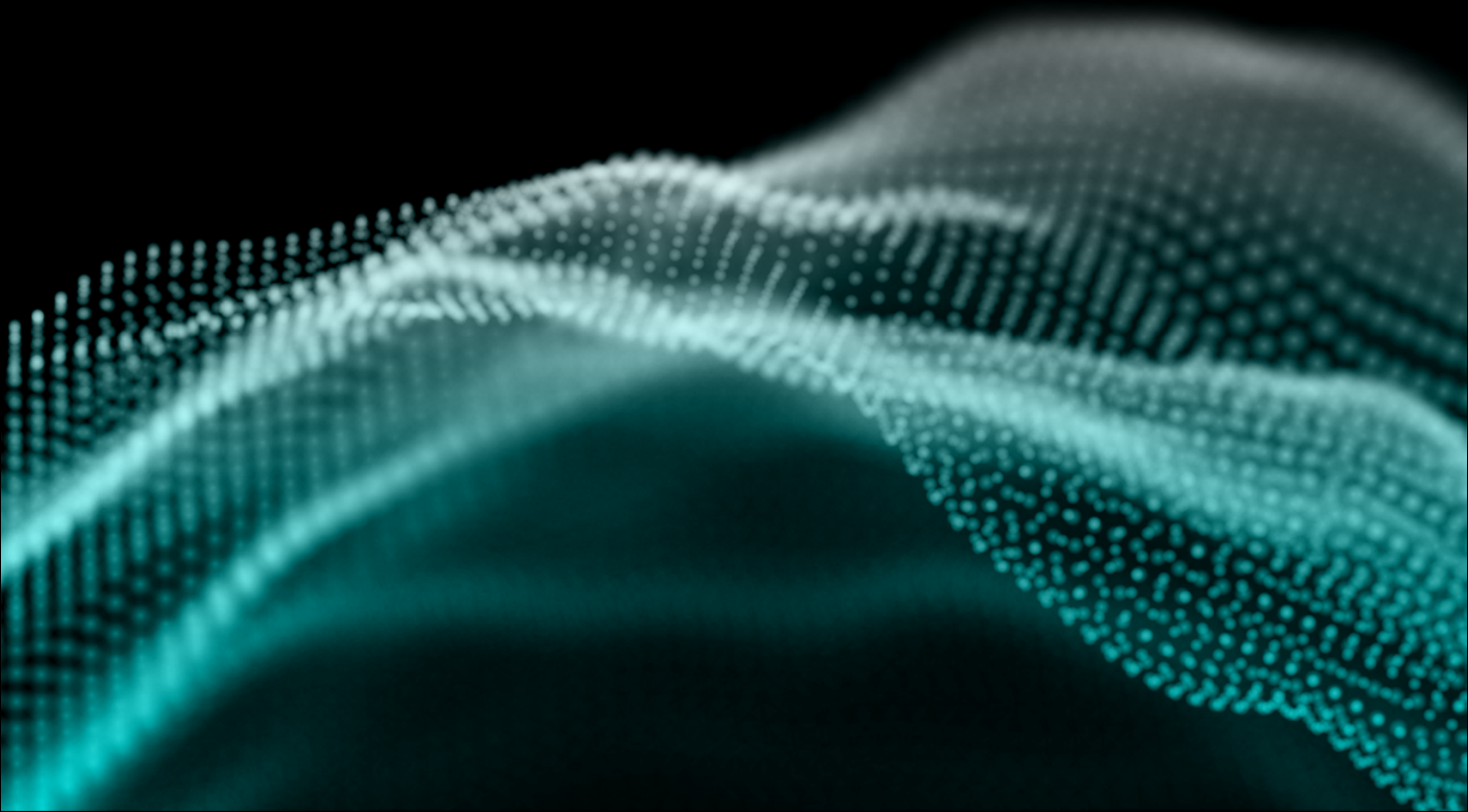
UNDERSTANDING
OBSERVABILITY
PIPELINES 03

Cribl

WHITE PAPER

Migrating Log Analytics to the Cloud

BY STEVE LITRAS, LEAD EVANGELIST, CRIBL



WHITE PAPER

Migrating Log Analytics to the Cloud

by Steve Litras, Lead Evangelist, Cribl

YOU WOULDN'T CONNECT A THIRD PARTY DIRECTLY INTO YOUR NETWORK WITHOUT SOME KIND OF FIREWALL, SO WHY WOULD YOU DO SO WITH YOUR DATA?

