Cribl

WHITE PAPER

Five Ways to Save on Splunk

WHITE PAPER

Five Ways to Save on Splunk

According to IDC's GlobalSphere Data Forecast, the world is going to create, capture, copy, and consume over 59 Zettabytes of data in 2020 – the equivalent of filling up a Terabyte hard drive every single day for 161 MILLION years! This statistic will only continue to increase as well; IDC projects a combined annual growth rate of 26% YoY through 2024. 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 begin to also balloon and become prohibitively expensive.

It's no secret that Splunk provides one of the most comprehensive observability experiences in the industry – this is why they've consistently been named a leader in data analytics. Splunk 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, what can be done to mitigate the impact of increasing data storage and analysis costs without impeding growth and security?

With a little bit of work and help from a product like Cribl Stream, administrators can typically trim licensing and storage costs dramatically 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*

THE WORLD IS GOING TO CREATE, CAPTURE, COPY, AND CONSUME OVER **59 ZETTABYTES** OF DATA IN 2020.

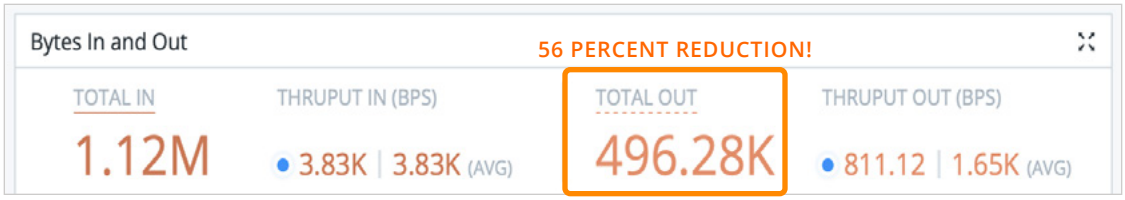
INDEXED DATA TAKES APPROXIMATELY FOUR TIMES MORE SPACE TO STORE THAN RAW MACHINE DATA.

FILTERING OUT THE NOISE

The first and easiest option to save money on Splunk licensing and infrastructure costs is to filter out extraneous data that is not contributing to insights. By employing a simple filter expression, an administrator can reduce the data destined for Splunk in the first place. You can apply the filter 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*
- **Dynamic Sampling:** *low-volume data of this type is sent to Splunk, 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*

In addition to the licensing savings from reducing the number of events sent to Splunk, filtering also leads to savings on infrastructure costs. Indexed data takes approximately 4X more space to store than raw machine data, and it is a good best practice to deploy Splunk for high availability, which usually replicates the indexed data 3 times. This means that for every bit of data sent to Splunk, it takes 12X more resources to store it there compared to inexpensive object-based storage.



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

ROUTING TO THE MOST COST-EFFECTIVE DESTINATION(S)

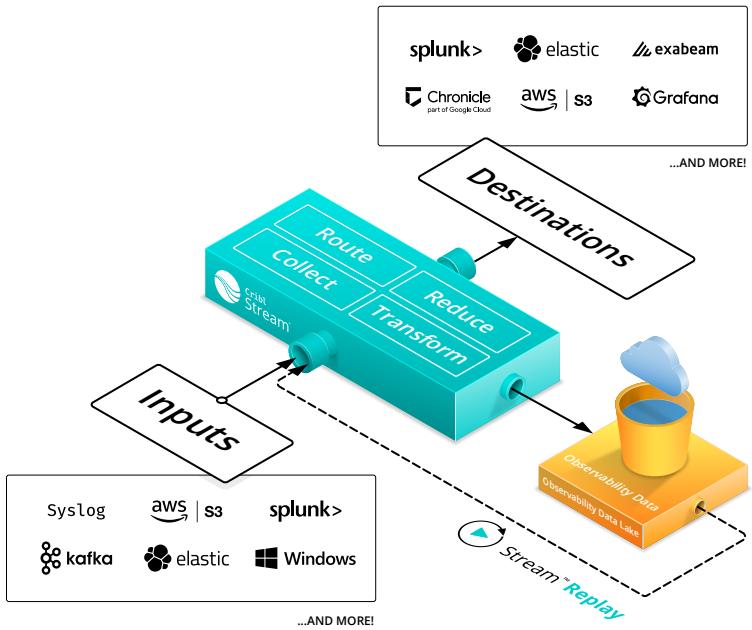
Routing data to the appropriate tool for the job is another way to save money on Splunk. As mentioned above, indexed storage can take about 12X the resources as object storage, with linear costs to match. Another easy way to reduce both licensing and infrastructure costs with Splunk is to route data to a more cost-effective location instead of storing it all in Splunk. An important enabler here is the ability to separate the system of analysis, Splunk, from the system of retention, which can be an inexpensive storage option, such as S3 or MinIO. 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 indexed storage.

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

CRIBL PROVIDES THE ABILITY TO EASILY AND COST-EFFECTIVELY RETRIEVE STORED DATA WITH STREAM REPLAY.

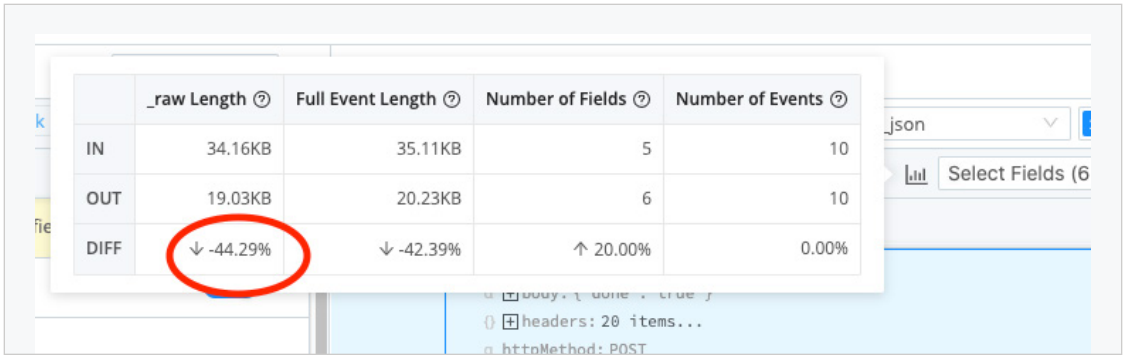
An important consideration when separating the system of analysis from the system of retention is ensuring there is a way to retrieve data from the system of retention without having to wait to thaw out cold storage or send someone to find it on a tape backup system. Cribl Stream provides the ability 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 or other tool for immediate analysis.



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 or other tool for immediate analysis.

REDUCING VOLUME WITH PRE-PROCESSING

In addition to filtering machine data to save money on Splunk, 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!



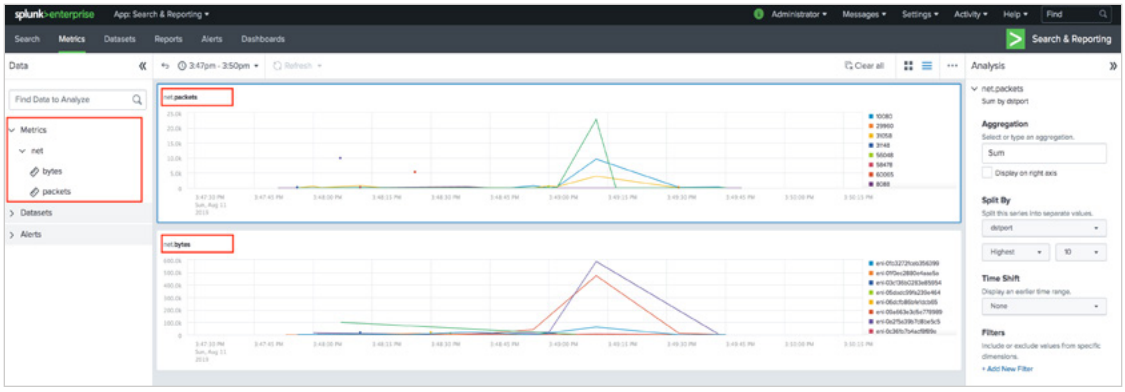
Administrators can see up to 75% reduction in log volume just by getting rid of fields that do not contain any data at all.

ADMINISTRATORS CAN ADOPT A DISCERNING DATA MANAGEMENT STRATEGY USING DATA STORAGE TECHNIQUES WHICH ARE FIT FOR PURPOSE.

COMPRESSING LOGS INTO METRICS

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 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, or potentially route the metrics instead to a dedicated time series database (TSDB), such as InfluxDB or Datadog for efficient storage and retrieval.



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.

DECREASING OPERATIONAL EXPENSES

One last way to reduce spend on Splunk 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 (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 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 is a leader in the data analytics industry for a reason, but that superior experience can often come with a hefty price tag. Cribl Stream gives administrators the power to control and shape all machine data, making it possible to realize the benefits of a world-class observability solution as well as stay within (or under) budget. The only solution of its kind, Stream was purpose-built to help customers to unlock the value of all machine data. Better yet, it's free up to 5TB/day for Splunk users, so why not give it a try? Implementing a couple of these options using Stream could cut your log volumes dramatically!

Download Stream for free **here**.

ABOUT CRIBL

Cribl makes open observability a reality for today's tech professionals. The Cribl product suite defies data gravity with radical levels of choice and control. Wherever the data comes from, wherever it needs to go, Cribl delivers the freedom and flexibility to make choices, not compromises. It's enterprise software that doesn't suck, enables tech professionals to do what they need to do, and gives them the ability to say "Yes." With Cribl, companies have the power to control their data, get more out of existing investments, and shape the observability future.Founded in 2017, Cribl is a remote-first company with an office in San Francisco, CA. For more information, visit **www.cribl.io** or our **LinkedIn**, **Twitter**, or **Slack** community.