

Data from Everywhere to Anywhere: Open Architecture

Table of Contents

Executive Summary1

What Works and What Doesn't Work..... 3

Open Standards and Open Architecture 3

Leveraging an Open Architecture and Use Cases..... 6

Use Cases within an Open Architecture 6

Working with Elastic, Splunk, and Hadoop..... 7

Appendix A: Elastic..... 8

Appendix B: Splunk 12

Appendix C: Hadoop..... 14

Summary.....25

Executive Summary

Large data sets have transformed the way enterprises ingest and interpret data. Specialized analytics tools and applications that interpret this data to make meaningful business decisions should utilize open architectures. A partial-definition for technological “openness” can be found on Open Referral’s website: “An open platform enables its data to be both accessed directly by users and also published in open formats.”¹ Investments in open technologies enable businesses to easily move data between applications and infrastructures according to need. Implementation of technologies that support common data models, formats, and standardized protocols are examples of open architecture in a corporate environment. Open platforms provide flexibility across toolsets ensuring the best one for the job is used. Increased productivity and maximized business value is achieved from large data sets. Closed systems make data sharing difficult and lead to data silos. Some vendors intentionally design products to be incompatible with open standards and technologies. This “vendor lock-in” strategy leverages a customized tool to make it difficult and expensive to switch to a competing product. This strategy has resulted in businesses retaining the status quo in the past, at times to the detriment of their own growth, performance, or stability.

In many implementations of data platforms, use cases are driving the need to have more than one database and in some instances more than one Big Data platform. The belief that one type of platform can be used to do every possible use case is misleading and incorrect. As data and technology move forward to help us make intelligent decisions, different formats, systems, and technologies must be used along the way. This is especially true for the modern security operations center that have multiple data repositories, systems, and tools.

Open architecture specifications are public through officially² approved standards as well as privately designed architectures whose specifications are made public. This visibility enables flexible integration with other technologies, allowing data to be transferred between front-end applications and analysis infrastructures easily. Businesses initially see a return on investment with lower cost when implementing open architectures and later realize additional value when integrating new systems or migrating to alternate infrastructures.

The second problem is that technology choices need to augment personnel and processes. Organizations that buy security products without investing the time to train analysts or configure and tune products find security noisy and full of false positives. Some misconfigured devices might not be sending any valuable data at all. Without proper installation, configuration, and turning some systems generate so many alerts it is impossible for an analyst to make sense of the data. Raw data, while useful for some situations, isn’t particularly useful for security use cases. Understanding the context, situation and circumstances is critical. Additionally, enrichment is also key—filling in missing data, making sure data is correct and adding critical information is important. A good example is around Microsoft Windows logs—while they tend to be verbose in detail, they can (and do) miss critical pieces of information such as an IP address or hostname on certain logs. This adds complexity and additional

In many implementations of data platforms, use cases are driving the need to have more than one database and in some instances more than one Big Data platform. The belief that one type of platform can be used to do every possible use case is misleading and incorrect.

-
1. openreferral.org/faqs/what-do-we-mean-by-open-platforms-whats-an-api/
 2. openreferral.org/documentation/

processing to understand their meaning. By having a sophisticated log collection process that understands the context and enriches the log data, large data sets are made easier to understand and process. Technology choices must be made to make the security team's performance easier and more effective. Before purchasing an expensive commercial solution, a proof of concept using open source solutions tied by automation might be deployed to show the value of use cases under consideration.

The third problem, alert fatigue, is a source of error and talent attrition. When an analyst is dedicated to pure alert triage, the work can be mind-numbing and dull valuable skills. Many talented analysts will pursue other roles where investment in security has been properly planned and prioritized by leadership. Another consequence of alert fatigue is human error. It's easy for someone to miss one step in a triage process that's done tens or hundreds of times each day. Security teams who hire the best people and make them do alert triage all day are not configuring their environment for sustainable management. Automation must reduce the manual work involved in triage if possible, or offload alert triage to parts of the business that do not burden upper tier security analysts in their investigative roles. The most effective teams involve their security analysts to provide ideas to improve their accuracy and efficiency, and empower them to make changes in the workflow, processes, and tools.

The last problem is that alerts without context are not actionable. Security monitoring is often built on alerts that are configured so only fragments of the necessary data is captured, along with a vague signature for context. These alerts may be categorized as a false positive when a minority of them are true positives that were not confirmable with the available evidence. Using a network security monitoring approach that involves full packet capture, host forensics, endpoint logs, app server logs, network analytics data, entity behavior and other context allows a SOC to monitor the security of the enterprise more effectively.

Getting all this data into a SIEM is difficult and sometimes impractical to do in a scalable way, yet remains a primary goal in securing the enterprise. Storing data and logs from multiple disparate systems provides an audit trail that can be used to understand the activity of the system, transactions, and interactions between users and systems. Under the combination of data from these different systems, statistical analysis may be performed and yield correlations between seemingly unrelated events on different systems. New insights obtained from security data provides business value beyond security.

Open architecture maximizes usage of the infrastructure and enables leveraging data over numerous applications. Connecting to data lakes for analytics, or other applications for uncovering innovative business insights, is designed into the open architecture. Moving data from anywhere to anywhere, securely, is the first design decision. Data on security posture and events may then be leveraged across the enterprise flexibly, integrating with solutions suitable to a particular environment or need.

Open architecture maximizes usage of the infrastructure and enables leveraging data over numerous applications. Connecting to data lakes for analytics, or other applications for uncovering innovative business insights, is designed into the open architecture.

What Works and What Doesn't Work

More data is not always better. A common trap in building security architectures is to design the system to collect all possible data in a semi-organized manner to sift through for clues after a breach has occurred. Two issues with this approach are a return to alert fatigue through irrelevant data, and conversely a lack of context built in to the collection process that makes the collected data useless for reconstructing the breach. Successful enterprises consider the value of their data streams when determining what to collect. Refining data streams provides an additional return on investment when considering long-term storage costs. The enrichment of non-security data with security context paves the way for faster and more accurate threat detection, allows the application of security expertise to raw data, and open architecture allows the collection of data from anywhere to be applied to various security use cases.

More data does not mean better security. Without understanding the meaning, situation, and circumstances of the logs, finding an attacker will be impossible. From the swamp of alerts to the lack of context, more data will require more people to process and react. Some alerts that fall outside of the normal playbook will need to be investigated. These could be suspicious alerts, and still need to be investigated. One example, when the username Administrator is used to install a patch on a Windows Server, it could be a legitimate log or alert, but it could be an attacker. Either way, it must be investigated. So, more data means more of these situations. While technology can be employed to do analytics, a person is still required to add intelligence to answer this question. And that means more people, more investigation, and usually less security because enough cannot be done.

Structured data provides context in an open architecture, allowing attribution of actions to be stored in crucial fields that are then available to integrated toolsets. The storage of structured security data allows complex alerts based on historical behaviors and attributions. By structuring its data, the enterprise prepares it for later data mining and analysis. Structured data is organized and can be made readily searchable by search engine algorithms or other search operations. Unstructured data makes compilation a time and energy-consuming task.

Structured data provides context in an open architecture, allowing attribution of actions to be stored in crucial fields that are then available to integrated toolsets. The storage of structured security data allows complex alerts based on historical behaviors and attributions.

