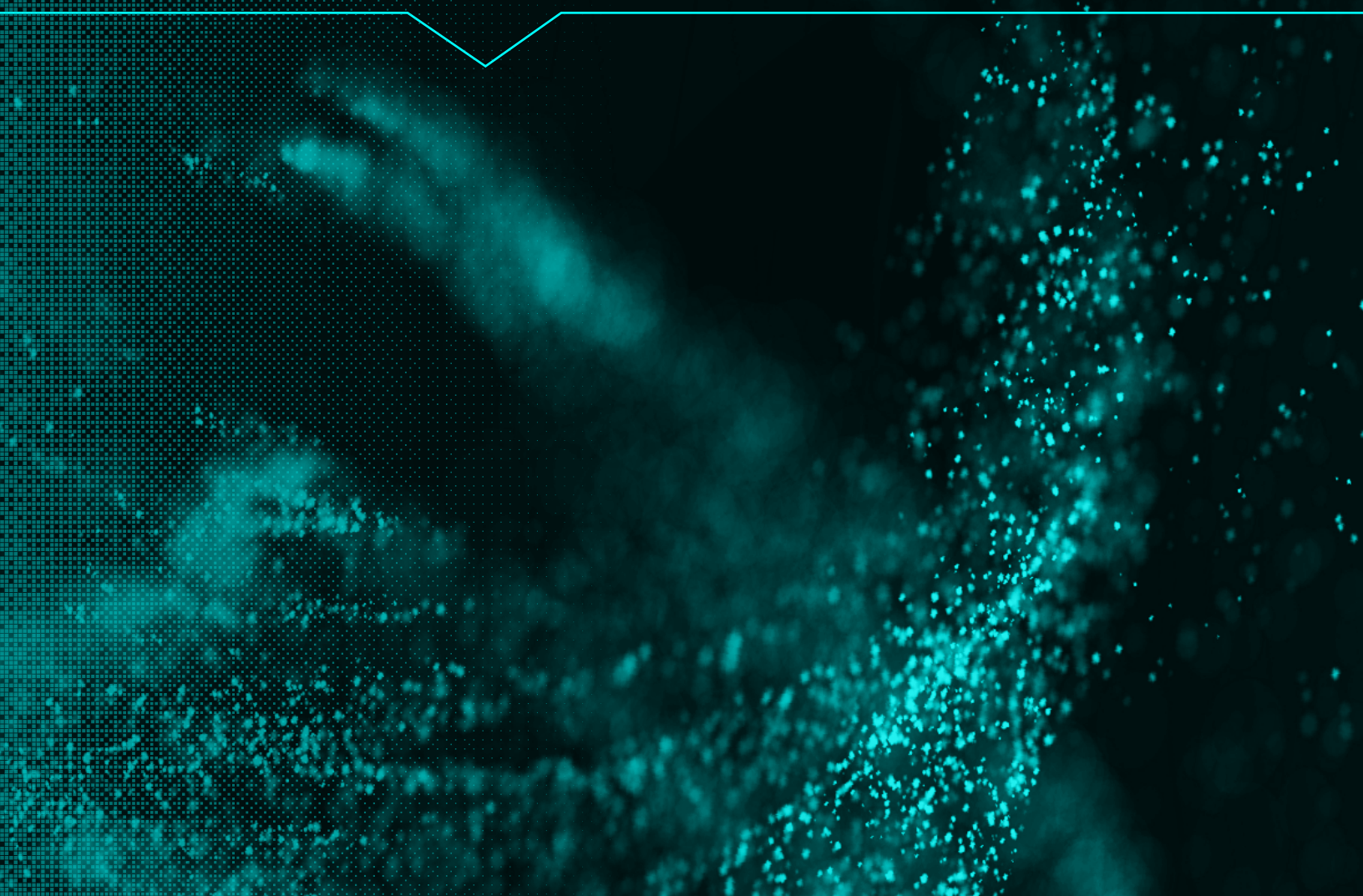
Cribl

>WHITE PAPER_

Improving Splunk software and lowering CPU usage with Cribl.



Cribl

>WHITE PAPER_

Improving Splunk software and lowering CPU usage with Cribl.

THE CHALLENGE

Enterprises need a versatile observability solution for seamless integration, flexible data routing, and scalable decision-making.

THE SOLUTION

Cribl's comprehensive collection, processing, routing solution enables optimized data analysis in Splunk software and informed business decisions at scale.

