riskfocus
INSIGHT

Splunk vs. ELK
The Risk Focus Way

riskfocus

Splunk vs. ELK – The Risk Focus Way

Table of Contents

Why aggregate logs?

Over the past decades, most enterprises and organizations have focused on automating their business operations. Manual business functions have been transformed into computational workloads. In parallel, we have seen technology move towards distributed systems, micro-services, and cloud deployments. These two trends have combined to exponentially grow the amount of log data and metrics at organizations. Not only do existing computational applications produce more log data when composed of micro-services or distributed architectures, organizations are constantly adding to compute workloads by migrating paper and/or other processes to machines. Furthermore, as applications become more distributed, the task of correlating application logging across components becomes more complex, requiring ever more sophisticated analysis to identify the state of these applications. Organizations have also realized this data allows them to optimize resource use and provide operational insights.

Organizations that wish to improve their increasingly complex and interdependent compute environments require mechanisms to manage, catalog, search and report on log data. Such data is produced by an increasingly diverse set of components and applications, often including different generations of technology that do not share common frameworks for logging or monitoring.

Prior generations of monitoring tools imposed tight specifications for log data to simplify analysis. This worked well for organizations where compute and data environments were homogenous. Log analysis is straightforward for applications built on one vendor's technology or from a single source type such as applications written in C on a commercial Unix variant writing to syslog. There are very few organizations that fit this profile today. Corporate datacenters in 2019 are populated with a diversity of compute and network hardware running an inter-connected matrix of open-source, vendor provided, and custom software. Retrofitting all these applications, operating systems and hardware components to a single monitoring standard is a Sisyphean task. This challenge is made more complex in organizations that are consuming and synthesizing log-data from public cloud environments with their own rich log ecology.

Instead of demanding a common format, the current generation of log-aggregation tools address the challenge of monitoring and managing diverse environments more forgivingly. These systems allow for data sources that support different protocols and have inconsistent structure. By making minimal assumptions about data format, log aggregation systems support a broader set of use cases. These tools apply a search engine approach to the problem: they consume data

in whatever form it is found and analyze it to create a meaningful result. Modern log-aggregation platforms deliver a common feature set:

- *Ingestion/Log-Aggregation:* Collect log or other machine data from disparate sources, accepting data in a wide range formats/protocol
- *Indexing:* Organize the data to enable fast search and analysis
- *Search:* Identify specific events/patterns in historic or current data
- *Visualization and Analysis:* Analyze data in multiple ways to gain insight, identify events, and measure KPIs.

Within the log aggregation and analysis space, two tools have captured the bulk of market share, Splunk and ELK/EFK (Elastic, Logstash/Fluentd, Kibana). Splunk is a commercial product, while ELK is an open-source project with enterprise support offered by Elastic. In this white paper, we compare Splunk and ELK across several different criteria to help CIOs and technical evaluators make a more informed decision between the two tools and eco-systems.

Executive Summary

Though Splunk and ELK offer similar features, choosing the ideal solution for a specific enterprise depends heavily on the use-case being contemplated and the organization's capacity to integrate and build on top of technology platforms.

Splunk offers robust commercial level support and a broader set of plugins/apps than ELK. These offerings can lower the effort required to integrate Splunk into an enterprise and use it across a broad set of use-cases. For organizations with narrower requirements, ELK offers a compelling solution, with support for numerous data sources with standard plugins and apps. ELK's open source nature allows teams capable of developing plugin code to build an excellent customized solution on the ELK platform.

The commercial decision between Splunk and ELK is not purely a choice between commercial software and open source. Total comparative cost depends on both the organization and the use cases to be supported. Most commercial enterprises will want to secure vendor support for ELK installations to get enterprise features and support. Acquiring this support lowers the relative difference in license acquisition cost between Splunk and ELK. CIOs conducting TCO analysis must carefully consider the amount of development that will be required to make effective use

of ELK in their environment. Training and maintenance costs will depend on the number and technical proficiency of users.

Infrastructure costs for large-scale Splunk and ELK clusters will depend on the mix of activity performed on the cluster and the requirements for search responsiveness and real-time data.

When we last did a side by side comparison of Splunk and ELK for enterprises several years ago, it was clear that Splunk was the more usable option for enterprise scale problems. Since that time, the Elastic offering has narrowed the gap significantly, making it worthwhile to revisit the comparison.

