.conf2015

Taming Your Data

Mark Runals

Lead Security Engineer, OSU

splunk>

Disclaimer

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 the 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.

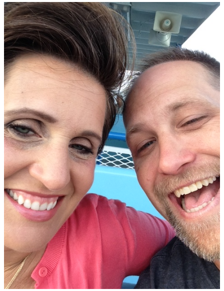
Agenda

- OSU Splunk Deployment – Environmental Background
- Quick Look at Data Curator Overview Dashboard
- Props & Field Extraction Score Methodology
- Views to help Splunk admins prioritize time

FYI - Splunk Admin Focused Presentation

About Me

- Using Splunk for 3 years
- ArcSight admin for 3 years
- Worked in InfoSec for 10 yrs+
- Motto - Solve for 80% and move on
- Getting data into Splunk isn't the end game



On ferry going to Survivor open casting call

Some Background & Program Drivers

OSU Environment

135 Distributed IT units around OSU

- Each group is autonomous
- No standardization
- Huge variety of technologies
- Splunk use not mandatory

Desired lightweight onboarding process

- For units & for SecOps team

+ splunk>

=

Incredible roll-on/ adoption rate

Fast Forward 3 Years +/-

- 2TB of data
- 2,800+ Splunk agents
- 16k devices
- 12 types of firewalls
- Multiple OS
- 90+ teams with data in Splunk
- 900+ sourcetypes – many ‘learned’
- 550+ accounts provisioned

Fast Forward 3 Years +/-

- 2TB of data
- 2,800+ Splunk agents
- 16k devices
- 12 types of firewalls
- Multiple OS
- 90+ teams with data in Splunk
- 900+ sourcetypes – many ‘learned’
- 550+ accounts provisioned

Is data being ingested correctly?

What fields have been defined?; where?

What types of data are in Splunk?

What’s not configured correctly?

Issue Overview

Out of the box Splunk will generally ingest data correctly

- Host names
- Sourcetypes
- Timestamp
- Line breaking
- Auto key-value fields

At best this isn't efficient.
At worst it can strain your deployment and may drop/lose events

Factors in play

- Hardware
- Data distribution
- Sourcetype velocity
- Ratio of indexers to total log volume
- Weird date/time information in your logs
- etc

