SOLARWINDS WHITEPAPER

Estimating Log Generation for Security Information Event and Log Management

Brad Hale

SOLARWINDS WHITEPAPER

Estimating Log Generation for Security Information Event Management

As more solutions enter the marketplace claiming to collect, analyze and correlate log data, it is becoming increasingly necessary to have the ability to estimate log generation for one's environment. This is required for two primary reasons: to estimate the amount of storage required for log data; and to estimate the cost of various solutions given their licensing model. This paper will discuss an approach to estimating the amount of log data generated in a hypothetical network environment.

Disclaimer

There is no one size fits all for estimating log generation. Factors that impact the amount of data generated include, but are not limited to: network complexity and design including number and type of devices on your network (switches, routers, firewalls, servers, etc…) and load on each device; device logging policies, especially the severity level for which logs are generated and which logs you actually want to collect and monitor; and size in bytes of the log generated.

First, The Basics

Every device in your IT infrastructure generates log data that can be used to analyze and troubleshoot performance or security related issues. What one does with the data depends on what one is trying to accomplish with the data and is usually categorized as either Log Management or Security Information Event Management.

According to Wikipedia (therefore, we know it must be accurate – queue the laugh track), **Log Management** (LM) comprises the approach to dealing with large volumes of computer-generated log messages. LM covers log collection, centralized aggregation, long-term retention, log analysis, log search, and reporting. LM is primarily driven by reasons of security, system and network operations (such as system or network administration) and regulatory compliance.

Security Information Event Management

SIEM, also known as security information management (SIM) or security event management (SEM), goes beyond LM by not only performing the data aggregation, but also including correlation, alerting and presentation in a graphical dashboard for the purpose of compliance and retention. Essentially, SIEM adds the intelligence to LM so that IT professionals can more pro-actively monitor and manage the security and operations of their IT infrastructure.

Under either approach, one needs the ability to collect the data from the various sources and that data will vary greatly in the amount and frequency of the data generated.

Events per Second

The most common approach to determining how much log data will be generated is to use **Events per Second** (EPS). EPS is exactly what it is called, the number of log or system events that are generated by a device every second.

$$EPS = \frac{\#\ of\ System\ Events}{Time\ Period\ in\ Seconds}$$

But, why is EPS important and how is it used? Using EPS will help you scope or determine:

An appropriate LM or SIEM – since many LMs or SIEMs are rated or licensed based on EPS or amount of logged data, it is critical that you have an accurate estimate of your EPS or else you risk oversizing (paying too much) or under sizing (losing data) your solution.

Your online and offline storage requirements – if you have compliance requirements then you will have some type of retention policy. Your retention policy along with the amount of log data generated will determine your storage requirement.

Your daily storage management – Storage costs money and you don't want to spend more than you have to, however, you do not want to run out of storage either. Understanding your EPS will better allow you to manage and plan your log data storage needs.

Normal vs. Peak

There are two EPS metrics that need to be factored into your planning and analysis: Normal Events per Second (NE_x), and Peak Events per Second (PE_x).

NE_x, just as its name implies, represents the normal number of events per second while PE_x, represents the peak number of events that are caused by abnormal activities such as a security attack. While PE_x is a theoretical, albeit impractical, measurement, it does need to be factored in as it could impact the performance of your SIEM/LM solution as well as your storage requirements.

Why should you be concerned about PE_x? Quite simply, a single security incident such as a worm, virus or DOS may fire off thousands of events per second from the firewall, IPS, router, or switch at a single gateway. Multiply this by your multiple subnets and it can quickly spiral out of control.

Log Volume

Now that we understand our EPS, we can estimate the amount of log data that is being generated per second and per day based on the following formulas:

$$\frac{GBytes\ of\ Data}{Second} = \frac{(EPS\ x\ Bytes\ Per\ Event)}{1{,}000{,}000{,}000}$$

$$\frac{GBytes\ of\ Data}{Day} = \frac{Gbytes\ of\ Data}{Second} \times 64{,}800$$