For most organizations with no installed base for either product, we expect the adoption cycle for ELK to require more effort than that for Splunk, but we can no longer say this is true for all organizations and all use-cases. Today, an organization's decision between the two is more likely to be guided by cultural fit rather than by feature, performance or cost difference. For example, both products have a rich ecology to manage large-scale, multi-site clusters. Elastic expects you to configure and maintain such an environment via APIs or configuration files/parameters. Splunk offers the additional option of a web-based configuration console.

Feature Set Comparison

Feature	Splunk	ELK
Schema on the fly	Yes	No
Search time field extraction	Yes	No
Index time field extraction	Yes	Yes
Open source	No	Yes
Machine learning App?	Yes (Requires Splunk License)	Yes
Search Languages	SPL	Lucene
Cluster health monitor	Yes (DMC)	Yes (X-pack, paid plugin)

Feature	Splunk	ELK
User Authentication	LDAP, SAML, AD require Splunk license	LDAP, SAML, AD supported with X-Pack (paid plugin)
App/Plugin Deployment	Admin console or bundles delivered by deployment server	Service console or bundles delivered to
Cluster Configuration Management	Config files, REST API and web console	Config files, REST API
High Availability	Yes, via cluster replication factors	Yes, via replica shards
Cluster replication for redundancy	Yes, with multi-site architecture	Yes, with leader/followers for each index
Cluster rescaling	Master/peer configuration on indexers, options to control bucket fixing during rescaling/restarts	Quorum based reconfig, settings to limit/control shard rebalancing during cluster rescaling
Object Stores (S3, etc.) for index storage	Smartstore offers ability to store index data on S3 for retrieval during search	Only as backup/restore
Backup/Restore	Requires backing up, restoring the Splunk file system	Incremental index snapshots can be stored on S3, HFS, etc. for restore

Both products can scale to handle very large amounts of data. Clusters holding peta-bytes of data are not uncommon. As the size of your cluster increases, so does complexity around administering and managing it. Tasks like re-indexing, deployment, configuration may require other tools to keep them manageable. Clients may wish to utilize configuration management frameworks (such as SALT or Ansible) to manage some aspects of cluster configuration. With larger clusters, tasks such as upgrades and migrations also become more complex and can vary from release to release.

Ingestion/Integration Add-Ons/Adapters

For organizations of any appreciable size, the largest challenge of adopting a new technology is integrating it into existing IT infrastructure and applications. In the log aggregation space, integration largely refers to data ingestion, though integration with an enterprise-wide

configuration management system can also be a significant effort. A large enterprise with a mix of installed technologies will require support for a variety of ingestion methods.

Both Splunk and ELK offer a variety of adapters and add-ons for users to ingest or stream data from a diverse set of sources. This ranges from the classic use case of reading from a log-file concurrently written to by a process to add-ons that enable streaming data from modern cloud logging services such as AWS CloudWatch or Google Monitoring/Stackdriver.

Ingestion Method	Splunk	ELK
Tailing files	Core	File plugin
SNMP	Core	snmptrap plugin
REST calls	Core	http and http_poller plugin
Syslog	Core	syslog plugin
Windows Event Manager	Windows Add-on	Wingbeat
JMS, JDBC, log4j, RabbitMQ, E-mail/IMAP, Unix sockets, TCP/UDP sockets, Chat/Messaging: Twitter, IRC, Jabber	Predefined source types and apps/add-ons available	Plugins available
Apache Kafka	Splunk plugin	Logstash integration
Active Directory	Splunk on Windows only	Winlogbeats to ingest events
AWS: S3 buckets	AWS Add-On	logstash-input-s3
AWS: CloudWatch	AWS Add-On	CloudWatch Plugin
AWS: SQS	AWS Add-On	SQS plugin
AWS: Kinesis	AWS Add-On	Kinesis plugin
AWS: CloudFront Access Logs, VPC Flow Logs, Billing Reports	AWS Add-On	Kinesis plugin
GCP: Pub/Sub, Monitoring/Stackdriver, Billing	Google Cloud Add-On	Logstash-input GCP Pub/Sub

Azure : Office 365 Management Metrics, Audit, VM Metrics	Microsoft Cloud Services Add-On	Azure Event Hubs Plugin

Applications/Pre-built reports

The difference between Splunk and Elastic at a usage level is neatly illustrated with the case of public cloud billing analyses. For AWS and Google Compute, Splunk has pre-built dashboards for Billing reports in its Add-Ons. That makes it simple to get visibility into billing within a few minutes. Elastic does have pre-built plugins to ingest the same AWS reports. However, you will either have to map the schemas and build dashboards or rely on a community plugin that is a few years old.