.conf2015

The Data Curator App

splunk>

Data Curator App

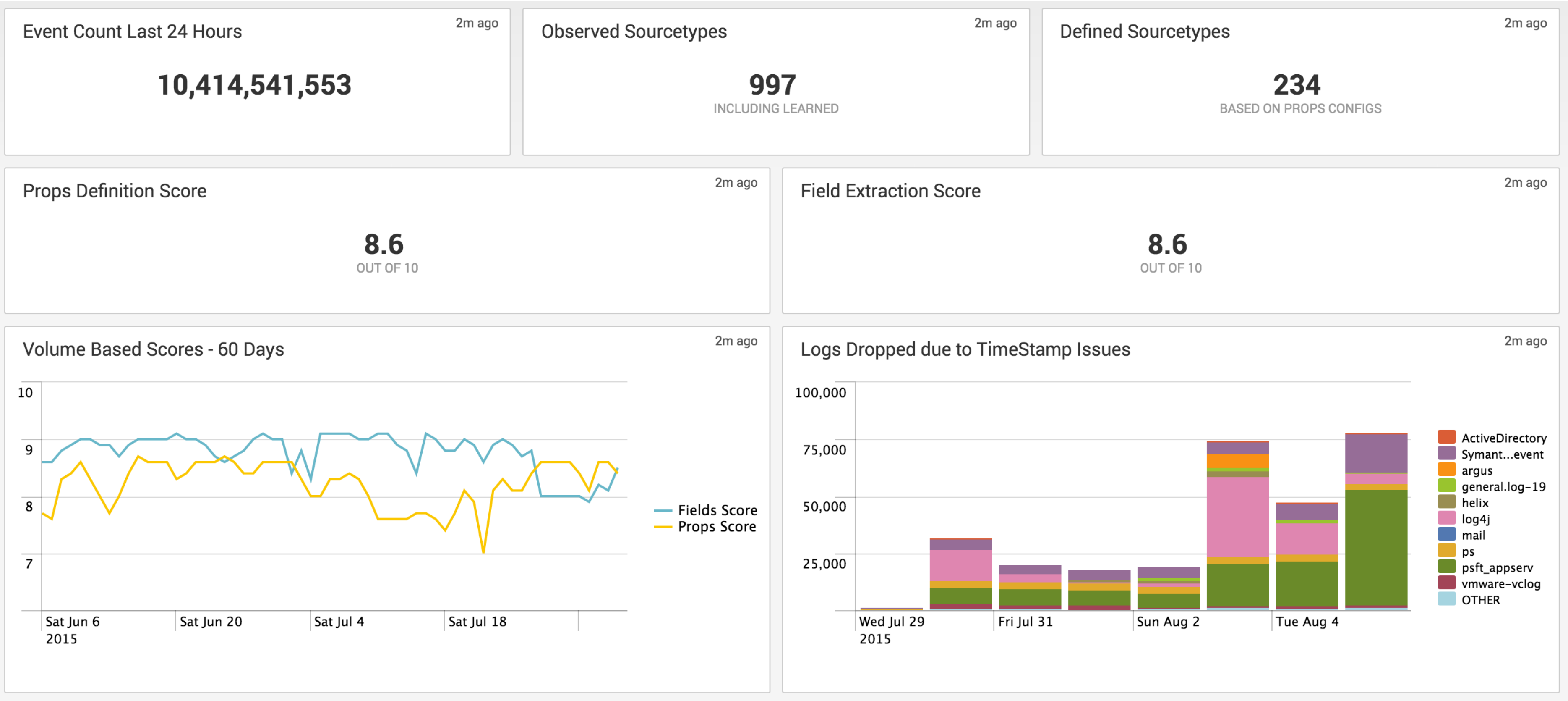
Goals

- Generate aggregate data onboarding maturity scores
- Generate ~accurate individual sourcetype maturity score
- Show what app/package contains props settings
- Show current props settings
- Highlight issues related to/solvable by props settings
  - Line breaking
  - Timestamp
  - Transforms issues

Take Note!

- App will NOT tell you what the settings should be
- Requires Splunk 6x search head
- Only able to work through issues I saw in my environment - you may have others.
- I can troubleshoot my app – not your deployment =)

Deployment At A Glance



Event Count Last 24 Hours — 2m ago

10,414,541,553

Observed Sourcetypes — 2m ago

997
INCLUDING LEARNED

Defined Sourcetypes — 2m ago

234
BASED ON PROPS CONFIGS

Props Definition Score — 2m ago

8.6
OUT OF 10

Field Extraction Score — 2m ago

8.6
OUT OF 10

Volume Based Scores - 60 Days — 2m ago

Logs Dropped due to TimeStamp Issues — 2m ago

Props Score Methodology

Props Score Methodology

- Based on *Getting Data in Correctly* presentation
- Individual scores reference the 7 primary props settings each sourcetype should have
- Aggregate score takes individual scores and factors in sourcetype volume
- Score converted to a 10 point scale (customizable)

Props Score

```
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE =
LINE_BREAKER =
TRUNCATE =
TZ =
```

Props Score

```
[mah_data_stanza]
TIME_PREFIX =              +1
MAX_TIMESTAMP_LOOKAHEAD =  +1
TIME_FORMAT =              +1
SHOULD_LINEMERGE =
LINE_BREAKER =
TRUNCATE =
TZ =
```

OR

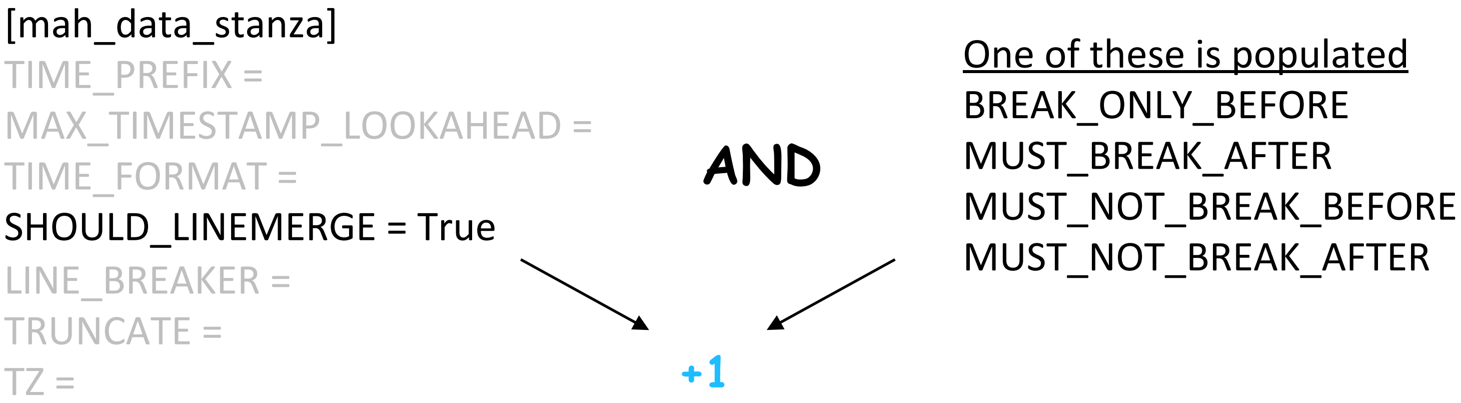
```
DATETIME_CONFIG =  +3
```

Props Score

```
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE = False   +1
LINE_BREAKER =
TRUNCATE =
TZ =
```

….but what if my data *should* be merged?

Props Score



```
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE = True
LINE_BREAKER =
TRUNCATE =
TZ =
```

AND

<u>One of these is populated</u>

```
BREAK_ONLY_BEFORE
MUST_BREAK_AFTER
MUST_NOT_BREAK_BEFORE
MUST_NOT_BREAK_AFTER
```

+1

Props Score

```
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE =
LINE_BREAKER =  +1
TRUNCATE =
TZ =
```

Default is ([\r\n\]+)

Don’t want to line break?

((?!)) or ((*FAIL)) are a couple options*

*http://answers.splunk.com/answers/106075/each-file-as-one-single-splunk-event

Props Score

```
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE =
LINE_BREAKER =
TRUNCATE =   +1
TZ =
```

Default is 10000 +0

Game your score!

