Cribl

>WHITE PAPER_

Data independence: 5 proven tactics to optimize Splunk software.

>WHITE PAPER_

Data independence: 5 proven tactics to optimize Splunk software.

According to IDC's Global Data Sphere and Global Storage Sphere, structured and unstructured data is forecasted to grow to 181ZB by 2025. With this tremendous growth of data in motion, it's only a matter of time before the tools designed to help manage and understand our large-scale distributed systems are overwhelmed and become prohibitively expensive. You need to plan for what will happen to your data long term if your vendor or requirements change to mitigate risk to your business and to take control of continuity.

The world is going to create, capture, copy, and consume over 181ZB by 2025.

– IDC WW Global DataSphere and Global StorageSphere Structured and Unstructured Data Forecast 2022-2026 166ZB of unstructured data – logs, events, metrics and traces.

It's no secret that Splunk software provides one of the most comprehensive observability experiences in the industry. It is known for ingesting data from multiple sources, interpreting data, and incorporating threat intelligence feeds, alert correlation, analytics, profiling, and automation/summation of potential threats. While all of these features are necessary to build a comprehensive Observability practice, they also put tremendous pressure on capacity and budget. Licensing and infrastructure costs quickly become prohibitively expensive, and force tradeoffs between cost, flexibility, and visibility. The real question is, can anything be done to mitigate the impact of increasing data storage and analysis costs without impeding growth and security?

The answer is "Yes". Data analysis is not just about the volume of data, it is about the value of the data (which could be different for different stakeholders); and with Cribl, administrators can optimize their licensing budgets and achieve dramatic storage cost reduction by following these 5 simple steps:

1. Filter out duplicate and extraneous events.
2. Route to more cost-effective destinations.
3. Trim unneeded content / fields from events.
4. Condense logs into metrics.
5. Decrease operational expenses.
6. Filtering out the noise.

Filtering out the noise.

The first and easiest option to optimize your Splunk software licensing and infrastructure charges is to filter out extraneous data that is not contributing to insights. By employing a simple filter expression, an administrator can reduce the volume of raw data destined for Splunk software in the first place. You can apply the filters to drop, sample, or suppress events. All of these filtering options can be configured based on meta information, such as hostname, source, source type, or log level, or by content extracted from the events, or both.

- Dropping: 100% of this type of data is discarded or routed to a cheaper destination.
- Sampling: If there are many similar events, only 1 out of a defined sample set is sent to Splunk software.
- Dynamic Sampling: low-volume data of this type is sent to Splunk software, but as volume increases, sampling begins.
- Suppression: No more than a defined number of copies of this type of data will be delivered in a specified time period.

The result is a reduction in total data, and a higher percentage of useful data. Indexed data takes approximately 4X more space to store than raw machine data, and it is a good best practice to deploy Splunk software for high availability, which usually replicates the indexed data 3 times. This means that for every bit of data sent, it takes 12X more resources to store it there compared to inexpensive object-based storage.

Indexed data takes approximately four times more space to store than raw machine data.

56 percent reduction! By employing a simple filter expression, an administrator can reduce the data destined for Splunk software in the first place.

Routing to the most cost-effective destination(s)