- **Route data from various sources to Splunk software with ease.**
- **Enrich data with third-party sources for additional context.**
- **Eliminate null values and duplicate fields, or aggregate logs intometrics for increased downstream performance.**
- **Analyze high-value data, and cost-effectively store the rest.**

Introduction: Doing more with your Splunk software environment.

In a world where data reigns supreme, Splunk software stands out as a powerful platform for turning vast amounts of unstructured data into actionable insights. As the volume of data skyrockets, organizations face challenges managing it effectively. Sound familiar?

That's where the Cribl suite of products emerges as a game changer, enhancing Splunk software's capabilities by optimizing data flows, reducing processing loads, and enabling more strategic data management practices. That way, you can manage data more efficiently while transforming that data into a strategic asset for the enterprise.

At the heart of the solution is Cribl Stream, a dynamic pipelining engine that facilitates efficient data routing into Splunk software while refining that data in transit, ensuring that only high-value, actionable information is ingested. This process significantly enhances its search performance and reduces the infrastructure's CPU load, optimizing operational costs and system responsiveness. But why stop there? With the addition of Cribl Edge and Cribl Search, we extend this capability right from the edge of your environments (where data is born) to the depths of your data lakes and storage (where insights are uncovered!).

This white paper will step through a new method for searches and how you can use Cribl's suite of products with Splunk software. Read on for a comprehensive overview of how each component — Cribl Stream, Edge, and Search — can be leveraged to maximize your environment's efficiency. From reducing unnecessary data ingestion to enriching data for deeper insights, we will explore how these tools collectively empower organizations to not just manage but regain control of their data landscape.

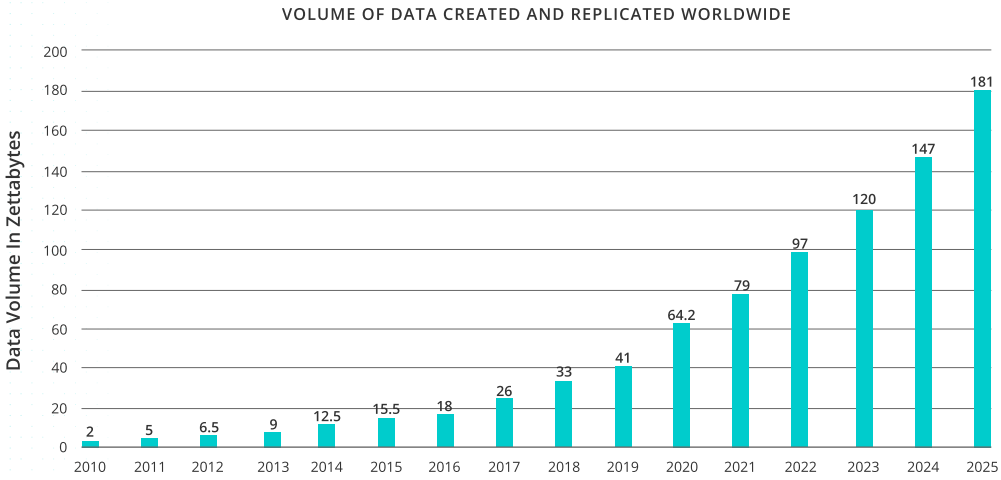
Opportunities to amplify the power of Splunk software with the Cribl Suite

Splunk software shines in its ability to turn complex, unstructured machine data into valuable insights, similar to how you'd search the internet for information. However, given its design for search-time analytics, where data structuring happens during the search, challenges arise when dealing with vast amounts of data. This can affect performance, especially when generating detailed reports or visualizations from large datasets.

As the reliance on Splunk software increases for a broader range of use cases, these performance issues can become more pronounced. This is compounded by the rapid growth of data across all industries, pushing the limits of what organizations can manage and analyze effectively.

The cost implications of using Splunk software for long-term data storage can also be restrictive, limiting the ability to fully utilize licenses due to the high expenses associated with storing massive volumes of data.

When you use Cribl Stream to optimize your Splunk software environment, you can ingest more data, reduce infrastructure size, and lessen hardware requirements and overall cost.



A new strategy for enhanced performance.

To address these challenges, a shift in strategy is essential. Implementing new approaches, such as populating index-time fields and employing time-series metrics databases, can significantly improve search efficiency and overall Splunk software performance. These methods not only speed up data retrieval but also optimize CPU usage, leading to more cost-effective and scalable deployments.

By strategically ingesting data and ensuring that only relevant, pre-processed data is analyzed by Splunk software, organizations can mitigate the challenges associated with large-scale data analysis. This approach allows for the accommodation of more data and users without correspondingly large increases in infrastructure or licensing costs.

