SANS ANALYST PROGRAM

Sponsored by Loglogic

SANS Annual 2009 Log Management Survey

A SANS Whitepaper – April 2009

Written by: Jerry Shenk

The Importance of Log Data

Who's Collecting Log Data

Why Is Log Data Collected?

Where Is Log Data Collected From?

Log Data More Valuable

Ongoing Challenges

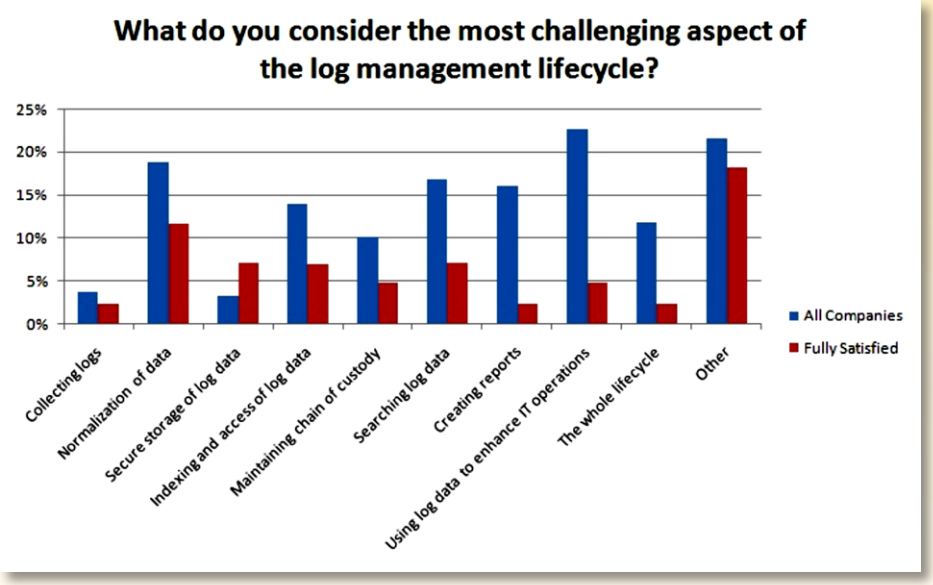
Traits of Successful Log Management Initiatives

Executive Summary

The SANS Analyst team has conducted a survey of the Log Management Industry every spring since 2005. These surveys have given us insight into why people use logs, what problems they encounter with log management, and what user organizations would like to see from vendors.

In January, Alan Paller of the SANS Institute commented on a Department of Defense study reporting that log management controls were ranked high among controls that could have blocked or kept attackers from getting a strong foothold in target networks.[1] Over the past two years, our surveys have shown that organizations are realizing the value of log management controls because they know that valuable information exists in log data for security and operationas. We are also seeing that mature organizations are beginning to use logs for these more advanced purposes.

Until this year, the majority of companies had trouble just collecting log data. In the 2007 survey, more than 50 percent of respondents reported having the most problems with collection, while this year only four percent cited collecting data as most difficult. Now their problems have shifted to higher-level concerns of normalization, indexing and access, creating reports and the whole lifecycle of log management.



[1] www.sans.org/cag

Another notable change in 2009 is that organizations are using their log data for more varied purposes than in the past. Top uses for log data this year were "Tracking suspicious behavior and user monitoring" and "Forensics and day-to-day IT operations." Proving compliance with regulations was also cited by more companies than in the past—53 percent this year versus 43 percent last year.

One problem that has been consistently difficult for companies to address is log normalization because most devices and operating systems log data differently (if at all). This makes it difficult for a log analyst to compare similar events without a lot of manual, multi-system expertise to read and interpret each log format. Vendors need to take logging seriously in their operating systems, applications and appliances and adopt a common syntax, such as Mitre's CEE, a developing, vendor-neutral common syntax for log expression, the first of which is expected to be released this summer.

With most users satisfied with their log management systems this year, the respondents indicated that the key factor to the successful log usage is stronger management buy-in of log management for security and operations. This is demonstrated by setting measurable goals and by making log analysis a part of the normal workflow, which we're seeing with log management integrations into SIEM (Security Information Event Management) and virtualization platforms, according to survey results.