The Splunk App Store has hundreds of free and paid apps that can be used to accelerate ingestion or analysis of data. Splunk's app selection covers most enterprise software such as databases (SQL Server, Oracle, etc.), messaging solutions (TIBCO), servers, network devices (Palo Alto, Cisco), security applications. Splunk also has a certification process for trusted publishers to make applications available on the app store.

ELK has fewer pre-built apps (plugins). Many valuable plugins suffer from the open-source curse of abandonment when an author or team moves on to something else. ELK also lacks a comprehensive certification process for plugins. Organizations adopting ELK and expecting to ingest a wide array of sources should budget for additional DIY development to extend it across the enterprise.

Of particular interest to security practitioners are the SIEM capabilities of log aggregation platforms. Both Splunk and ELK offer toolkits that utilize the analytics capabilities of the platforms to enable Security Incident and Event Management. As with most of the plugins, Splunk Enterprise Security is more refined, delivers several actionable reports immediately and has a more finished feel. ELK's Security Analytics dashboards offer similar capabilities, but generally requires more work to get to the point where an organization has actionable intelligence. For existing ArcSight users, integration between ELK and ArcSight can help shorten the adoption path.

Performance

There are typically two considerations for log analysis performance: ingestion/indexing and search.

We performed the following test on a standalone Splunk server and an ELK instance on similar benchmark hardware with out-of-the-box configuration:

- Ingest 27GB of data, roughly 95M events
- Execute query to return all events

Criteria	Splunk	ELK
Infrastructure	AWS m3.large + EBS	AWS m3.large + EBS
Version	7.1.3	6.3.2
Indexing rate	20,000-25,000 EPS	800-100 EPS
Query Execution Time	800-1000 sec	8-10 sec

Splunk exhibits significantly faster ingestion times out of the box, accepting 100x more events than ELK does. ELK is significantly faster at returning complete search results once ingestion and indexing are complete.

This is largely a result of the way that Splunk/ELK tackle ingestion, with the Splunk approach being weighted towards optimizing ingestion and continuous indexing over time employing the "schema on the fly" model. ELK requires a schema, key-value pair be available for data at the time of ingestion, and indexes all data as it comes in. This then results in much quicker search performance across large data sets. This Elastic blog post contains an overview of why their product opts for a “schema on write” approach.

Of course, this comparison is very general, and relied on a single node cluster. Results can vary significantly depending on architecture. Both platforms have rich recommendations to improve performance:

- Elastic recommends several configuration changes to tune performance if the use case involves infrastructure logs with very small sized events. Users should carefully consider these for their use-case. In addition, Enterprises committing to Elastic should consider using Kafka or an alternative message broker to handle spikes in log traffic.
- Splunk overcomes the relatively slow search performance by providing results as they're collated and recommending that users concerned about search performance focus on narrower windows of time. For its SIEM product, Splunk relies heavily on Data Model Acceleration to speed up results.

Actual performance in your environment can be tuned by controlling I/O performance, network performance, source system configuration, aggregating forwarders/beats, concurrent user activity, compute resources available to the log-aggregation cluster, index configuration, ingestion configuration, etc.

Which solution is more suitable depends on your specific application and whether you are optimizing for ingestion (i.e. large, variable ingestion flows) or real-time and speedy search. However, the distinction of schema-on-the-fly vs. Pre-defined key-value pairs has several implications that may not be immediately apparent. In particular, consider the fact that re-indexing may be required in Elastic if a metric is not initially defined, resulting in a significant computational cost.

Maintainability and Scaling

Both Splunk and ELK are used in large environments with several terabytes of data ingested and searched each day. When dealing with very large clusters, organizational challenges are more likely to revolve around maintenance tasks associated with dozens or even hundreds of nodes.

Deployment automation and DevOps practices are an important area for organizations to consider. and we explore them in some detail for Splunk in our Splunk Automation post. Many of the same considerations apply to ELK clusters. These issues are much less challenging for clients using AWS' managed ElasticSearch offering.

Data Archiving

For most organizations, storage will be the largest cost for a sizable log aggregation cluster. Managing the costs of this expensive resource requires some sort of data archiving strategy. Both ELK and Splunk offer usable approaches to delete data or archive older data to a less costly, performant storage medium. The exact process used to perform this task will impact performance.