Benefits of optimizing data for Splunk software.

Ingest more data.

Optimizing data before ingestion enables the handling of more data, supporting growth, and enabling more extensive user access without significant increases in costs.

Reduce Splunk software infrastructure size.

Enhanced search performance and efficiency can lead to a reduced need for indexers and search heads, leading to cost savings and more streamlined operations.

Lower overall cost.

With the move towards workload-based licensing, focusing on CPU usage, optimizing data and searches becomes crucial for managing costs effectively.

Lessen hardware requirements.

Optimized data processing can also reduce the hardware footprint, particularly for Splunk software in the cloud users, by ensuring that searches are more efficient and less resource-intensive.

Adopting these strategies can significantly enhance the value derived from Splunk software, enabling organizations to overcome the challenges of data volume growth and complexity. This approach lays the groundwork for a more efficient, cost-effective approach, even as data demands continue to escalate.

Improving Splunk software performance for search and lowering CPU usage with the Cribl Suite.

Stream can convert logs to metrics, aggregate data, suppress unwanted data, and more.

Optimizing search performance while managing CPU usage is critical in Splunk software ecosystems. To directly address the challenges faced by growing data volumes and the inherent limitations of traditional data management within these environments, teams need to shift towards more advanced and efficient data handling strategies. This new approach, which encompasses techniques such as populating index-time fields and utilizing time-series metrics databases, promises to significantly enhance data processing speeds, leading to improved performance and reduced CPU usage.

The Cribl Product suite — including Cribl Stream, Edge, and Search — offers a comprehensive solution to refine and streamline data before it's processed by Splunk software, ensuring efficiency and scalability.

Cribl Stream

Stream plays a pivotal role by preprocessing data, enriching and filtering it to ensure Splunk software processes only the most relevant information. By performing real-time data shaping—such as enrichment, reduction, and obfuscation—Stream ensures that Splunk software ingests only optimized data, significantly improving query performance and reducing storage and processing requirements.

Cribl Edge

Cribl Edge extends this optimization further by processing data at the source. By filtering and reducing data volume at the edge, Edge minimizes the data sent over the network, reducing bandwidth usage and easing the ingestion load on Splunk software. This means even faster processing times and lower infrastructure demands. This approach has the added bonus of streamlining data flows and enhancing security by reducing the attack surface through data minimization.

A data collection, reduction, enrichment, and routing system for observability data.

An intelligent, scalable, edge-based data collection system for logs, metrics, and application data.

Perform federated "search-in-place" queries on any data, in any form.

A simplified data lake solution to easily store, manage, and access data.

Cribl Search

Cribl Search complements this ecosystem by providing advanced search capabilities, enabling more efficient data exploration and analysis. By leveraging Search, users can perform granular queries across diverse data sets, uncovering insights more rapidly and reducing the computational load on Splunk software's search heads (further contributing to lower CPU usage!).

Recommended deployment option: Cribl.Cloud.

Cribl's got a ton of flexible deployment options, including single-instance and distributed deployment of the solution. Deploy with Cribl.Cloud to quickly launch a Cribl-hosted deployment of the combined Cribl solution (Stream, Edge, and Search). With this option, Cribl assumes responsibility for provisioning and managing all infrastructure, on your behalf.

Incorporating Cribl.Cloud into this suite brings the added advantage of cloud scalability and management simplicity. As a fully managed service, Cribl.Cloud allows organizations to maintain optimal performance of their Splunk software environments without the overhead of managing physical infrastructure. As data volumes grow, you can rest assured your system can scale seamlessly without compromising performance.

Integrating Cribl Stream, Edge, Search, and Cloud into your Splunk software ecosystem translates to significantly improved search performance, reduced CPU usage, and a more scalable, cost-effective data management strategy.

In the sections that follow, we'll explore the practical applications of these tools in enhancing Splunk software's performance, demonstrating how they can be integrated seamlessly into existing workflows to drive efficiency, reduce costs, and unlock new levels of insight from your data.

Improving Splunk software performance for search and lowering CPU usage.

If you're a Splunk software enthusiast, you probably know the value of using the tstats command to achieve performant searches. If you probably also know the value of time-series databases, like a metrics index.

Using a set of Docker containers, Cribl Stream was shown to:

- Improve the performance of Splunk software searches significantly by populating index-time fields and searching via tstats.
- Improve performance in scenarios where a metrics index was populated instead of a traditional event index. Additionally, the search was simplified by using the analytics workspace.
- Performance improvements are even larger in production environments where billions or trillions of events are searched.

Getting started: how to improve Splunk software performance with Cribl Stream.