[2] http://cee.mitre.org

The Importance of Log Data

Companies typically find out there is a problem on their networks when something breaks. Often, their systems haven't been configured to log all the available data. So to pinpoint and then repair the problem, they need to reassemble what went wrong. That's not very efficient.

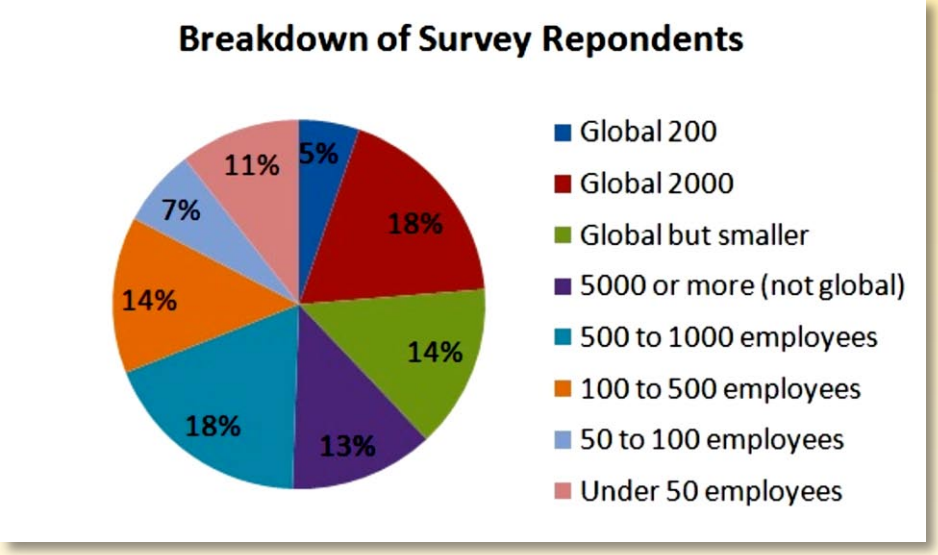
Most current computer systems generate some type of log data—data that could announce the onset of a problem long before there is a system outage. Because many organizations are now paying more attention to deriving the most value for their dollars, IT departments can also use logs to increase efficiencies. In addition to identifying outages before they affect the bottom line, logs can also be used to improve the overall health of the network by consolidating log data to create alerts and troubleshoot vulnerabilities in need of patching and repair.

Regulatory requirements also place significant importance on log data. According to past surveys, many companies start down the log management path due to requirements such as PCI or SOX. Then they soon find out, thanks to the information provided by those same logs, that there are things going on they weren't aware of. This awareness at the management level often leads to a more efficient network, better repair processes, fewer outages and a more timely resumption of operation when there is an incident.

Who's Collecting Log Data

Respondents to this year's survey include a balanced mix of IT management/security and IT staff/security positions from companies of a variety of sizes in a wide range of sectors. Many respondents indicated that their duties include both managerial and administration duties, with 30 percent indicating that they only have management responsibilities. The most heavily represented industry was finance (20 percent) with education, government and telecommunications garnering 16, 13 and 10 percent, respectively.

