Scalability in Log Management

Research 010-021609-02
**Introduction**

In the last few years, log management has become increasingly relevant to multiple groups within any organization. Audit teams leverage logs to automate compliance reporting and detect policy breaches. Security teams monitor log data to detect internal and external threats as well as for forensic investigations. Logs are also widely used by IT operations or helpdesk teams for faster troubleshooting and better adherence to service level agreements.

Companies typically begin their search for a log management solution with a given driver in mind, such as security threat monitoring. Yet, over time most companies will expand the scope of the original use case across more locations and devices and also grow into new use cases such as regulatory compliance or IT operations.

This trend is validated by the SANS Institute’s 2008 Log Management survey, which shows that most companies are leveraging logs for multiple use cases and it highlights the importance of evaluating the scalability of log management solutions up front. To that end, this whitepaper provides an overview of drivers for log management scalability and the resulting requirements that should be considered in any log management evaluation.
Dimensions of Log Management Scalability

Scalability in Log Management can be evaluated along five primary dimensions, which in turn point to specific requirements. Factors like administrative and architectural scalability are addressed within these dimensions, as appropriate.

Event Volume

As organizations broaden the scope of their log management initiative, event volumes can increase significantly. Any large organization is likely to collect thousands of events every second (equivalent to millions of logs each day) just to support a single use case around security and network device monitoring. But consider what happens when regulatory drivers (PCI, SOX, HIPAA, GLBA, etc.) kick in? The organization must now collect more logs from existing devices (especially logs capturing user activity) and will also have to expand collection to operating system, database, and application logs, which actually store the regulated data. Effectively, across use cases a large organization may easily require its log management investment to scale upwards of tens of thousands of incoming logs each second.

Expanding use cases can also drive the need for scale in outbound log event volumes. For example, many organizations invest in log management as a first step towards real-time use cases, such as fraud detection acquired through a SIEM solution. Alternatively, large distributed organizations may need to deploy log management in a tiered architecture depending on their organizational structure. In both cases the first tier of the log management architecture will need to support very high collection rates while simultaneously filtering and forwarding logs to the next tier or to a SIEM destination.

Scalability considerations:

- Can the solution easily expand to handle the total inbound and outbound event rates across devices and use cases (not just the immediate scope of deployment)?
- Are there sufficient buffers to accommodate peak collection rates over the expected sustained event rates?
- Is the solution able to support long-term event volumes (inbound and outbound) while leaving sufficient processing room for all the log analysis activity? Or conversely, what tradeoffs on analysis performance and capacity does scalability in collection impose?

Assets

In the context of log management, asset scalability is often overlooked because organizations tend to only count assets from which logs are actually being collected. Most use cases however require much greater scale when it comes to assets.

Consider a large financial institution that is currently only monitoring logs from a few hundred security and network devices for perimeter threat detection. Even for this scenario you would need an asset model that can scale to tens or even hundreds.
of thousands of assets (desktops, printers, servers, network devices, security devices, etc.), which are generating the traffic across the network and security devices being monitored. This requires more than just a static object list. Each asset has to be modeled along numerous attributes (known vulnerabilities, criticality, asset role, related regulations, etc.) that must be referenced by real-time correlation rules and by reports. For example, if an IDS source suggests that a specific IP address is being attacked, the criticality and susceptibility (received through a vulnerability scanner) of the host provide important context needed for prioritization of the threat and the correlated alert in real-time.

Use cases such as regulatory compliance, sensitive data protection, and fraud detection increase the need for asset scalability by another order of magnitude. To support these use cases the solution must support modeling of more objects including users, applications, and data - along with all their attributes. As a result the asset model may well need to scale to over a million entities.

Scalability considerations:

- Can the solution scale adequately to model all assets (hosts, users, applications, data, etc.) in the organization while accounting for growth?
- Can the solution scale to support a large number of attributes per asset and can those attributes be referenced dynamically in real-time correlation rules and in reports?
- Does the solution provide a way to import common sources of asset context out-of-the-box (vulnerability scanner test results, user identity and role based sources, etc.)?

Locations

Although any large organization is likely to have multiple locations and data centers, highly distributed environments are the norm in verticals like retail, entertainment, hospitality, transportation, and banking. Research suggesting that most breaches occur at the point of sale or customer interaction is driving the need to monitor assets at remote locations. Expanding the footprint of log management across a large number of locations introduces more scalability challenges.

The basic problem is that simply streaming logs back to the data center is often not an option. First, bandwidth is usually limited and must be prioritized for actual business traffic, so log traffic needs to be cached and batched during off hours or must use a specified fraction of available bandwidth. Second, regulatory compliance requires logs to meet audit quality, so streaming logs over unreliable and unsecured connections is usually unacceptable. These constraints necessitate a local collection option at each individual store or branch location, which in turn creates deployment- and administration-related scalability challenges.
Deployment scalability is a challenge because there is typically no IT staff to rollout localized log collection at each remote location. Additionally, at branch offices rack space is often limited and existing critical servers can not be overloaded with additional log collection software. So any store / branch level collection component of a log management solution must either be available as a plug and play appliance or should be remotely deployable as software.

Even past the deployment stage, there are very real challenges around administrative scalability associated with managing updates, upgrades, configuration changes and general maintenance of distributed log collection infrastructure. To address this, log management solutions must offer centralized management capabilities. For example, if 1500 stores share a common application that processes credit card data, there must be a simple way to rollout the collection parameters to all the locations simultaneously.

