Extracting More Value from SIEM Deployments:
Integrating Splunk with ArcSight
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 events. These security events are expressed as correlation rules implemented on the SIEM. The primary goal of these rules was to automate the correlation of security log data coming from many sources in the enterprise relieving the analyst of manual activity. The security team’s primary task was to protect networks and computers. Now it’s all about protecting the data.

Security events can start anywhere and it’s become impossible to predict what data sources will be relevant when analyzing possible threats. It’s no longer enough to apply correlation rules just across traditional security data. Data generated by custom and off-the-shelf applications might not be collected or viewed as part of a security event. Responding to data from traditional security sources without being able to view its effect on an application is like seeing an auto accident on the side of the road and having no visibility into the health of the occupants. This lack of visibility also affects the response to incident: Do I send a tow truck, an ambulance or both. To truly mitigate business risk, traditional security data and application data must be viewed together and on the same timeline.

Also, attackers have become better at “flying under the radar” hiding from deployed security products and making it difficult for machines to distinguish the security relevant events from other machine-generated data that otherwise might have been considered “noise.” For those security teams focused on data protection and root-cause analysis, all machine-generated data has security relevance. Security professionals are caught in a data deluge and are looking for new tools to organize all the data and find relevant but less obvious relationships in in this data.

Recent prominent security events have illustrated that often only human beings can pick out the patterns necessary to spot sophisticated attacks. For example, the advanced persistent threat (APT) that targeted Google in December 2009 could not have been detected by a SIEM ingesting anti-virus, IPS, web proxy, and firewall data. “The persistence in APT intrusions is manifested in two ways: maintaining [an undetected] presence on your network, and repeatedly attempting to gain entry to areas where presence is not established.”

SIEM Challenges

The SIEM approach to data deluge is to automate the decision-making process and reduce the amount of data the analyst needs to review. SIEM complexity can make this difficult and costly to get right. As SIEMs have bulked up with features, complex SIEM deployments are taking longer to tune. Industry analysts report speaking with customers for whom, “...a year of tuning was required.” Traditional methods of correlating security data that involve data reduction or “removing the noise” are no longer effective. There are several challenges for proper deployment and implementation of a SIEM:

- **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. Some SIEM deployments require a database administrator (DBA) for continuous maintenance and performance optimization.

- **Data collection and structure:** Unstructured data is a constant challenge. SIEMs are dependent on custom parsers or connectors to normalize data at collection time into a fixed schema. Application logs (custom and off-the-shelf) consist of unstructured data, do not follow a standard format and are subject to frequent change. To add support for a new data source some SIEM vendors require a professional services engagement while some challenge the user to create their own parsers. Both add cost and take time to create.

- **Implementation and tuning:** Most SIEMs only support a go-forward view of security data. Once the rules engine has processed an event there is no going back. If you get a correlation rule wrong, you can’t re-analyze the data already processed and replay the analysis. Getting it “the next time” isn’t a good response to the CIO or CSO. Creating new correlation rules and tuning for false-positive reduction often requires a professional services engagement and on-going expenses.

- **Trending and analytics:** SIEMs don’t adequately support long-term metrics nor are they flexible enough to adjust to changing conditions. Historical search, long-term trending and analytics are increasingly important for a risk-based approach to security and compliance.

- **The SIEM “rule-based” approach:** Humans tend to ignore subtle signs or signals that don’t fit into or support a pre-existing model. Simply put, when a rule-based security event is presented to the security analyst, it’s reasonable to expect the analyst to have a tendency to limit their investigation to the data sources the SIEM used for correlation and not other data in the same time window. 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.

Why Splunk

Splunk Enterprise allows users to quickly work with time-stamped ASCII data without filtering or reduction. Events are written to a flat file data store using a real-time indexing algorithm invented for collection of machine-generated data. Users create queries in a natural search language combined with a rich layer of analytical commands. This approach supports metrics tracking, which can be used to create visualizations, alerts and dashboards that support real-time data capture and display. Splunk’s key benefits are:
Broader focus on data beyond traditional security logs: Splunk can collect data in virtually any format without requiring parsers or connectors. Splunk can be used to access machine-generated data including custom application logs and other non-traditional data sources including registry changes, performance metrics, process tables, file system changes, etc. “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.”