Open Standards and Open Architecture

Key pieces of enterprise architecture in disparate implementations solve for the same business needs: utilizing extremely large data sets. Technologies like Apache's Kafka,³ Spark,⁴ and Hadoop⁵ are open, utilizing standards for flexibility in handling data and communicating with other platforms. These technologies are widely adopted in part because their open, standards-based approach facilitates quick adoption into existing infrastructures. This technology stack provides a stable framework through which to handle the near-limitless data sets created by the modern enterprise. The open standards free security providers to focus on solving the Sisyphean challenge of keeping the enterprise safe.

3. kafka.apache.org/
 4. spark.apache.org/
 5. hadoop.apache.org/

Security data is rightfully categorized as sensitive to the enterprise. The grouping of sensitive data together makes the data store itself at the highest level of risk and becomes a focus for attack and reconnaissance. This has traditionally led security data to be purposefully kept in a silo, doled out in stingy sums on an as-needed basis. Security logs could take time to produce when requested, especially when outsourced. In today's fast-moving enterprise business decisions will be made without good security data, if that data can't be made available quickly.

At the point of collection data must be structured and of value. Once logs, sensors, stream network traffic, security devices, web servers, custom applications, cloud services, and other security data is recorded it becomes immutable. Immutable data is verifiably unchanged, a necessary attribute for reconstruction of breaches from databases. Valuable business information can also be mined from massive security data sets, such as personnel movements from active directory in traditional IT, security tools like endpoint analytics, cloud IT such as Salesforce, IoT/OT like badge readers, or mobile like iOS and Android.

The ArcSight Security Open Data Platform (SODP) by OpenText™ architecture leverages these open standards to provide security data across the enterprise, lowering barriers to address business and security needs across teams, projects, and business units.

The ArcSight Security Open Data Platform (SODP) by OpenText™ architecture leverages these open standards to provide security data across the enterprise, lowering barriers to address business and security needs across teams, projects, and business units.

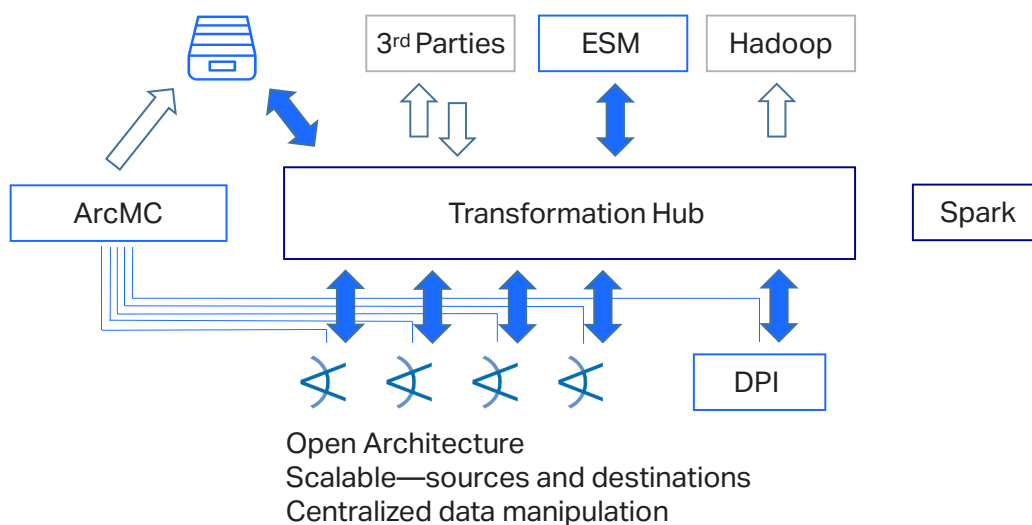


Figure 1. ArcSight Security Open Data Platform (SODP) features

ArcSight Security Open Data Platform (SODP) delivers the industry's first open security architecture that seamlessly connects to third-party platforms, including Hadoop.

ArcSight SODP components include:

Transformation Hub: The Kafka-based Transformation Hub (TH) allows for the consumption of up to one million events per second. The TH is a component of ArcSight SODP that provides a message bus for the scalable distribution of data between multiple destinations. It has an enterprise capability to receive and distribute data, such as log data, across multiple systems and technology with ease, scalability and resiliency.

ArcSight Management Center: ArcSight Management Center (ArcMC) provides one centralized view for end-to-end monitoring and simplified processing of bulk operations.

ArcSight Logger: ArcSight Logger by OpenText is a log management solution that is optimized for high event throughput, efficient long-term storage, and rapid data analysis.

SmartConnectors: More than 480 pre-built connectors to easily extend data collection sources without manual customization. Data is enriched with context and meaning making it complete and relevant.

Load Balancer: SmartConnector Load Balancer provides a “connector-smart” load balancing mechanism by monitoring the status and load of SmartConnectors.

ArcSight SODP transforms the data collection process, simplifying administrative tasks and improving the effectiveness of monitoring with the ability to collect from a variety of data sources with high velocity and high volume of data ingestion. Kafka is an open source project from Apache designed to handle thousands of clients at hundreds of megabytes per second. It distributes data at scale within the clustered and high availability (HA) environments enterprises need so was selected as the underlying technology powering the TH.

The ArcSight SODP TH centralizes event processing, makes scaling your ArcSight by OpenText environment even easier, and opens up ArcSight security data to additional applications and third-party solutions. It enables enterprises to take advantage of scalable, high throughput, multi-broker clusters for publishing and subscribing to event data.

ArcSight SODP transforms the data collection process, simplifying administrative tasks and improving the effectiveness of monitoring with the ability to collect from a variety of data sources with high velocity and high volume of data ingestion.

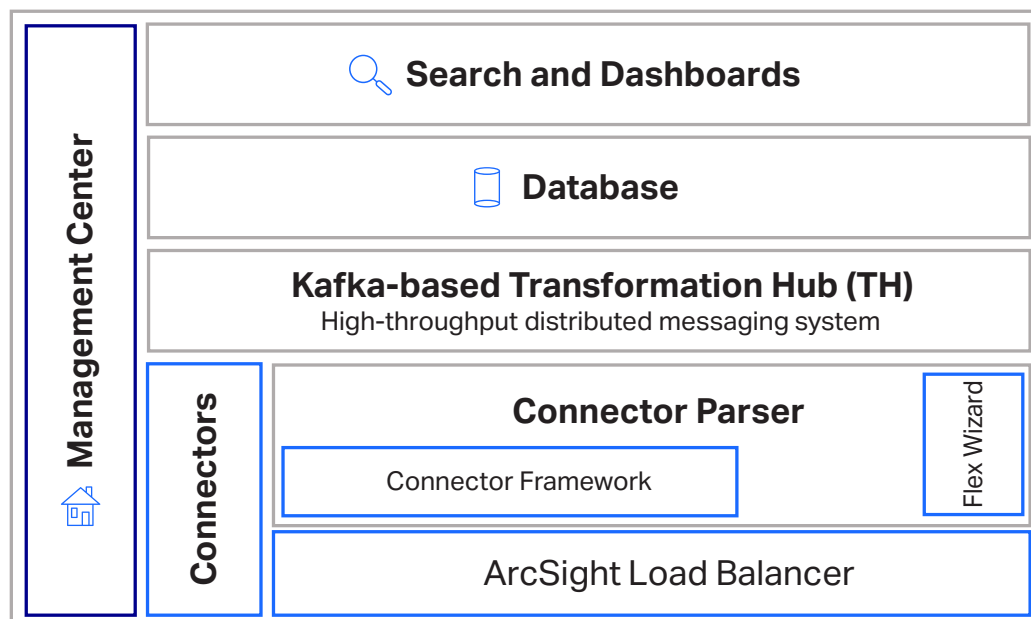


Figure 2. ArcSight Security Open Data Platform (SODP)

The ArcSight SODP TH includes a pre-packaged version of Apache Kafka. After installation and configuration of an TH, or cluster of brokers, ArcSight SODP SmartConnectors begin publishing security data. ArcSight SODP Logger and Apache Hadoop or other consumers can subscribe to that security data.

