PURESTORAGE®

WHITE PAPER

Data Storage Considerations for Security Analytics

An Excerpt From O'Reilly's "*Understanding Log Analytics*" Report

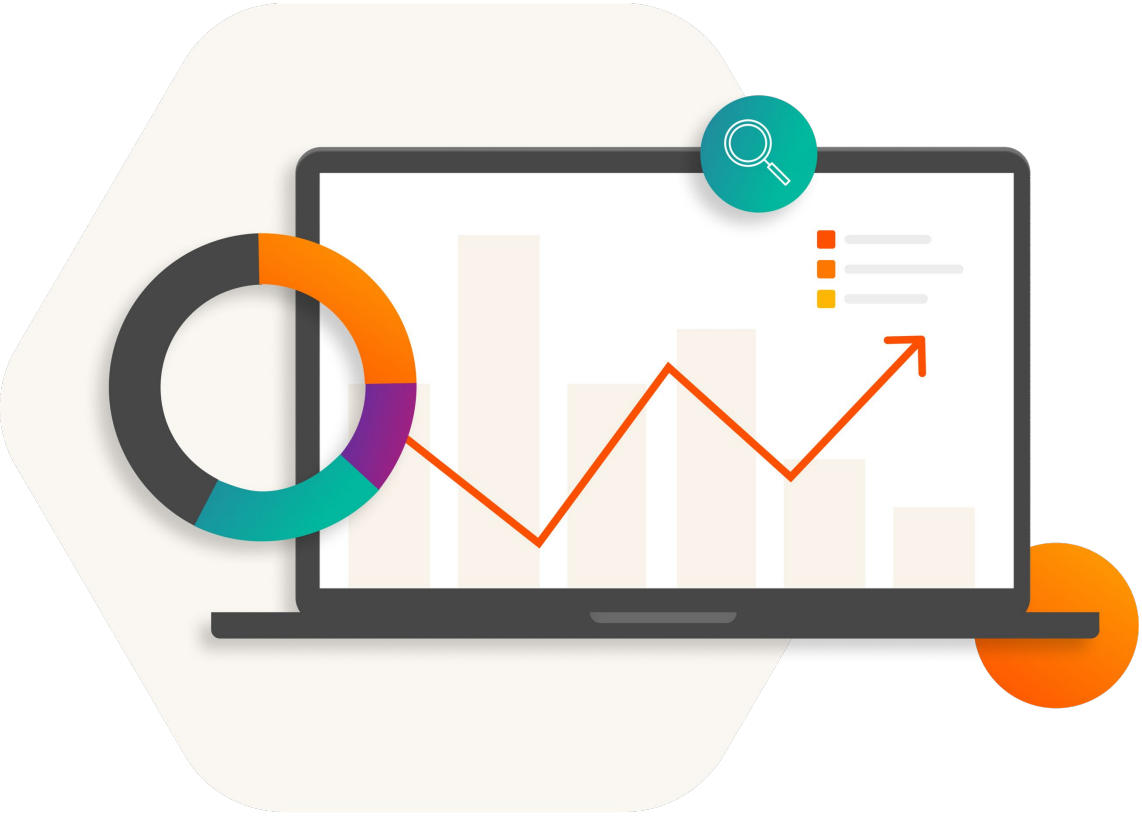
Contents

Note: Chapters are aligned to O'Reillys "*Understanding Log Analytics*" report for consistency.

Overview

Log analytics has taken increasing importance over the years. In particular, applications that use log analytics have provided key new capabilities for security operations teams within IT organizations. The many log analytics use cases can all benefit from scalable, performant data delivery through a modern storage platform. Security analytics in particular have special considerations that can be met with effective planning and deployment of data storage infrastructure.

This excerpt from O'Reillys "***Understanding Log Analytics***" a concise overview of the security analytics use case. Additionally, it provides a reference for performance and scalability considerations, and planning that aids the deployment and operation of data storage for security analytics.



Chapter 2

Log Analytics Use Cases

Technologists have been analyzing machine logs for decades, from the earliest days of tuning or troubleshooting their environments. Over time, the industry has found ways to increasingly automate that analysis, leading to the emergence of log analytics as we know it today. Now more than ever, log analytics can help businesses run more efficiently, reduce risk, and ensure continuity of operations.

The use case described in this section illustrate an example of how log analytics has taken on new importance in the past several years, demonstrating how it can deliver unprecedented value to organizations of all types and sizes. Factors that have contributed to that growing importance include the following:

- *Machine data* provides greater opportunities for log analytics as well as challenges. The scale of data analysis will grow further as we continue to drive intelligence into the world around us. A single self-driving car is estimated to generate multiple terabytes of data each day, while a smart factory might generate a petabyte per day.[1]
- *Greater variety* among types of endpoints has already reached unprecedented levels as the IT environment has become more complex. As the pace of change accelerates and the Internet of Things (IoT) adds billions of new devices online, the insights to be gained by bringing together multiple data sources will continue to increase.
- *Technology evolution*, making log analytics feasible at greater scale than before. In particular, the mainstream emergence of flash storage offers faster read/write speed than conventional spinning hard disk drives (HDDs), and low-cost compute capacity offers high performance with commodity servers.