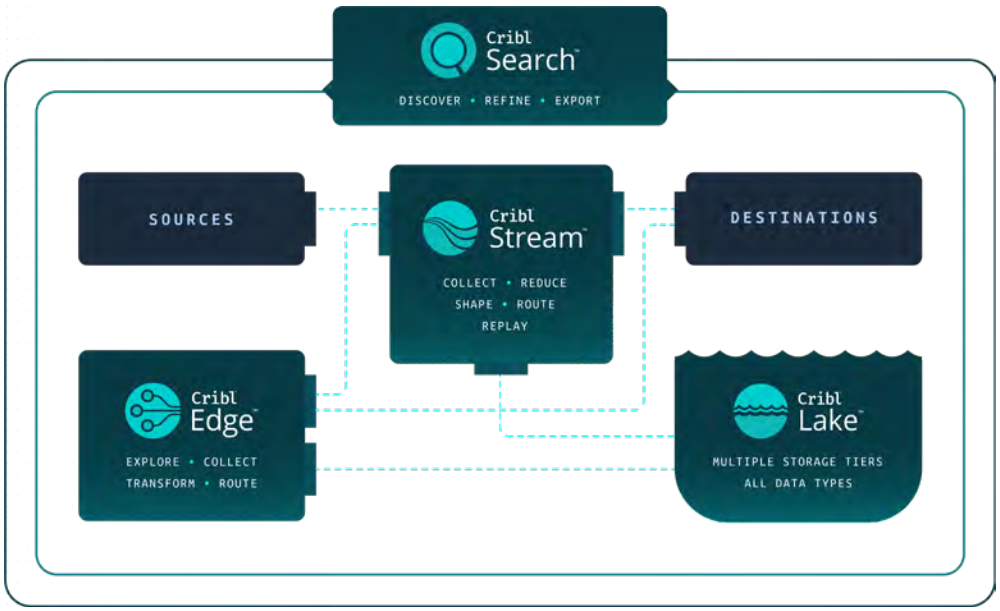
Routing data to the appropriate tool for the job is another way to save money on Splunk software. As mentioned above, indexed storage can take about 12X the resources as object storage, with linear costs to match. Another way to potentially reduce licensing and infrastructure costs with Splunk software is to route data to a more cost-effective location instead of storing it all there. An important enabler here is the ability to separate the system of analysis from the system of retention, which can be an inexpensive storage option, like Cribl Lake, Amazon S3, or a host of other storage options. These stores are generally pennies on the dollar compared to indexed data in block storage, and allow administrators to capitalize on the lower cost and increased compression ratios while still complying with data retention requirements.

This solution also allows an administrator to retain a full-fidelity copy of the original logs, in vendor-agnostic raw format (in case of future tooling changes), and concurrently deploy the filtering options from above to significantly reduce the data sent and retained in Splunk software indexed storage.

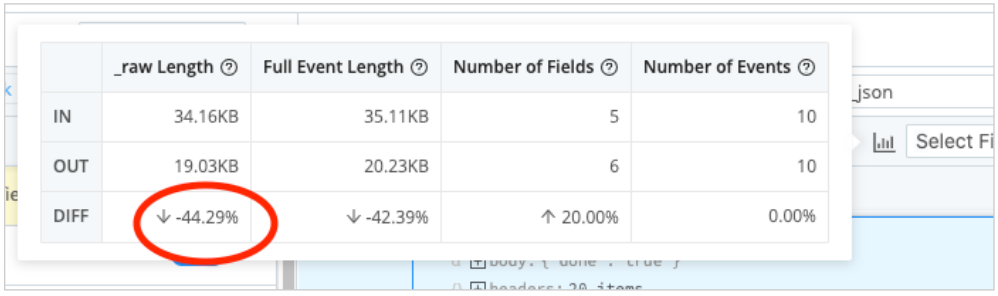
Many Splunk software customers see 30% savings or more just by employing these first two steps — filtering and routing — since the cost of retaining data is linear and continuously increasing as data is added.

An important consideration when separating the system of analysis from the system of retention is ensuring there is a way to retrieve data from the system of retention without having to wait to thaw out cold storage or send someone to find it on a tape backup system. Cribl Stream provides the ability to easily and cost-effectively retrieve stored data with our Replay feature. Replay gives administrators the power to specify parameters, such as user, date/time, or other information, that identify which data to retrieve and send to Splunk software or other tool for immediate analysis.

Cribl provides the ability to easily and cost-effectively retrieve stored data with stream replay.



Stream™ Replay gives administrators the power to specify parameters, such as user, date/time, or other information, that identify which data to retrieve from object stores and send to Splunk software or other tool for immediate analysis.