Leveraging an Open Architecture and Use Cases

By leveraging an open architecture, ArcSight SODP is on the forefront of transforming data consumption and distribution. Its unique capability to scale, and powerful log and event distribution system for the enterprise makes it the strategic security platform of the future. Data sent to and ArcSight SODP, Hadoop, and an enterprise's internal security data warehouse, may be filtered, have controls added, or provide in-stream processing. ArcSight SODP produces an enterprise-class secure data environment with flexible formats and transformations for a number of use cases.

By leveraging an open architecture, ArcSight SODP is on the forefront of transforming data consumption and distribution. Its unique capability to scale, and powerful log and event distribution system for the enterprise makes it the strategic security platform of the future.

Use Cases within an Open Architecture

HIPAA Compliance and Patient Privacy

Urgent care clinics, labs, medical offices, and health care networks are particularly hard to protect because of diverse systems and different equipment connected within the organization. These networks are faced with the daunting task of adding new Figure 2. ArcSight Security Open Data Platform (SODP) Data from Everywhere to Anywhere: Open Architecture 8 technology to their networks in order to remain competitive while keeping sensitive data flowing through those networks secure. PCI DSS and HIPAA are the foundation of security in these networks, and require a wide variety of mitigating controls to protect cardholder data and patient health information. ArcSight SODP leverages data over third-party applications and provides visibility into HIPAA concerns through compliance packs, delivering a broad spectrum of coverage and enabling quick time to response when privacy compromise is detected.

Malicious Insiders and Vulnerable Endpoints

Oil and gas explorers and gas/seal pipe manufacturers deal with important telemetry data, often in harsh environments. ArcSight by OpenText's comprehensive suite of products work to protect this data in a distributed environment. Different models meet the needs of specific data, whether it be in a remote site or placed in a data center. ArcSight SODP's open architecture sets up a security operations center with a mature infrastructure to allow a wider range of applications of data from any source. ArcSight SODP's can log this data and ArcSight ESM by OpenText produces security analytics to identify the bad guys.

Payment Card Industry Data Security Standard (PCI DSS)

Small to medium businesses in the retail industry have a complex time managing devices on the network that are disparate across cities or states. Simplifying the ability to audit and manage PCI DSS compliance is a must, given the rules and regulatory penalties for non-compliance. ArcSight SODP's syslog event management and reporting on compliance

issues protects credit card information and POS systems. ArcSight SODP with the ArcSight Compliance Insight Package for IT Governance helps organizations perform the following automatically: review event and log data on an ongoing basis, store logs online and in long-term archives, meet forensically sound practices, and prepare and demonstrate compliance for auditors. ArcSight Security Open Data Platform is an all-in-one log collection, storage, analysis architecture for cost-effective automation of log management, and can connect to existing data repositories and data lakes.

Security and Exchange Commission (SEC) and Federal Trade Commission (FTC)

Credit Unions, pay day loan services, and small banks all need to meet regulatory compliance for SEC and FTC. ArcSight and Data Security products offer platforms to control and manage user and application access in addition to threats. With the control to lock down environments and the logging capability of SODP, large volumes of historical data are available for analysis, compliance, and assist with internal audits.

Focus Use Case: Malicious Insiders and Vulnerable Endpoints— Proxy Log Analysis on Prohibited Websites

Tracking resource access to find insider abuse and sometimes fraud is a common security use case. This information is valuable during incident response for determining which resources an attacker has accessed and possibly corrupted or modified. One investigative approach is top internal users blocked by proxy from accessing prohibited sites and known malware sources. This web access data report can be used for multiple purposes from tracking compromised systems, to data leakage tracking, to improved productivity through exclusionary evidence.

ArcSight SODP's open architecture for moving data to and from other platforms and applications for analysis enables leveraging additional tools for a specific investigation. Many SOCs leverage a plethora of tools, including ELK, Splunk, Hadoop, and others. Whether to prove business justification for resource expenditures, provide the final audit trail for post-breach review, refine an incident response plan, or enrich a hunt portfolio, enterprise security data can be accessed by the proper tool through open standards.

The following are steps to use ArcSight SODP and TH with selected other platforms.

Working with Elastic, Splunk, and Hadoop

Elastic

The Elastic platform is used for log management and analysis. In this use case, we setup a single node Elastic 5.0 server to ingest ArcSight SODP events from the TH through Logstash, then use Kibana to create a visualization of the data:

1. Elastic is installed and configured via instructions found on the Elastic website⁶.
2. For applying structure to events, a Common Event Format (CEF) codec plugin is available for download⁷ with rich CEF mapping. Alternatively, a unique CEF filter can be created by hand or through a Professional Services engagement. (An example CEF filter has been provided in Appendix A.)

ArcSight SODP's open architecture for moving data to and from other platforms and applications for analysis enables leveraging additional tools for a specific investigation.

6. elastic.co/guide/index.html
7. elastic.co/guide/en/logstash/current/plugins-codecs-cef.html

3. Install and configure Kibana (see *Appendix A*).
4. Create an initial index making sure ALL fields (CEF or not) are available for searching and aggregation. This is required for searching and visualizations.
5. Now the CEF fields are extracted from the event ingestion in Logstash.
6. Visualizations such as top users with the most denies from a proxy device can be shown.

Splunk

Splunk is a data platform that captures, indexes, searches, and analyzes machine-generated data gathered from devices. With the open architecture of ArcSight SODP, events can flow into a Splunk environment.

1. Splunk is installed and configured via instructions found on the Splunk website⁸.
2. Configure the new data input for Kafka messaging.
3. Provide the ZooKeeper port rather than Kafka port. The ZooKeeper host points to the first node of the TH (see *Appendix B*).
4. Events are seen flowing into the Splunk environment.

Hadoop

Hadoop is a Java-based programming framework designed to support the processing of large data sets in a distributed computing environment. Data growth has been a critical component of many enterprise strategies, and Hadoop has eased modernization costs with great efficiency. ArcSight SODP integrates with Hadoop through the ArcSight SODP Transformation Hub, opening security data to business applications.

Integrating Cloudera Hadoop with ArcSight SODP relies on several services, including HDFS for event storage, HUE for visualization and querying, and FLUME to pull events from ArcSight SODP TH in to HDFS:

1. Cloudera Hadoop is installed and configured via instructions found on the Cloudera website⁹.
2. Add, start, and configure FLUME service (see *Appendix C*).
3. Configure HUE to view events from HDFS (see *Appendix C*).
4. Once the FLUME and HDFS services are started, a directory called “flume” will appear and it will store the CEF files and its contents.

Appendix A: Elastic

Elastic is a suite of applications and tools and a great platform for log management and analysis. It is a strong competitor to Splunk and being open source has a strong community following. It is certainly a strong competitor for ArcSight Logger too. ArcSight Logger has the distinct advantage of being relatively easier to set up and get value from. Elastic on the other hand has incredible flexibility if you have the time to configure it.

Data growth has been a critical component of many enterprise strategies, and Hadoop has eased modernization costs with great efficiency. ArcSight SODP integrates with Hadoop through the ArcSight SODP Transformation Hub, opening security data to business applications.

8. splunk.com/en_us/download.html splunk install instructions
9. cloudera.com/products/apache-hadoop.html

The steps below are for a single node Elastic (5.0) ingesting events via Logstash which is configured with the Kafka plugin for reading EventBroker topics.