With the increased scope and prevalence of log analytics as a whole, a growing set of common use cases have emerged, including cybersecurity.

Cybersecurity

Securing IT and other systems is a classic application of log analytics based on the massive numbers of events that are logged and transmitted throughout a large organization. Cyber protection in this area draws from log data and alerts from security components such as firewalls and intrusion detection systems, general elements of the environment such as servers and applications, and activities such as user login attempts and data movement. Log analytics can play a role in multiple stages of the security life cycle:

- *Proactively identifying and characterizing threats*
  - Log analytics can iteratively search through log data to detect unknown threats that conventional security systems are not designed to identify, creating testable hypotheses.
- *Detecting and responding to attacks and other security events*
  - When abnormal indicators arise, log analytics can help to identify the nature and scope of a potential breach, minimize exposure, and then neutralize and recover from the attack.

1 Richard Friedman. Inside HPC, May 31, 2019. "Converging Workflows Pushing Converged Software onto HPC Platforms." https://insidehpc.com/2019/05/workflows-converged-software-hpc/.

- *Performing forensic analysis after a breach has occurred*
  - A robust log analytics platform helps identify the point-in-time log information that should be brought into a post-mortem investigation as well as making that data available and acting on it.

Speed Matters

In recent years, there has been a growing recognition that effective cybersecurity measures are not just focused on preventing intrusions, but rather detecting intrusions as rapidly as possibly and ideally automatically acting to remediate the breach. The metrics mean time to detect (MTTD) and the mean time to remediate (MTTR) are now all standard KPIs for sophisticated cybersecurity organizations.

Reducing MTTD is vital because the less time a malicious actor has to operate in a corporate environment, the less damage they can do. In many cases, attackers lurk in the environment for days or months before picking a target. It is essential that security teams are able to retain searchable logs and correlate events quickly across several days or months to piece together an attack. Inability to look back quickly or slow historical queries can lead to blind spots.

Calculating this metric is actually quite complex in that an organization must record when a breach first occurred and then the time delta for when it was actually reported. Doing so usually will involve multiple logs as well as some records from the SIEM that actually records when the breach was detected.

MTTR is the obvious follow on to MTTD and for which there are many moving pieces. In addition to detecting the incident, it must be remediated. After detecting a potential incident, the best systems will automatically act to remediate the incident. Whereas a security Incident and event management (SIEM) platform can detect and monitor security incidents, a security orchestration, automation and response (SOAR) platform is necessary to remediate threats while minimizing human intervention. A SIEM will detect the threat and fire an alert, but the SOAR platform can take actions such as isolating an infected device, automatically without human intervention. SOAR platforms can significantly reduce the MTTR by removing or greatly reducing the human element from the equation of security incident remediation.

With all that said, speed is of the essence in both detection and remediation. Platforms which can rapidly access and analyze the data will ultimately be the most effective in reducing risk to organizations.

Detecting Anomalies

Cyber processes often use analytics to define a typical working state for an organization, expressed as ranges of values or other indicators in log data, and then monitor activity to detect anomalies. For example, an unusual series of unsuccessful authentication attempts might suggest attempted illicit access to resources. Unusual movement of data could indicate an exfiltration attempt.

The sheer volume of log data makes it untenable for humans to interpret it unaided.

With thousands of events per minute being documented by hardware and software systems all over the computing environment, it can be difficult or impossible to determine what is worthy of attention. Machine learning models can help analytics engines cull through these huge amounts of log data, detecting patterns that would not be apparent to human operators.

Those processes can occur automatically, or they can be initiated by ad hoc queries by analysts or others. Their outputs can be used to identify items of interest for human analysts to investigate further, allowing them to focus their attention where it is the most valuable. A common application is that threat hunters often use log analytics to help identify potential threats, look more deeply into them, and determine what response, if any, is required.

Machine Learning Is Invaluable to Anomaly Detection

The twin limiting factors in detecting anomalies in log data for security usages are massive data volumes and the necessity of looking for undefined patterns. The data itself is messy, consisting of many different formats and potentially containing misspellings, inconsistencies, and gaps. The anomalous patterns being looked for can be subtle and easy to overlook.

All of this makes humans poorly suited to anomaly detection at scale. Sustained massive levels of event volumes quickly become overwhelming, and a significant proportion of those events are irrelevant. At the same time, software tools might also not be successful, given that the effectiveness of its detection is limited by the accuracy of its assumptions, which are likely to be predetermined and static. Over the past five to 10 years, the industry has developed sophisticated dashboards to provide real-time views that help identify potential security incidents. These techniques can also be applied to insider threat detection and user behavior analytics (UBA). Vendors such as Exabeam, Securonix, and Splunk User Behavior Analytics (UBA) have become leaders in the space.

