splunk> .conf2017

Data Onboarding

Where Do I begin?

Luke Netto | Senior Professional Services Consultant @ Splunk

September 26, 2017 | Washington, DC

Forward-Looking Statements

During the course of this presentation, we may make forward-looking statements regarding future events or the expected performance 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 could differ materially. For important factors that may cause actual results to differ from those contained in our forward-looking statements, please review our filings with the SEC.

The forward-looking statements made in this presentation are being made as of the time and date of its live presentation. If reviewed after its live presentation, this presentation may not contain current or accurate information. We do not assume any obligation to update any forward looking statements we may make. 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 functionality described or to include any such feature or functionality in a future release.

splunk> .conf2017

Who Are You?

- You have Splunk installed, either in your datacenter or your laptop
- You have data you want to onboard into Splunk
- Your data comes from syslog, wineventlog, and/or flat files such as .csv | .log | .json | .txt

Who Am I?

- 3+ years of Splunk experience
- 7+ years of systems engineering
- 5+ years of data analytics
- systems engineering + data analytics = Splunk

Agenda

- Importance
- Splunk Terms/Components
- Pre-onboarding/Data Discovery
- Splunkbase
- Creating your own sourcetype
- Onboarding – inputs.conf
- Optimizing for performance
- Normalizing

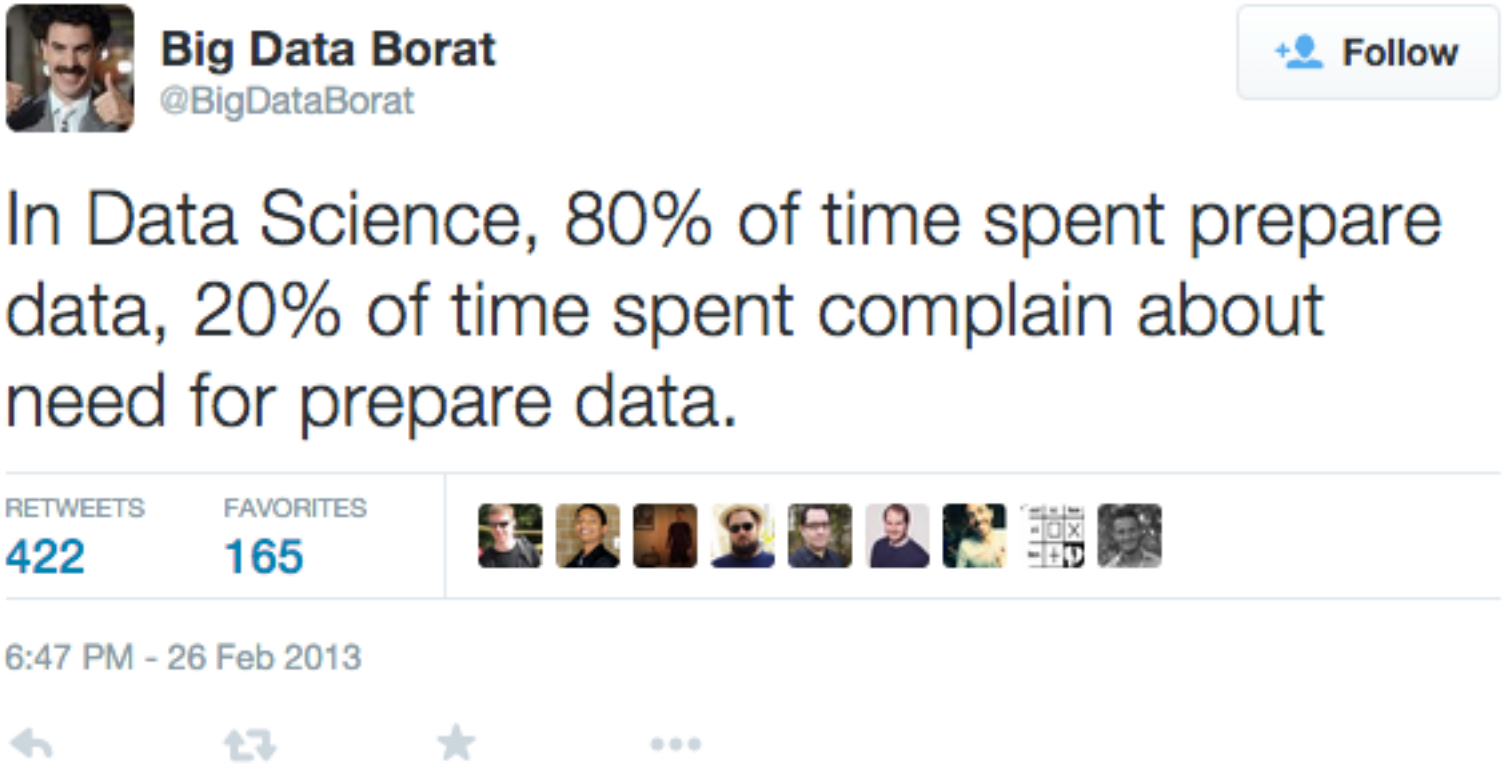
splunk> .conf2017

Why Is This Important?