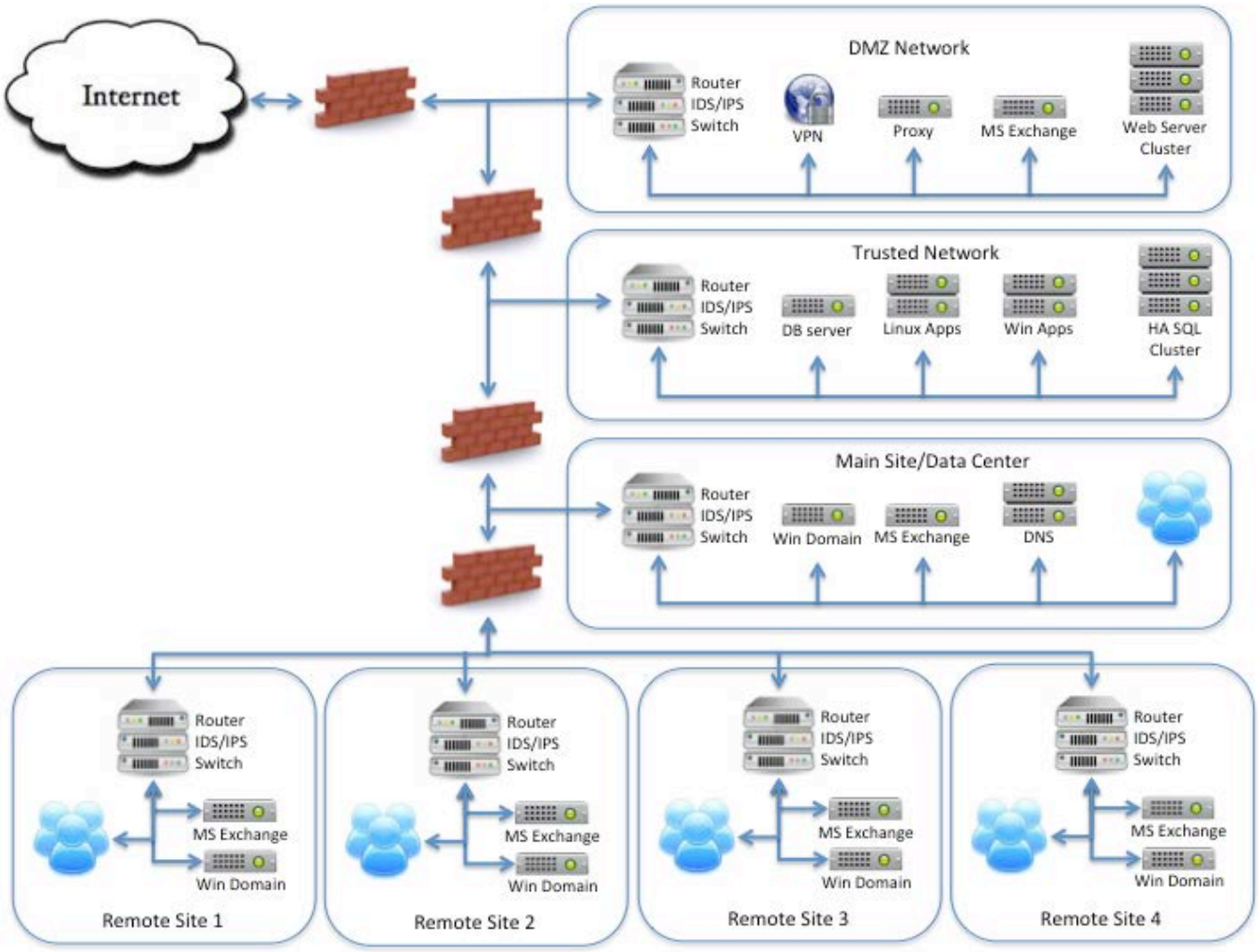
Some SIEM and LM solutions in the market license by the amount of log data collected, or indexed, on a daily basis. This calculation will allow you to estimate the size of the license required under that model.

In addition, by applying the above calculation to your data retention policies, you can estimate the amount of storage required for your log data.

$$Gbytes\ Storage = \frac{(EPS\ x\ Bytes\ Per\ Event)}{1{,}000{,}000{,}000}\ x\ 64{,}800\ x\ Retention\ Period\ (in\ days)$$

Our Hypothetical Infrastructure

Now let's apply what I have discussed so far to a hypothetical mid-sized organization with 1000 employees located across 5 sites and containing one data center (see network diagram).



DISCLAIMER AGAIN: The estimates in the following table are simply best estimates for EPS and should be used only for illustrative purposes. The most accurate measurement of EPS is to use a simple syslog server, such as Kiwi Syslog Server, and measure actual EPS over a period of time.

Quantity	Type	Description	Avg. EPS	Peak EPS	Avg. Peak EPS
1000	Employee Endpoints	Desktops & Laptops (.005 EPS/Employee)	5	50	20
5	Network Switches	One @ each location - NetFlow Enabled	150	1500	750
5	Network Gateway/Router	One @ each location	5	50	25
5	Windows Domain Server	One @ each location	35	350	100
2	Windows Application Server	at Data Center	4	900	450
2	Linux Server	at Data Center	4	900	450
6	Exchange Server	One @ each location, 2 @ Data Center	3	1200	600
3	Web Servers (IIS, Apache, Tomcat)	High availability cluster @ Data Center	1.5	2250	1125
2	Windows DNS Server	at Data Center -failover	1	100	100
4	Database Server	MSSQL, Oracle, Sybase, etc…):	2	40	40
2	Firewall	Trusted	10	1000	1000
2	Firewall	DMZ	60	3000	3000
7	IPS/IDS	1 @ each location, 1 in DMZ, 1 in Trusted	70	10500	5250
1	VPN	at Data Center facing the internet	2	150	150
1000	AntiSpam/Proxy	.005 EPS/employee	5	50	20
1000	Antivirus Server	.005 EPS/employee	5	50	20
		Totals	363	22,090	13,100
		Avg Log Size (bytes)	100	100	100
		Bytes/Sec	36,250	2,209,000	1,310,000
		GBytes/Day	3.13	190.86	113.18

As you can see from this example, it is quite easy to be generating multiple GBytes of log data per day with just normal activity. If one were to scale their SIEM, LM or storage based on the peak load or average peak load, then it can get quite expensive.

Summary

As stated at the beginning of this paper, there is no simple "rule-of-thumb" approach to estimating the amount of log data that can be generated by an organization. There are simply too many factors that have an impact. When scoping a SIEM or LM solution, the most accurate method to determine log data generation is to take a sample over a given time period using a simple syslog server tool that can tell you exactly how much data his being generated.

If you are concerned about the uncertainty of volume based licensing models for a SIEM or LM solution, then you can, alternatively, evaluate products that license based on the number of nodes that are monitored. Node based licensing will offer a more predictable cost without having to go through the exercise of estimating log volume. SolarWinds Log & Event Manager is an example of a low-cost, easy-to-use, software based Security Information Event Management/Log Management solution that collects, correlates, and analyzes log data in real-time. Learn more about SolarWinds Log & Event Manager.