Machine learning (ML) and artificial intelligence (AI) are increasingly viable for improving those user monitoring approaches, overcoming some key limitations and turning massive data stores from a liability into an asset for helping to train ML models. Algorithms can use both historical and real-time data to continually update their vision of what “business as usual” looks like and use that moving baseline as the standard against which they interpret emerging log events.

In recent years, predictive analytics have become more common in a variety of usages. Based on all data received up to the current moment, a machine learning model can predict expected parameters of future events and then flag data that falls outside those ranges.

Working from that set of detected anomalies, the algorithm can correlate them with other incidents to limit the universe of events to be considered and to illuminate patterns in real time. By alerting security analysts to those anomalies and patterns, the system can pare the scope of alerts that must be investigated by human operators down to a manageable level. As a result, IT staff can focus on innovation and adding value to the organization rather than just maintaining the status quo.

In addition to alerting, ML models can be included as a part of an automated playbooks to not only identify malicious activity in a corporate environment, but remediate this. For example, a ML model can be built to detect anomalous behavior of a particular class of networking device. If it detects unusual behavior, the automation can then check the change management logs to see if that device has open change tickets. If it does not, the device can be isolated from the network, thereby preventing damage.

Identifying and Defeating Advanced Threats

One of the issues confronted by modern security teams is the subtlety and long time horizons associated with today's stealthy attacks. Advanced persistent threats operate by moving laterally as quietly as possible through an organization to gain access to additional resources and data, in a process that often elapses over a matter of months. The key to detection often lies less in identifying any specific event than in overall patterns.

In practice, a security analyst might begin with a suspicious activity, such as a service running from an unexpected file location and then use various log data to uncover additional information surrounding that event to help discover whether it is malicious or benign. For example, other activities during the same login session, connections from unexpected remote IP addresses, or unusual patterns of data movement captured in logs can be relevant.

Treating log data as a coherent whole rather than natively as a disparate collection of data points enables analysts to examine activities anywhere across the entire technology stack. This approach also enables analysts to traverse events backward and forward through time to retrace and analyze the behaviors of a given application, device, or user. This capability can be vital in cases such as understanding the behaviors of persistent cyber threats that operate over the course of weeks or months.

Data context consists of information about each data point's connections to others, which must be encoded along with the data itself, typically in the form of metadata created to describe the main data. This context enables analysis to identify the significance of a given data point in relation to the greater whole.

Statistical analysis against bodies of machine log data can reveal relationships that would otherwise remain hidden. Those insights help security teams more confidently categorize events in terms of the levels of threat they represent, enabling faster, more precise responses that help limit negative impacts on operations, assets, and reputations. In the context of a smart factory, for example, that analysis can help avoid unplanned outages that would otherwise lead to lost productivity and profitability.

Chapter 6

Performance Implications of Storage Architecture

Growing IT complexity is a fact of life as more business processes become digitized, higher levels of automation are enabled, and new technologies enter the datacenter. That growing complexity drives increasing volumes of log data that can potentially be used for log analytics; a company of a given size would generate far more logs today than a company of similar size a decade ago.

The availability of more log data generates the potential for more sophisticated log analytics, placing more extensive demands on the underlying systems. Specifically, even as the amounts of data that need to be ingested, handled, and stored rises exponentially, so do the numbers of queries being made against it, both by automated systems running reports and dashboards, as well as by human users placing ad hoc queries to generate business insights.

The emergence of flash storage for the enterprise in the past 10 years or so represents a sea change in storage architecture because of its dramatically higher speed and longevity. Notwithstanding those advantages, cost constraints mean that spinning disks still dominate in the datacenter. As the cost of flash storage decreases, spinning disk usage will decrease.

Conventional HDDs are appropriate for dumping large amounts of data that won't be searched against often, but the random reads and writes are much slower than flash. Searching a huge dataset to support real-time or near-real-time log analytics requirements can be prohibitively time consuming with spinning disks.

Log analytics operations depend on rapid, dependable access to stored data, which places growing performance and scalability demands on the storage hardware. Fast response rates are critical to business use cases, both to optimize efficiency and to provide a good user experience. These requirements are driving flash adoption in the enterprise; in fact, flash storage has become the standard implementation for many use cases.

Even as the volumes of log data being generated are growing rapidly, many companies are extending their standard retention periods, requiring longer-term preservation of data. A key driver behind this trend is the fact that analytics models can often be made more powerful by making larger sets of historical data available to them. Querying against several years of data allows tracking of long-term trends, and as AI models are increasingly adopted for analytics of all kinds, those large datasets can also be useful for training deep learning models.

Log analytics presents substantial performance challenges. Throughput must be optimized from the point of ingest, through all stages of processing, to outputs such as alerts, dashboard visualizations, or reports. Data pipelines and storage must be consolidated, eliminating data silos and the practice of storing multiple copies of data for different usages.