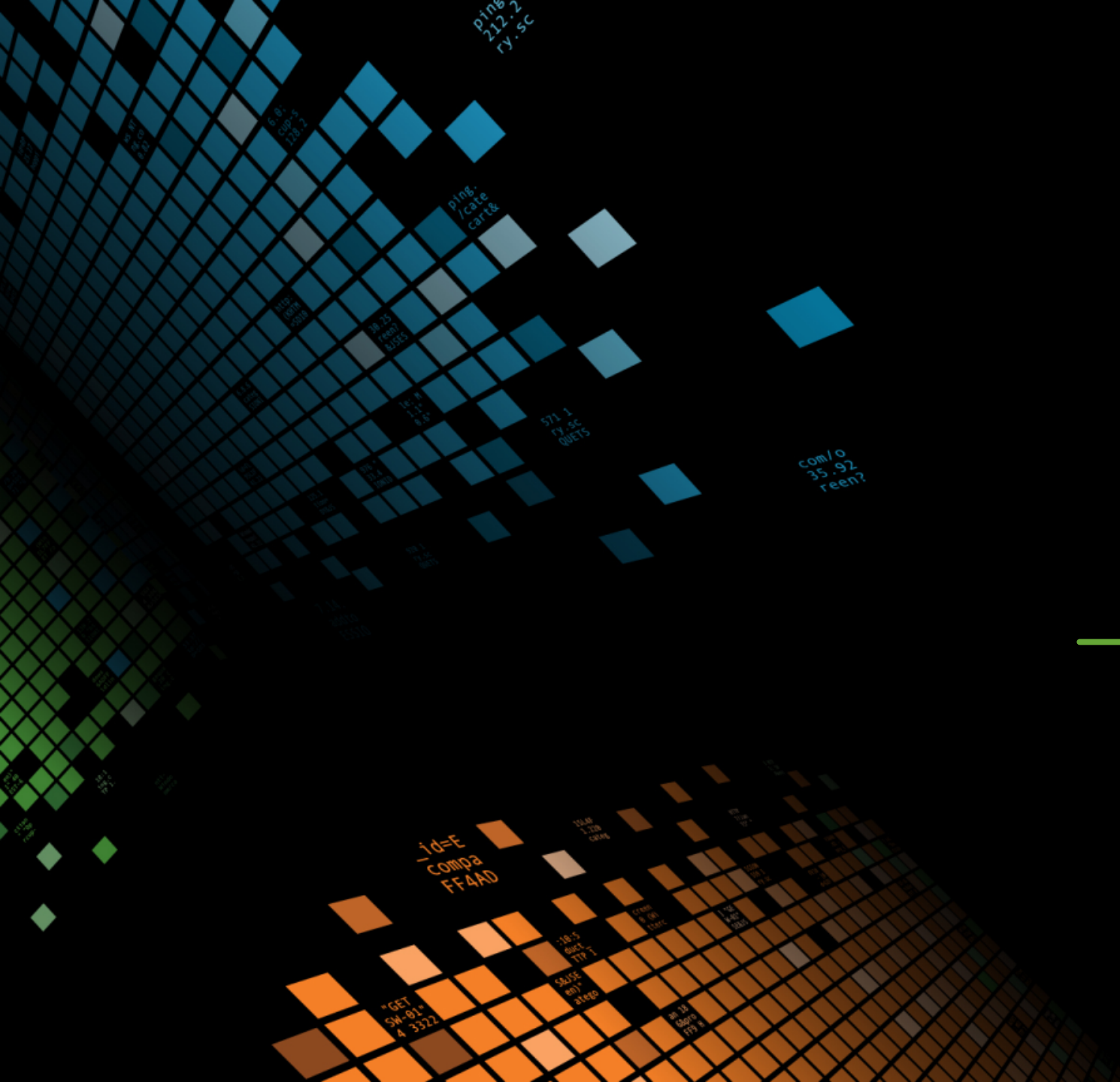
- Your organization wants to become data-driven
  - Data collection
  - Data access
- Decisions without ***quality*** data is simply guessing

