Data Onboarding Where Do I Begin?

Luke Netto
Staff Professional Services Consultant | Splunk

splunk> .conf19

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or plans of the company. We caution you that such statements reflect our current expectations and estimates based on factors currently known to us and that actual events or results may differ materially. The forward-looking statements made in the this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, it may not contain current or accurate information. We do not assume any obligation to update any forward-looking statements made herein.

In addition, any information about our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only, and shall not be incorporated into any contract or other commitment. Splunk undertakes no obligation either to develop the features or functionalities described or to include any such feature or functionality in a future release.

Who Are You?

You have Splunk installed, either in your datacenter or on your laptop

You have data you want to onboard into Splunk

Your data comes from syslog, wineventlog, custom modular inputs, and/or flat files such as .csv | .log | .json | .txt

Who Am I?

5+ years of Splunk experience

10+ years of systems engineering

7+ years of data analytics

systems engineering + data analytics = splunk>

Agenda

Why is this important?

Basic Splunk overview

What do I ask before I onboard data?

How Splunk makes it easier & faster

Creating your own sourcetype

Onboarding data (inputs.conf/props.conf)

The importance of normalizing (props.conf/eventtypes.conf/tags.conf)

Why is This Important?

Many Reasons!

1. Your organization wants to become data-driven
2. Data needs to be collected, accessible, and queryable
3. Decisions without **quality** data is simply guessing

In Data Science, **80%** of time spent **prepare data**, 20% of time spent complain about need for prepare data.”

@BigDataBorat

https://twitter.com/bigdataborat/status/1065963529918300 16

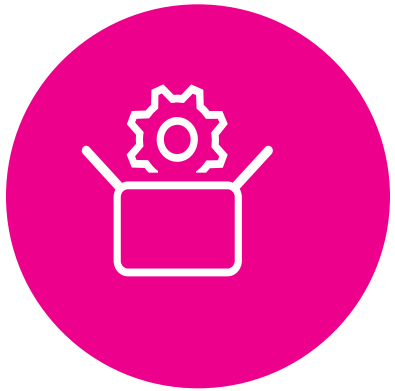
Data Life Cycle

What your executives want

Making Machine Data Accessible, Usable and Valuable to Everyone.

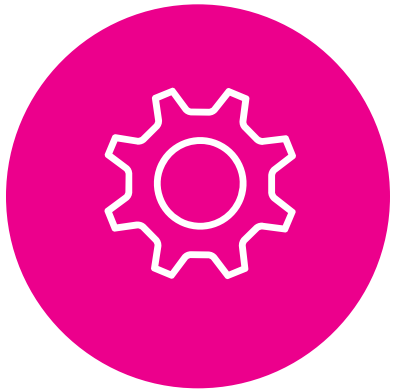
Reducing That 80%

How Splunk is helping you!



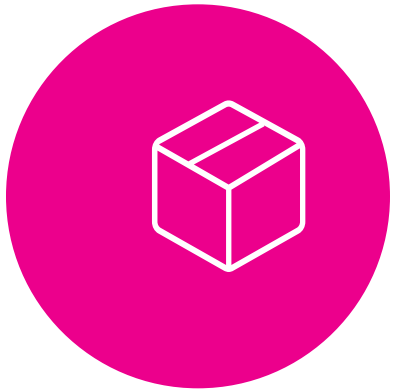
Splunkbase Apps & Add-ons

Guided Data Onboarding (GDO)



Splunk Connect

Splunk Data Stream Processor (DSP)



Splunk Add-on Builder

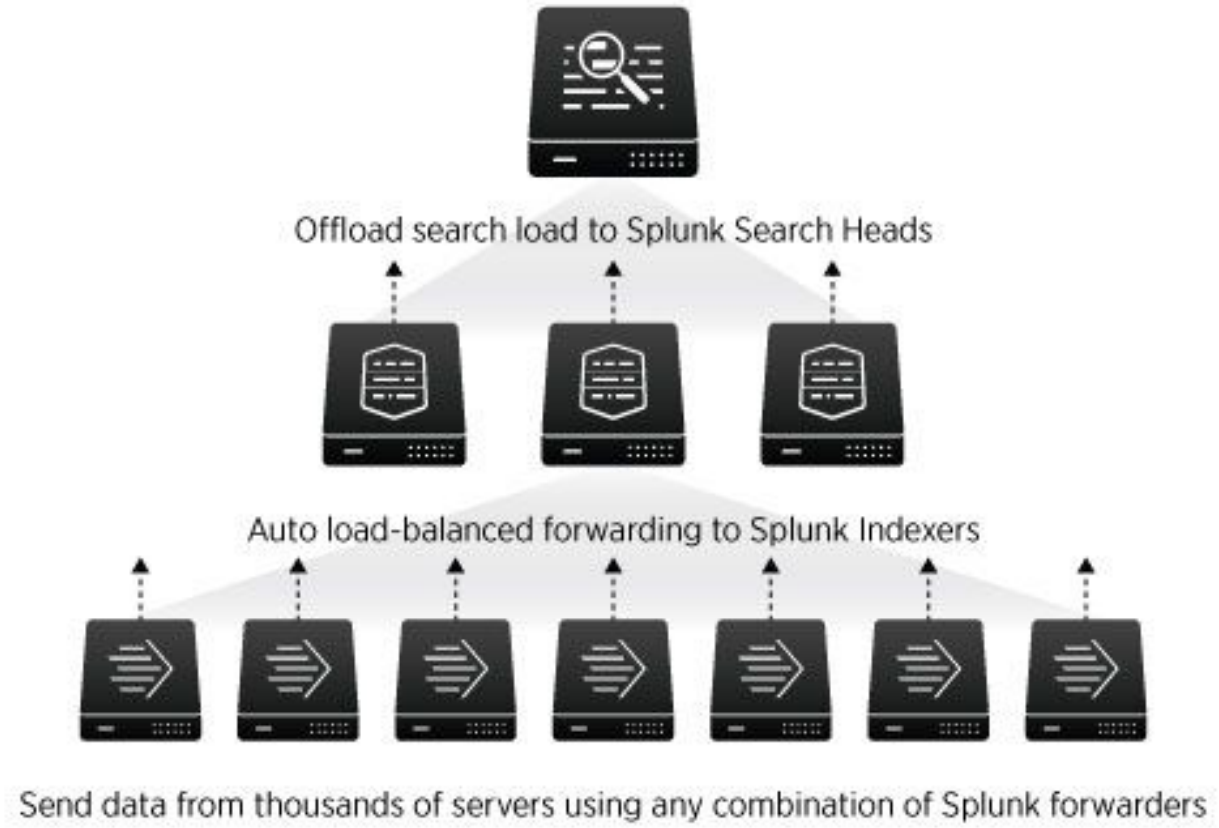
Splunk's Data Collection

Basic Architecture Refresher

splunk> .conf19

Basic Architecture

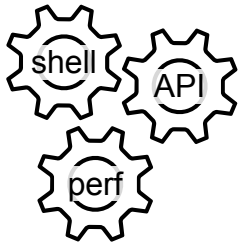
How Splunk works at a high level



What Can Splunk Ingest?