Tiered storage is a common approach to handling large amounts of historical data, including log data. By providing multiple areas of storage, each with a different balance between cost and performance, tiered storage allows the medium to be tailored to different needs, as illustrated in Figure 6-1. In this conception, the performance tier houses the data most likely to be queried frequently and needed for real-time analytics scenarios (i.e., the "hottest" data). The archive tier is for long-term storage of infrequently accessed "cold" data, and the capacity tier in between strikes a balance between the two for "warm" data.

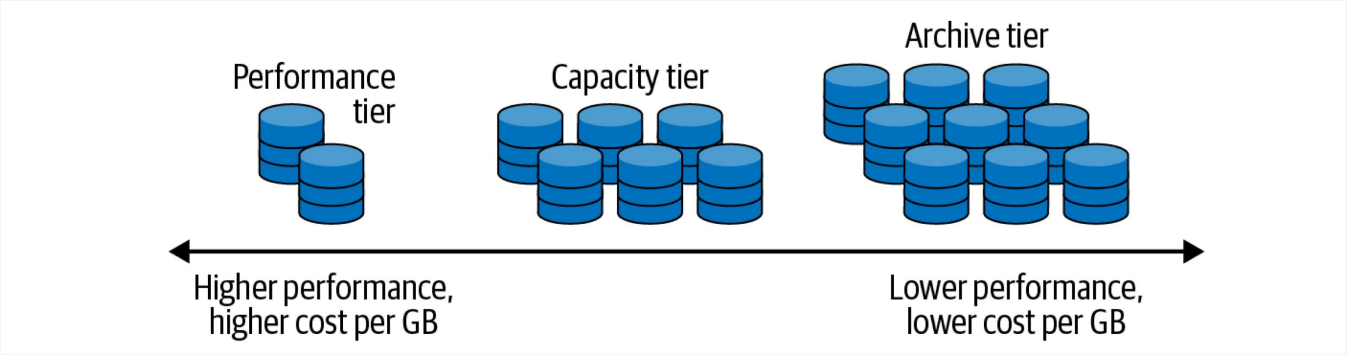


Figure 6: Tiered storage balances cost and performance for different data types.

In practical terms, this range might correspond to the age of log data. For example, the most recent 30 days of log data might be stored in the performance tier, data 31 days to a year old in the capacity tier, and data that is older still in the archive tier.

As organizations continue to stretch the limits of what they can monitor, accomplish, and predict with log analytics, query volumes will continue to increase, including searches against older data. Increased requirements to perform analysis and extract insights from data stored in the capacity and even archive tiers compels architects to increase the performance of those tiers, which is largely accomplished by the addition of flash storage.

Drive More Value from Storage with Data-Reduction Technologies

Because flash is an order of magnitude faster than spinning disks, it has become all but essential to data-intensive workloads such as log analytics. Particularly for real-time results that draw on historical data, the storage tier requires the speed that only flash can offer.

Storage vendors have devised a range of data reduction features that reduce the amount of capacity needed to store a given body of data. While not all are available from all providers, the following capabilities have been developed by the industry:

- *Pattern removal* detects and consolidates simple binary patterns within datasets (e.g., summarizing a string as "1,000 zeroes" instead of actually encoding the 1,000 identical values). These measures can reduce storage requirements, as well as processing for other data-reduction measures such as deduplication and compression.
- *Deduplication* ensures that only unique blocks of data are committed to flash storage. Ideally, deduplication works globally (rather than within a volume or a pool) and on variable block sizes for maximum efficiency.
- *Inline compression* reduces the number of bits needed to represent a given piece of data. The use of multiple algorithms is desirable, to optimize compression ratios among different types of data that have different requirements.
- *Post-process* compression applies additional compression algorithms to data after the process is complete, increasing the data reduction result achieved with inline compression.
- *Copy reduction* handles data-copy processes within the flash storage medium using only metadata, producing snapshots and clones that offer greater efficiency than actual copies of the data.

To get the full value of these capabilities in the datacenter, they should be available right out of the box, without requiring extensive configuration or tuning. High-performance data-reduction measures should be always-on and suitable for use across workloads, regardless of size, type, or criticality.

Additional Considerations for Security Analytics

Security analytics can place additional, unique requirements on IT organizations. A review of special considerations offers perspective on common security analytics challenges and the role of scalable data storage that can adapt to the need for real-time and historical data analysis. Effective real-time data processing is key to achieving organizational MTTD and MTTR goals. Easy access to historical data enables needed insights into threat landscapes for root cause analysis and persistent threat detection. Scalable, resilient architectures enable the agility to keep pace with not only the growth of data, but to new and changing sources of log and event data.

Responding to and Detecting Threats in Real Time

In order to minimize the MTTD and MTTR, an organization must have the ability to query and perform statistical analysis on log data in real time or near real time. Simply detecting and blocking malicious activity isn't enough to adequately protect an organization. In order to respond to more sophisticated attacks, it is often necessary to conduct historical data analysis, as well as real-time detection.

