Onboarding data to Splunk Enterprise Security

Ingesting data correctly is a foundational step in your Splunk security implementation that, if done correctly, allows you to get the most value across your entire Splunk environment.

Splunk Enterprise Security works most effectively when you send all your security data into a Splunk deployment to be indexed. You should then use data models to map your data to common fields with the same name so that they can be used and identified properly.

The Splunk Common Information Model (CIM) is a “shared semantic model focused on extracting value from data.” It is used to normalize your data to match a common standard. For example, when you search for an IP address, different data sources may use different field names such as `ipaddr`, `ip_addr`, `ip_address`, or `ip`. The CIM normalizes different data sources to use the same field name for consistency across all sources. This normalization is especially important when you are ingesting data from multiple sources, which can cause problems if they are not standardized with a time synchronization mechanism.

The volume, type, and number of data sources influence the overall Splunk platform architecture, the number and placement of forwarders, estimated load, and impact on network resources. The Splunk platform can index any kind of data, for example any and all IT streaming, machine, and historical data, such as Microsoft Windows event logs, web server logs, live application logs, network feeds, metrics, change monitoring, message queues, or archive files. Getting Data In (GDI) is the process that you'll follow to ingest machine data into Splunk.

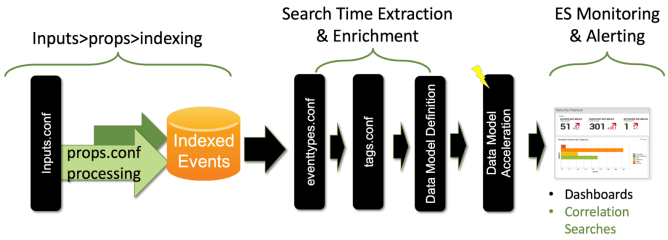
First, let’s discuss some of the terminology and concepts that are important to bringing the right data in the right way.

This article is part of Splunk's Use Case Explorer for Security, which is designed to help you identify and implement prescriptive use cases that drive incremental business value. In the Security maturity journey described in the Use Case Explorer, this article is part of Normalization.

Data models, normalization and the Common Information Model (CIM)

Data models

After you configure data ingestion in your Splunk environment, you'll need to work with data models in order for Splunk Enterprise Security to take advantage of that data. Data models are an abstraction that sit over the top of existing data or searches and apply an information structure to raw data. They are most commonly used for normalization and acceleration or summarization. Using a data model makes your searching faster and more efficient by subdividing your events into separate classes of events (similar to eventtypes).



Common Information Model (CIM)

The CIM is a shared semantic model focused on extracting value from data. It contains a collection of data models, documentation, and tools that support data normalization for maximum search time efficiency. It also contains multiple security-related data models, such as Authentication, Network Traffic, Malware, Vulnerabilities, Endpoint, and more. Splunk Enterprise Security includes the CIM add-on out of the box.

Deploying add-ons to normalize data for Splunk Enterprise Security

When onboarding data, you can choose from a few different ways to collect and process it. Because of the versatility of Splunk, you can manually build your own field extractions and tags. However, Splunk has coverage and support for the vast majority of vendors and products. A lot of the work of extracting fields and onboarding is already done.

The Splunk-preferred option is to use an existing Splunk Certified or Splunk Built and Supported add-on from Splunkbase. These add-ons include extensive documentation and are generally easy to configure. For example, some configurations might ask you only which index you want the data to go into.

There are some exceptions for which it may make more sense to build a custom add-on from scratch. Some sources might not be as high quality or require significant modifications. For these cases, you can use the Splunk Add-on Builder to build a CIM-compatible add-on.

Why is normalized data important?

Splunk supports machine data from any data source in any format. Normalization helps to match your data to a common standard by using the same field names and event tags for equivalent events from different sources or vendors. This acts as a search-time schema while leaving the raw machine data intact.

Another benefit of normalization is that it allows you to present a unified view of a data domain using reports, correlation searches, or dashboards. Almost all of the existing reports, dashboards, and correlation searches included in Splunk Enterprise Security are designed to reference a data model rather than any specific index. This allows you to add any additional data from different vendors without having to redesign your searches and reports. Adding a new data source to a data model automatically includes the source in results.

Next steps

These resources might help you understand and implement this guidance:

- Getting data into ES
- Don't let security go up, up and away (in the clouds), start with data
- Modernizing your SOC for the cloud age starts with security foundations

Still need help with this use case? Most customers have OnDemand Services per their license support plan. Engage the ODS team at OnDemand-Inquires@splunk.com if you require assistance.