- Set this to anything other than the default
  i.e. 10001 or 999999

Props Score

```
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE =
LINE_BREAKER =
TRUNCATE =
TZ =    +1
```

If setting this across your environment isn’t possible/practical reduce the max score macro in the app. It’s used as a variable.

Macro: props_score_upper_bounds = ~~7~~ 6

Props Score

```
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE =
LINE_BREAKER =
TRUNCATE =
TZ =
```

Max Score = 7

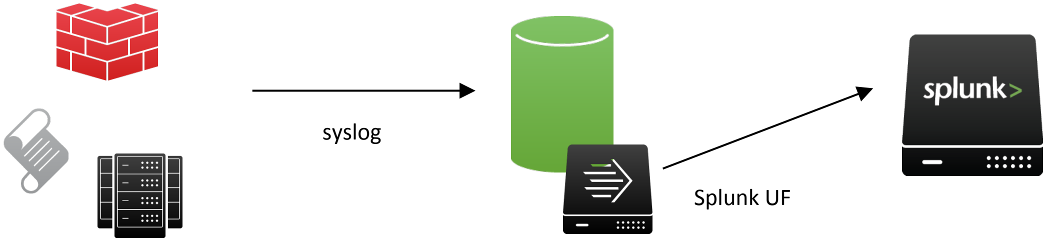
(st_score * `props_score_scale`) / `props_score_upper_bounds`

10

Props Score Caveats

There are a lot of additional props settings that could be applicable for your data/environment.

This method/app doesn't address host fields that are incorrect

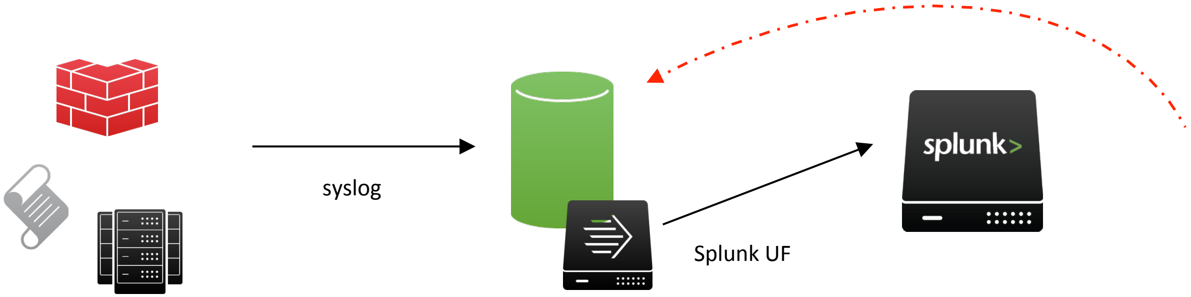


Default host field?

Props Score Caveats

There are a lot of additional props settings that could be applicable for your data/environment.

This method/app doesn't address host fields that are incorrect



Default host field?