Scalability considerations:

- Can agent-less collection capabilities be deployed easily to a large number of branch offices as a remotely configurable appliance or as remotely deployable software?
- How easy is it to expand collection to new devices or device types at remote locations?
- Can log collection be administered across hundreds or even thousands of remote locations simultaneously (this includes updates, upgrades, and collection parameter updates)?

**Capacity**

While capacity can have multiple connotiations in Log Management, the reference here is to scalability around “storage” capacity. Given regulatory requirements for data retention, any large organization can end up with hundreds of terabytes of logs over time. For example PCI requires log data to be available for 90 days online and a year offline. Other regulations like SOX, GLBA, and HIPAA are less prescriptive but common interpretation suggests multiple years of log data need to be retained.

For scalability in storage capacity, log management solutions must efficiently compress logs without a major compromise in collection or analysis performance. Yet, many commercial log management solutions can only optimize storage capacity and compression levels by compromising log analysis and collection rates.

Log storage capacity also needs to scale architecturally and support a tiered deployment. For example each subsidiary or geographic division of a large organization may have separate IT teams that need to manage logs locally but the corporation as whole will still need universal visibility across locations for regulatory compliance. It is common for such organizations to have a tier of log storage at the division level for local analysis and another tier of aggregated storage at the corporate level.

Log capacity should scale without imposing significant administrative overhead and this is where storage formats become important. Some solutions rely on relational databases even for long-term storage, which will require DBA management expertise to manage large log volumes.

Scalability considerations:

- What is the compression ratio of logs in long-term storage and how is it reduced when optimizing for high-performance analysis (typically through indexing)?
- As additional capacity is added to a given location or at multiple data centers, can analysis (forensic search for example) still be conducted across all logs or is it limited to each data store?
- How much administrative effort is required to manage the data (archival, retention policies, etc.)? Does old log data first have to be restored for analysis?
Analysis
With organizations leveraging logs for more use cases, log management systems need to scale in terms of the analysis load that can be supported. While regulatory compliance introduces significant reporting requirements, the forensics use case requires quick search capabilities. On the other hand security-related monitoring requires scalability in real-time processing or correlation of logs. Yet most log management solutions can only scale historical or real-time analysis performance by significantly compromising collection performance and storage efficiency. This forces companies to bear the cost of additional hardware (or appliances) for greater computing power and faster analysis.

The scan rate per instance of a log management platform is one metric for assessing scalability of log analysis. Organizations should assess whether their reporting requirements can be satisfied in a timely manner. Some regulations such as the EU Data Retention Directive also come with requirements to respond to a forensic search request within a certain time frame. Generally large organizations will need a solution that can search hundreds or even millions of logs each second.

Concurrence in log analysis represents a second key metric for log analysis scalability. Across use cases a log management solution must concurrently scale the number of users, real-time correlation rules, reports, or searches. For example a solution that can only process a handful of correlation rules will limit the number of threat vectors that can be detected. Ideally the real-time correlation engine alone must be able to simultaneously evaluate hundreds of correlation rules against the incoming log data stream.

Scalability considerations:

- What scan rate does one instance of the log management solution support for searches and reports?
- How many real-time correlation rules can be processed concurrently by one instance of the real-time correlation layer?
- What compromise in capacity and collection rates are imposed when analysis performance is optimized? How does this impact the investment in terms of total hardware cost (including power, management, etc.)?

Summary
Log Management evaluation criteria have largely been focused on functional aspects of log collection, storage, and analysis. Functional criteria are certainly important but as more and more organizations leverage logs across broader use cases, scalability becomes more important to the initial evaluation process. As reviewed in this whitepaper, at a minimum log management scalability should be assessed along the lines of event volume, assets, locations, capacity, and analysis. Organizations should evaluate candidate solutions across all of these scalability dimensions and most importantly should base requirements on the long-term scope of their log management initiative and not just the immediate drivers. A summary of the key scalability questions to consider as part of any log management evaluation follows.
### Scalability Dimension

<table>
<thead>
<tr>
<th>Scalability Dimension</th>
<th>Considerations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event volume scalability</td>
<td>Can the solution handle the total event rates across use cases?  Are there sufficient buffers to accommodate peak collection rates?  What tradeoffs on analysis and capacity does collection scalability impose?</td>
</tr>
<tr>
<td>Asset model scalability</td>
<td>Can the solution model all assets (hosts, users, applications, data, etc.)?  Can the asset model store and reference attributes of all assets?  Is there an easy way to import common sources of asset context?</td>
</tr>
<tr>
<td>Location based or geographic scalability</td>
<td>Can agent-less collection be deployed easily to a large number of locations?  Is a tiered architecture for major divisions or data centers available?  How will de-centralized log management infrastructure be administered?</td>
</tr>
<tr>
<td>Capacity or storage scalability</td>
<td>How does changing the compression impact analysis performance?  Can analysis be conducted across all locations and data stores?  To what extent are retention policies automatically enforced?</td>
</tr>
<tr>
<td>Analysis scalability</td>
<td>Do the scan rates satisfy internal or regulatory forensic requirements?  How many real-time correlation rules can be processed concurrently?  What impact does optimizing analysis performance have on storage efficiency or collection rates?</td>
</tr>
</tbody>
</table>