Just Remember

Data preparation is 80%



What Your Executives Want

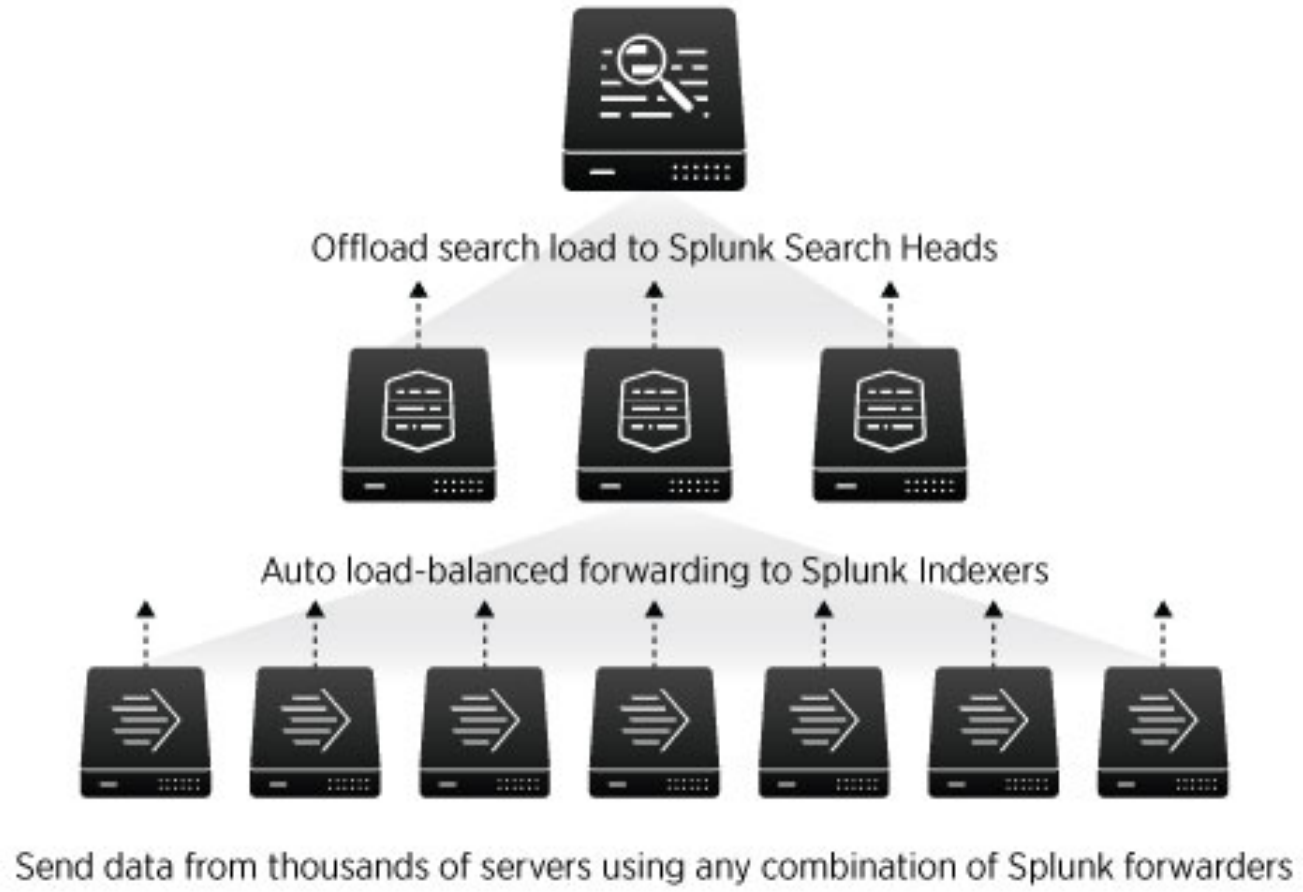


Splunk Data Collection Architecture

splunk> .conf2017

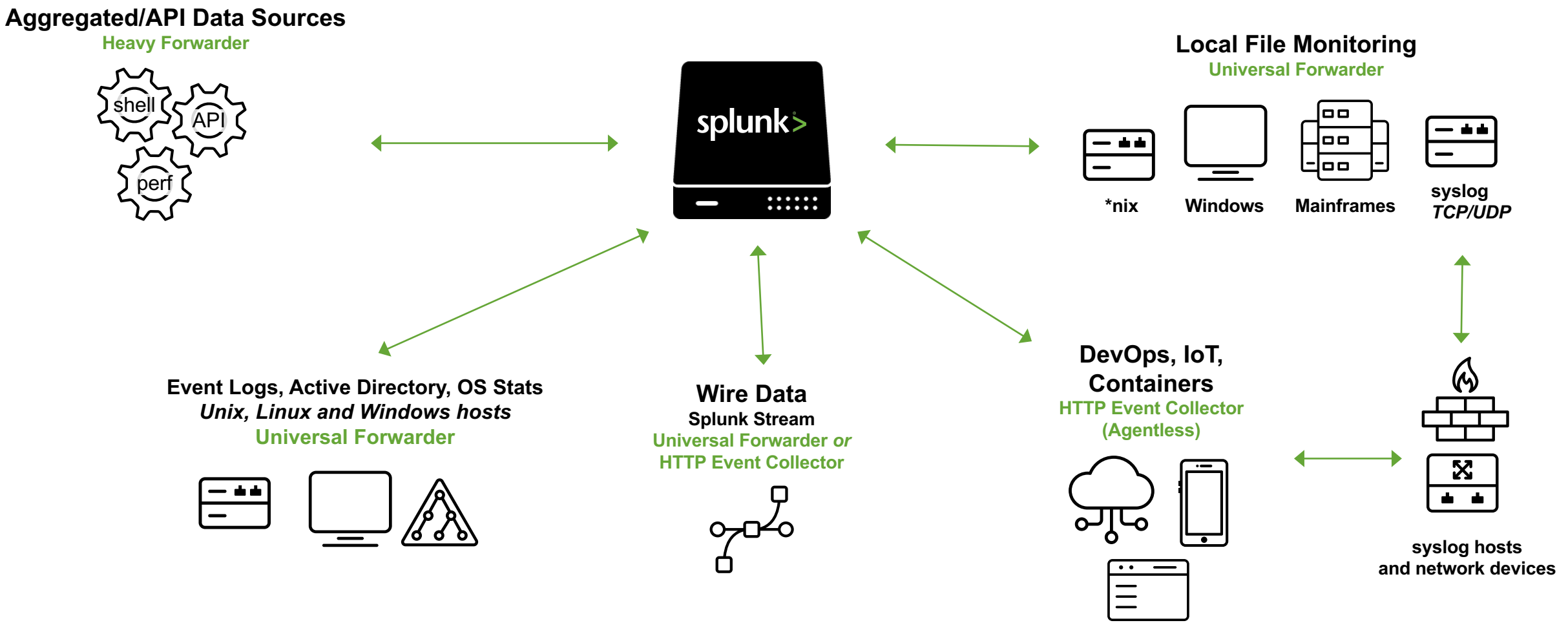
Basic Architecture Refresh

How Splunk works at a high level

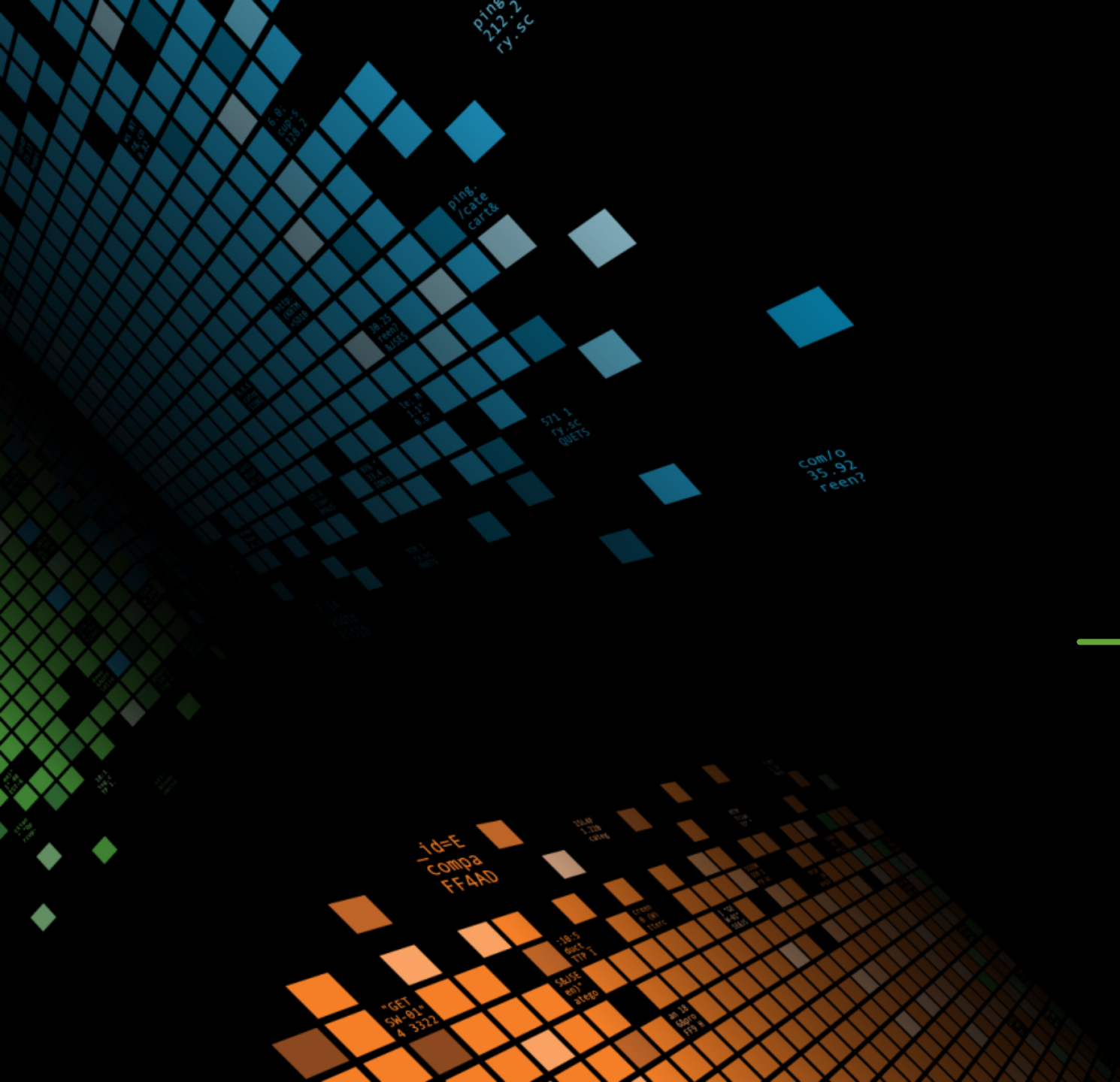


What Can Splunk Ingest?