Installing and configuring a basic Elastic node is very straight forward. Simply follow the instructions on their website: elastic.co/guide/index.html

The key is pulling events from EventBroker, and massaging them in to something that is structured and easily searchable in Kibana. Below is a CEF filter which gives us all the basic/necessary CEF fields for BlueCoat events in order to do some analysis. Keep in mind though there is a CEF codec plugin available: elastic.co/guide/en/logstash/current/plugins-codecs-cef.html (manual installation) which has a rich CEF mapping already done for you. So, you have a clear choice on how you want to get your CEF events in.

```
Input {
  kafka {
    topics => ["BlueCoatEvents","ProductionEvents"]
    bootstrap_servers => "eventbroker:9092"
  }
}
The above code simply pulls two topics from the EventBroker.
filter {
  # Manipulate the message
  mutate {
    # Saved the original message into a temporary field
    add_field => { "tmp_message" => "%{message}" }
    # splits message on the "|" and has index numbers
    split => ["message", "|"]
    # generate fields for the CEF header
    add_field => { "cef_version" => "%{message[0]}" }
    add_field => { "cef_device_vendor" => "%{message[1]}" }
    add_field => { "cef_device_product" => "%{message[2]}" }
    add_field => { "cef_device_version" => "%{message[3]}" }
    add_field => { "cef_sig_id" => "%{message[4]}" }
    add_field => { "cef_sig_name" => "%{message[5]}" }
    add_field => { "cef_sig_severity" => "%{message[6]}" }
  }
}
```

Figure 3. How you want to get your CEF events in

The above code uses the “mutate” plugin to slice and dice the CEF header in to useful information. After this point, CEF is of course key/value, which is what our “kv” plugin on the following page does.

```

# Parse the message with field=value formats
kv {
  # Note: values with spaces are lost (still getting there)
  field_split => " "
  trimkey => "<>\\[\\], "
  trim => "<>\\[\\], "
  # Only included the fields which are of interest (dont need everything)
  include_keys =>
  ["cat", "act", "proto", "dst", "dpt", "src", "spt", "dhost", "suser", "request", "catdt"]
}

```

Figure 4. “kv” plugin

“kv” is only pulling certain field names here. You could easily extend this array of field names to include literally all CEF fields if you’re interested.

We call mutate again simply to rename the CEF fields in to something a little easier to read and work with in Kibana.

```

mutate {
  # Rename fields to cef_field_names
  rename => [ "cat", "cef_traffic_category"]
  rename => [ "act", "cef_traffic_action"]
  rename => [ "proto", "cef_traffic_proto"]
  rename => [ "dst", "cef_traffic_dst_ip"]
  rename => [ "dpt", "cef_traffic_dst_port"]
  rename => [ "src", "cef_traffic_src_ip"]
  rename => [ "spt", "cef_traffic_src_port"]
  rename => [ "dhost", "cef_traffic_dst_host"]
  rename => [ "suser", "cef_traffic_src_user"]
  rename => [ "request", "cef_traffic_request"]
  rename => [ "catdt", "cef_traffic_category"]
  # Revert original message and remove temporary field replace => { "message" =>
  "${tmp_message}" } remove_field => [ "tmp_message" ]
}

```

Figure 5. Rename the CEF fields

Lastly, we obviously want to push the events in to Elastic. Remember, by default Elastic (and most of the other pieces to it) assume loopback only binding and communication. So, this needs to be fixed this first.

```

output {
  elasticsearch {
    hosts => ["elastic-1:9200"]
  }
}

```

Figure 6. Push the events in to Elastic

Once you have Kibana installed and configured (elastic.co/guide/en/kibana/current/install.html), create an initial index and make sure you include ALL your fields (CEF or not) available for searching and aggregation. This is needed to search and also create visualizations. A great place to quickly learn how to do visualizations is here: digitalocean.com/community/tutorials/howto-use-kibana-dashboards-and-visualizations.

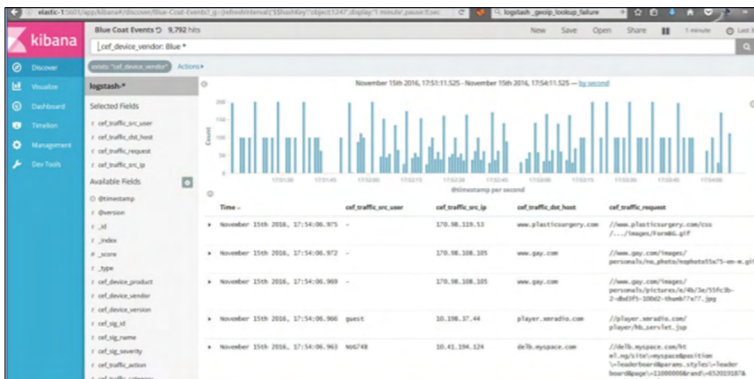


Figure 7. Kibana

You can see all the “CEF” fields we pulled out from the event ingestion in Logstash. Again, this is a manipulation of only a dozen fields. CEF of course is hundreds of fields. The URLs in the above come from demonstration events.

Pulling open each event, you can see all the details we need to analyze and visualize.

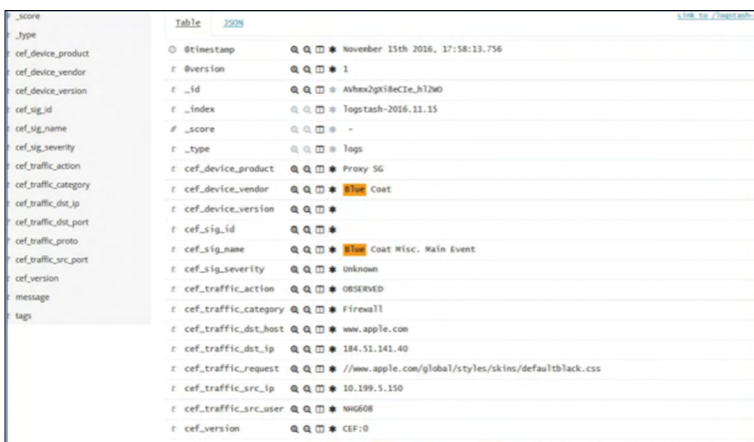


Figure 8. Pulling open each event

From here, you can demonstrate searching and of course, visualization. On the following page is a quick stacked bar graph to show Top destination hosts and their associated top users requesting connection to them (Blue Coat).

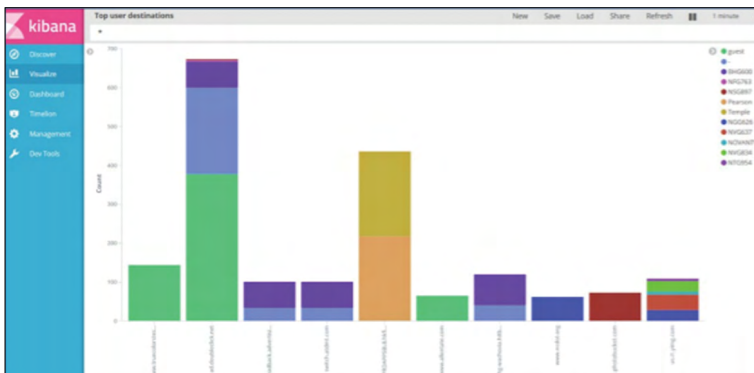


Figure 9. Top destination hosts bar graph

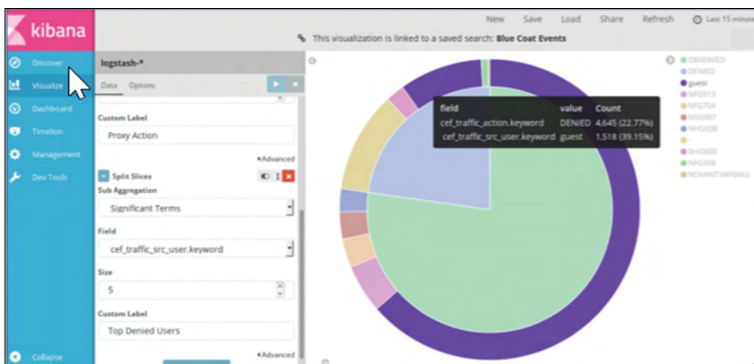


Figure 10. Denies pie chart

The visualization above shows top user names having the most DENIES from a monitored proxy device.