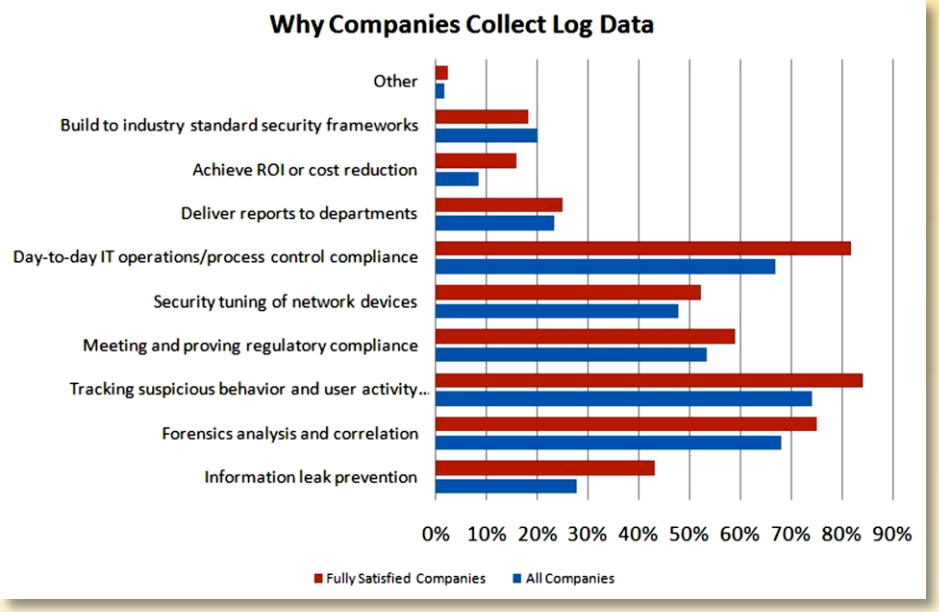
With respect to size, 24 percent of respondents are from the "Global 2000" group (includes global 200), with 14 percent representing "Global but smaller" companies, 13 percent from organizations that are not global but have 5000 or more employees, and 32 percent representing small and mid-sized organizations (100 to 1000 employees).



Why Is Log Data Collected?

The response choices used in the question about why log data is collected have been expanded in this year's survey, so it isn't possible to directly correlate the responses from this year's survey with the 2008 responses. In 2008, however, the top issue was "Detection and analysis of security and performance incidents."

This year, respondents picked "Tracking suspicious behavior and user activity monitoring," followed by "Forensic analysis and correlation," and "Day-to-day IT operations/process control compliance," which are similar to the responses in the 2008 survey. The chart below compares the fully satisfied companies against the total group's reasons for collecting log data.



Compliance was not as high a driver this year, with just under 60 percent collecting logs for this reason. Previous surveys have shown that compliance has been the primary reason organizations begin to pay attention to their logs, with regulatory requirements such as PCI, SOX, GLBA and others mandating the collection of logs to protect personally identifiable information (PII). Now we're seeing that organizations then grow in their use of log data from merely checking a box on a compliance survey to finding ways to derive other value from their logs for security, network and business operations.

"Information leak prevention" is another new option in this year's survey. It was selected by 28 percent of respondents, which shows growing interest in integrating logs with information-centric protections.

Another new option in this year's survey, "Deliver reports to departments," was chosen by 23 percent of respondents. Those organizations that are delivering reports outside the IT organization are the same organizations that responded as the most satisfied with their log management—an indicator of more mature log management integrations. As log management has matured in these organizations, the data has become valuable to a variety of different groups for security and operational purposes.

Good log reporting features enable IT departments to distribute useful data to other departments without revealing raw log data that could have sensitive security information. An even bigger advantage is that the complexity of the raw log data is reduced to reports and charts that can be useful to departments without forcing them to learn more about the IT intricacies and the format of different log generators.

Where Is Log Data Collected From?

Respondents are collecting log data from a growing list of applications and devices. This year, 35 percent reported collecting logs from between 10 and 100 sources, and 22 percent collected logs from 101 to 500 sources. Similarly, last year, the most popular choice was "Under 100" sources, with the second most popular choice being 100-250.

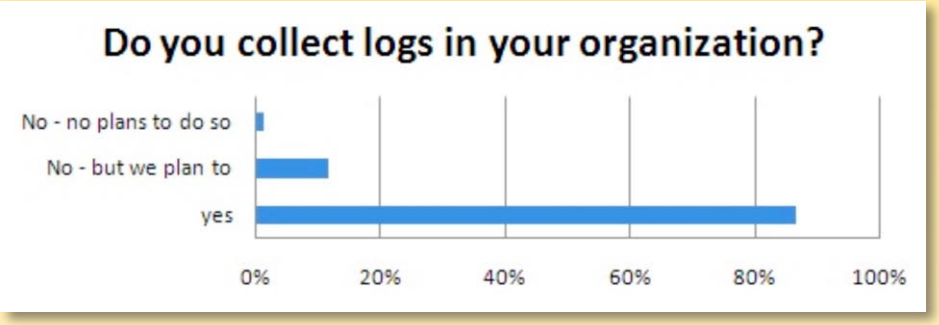
Most commonly, respondents are collecting logs from operating systems (92 percent of respondents collect this type of data), followed by switch, router and firewall data (90 percent). Those were the top two choices in last year's survey as well, although a slightly lower percentage chose each option. The largest decrease this year was in the number of organizations collecting log data from mainframes—down to 18 percent from 25 percent in 2008.