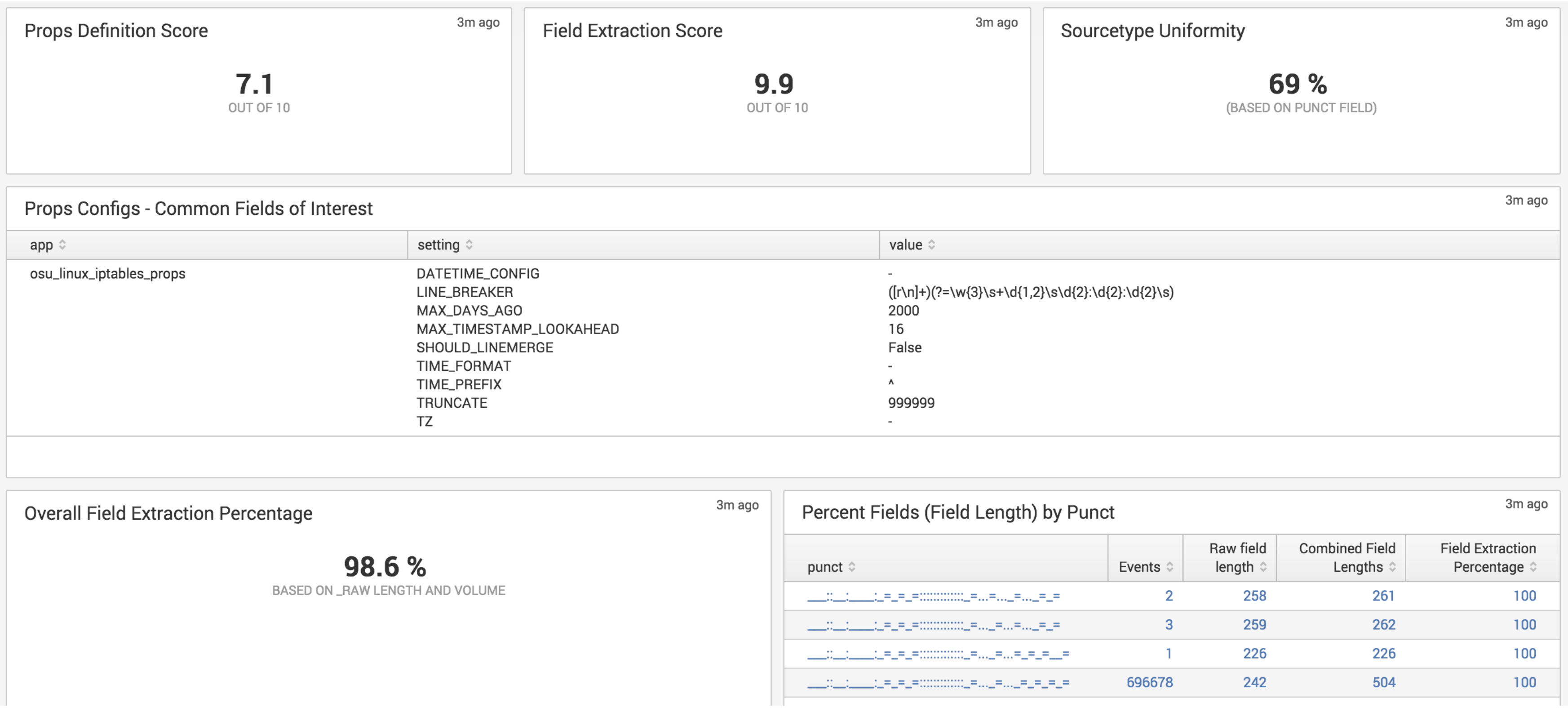
Props Score Macro’ed Query

```
rest splunk_server=local /servicesNS/-/-/configs/conf-props
| eval sourcetype = if(isnull(sourcetype) OR len(sourcetype)<1, title, sourcetype)
| rename eai:appName as App
| search App!=system App!=learned TIME_FORMAT=* OR TIME_PREFIX=* OR MAX_TIMESTAMP_LOOKAHEAD=* OR
LINE_BREAKER=* OR TZ=* OR TRUNCATE=* OR DATETIME_CONFIG=*
| eval datetime_set = if(DATETIME_CONFIG!="/etc/datetime.xml", "yes", "no")
| foreach TIME_FORMAT TIME_PREFIX LINE_BREAKER TZ BREAK_ONLY_BEFORE MUST_BREAK_AFTER MUST_NOT_BREAK_AFTER
MUST_NOT_BREAK_BEFORE [eval <<FIELD>> = if(isnull(<<FIELD>>) OR <<FIELD>>="","0","1")]
| eval multiline_settings = BREAK_ONLY_BEFORE +MUST_BREAK_AFTER +MUST_NOT_BREAK_AFTER
+MUST_NOT_BREAK_BEFORE
| eval line_merge = case(SHOULD_LINEMERGE=0, 1, SHOULD_LINEMERGE=1 AND multiline_settings=0, 0,
SHOULD_LINEMERGE="1" AND multiline_settings>0, 1)
| eval max_timestamp_lookahead = if(MAX_TIMESTAMP_LOOKAHEAD=150, 0, 1)
| eval truncate = if(TRUNCATE=10000, 0, 1)
| eval time_score = if(datetime_set ="no", max_timestamp_lookahead + TIME_FORMAT+ TIME_PREFIX, 3)
| eval props_score_raw = time_score + LINE_BREAKER + TZ + truncate + line_merge
| table sourcetype props_score_raw
```

Individual Sourcetype Deepdive

Sourcetype Deep Dive Dashboard

Linux:iptables



Props Definition Score

3m ago

7.1

OUT OF 10

Field Extraction Score

3m ago

9.9

OUT OF 10

Sourcetype Uniformity

3m ago

69 %

(BASED ON PUNCT FIELD)

Props Configs - Common Fields of Interest

3m ago

app	setting	value
osu_linux_iptables_props	DATETIME_CONFIG	-
	LINE_BREAKER	([r\n]+)(?=\w{3}\s+\d{1,2}\s\d{2}:\d{2}:\d{2}\s)
	MAX_DAYS_AGO	2000
	MAX_TIMESTAMP_LOOKAHEAD	16
	SHOULD_LINEMERGE	False
	TIME_FORMAT	-
	TIME_PREFIX	^
	TRUNCATE	999999
	TZ	-