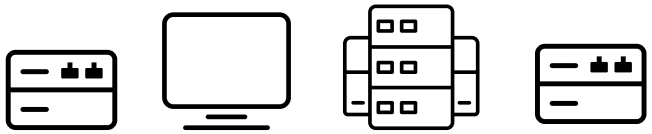
Data from anything!

API Data Sources
Heavy Forwarder



splunk>

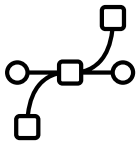
Local File Monitoring
Universal Forwarder



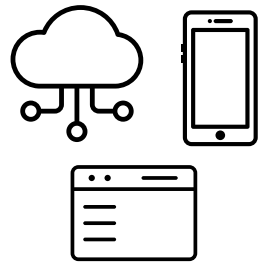
*nix | Windows | Mainframes | syslog

Event Logs, Active Directory, OS Stats
Unix, Linux and Windows hosts
Universal Forwarder

Wire Data
Splunk Stream
Universal Forwarder *or*
HTTP Event Collector



DevOps, IoT, Containers/Kubernetes, Applications
HTTP Event Collector
(Agentless)



splunk> .conf19

Let's Begin

Index Time

Metadata

Your Data's Data

- Index
- Host
- Source
- Sourcetype
- Date/Timestamp
- Event Boundary/Line Breaker

Index

The repository for data

When the Splunk platform indexes raw data, it transforms the data into searchable events

Indexes reside in flat files on the indexer

Used to organize your data, role-based access control (RBAC), and retention

- winlog, winlogsec, winperf, nixlog, nixlogsec, nixperf
- firewall, proxy, dhcp, dns, dblog, web, app

Host

A default field that contains the hostname or IP address of the device that generated the event

Use the host field in searches to narrow the search results to events that originate from a specific device

Allows you to locate the originating device

If you are forwarding data from many hosts through a single host, you may need to create a new ”indexed field” called orig_host or something similar

Source

A default field that identifies the source of an event (where the event originated)

For data monitored from files and directories, the source consists of the full pathname of the file or directory

- /var/log/messages
- /var/log/messages.1
- /var/log/secure

For network-based sources, the source field consists of the protocol and port

- UDP:514
- TCP:1514
- HTTP:<token name>

Timestamp

Splunk uses timestamps to

- correlate events by time
- create the timeline histogram in Splunk Web
- set time ranges for searches

Usually automatic or defined by sourcetype (recommended)

Event Boundary/Line Breaker

Allows Splunk to break the incoming stream of bytes into separate events

Supports single-line and multi-line

Splunk can usually do this automatically or it can be defined by sourcetype (recommended)

Sourcetype

Defining your data's data

A default field that identifies the data structure of an event including datetime extraction and linebreaking

Determines how Splunk extracts & calculates fields during search time

Use the sourcetype field in searches to find all data of a certain type (as opposed to all data from a certain source)

Important – syslog, csv, json, xml are not sourcetypes!

Examples: cisco:asa, cisco:ios, pan:firewall, wineventlog, linux_secure

Source vs. Sourcetype

Events with the same sourcetype can come from different sources

- /var/log/messages
- /var/log/messages.1
- udp:514

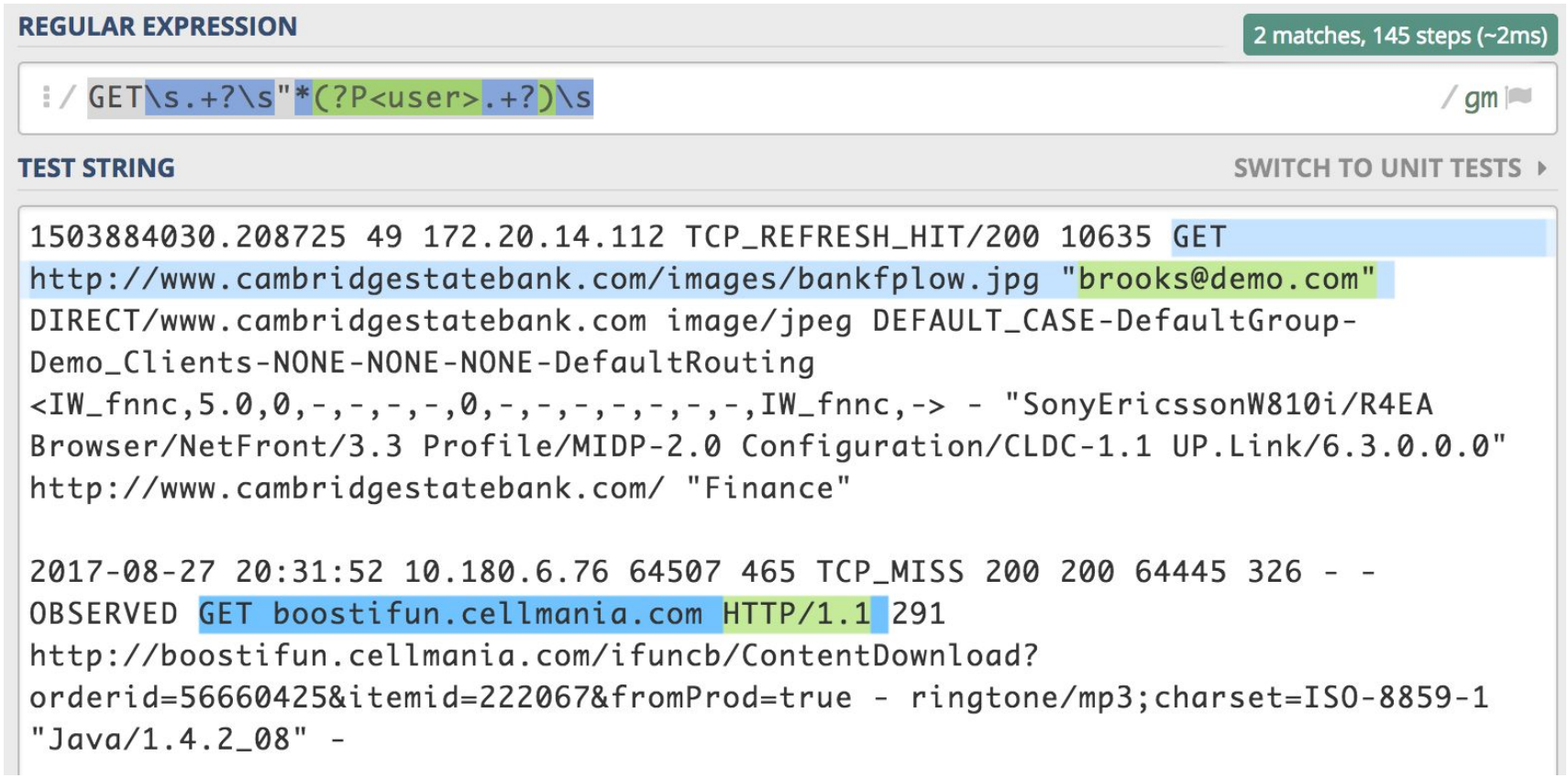
sourcetype=linux_messages_syslog may retrieve events from both of those sources