Agent-Less and Forwarder Approach for Flexibility and Optimization

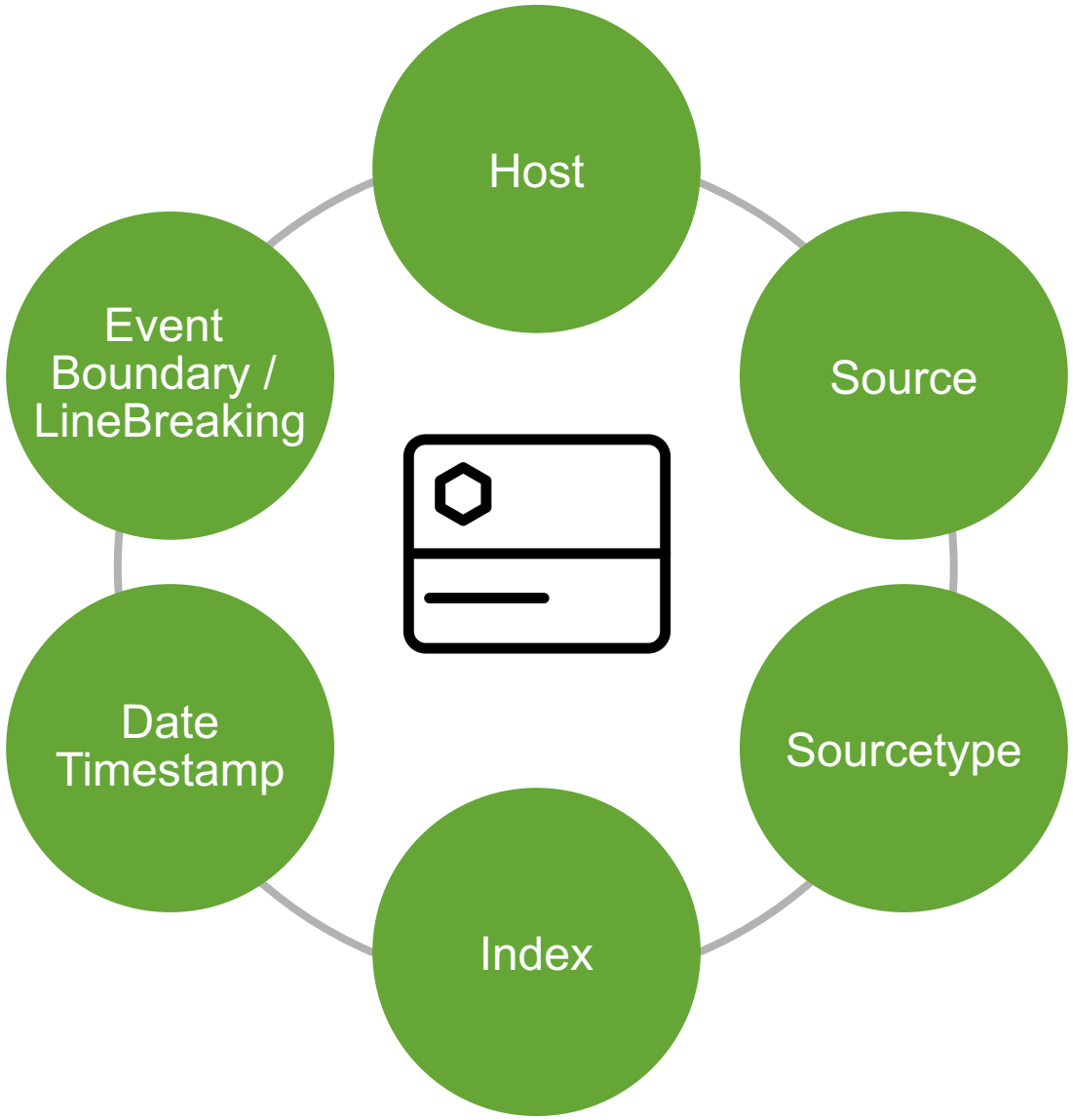


splunk> .conf2017



Default Fields

Six Things to Get Right at Index Time



Host

- A default field that contains the hostname or IP address of the network 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 located the originating device

Source

- A default field that identifies the source of an event, that is, 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

Sourcetype

- A default field that identifies the data structure of an event
- The format of the data input from which it originates
  - access_combined
  - cisco:asa
- 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!

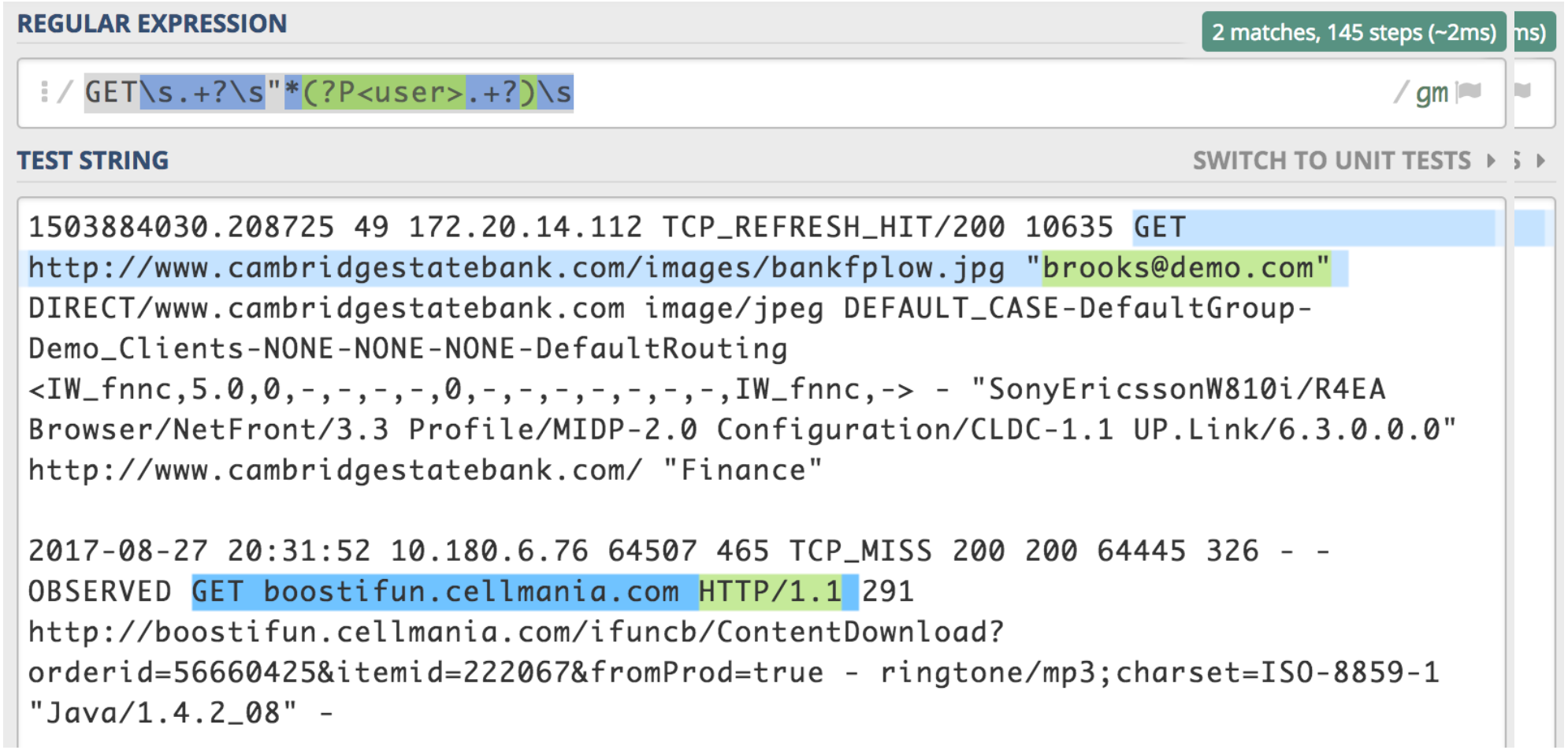
Source vs Sourcetype

- 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

Bluecoat Cisco Squid



splunk> .conf2017

Index

- The repository for data in Splunk Enterprise
- Indexes reside in flat files
- Similar to a folder
- Used for data access, retention, or organization