Appendix B: Splunk

Splunk comes with a small license that allows you to get started. If you need something bigger, try the Developer portal dev.splunk.com. You can obtain a 10 GB/day license for 6 months.

Simply install Splunk on Windows or Linux. On Windows it's a single executable. Now to get events in from Transformation Hub (KAFKA), that's arguably as simple.

1. Simply go to Settings --> Data Inputs
2. Choose "KAFKA Messaging"
3. Choose "New"
4. Fill in the appropriate fields. You can use the screen shot on the following page for hints.

Output Settings	
Data Output	STDOUT
Consumer Connection Settings	
Kafka Topic Name	bluecoat
<i>Kafka Topic Name</i>	
Kafka Group ID	bluecoat
<i>Kafka Group ID</i>	
Zookeeper Host	eb-1.aus.hp.com
<i>Zookeeper Host. IP or resolvable hostname.</i>	
Zookeeper Port	2181
<i>Zookeeper Port. Will default to 2181.</i>	
Zookeeper CHROOT path	
<i>Zookeeper CHROOT path ie: foo/bar</i>	
Zookeeper Raw Connection String	

Figure 11. Splunk form

Topic name is of course the Topic. Group ID is your “Consumer Group.” Zoo Keeper host is of course your Transformation Hub node #1. Remember to give it the ZooKeeper port, NOT the Kafka node port (9092) as the Splunk app uses a slightly older implementation of the technology. For Source Type, I used “generic,” as per screen shot on the following page.

For Source Type, I used “generic,” as per screen shot on the following page.

Figure 12. Source Type

From here you will see events start to stream in from Transformation Hub.

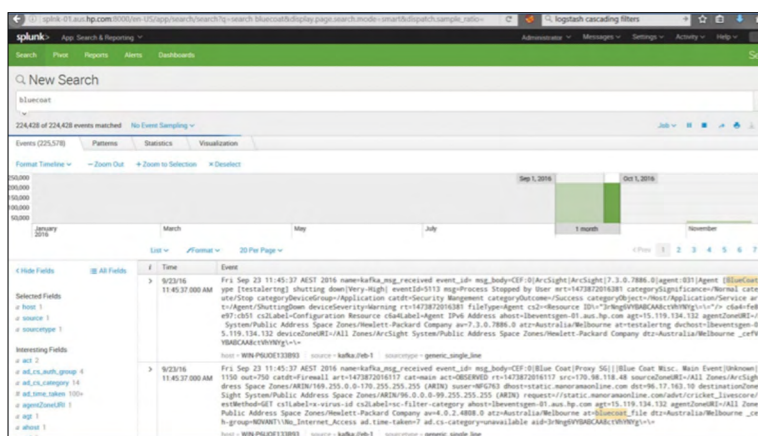


Figure 13. Transformation Hub

Appendix C: Hadoop

Hadoop has a very easy method to demonstrate integration which is covered step by step in the SODP EventBroker administration guide. To look for something more refined and visual, here are some steps to implement the free Cloudera Hadoop virtual machine.

First, download the latest Cloudera Hadoop trial VM from **Cloudera.com**. Install it, make sure you meet the requirements for the “Express” version. It’s more than enough to experiment with and demonstrate. If you choose “Enterprise,” it’s a true trial and also has larger resource requirements like RAM and CPU.

When you’re up and running, double click on Express icon as per screen shot on the following page. It will kick off a heap of script installs and get the VM ready for you.

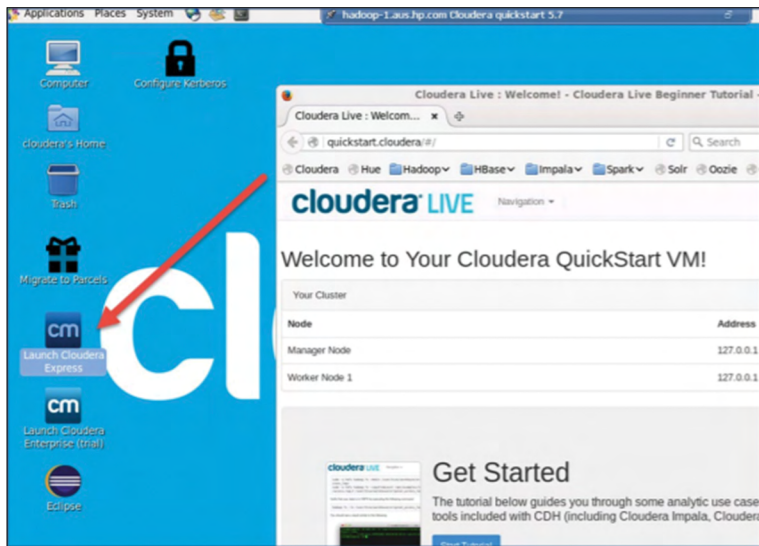


Figure 14. Cloudera Express

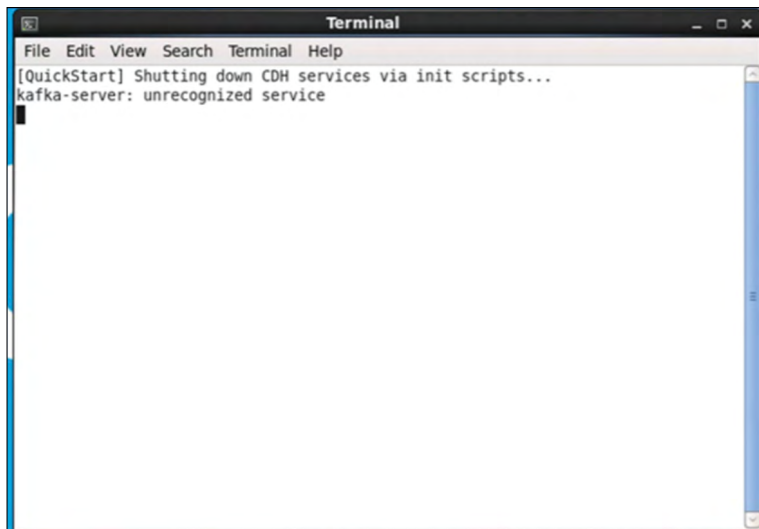
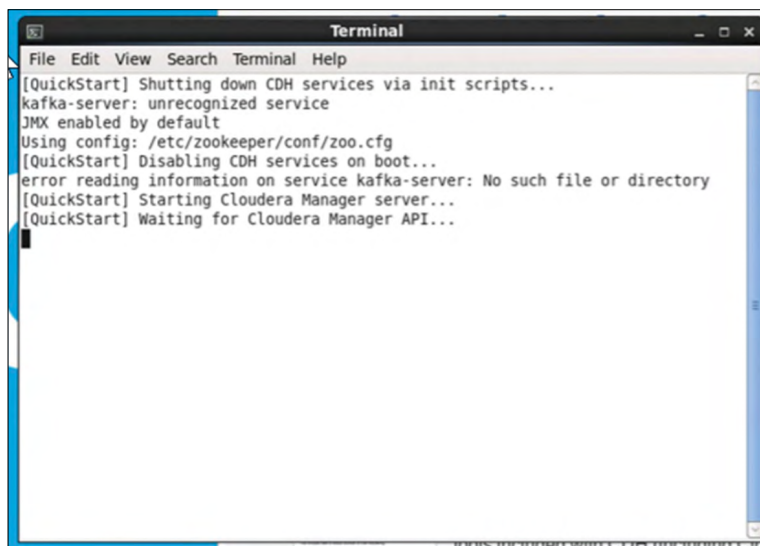