	_raw Length	Full Event Length	Number of Fields	Number of Events
IN	34.16KB	35.11KB	5	10
OUT	19.03KB	20.23KB	6	10
DIFF	↓ -44.29%	↓ -42.39%	↑ 20.00%	0.00%

Administrators can see up to 75% reduction in log volume just by getting rid of fields that do not contain any data at all.

Reducing volume with pre-processing

In addition to filtering machine data to optimize your Splunk software deployment, another option is to reduce the volume of the events themselves. While a verbose set of logs can aid in troubleshooting, it's fairly common to see many unnecessary or unwanted fields within a specific event. By using pre-processing capabilities in Cribl Stream, it is possible to trim the event itself by removing NULL values, reformatting to a more efficient format (XML to JSON, for example), dropping duplicate fields, or even changing an overly verbose field to a more concise value. While the number of individual events may be the same when using pre-processing, depending on the dataset, administrators can see up to 75% reduction in log volume just by getting rid of fields that do not contain any data at all!

Compressing logs into metrics

Administrators can adopt a discerning data management strategy using data storage techniques which are fit for purpose.

Many of the highest-volume data sources come from having to ingest extraneous information just to access a single useful statistic, otherwise known as a metric. Web activity logs, NetFlow, and application telemetry are great examples of this type of event, and another way to realize significant savings in Splunk software is to aggregate logs like this into summary metrics. Since a metric usually contains only a name, a value, a timestamp, and one or more dimensions representing metadata about the metric, they tend to require much less horsepower and infrastructure to store than log files.

Stream gives administrators the power to extract fields of interest, using built-in Regex Extract or Parser functions, and then publish the result to metrics. Once aggregated, administrators will see a major reduction in event counts and data volume, and then can choose whether to send those metrics to Splunk software, or potentially route the metrics instead to a dedicated time series database (TSDB), such as InfluxDB or Datadog for efficient storage and retrieval.

A high standard of speed

When it's literally a matter of national security, rapid identification and resolution of issues is of critical importance. Analysts at many federal agencies use Splunk software to dig into and clarify potential anomalies, and AFS brings Stream into the mix to ensure those analysts get the best performance possible.

> **"Analysts were building many searches just to build lookup tables; we had hundreds of searches scheduled just to build out IP lookups. Using Stream makes Splunk software more efficient by letting you save your search resources for faster searching instead of having to build metrics to search."**
>
> — **Gared Seats**, Security Engineer, Accenture Federal Services

Decreasing operational expenses

One last way to reduce spend on Splunk software is simply to reduce the number of hours and resources dedicated to supporting it. By employing functions such as filtering, parsing, and reformatting in Cribl Stream, it is possible to reduce the overall noise to such a degree that finding the valuable and necessary information takes far less time in Splunk software (or any other data analysis platform). Once events are optimized before being indexed, crafting the necessary search is easier, and the actual search itself runs faster as well. This not only reduces the time to insight, but it also removes the burden of bloated infrastructure, constant juggling of content and compliance requirements, and building out custom solutions to solve a point problem.

In addition, consolidating multiple tools into a single, centralized interface further reduces the operational overhead associated with observability deployments. Administrators can replace the functionality of intermediate log forwarders, like Splunk software's heavy forwarder or Logstash, and other open source tools such as syslog-ng or NiFi, with Stream. The obvious advantage here is fewer tools to install, manage, and maintain, but Stream also delivers increased efficiency by consolidating ingestion, processing, and forwarding of data streams for centralized visibility and control.

Summary

Splunk software is a leader in the data analytics industry for a reason, but that superior experience can get bogged down by noisy data driving up processing and storage requirements. To separate the signal from the noise, Cribl provides a data engine purpose-built for IT and Security teams capable of exploring and processing billions of events per second, extract the valuable data teams need, when they need it, in the desired format, with the ability to recall and rehydrate that data at will giving you full data independence and control to optimize your Splunk software deployment. Implementing a couple of these options using Stream could cut your log volumes dramatically!

Download the Cribl suite of products, including Stream, here.