reason or reasons will drive strategy for enabling, collecting, storing, providing access to, analyzing, and archiving every log in the environment. For example, if one of the logging goals is compliance, many logs will need to be collected that no one will look at on a daily, weekly, or even monthly basis. Should those logs be stored in the SIEM and counting against your license, or should they be sent to a separate data lake? That is a decision that needs to be made in that use case.

If there are multiple goals for which you are collecting logs, does the chosen logging solution provide the ability to meet all of them? If not, is there a road map by which the solution can evolve to meet those goals? One thing NIST 800-92 recommends is to create an organizational policy and/or procedure dictating what must be logged and give direction for enabling and collecting those logs. This recommendation ensures that the organization created a road map to meet the goals set.

What are your 'crown jewels'?

Next, you need to decide what it is you are protecting. To do that, you must understand the business of the organization. What is it about your organization that is its reason for existing? From the perspective of security, this is what must be protected. That protection then includes the people (users), processes, and data critical to those crown jewels.

For example, if the organization conducts R & D, log the activity surrounding the testing, results, analysis, and researchers, which are, in essence, the crown jewels of your organization. It could be IP, customer data, or financial processes, and may even be more than one thing.

For example, a manufacturing plant may have IP information regarding the unique component it is building, and it could have financial bookkeeping/ billing processes that need protection.

Knowing what needs to be protected is closely aligned with learning your threat vectors. That is, do you know what are the most likely avenues of attack and who is most likely to attack? You can determine this by conducting a threat modeling exercise. However, don't get so focused on only activity surrounding the crown jewels. Remember that attackers can move laterally once they gain a foothold. So, understand the paths to your crown jewels and ensure you can monitor them as well. This could mean monitoring for such activity as lateral movement and unauthorized network scans from administrative assistants, security guards, interns, and others not directly connected to your crown jewels.

Evaluate Monitoring Devices

Now that you have determined your logging goals, what your crown jewels are, and the potential threats against your organization. It's time to evaluate your security stack to determine if the devices deployed can provide the logs to monitor the activity needed to fulfill those requirements. **Some example questions to ask are:**

- Can the devices log attributable user activity?
- Can they log changes to critical data?
- Can they log who accessed the system housing your crown jewels?
- And most importantly, do the devices protect the logs they collect both on the device itself and when sending it to a centralized location?

It's best to prioritize attributable sources over non-attributable ones. But if the only logs of an event you have do not attribute information, you should collect them nevertheless. It is better to have a record for an event than have nothing at all, plus there are manual methods of attribution. Attribution can be "`user - event`" or "`machine - event`" or "`user - machine - event`." Since attribution can come from many sources, all logs should be appropriately synchronized with a centralized time server.

Decide on your Roles and Responsibilities

The planning (or review) process for your logging solution must include evaluating roles and responsibilities. All individuals, contractors, and consultants working on an organization's security must understand their roles and responsibilities regarding logging and incident response. Who is responsible for enabling logs, creating filters, creating use cases, and/or investigating alerts?

Understanding the roles and responsibilities and having them laid out before setting up logging is essential. However, be careful that the processes to change logging and collect new logs do not become overly complicated and filled with red tape. Due diligence is prudent, but onerous approvals, reviews, and paperwork prevent needed changes and often cause devices and applications not to be monitored for months. All the while, the approvals languish in someone's inbox.

Whom to Tell and What NOT to Do

Make sure roles and responsibilities have been documented before dealing with logging. Understand what is being protected and what options defenders have when they see activity in logs that indicates something is wrong.

From there, determine who needs to be informed about the logs and which individuals can take action. It may be the group dedicated to reviewing records or another group or division. Have a decision process created beforehand to determine if and how somebody should notify the Insider Threat team or others because they may not have access to the information. Engage with HR or the legal team, depending on what is going on. And if a current attack is discovered, it is NEVER a good idea or ever recommended to hack back.