Timestamp

- Splunk uses timestamps to:
  - correlate events by time
  - create the timeline histogram in Splunk Web
  - set time ranges for searches
- Usually automatic

Line Break

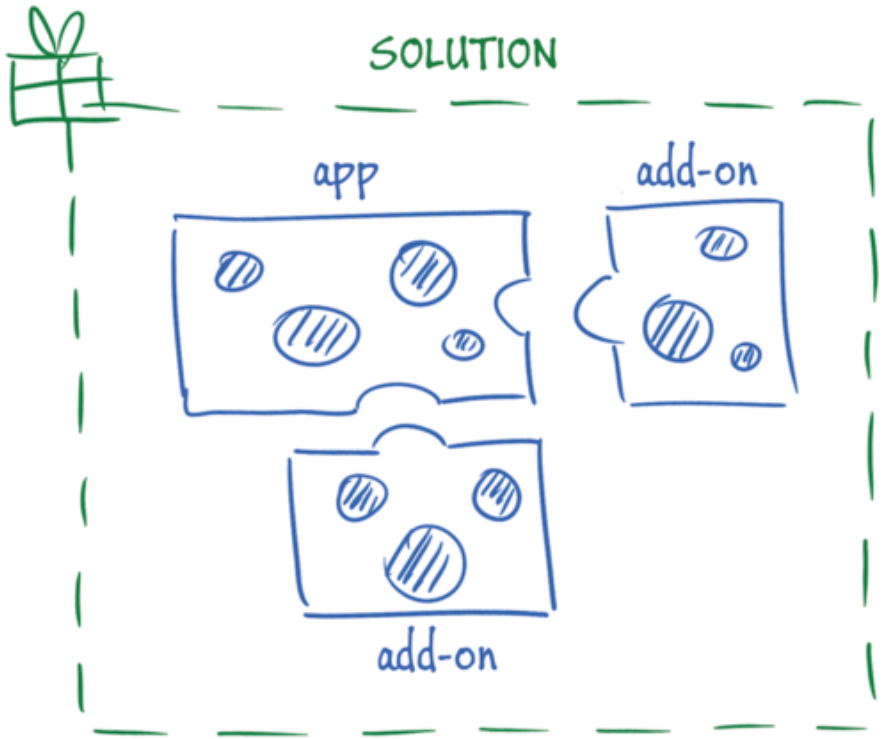
- Allows Splunk to break the incoming stream of bytes into separate events
- Supports single-line and multi-line
- Splunk can usually do this automatically

Data Discovery

Find the Data

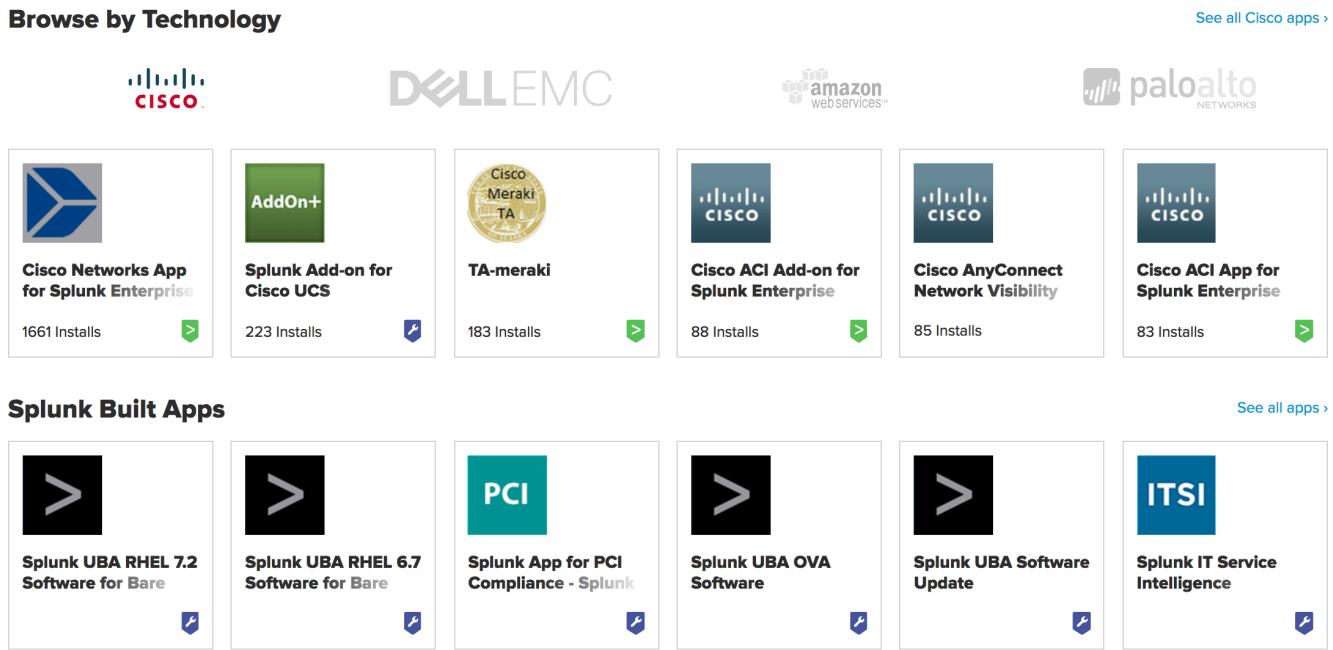
- What is producing the data?
  - Appliance
  - Application
- Where is the data?
  - Flat file
  - Network/Syslog feed
  - REST API
  - Database
  - Wineventlog
- Can you get a sample?

