Splunk Technical Paper:
Extracting More Value from SIEM Deployments:
Integrating Splunk with ArcSight

Co-Authors:
Alex Raitz, Security Specialist
Dan Goldburt, Federal Solutions Architect
Mark Seward, Director, Security and Compliance Solutions Marketing
The Challenges of Increased Security Scope

Over the last decade, many organizations have invested in Security Information and Event Management (SIEM) systems to automatically identify and manage security incidents without having analysts manually correlate events. Today, it’s no longer enough to apply correlation rules across traditional security data. Security has moved from monitoring the perimeter to protecting the user, and mitigating risks means having to look at data from the network, server, OS, and application layers. It’s become difficult to predict what data sources will be relevant to the security analyst, and difficult to scale a SIEM to get all the data. At the same time, attackers have become better at flying under the radar, making it difficult for machines to distinguish the security relevant events from the ‘noise’, and making it harder to tighten SIEM correlation rules without getting thousands of false positives a day. Security professionals caught in the data deluge are looking for new tools to organize all the data in a way that allows the human to determine what is relevant.

The Splunk + SIEM Solution

Splunk integration with SIEM lets you leverage all your IT data for a view into the entire infrastructure, while continuing to send a subset to the SIEM for further correlation. Splunk can collect and retain events from any machine without the overhead of having to normalize events into a fixed schema. Then, Splunk can send either raw data or structured events to the SIEM. Working with ArcSight, Splunk can provide real-time data streams in Common Event Format (CEF). A Splunk + ArcSight solution offers an organization the ability to look for patterns and relationships across terabytes of data, while consolidating their Connector footprint, offloading processing from ESM, and eliminating the need for separate Logger storage appliances.

In the illustration above, data is collected by Splunk, which provides long-term retention, the ability to drill into events, and the ability to dynamically impose structure to provide metrics and reports. Splunk also forwards a real-time data stream to ArcSight in CEF format. These events are correlated by the SIEM and actionable incidents are presented as alerts in the ESM console. Analysts review the alerts and when necessary, open tickets to manage a response. To perform incident investigation and analysis, the security teams have quick access to the raw data via Splunk.
SIEM Challenges

The promise of a SIEM is to automate decision-making and to reduce the amount of data the analyst needs to review. To achieve this SIEMs implement state machines and sophisticated rules that come with overhead. Extending the SIEM to cover larger volumes of data and to detect weaker signals of an attack presents several challenges:

- **Scalability** – SIEMs are expensive to scale and often reach architectural scalability limits relative to the volume of data an organization needs to collect and process. SIEM deployments are likely to require a database administrator (DBA) for continuous maintenance and performance optimization.

- **Data collection and normalization** – Handling unstructured data is a constant challenge. SIEMs are dependent on custom parsers (e.g. ArcSight Connectors) to normalize data into a fixed schema. Application logs (custom and off-the-shelf) do not follow a standard format and are subject to frequent change. To add support for a new data format, some SIEM vendors require a professional services engagement, while some challenge the user to create their own parsers.

- **Implementation and tuning** – It is difficult to strike the right balance between correlation rules that catch all possible attacks and correlation rules that produce too many false-positive alerts. Tuning often requires a professional services engagement and on-going expenses, and industry analysts report speaking with customers for whom, “…a year of tuning was required.”

- **Trending and analytics** – SIEMs often have a selection of canned reports but new report creation is not flexible enough to adjust to changing conditions. Canned reports can be useful, and may look great initially, but relying on a canned report to understand the end-to-end implications of a security event from the edge router to the application simply doesn’t work.

- **The “Rules-based” approach** – When a correlated security event is presented to the security analyst, it’s reasonable to expect the analyst to limit his or her investigation to the data sources reported by the alert. Humans tend to ignore subtle signs or signals that don’t fit into a pre-existing model, and SIEMs reinforce that tendency. A “Rules-based” approach supports only a go-forward view of security data; if you get a correlation rule wrong, you can’t adjust the model and re-analyze the data, because events that didn’t match the old rule have already been discarded.