Overall Field Extraction Percentage

3m ago

98.6 %

BASED ON _RAW LENGTH AND VOLUME

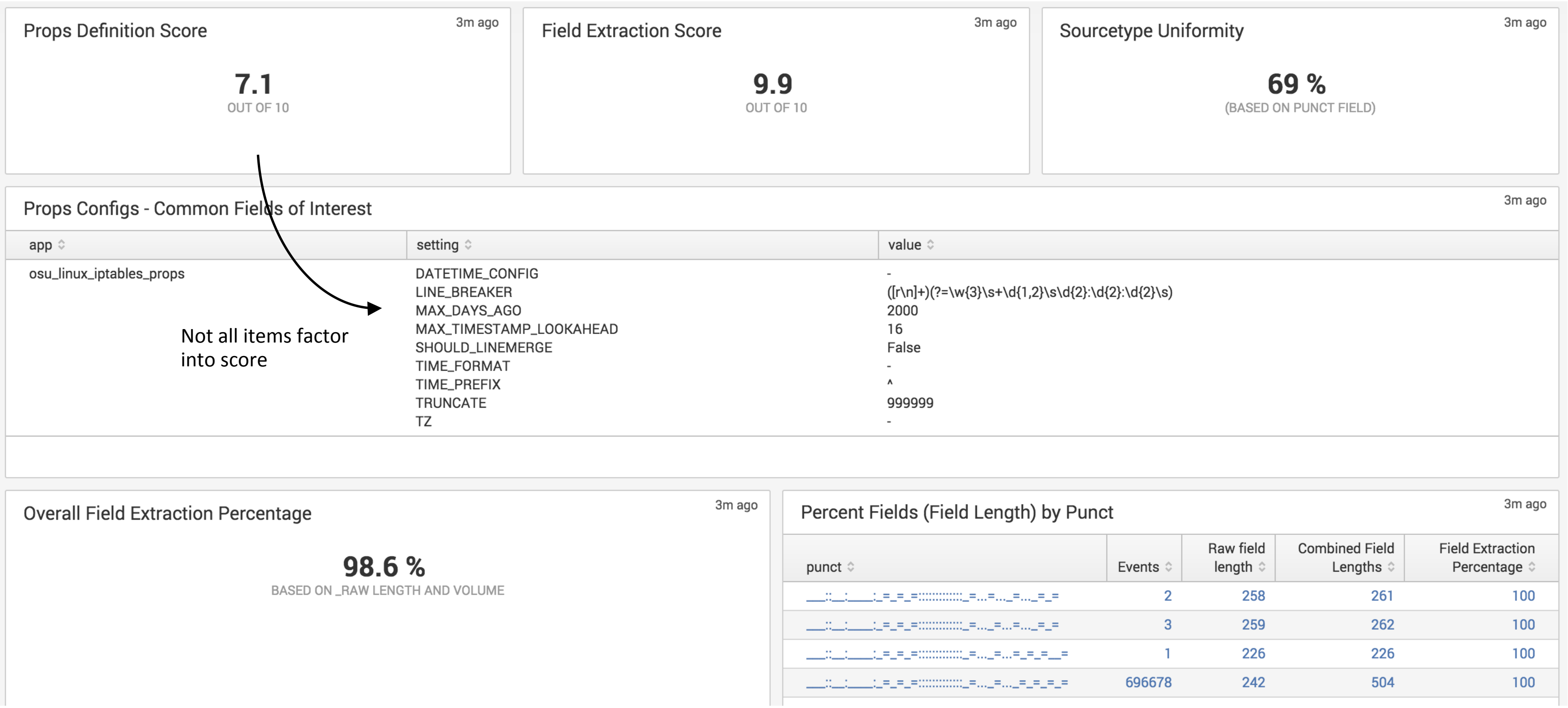
Percent Fields (Field Length) by Punct

3m ago

punct	Events	Raw field length	Combined Field Lengths	Field Extraction Percentage
___::__:_____:_=_=_=:::::::::::::_=...=..._=..._=_=	2	258	261	100
___::__:_____:_=_=_=:::::::::::::_=..._=...=..._=_=	3	259	262	100
___::__:_____:_=_=_=:::::::::::::_=..._=...=_=_=__=	1	226	226	100
___::__:_____:_=_=_=:::::::::::::_=..._=..._=_=_=_=	696678	242	504	100

Sourcetype Deep Dive Dashboard

Linux:iptables



Props Definition Score 3m ago

7.1
OUT OF 10

Field Extraction Score 3m ago

9.9
OUT OF 10

Sourcetype Uniformity 3m ago

69 %
(BASED ON PUNCT FIELD)

Props Configs - Common Fields of Interest 3m ago

app	setting	value
osu_linux_iptables_props	DATETIME_CONFIG	-
	LINE_BREAKER	([r\n]+)(?=\w{3}\s+\d{1,2}\s\d{2}:\d{2}:\d{2}\s)
	MAX_DAYS_AGO	2000
	MAX_TIMESTAMP_LOOKAHEAD	16
	SHOULD_LINEMERGE	False
	TIME_FORMAT	-
	TIME_PREFIX	^
	TRUNCATE	999999
	TZ	-