Splunk supports this new strategy by facilitating patterns as automated search queries. Users can seek and discover patterns of activity in log data that can uncover risks to the business. Using Splunk’s analytical commands, users can model those risk patterns in Splunk and adapt them to mitigate new risks over time. A pattern-based strategy is the complement to the rules-based state machine technology offered by a SIEM.

Unparalleled analytical capabilities: Splunk allows users to quickly discover new relationships in the data and provides over 80 search commands to manipulate and discover meaning in data. These commands range from statistical operations like ‘k-means’ to session-analysis operations like ‘transaction’. The user can immediately 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. The same query language that applies to historical data can be run in real time.

The Solution
Integrating Splunk with a SIEM lets you leverage all your machine-generated data for broader views across possible attack vectors. Splunk can send a subset of either raw data or search-correlated events to the SIEM for further correlation. These events can contain data that might otherwise go uncaptured without custom parsers. Specifically with ArcSight, a longtime SIEM leader, Splunk can now provide a real-time data stream converted to Common Event Format (CEF)—an open event standard supported by ArcSight ESM. A Splunk+ArcSight solution offers the user the ability to create saved searches that look for patterns and relationships across terabytes of structured and unstructured data thus eliminating the need for a separate log collection appliance. When a relationship is found, the search can be saved as an event and sent to the SIEM taking advantage of legacy workflows already in place.

Splunk Feeding ArcSight ESM
In the illustration below, logs are sent to Splunk, which provides log term trending for metrics with the ability to drill into the log data. Splunk also presents a real-time data stream to ArcSight in CEF format. These events in the stream are correlated, and, as needed help desk tickets are created and alerts are sent to security analysts for review. The help desk and security teams have access to log data for troubleshooting and analysis.
Splunk can also collect data from a variety of other data sources such as content aware DLP, database monitoring tools and other access management tools and forward events to the SIEM to create a view of possible fraud activities in the enterprise.

**ArcSight ESM Feeding Splunk**

While Splunk is capable of forwarding a real-time CEF feed to ArcSight ESM, it should be noted that there are current Splunk customers that forward ArcSight ESM data to Splunk. Customers have found that Logger was not able to provide a review of historical data over a long enough period of time. ArcSight ESM can be set up to extract certain fields from its database to a CSV file as a scheduled task. Splunk in turn can read the CSV and store the information for forensics use cases and has the scalability to search over a much longer time. There are also customers that see the value of archiving ArcSight ESM alerts in Splunk. The combination of scalability and the flexibility of Splunk’s command language make it easier to perform ad-hoc searches for analysis of zero-day attacks and discovering attack vectors for Advanced Persistent Threats. The diagrams below illustrate an active integration between Splunk and ArcSight with Splunk acting as a drop-in replacement for ArcSight Logger.

**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.

**Connectors:** Connectors are responsible for collecting and normalizing events into CEF and for sending these normalized events to ArcSight components up 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 (approximately 275). For unsupported formats, users can program a FlexConnector, which requires a developer license and can be a lengthy project, particularly for multi-line application logs. For the purposes of this paper, we highlight the CEF connector, which is used for events natively in CEF, and the syslog connector, which is used for CEF events that are transmitted in the body of a syslog message.

**Logger:** This is ArcSight’s event collection and reporting component, available as an appliance. It can receive normalized data from connectors or raw data from syslog or a file. Depending on whether the data is structured or unstructured, Logger stores events in either a back-end database or indexes them with 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 flagship SIEM component. It receives CEF formatted events, performs event correlation and alerting to allow analysts to manage incident response workflow and also provides some reporting capabilities. 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 a database, prioritized, and displayed on a console; events that don’t correlate are discarded. Tuning can minimize but not eliminate false positives, and security analysts must manage events 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.

- **Limited complexity with a single-tier 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.