Let's Talk Strategy

Next, decide on a strategy to collect logs. Justin Henderson, the author of SANS SEC555[1], believes there are three general strategies, Input, Output, and Hybrid.

1 Input Method

When you think of the Input Method, envision collecting everything from anything that is generating a log.

Pros

In one sense, this makes some things easy. You don't have to worry about deciding which log is more important than the other. There are some advantages to this method because if everything is collected and an analyst needs to find a particular log or item, that information will be there. For example, if a new security threat is discovered, you probably will be able to conduct an incident response because you have the data and the information.

If your primary motivation for logging is compliance, you will almost always be compliant because everything has been collected and retained.

Cons

The main challenge associated with the Input Method is the cost. Collecting all the data will incur a high price for storage since all the amassed information will need to be warehoused. Further, the organization will pay for the significant expense of a license.

No matter what license model is being used (by compute, by EPS, or by storage per day), collecting everything will drive up the cost. Compute is a sneaky cost, because to efficiently search that much data, a lot more power is needed to process all of it to prevent **v e r y s l o w** searches. Therefore, those related costs will also rise to compensate.

2 Output Method

The Output Method is when logs are collected based on what you determine you want. The organization only obtains records for those use cases, threats, attacks, and other behaviors that the organization has decided are most urgent and nothing else.

Pros

One of the significant advantages of this method is that it provides a speedy, very efficient logging solution. You can keep your license costs, compute costs, and storage expenses down, while the process continues to chug along. Searches will remain very fast because the indexes will remain relatively small.

Cons

There is a vulnerability issue that comes with this method. If there is ever a threat or incident that does not meet the requirements of predetermined use cases or threats, you probably will not have the data to respond and conduct a response or investigation.

1) https://www.sans.org/cyber-security-courses/siem-with-tactical-analytics/

3 Hybrid Method

The Hybrid Method is a compromise and a merging between the two previous methods, which will result in a “best practice” outcome for most.

Pros

This method allows you to build a customized solution for your enterprise. For each category or device, start by collecting everything, then review that data to determine what is actually useful for your goals, protection of the crown jewels, and detection of determined threats. Some of the data may align to the current use cases, but some may not.

Any data that does not serve to help with the above can be discarded prior to collection. You may choose to leave it on the device itself or to disable logging of that particular event completely. This creates a more streamlined collection of information for the organization.

Cons

The method requires a continuous time and effort investment, and you will have to decide what the purpose of your tool is going to be, whether it’s geared toward compliance, or tactical processing. This method also REQUIRES a constant review of the data you ingest so that you can filter out (i.e., stop collecting) data determined as not useful. This method means your costs will be less than the Input Method, but unlike the Output Method, there will be a higher likelihood that with new attacks there will be logs of evidence collected for investigation.

The Hybrid Method will create the best collection of logs for monitoring and protecting your environment, but it requires you treat the log collection effort as a program, not a project. That is, view this effort as ongoing with continual maintenance and improvements. Remember, nothing in your network is static and needs change over time, as do rules, regulations, laws, applications, and networks. This consistent review will make sure that the logging you are doing today is still relevant, no matter when you began logging.

The Tactical SIEM: A type of logging solution where the SIEM is very narrowly focused and precisely tuned to collect what you want.

The Compliance SIEM: A Type of logging solution where the goal is to collect everything and store for retention.

Most organizations have to choose between one or the other types of SIEM. There is another option, but this one is usually only available to larger and perhaps more well-funded security organizations. These organizations could create two logging solutions. One for compliance and one for tactical response. This is often the case where an organization needs to collect everything to meet some sort of regulation, but their security team needs only a portion of that data to protect the environment.

In this case, those organizations that can afford it can set up two destinations to send logs. The compliance solution is a data lake where everything is dumped, with the expectation that rapid investigations cannot happen due to the large amount of data. However, that amount of data involved makes for a great resource for Machine Learning and other advanced analysis tools.