Apps & Add-ons (TA's)

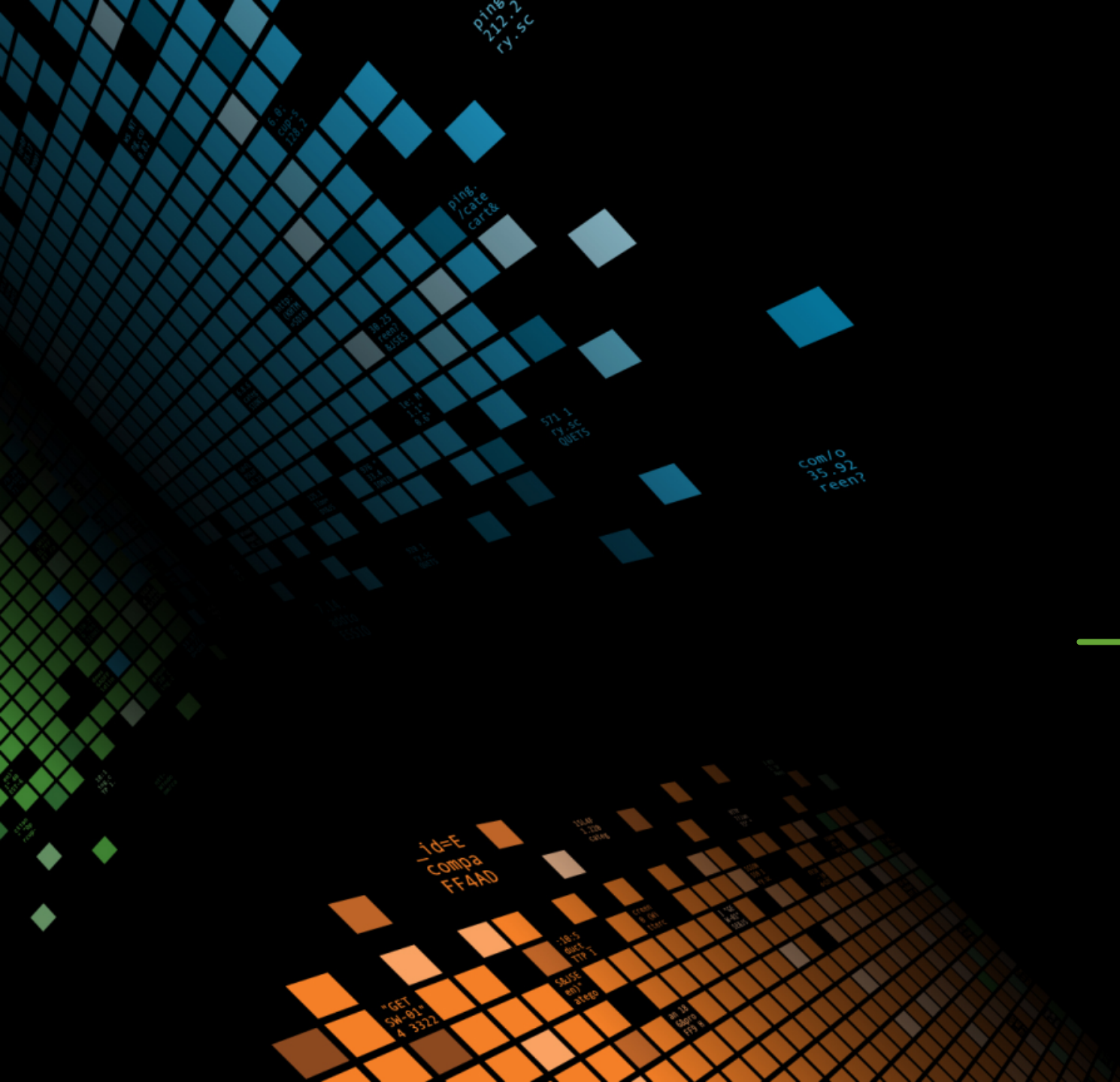


- Your *first* choice when onboarding new data
- Includes relevant config files (props/transforms) and ancillary scripts & binaries

Where do you get Apps? Splunkbase!



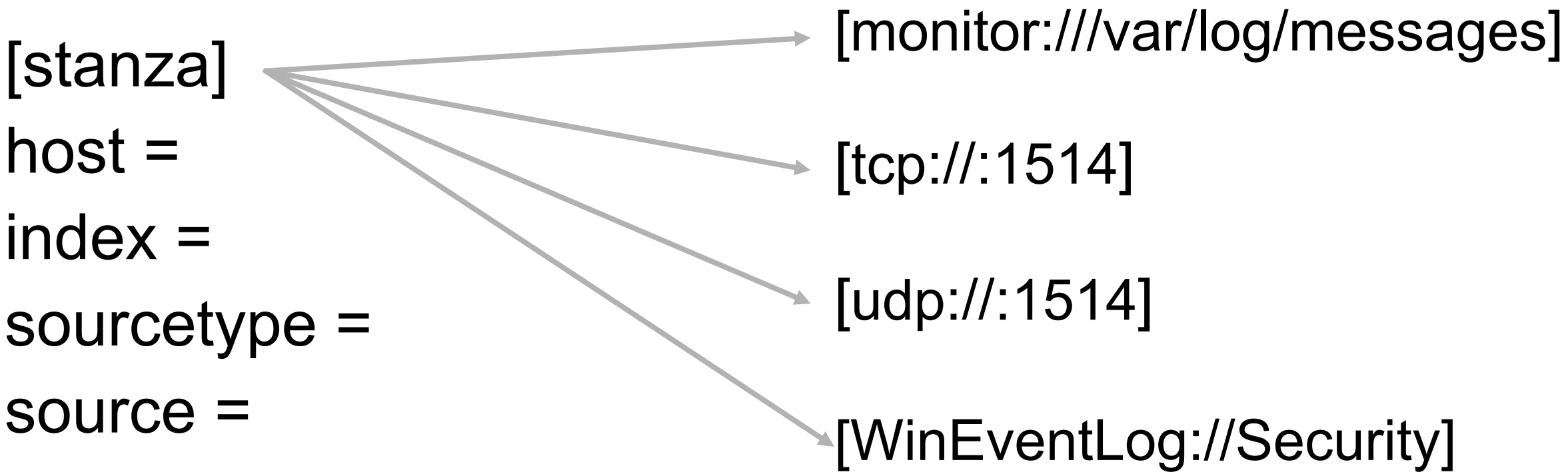
What If There Is No App?



Let’s Do It

inputs.conf

Tells Splunk to monitor a file, directory, listen on a port, etc.