On the surface, log collection from databases appears to have decreased as well. Of the respondents, 57 percent indicated that they are collecting log data from their database systems, compared with 61 percent in the 2008 survey. However, another 11 percent of this year's respondents selected Database Activity Monitoring (DAM), a new category. So if you combine these two categories, the number of organizations gathering information from database systems actually goes up. Furthermore, this category indicates that organizations will increasingly be collecting logs as they add additional database security tools.

In another new category this year, 49 percent of respondents indicated that they collect log data from virtual machines. In another question, 68 percent predicted that nearly 70 percent of their logs will be coming from virtual machines by 2010. Collecting virtual logs should be a nonissue for good log management systems, even (and particularly) in co-hosting environments where different events belong to different clients.

Log Data More Valuable

In recent years, respondents have placed increased value on their log data. As recently as 2007, 44 percent of survey respondents indicated that their IT groups did not collect log data. In 2008, that number shrank to 27 percent. This year, 87 percent indicated that they collect logs, and 12 percent have log collection in their plans. Clearly, companies now understand that collecting log data is important.

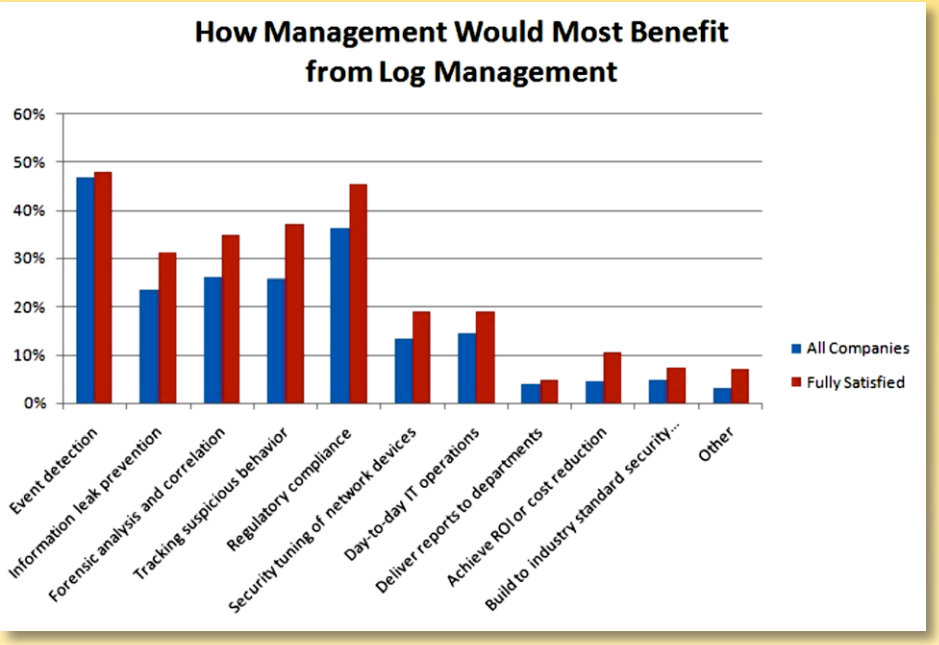


In a recent Department of Defense report[3] about things that could have prevented or minimized the impact of recent cyberattacks on its systems, log management was included as a critical control recommendation. Other regulatory and control initiatives such as SOX (Sarbanes Oxley) and Payment Card Industry (PCI) are also demanding log collection, storage and availability.

Once companies start down the path of log collection and management, organizations often discover there are things going on in their networks that they weren't aware of. They then begin to use their logs to pinpoint repairs. Ultimately, they expand the use of those logs for other purposes, leading to a more efficient network, fewer outages and a more timely resumption of operations when there is a problem.