What is clear is that security use cases require a considerable amount of data, regardless of whether the system is real time or not. For a real-time system to be effective, it must have rules which the security team develops by analyzing a large corpus of data. Additionally, real time systems frequently employ machine learning models which must be trained (and retrained) on large collections of data. In any case, an organization will need an efficient and effective file system to enable their security professionals to access and analyze the various forms of security data they collect.

Let us examine several use cases for real-time threat detection and remediation.

Ransomware

Ransomware is a type of malware which blocks a user or organization from accessing their data, usually by encrypting it and simultaneously demanding payment for releasing the data. In the last year, ransomware has crippled several large organizations including hospitals, school systems, and government organizations.

If an organization has effective backups, it is possible to reverse the damaging effects of ransomware. However, some ransomware also encrypts backups making restoring impossible. The best approach of course is to detect and defeat ransomware as quickly as possible. Unfortunately, it is very difficult to detect ransomware fast enough to prevent damage. Ransomware can be detected in log files by looking for anomalous file system activity, abnormally high CPU usage, or disk activity, and suspicious network communication. If these patterns are identified, it is vital to isolate the infected systems from the rest of the corporation to prevent the further loss of data.

The most common ways organizations get infected by ransomware are from spam, malicious websites, infected removable drives and malicious applications. Proactively monitoring logs for these attack vectors can reduce the risk of a ransomware infection.

Anomaly Detection

Anomaly detection is a broad grouping of machine learning techniques used to detect things that are outside the ordinary. In a security context, usually this means something bad happened, but not always. For example, a router may be exhibiting abnormal behavior. The cause could be that the router has been infected with malware, or it could be that a technician is reconfiguring the device. Treating anomalies as automatic alerts can result in flooding operators with useless alerts, which will annoy your SOC and possibly worse, cause them to miss a real alert.

Anomaly detection can be broken down into novelty and outlier detection. There is a subtle distinction between the two, which is that novelty detection is when you start with a data set that contains only good observations. This data is then used to train a machine learning model which can then identify inliers and outliers. Outlier detection starts with a dataset with both inliers and outliers. Those familiar with machine learning will likely map these to supervised and unsupervised learning. Anomaly detection is not a single technique but rather a collection of techniques which could include simple statistical analysis, all the way to sophisticated machine learning techniques.

At the high level, most anomaly detectors work in a similar manner where they calculate a score which is used to represent "normal". This score is then calculated for an unknown observation and if the difference exceeds a threshold, it is treated as an anomaly.[2]

In order to perform anomaly detection, it is necessary to have a logging infrastructure that supports historical querying. While it may be possible to use summary data for this purpose, this will only work if the summaries contain the correct features for the model. Therefore it will be necessary to have a decent volume of historical data from which an analyst can build their model. A data scientist working on such a problem will need to perform aggregate queries over long time periods and thus the logging infrastructure should have enough data to support this as well as a reasonable latency for queries which span long time periods.

IDS/IPS Considerations

Intrusion detection systems (IDS) and intrusion prevention systems (IPS) are related systems which detect and prevent intrusions into networks. An IDS sits behind a firewall and scans packets, sending alerts in the event of a suspected breach. An IPS is similar to an IDS in that it also sits behind a firewall and scans packets, but an IPS can take action to prevent damage to corporate networks. In addition to simply sending alerts, an IPS can drop potentially malicious packets, block malicious traffic or reset network connections.

While IDS/IPS can reduce the risk to organizations by reducing the MTTD/MTTR, an organization still must analyze the data contained in their logs to understand the root causes for the entries. For example, if a user has consistent failed login attempts, or if there are unknown devices on the network in particular areas or times, these could all be indications of an orchestrated attack on the organization.

2 A simple example might be to take the last 100 years of temperature data for a given location on a given day. Then if that temperature varied by more than a number of degrees, there is an anomaly.

Analyzing the Threat Landscape

Another facet of security data analysis is longer term analysis. Sophisticated actors will gain unauthorized access to an organization with the goal of maintaining undetected but unauthorized access for a lengthy period of time. These actors' goals are usually not immediate profit but rather industrial or political espionage, financial crimes or serious disruption of their target. These kinds of attacks are labeled as advanced persistent threats (APT).

Detecting Advanced Persistent Threats

In order to understand how to detect APTs and the role log analytics plays in this, you must understand the high level pattern which APTs will follow.

Most APTs follow a pattern of: initial compromise whereby they gain access to some portion of a network. This first step can be accomplished by gaining access to an individual's user account via phishing email, web hijacking or some other technique. Once that has been accomplished, the goal is to make this access persistent, which is known as establishing a foothold. After the foothold has been established, the next goal is to escalate privileges. Usually the attackers will not gain access to root level or other administrative accounts. They commonly will break in via some weakly secured system and gain user access. However, since their goal is to gain persistent access to valuable information, the actor will attempt to gain access to an administrator or root level account in the network. This will allow the malicious actor to perform activities such as reconfigure the network, disable security provisions, and create new user accounts. The final step in the process is lateral motion which is moving from one internal device or network segment to another and repeating this process.

