Capacity Planning for ArcSight Logger Software

ArcSight Logger is a highly scalable log management product. It can be deployed to meet almost any availability and scalability needs. The purpose of this brief is to assist in capacity and deployment planning for the software version of ArcSight Logger.

Hardware Reference

We will present two reference systems for capacity planning and discuss variations as necessary. These are meant to be good reference architectures to give an idea of what the capacity is for different hardware configurations. They are not minimum requirements.

Reference System 1: Small to Medium

- Processor: 4-core 64-bit: 1 x current Intel Xeon Quad Core or equivalent
- Memory: 12 GB
- Disk:
    - Approximately 850 GB available for log data storage
    - Root partition: 120 GB
- OS: Linux 5.4 64-bit

Reference System 2: Large

- Processor: 8-core 64-bit: 2 x current Intel Xeon Quad or equivalent
- Memory: 24 GB
- Disk:
    - Approximately 4.5 TB available for log data storage
    - Root partition: 400 GB
- OS: Linux 5.4 64-bit

When scaling outside or between these configurations, keep the proportion of power in CPU, memory and disk relatively fixed to give a good balance of performance for both load capacity and searching.

The recommendations leave approximately half the CPU available for searches. Searching will likely be primarily disk I/O intensive, but can be CPU intensive also, depending on the search query. If you will be running many concurrent searches, allow more CPU.

Indexing is primarily disk I/O intensive, but can impact CPU utilization as well, depending on the data and index settings used.

Note that some of the disk storage will be used by the OS for files and swap.

Capacity Planning

First, decide what data you would like ArcSight Logger to store and how to send that data to ArcSight Logger. There are two ways data can be imported:

- **Directly (as raw data):** such as syslog sent directly over the network or log files on disk
- **Through ArcSight Connectors (as CEF data):** You can leverage ArcSight Connectors to collect logs from an enormous variety of sources. Using ArcSight Connectors is an easy way to collect data from almost any source, standardize it's format to make searching easier and faster, and add information to help make searching more powerful.

Next, determine which reference hardware system is appropriate for this load and your storage space needs.

For partner-fed CEF data, treat it as imported directly for the purposes of storage size, and through ArcSight Connectors for the purposes of daily capacity.

Determine Daily Capacity Needs

Daily capacity is primarily determined by the disk I/O speed, but parsing and indexing also use some CPU.

Note that the reference systems can handle a peak burst load much higher than the sustained load. That is, the system will not lose data if you send in a burst. After the burst, the indexing will catch up. Searches over recent data may be slower until it does catch up, but no data is lost.

Step 1: Compute the EffectiveIn GB/day

Compute:

EffectiveIn GB/day = Direct GB/day + [ConnectorIn GB/day * Connector Storage Conversion Factor]

Where:

- **Direct GB/day**: the size, in GB, of raw logs sent directly to ArcSight Logger
- **ConnectorIn GB/day**: the size, in GB, of raw logs sent through ArcSight Connectors to ArcSight Logger
- **Connector Storage Conversion Factor**: 2.95, a factor to convert the size of the raw logs to the log size after enhancement by the ArcSight Connector

Example 1: All Data Direct

In this example, all the data is sent to ArcSight Logger for analysis and storage; it is directly received by ArcSight Logger without any connectors. Consider the case where the data is from syslog files created by a set of local syslog concentrators that already log to disk from several UNIX systems, as well as devices. You would configure the syslog file receiver in ArcSight Logger to pull those files, or configure those concentrators to send to a Logger syslog receiver. You can understand how large the files are just by looking at their sizes on disk, and then determine the direct GB/day based upon those sizes.