Stream is a data pipeline solution that transforms unstructured data to be more structured before it persists to disk. This improves sending to Splunk software, as well as sending to other observability solutions like Datadog, Wavefront, the Elastic Stack, or Sumo Logic. Writing to an S3-compliant API, GCP Cloud Storage, or Azure Blob Storage is also simplified.

1. Stream improves your Splunk software search performance using the following methods:
2. Populating data into index-time fields and searching with the tstats command.
3. Converting logs into metrics and populating metrics indexes.
4. Aggregating data from multiple events into one record. This is beneficial for sources like web access and load balancer logs, which generate enormous amounts of "status=200" events.

Suppressing unwanted data, thereby reducing the amount of data searched and resulting in faster speeds.

Cribl engineers tested these methods by running various tests generating 2GB/day in Docker containers on a Macintosh with 8 CPU cores and 32GB RAM. For this test, Tomcat application logs were used; it should be noted that these have high variability with several event formats in one log file. The Cribl Stream pipeline processes the Tomcat application logs. In this case, the pipeline is doing transformations for four distinct types of identified events:

1. Events reporting response times as key-value pairs, in addition to some less-structured data.
2. Multi-line events reporting Java exception errors.
3. Events reporting statuses and metrics inside a JSON payload that takes up part of the event.
4. All other, mostly unstructured, events of info or error severity.