What Happens with Bad Sourcetypes

Same regex, same sourcetype, bad results ☹

Squid Proxy

Blue Coat ProxySG



With Correct Sourcetypes

Same regex, different sourcetypes, good results ☺

Cisco Squid

REGULAR EXPRESSION

1 match, 95 steps (~3ms)

/ GET\s.+?\s"*(?P<user>.+?)\s / gm

TEST STRING

SWITCH TO UNIT TESTS

```
1503884030.208725 49 172.20.14.112 TCP_REFRESH_HIT/200 10635 GET
http://www.cambridgestatebank.com/images/bankfplow.jpg
"brooks@demo.com" DIRECT/www.cambridgestatebank.com image/jpeg
DEFAULT_CASE-DefaultGroup-Demo_Clients-NONE-NC
<IW_fnnc,5.0,0,-,-,-,-,0,-,-,-,-,-,-,-,IW_fnnc
"SonyEricssonW810i/R4EA Browser/NetFront/3.3 P
Configuration/CLDC-1.1 UP.Link/6.3.0.0.0"
http://www.cambridgestatebank.com/ "Finance"
```

Bluecoat

REGULAR EXPRESSION

1 match, 48 steps (~1ms)

/ GET\s.+?\s"*(?P<protocol_version>.+?)\s / gm

TEST STRING

SWITCH TO UNIT TESTS

```
2017-08-27 20:31:52 10.180.6.76 64507 465 TCP_MISS 200 200 64445 326 -
- OBSERVED GET boostifun.cellmania.com HTTP/1.1 291
http://boostifun.cellmania.com/ifuncb/ContentDownload?
orderid=56660425&itemid=222067&fromProd=true -
ringtone/mp3;charset=ISO-8859-1 "Java/1.4.2_08" -
```

Configuring the Metadata

- props.conf

```
[your:sourcetype]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =

SHOULD_LINEMERGE = false
LINE_BREAKER =
EVENT_BREAKER_ENABLE = true
EVENT_BREAKER =


TRUNCATE = 10000
```

- inputs.conf

```
[monitor:///your/logs]
host = <optional/depends>
index = <your index>
sourcetype = your:sourcetype
source = <optional/depends>
```

host & source are usually automagically set

Time to Explore!

Everything else happens at search!

Start exploring and understanding your data

Apply different lenses to your data for use cases not yet discovered

Maintaining the original log ensures integrity

Data Discovery

Using what we now know, what should we ask before data onboarding?

splunk> .conf19

Find the Data

What is producing the data?

- Appliance
- Application

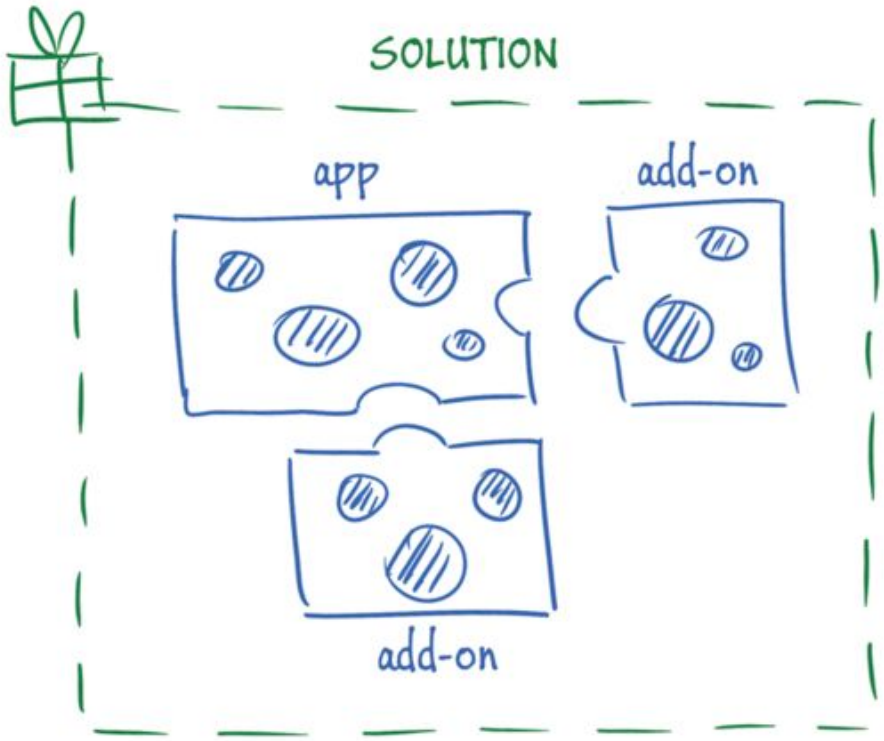
Where is the data?

- Flat file
- Network/Syslog feed
- REST API
- Database
- Wineventlog

Can you get a sample?

Is there a logging document or manual? – these sometimes do exist!