Say there are approximately 100 GB/day of syslog files currently stored by the concentrators. When ArcSight Logger is configured to load those files, the following would apply:

EffectiveIn GB/day = Direct GB/day = 100 GB/day

Example 2: All Data via an ArcSight Connector

In this example, all data processed by ArcSight Logger is sent through an ArcSight Connector. Consider the case where you have several different firewall types – firewalls from different vendors, but all of which use syslog logging. Configure these sources to send logs to a syslog daemon connector, which sends to ArcSight Logger. One advantage of this case is that you could query across all the firewalls more easily, since the ArcSight Connector will normalize the log entries and attach categorization information from the categorization content. Determine the ConnectorIn GB/day by using a Syslog concentrator to log the firewall types to disk, in addition to sending to the ArcSight Connector for a few days and get an average. Then, you can disable that concentrator. Say there are approximately 100 GB/day of syslog data in the files. The following would apply:

EffectiveIn GB/day = ConnectorIn GB/day * Connector Storage Conversion Factor
= 100 GB/day * 2.95
= 295 GB/day

Note that ArcSight Connectors would also need to be used if the data acquisition method is not through syslog or file access. Consider the case where you have a Checkpoint firewall as well as some Oracle database auditing logs. You would install ArcSight Connectors for each of those sources. Determining the raw size would require using the application (Checkpoint or Oracle), rather than a disk size check.

Example 3: Mixed Data Acquisition Case

In this example, some data processed by ArcSight Logger is sent directly to ArcSight Logger syslog or file receivers and some is sent through ArcSight Connectors. Consider the case where one data center (DC1) has several devices that log through syslog and can have a local ArcSight Connector, while another data center (DC2) has many UNIX systems already logging to a concentrator, but there is no ArcSight Connector. To determine the collective raw syslog log file sizes per day from DC1, let's assume there are 30 GB/day from its devices being sent to an ArcSight Connector. Now determine the collective raw syslog log file sizes per day from DC2. Assume that there are 40 GB/day from systems in DC2 that are sent to ArcSight Logger directly using an ArcSight Logger syslog receiver. The following would then be true:

EffectiveIn GB/day = Direct GB/day + [ConnectorIn GB/day * Connector Storage Conversion Factor]
= 40 GB/day + [30 GB/day * 2.95]
= 128.5 GB/day

Step 2: Determine Hardware Specifications to Use

EffectiveIn GB/day	Specification to Use
< 70 GB/day	Reference System 1: Small to Medium
> 70 GB/day and < 200 GB/day	Reference System 2: Large
> 200 GB/day	Linearly add more servers of either reference system type, depending on the size of the expected load

For example, if the EffectiveIn GB/day computed in Step 1 is 50 GB/day, then use Reference System 1: Small to Medium. Another example: if the EffectiveIn GB/day computed in Step 1 is 120 GB/day, then use Reference System 2: Large.

Determine Retention Needs

Once you know how much of each type of data per day you wish to process, determine how long you wish to retain the data.

This section follows a simplified model where all the data is stored with the same retention policy. However, ArcSight Logger allows you to set multiple retention policies, one per storage group, so that some data can be stored longer with the same space requirements. If you want to retain different data sets for different lengths of time, determine the daily load for each data type separately and add up the retention needs for each type.

Compute the space required for a day of retention using this formula:

One Day of Storage GB = [Direct GB/day * Direct Storage Factor] + [ConnectorIn GB/day * ConnectorIn Storage Factor]

Where:

- ***Direct GB/day***: defined in Step 1 of "Determine Daily Capacity Needs" section
- ***ConnectorIn GB/day***: defined in Step 1 of "Determine Daily Capacity Needs" section
- ***Direct Storage Factor***: 0.29, a factor to convert the size of raw logs to the storage size, given typical storage configurations
- ***ConnectorIn Storage Factor***: 1.34, a factor to convert the size of raw logs sent to the ArcSight Connector to the storage size in ArcSight Logger, given typical storage configurations

Multiply the resulting number by the number of days retention required to determine the total storage requirements:

Total Storage Requirement (GB) = One Day of Storage GB * Number of Days Retention

Note that you can use local or SAN storage.

Effective compression rates will vary with the data, as well as with the index settings in ArcSight Logger. However, this is a good guideline for planning for most use cases.

You can use NFS or CIFS for primary storage, but the numbers above are based on local or SAN storage. The performance will be negatively impacted with NFS or CIFS.

Note: Calculations contained in this brief are based on actual averaged data from ArcSight customers. Individual company factors or computations may vary, depending on the type of data that is sent to ArcSight Logger.