Not all items factor into score

Overall Field Extraction Percentage 3m ago

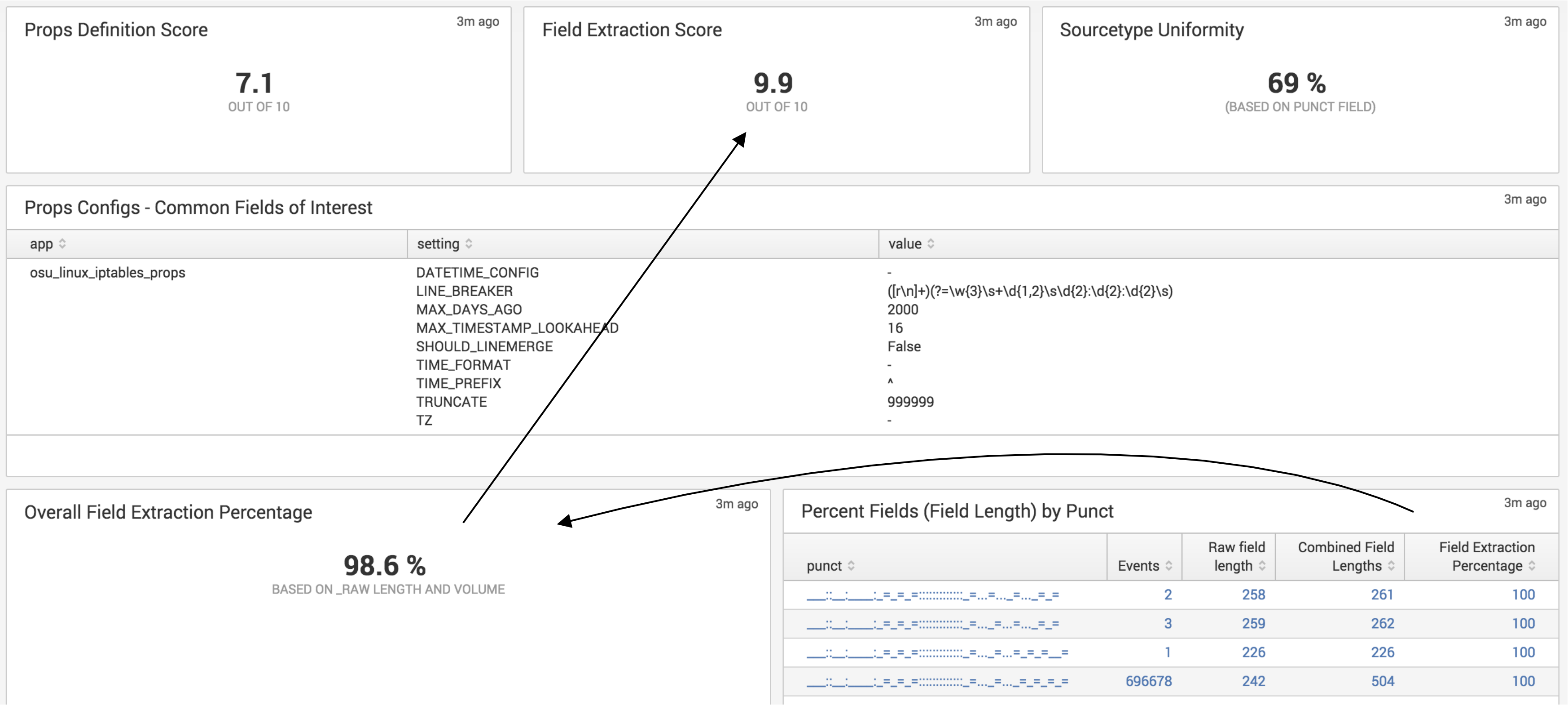
98.6 %
BASED ON _RAW LENGTH AND VOLUME

Percent Fields (Field Length) by Punct 3m ago

punct	Events	Raw field length	Combined Field Lengths	Field Extraction Percentage
___::__:____:_=_=_=:::::::::::::_=...=..._=..._=_=	2	258	261	100
___::__:____:_=_=_=:::::::::::::_=..._=...=..._=_=	3	259	262	100
___::__:____:_=_=_=:::::::::::::_=..._=...=_=_=__=	1	226	226	100
___::__:____:_=_=_=:::::::::::::_=..._=..._=_=_=_=	696678	242	504	100

Sourcetype Deep Dive Dashboard

Linux:iptables



Props Definition Score

3m ago

7.1

OUT OF 10

Field Extraction Score

3m ago

9.9

OUT OF 10

Sourcetype Uniformity

3m ago

69 %

(BASED ON PUNCT FIELD)

Props Configs - Common Fields of Interest

3m ago

app	setting	value
osu_linux_iptables_props	DATETIME_CONFIG	-
	LINE_BREAKER	([\r\n]+)(?=\w{3}\s+\d{1,2}\s\d{2}:\d{2}:\d{2}\s)
	MAX_DAYS_AGO	2000
	MAX_TIMESTAMP_LOOKAHEAD	16
	SHOULD_LINEMERGE	False
	TIME_FORMAT	-
	TIME_PREFIX	^
	TRUNCATE	999999
	TZ	-