Log Management vendors over the last several years have created compelling Software as a Service (SaaS) offerings which are driving a big shift from on-premises installs to SaaS. SaaS shifts the burden of running the core engines of these systems to the vendors. This is not a silver bullet for all use cases, and care still needs to be taken to ensure data is readily ingested and available in an economically viable way. Additionally, SaaS introduces new challenges: costs such as ingestion charges and cloud service egress charges, limitations on flexibility, and vendor lock-in.

This document offers options to meet those challenges. Many IT shops look to compartmentalize change during transitions like this, to minimize risk. It's our contention at Cribl that moving to SaaS without putting protective measures in place adds to your risk. **You wouldn't connect a third party directly into your network without some kind of firewall, so why would you do so with your data?**

The Challenges

COSTS

SaaS log systems take infrastructure responsibility and cost off the table, but the potential cost of sending "low-value data" to them escalates quickly. Costs include:

- *Ingestion costs (cloud logging providers' fees are structured on the amount of data ingested per day).*
- *Increased network bandwidth requirements.*
- *Possible network-data egress fees.*

Add these up, and it's critical that enterprises get the most value possible out of every byte they send to log system(s).

The landing zone for your data moves to the vendor's cloud environment, which means your data has to leave your environment. If you are a cloud shop, that means egress charges. For example, in an AWS shop that's sending 5TB/day to your log analytics system, egress charges will add up to about $132,000/year, according to the AWS Pricing Calculator. That's just the cost of moving the data, not including compute or storage. Of course, most enterprises will reduce that cost through discounts, etc., but it remains a cost of moving log analytics to a SaaS environment.

FLEXIBILITY LIMITATIONS

When you pass data to a SaaS environment, you sacrifice control of that data. Redirecting Elastic Beats Agents or Splunk Universal Forwarders to feed a cloud service is simple, but it complicates doing anything with that data – like enrichment, shaping, or routing to other tools. For example, if the same data that's going to the cloud log provider also needs to be sent to your Snowflake-based data lake, you need a way to facilitate that delivery.

VENDOR LOCK-IN

One benefit of moving to SaaS services is that the enterprise no longer needs to maintain deep expertise in day-to-day running of the migrated system. This, in theory, helps the enterprise from lock-in to that systems vendor. But if that system becomes a black box that controls your data, another form of lock-in comes to light: dependence on the vendor for both data operations and stewardship of your data.

The Solution: An Observability Pipeline

INTRODUCING LOGSTREAM™ WILL RESULT IN A SINGLE, UNIFIED PLATFORM FOR ALL DATA TASKS: ROUTING, TRANSFORMING, ENRICHMENT, AND SECURITY.

We at Cribl believe that the best way to mitigate these challenges is through the implementation of what we call an **observability pipeline**. Consolidation of routing, transformation, and enrichment into a unified observability pipeline gives the enterprise a significant capability that can be the difference between great analysis experience and escalating costs.

Additionally, the pipeline simplifies configuration, deployment, and management of data operations. Moreover, the instantiation of the pipeline concept during the migration gives the enterprise the control it needs over its data, at the time that control is most needed. The pipeline can be built with open-source tooling, such as Fluentd + Apache NiFi (see our blog post about building this type of architecture for details). But we offer a product, Cribl LogStream, that provides this capability out of the box, and for pennies on the dollar compared to your log analytics spend.

We provide a unified control plane for log data operations, a simple way to prototype your rules and test them against real data before putting them in production, and an extremely scalable and performant log control and routing capability. Cribl LogStream is purpose-built for handling logs and metrics. Introducing LogStream will result in a single, unified platform for all data tasks: routing, transforming, enrichment, and security.

MANAGING COSTS

Logging systems have organically become both analysis systems and data retention systems, to the detriment of both purposes. We hear from many of our customers that most of the data in their logging systems is never analyzed or acted on, and is only there to satisfy their retention requirements.

LOGSTREAM ALLOWS YOU TO RETAIN FULL-FIDELITY DATA IN A LOWER-COST CLOUD STORE, BUT STILL BE ABLE TO ANALYZE IT AT WILL.

At the same time, emergent cloud services, such as AWS's S3 and Glacier and Azure's Blob Store, create opportunities to retain data for cents on the dollar compared to traditional storage or datastore technologies. By implementing a separate retention store, and offloading the "firehose" of data there before sending the data to the log system, you can send only the data you need for analysis.