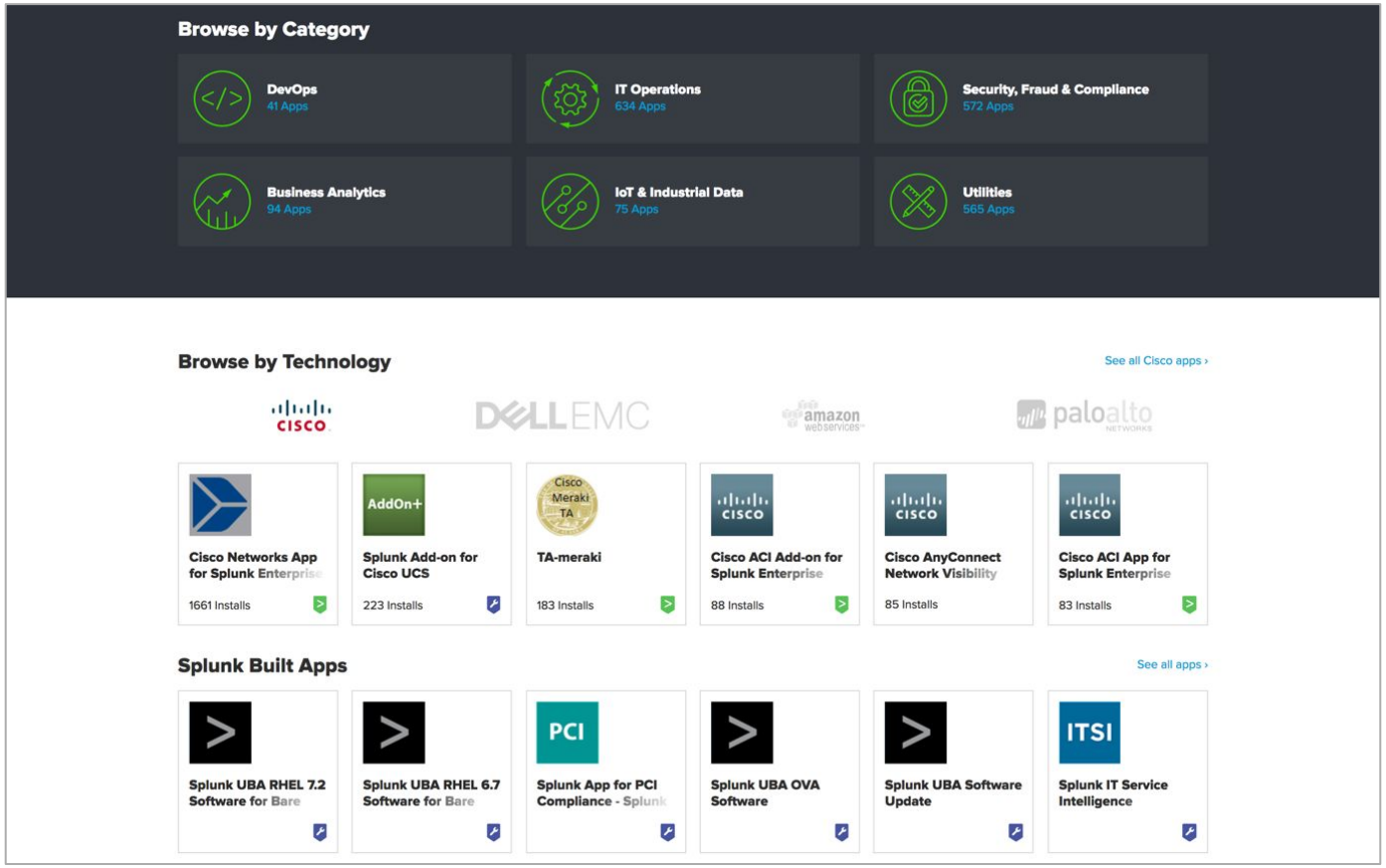
Apps & Add-Ons (TA's)



Your **first** choice when onboarding new data

Includes relevant config files (props/transforms) and ancillary scripts & binaries

Where Do You Get Apps? Splunkbase!



splunk> .conf19

What If There is No App/Add-On?

Let's Do It

Getting Data In!

Data Discovery

Understand the data source

Your networking organization wants to bring their firewall logs into Splunk

Ask yourself…

- Are the logs coming from an appliance or application?
- How will I access the logs?
- Who needs access to the logs in Splunk?
- How long do I need to retain the logs?
- Can I get a log sample?

Data Discovery

Our answers…

Your network team uses Cisco ASA and Palo Alto firewalls

Both firewalls are appliances, outputting their logs using syslog

The networking team also has a syslog server

Your networking team is the only team that needs access until your security team becomes aware of the data in Splunk (same data, difference lenses, multiple use-cases)

Your organization would like 12 months of logs for any network related equipment for compliance purposes

A log sample has been provided

Questions For Us

What index do I use?

Do you have any existing indexes for the network team?

- Yes, but all of their other data is in there and security will probably want to see the firewall logs

Do their existing indexes have the required retention?

- No, we keep their other logs for only 90 days

Sounds like we need a new index…maybe “firewall”?

Questions For Us

How am I going to receive the data?

Leverage the networking team's existing syslog server?

Use my own syslog server?

Use Splunk Connect for Syslog?

Any answer is acceptable!

Questions For Us

What sourcetype do I use?

Are there any existing apps on Splunkbase? YES!

Source type	Description
`cisco:asa`	The system logs of Cisco ASA record user authentication, user session, VPN, and intrusion messages.
`cisco:fwsm`	The system logs of Cisco FWSM record user authentication, user session, and firewall messages.
`cisco:pix`	The system logs of Cisco PIX record user authentication, user session, and intrusion messages.

Log source	SourceType
Only Firewall logs	pan:firewall
Only Traps Management Service logs	pan:traps
Only Traps 4.x logs	pan:traps4
This input receives Firewall and Traps logs	pan:log

Read the docs, we did most of the 80% already for you!

What Are Our Inputs?

Making some assumptions here…

[monitor:///var/log/cisco_asa/<host>]

index = firewall

sourcetype = cisco:asa

disabled = false

host_segment = 4

[monitor:///var/log/pan_firewall/<host>]