---

1 Andrew Hay, 451 Group, Interview, August 2010
Why Splunk

Splunk allows the user to quickly work with all machine data without filtering or reduction. Events are stored in flat files using a real-time indexing algorithm invented specifically for IT data, and users issue queries in a natural search language with a rich layer of analytical commands. The mechanism of search supports the ability to extract knowledge from events, create visualizations, alerts, and dashboards and enable real-time data capture and display. Splunk’s key benefits are:

- **Scale beyond what is achievable with a SIEM** – Splunk is the industry’s first product to use ‘Map-Reduce’ (a paradigm Google pioneered) to enable massively parallel reporting across commodity machines, from one place and in real-time. Splunk deployments exceed indexing rates of a terabyte per day.

- **Broader focus on data beyond traditional security logs** – Splunk can collect data in any format, without requiring parsers or connectors. Get all the data including custom application logs and other non-traditional data sources including registry changes, performance metrics, process tables, and file system changes.

  “Splunk is very adept at handling unstructured data sources, providing strong reporting and statistical analysis, whereas many other solutions require that unstructured data be normalized before any reporting or analysis can begin.”

- **Reduced time to investigate** – The SIEM creates an alert, then what? With Splunk, security analysts can access all IT data in the organization and follow an investigation wherever it leads, from the firewall and IDS to the server, OS, and application – without jumping to other tools. By reducing forensic investigations to minutes instead of hours or days, Splunk lowers incident response cost, but also lowers risk as attack windows are shortened.

- **Unparalleled analytical capabilities**. Splunk allows the user to quickly discover new relationships in the data and provides over 80 analytical commands to manipulate and discover meaning in data. These commands range from statistical operations like ‘k-means’ to session-analysis operations like ‘transaction’. The full query language can be applied to historical data and can also be run in real-time. The user can ask

any question of the data, without having to plan ahead during data-ingest or having to tune a back-end schema. Result previews are generated instantly, so the analyst can get insight without waiting for the query to fully complete.

- **Support for a ‘Pattern-based’ security strategy.** In 2009 Gartner began a new research theme on "Pattern-based Strategy." This research is applicable for all business activities including security. The strategy consists of three parts:
  1. Seek – analyze activity and access patterns that contain the signals of a potential threat
  2. Model – implement analytics and assessment to determine which patterns present greater risk to the organization by qualifying and quantifying the impact
  3. Adapt – action to protect users, accounts, data and infrastructure from the threat that was discovered and assessed in the previous phases

Splunk supports this new strategy. Users can ‘seek’ new patterns of activity in log data that indicate a risk to the business. Using Splunk’s analytical commands, users can ‘model’ those risk patterns and save them as a Splunk query, and ‘adapt’ to mitigate new risks by scheduling the query to monitor for that pattern in real-time. A pattern-based strategy is the complement to the rules-based state machine technology offered by SIEM.
ArcSight Components Overview

This section undertakes a brief review of ArcSight components that will be important in later sections. ArcSight is typical of a SIEM that has a log collection component to store-and-forward data to a correlation engine. Other vendors have similar architectures with different names for each component.

CEF – Common Event Format is used as the standard normalization format for the ArcSight platform. It is also promoted as an open log standard. The basic format is:

CEF:Version|Device Vendor|Device Product|Device Version|SignatureID|Name|Severity|Extension

The Extension section contains the bulk of the application specific message. Some of the extension fields are standardized, such as ‘user’ and ‘src’, while others can be custom and are assigned names such as cs1, cs1Label, cn1, cn1Label – where ‘cs’ stands for custom string and ‘cn’ stands for custom number. See page 7 for an example.

Connectors – Connectors are responsible for collecting and normalizing events into CEF and for sending normalized events to ArcSight components down the line. Connectors can also perform event filtering, event message caching, and network bandwidth throttling. There are many versions of Connectors for different event formats (275+ as of 2010). For unsupported formats, users can program a FlexConnector, which requires a developer license and can be difficult, particularly for multi-line application logs. For the purposes of this paper, we highlight the CEF Connector and the Syslog Connector, the latter is used for CEF events transmitted in the body of a syslog message.