Cribl LogStream 2.2 brings a new data collection capability which allows “replaying” of raw logs from a datastore, such as S3, through your observability pipeline. This allows you to retain full-fidelity data in a lower-cost cloud store, but to still be able to analyze it at will – even long after the real-time events occur. This is particularly interesting for security breaches, which are often discovered months after the fact. For more info, see our blog post on this approach.

CREATING FLEXIBILITY

Log data holds a lot of power, but it is rarely pretty – it’s largely unstructured or semistructured, and is not very valuable without context or cleaning. For example, a switch reporting in its log that a port is flapping doesn’t really help anyone – not unless they can also understand what server is connected to that port, what application is running on that server, and what business process that application supports.

If that contextual data is available alongside the log entry, it helps an observer decide how much priority to put on the error. Moreover, that same set of “switch port flapping” log entries uses a lot of space to identify a simple problem: The single log entry is not relevant on its own, but a collection of, say, 50 of them over a short period indicates a real problem – at the cost of ingesting and storage 50 entries.

Reducing that collection to a single item that summarizes the problem is far more efficient. Reshaping and enriching data in the pipeline increases the value of the data being sent to the analytics system, while potentially reducing the cost of sending it.

FIGHTING VENDOR LOCK-IN

YOU NEED A PLAN TO MITIGATE CHALLENGES LIKE NEW COSTS, LIMITATIONS ON FLEXIBILITY AND VENDOR LOCK-IN.

Instead of the control that an enterprise had over its on-prem data, a SaaS environment reduces that control to whatever APIs the log analytics provider chooses to expose. Introducing an observability pipeline reasserts control over data, reducing the end analytics system to a replaceable analysis system.

Additionally, with the trend of moving traditional ops teams to devops roles, tooling is becoming more distributed: While you might have bought Splunk for the enterprise, a couple of teams may want to use Elasticsearch or Grafana as their analysis tool. Without a control point in front of the log analytics systems, supporting multiple tools gets complex and difficult. Add an observability pipeline, and it gets almost trivial.

Putting It All Together

As we've discussed, there are many good reasons to move log analytics to the cloud, but you need a plan to mitigate some of the accompanying challenges including new costs, limitations on flexibility and vendor lock-in. Implementing an observability pipeline in front of your SaaS log analytics environment is such a plan. An observability pipeline allows you to use your analysis tools more effectively, shape and enrich your data to provide analysts with the context they need to derive insights, and most importantly, get more value out of your data. Cribl LogStream provides all of this capability "out of the box", and gives our customers the superpowers they need to thrive in a cloud environment.

We have experience helping customers dramatically reduce the amount of data they send to their analytics systems - a critical capability to control costs that come with logging in the cloud. One of our customers, a leading financial services organization, uses LogStream to enrich their logs with third party data, enabling them to identify and drop less interesting logs, only sending high value data to their log analytics system. "This is a must-have tool to complement massive seas of data."

We've also helped several customers parse and restructure their existing data so that it can be routed to multiple destinations, including not only log analytics systems, but also to schema-ful destinations like metric stores or analytics databases. If you are migrating log analytics to the cloud, this is critical for avoiding vendor lock-in and restoring the flexibility that you often sacrifice with a move to the cloud. A global, cloud-based financial firm uses LogStream to manage and route their observability data to multiple tools. "We've been looking at a lot of things. While you can never expect to find a silver bullet, this is pretty close."