index = firewall

sourcetype = pan:firewall

disabled = false

host_segment = 4

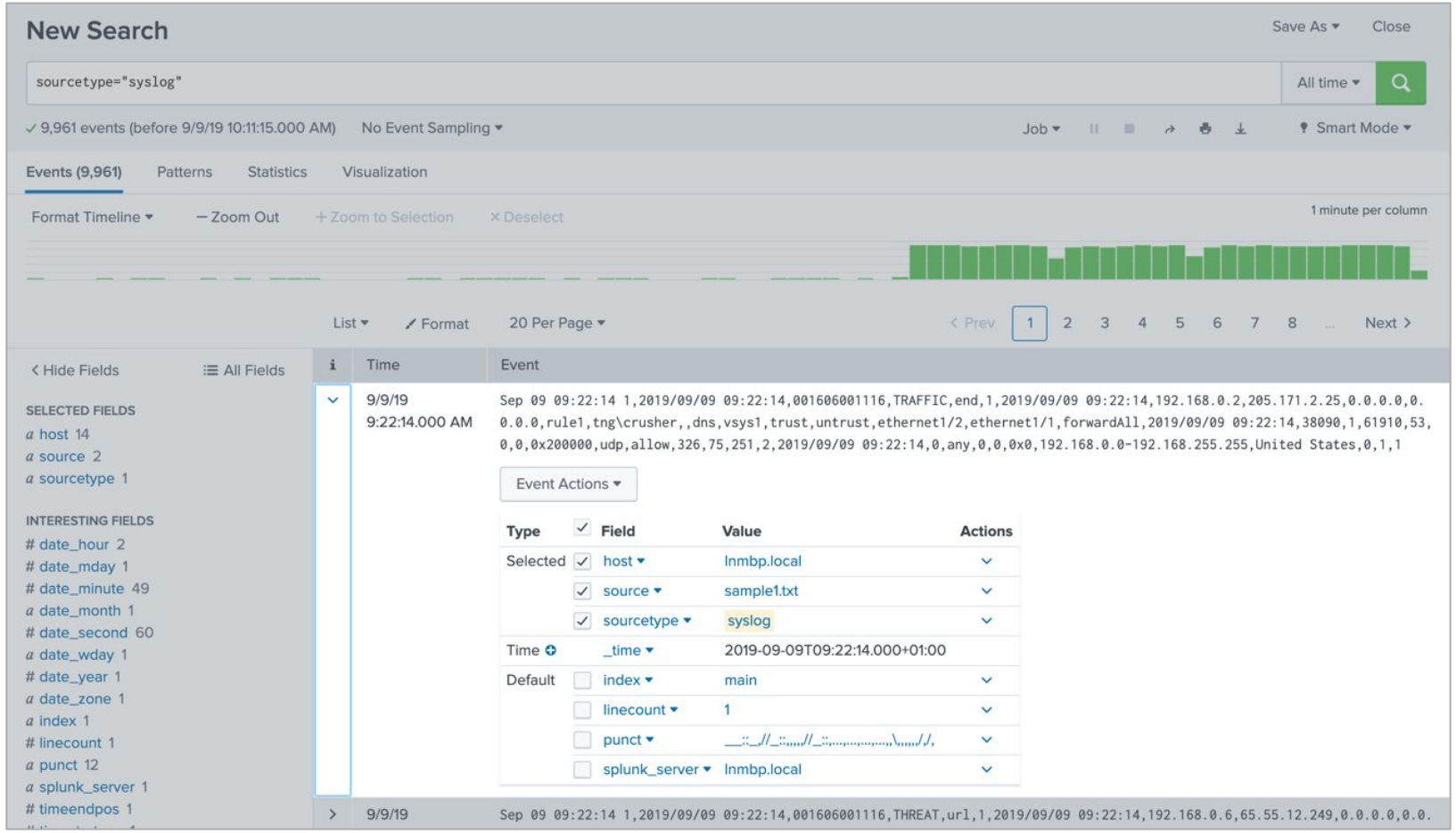
Value Extraction

Exploring Our Data!