Most corporate networks are well protected from unauthorized activity from outside the firewall. However APTs are particularly dangerous because the attackers will use an organization's internal network to attack other internal systems. APT actors will use extremely sophisticated software for command and control which carefully masks its external communications to blend in with the normal activity of an enterprise network.

The challenge with detecting APTs is largely separating the noise from the actual attack traffic. Automated tools are usually ineffective in detecting APTs and as a result, organizations use cyber threat hunting teams to detect and defeat APTs. In order to do their work, hunting teams must have access to large amounts of log data from a variety of sources. Additionally many cyber threat hunters will build custom machine learning models to detect threats. All of this depends on having meaningful access to large amounts of diverse security data, often spanning long time periods.

FireEye reports that the median time an APT attack goes undetected in the United States is 71 days, in EMEA 177 days and APAC, 204 days. [3] With these numbers in mind, it is clear that an organization should retain enough historical data to detect these breaches and also to be able to analyze the actor's activity to remediate the damage.

Legal Discovery (e-Discovery)

In the event of a legal proceeding or dispute, both parties to the dispute are subject to discovery—a process in which information is gathered with the intent of using it as evidence. As an ever-increasing amount of human activity and interaction takes place on networked computers, when legal or criminal issues arise, it is increasingly common to gather evidence from computer systems. And that evidence can include emails, computer activity from logs, malware, documents, chat histories and more.

3 https://www.fireeye.com/current-threats/annual-threat-report/mtrends.html

In an enterprise, most data is stored on central servers rather than on an individual's machine. Therefore, the process and requirements for e-discovery are similar to threat hunting, in that the investigator must have access to a large corpus of log (and other) datasets and must be able to effectively search that data for items which are relevant to the investigation. As disputes can cover years worth of activity, effective legal discovery depends on organizations retaining considerable amounts of data and storing this data in a way which enables effective analysis.

Root Cause Analysis

When a cyber or IT operations incident occurs, logs can provide the first line of defense in identifying the problem and possibly, via a SOAR platform, remediating it. However, understanding the root cause of the issue is helpful in order to prevent the issue from recurring in the future.

Root cause analysis (RCA) is a process through which an analyst will identify the problem, establish a timeline from the normal activity which preceded the incident and the problem, and finally identify the root cause, as well as any other causal factors. The usual result of RCA is identifying actions to prevent the problem from recurring.

Similar to threat hunting, RCA depends on having a large corpus of historical data, stored in a platform which will allow the analyst to reconstruct the timeline of events. This will necessarily draw from multiple log sources and formats. An analyst must be able to execute queries over long time spans, in order to reconstruct the timelines and sequence of events.

Seamless Scalability

One of the main challenges in log analytics is determining how much data to retain. The drivers for this can be cost, compliance, risk appetite, use cases, and other factors. Inevitably, a log analytic platform will need to grow as new systems are brought online, or new use cases are developed which require additional data.

Log analytics are frequently mission critical, in that they can identify (and prevent) breaches, incidents, and other issues, so a system must be able to scale without issues. Some things to consider are dynamic data, system disruption, and the process for adding data sources.

Adding Data Sources

As log infrastructure scales, it will be necessary to add new data sources. For analytic systems such as Splunk, it is necessary to create new data flows from the source system into the log analytic system. Once that flow is established, it is necessary to configure that system to parse the data into indexes. Frequently however, there are many organizational hurdles to make this happen. For instance, there may be legal, privacy, policy, or compliance approvals which need to be obtained prior to starting data flows.

Having a standardized infrastructure, such as a Kafka messaging bus, can greatly reduce the technical complications of sending data into the log indexer. Kafka also allows an organization to split feeds into multiple destinations. Lastly, a messaging bus, such as Kafka, allows an organization to track data sources and how they are being used to some degree.

Dynamic Data Formats

A very common issue as log analytics platforms scale is dynamic data, or data in which the schema changes. Often schema changes will cause corrupt records in SIEMs which can lead to incidents being missed entirely. As many log analytic platforms use regular expressions to parse logs into fields, if the log format changes, it will no longer match the regex.

To some degree, unexpected changes can be mitigated with effective change management procedures. Ensuring that the SIEM engineers are aware of changes made to log generating systems will allow them the time to prepare the SIEM to accept and process new log formats. However, a better approach is to log data in a format which contains the schema itself, such as JSON or Parquet, and that way the query engine will automatically detect and parse any new fields. This approach will prevent corrupt data from entering the SIEM, but it is still only a partial solution. If downstream visualizations or other analytics depend on particular fields, and if those fields are renamed, those analytics will break.

In recent years, an alternative approach of using machine learning to parse logs and there have been some models which can parse logs into fields with greater accuracy than regex. The machine learning approach is likely the most effective for organizations that have highly dynamic data.

Chapter 8

Nine Guideposts for Log Analytics Planning

The benefits that log analytics can provide vary dramatically among different organizations, as do the infrastructure and techniques best suited to enabling those benefits. Nevertheless, the common set of best practices and considerations described here can help guide architects during the planning process.