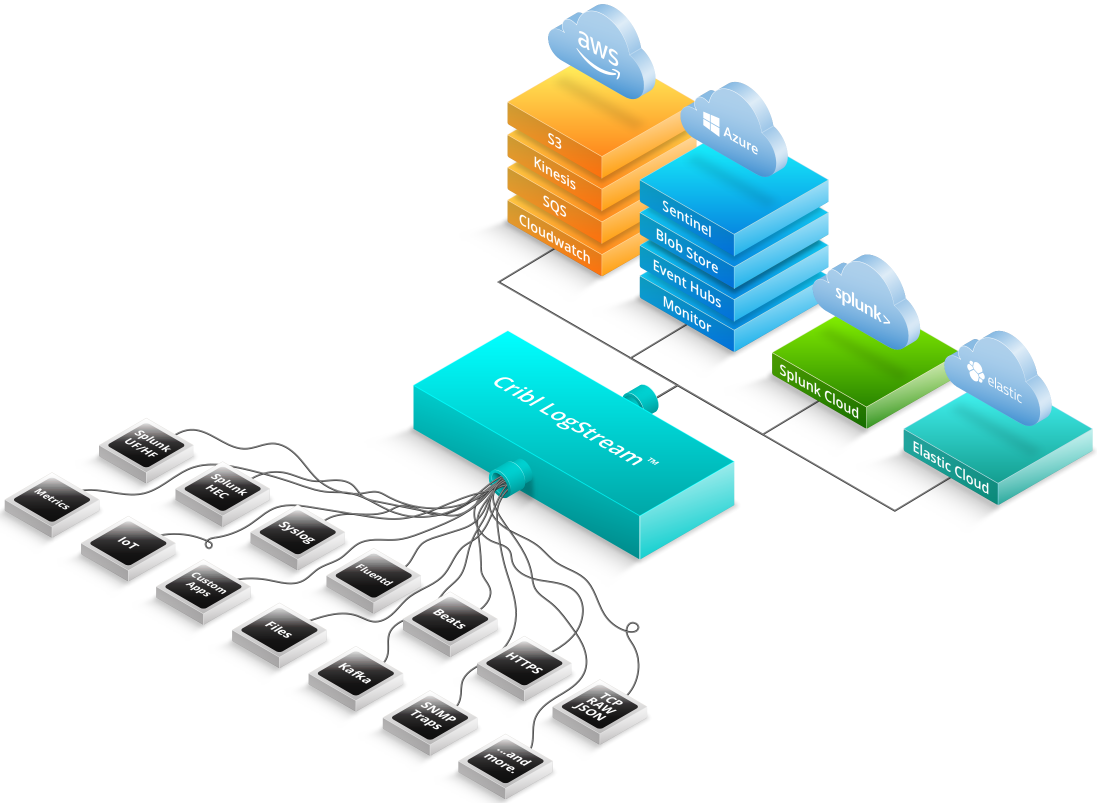
Don't take our word for it, though — try it for yourself. Learn about the capabilities of LogStream via the LogStream Fundamentals sandbox course, and then download the product to try in your own environment.

ABOUT CRIBL

Cribl is a company built to solve customer data challenges and enable customer choice. Our solutions deliver innovative and customizable controls to route security and machine data where it has the most value. We call this an observability pipeline, and it helps slash costs, improve performance, and get the right data, to the right destinations, in the right formats, at the right time. Join the dozens of early adopters, including market leaders such as TransUnion and Autodesk, to take control and shape your data. Founded in 2017, Cribl is headquartered in San Francisco, CA. For more information, visit www.cribl.io or our LinkedIn, Twitter, or Slack community.

APPENDIX

The Cribl LogStream™ Observability Pipeline and the Cloud



CRIBL LOGSTREAM OBSERVABILITY PIPELINE

- *Single product capable of delivering to both Splunk and Elastic, and other destinations on-demand.*
- *Allows you to maintain your existing agent footprint.*
- *Native support for all Splunk, Elastic and Syslog agents.*
- *With a single Observability Pipeline, you can simplify your configuration and remove Splunk Heavy Forwarders, Logstash Systems, and SyslogNG Servers. This not only minimizes infrastructure, but simplifies the environment, and reduces management and troubleshooting burdens on your team.*
- *With the introduction of data collection in LogStream 2.2, you can ingest data from other repositories (like a retention S3 bucket, or your existing log management tool) at will.*
- *Wide range of protocol support for many sources and destinations.*
- *Easier to configure, work with, monitor, and maintain.*
- *Built-in troubleshooting features such as introspection and data capturing*

GREAT PERFORMANCE & SCALABILITY

- *Cribl LogStream can be up to 7x more efficient than Logstash and Fluentd; this will shrink infrastructure footprints, yielding cost savings*
- *Scalable – LogStream has been deployed in multi-TB/day environments and tested at tens of PBs/day.*

ENTERPRISE-READY

Simple, Unified Management Interface

- *Manage a widely distributed deployment from a single master instance.*
- *Prototype pipelines in the UI before deploying – minimizing production configuration errors.*

Resiliency

- *Git-based configuration, allowing simple reloads of prior configurations.*
- *Built-in monitoring dashboards to track status and data flow across sources, destinations, and pipelines.*
- *Disk-backed queueing for Elastic and Splunk destinations, minimizing data loss during destination service outages.*

Supporting Your Compliance Regimen, Out of the Box

- *Change tracking via Git configuration.*
- *Support of LDAP and SAML for authentication.*
- *Ability to archive all data to "retention storage," meeting retention requirements without burdening log analytics tools.*