splunk> .conf19

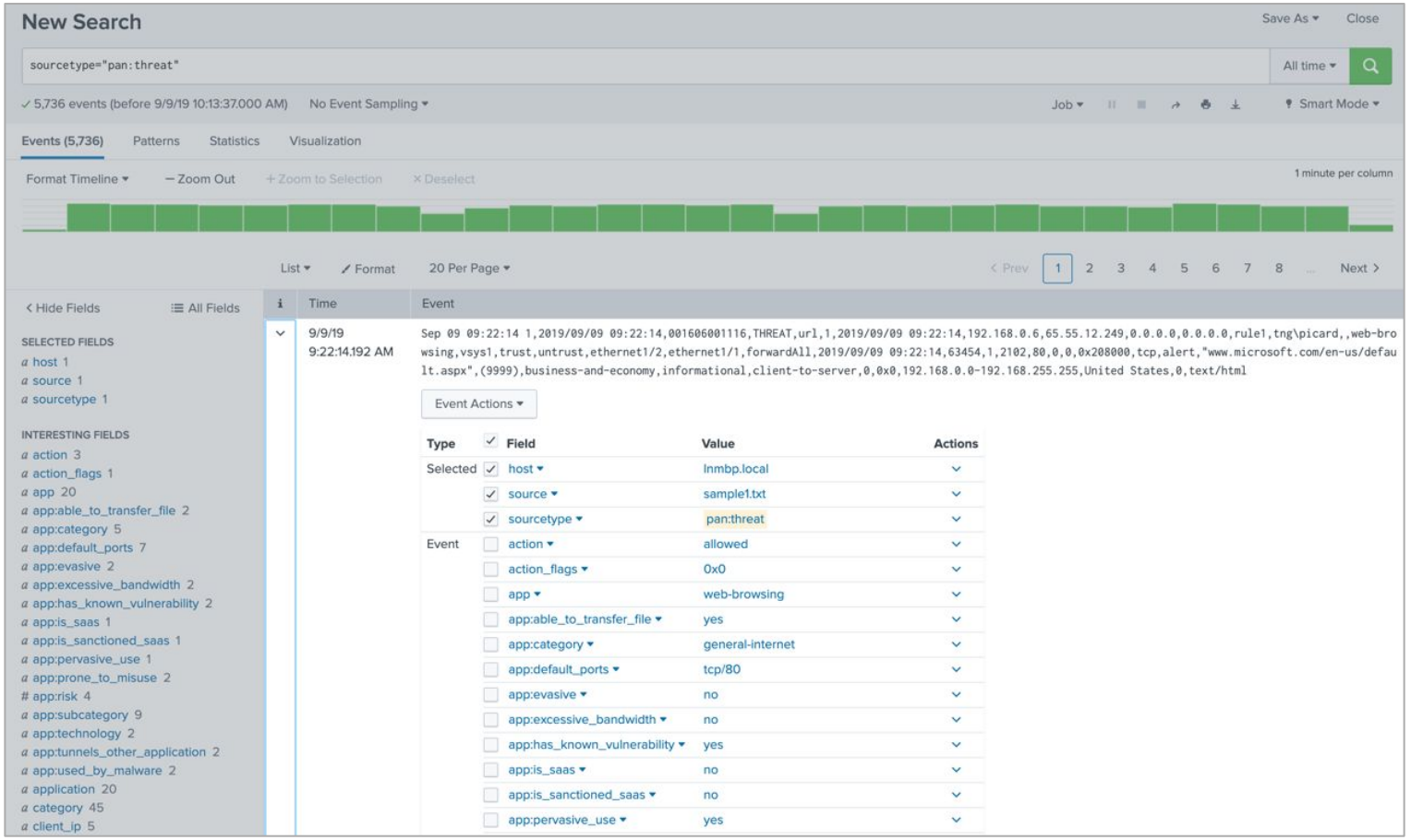
If We Used the Incorrect Sourcetype

We would have a lot of work ahead of us ☹



Using the Add-On: Palo Alto

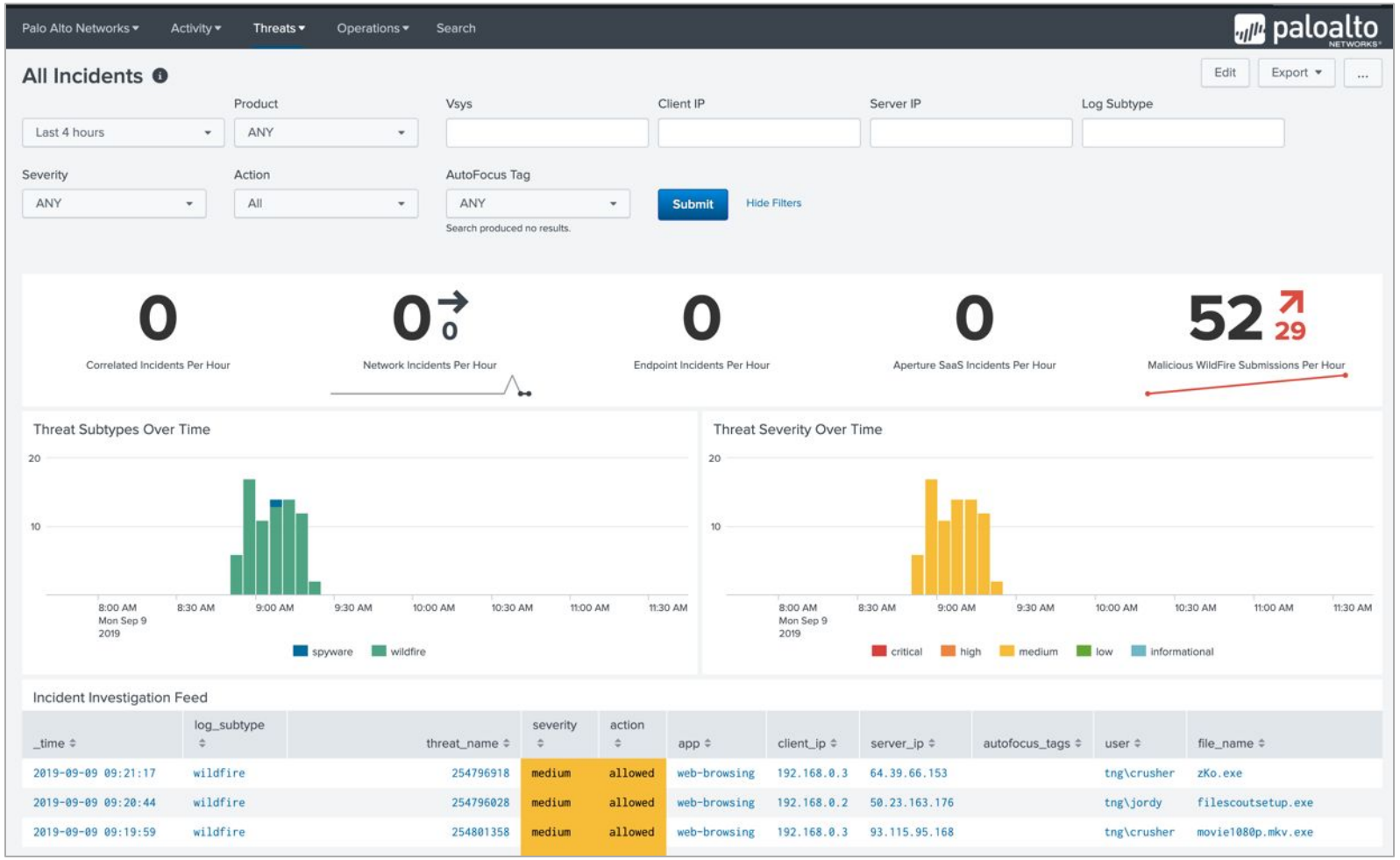
https://splunkbase.splunk.com/app/2757/



splunk> .conf19

Using the App: Palo Alto

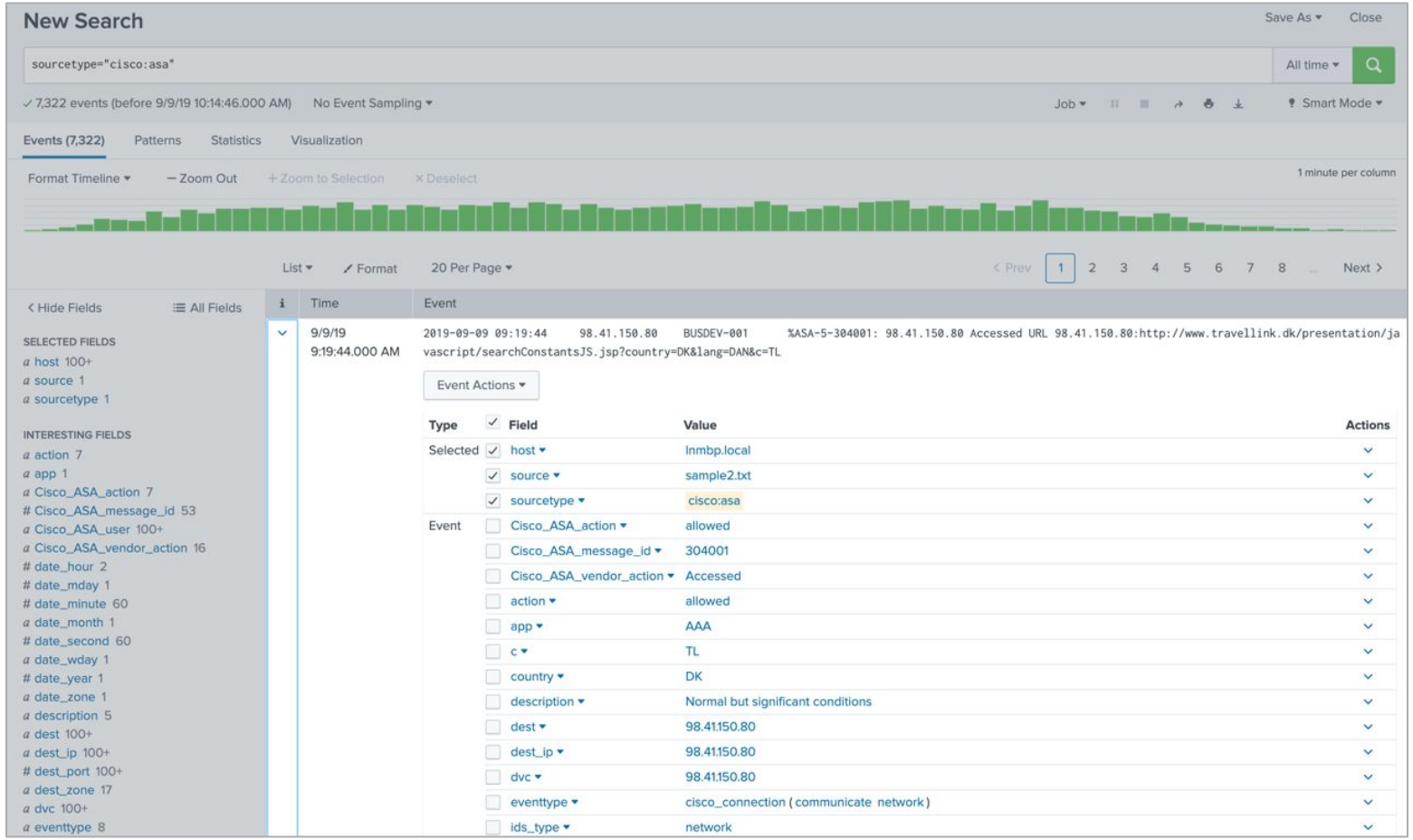
https://splunkbase.splunk.com/app/491/



splunk> .conf19

Using the Add-On: Cisco ASA

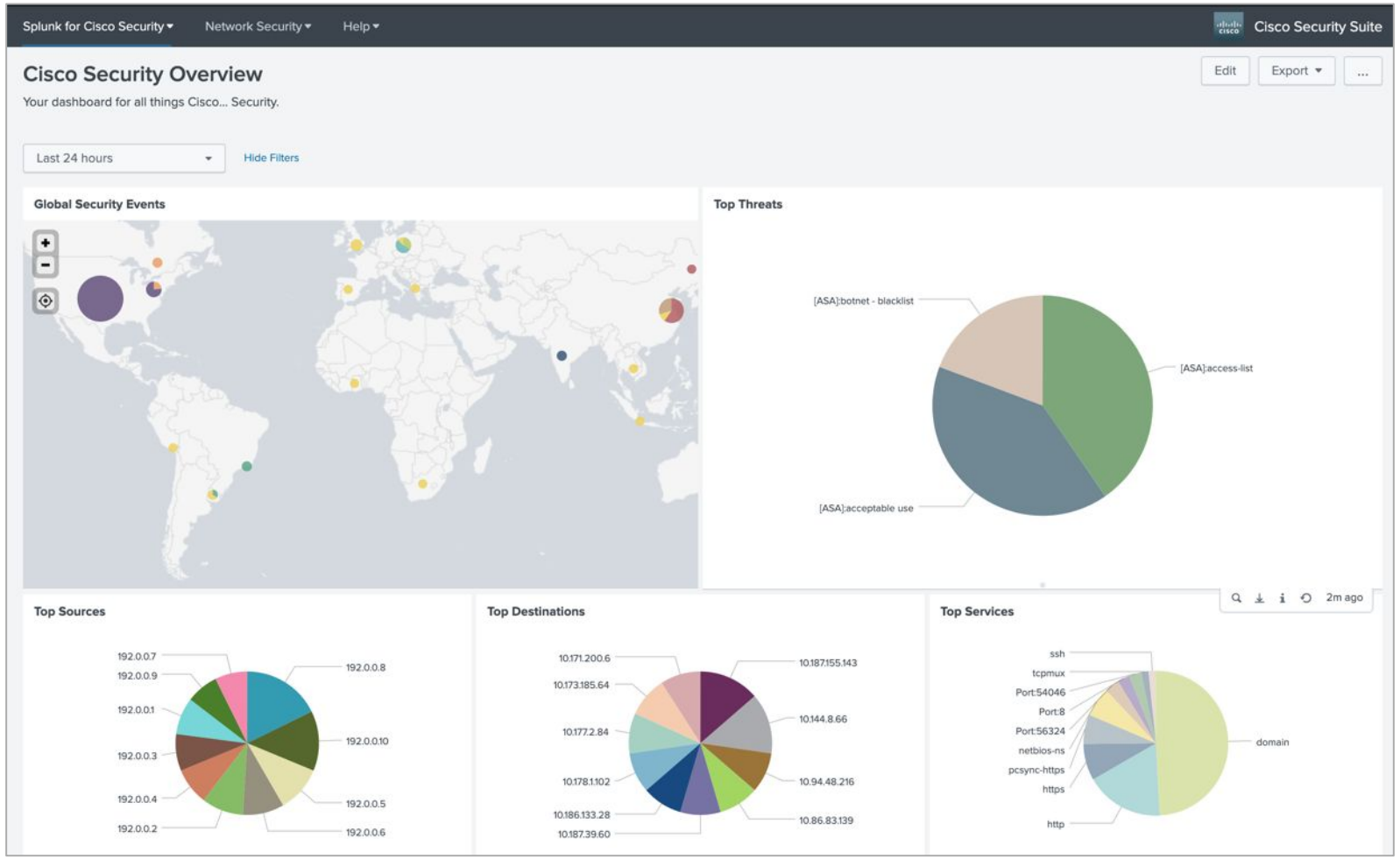
https://splunkbase.splunk.com/app/1620/



splunk> .conf19

Using the App: Cisco Security Suite

https://splunkbase.splunk.com/app/525/



splunk> .conf19

What If There is No Add-on?

That’s OK!

Data Discovery

Understand the data source

Your networking organization wants another set of logs into Splunk

Ask yourself…

- Are the logs coming from an appliance or application?
- How will I access the logs?
- Who needs access to the logs in Splunk?
- How long do I need to retain the logs?
- Can I get a log sample?

Data Discovery

Our answers…

You discovered these are logs from a new application called “Acme Photon”

The logs are written to a directory on the application server

There is no add-on available on Splunkbase (yet)

Sample logs and the complete file path have been provided

/opt/acme/photon/audit.log

/opt/acme/photon/db.log

/opt/acme/photon/scan.log

/opt/acme/photon/web.log

What Are the Sourcetypes?

https://docs.splunk.com/Documentation/AddOns/released/Overview/Sourcetypes