From a management perspective, respondents to this year's survey indicated that managers' most critical need for log data was for "Event detection," with 47 percent of companies reporting it as absolutely critical and an additional 42 percent rating that use as important. "Meeting and proving regulatory compliance" ranked second among management concerns, with 36 percent reporting that use as absolutely critical. An analysis of the responses from those who were in management compared with those in more technical roles showed almost no difference in the value the two groups placed on their log data or what items were most critical. The only area that was statistically different was "Forensic analysis and correlation," which the management group ranked as more critical than the full group of respondents did.

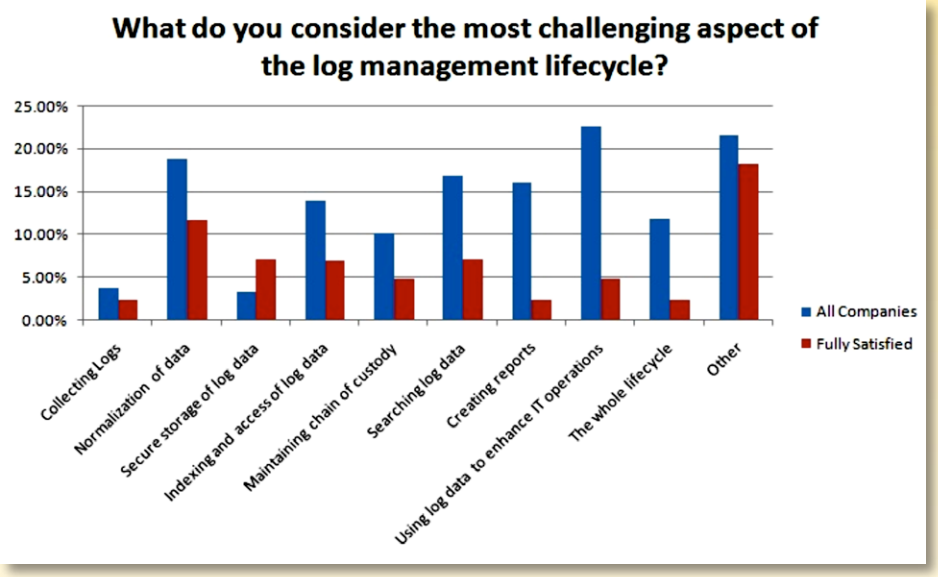
[3] www.sans.org/cag



Log data has also become valuable to Security Information Event Management (SIEM). In another new question this year, 32 percent of overall respondents and 43 percent of the fully satisfied group indicated that they are actively incorporating log management with SIEM. An additional 26 percent intend to move in that direction in the future. This is a logical market progression that analysts have been predicting. Log data has value both from a security standpoint and for IT operations; so it makes sense that SIEM systems use log data as part of their event indicators.

Ongoing Challenges

While surveys indicate consistent growth in the perceived value and usage of log data, this year's survey also shows that respondents are still having difficulty with many aspects of their log management functions. Last year, 51 percent of respondents indicated that collection was their biggest problem. This year, only four percent selected collection as their top problem. Part of this dramatic change may be related to the addition of normalization as a choice, which some respondents might have seen as part of the collection process in our past survey. Two years ago (2007), respondents indicated that the second and third most difficult problems were searching log data and reporting on log data. This year, we added some additional options, but "Using log data to enhance other IT operations" was the top-ranked problem, and "Normalization of data" ranked the second most difficult.



Normalization

Our prior surveys did not ask specifically about normalization, but they did ask about searching data and reporting. Because some of the comments from last year's survey pointed to normalization problems, we asked more questions about those problem areas this year, and normalization and searching log data turned out to be among the most challenging aspects of management.

Normalization is the process of converting data from different formats into a common format that the log management system can interpret, correlate and report against. Most devices log data in different formats (if they log events at all). Log data formats are proprietary to their different system manufacturers, who in many cases use different formats even within their own product lines or between versions. For users upgrading from Windows Server 2003 to Windows Server 2008, the log format changes go way beyond subtle: The event codes have virtually all changed so log management tools that worked for Windows Server 2003 need to be modified or replaced to work with Windows Server 2008. So now, many of the log management systems include options to update the data parsing and normalization features of their products.