- **Splunk offers a way to get at any data being 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 and index the STDOUT.

In this way Splunk can be extended to capture binary data such as netflow or 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.

- **Splunk offers an opportunity for Collector consolidation.** 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 TCP/SSL. 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 solution 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 is appropriate for data that is natively supported by ArcSight, and events must be routed to the appropriate ArcSight Connector. Splunk forwards raw events in a way that is natively supported by an 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 (i.e., 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. Note that comments must be removed to constitute a valid config. 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

```bash
[syslog] #initiates the syslog output processor
defaultGroup=syslog_receiver #specifies the target group name where all events will be sent
```
For example, this is the CEF definition for DHCP events:

```
CEF:0|Microsoft|DHCP Server|{EventID}|{EventName}|Unknown
n|cn|l|{leases expired}|c|n|l|{leases deleted}|cs|a|t|l|{MAC Vendor Prefix}|cs|5|{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 real-time search looks at events matching `dhcdp request` and asks for extraction on all known fields. All necessary fields in the “Extension” portion of the message, such as src and smac, must be extracted at search time, and a lookup to a list of Microsoft event IDs must be configured to add the ‘EventId’ and ‘EventName’ 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 (‘-t’ is supported for output to a file). ‘-I’ allows for interactive Splunk password prompts.

An example result is: Jul 14 11:42:14 localhost CEF:0|Linux|Dhcp Server|{EventID}|{EventName}|Unknown
n|cn|l|{leases expired}|c|n|l|{leases deleted}|cs|a|t|l|{MAC Vendor Prefix}|cs|5|{Ethernet Vendor} rt|{Date, Time} src|{Address} shost|{HostName} smac|{sourceMAC}

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 searches 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. Writing alerts in Splunk does not require a developer license as it does in ESM. Splunk also has different strengths than ESM in the patterns it can apply. For example, Splunk can look for “low and slow” persistent threats by examining a large time-window of events, something difficult to do with a SIEM in-memory state engine. Splunk also excels at reconstructing a transaction that crosses multiple systems and multiple log files, and at enriching data with external information such as geo-location or asset information.

For example, the search below looks over the last day of data for repeat authentication errors outside of a statistical deviation.

```
‘failed password startdaysago|1 | stats count by src_ip | eventstats avg(count), stdev(count) | where count > ‘avg(count)+‘stdev(count)’
```
This search would be saved as a scheduled search, and if any events are found they would trigger a scripted alert that outputs CEF or sends an SNMP trap. The result would be to trigger an alarm on the ESM console. For more documentation and example alert scripts see:

http://www.splunk.com/base/Documentation/latest/admin/ConfigureScriptedAlerts

For additional information about the framework for streaming real-time data from Splunk to ArcSight using CEF, please contact:

Alex Raitz araitz@splunk.com
Dan Goldburt dan@splunk.com

Other Integration Workflows

Finally, many security analysts want a bi-directional integration between the ESM console and the raw data. 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 Logger appliance holds the raw events, then manually log in to that Logger interface and construct a query that returns the events of interest. With additional professional services work can be done to implement a one-click drill down into Splunk, using the query string-driven permalink feature. For example, the link below will open a browser with the query already running, thus making it easy to quickly conduct an incident investigation in Splunk.


Conclusion:

For those ArcSight ESM customers wishing to have a more complete and schema-less real-time log management solution with the ability to customize dashboards with analytics, correlation capabilities, and no limits on alerting, Splunk can act as a Logger replacement while enhancing the user experience with ArcSight ESM. For customers with other SIEM products, Splunk can also stream data to a SIEM as syslog or generic TCP stream. For additional information about the framework for streaming real-time data from Splunk to ArcSight using CEF or other SIEM product, please contact a Splunk sales engineer for more details.

2 Andrew Hay, 451 Group, Interview, August 2010
3 http://www.infoworld.com/d/data-explosion/log-management-review-splunk-4-678
4 Pattern Discovery With Security Monitoring and Fraud, Detection Technologies, Mark Nicolett, Gartner Research, September 2, 2009