The second destination is a tactical SIEM, which could be configured per the Output Method or better, the Hybrid Method of data collection. Indeed, the primary factor for logs sent to this solution will be to meet known use cases. This is where security analysts would spend their time ensuring the organization is secure while they monitor for attacks. Then, if an event occurs that is outside of the use cases designed for tactical solution, the investigative team can retrieve the information they need from the compliance solution and continue with their response.

It is imperative to understand an organization's resource capacity when designing or evaluating the logging solution. The key components to review are:

- ✓ **License**
- ✓ **Storage**
- ✓ **Compute power**
- ✓ **Personnel**

License

How much license is needed for the method you have chosen? How much license can you afford? And can you afford to grow that license by at least 20% year over year? Plan for that amount and map the increase over time as more information will be collected.

Compute Power and Storage

Determine the amount of compute power needed to process all the information you are collecting and calculate whether you have enough storage for the data. This is especially important if you are going to host the logging solution in the cloud. If not carefully calculated and monitored, costs may skyrocket out of control. You do not want to have any surprises in the process, such as believing you have a year's worth of logs only to learn later that only four months of records have been retained due to a storage limit.

Personnel

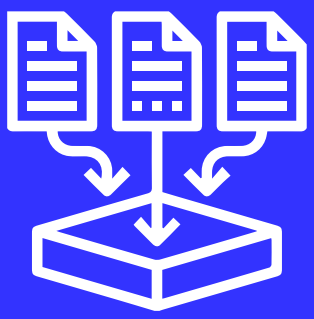
An often overlooked, but vital aspect to consider is the amount of personnel assigned to the system. The larger and more complex the logging solution, the more resources it takes to maintain. This applies to both the administration of the system and to those monitoring dashboards and responding to alerts. There is a tough challenge here that many organizations struggle with as they work to balance FTEs vs skills vs budget. Carson Zimmerman, in his book Ten Strategies of a World-Class Cybersecurity Operations Center, recommends quality over quantity in personnel.[2] A person with the right attitude, skill set and is trainable (meaning they don't need to start out as SMEs but can grow into that role) can outperform multiple people who are just butts in seats and are more cost effective in the long run despite requiring a higher salary.

Design and Implementation

Once a collection strategy has been decided upon, the organization needs to design and implement technology to collect the logs. There are multiple technical solutions for gathering logs. Many logging solutions provide their own agents or applications. These have varying degrees of capabilities and come with the assurance of technical support.

There are also several open source solutions that may offer just as many, if not more, capabilities as the vendors. Additionally, some open source solutions do offer commercial support contracts (i.e. nxlog). Deciding which is best for the organization will involve understanding the architecture of the enterprise and the tolerance for modification to current infrastructure.

2) https://www.mitre.org/publications/all/ten-strategies-of-a-world-class-cybersecurity-operations-center



Some critical factors to consider when deciding on collection technologies include:

- How does the collector operate? Via Agent, Agentless, or API?
- What capabilities does the collector have?
  - Queuing
  - Filtering
  - Routing
  - Extraction/Parsing
- Is there a clear and simple network path between source and indexers?
- Is there a method of load balancing data across the indexers? Across the collectors?
- Can the logs securely transmit all data?

COLLECTION TECHNOLOGY FACTORS

Agent/Agentless/API

You may opt for Agents, Agentless, APIs scripts, or ODBC, maybe even a cloud trail or some other capability in the cloud arena. These are all great methods and the plethora of choices may have the reader wondering which to employ. In this case, it is my opinion that best practices are whatever process works to get the logs you need into the log collection solution so the analysts can analyze the data and work to improve security.

For example, at one organization, I started out collecting Microsoft Windows logs via WMI calls, which is an agentless method of remotely 'pulling' log data. I used this method because I was not allowed to place an agent on the Domain Controllers at the time, but I needed the logs from those devices as they were key to populating several of the critical use cases I had created. So, I used the agentless, remote log pull method of WMI calls. There are several challenges with this method, among them is the fact you are passing administrator credentials over the wire multiple times an hour and if you are monitoring a busy server (for example, a DC), you can miss logs. Eventually I was able to show the value in the logs I collected and that by continuing to miss logs, I was placing the organization at additional risk. After the realization, I was given permission to place an agent on the DC to ensure that all logs were gathered in a timely manner.