- **Guidepost 1:** *What are the trends for ingest rates?*
  - Accommodate future needs for performance and capacity.
- **Guidepost 2:** *How long does log data need to be retained?*
  - Fine-tune data retention policies to optimize costs and minimize liability.
- **Guidepost 3:** *How will regulatory issues affect log analytics?*
  - Provide verifiable measures to govern data usage, transport, and storage.
- **Guidepost 4:** *What data sources and formats are involved?*
  - Forecast and prepare for upcoming requirements for changes in data-transformation pipelines.
- **Guidepost 5:** *What role will changing business realities have?*
  - Align infrastructure planning for log analytics with broader corporate strategy.
- **Guidepost 6:** *What are the ongoing query requirements?*
  - Identify future query volumes among different types such as ad hoc versus point queries.
- **Guidepost 7:** *How are data-management challenges addressed?*
  - Plan for impacts on log data formatting and delivery from changes in the environment.
- **Guidepost 8:** *How are data transformations handled?*
  - Ensure that tools and applications in place to transform data are sufficient for the future.
- **Guidepost 9:** *What about data protection and high availability?*
  - Designate log data's criticality and sensitivity, reflected in security and backup/restore policies.

In particular, it is important to keep in mind that key considerations and concerns that bear on planning infrastructure for log analytics will intensify as log data continues to grow in volume, velocity, and variety. Architects must therefore plan for flexible scalability of capacity and performance in their storage systems to support the log analytics function as it continues to become more demanding, as well as more valuable to the enterprise as a whole.

Guidepost 1: What Are the Trends for Ingest Rates?

Log data originates all over the environment, and its volume grows continually as the environment becomes more complex over time. Even as multiple terabytes per day inundate the log analytics platform initially, the sheer scale of the data is unbound in the future. Determining how those data volumes are likely to grow over time is critical to understanding the future state of the environment.

In particular, the storage infrastructure must be designed to accommodate future needs from both the capacity and performance perspectives. The capacity aspects of this requirement speak to the value of decoupling compute and storage so

that the latter can scale independently of the former. The storage infrastructure must be able to scale in terms of performance to support the larger numbers of more complex queries against ever-growing log data volumes.

Guidepost 2: How Long Does Log Data Need to be Retained?

Corporate standards, audit provisions, and regulatory requirements can all affect the required retention period for log data. As volumes continue to grow exponentially, the cost and complexity associated with storing it can become burdensome or even untenable. At the same time, growing data stores make a structured approach to tracking the data life cycle more vital so that it is not retained longer than necessary.

Best practices in this area include assessment and tuning of data-retention policies to ensure that they are appropriate both to meet requirements and to ensure that retention periods are maintained at the shortest appropriate level. If possible, retention requirements should be projected forward to discern whether they are likely to increase or decrease in the future. Nearer-term requirements include data reduction techniques such as deduplication and compression to reduce the burden on storage system capacities as much as possible.

Architects should also consider options for cost-effectively storing archival data. For example, if performance requirements associated with older data stored as Amazon S3 objects are lower than for current operating data, it might be desirable to push them out to Amazon Glacier or a similar cost-optimized service.

Guidepost 3: How Will Regulatory Issues Affect Log Analytics?

Setting data-retention standards is a clear issue associated with meeting regulatory requirements, but the full scope of considerations in this area is far broader. Personally Identifiable Information (PII) and other sensitive data must be controlled and protected with verifiable measures that govern how it is used, transported, and stored. This set of concerns can affect issues, such as how specific data can be used in public cloud infrastructures or shared with partners, for example.

Particularly for organizations that operate in multiple geographic areas, data sovereignty can be a complex issue. Because data is governed by the laws of the jurisdiction where it is located, organizations must be concerned with the physical locations of their data, particularly in cases for which public cloud resources are used. For example, data that was collected through perfectly legitimate means in one country can be in violation of the privacy laws in another.

As a related matter, confidential data might be subject to subpoena or other unwanted inspection by government or legal entities in the jurisdiction where it is stored. Response times potentially required by subpoena actions can be a challenge in the common case where it requires weeks to restore older data from backup and then days to query against the associated large volumes of data. For organizations that must regularly respond to law-enforcement requests for archival data, architects might need to accommodate streamlined access.

In a world in which regulations over data privacy and usage are evolving, forecasting legal requirements can be problematic. One approach is to consider the measures taken to date by government entities that have provided early leadership. Frameworks such as the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) suggest the types of regulatory patterns that might later be adopted elsewhere.

Guidepost 4: What Data Sources and Formats Are Involved?

The semi-structured nature of log data means that the sources of those logs play an important role in determining how to handle the data. As IT complexity and the scope of sources grow and change over time, the associated challenges can become more complex. Particularly as IoT topologies are built out over the next several years, many organizations will find themselves needing to accommodate a vast assortment of new sensors and other endpoints, and that variety will also usher in an expanded diversity of log types and formats.