Logger – This is ArcSight’s event retention component, available as an appliance. It can receive normalized data from Connectors or raw data from syslog or files. Depending on whether the data is structured or unstructured, Logger stores events in a back-end database or in a flat-file index (using a modified search engine based on the open-source Lucene project). Reporting is not possible on unstructured data, and reporting across appliances is also not possible. Logger supports alerting but is limited to a total of five alerts.

ESM – Enterprise Security Manager is the SIEM. It receives CEF formatted data, performs event correlation and alerting, allows analysts to manage incident response workflow, and provides reporting. ESM receives normalized events and determines which are relevant, employing algorithmic techniques such as Bayesian logic to quantify uncertainty and decision trees (known as “correlation rules”) to ultimately accept or ignore data. Events that correlate are retained in an Oracle database; Events that don’t correlate are discarded. Tuning can minimize but not eliminate false positives, and security analysts spend much of their time prioritizing and managing investigations by using the workflow capabilities of the ESM console.
Getting Data into Splunk

We start the discussion at the data collection layer. If an existing ArcSight collection infrastructure is in place, Splunk can integrate by receiving raw syslog or CEF normalized events from any ArcSight Connector. For data not currently collected, and even for data that is currently collected with other tools, using Splunk offers several unique capabilities:

- Splunk always maintains events in original form, both in-transit and on disk. Multi-line events such as Windows Event Logs and Java stack traces are not converted to single-line messages during transport. Events are written to flat files on disk, not inserted into a database or locked in any proprietary format.

- Out-of-the-box support for any and all formats, with no marginal cost for new data types. No parsers to buy or update as data formats change. No per-source licensing model.

- Simpler data collection architecture. There is no need to deploy and maintain a specialized normalization tier (i.e. ArcSight Connectors) between the systems generating the data and Splunk. Events only need to be transported to a Splunk indexer, where reporting is supported even for unstructured data.

- Get at any data generated by any machine. To get data, Splunk offers the following input capabilities: a) listen for TCP and/or UDP data streams over the network; b) tail local log files or directories; c) index the entire contents of configuration files; d) monitor the local disk for creation, deletion and modification of files; and e) gather performance metrics from system utilities. Windows-specific input capabilities allow Splunk to: f) collect WinEventLog and Performance Counters; and g) detect changes to the Registry or Active Directory schema. Finally, Splunk offers h) a scripted input option, which allows the user to specify any program which generates output. Splunk will execute the program and index STDOUT. With scripted inputs, Splunk can be extended to capture binary data such as netflow and snmp traps, or to poll events from a database. Other common examples are to index the output of system utilities like ps, top, vmstat, and netstat.

- An opportunity to consolidate data collectors. In many cases, Splunk can retrieve data without an agent. This works for anything that can be exported using a network protocol like syslog, or retrieved using a programmatic interface like WMI. For all other data, Splunk can be installed in agent mode (as a “lightweight forwarder”), on any modern operating system. Forwarders feature all the input options listed above and route data securely and reliably using SSL over TCP. Lightweight forwarders have a footprint of 50MB RAM, 1% CPU, and support network throttling. Furthermore, events collected by Splunk can be sent to a third party in any format, allowing one Splunk agent to serve data to multiple tools. The rest of the paper will address how this is accomplished.
Streaming Data from Splunk to ArcSight

Splunk can stream real-time data to ArcSight in two ways: by forking off raw events at index-time or by forwarding transformed events at search-time. Index-time output can be used only for data that is natively supported by ArcSight, and events must be routed to the appropriate ArcSight Connector. The index-time method is less flexible but simple to setup and maintain. Search-time output offers the maximum flexibility—the full Splunk search language can be used to select the subset of events to be forwarded—with the added ability to configure how an event should be re-formatting before it’s released. The two methods can also be combined; for example, Splunk can collect multiple data streams from a Linux server, such as CPU metrics, changes to the passwd file, and updates to the audit log, then route only syslog events (e.g. the audit log) at index-time, while using search-time routing for a subset of the other events (e.g. processes that are consuming more than 50% of CPU), sending those as CEF.