Queuing

As the logs are collected, the ability for the log collection to have some queuing capabilities is vital, particularly if the network gets congested or has some kind of hiccup. You do not want to lose logs in transit from collection at the source or to storage at the index. This can be part of the vendor log collection application, if you are playing within the walled garden of a particular solution (i.e. Elastic with Beats and LogStash or Splunk's Universal Forwarders). These applications are built to communicate with each other and say "slow down" or "stop for a while" and the log collector will automatically queue logs until it receives the go-ahead.

However, suppose you are not playing totally within the walled garden. If a combination of solutions for collecting and transmitting those logs has been chosen, the organization should find a way to build some kind of queuing capability into the pipeline. Whether it be with RabbitMQ, Kafka, or another solution, these provide the assurance that if the network gets congested and/or connections are dropped, logs would not be lost in transit. Along these same lines, you will want to make sure that the logs are stored and not deleted during reboots for any of these log collectors.

Filtering

Filtering is another item that deserves your attention. This speaks directly to using the Hybrid Method, where you decide what logs and/or events are not relevant and therefore do not need to be collected. The closer to the log source the filter can be applied to drop unwanted logs the better. This lowers the processing load on the servers handling the transmission and indexing.

Routing

Routing is akin to Filtering and is also vital to the entire process. If you have a particular set of logs that you need and only some useful for operational purposes, just send that subset to the system's logging solution. The rest can be sent to an enterprise data lake. If you can route at the source, where your collection point is, only send some of the information to the actual log solution. This can cut down on network traffic significantly.

Extraction/Parsing

Where possible, choose a log format where data can be extracted based on the type. The most common log formats that provide this capability are JSON, KV, and CSV (sometimes XML, but that format has its own challenges). Most other log formats require some level of parsing. Parsing usually means using regular expressions (REGEX) to analyze the log, compute what parts of the log represent fields and values, and then assemble the metadata in the index for searching. This is a lot of processing and slows down the pipeline if many of the logs require their data parsed out vice extracted. Extraction saves processing capabilities for the advanced tasks and analytical calculations.

Examine the network architecture between the different sources of log data and the final destination. Determine how to send the logs via a clear and straightforward network path to the log collection destination (often the indexers or database cluster servers) wherever possible. The simpler the path between the source and the indexers, the better because you will probably have multiple indexers, and you will want to load balance along those.

Usually, the vendor log collection application will do this automatically (as long as you have configured it properly). However, if they do not, or if the architecture does not use the vendor software, an actual load balancer can be employed (i.e. F5). Avoid using DNS Round Robin, since it does not know if one of the servers on the other end goes down.

Whether doing load-balancing and/or routing, make sure to use a DNS name or names for the targets, not IP addresses. It is much easier to swap out the endpoints because it is a simple DNS change and not modifying IP addresses on four to five or even 400 different systems.

Configure the transmission of logs so that it is secure and/or encrypted wherever possible. It will do no good to lock down access to the logs at the source and at the log collection destination if an adversary can read all the data in the clear off the wire. Again, many vendor solutions have this capability natively, but always check regarding the exact method of configuration. Mistakes can result in data not being sent at all or default to be in the clear. Additionally, both rsyslog and syslog-ng have the means to encrypt transmissions via tls.

Configure a centralized time server so the events are synchronized across your enterprise. Unless the enterprise is entirely based in one time zone, synchronize your time to UTC. This makes it easier to correlate events during an investigation. If possible, the synchronization should be at the endpoint. If not possible (people like their Windows workstations set to local time), convert the time to UTC during indexing when that data is put into storage. Don't worry about confusing your analysts because most front end graphical user interfaces for the logging solutions will convert time stamps for users. They will change the view of the logs to the user's time zone when they are doing searches, but on the back end, it remains unaltered in UTC.