Naming format: vendor:product:technology:format

/opt/acme/photon/audit.log

- acme:photon:audit

/opt/acme/photon/db.log

- acme:photon:db

/opt/acme/photon/scan.log

- acme:photon:scan

/opt/acme/photon/web.log

- acme:photon:web

Sample Logs

2019-09-24 13:04:04 device=something1 user login (paynealbert)
result="<code>1</code>" from 192.168.1.10.136:51750 to 192.168.10.1:443
uri="https://acme.com/"

2019-09-24 13:04:06 device=something2 user login (steven81)
result="<code>2</code>" from 192.168.1.14:10607 to 192.168.10.2:443
uri="https://acme.com/"

Configuring the Metadata

- props.conf

```
[acme:photon:audit]
TIME_PREFIX = ^
MAX_TIMESTAMP_LOOKAHEAD = 19
TIME_FORMAT = %Y-%m-%d %H:%M:%S

SHOULD_LINEMERGE = false
LINE_BREAKER = ([\r\n]+)
EVENT_BREAKER_ENABLE = true
EVENT_BREAKER = ([\r\n]+)

TRUNCATE = 10000
```

- inputs.conf

```
[monitor:///opt/acme/photon/audit.log]
index = myindex
sourcetype = acme:photon:audit
```

host & source are usually automagically set

Where are the Fields?

auto vs custom

2019-09-24 13:04:04 device=something1 user login (paynealbert) result="<code>1</code>" from 192.168.1.10.136:51750 to 192.168.10.1:443 uri="https://acme.com/"

- _time = 2019-09-24 13:04:04 TIME_PREFIX, TIME_FORMAT, etc.
- device = something1 auto extracted
- user = paynealbert ???
- result = <code>1</code> auto extracted
- result_code = 1 ???
- src_ip = 192.168.1.10.136 ???
- src_port = 51750 ???
- dest_ip = 192.168.10.1 ???
- dest_port = 443 ???