Figure 15. Script installs

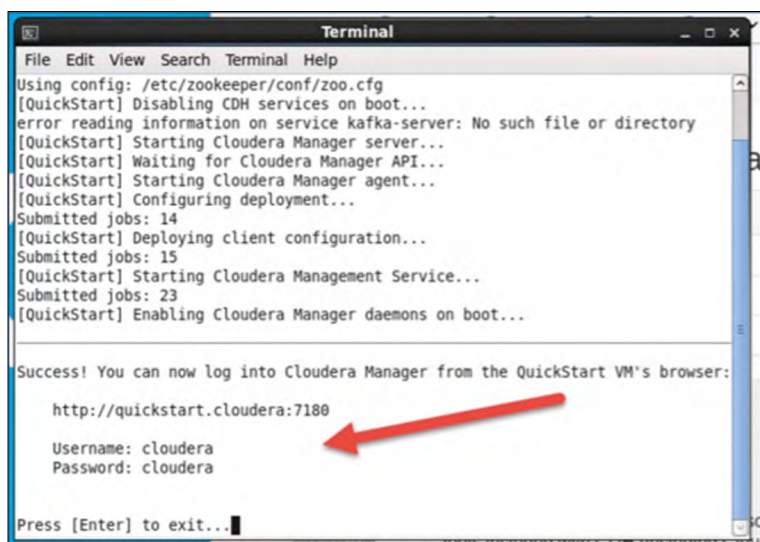
This can take several minutes depending on your machine.



```
Terminal
File Edit View Search Terminal Help
[QuickStart] Shutting down CDH services via init scripts...
kafka-server: unrecognized service
JMX enabled by default
Using config: /etc/zookeeper/conf/zoo.cfg
[QuickStart] Disabling CDH services on boot...
error reading information on service kafka-server: No such file or directory
[QuickStart] Starting Cloudera Manager server...
[QuickStart] Waiting for Cloudera Manager API...
```

Figure 16. Preparing VM

Once completed, it will give you instructions on how to connect to the Manager, as per screen shot below.



```
Terminal
File Edit View Search Terminal Help
Using config: /etc/zookeeper/conf/zoo.cfg
[QuickStart] Disabling CDH services on boot...
error reading information on service kafka-server: No such file or directory
[QuickStart] Starting Cloudera Manager server...
[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Starting Cloudera Manager agent...
[QuickStart] Configuring deployment...
Submitted jobs: 14
[QuickStart] Deploying client configuration...
Submitted jobs: 15
[QuickStart] Starting Cloudera Management Service...
Submitted jobs: 23
[QuickStart] Enabling Cloudera Manager daemons on boot...

Success! You can now log into Cloudera Manager from the QuickStart VM's browser:

    http://quickstart.cloudera:7180
    Username: cloudera
    Password: cloudera

Press [Enter] to exit...
```

Figure 17. Instructions

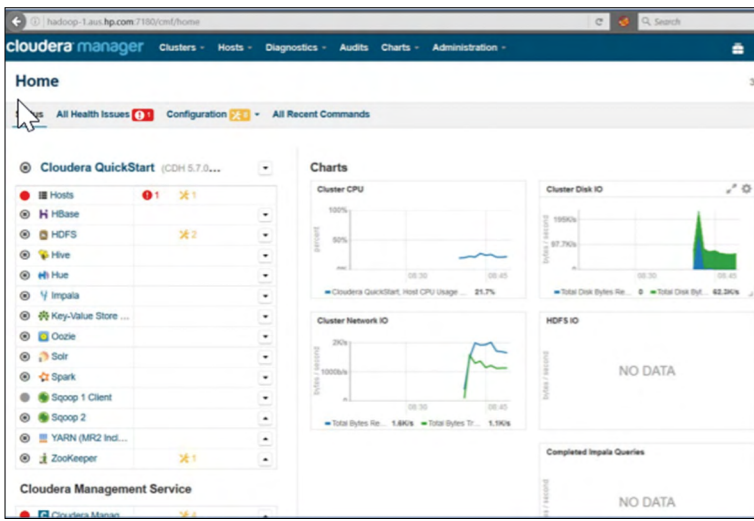


Figure 18. Set up services

You need to set up several of the services—HDFS for event storage, HUE for visualization and querying, and FLUME to pull the events from Transformation Hub in to HDFS. Below are the steps to do this.

FLUME

First thing is to install Flume, but we're not going to configure it just yet. Let's just get the service installed.

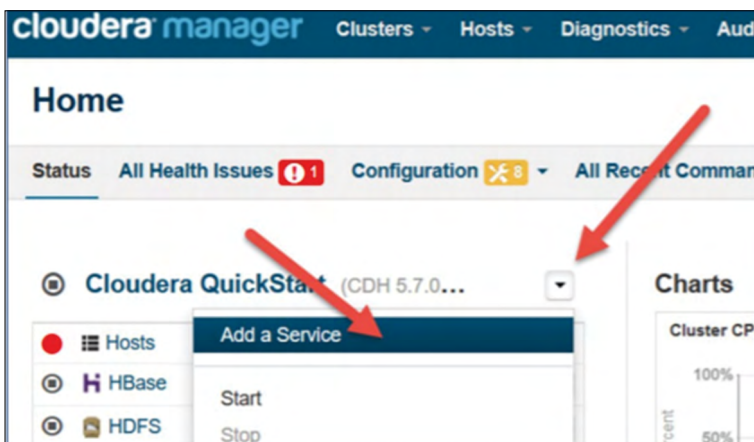


Figure 19. Add a service

Click on the Down Arrow on Cloudera QuickStart, and select Add a Service.

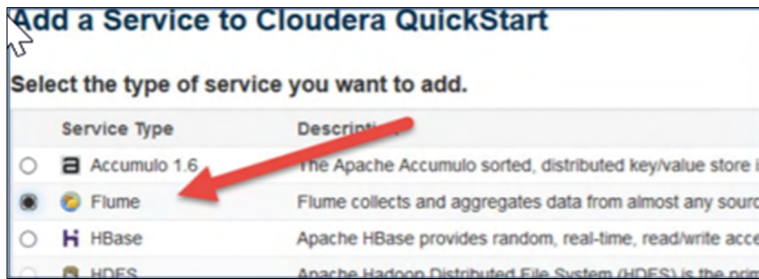


Figure 20. FLUME as a service

Select FLUME as the service.



Figure 21. HDFS and ZooKeeper dependencies

The dependencies should be HDFS and ZooKeeper. Select this option as it keeps the amount of work to complete to a minimum. At this point we have no need for HBase or Solr.