Normalization

Once you start collecting logs, the most critical action that needs to be done is to normalize that data.

Normalization means changing the fields and the field names so that they are similar between log types and event types. Some log collection solutions perform this normalization for you (i.e. ArcSight), others require that it be applied through administrative configuration using their supplied schema (i.e. Splunk CIM, Elastic ECS) or one the organization has designed itself.

This ensures that no matter what a monitoring vendor has named a field, analysts only need to search for a single term for each field type.

For example, when a desktop sends data to a server, a firewall may label the desktop IP as src, but the router may label it as srcip, while the network traffic analysis tool may label it as `source_address`. This makes searching complex. Normalization renames all those labels with a common name, perhaps `source_ip`. Now the analyst can use one term and retrieve all related events.

Logging Enrichments

You may not think that log enrichment is part of best practices, but the opposite is true.

By adding context to logs, they will be customized to your specific organization. This allows analysts to quickly understand the context surrounding alerts or reports and investigations. With enrichment comes understanding, knowledge and a means for making better decisions concerning the data and information collected.

Ways to enrich your logs include adding:

- ✓ **GeoIP / ASN information**
- ✓ **Frequency Analysis of fields**
- ✓ **Domain Information (i.e. Creation Date, Top 1 Million Rank)**
- ✓ **Threat Intel**
- ✓ **User Context (i.e. admin, user, department, area of responsibility)**
- ✓ **Other tags regarding specifics about your environment (i.e. office location, department, subnet description)**

There is a wide variety of methods that can be used to enrich and add more value to logs. It is not just, for example, noting that an IP was from Russia or France, but it is adding additional items such as an ASN number that brings so much more value in terms of information. GeoIP may find an IP that comes from China, but with the additional information in the ASN, you can determine that it belongs to Netflix or Amazon. While you still may be cautious, you at least now have more than just a country code, you know who owns the IP block and not just where it is from. Additionally, this information provides an excellent context for conducting frequency analysis.

For years, we have heard how malicious actors create domains with random domain names, and that is how they can get past security using DGA (Domain Generation Algorithm) hostnames. They quickly create and register new domains with random numbers and letters and rotate connections through those domain names (More details here). Frequency analysis enables the detection of those random domains, and is not limited to just domain names, but can be done on user names, service names, and machine names for when attackers create malicious versions of those entities via automated tools.

For example, consider X-509 certificates. Network protocol analysis tools will carve out your X-509 certificates from TLS traffic. Those certificates have common names and issuer names which should be similar to domain names and English type names. Attackers will create certificates to encrypt data that they will exfiltrate or use to encrypt their C2 traffic (command and control communication). In generating their certificate, they need something to fill in those fields, so they just use random information. Frequency analysis can be used to detect those certificates, because they are traveling in the clear across the wire. All that is needed is to pull out that information, run it through a frequency analysis tool, and boom! Something interesting to investigate was found right away!

Free frequency analysis tools include: freq_server written by Mark Baggett from SANS. There are also ways to integrate this tool with multiple logging solutions.

Domain information can add contextual value to http and dns traffic. By adding `whois` information such as the creation date and the domain ranking, according to the top one million rankings, analysts can get an additional insight when reviewing alerts.

For example, if an analyst received an alert and the information on the domain shows the domain was created last week and it is not in the top one million plus has a meager frequency analysis score, that scenario implies the domain is more likely to be malicious and an investigation should be started.

Tags are another great way to add customized environmental context to your information. You should tag as much of the data collected as possible. Some examples include whether the IP is an internal IP or an external IP or belongs in the DMZ. Does the IP belong to the R & D group, or Finance or HR? Is it a server or a workstation or a router or something else? Users can be tagged with their role, their office, and/or their department. Tags will help provide context if the event is normal, expected, unusual, suspicious, or alarming. That context helps analysts, automation, alerts, and reports filter data because it is customized to your environment and will provide better insight.

For example, should the HR supervisor be logging onto the research machine located in the DMZ? That kind of context is much more suspicious than jdoe logged into Server X with IP 1.2.3.4.