Normalization of data makes reporting easier because the raw data from each device is converted into a common format that can then be processed by a single query. For example, a PIX firewall can block data from an attacking IP address, while a Web server can report that nonexistent pages were requested by the same IP address, and an SSH server can report 50 failed login attempts from that same IP address. Getting a single report to include all those failures requires a reporting program that understands the way each application reports a failed connection. An experienced analyst who is familiar with the differing formats of log data can do some correlation of the raw events, but trying to automate correlation before the data has been normalized is difficult for software and humans, alike.

The Common Event Expression (CEE) by The Mitre Corporation may lead to standardizing the way events are logged. The emerging CEE syntax is supported by LogLogic, Splunk and other log management and SIEM vendors, and is also supported by Oracle, Cisco, Microsoft and the Syslog working group, according to Bill Heinbockel, senior security engineer at Mitre who runs the CEE working group. The CEE working group is also in talks with the Open Web Application Security Consortium (OWASP) to cover web application log expressions, and is on target to deliver a first syntax around firewall logs this summer, followed by a syntax for IDS logs. CEE, which is being designed to be customizable and reportable per user requirements, is gaining adoption among international (NATO) and US government agencies, including the National Institute of Standards (NIST). Mitre, which is behind the widely adopted Common Vulnerability Expressions (CVE), has a good chance of fully integrating common expressions over the years. However, progress on the CEE standard depends on vendor adoption rates, Heinbockel adds, which depends on development lifecycles and the necessary re-education of developers.

[4] http://cee.mitre.org

Below are outputs from two firewalls that are not put into a common format: One comes from a Linux Iptables firewall, the second from a Cisco ASA firewall. The connection comes from 123.123.123.123 to a server at 223.223.223.223. Both show the source and destination IP addresses and ports and timestamps, but clearly, the information is in different formats.

```
Feb 22 13:16:31 linux_firewall kernel: IN=eth1 OUT=
MAC=00:60:97:a7:9f:97:00:01:5c:24:0f:82:08:00 SRC=123.123.123.123 DST=223.223.223.223
LEN=40 TOS=0x00 PREC=0x20 TTL=111 ID=256 DF PROTO=TCP SPT=12200 DPT=25
WINDOW=8192 RES=0x00 SYN URGP=0
```

```
Feb 21 2009 22:24:38: %ASA-2-106001: Inbound TCP connection denied from
123.123.123.123/59137 to 234.234.234.234/25 flags SYN on interface outside
```

These simple examples point out the difficulty in normalizing data between two common devices. Some of the data is reported by both devices; however each firewall reports some data that the other firewall did not report. If the log analyst is familiar with both formats, the idiosyncrasies of the log data are not hard to understand. But this one simple blocked packet is difficult to pass on to management without a rather lengthy explanation of what the different fields mean.

Using common syntax, each log entry looks the same, as in the following example of a Common Event Format (CEF)-compliant log entry.[5]

```
CEF:0|Unix|Unix|| |IPTables Event|Medium| eventId=1393304 proto=TCP
categorySignificance=/Informational categoryBehavior=/Access categoryDeviceGroup=/Firewall
categoryOutcome=/Attempt categoryObject=/Host/Application/Service art=1228442591805
deviceSeverity=warning rt=1228442591805 src=123.123.123.123 sourceZoneURI=/All
Zones/System Zones/Public Address Space/198.20.0.0-222.255.255.255 smac=00:60:97:a7:9f:97
spt=12200 dhost=  dst=234.234.234.234 destinationZoneURI=/All Zones/System Zones/Dark
Address Space dmac=00:01:5c:24:0f:82:08:00 dpt=25 cs1=kernel cs2=kern cn1Label=File
Descriptor cs2Label=Facility cs1Label=Module ahost=10.1.1.242 agt=10.1.1.242
agentZoneURI=/All Zones/System Zones/Private Address Space av=4.6.5.5134.0
atz=America/New_York aid=C+tRjB0BABCLCt-wrkHRrg\=\= at=syslog dvc=10.1.1.2
deviceZoneURI=/All Zones/System Zones/Private Address Space dtz=America/New_York
deviceInboundInterface=eth1 deviceProcessName=iptables ad.datagramType=08:00
ad.TOSPrecedence=0x20 ad.Comment= ad.TTL=111 ad.Length=40
ad.DatagramID=256 ad.TOSType=0x00
```

[5] www.arcsight.com/solutions/solutions-cef

Neither the vendor neutral CEE by Mitre or the CEF by ArcSight can resolve the normalization problem overnight. Also, there's some question about a single vendor-specified standard driving the process. So, until the day when all systems are processing log data in common syntax, log management tools must continue to expand their normalization capabilities to support more applications and uses.