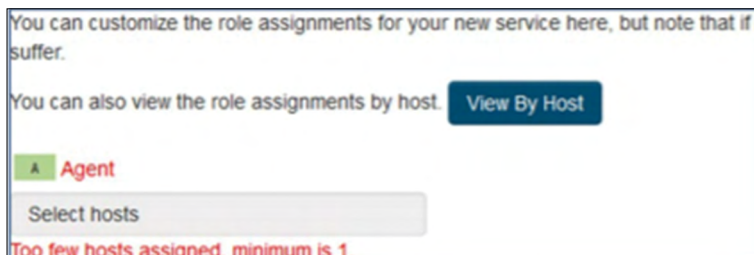


Figure 22. Select your VM

Click on Select hosts and then select your VM, as per screen shots.

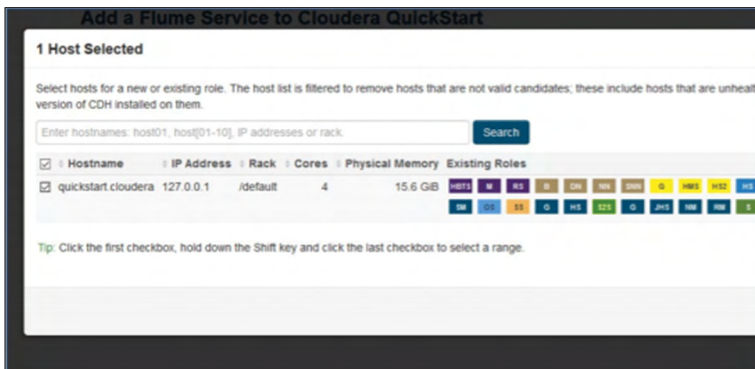


Figure 23. Finish

Click Finish to save this configuration.

Then go to Configuration for the Flume service. Here we are going to fill out our source/sink configuration so we can subscribe to Transformation Hub events and publish them back to HDFS.

1. Click on Flume --> Configuration
2. Scroll to "Configuration File" and delete ALL the content. Replace it with the information below.
3. Save the configuration and RESTART the service, or "Refresh the cluster." Either way, we want to reload the configuration file.

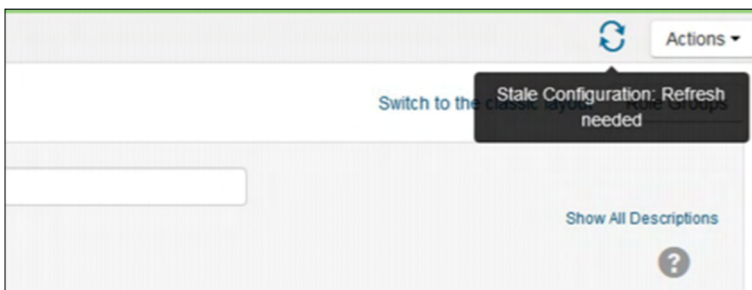


Figure 24. Stale Configuration

```
#####
#Working Flume/Kafka configuration file for Cloudera / Hadoop / Event Broker 1.0
#####
#defined Kafka Source, Channel, and Destination aliases
tier1.sources = source1
tier1.channels = channel1
tier1.sinks = sink1
#Kafka source configuration
tier1.sources.source1.type = org.apache.flume.source.kafka.KafkaSource
tier1.sources.source1.zookeeperConnect = eb-1.aus.hp.com:2181
tier1.sources.source1.topic = BlueCoatEvents
tier1.sources.source1.groupId = flume
tier1.sources.source1.channels = channel1
```

Figure 25. Delete ALL the content. Replace it with the information above

```
tier1.sources.source1.interceptors = i1
tier1.sources.source1.interceptors.i1.type = timestamp
tier1.sources.source1.kafka.consumer.timeout.ms = 150
tier1.sources.source1.kafka.consumer.batchsize = 100
#Kafka Channel configuration
tier1.channels.channel1.type = memory
tier1.channels.channel1.capacity = 10000
tier1.channels.channel1.transactionCapacity = 1000
#Kafka Sink (destination) configuration
tier1.sinks.sink1.type = hdfs
tier1.sinks.sink1.channel = channel1
tier1.sinks.sink1.hdfs.path = hdfs://localhost:8020/user/cloudera/flume/events/%y/%m
tier1.sinks.sink1.hdfs.rollInterval = 360
tier1.sinks.sink1.hdfs.rollSize = 0
tier1.sinks.sink1.hdfs.rollCount = 0
tier1.sinks.sink1.hdfs.fileType = DataStream
tier1.sinks.sink1.hdfs.filePrefix = cefEvents
tier1.sinks.sink1.hdfs.fileSuffix = .cef
tier1.sinks.sink1.hdfs.batchSize = 100
tier1.sinks.sink1.hdfs.timeZone = UTC
#####
```

Figure 26. Save the configuration and RESTART the service

You'll need to update some of the lines, which are detailed below:

```
tier1.sources.source1.zookeeperConnect = eb-1.aus.hp.com:2181
Your Event Broker node #1-with ZK port (not KAFKA port)
tier1.sources.source1.topic = BlueCoatEvents
```

Figure 27. Update lines

Your first topic

```
tier1.sources.source1.groupId = flume
```

Figure 28. First topic

This is your Consumer Group name

```
tier1.sinks.sink1.hdfs.rollInterval = 360
```

Figure 29. Group name

How often do you want your event files to roll over in HDFS? You'll need to adjust this depending on your expected ingestion rate and own needs. 360 is just fine for demonstrations.

```
tier1.sinks.sink1.hdfs.filePrefix = cefEvents
tier1.sinks.sink1.hdfs.fileSuffix = .cef
```

Figure 30. Frequency

The above are simply a HDFS file prefix and suffix. Change to suit your needs.

```
tier1.sinks.sink1.hdfs.path = hdfs://localhost:8020/user/cloudera/flume/events/%v/%m
```

Figure 31. HDFS file prefix and suffix

This is our sink—where in HDFS our files are going to be written. What I've done here is to place them under the cloudera users home directory with a structure of flume/events/ and then a month/year directory at the end. This allows files to be partitioned by month and year. Set this to your own preferences.

Of course, Flume is run as user flume and will typically not have write access to /user/cloudera/ so, update its permission to o+w. Obviously this is not best practice, but it allows us to write files for demonstrations. See the HUE configuration below for steps on how to achieve this in the WebGUI.

HDFS

Click on HDFS from the manager.

Click on Configuration.

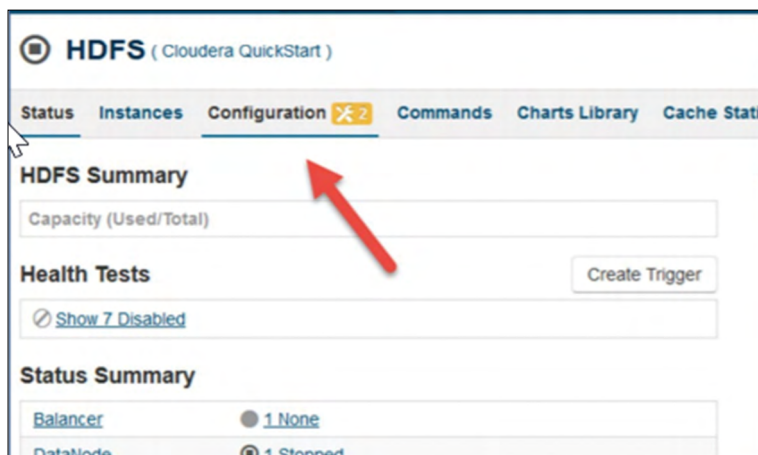


Figure 32. Configuration

Note that in the configuration you can specify multiple locations for event data. By default, the VM provides very little space, so consider adding another virtual disk if you need longevity. The disk can fill up and you'll find some services fail to work correctly.