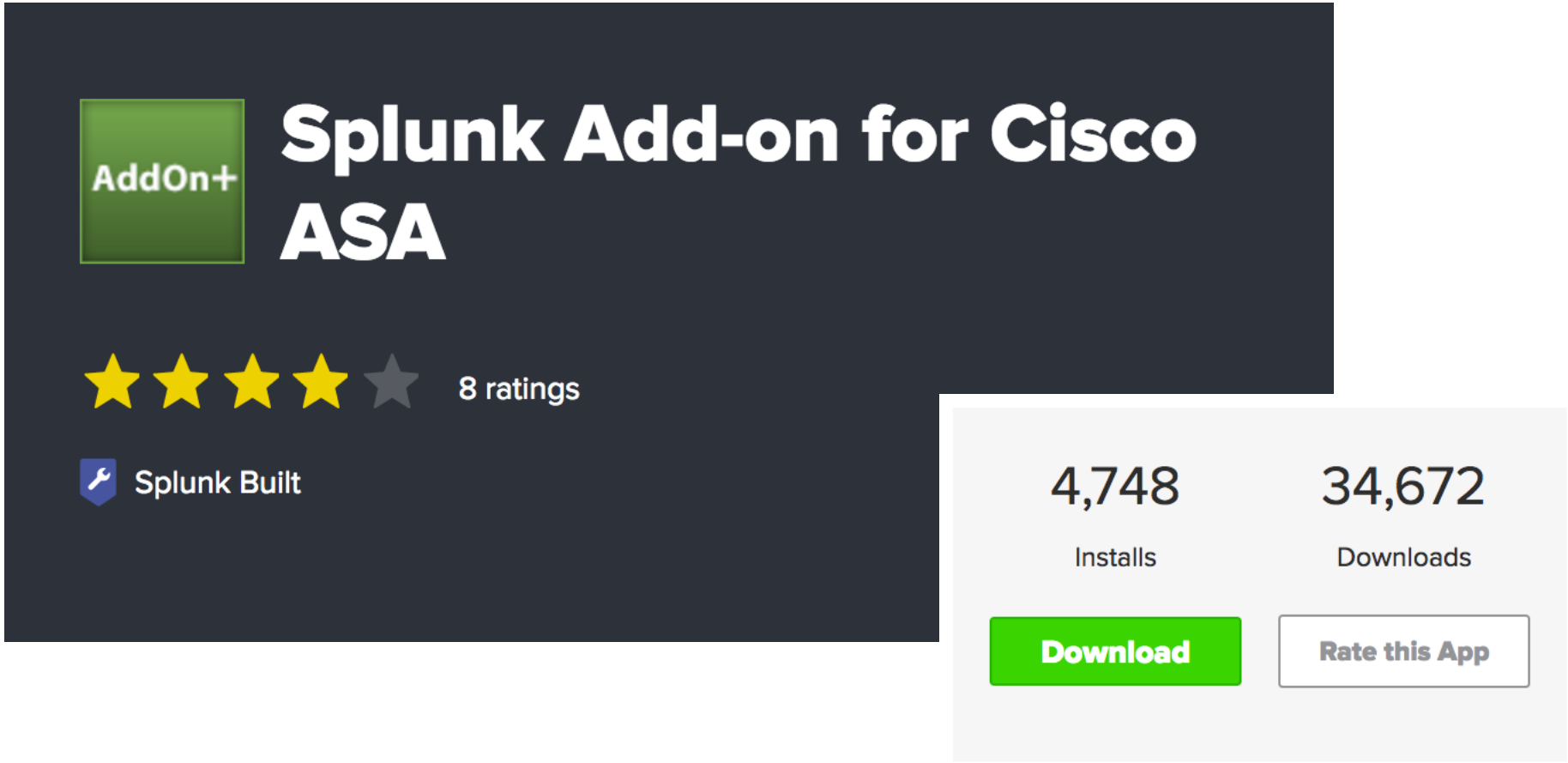


Example: With An App

- You have an appliance owned by the network team.
- The appliance is a Cisco ASA.
- Now what?

Is There An App?

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



Install the TA, typically on your Indexers and Search Heads.

Let's Make It Work

Not the best way…

- Make an inputs.conf.

```
[udp://:1515]
index = thenetworkindex
sourcetype = cisco:asa
```

- Where are the other fields?
  - They are automatic!

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.

Why Is It Not the Best Way…

- Disruption of Data Collection
- Splunk Metadata
- Syslog-ng/Rsyslog Flexibility

The Better Way

Use syslog-ng or rsyslog with a UF.

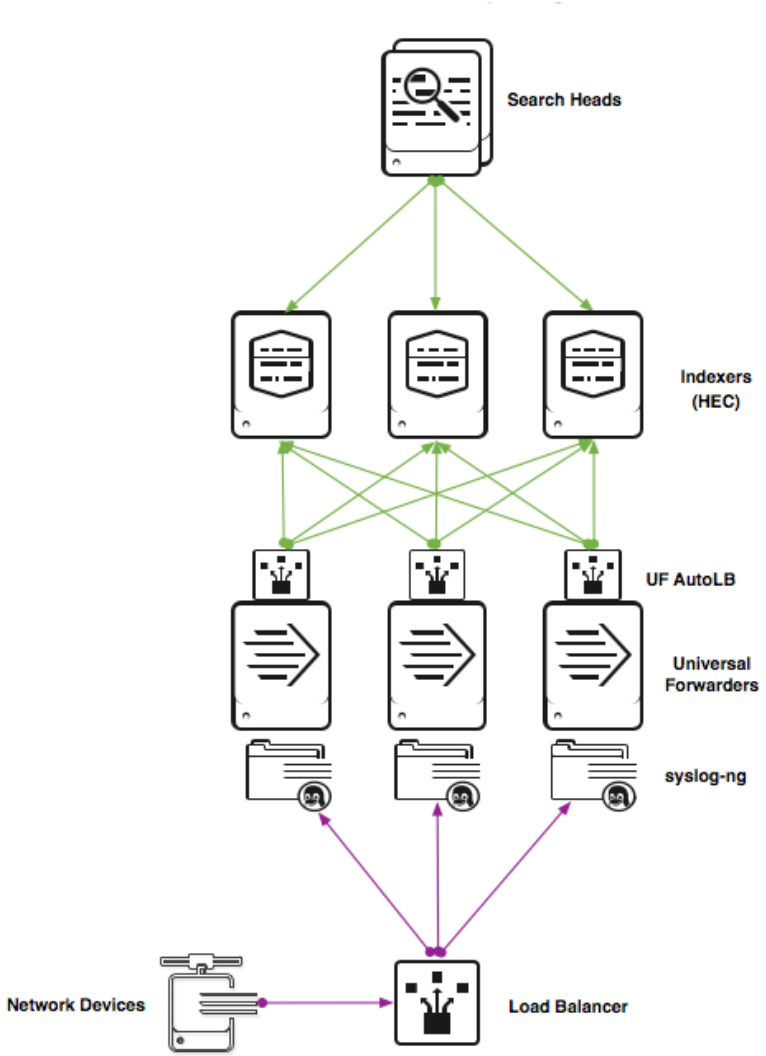
- Use syslog-ng or rsyslog, write the syslog to a file, and allow Splunk to monitor it.
- /data/syslog/<sourcetype>/<host>/<host>.log
  - /data/syslog/cisco_asa/<host>/<host>.log
  - /data/syslog/pan_log/<host>/<host>.log

```
[monitor:///data/syslog/cisco_asa]
index = thenetworkindex
sourcetype = cisco:asa
host_segment = 4
```

A Special Note On Syslog

- Syslog is a *protocol* – not a sourcetype
- Syslog typically carries multiple sourcetypes
- Best to pre-filter syslog traffic using syslog-ng or rsyslog
- Do *not* send syslog data directly to Splunk over a network port (514)
- Use a UF (next slide) or HEC to transport data to Splunk
- Remember to rotate your logs

A Recommended Syslog Architectures



What About CSV, JSON, XML, etc.