The advantages of normalizing log data is to make it possible for a single tool to understand data in a variety of formats (collected from a variety of supported logging devices) without having to code the reporting engine specifically for each log type. Problems with normalization include the possibility that data could be misinterpreted by the normalizers, or that detailed information is not available when needed. Some log managers allow for the normalized data to be stored alongside the raw data. This is helpful from a forensics standpoint and also increases storage space requirements. That said, storage was not an issue this year, with only three percent citing it as a significant problem and 43 percent citing it as least difficult. This is an improvement from last year, in which 24 percent cited secure storage as a problem. According to a new question added this year, organizations are most commonly storing data from three months to a year: This doesn't meet basic PCI and other log retention requirements, which call for maintaining records for a year or longer.

Reporting

Normalization has a bearing on the creation of reports, and reporting is still a challenge for organizations, according to this year's responses. The survey broke reporting into a number of different functions. Of those, "Using log data to enhance other operations/cost reduction," was considered most difficult by 23 percent of survey respondents and difficult by an additional 35 percent. Other related items, including "Searching data" and "Creating reports" also ranked relatively high on in the difficulty question.

When asked to rank their satisfaction with their log management systems, those systems ranking high among the most satisfied users had automated reporting. Log management can be time-consuming, so anything a user can do to make the review of log data more efficient will increase overall efficiency and satisfaction. The most common method for sending those reports is e-mail. Some systems also have an option to allow an analyst to log onto the system and retrieve the previously compiled reports. Many log management systems allow stored reports to be accessible only to those parties who have previously been granted access—and then only to specific reports. As business use for log data dictates, reporting tools should also be accessible based on area of operation such as audit, security or operations.

Log Data Collection

When asked what kind of logservers they are running, respondents indicate they are using both TCP and UDP log servers. Of this year's respondents, 80 percent are using the defacto standard UDP syslog server, and 54 percent of respondents are using TCP syslog servers. There are a number of advantages to TCP-based syslog, including reliability, increased log data throughput and the ability to encrypt the data.

Also in this year's survey, some 58 percent of respondents indicated they have logservers that pull data from their hosts. Many applications log data to flat files, which can sometimes be difficult to collect because collectors have to figure out how to individually handle each application from which they collect logs. Windows servers are one example of a widely used operating system that does not include the ability to send syslog data. On such systems, either an agent needs to be installed on the Windows server or the logserver needs to pull log data from the Windows server.

Traits of Successful Log Management Initiatives

In order to determine overall satisfaction with the log management industry and keys to successful initiatives, we asked respondents to rank satisfaction levels with their current log file analysis. In this year's survey, 70 percent were satisfied, 58 percent were somewhat satisfied, and 12 percent were fully satisfied. In 2008, the question only included options for satisfied and not satisfied, and 36 percent indicated that they were satisfied.

Evaluate Effectiveness

In most categories, the responses were similar between the satisfied and the unsatisfied users. The biggest difference between the two groups was in responses to a new question about the use of organizational measures for effectiveness. Out of the total group of respondents, 37 percent said they measure security effectiveness. Of the group that was fully satisfied with their log management, 64 percent measured the effectiveness of security. Of the group that indicated either full or partial satisfaction with their log management, 47 measured security effectiveness. This is another indicator of the maturation taking place in the most satisfied survey base.

It's also interesting that the same type of controls that work for other business practices work for IT initiatives. The measure that the most satisfied respondents used to gauge security effectiveness was "Time to respond to incidents." The remainder of respondents measured their security effectiveness most heavily by "Incident prevention." Free-form comments by the most satisfied users included number of incidents by class (disclosure, compliance, malware, etc.), cost and impact. Those comments indicate that next generation logging systems and management tools should concentrate on demonstrating these and other measures of effectiveness, including the ability to extrapolate events from log data to decrease response time.