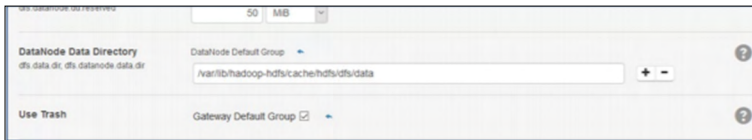


Figure 33. Save changes

Save any changes you make here. All other items can be left as default. Now start the HDFS service as per screen shot below.



Figure 34. HDFS service

HUE

Now we want to configure HUE. This will allow us to easily look at the events from HDFS. It really does not do anything for good visualizations, but it will at minimum demonstrate the ability to capture events in the entirety.

Start the HUE service from the Manager.

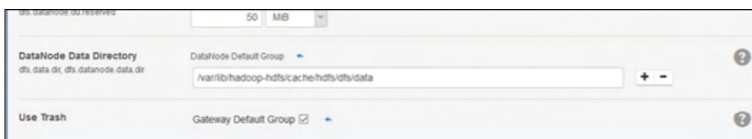


Figure 35. HUE service

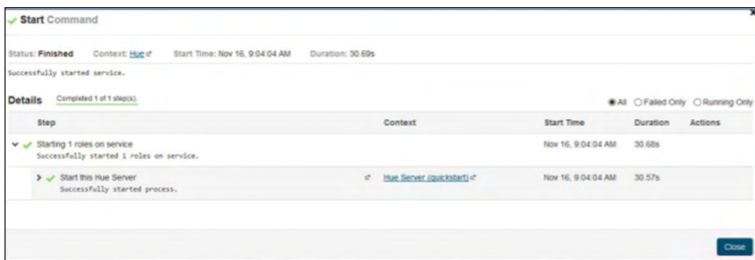


Figure 36. Connect HUE

Once started, you can connect to HUE on port 8888 (default) over HTTP. Login again as user cloudera. Skip the wizard, we don't have a need for it.

1. Click on File Browser from the top menu
2. Browse to /user/
3. Right click on "cloudera"

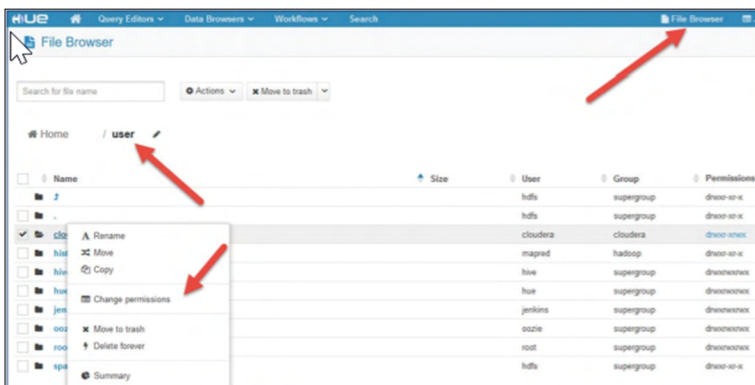


Figure 37. Write files to HDFS

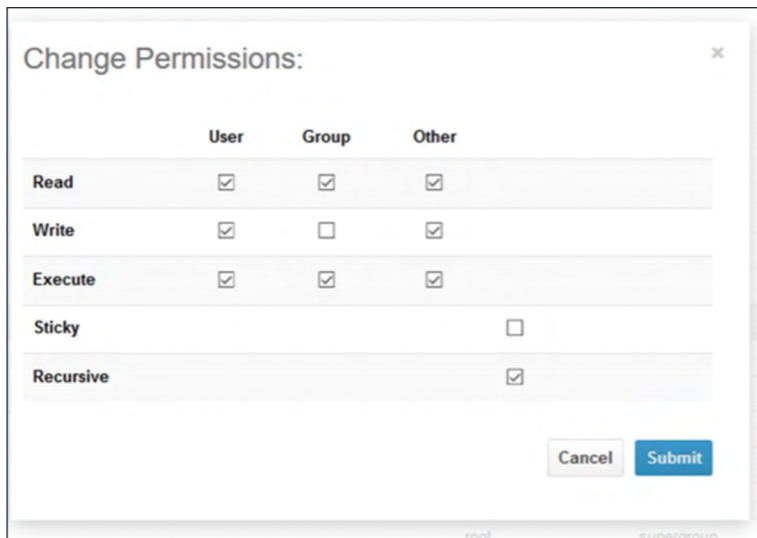


Figure 38. Change permissions

4. Select “Change permission”

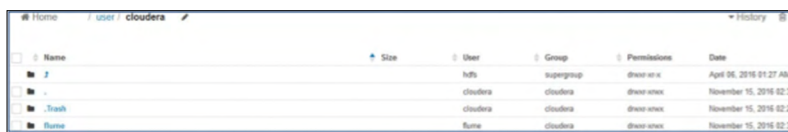


Figure 39. Flume directory

5. Give “other” write access and make it recursive. Now Flume can write the files to HDFS

Make sure Flume and HDFS services are started if they have not been done so already. Once they are started, you should see a directory called “flume” appear under /user/cloudera/ as per screen shot below.

Notice the owner is user flume—this is to be expected. The steps are minimal to achieve a working model.

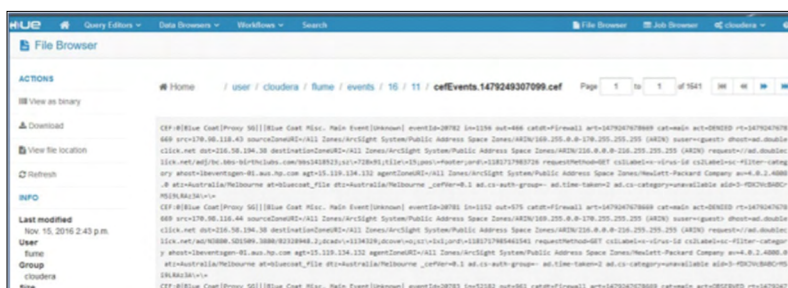


Figure 40. Stop Flume service from manager

Now traverse this directory to get to the .cef files stored. From here you can select a file and it will open, showing you the CEF contents of the file. At this point, the events will continue to be ingested and fill up your VM disk. To stop this, simply stop the Flume service from the Manager.

Summary

Security analysts struggle to detect and respond to threats using all available data connected to the network. ArcSight Security Open Data Platform leverages an open architecture to make data visible without boundaries, to be used over multiple use-cases. It enriches data in real time and makes it possible to move data from anywhere to anywhere for better detection, investigation, and response to threats. It lays the foundation for intelligent security operations by providing a reliable open architecture data collection platform that large environments demand.

Resources and Additional Links

ArcSight Security Open Data Platform: www.microfocus.com/sodp

ArcSight for Security Operations: www.arcsight.com

Learn more at

www.microfocus.com/sodp

Connect with Us

www.opentext.com



opentext™ | Cybersecurity

OpenText Cybersecurity provides comprehensive security solutions for companies and partners of all sizes. From prevention, detection and response to recovery, investigation and compliance, our unified end-to-end platform helps customers build cyber resilience via a holistic security portfolio. Powered by actionable insights from our real-time and contextual threat intelligence, OpenText Cybersecurity customers benefit from high efficacy products, a compliant experience and simplified security to help manage business risk.