Overall Field Extraction Percentage

3m ago

98.6 %

BASED ON _RAW LENGTH AND VOLUME

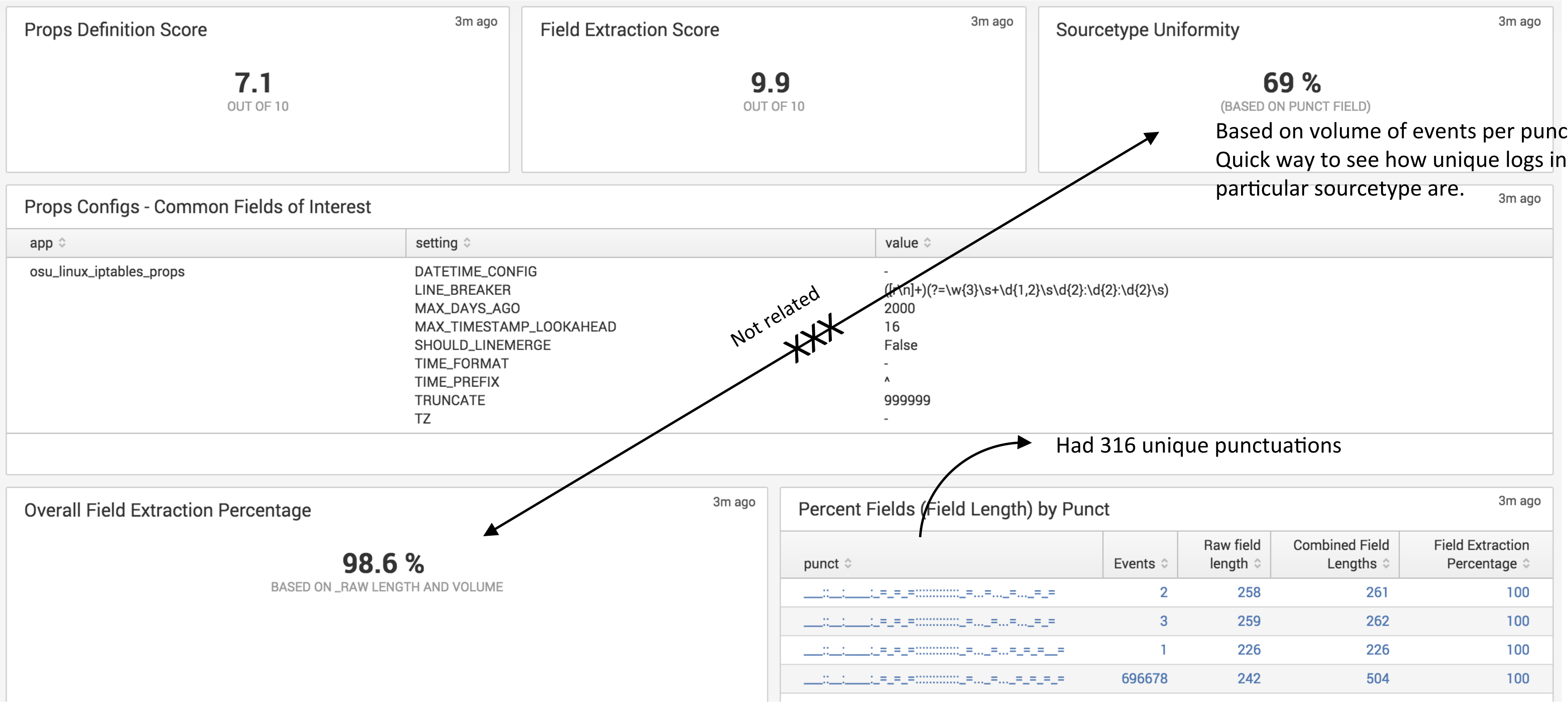
Percent Fields (Field Length) by Punct

3m ago

punct	Events	Raw field length	Combined Field Lengths	Field Extraction Percentage
___::__:_____:_=_=_=::::::::::::_=...=..._=..._=_=	2	258	261	100
___::__:_____:_=_=_=::::::::::::_=..._=...=..._=_=	3	259	262	100
___::__:_____:_=_=_=::::::::::::_=..._=...=_=_=__=	1	226	226	100
___::__:_____:_=_=_=::::::::::::_=..._=..._=_=_=_=	696678	242	504	100

Sourcetype Deep Dive Dashboard

Linux:iptables



Props Definition Score — 3m ago

7.1
OUT OF 10

Field Extraction Score — 3m ago

9.9
OUT OF 10

Sourcetype Uniformity — 3m ago

69 %
(BASED ON PUNCT FIELD)

Based on volume of events per punct. Quick way to see how unique logs in a particular sourcetype are.

Props Configs - Common Fields of Interest — 3m ago

app	setting	value
osu_linux_iptables_props	DATETIME_CONFIG	-
	LINE_BREAKER	([\r\n]+)(?=\w{3}\s+\d{1,2}\s\d{2}:\d{2}:\d{2}\s)
	MAX_DAYS_AGO	2000
	MAX_TIMESTAMP_LOOKAHEAD	16
	SHOULD_LINEMERGE	False
	TIME_FORMAT	-
	TIME_PREFIX	^
	TRUNCATE	999999
	TZ	-

Not related XXX

Overall Field Extraction Percentage — 3m ago

98.6 %
BASED ON _RAW LENGTH AND VOLUME

Had 316 unique punctuations

Percent Fields (Field Length) by Punct — 3m ago

punct	Events	Raw field length	Combined Field Lengths	Field Extraction Percentage
___::__:_____:_=_=_=:::::::::::_=...=..._=..._=_=	2	258	261	100
___::__:_____:_=_=_=:::::::::::_=..._=...=..._=_=	3	259	262	100
___::__:_____:_=_=_=:::::::::::_=..._=...=_=_=__=	1	226	226	100
___::__:_____:_=_=_=:::::::::::_=..._=..._=_=_=_=	696678	242	504	100

.conf2015

Fields Score Methodology

splunk>

Field Extraction Score Methodology

```
10.10.10.10 - - [20/Aug/2014:13:44:03.151 -0400] "POST /services/broker/phonehome/
connection_10.10.10.10_8089_10.10.10.10_TEST-TS_68D82260-CC1D-4203-83CA-6E24F9FE6538 HTTP/1.0" 200
24 - - - 1ms
```

Length of Fields

1. Account for any autokv field names
2. Do convoluted search to get length of fields
3. Account for timestamp in log
4. Get total length