- Same rule as syslog – these are data formats and carry multiple sourcetypes

splunk> .conf2017

Wait…What If There Is No App?

Example: With No App

- You have an application from *Super Awesome Apps* called *Incredible*.
- The app has several logs it outputs:
  - /opt/saa/incredible/useractivity.log
  - /opt/saa/incredible/dbactivity.log
  - /opt/saa/incredible/webui.log

We Have This!

[monitor:///opt/saa/incredible/useractivity.log]

index = theindex

sourcetype = ????

- UF already knows the hostname, usually.
- What is the sourcetype?

Let's Talk

- Naming format: vendor:product:technology:format

- /opt/saa/incredible/useractivity.log
  - sourcetype = saa:incredible:useractivity
- /opt/saa/incredible/dbactivity.log
  - sourcetype = saa:incredible:dbactivity
- /opt/saa/incredible/webui.log
  - sourcetype = saa:incredible:webui

splunk> .conf2017

With Correct Sourcetypes

Same Regex, Different Sourcetype

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
<IW_fnnc,5.0,0,-,-,-,-,0,-,-,-,-,-,-,-
"SonyEricssonW810i/R4EA Browser/NetFr
Configuration/CLDC-1.1 UP.Link/6.3.0.
http://www.cambridgestatebank.com/ "Fi
```

Bluecoat

REGULAR EXPRESSION

1 match, 48 steps (~1ms)

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

TEST STRING

SWITCH TO UNIT TESTS

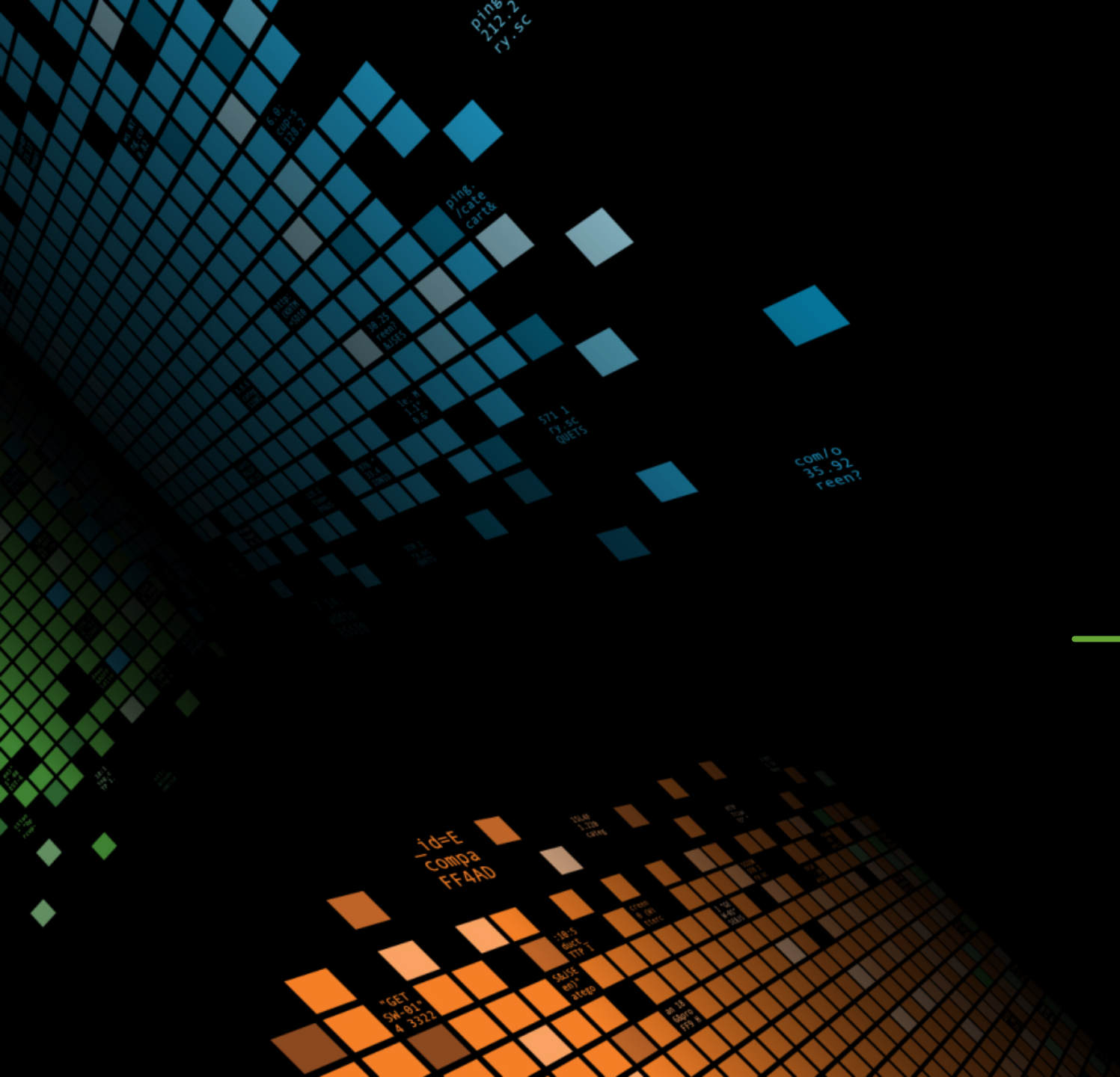
```
2017-08-27 20:31:52 10.180.6.76 64507 465 TCP_MISS 200 200 64445 326 -
- OBSERVED GET boostifun.cellmania.com HTTP/1.1 291
http://boostifun.cellmania.com/ifuncb/ContentDownload?
orderid=56660425&itemid=222067&fromProd=true -
ringtone/mp3;charset=ISO-8859-1 "Java/1.4.2_08" -
```

splunk> .conf2017

For Performance

Configuring these 6 settings for each sourcetype, on your indexers.

```
[saa:incredible:useractivity]
TIME_PREFIX = ^
SHOULD_LINEMERGE = false
LINE_BREAKER = ([\r\n]+)
MAX_TIMESTAMP_LOOKAHEAD = 30
TIME_FORMAT = %Y-%m-%d %H:%M:%S.%f%z
TRUNCATE = 10000
```



Data Usability

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

splunk> .conf2017

What Do You Do?

Making your data usable.

- Already have a proper sourcetype
- Extract your fields
- Create field aliases
  - username AS user
- 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

splunk> .conf2017

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
- index=* (sourcetype=*-* OR sourcetype=*too_small) | stats count by sourcetype
- Don’t make the puppy sad!

What's Next?

- Splunk Enterprise Data Administration
  - https://www.splunk.com/view/SP-CAAAPSE

- Go see these sessions….
  - Dashboards, Alerting, Reporting and Visualization - What's New
  - Next Generation Dashboards
  - Lesser Known Search Commands
  - Creating Your Own Splunk Learning Environment

splunk> .conf2017

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

Don't forget to rate this session in the .conf2017 mobile app