Making Log Analysis a Priority

Another key differentiator between the group of fully satisfied respondents and the total group is that the satisfied group consistently spent time on log analysis and had integrated log analysis into their overall workflow. The survey also indicated that the fully satisfied users knew how much time they were spending on their log management—an average of between a few hours a day and a few days a week, according to this year's survey. Some of the least satisfied respondents spent almost no time, while other unsatisfied users spent a large amount of time with poor results. On average, most companies are spending about the same amount of time on logs as last year, which is a few man-hours per week was the option selected by 45 percent of last year's survey respondents.

Out of the total group, ten percent of respondents didn't know how much time they spent in log management, while none of the fully satisfied group chose that response. Of the fully satisfied group, 32 percent indicated that log management was integrated into the workflow, while 16 percent of the rest of the respondents indicated that log analysis was integrated into the workflow. This is a measure of maturity of log management systems themselves. It also suggests that support for and following through with a plan are contributors to successful log usage.

This same pattern was apparent in the frequency of reports being generated by log analysis. Of the companies that indicated full satisfaction, 43 percent generated weekly and daily reports, while only 29 percent of the remainder generated routine reports. With more actionable data provided by the log management system, business units should expect their own, segregated access to data analysis related to their functions.

Automation Helps

The fully satisfied users indicated that over 90 percent of their collection and storage is automated. These functions were automated by 65 percent of the remaining respondents. Just under half of the fully satisfied group reported that search/analysis and correlation were automated, while only ten percent of the remainder have automated search and analysis.

The fully satisfied users also use either a single third-party tool or a combination of third-party tools and homegrown tools. Of the fully satisfied group, 39 percent use a single third party tool, while 19 percent of the remaining respondents use a single third-party tool. This can be interpreted in a number of ways, one of which might be a higher level of satisfaction with a single vendor system as opposed to a mixed environment. Of course, there's always the trade off with best of breed versus single vendor solutions. In both groups, about one third of respondents use a combination of third party and home grown tools.

Summary

The tide has turned. Log management is no longer a toy for just the geeks. This year's survey included responses from management that demonstrated perspectives similar to respondents in day-to-day IT operations, security analysis and audit. The companies that are most satisfied with their log management initiatives have strong management support. As this year's survey indicates, some keys to success in a log management initiative include getting started, integrating log management into the normal workflow or process, measuring effectiveness, and automating functions like normalization and reporting.

Another implication highlighted by the survey is that while a large percentage of respondents didn't find log storage to be difficult, many weren't storing the data in a way that would be compliant with regulatory standards. Log data storage requirements are implied or explicitly defined by numerous regulations. There are many reasons for these requirements: historic log data has proven valuable for tracking suspicious behavior, for tuning networks and systems, and for forensic analysis.

Log management has gained legitimacy as an important piece of both security and operations on many levels. A recent FTC ruling[6] against Geeks.com is a good illustration. The FTC identified the lack of effective monitoring as a contributing factor to the prolonged leakage of personally identifiable information (PII) including credit card data. Geeks.com had stored the PII in the clear—an explicit violation of PCI standards. Log management, in this case, could have been used to connect the dots between the network and network security devices, applications and operating systems that were being exploited.

In the future, expect to see log management to be of growing importance among business units and IT departments, driven in part through continued regulatory requirements, then leading to improved operational and security efficiencies. This puts more demands on automation, particularly around correlating and reporting to reduce incidents, to enable faster response and to support multiple operational, risk management and regulatory objectives.

[6] www2.ftc.gov/os/caselist/0823113/index.shtm

About the Author

Jerry Shenk currently serves as Senior Analyst for the SANS Institute and is the Senior Security Analyst for D&E Communications. Since 1984, he has consulted with companies and a variety of financial and educational institutions on issues of network design, security, forensic analysis, and penetration testing. His experience spans small home-office systems to global networks. Along with some vendor-specific certifications and a CISSP certification, Jerry holds five GIAC GOLD certifications: GCIA, GCIH, GCFW, GSNA and GCFA—all completed with honors.

SANS would like to thank this paper's sponsor: