Logging Modernization

Cloudera's log analytics solution significantly lowers cost and scales up log processing, delivering real-time insights.

What is Logging Modernization?

Logging Modernization is a holistic approach towards unlocking the value of machine generated data by lowering processing costs and enabling a whole range of new analytics use cases. This is achieved through:

- Edge processing for cost-effective data movement.
- Intelligent, content-based routing, transformation and enrichment.
- Sending data to alternative systems based on value, content, and use case.

Unlock the Value of Machine Data

Organizations across the globe, irrespective of size and market, continuously improve their businesses and operations using machines. These range from power generators to mobile apps and legacy mainframes with every remote sensor, email, video, and phone call in between. Millions of events are logged, but only when they are aggregated and interpreted successfully do they mitigate operational risks and improve business outcomes.

The value in log files is derived not by stockpiling huge volumes and wide varieties of machine generated data, but by collecting, curating, and analyzing the data so that your businesses and operational teams are able to gain and leverage actionable intelligence.

The digital world has reshaped market dynamics through rising customer expectations, tighter profit margins, and increased competition. Logging is critical to maintaining and leveraging your business and operational infrastructure, but Logging Modernization could be the difference between new revenue streams and delayed market response.

Logging Modernization is a holistic approach that unlocks the value of machine generated data by using a **comprehensive streaming platform**. This platform should include everything from real-time data ingestion,edge processing, transformation, and routing through to descriptive, prescriptive and predictive analytics. All of which should be securely shared across on-premises, public, or hybrid cloud environments.

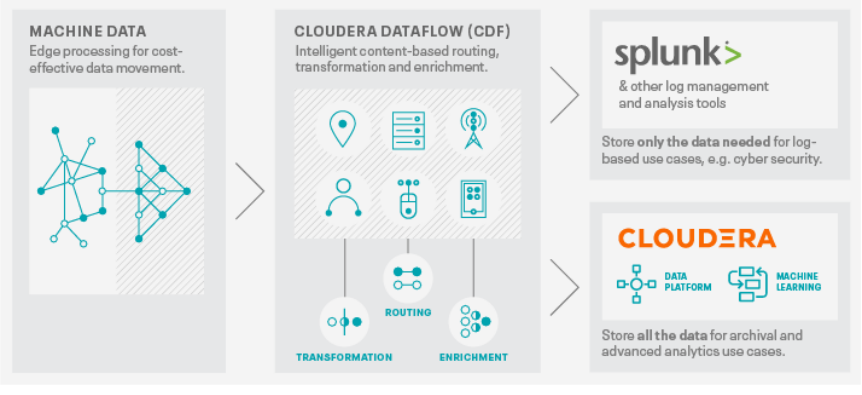
Challenges: Costly and Complex Integration

According to Gartner Research, Splunk is one of the leaders in the log heavy disciplines of IT Operations Management (ITOM) and Security Information and Event Management (SIEM). Unfortunately, limited data filtering and routing capabilities at the data source and downstream often lead to excessive collection of data, of which 60% is left unused.

In addition to the cost of indexing data that will not be used, there is an opportunity cost. Offloading data to non-Splunk systems is challenging, which limits the ability to integrate log data with other enterprise data sources to glean insights and drive better business outcomes.

Logging Modernization Lowers Cost and Unlocks Value

The diagram below illustrates how Cloudera log analytics will lower cost and unlock the value of machine generated data. See the next page for how this is done in an iterative and scalable manner for Splunk and other market leading log management and analysis tools.



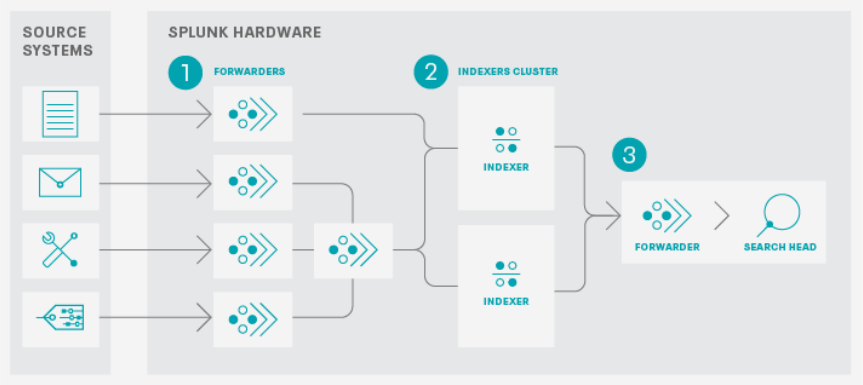
Challenges of Traditional Log Processing

Excessive header information, suboptimal file formats, and poor data quality checks on ingest mean customers store and pay for data that adds little business value.

- **Indiscriminate data ingestion**: Filtering/routing mechanisms are very limited, which leads to indexing and storage of data that may go unused.
- **Cost**: Traditional log analytics solutions charge by the volume of the data that is being indexed. If you index everything, you pay for everything.
- **Complexity**: Configuration and management are complex and offloading data to other critical business and operational systems can be challenging.

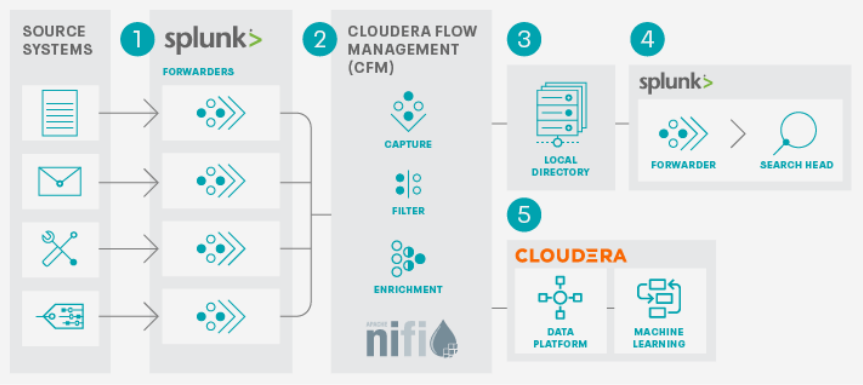
Traditional Architecture

Splunk components and architecture include a series of Forwarders, Indexers, and Search Heads. All data, regardless of utility, is indiscriminately ingested, indexed, and stored in one data silo. This scenario adds complexity and cost to enterprise-wide data integration downstream. Read more on the challenges of traditional log processing in the sidebar.



Stage One: Optimize Indexing and Load Only the Data Needed

Introduce Apache NiFi to capture, filter and enrich logs prior to sending them to Splunk or other downstream apps. This enables significant reduction of noise into downstream apps. Cloudera Data Platform (CDP), an enterprise data platform that supports shared analytics and collaboration across teams can be introduced here to extend the data for new and extended use cases like data mining, machine learning etc. This might seem like a simple approach, but it is highly effective. See the left side bar for an example.



1. Splunk Forwarders continue to ingest data but now push it to NiFi over TCP.
2. NiFi receives, compresses, filters, and transforms the data based on content and/or attributes prior to routing it. This is Cloudera Flow Management (CFM).
3. NiFi writes data to a local directory.
4. Splunk Forwarder continues to send data (that is now filtered) from the local directory to Splunk, storing only the data required.
5. NiFi also sends a larger set of data to CDP for deeper analysis, machine learning, and other use cases.

Stage Two: Enable Edge Processing and Filter Data at the Source

This next iteration replaces the Splunk Forwarders with Apache MiNiFi Agents while the rest of the architecture is the same as above. Replacing the Splunk Forwarders with MiNiFi agents delivers several advantages. Unlike Splunk, MiNiFi can filter at the source so complexity, load,

Cloudera Customer Success Story

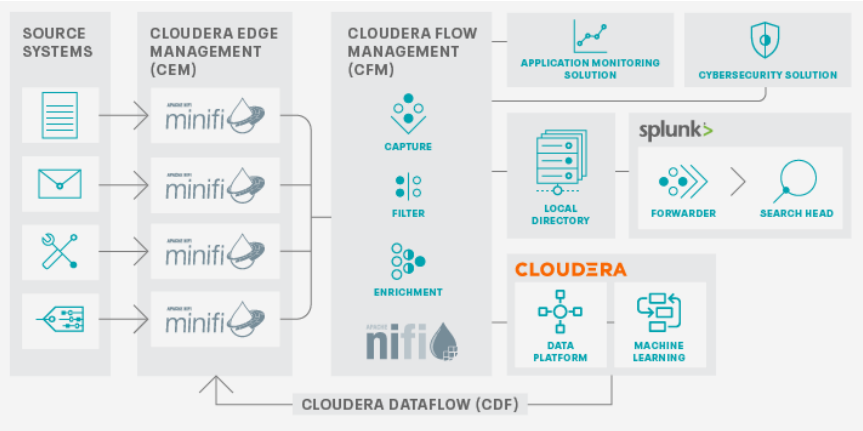
Challenge: A major oil and gas company was ingesting over 800GB of data per day and spending over $2 million over five years in Splunk perpetual license and annual maintenance costs.

Solution: By simply introducing NiFi to compress, filter, transform, and route log data based on the use case (see #2 in Stage One Architecture on the right), this Cloudera customer saw a remarkable decrease in storage needs and cost.

Impact:

- **60% reduction** in Splunk log ingest
- **$1.2M reduction** in Splunk costs over 5 years
- **40% savings** on estimated hardware costs

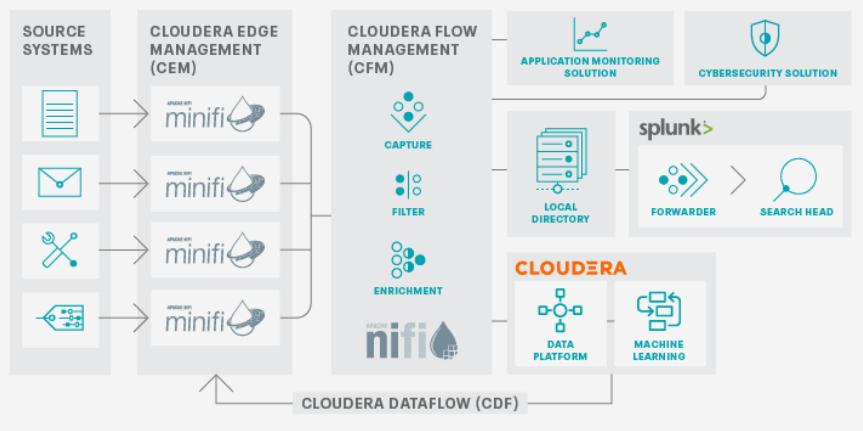
and costs are reduced downstream. While Splunk Forwarders are comprised of manual configuration files, MiNiFi is managed with Cloudera Edge Management (CEM), which provides a centralized graphical enterprise-grade configuration and management solution.



1. Splunk Forwarders are replaced with MiNiFi which collects, filters, and processes data at the source.
2. MiNiFi then sends the data to NiFi over a secure protocol (as opposed to TCP in the Stage One iteration) and continues to be managed with CEM.
3. The rest of the architecture is identical to Stage One, above.
4. In this scenario, data collected into CDP can be leveraged for building machine learning models.
5. Machine learning models can then be deployed back to MiNiFi agents at the edge to make the edge smart enough to make decisions.

Stage Three: Enable Enterprise-wide Data Movement

Once the architecture is set up to ingest log data from thousands of endpoints via MiNiFi into NiFi, the architecture can scale up or out. The needs for scaling are largely driven by extended use cases. Given that NiFi can act as a neutral data movement engine between Splunk and CDP, it can also be extended to push the data into other applications such as cybersecurity engines. The value of log data from thousands of machines into a cybersecurity solution can prove to be extremely valuable in providing operational insights about threats, vulnerabilities and identity information. In a traditional architecture, this type of log data is not readily available to a cybersecurity solution. But with NiFi serving up this data to all applications, this data can be made available in near real-time. This can help in alerting the IT Ops teams with critical and crucial information before any real damage is done.



The components of logging modernization

Cloudera DataFlow (CDF) is a scalable, real-time streaming data platform that collects, curates, and analyzes data, providing immediate actionable intelligence. CDF enables organizations to:

- Ingest and process real-time data streaming at high volume and high scale.
- Drive stream processing and analytics on data-in-motion.
- Track data provenance and lineage of streaming data.
- Manage and monitor edge applications and streaming source.

Cloudera Data Platform (CDP) is an enterprise data platform that supports shared analytics and collaboration across teams. Attributes unique to CDP include:

- Hybrid and multi-cloud – provides choice to manage, analyze and experiment with data in the data center and/or in any public or private cloud environments for maximum choice and flexibility.
- Multi-function – solves the most demanding business use cases – applying real-time stream processing, data warehousing, data science and iterative machine learning across shared data at scale.
- Secure and governed – simplifies data privacy and compliance for diverse enterprise data with a common security model and governance to control data on any on-premise, cloud – public, private - or hybrid environments.
- Open – facilitates continuous innovation from the open source community, the choice of open storage and compute architectures, and **the confidence and flexibility of a broad ecosystem.**

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.

Learn more at cloudera.com

Success is More Than Cost Reduction

Although reducing infrastructure costs is extremely important, value and success is also created by enabling a whole range of new analytics use cases

In the architecture scenarios above, all data is routed and stored in CDP, but can alternatively be sent to other destinations such as AWS, Azure or Google Cloud Platform alongside other enterprise data for holistic contextual analysis.

With so much data-in-motion from source to destination, it's important to point out the streaming analytics capabilities within Cloudera Data Flow (CDF) (see side bar). Cloudera Streaming Analytics (CSA), powered by Apache Flink, is a distributed processing engine and scalable data analytics framework that processes millions of data points or complex events very easily and delivers predictive insights in real-time. This makes CDF an extremely compelling platform for processing high-volumes of streaming data at high-scale.

An additional level of advanced data analytics, as noted in the Stage Two architecture, is Cloudera Machine Learning (CML). CML accelerates machine learning from research to production in the following ways:

- Facilitating data science at scale to build, test, iterate, and deploy machine learning models in production.
- Experimenting faster, using R, Python, or Scala with on-demand compute and secure data access.
- Enabling data scientists to push these models out to the edge (MiNiFi and NiFi) in order to continuously monitor digital signatures from connected data sources and drive action in real-time.

Summary

Splunk is a powerful solution for putting machine-generated data to use and is a cornerstone for many companies that use it to both protect, stabilize and analyze what is happening across their IT operations. However, Splunk's indiscriminate ingestion of raw data of any type means that costs and inefficiencies can quickly mount up.

By filtering and routing data based on Cloudera's content inspection and data flow automation, customers of Splunk and other market leading log management and analysis tools will gain the dual benefits of managing expenditure while enhancing their ability to extract value from data.

Cloudera, Inc. 395 Page Mill Road Palo Alto, CA 94306 USA cloudera.com