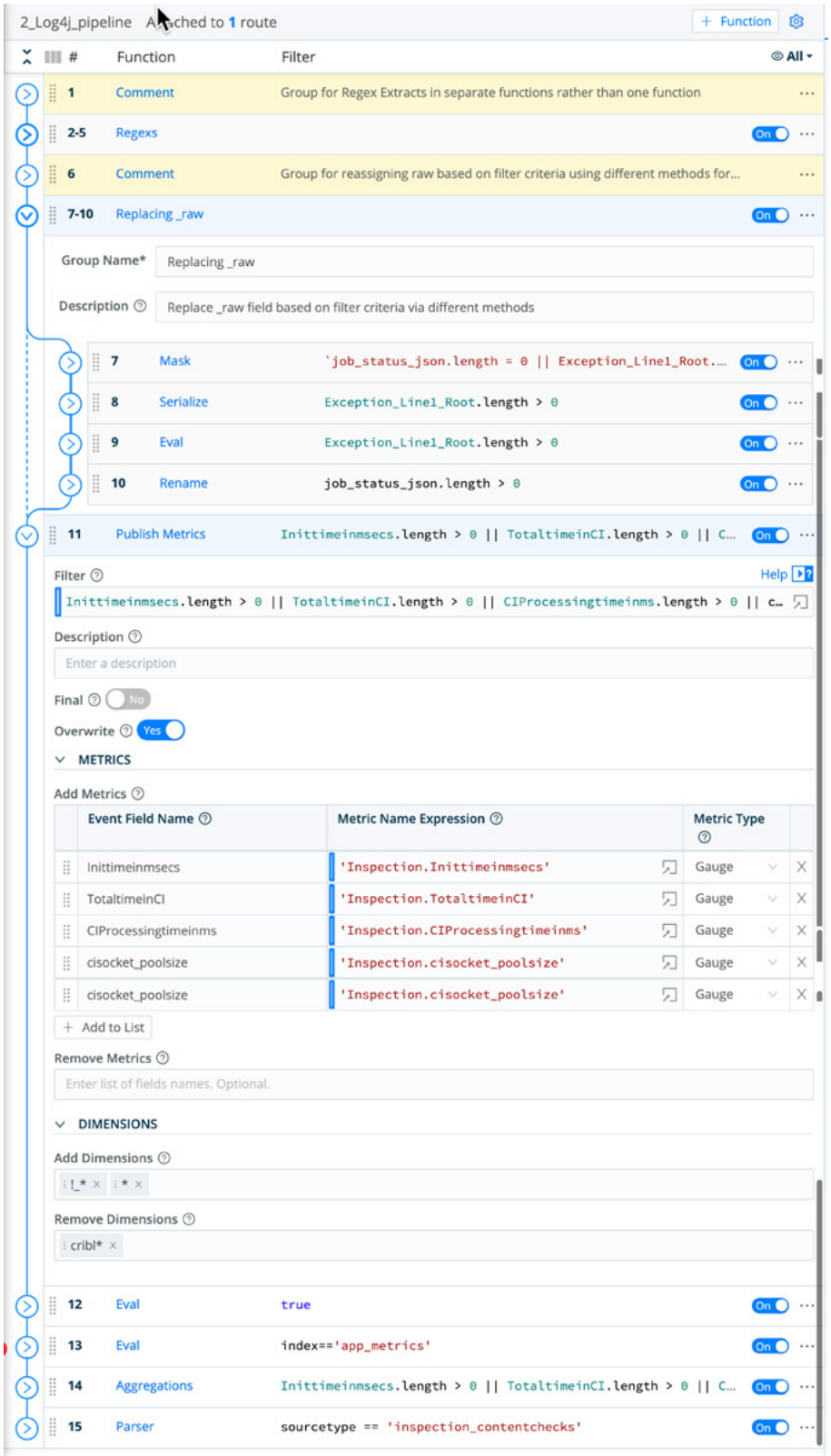


Fig. 1 The Stream pipeline.
Before Stream processing, Tomcat logs typically have JSON buried in some events, metrics buried in others, and additional Java exceptions – making for lengthy logs that are tough to process. Using Stream, admins can easily restructure their events and transform Java exceptions, reducing log size overall.

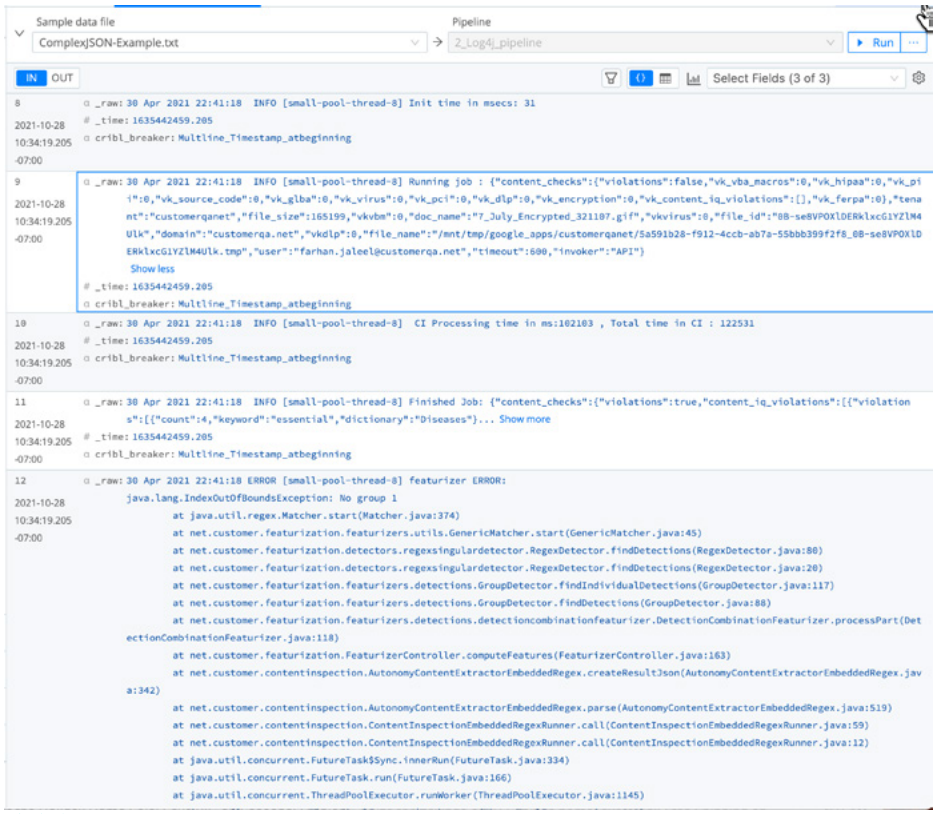


Fig 2: Before processing by Stream.

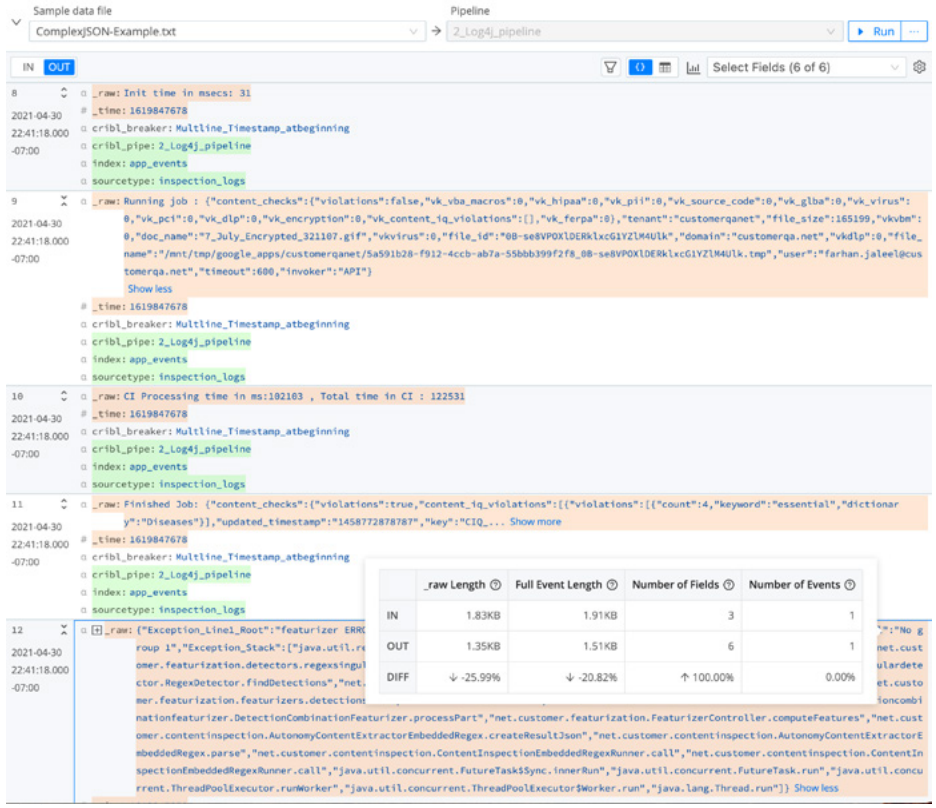


Fig. 3: After processing by Stream.

For our test, Cribl engineers sent two versions of this dataset to a Splunk software instance. One had raw, unprocessed logs from a Stream datagen targeted to index=main. The other? Processed logs, with events targeted to index=app_events, with converted metrics targeted to index=app_metrics.

Here's what the unprocessed events look like in Splunk software:

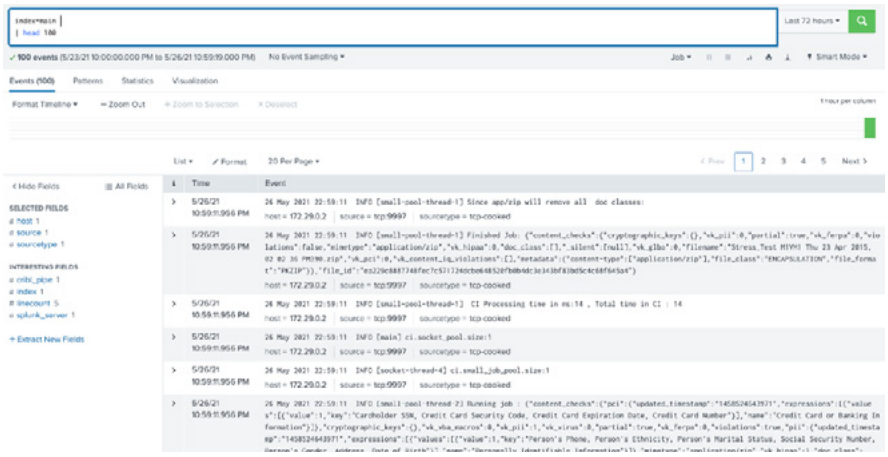


Fig. 4: A Tomcat event in Splunk software, before processing by Cribl Stream.

And here are some of the Stream-processed events in Splunk software:

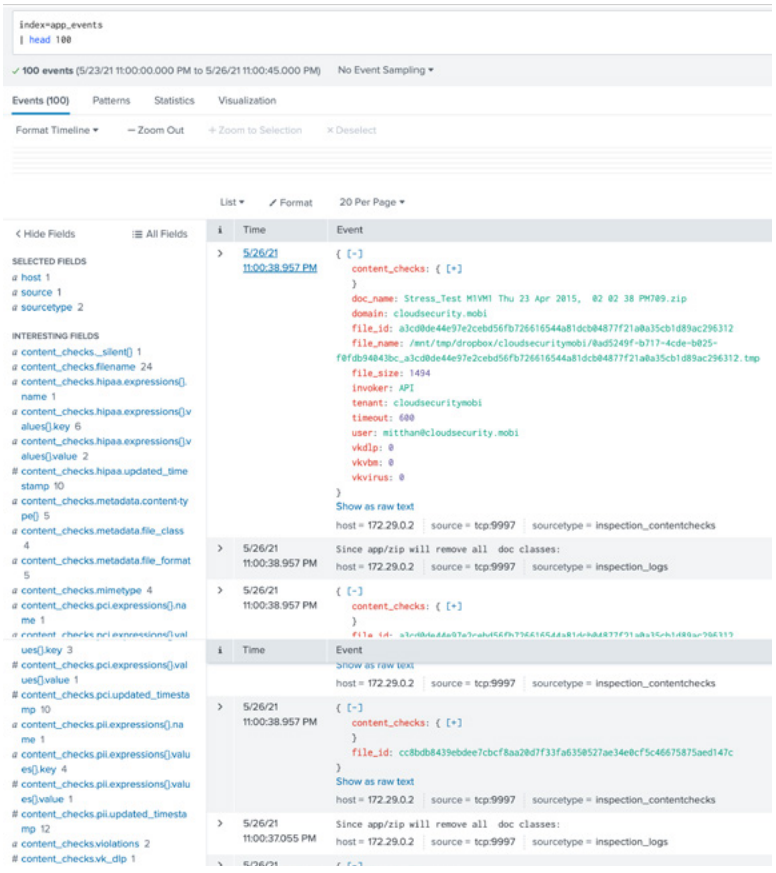


Fig. 5: A Tomcat event in Splunk software, after processing by Cribl Stream.

As you can see in the figure below, the metrics converted by Stream are all visible under the analytics workspace.

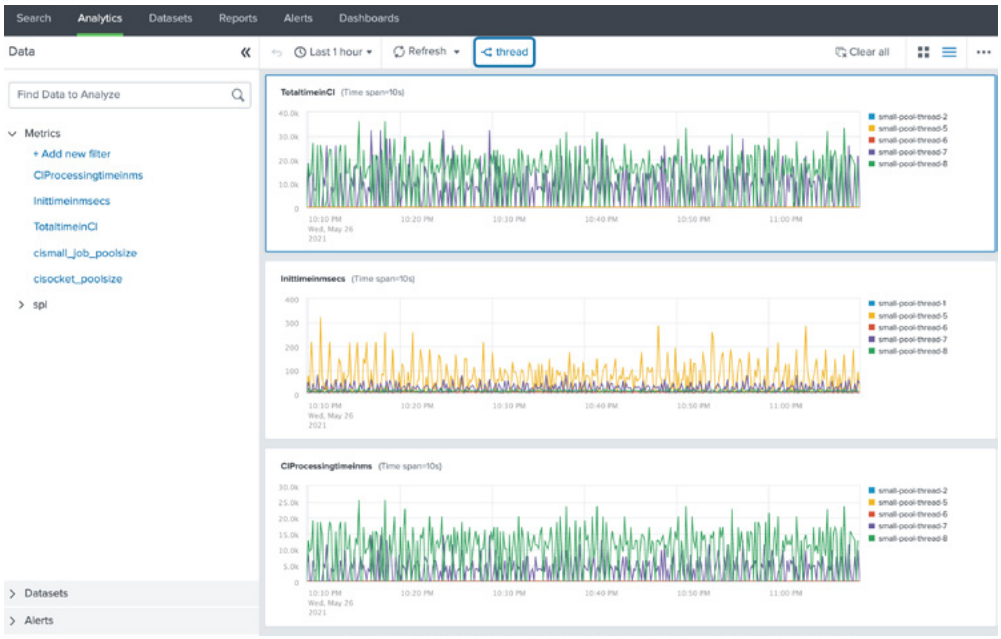


Fig. 6: Stream makes room for additional ingest.

Using the methods above, Stream was able to transform this combined dataset to make room for additional data ingest by nearly 20%.

Fig. 7: More benefits from faster search performance.

However, there are more benefits from faster search performance, which translates to lower CPU usage on Splunk software infrastructure.

There are two search performance comparisons:

1. Search-time field vs. index-time field within event indexes:
• |stats count command on the raw events in index=main over 24,48, and 72 hours of data • |tstats command on the raw events in index=app_events over 24,48, and 72 hours of data
2. Search-time field in event index vs. data in a metrics index:
• stats-average of a metrics in one of the events in index=main over 24,48, and 72 hours of data • mstats/analytics workspace rendering of the same metric in index=app_metrics over 24,48, and 72 hours of data

USE CASE	SEARCH-TIME SEARCH	OPTIMIZED SEARCH
Search-Time Field vs. Index-Time Field	index = main | rex "\:\d{2}\s+(?<severity>\w+)\s+\[(?<pool thread>[^\]]+)\]" | stats count by pool_thread	|tstate count where indexsapp_events by thread
Search-Time Field vs. Metrics Index	index = main "Init time" | rex "\:\d{2}\s{2}(?<severity>\w+)\s+\[(?<pool thread>[^\]]+)[^\:]+\:\s+(?<init_time_ms) by pool_thread |timechartt span=5m avg<init_time_ms) by pool_thread	|mstats avg ("Inittimeinmsecs") prestats=true WHERE "index"="app_metrics" span=5m BY thread |timechart avg("Inittimeinmsecs") span=5m useother=false BY thread WHERE max in topl0 I\ fields - _span*

Fig. 8: Searches used.

Because Stream can convert events and logs to metrics, transform and optimize data, and searches with the tstats command, you end up with better search performance and lower Splunk software CPU usage.

Because Cribl Stream can convert events and logs to metrics, transform and optimize data, and search with the tstats command, you end up with better search performance and lower Splunk software CPU usage.

TIME WAITED	TRADITIONAL SEARCH METHOD	RECORDS SEARCHED	SEARCH TIME (SECS)	OPTIMIZED METHOD SEARCH	RECORDS SEARCHED	SEARCH TIME (SECS)	STREAM PROCESSING SEARCH SPEED BENEFIT
4 Hours	Stats-count on search time field in index = main	288,346	1.804	Stats-count on search time field in index = apps_events	112,180	0.034	32x
16 Hours		1,154,664	3.738		448,721	0.047	80x
24 Hours		1,727,868	5.383		699,884	0.082	65x
48 Hours		3,154,729	9.179		1,170,021	0.10.127	92x
72 Hours		4,398,299	9.179		1,710,399	0.127	103x
4 Hours	Stats-count on search time field	46,599	0.41	Querying Metrics index (mstats)	46,529	0.075	5.5x
16 Hours		188,023	1.124		187,829	0.179	6.3x
24 Hours		280,580	1.602		280,762	0.2	8.1x
48 Hours		486,535	1.602		486,590	0.311	8.3x
72 Hours		710,048	3.869		710,090	0.426	9.1x

Fig. 9: Test results.

Conclusion

While the performance difference is expected to be even better in production environments, there is no debate: You can optimize your data with Cribl solutions and improve your search performance, even when you process petabytes of data daily.

- Cribl Stream enhances how data is sent to Splunk software, transforming and enriching it along the way. This means Splunk software only deals with the data it needs, leading to better search performance and less strain on system resources.
- Cribl Edge takes this optimization further, processing data right at its source. This cuts down on unnecessary data transmission and processing, lightening the load on Splunk software and improving overall efficiency.
- Cribl Search boosts Splunk software's search capabilities, making data retrieval quicker and more accurate. This reduces the time and computational power needed for searches, streamlining operations.
- Cribl.Cloud anchors the entire suite in a cloud-based environment, streamlining the orchestration and scalability of data operations. This cloud platform empowers organizations to effortlessly manage their data infrastructure, ensuring that performance and cost-efficiency are optimized across Splunk software ecosystems.

Regardless of destination, transforming the data first with Stream helps reduce infrastructure costs and storage costs, enabling you to do more with your Splunk software license.

By integrating these tools, organizations can effectively address the key issues of scalability, performance, and cost associated with big data environments in Splunk software. The Cribl suite offers a way to navigate the complexities of data analytics today, enabling deeper insights and greater efficiency.

Looking ahead, the need for adaptable, efficient data management tools like the Cribl suite is clear. In an era of ever-growing data volumes, optimizing data pipelines will be key to leveraging the full power of platforms like Splunk software. With Cribl, businesses are well-equipped to meet these challenges, driving innovation and extracting maximum value from their data!

If you want to get started improving Splunk software performance with sample data, try our hosted sandbox, absolutely free.

ABOUT CRIBL

Cribl, the Data Engine for IT and Security, empowers organizations to transform their data strategy. Customers use Cribl's vendor-agnostic solutions to analyze, collect, process, and route all IT and security data from any source or in any destination, delivering the choice, control, and flexibility required to adapt to their ever-changing needs. Cribl's product suite, which is used by Fortune 1000 companies globally, is purpose-built for IT and Security, including Cribl Stream, the industry's leading observability pipeline, Cribl Edge, an intelligent vendor-neutral agent, Cribl Search, the industry's first search-in-place solution, and Cribl Lake, a turnkey data lake. Founded in 2018, Cribl is a remote-first workforce with an office in San Francisco, CA.

Learn more: www.cribl.io | Try now: Cribl sandboxes | Join us: Slack community | Follow us: LinkedIn and Twitter

WP-0002-EN-3-1224