Index-time output

To configure index-time output, an extra output processor can be inserted into the index-time pipeline (the pipeline handles events from when Splunk receives them to when they are written to disk). Two output processors are available: syslog and tcpout. The tcpout processor sends completely raw events over tcp; it is the same processor that a lightweight forwarder uses when sending data to a Splunk indexer. The syslog processor adds a RFC 5424 compliant header to events before it sends them over TCP or UDP.

The following example configures a syslog and tcpout processor, and routes all events to both. To insert a processor, the user must modify outputs.conf in the correct location, as described here: http://www.splunk.com/base/Documentation/latest/Admin/Aboutconfigurationfiles. For other examples, including an example of conditionally routing subsets of data to different places, see: http://www.splunk.com/base/Documentation/latest/Admin/Forwarddatatothird-partysystems. Note that comments must be removed to constitute a valid config:

```
[syslog]
defaultGroup=syslog_receiver #specifies the target group name where all events will be sent
indexAndForward=true #forks events so a copy is indexed as well as forwarded
type=udp #specifies whether to send syslog over TCP or UDP

[tcpout]
defaultGroup=raw_tcp_receiver #a copy of events will be sent here as well as to the syslog processor above
indexAndForward=true #forks events so a copy is indexed as well as forwarded
sendCookedData=false #leaves the data untouched, otherwise will be optimized for a Splunk receiver

[syslog:syslog_receiver]
server=my.ip.or.name:port #specifying a comma separated list here will load-balance

[tcpout:raw_tcp_receiver]
server=my.ip.or.name:port2 #specifying a comma separated list here will load balance
```
Search-time output

For non-standard formats such as application logs Splunk can convert data to CEF in real time and export it to an ArcSight Connector. This allows Splunk to help the user avoid building his own FlexConnector. The CEF-output framework functions as follows:

- First, the real-time search API is used to inspect events just before they are indexed. Because the full Splunk search language works here, it is easy to intelligently filter what gets forwarded.

- Once an event occurs, the framework reformats it into CEF (or other defined format). To structure the output, information about the event must be made available by using Splunk functionality such as field extraction and lookups. Fields extracted to the Splunk Common Information Model\(^3\) are automatically translated to standard CEF fields. Lookups enrich events with the CEF meta-information, like ‘Device Vendor’ and ‘Product’. For help formatting events as CEF, please see the concluding section of this document.

- Finally, the event is sent to an ArcSight Connector. The default is to send the message as syslog to a Syslog Connector, but Splunk can also write the event to a local file for a software-based CEF Connector to read.

As an example, this is the CEF definition for DHCP events:

```CEF:0|Microsoft|DHCP Server||{EventID}|{EventName}|Unknown|cn1={leases expired} cn2={leases deleted} cs4={MAC Vendor Prefix} cs5={Ethernet Vendor} rt={Date, Time} src={Address} shost={HostName} smac={sourceMAC}```

The CEF output framework is invoked like this:

```
splunk cmd ./rtoutput.py -t CEF -S 'CEF.Connector.ip' -P 'port' -l "dhcpd request | fields *"
```

The flag ‘-t’ is set to ‘CEF’, which means the command will output in CEF (‘KV’ is also supported, which will output in key-value pairs.) ‘-S’ and ‘-P’ specify the CEF Connector destination (‘-F’ is supported for output to a file). ‘-I’ allows for interactive Splunk password prompts. The real-time search ‘dhcpd request | fields *’ looks at matching events and request for extraction on all known fields, which must include all necessary fields in the “Extension” portion of the CEF, such as src and smac. A Splunk lookup table must be configured to enrich events with the ‘EventId’ and ‘EventName’ fields.

An example result is: Jul 14 11:42:14 splunkindexer CEF:0|Linux|Dhcp Server|1.0|15|Lease requested|5|smac=001B63CDA156 rt=1279132934 src=192.168.0.1 cs5=apple computer inc cs4=001B863

A word on scalability of the integration: The speed at which Splunk can stream data can overwhelm a single ArcSight Connector. Furthermore, Splunk’s real-time search works in distributed mode, so running the search-time output framework from one place will pull events from all Splunk indexers in the distributed environment. To scale the output for ArcSight, either restrict the subset of events Splunk will forward, or set up a cluster of Connectors that Splunk can load-balance across.

---

Sending Splunk Alerts to ArcSight

Splunk search language supports queries that can look for arbitrarily complex patterns in one or more data sources over time. This allows not just for routing specific subsets of data to ESM for further correlation, but for triggering Splunk alerts to be sent to the ESM console. Splunk has different strengths than ESM in the patterns it can apply; For example, Splunk can look for “low and slow” threats by examining a large time-window of events, something difficult to do with the SIEM’s in-memory state engine. Splunk also excels at reconstructing a transaction that crosses multiple systems and multiple log files, at enriching data with external information such as geo-location or asset information, and at using statistical methods to detect anomalies. For example:

```
failed password startdaysago=1 | stats count by src_ip | eventstats avg(count), stdev(count) | where count > 'avg(count)'+stdev(count)
```

The above search looks over the last day of data for repeat authentication errors outside of a statistical deviation. This would be saved as a scheduled search, and if any new events were found they would trigger a scripted alert that would output CEF or send an SNMP trap to the SIEM. For more documentation and example alert scripts see:

Other Integration Workflows

Finally, many security analysts want bi-directional integration between the SIEM console and the raw data. With ArcSight, it is not easy to go from an ESM alert back to raw data in Logger: the analyst must find the alert field that reveals which appliance holds the raw events, then manually login to that Logger interface, and finally re-construct a query that returns the events of interest. With Splunk, future work can be done to implement a one-click drill down, using Splunk’s query-string driven permalink feature. For instance, the following link would open a browser with the query ‘failed password’ already running, making it easy to quickly conduct an incident investigation in Splunk: http://splunkserver:8000/en-US/app/search/flashtimeline?q=search failed password

Conclusion

Splunk is not a SIEM. It doesn’t provide a console for displaying alerts nor allow analysts to create tickets to manage response. Splunk also doesn’t implement an in-memory state machine to perform event correlation, but that’s not to say that Splunk can’t run complex analytics in real-time. Still, some questions are better asked of a SIEM, and some are better asked of Splunk.

The functional innovation of Splunk is that it is built from the ground up on search and indexing technology, while SIEMs are built on parsers, schemas, and database technology. Splunk provides a simple Google-like interface to find information in terabytes of data. Using the mechanism of search, Splunk can also perform statistical analysis, alert on complex conditions, and generate reports – all without a back-end database.

What that means for the architect is that he doesn’t have to worry about parsing the data to fit a proprietary schema (i.e. no need for ArcSight Connectors). With Splunk, events just need to be transported to an indexer in their native format. For the end user, this is of particular importance when facing new reporting requirements. With a SIEM tool, the user would be relying on the relevant events having been parsed correctly (since Logger and ESM can’t report on unstructured data), with no way to generate the report if a parsing mistake had been made. With Splunk, the user has the ability at search-time to impose ad-hoc structure, enrich events with outside information, and create new reports and dashboard definitions.

For those ArcSight ESM customers wishing to have a more complete, scalable, and flexible event analysis solution, Splunk can act as a Logger replacement while continuing to support the analyst by feeding CEF data to ArcSight ESM. For those customers with other SIEM products, Splunk can also stream data as syslog or raw TCP.

For additional information, including information about the framework for CEF streaming from Splunk to ArcSight, please contact a Splunk Federal or Commercial Solution Architect: fed@splunk.com, sales@splunk.com

About Splunk
Splunk is the world’s leading Operational Intelligence software used to monitor, report and analyze live streaming IT data as well as terabytes of historical data – located on-premise or in the cloud. Almost half of the Fortune 100 and more than 1,800 enterprises, service providers and government organizations in 70 countries use Splunk to improve service levels, reduce IT operations costs, mitigate security risks, and drive new levels of operational visibility.

For a new approach to IT, visit http://www.splunk.com, or visit http://www.splunk.com/download to download a free copy.