Note: The most recent Splunk versions offer object-based storage for indexed data in a feature they call SmartStore. This is designed to leverage services like AWS S3 or Google Cloud Storage. In AWS, for example, ingested data is indexed into files of a specific size, and when complete, they are stored as immutable objects in S3. These objects are retrieved as required to meet search demand. This approach can significantly reduce the cost of storage for a large Splunk environment. Elastic does have an object storage offering in beta, but it is initially planned to support the snapshot/restore functions only, not as a full-fledged store for indexed data.

Cloud Offering

Both Splunk and ELK offer AMIs with software pre-installed which organizations can use to run POCs or quickly create testbeds.

Splunk also offers a managed/SaaS AWS hosted solution. In general, the Splunk Cloud offering is suitable for small single node installations with a narrow set of input formats. Use cases that rely on a large or growing set of plugins will find themselves relying on a cumbersome Splunk support workflow to install these plugins. This coupled with the inability to host Enterprise Security on Splunk Cloud makes the solution unusable for large enterprises.

Elasticsearch offers a hosted solution, Elastic Cloud. In addition, AWS offers a managed Elasticsearch service. AWS's managed service is highly scalable and integrated with other AWS services. It allows fully automated deployments and a simple to use management framework. Multi-AZ deployments with replica shards for High-Availability are similarly available out of the box.

Support

For customers implementing mission-critical IT monitoring tools, appropriate support level agreements (SLAs) and enterprise-level engineering processes are mandatory.

Splunk is a fully integrated indexing and analytics package with enterprise-level support from both Splunk, Inc. and the sizable Splunk developer community. Splunk currently supports thousands of installations worldwide at some of the largest enterprises in the world. Annual support costs are around 20% of a perpetual license (all licenses are annual subscription starting in 2020).

Elastic also offers paid support with defined SLAs. These services are available at an additional cost, moving ELK into a "freemium" pricing model.

"What if something goes really wrong?"

Both Splunk and Elastic offer support contracts with access to experienced personnel capable of troubleshooting problems with clients.

Ease of Use and Training

Both Splunk and ELK are modern tools with intuitive interfaces. Most technically adept users should be able to set up a single node Splunk/ELK installation, ingest data and begin rudimentary searches within an hour or two.

Though these search functions can be very valuable (allowing users to search across large numbers of logs quickly), they rarely allow the identification, monitoring and resolution of underlying issues. To do that, users will need to utilize the analytical capabilities of Splunk/ELK to generate ongoing, actionable intelligence into their operations.

Enabling large numbers of users (typically hundreds within an IT organization) to begin creating customized dashboards and reports requires some sort of formal training program. Having

worked with several clients across log aggregation platforms of various sizes, we generally see users begin to do meaningful analysis using Splunk/ELK within the first month of use.

Splunk offers a rich education program, a Professional Services group and an expansive network of skilled consulting partners. Getting a team trained and certified by Splunk to the “power user” level can be accomplished within a month.

Some subset of users will require in-depth and advanced training on the log-aggregator chosen. This is required for those who are:

- working with very large datasets
- developing organizational standards/best practices for data ingestion
- administrating the log aggregation clusters
- serving as subject matter experts for search optimization or dashboarding

The learning curve to achieve these higher capabilities gets progressively steeper, such that Architect-level training can take significantly longer than a month. Time-to-value can be reduced by hiring a Splunk partner firm to roll out capabilities quickly and build advanced correlation apps.

Elastic's education offering is not nearly as extensive as Splunk’s for users. In keeping with its open source roots, Elastic relies on approachable documentation and short training videos for user education. Elastic does, however, have a handful of courses and two certification tarcks aimed at engineers. These are sufficient for almost all use cases. Organizations looking to staff an Elastic team will find it easier to recruit from a talent pool of younger engineers who've had some interaction with the tool via course-work or open-source projects.

Total Cost of Ownership/Use

As is appropriate, most enterprises will evaluate the total cost of ownership (TCO) when deciding on a log aggregation platform. TCO models estimate the true lifecycle cost of adopting and using the technology to meet the expected business need. We will consider how adopting this approach informs some common misconceptions about cost and log-aggregation platforms.

Licensing models

Perhaps the most-common sentiment expressed to us about the comparison between Splunk and ELK is that "Splunk is expensive". Like much intuited sentiment, the truth is more complex. Though Splunk license costs are material for most organizations, they form a single component of the complete TCO. The additional cost of licensing Splunk has to be weighed against the cost of infrastructure, support, maintenance and training for ELK.

Elastic licensing for the base product is free, though enterprise support and the security features in X-Pack are not. Splunk licensing is based on the amount of data indexed per day. This license is rigorously enforced: exceeding license usage more than five days in a month triggers license events that can result in the product being disabled until additional license capacity is acquired, or an exception acquired from Splunk support is applied. In contrast, Elastic prices for support and premium plugins are based on the number of nodes running ELK. If licenses expire, the product feature set reverts to those available in the open source version, and advanced security and monitoring features are disabled.

Infrastructure Costs

A major consideration when running very large clusters is the cost of compute resources required to ingest and search data. To understand the implications on hardware usage for the two products, we take a look at a hypothetical example of a 1TB/day data ingestion workload. If using Splunk, organizations should budget for the equivalent of 12 cpu cores and 12GB of RAM for this workload. ELK users should expect to budget for 24 cores and 24GB of RAM, and expect to expend effort tuning for ingestion to yield this throughput. When search workloads are added, Splunk's ingestion capacity will drop significantly. If the 12 core/12 GB ram cluster described above also experiences some interactive use, Splunk predicts an ingestion throughput of 400GB/day. With very heavy real-time search/dashboard updates this throughput drops even further. For example, Enterprise Security implementation ingestion workloads should be throttled to 100-150GB/day per node.

By contrast, ELK does not experience such a drop off in ingestion throughput when search workloads are added. For moderate search workloads, the performance of ELK and Splunk will be similar on similar hardware, as shown below

Low search volume use case	1TB/Day	10 TB/Day
Splunk	1x (12 cores + 12GB)	10x (12 cores + 12GB)
Elastic	2x (12 cores + 12GB)	20x (12 cores + 12GB)

Security, real-time search	1TB/Day	10 TB/Day
Splunk	8x (12 cores + 12GB)	80x (12 cores + 12GB)
ELK	8x (12 cores + 12GB)	80x (12 cores + 12GB)

Over time, the largest infrastructure cost for log-aggregation platforms is likely to be persistent storage. Both Splunk and Elastic compress log data to some degree and organizations should carefully consider the cost/benefits of compressing/archiving certain data or simply not indexing it.

Maintenance/Management costs

Enterprises should also consider the aggregated on-going costs of deploying, managing and maintaining a cluster to meet the organization's log aggregation and analysis needs. For larger clusters with dozens of hosts, due consideration needs to be given to cluster design and scalability patterns. Without robust DevOps practices around deployment and configuration of new nodes, organizations can easily find themselves facing a cascading set of failures involving configuration drift, index redistribution and performance degradation. Our blog post on DevOps practices for Splunk charts out a path to address these concerns.

Adoption, training and cultural fit

Compared to Splunk, adopting ELK can be more time-consuming and costlier for traditional IT organizations that are adept at open-source technology or have rigid development practices.

Conversely, small organizations with basic needs, or agile groups more comfortable with open-source products will find ELK to be a comparatively good investment. For IT development organizations seeking to build search capabilities in their own applications, ELK is a natural choice from a licensing and integration standpoint.

For large firms such as financial institutions, Splunk, with its wide support for different technologies and extensive set of plugins/apps can easily turn out to be cheaper in the long run despite its high license costs. Such organizations will find that using ELK effectively within their

environment involves building infrastructure around support or knowledge transfer and developing or modifying connectors to underlying log sources. Such organizations can address this requirement by using ELK on public cloud, effectively outsourcing infrastructure management.

Costs have to be considered against benefits. Organizations with a complex ecology of compute workloads will find that the benefits of using a modern log-aggregation tool significantly outweigh the costs associated with deploying and maintaining it. There is simply no other way to adequately manage and monitor operations across a heterogenous compute environment.

Conclusion

ELK and Splunk offer increasingly equivalent capabilities with implementation differences that can impact fitness for a specific use case. For many usage scenarios, the relative cost of implementing Splunk or ELK into a large enterprise is comparable.

The choice between these two solutions will often depend on the kind of environment they are being deployed into and specifics of the use case. Choosing the right log aggregation and analytics platform for your organization involves considering a variety of factors we've touched upon in this post and is a decision that requires some foresight. The Risk Focus data infrastructure team can help you make the right decision.