As the universe of data sources and log types becomes broader and more complex, novel transformation pipelines will be needed to create a coherent whole from that data. Forecasting those changes in advance and accommodating them in how the infrastructure scales is an important set of requirements for architects. The ability to cost-effectively scale out compute and storage, independently of each other, is central to successfully meeting this set of evolving requirements.

Guidepost 5: What Role Will Changing Business Realities Have?

The changing needs of a business and its use of log analytics to guide business change are deeply intertwined. Planning should include considering what business questions log analytics can answer, such as how to allocate resources for maximum efficiency or whether to launch a marketing initiative geared toward increasing revenue. Architects must consider how intelligence generated through log analytics can inform business strategy.

Just about any major business event can have an impact on infrastructure planning for log analytics. That reality makes it valuable for infrastructure planning to draw on the business's larger corporate strategy. For example, architects should consider the potential impacts if the company were to enter into new market segments or geographies. Either type of change would be likely to increase the volume and variety of log data. Likewise, mergers and acquisitions could introduce new and unpredictable infrastructures alongside what the company already operates. Because every organization is subject to unforeseen circumstances, architects must design flexible, scalable frameworks that can accommodate uncertain future needs.

Guidepost 6: What Are the Ongoing Query Requirements?

The central issues around query requirements are what types of searches are being done against the data and how many. In addition to searches by human users and machine-to-machine systems, planning must take into consideration factors such as executive dashboards, data visualization systems, reporting, and real-time alerting. Architects must account for potential addition of such new loads on the log analytics infrastructure as well as the expected growth in workloads generated by existing systems.

The nature of the queries being made against the log data store is a key design factor for log analytics. In practical terms, for example, architects should identify what data is most likely to have ad hoc queries made against it, as opposed to point queries. They also need to consider how often each type is likely to occur as well as how much historical data is likely to be searched and how frequently. These capabilities should be tuned to avoid slow or cumbersome search, avoiding lost opportunities or missed deadlines.

Likely usages such as ad hoc queries being done in conjunction with threat hunting activity should be considered to help guide this process. The answers to those questions can help guide design decisions about factors such as how to implement tiered storage or which data to index at the point of ingress.

Guidepost 7: How Are Data-Management Challenges Addressed?

Effectively managing data is a cornerstone of driving value from the massive volumes of log data constantly being generated by hardware, software, and processes of every description. In particular, the inherent variety within log data adds complexity to matters of schema evolution and back-compatibility as log data sources change and multiply. Solution architects should identify strategy for maintaining coherence and functionality in the face of such changes.

Data-management concerns associated with log analytics extend to changes in the operating environments in which log data is generated. For example, the firmware or software running on sensors, systems, and other entities that are transmitting logs might be upgraded or reconfigured. Those changes may alter the formatting of log files, requiring adaptation by the log analytics platform. Proper planning requires establishing processes to anticipate such changes in advance and setting standards for addressing them when they arise.

Guidepost 8: How Are Data Transformations Handled?

Transforming log data as it is collected from a wide variety of sources places significant demands on the underlying systems that must typically be satisfied in real time on steadily growing data streams. The alternative of simply transmitting and storing logs as flat files puts untenable processing burdens on analytics processes. The transformation workload itself can be highly compute intensive, and it takes place over a complex and varied compute layer. It also places significant demands on the storage layer.

The tools and applications in place to perform those transformations must be capable of handling any foreseeable data requirements as well as integrating and interoperating effectively with the other components of the log analytics pipeline. Likewise, the underlying infrastructure must be designed to provide storage and compute architectures that can accommodate the associated performance and scalability requirements.

Guidepost 9: What About Data Protection and High Availability?

As log analytics evolves within an organization, it enables increasingly sophisticated usages that deliver increasingly significant value. Over time, those usages can become business critical or even mission critical, elevating the value of the underlying log data as well as the requirements for assuring its accessibility. Foreseeing and planning for that transition involves providing high availability for a subset of log data, without interfering with the smooth operation of analytics based on mixed data streams.

Designating specific bodies of log data as critical is also tied into other aspects of IT planning. From a security perspective, this status must be considered when identifying requirements for how it should be protected and how sensitive information in the log data should be masked as well as its recoverability after tampering or other interference as the result of a breach. Backup and restore processes for protecting that data should also reflect its potentially changing value to the organization.

Conclusion

For a comprehensive look at additional log analytics use cases and considerations, be sure to read O'Reilly's "*Understanding Log Analytics*" report be. Pure Storage offers proven solutions for log analytics with leading vendors including Splunk and Elastic. You can learn more about our solutions for IT Operations and Security Analytics below. Our FlashBlade storage platform is the industry's most advanced all-flash storage solution for consolidating unified fast file and object data.

More Information:

- Read the full O'Reilly "*Understanding Log Analytics*" report
- Discover how to modernize your storage with a unified fast file and object storage platform
- Explore our solutions for IT Operations and Security Analytics

Pure Storage, Inc.
650 Castro Street, #400
Mountain View, CA 94041