Where are the Fields?

Custom extractions

```
[acme:photon:audit]

# continued…

EXTRACT-username = user\slogin\s\((?P<username>[^\)]+)\)

EXTRACT-result_code = result="<code>(?P<result_code>\d+)\<

EXTRACT-src_ip_port = from\s(?P<src_ip>.+?):(?P<src_port>\d+)\s

EXTRACT-dest_ip_port = to\s(?P<dest_ip>.+?):(?P<dest_port>\d+)\s
```

Common Information Model (CIM)

http://docs.splunk.com/Documentation/CIM/latest/User/Overview

A way of normalizing your data for maximum efficiency at search time

Splunk Certified TA's typically include necessary normalizations

Allows end-users to search using common fields such as "user" across many sourcetypes

Extract your fields, normalize, then tag your data

CIM Data Models

- Alerts
- Application State
- Authentication
- Certificates
- Databases
- Data Loss Prevention
- Email
- Interprocess Messaging
- Intrusion Detection
- Inventory
- Java Virtual Machines
- Malware
- Network Resolution (DNS)
- Network Sessions
- Network Traffic
- Performance
- Ticket Management
- Updates
- Vulnerabilities
- Web

What Do You Do?

Making your data usable

Already have a proper sourcetype

Extract your fields

Create field aliases

- username AS user
- src_ip as src, dest_ip as dest

Create calculations

- action=if(action="OK","success","failure")

Create tags

Use the Data Models as a guide

Fields for Authentication event datasets

The following table lists the extracted and calculated fields for the event datasets in the model. The table does not include any inherited fields. For more information, see How to use these reference tables.

Dataset name	Field name	Data type	Description	Expected values
Authentication	`action`	string	The action performed on the resource.	`success`, `failure`
Authentication	`app`	string	The application involved in the event (such as `ssh`, `splunk`, `win:local`).	
Authentication	`dest`	string	The target involved in the authentication. You can **alias** this from more specific fields, such as `dest_host`, `dest_ip`, or `dest_nt_host`.	

Tags used with Authentication event datasets

The following tags act as constraints to identify your events as being relevant to this data model. For more information, see How to use these reference tables.

Dataset name	Tag name
Authentication	authentication
|____ Default_Authentication	default
|____ Insecure_Authentication	cleartext OR insecure
|____ Privileged_Authentication	privileged

Applying CIM

Normalizing your fields

```
[acme:photon:audit]

# continued…

FIELDALIAS-user = username AS user

FIELDALIAS-src = src_ip AS src

FIELDALIAS-dest = dest_ip AS dest

EVAL-action = case(result_code==1,”success”,result_code==2,”failure”,1=1,null())
```

Tagging Your Data

- eventtypes.conf

[acme_photon_audit_authentication]

search = sourcetype=acme:photon:audit “user login”

- tags.conf

[eventtype=acme_photon_audit_authentication]

authentication = enabled

Become Famous

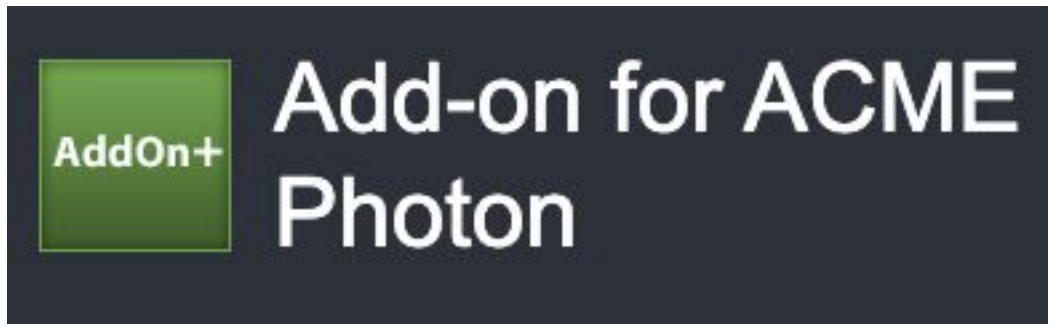
Submit your TA to Splunkbase ☺

Submit content to Splunkbase with the web UI

The user account that you use to submit your content becomes the account that owns the app. Log in to that user account to respond to questions and manage Splunkbase content.

1. Navigate to https://splunkbase.splunk.com/develop/ and click **Submit Your App**.
2. Review and agree to the Splunk Apps Developer Distribution License.

You are now on the Hosting page.



Recap

Don’t Forget!

Recap

Always set a sourcetype!

Don’t use syslog as a sourcetype!

Don’t use csv, json, xml as a sourcetype!

Use Splunkbase as a starting point!

Make your own sourcetype if you must, but use the naming format:
vendor:product:technology:format

Make the data usable – CIM!

Do You Have Bad Sourcetypes?

inputs.conf without a sourcetype defined

- | tstats count where index=* (sourcetype=*-* OR sourcetype=*too_small) by sourcetype

Don’t make the puppy sad ☹…

…so you can sleep good at night ☺

What's Next?

Splunk Enterprise Data Administration

- https://www.splunk.com/view/SP-CAAAPSE

Go see these sessions….

- SEC2280: Break Down Silos By Ingesting Multi-Purpose Data Sources into Splunk
- BA2766: Machine Data Alchemy: Transforming Digital Exhaust into Campus Gold
- FN2103: Exploratory Data Analysis on Aviation Safety Data
- FN1921: Saving the Nation's Food Supply with Data-Driven Analytics
- DEV1140: Next Generation Data Ingestion and Preparation with Splunk

Resources

https://docs.splunk.com/Documentation/Splunk/latest/Data/Whysourcetypesmatter

https://docs.splunk.com/Documentation/AddOns/released/Overview/Sourcetypes

https://www.splunk.com/blog/2012/08/10/sourcetypes-whats-in-name.html

https://www.splunk.com/blog/2010/02/11/sourcetypes-gone-wild.html

.conf19

splunk>

Thank You !

Go to the .conf19 mobile app to

RATE THIS SESSION