Log Integrations

Integrating your logging solution with other capabilities may not be considered in the best practices area. Still, it should be something to be considered when either purchasing a logging solution or looking at doing something different with current logging solutions. Suggested integrations include:

- Ticketing Applications / Case Management
- SOAR (Security Orchestration Automation and Response)

Ticketing and Use Case Management integrations can speed up the process of turning alerts into actionable tickets. This can ensure that all the data from the logging solution that generated the alert is populated in the tickets. It will also make it easier for investigators to pull the alert data or information related to the alert into your ticketing or use-case management system as they conduct further analysis.

Security Orchestration Automation and Response (SOAR) is quickly growing more prevalent as organizations look for ways to handle the high number of tickets being created and ease alert fatigue. Usually integrated with the SIEM, organizations use the tool to perform additional alert enrichments (adding user data, performing related searches, querying asset databases, gathering threat intelligence). If your organization does not have a SOAR, ensuring the logging solution can support a SOAR will position the organization better for this future enhancement. The more integration you have between your SOAR and your logging solution, the more your analysts will be freed up to do some more complex activities that will enhance your security posture. By freeing up your analysts, they can focus on higher-level, more analytical tasks that cannot be automated.

Log Review

Regardless of which logging method chosen above (input, output, or hybrid) logs should be reviewed regularly. Just because a review was completed when logs were first configured to be ingested is not good enough. In larger organizations, there are times when multiple devices may be logging the same activity. This is good for defense in depth, but may also tax log collection systems. In those cases, the organization needs to decide if they are going to collect only one set or multiple sets of those logs. Reviewing logs as compared to use cases, current threats, changing regulations and requirements ensures that only the logs needed are being collected. This prevents license exhaustion and overuse of resources such as compute and storage.

Finally, a corporate decision must be made regarding how long to keep your logs. This may be determined by storage availability, compliance and/or possibly legal requirements. Don Murdoch's publication Blue Team Handbook has a summary chart of log retention requirements in the Log Management section for some of the major regulations in the US that could affect your industry.[3] Outside of these requirements or resource restrictions, you must decide how long past events need to remain online? One week, three months, one year? This question is not as simple as it seems. The timeframe can change per event type and use case.

For example, DHCP logs may not be very valuable after a week and therefore can be purged from online storage. However, authentication logs may have value for several months or even years. Be aware though if you are using machine learning (ML) to find anomalies, you may need several months of data in order for the analysis to be done right. Then, once logs are removed from the online repository, do they get deleted or moved to archived storage? That storage could be compressed files locally, in the cloud or even tape and stored completely offline (yes, tape still exists).

3) https://www.amazon.com/Blue-Team-Handbook-condensed-Operations/dp/1726273989

Conclusion

Ultimately having a logging solution is not a one-time option or even a six month project. It is not something you want to implement and simply forget. You will need to make a constant effort to improve and maintain it continuously. The benefits will be worth all the time, resources, and energy you have invested in the end. You will have a customized solution that will meet your organization's specific needs while providing a safer cyber environment.

To implement the best logging practices, you should:

- ✓ Consider implementing the Hybrid Method of collection
- ✓ Review your logs regularly
- ✓ Manage your resources
- ✓ Perform log hygiene
- ✓ Add context to your logs

ABOUT THE AUTHOR

Craig Bowser

Craig Bowser is the Federal Practice Director, Data Analytics for GuidePoint Security. He has some 20 years of experience monitoring and securing networks and performing in-depth analysis and providing the types of solutions that allow organizations to complete their missions and security objectives. Craig also teaches SANS SEC555 SIEM with Tactical Analytics, where he instructs students on learning to extract actionable intelligence for a tactical SOC.

GUIDEPOINT SECURITY

2201 Cooperative Way, Suite 225, Herndon, VA 20171
guidepointsecurity.com • info@guidepointsecurity.com • (877) 889-0132

WP-LOGGING-102020-01