÷

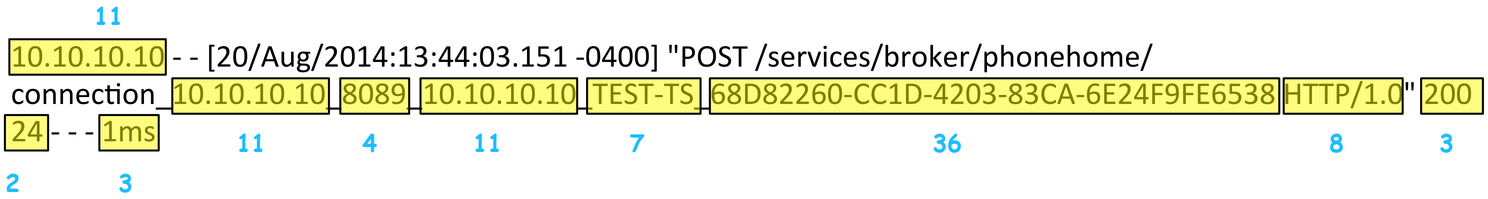
_raw length

1. Remove spaces
2. Remove newline characters
3. Get _raw length

=

% of Event has Fields Defined

Field Extraction Score Methodology



```
10.10.10.10 - - [20/Aug/2014:13:44:03.151 -0400] "POST /services/broker/phonehome/
connection_10.10.10.10_8089_10.10.10.10_TEST-TS_68D82260-CC1D-4203-83CA-6E24F9FE6538 HTTP/1.0" 200
24 - - 1ms
```

Length of Fields

1. Account for any autokv field names
2. Do convoluted search to get length of fields
3. Account for timestamp in log
4. Get total length

÷

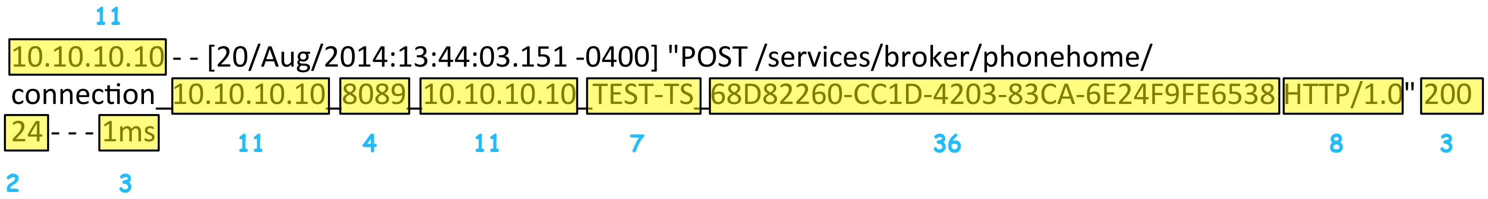
_raw length

1. Remove spaces
2. Remove newline characters
3. Get _raw length

=

% of Event has Fields Defined

Field Extraction Score Methodology



```
10.10.10.10 - - [20/Aug/2014:13:44:03.151 -0400] "POST /services/broker/phonehome/
connection_10.10.10.10_8089_10.10.10.10_TEST-TS_68D82260-CC1D-4203-83CA-6E24F9FE6538 HTTP/1.0" 200
24 - - - 1ms
```

Length of Fields = 96

1. Account for any autokv field names
2. Do convoluted search to get length of fields
3. Account for timestamp in log
4. Get total length

÷

_raw length = 171

1. Remove spaces
2. Remove newline characters
3. Get _raw length

=

= 56%

% of Event has Fields Defined

* Not a great example – Splunk forwarder phonehome logs actually have +100% field length compared to _raw

Field Extraction Score Methodology

Caveats / Considerations

- Doesn’t account for field alias (will artificially inflate score)
- If field extraction % is over 100 the score is set to 100
- Directionally correct is about the best this will get

➢ Fields extracted != field value

Field Extraction Macro’ed Query

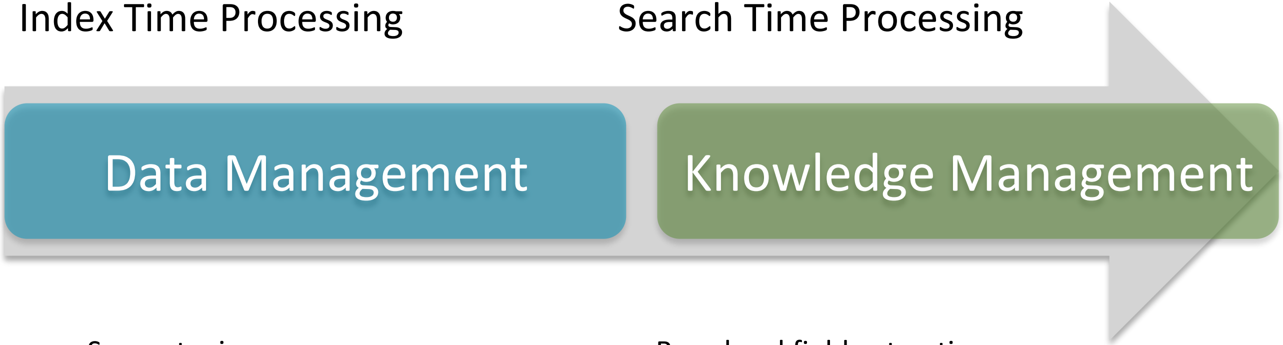
```
fields - date_* linecount eventtype source host splunk_server timestartpos timeendpos tag* index | rex max_match=100 "(?:\n|\s)?
(?<key_value_fields>\S+)=\s?" | nomv key_value_fields | rex mode=sed field=key_value_fields "s/ //g" | rex mode=sed
field=key_value_fields "s/\n//g" | eval kv_field_len = len(key_value_fields) | eval kv_field_len = if(isnotnull(kv_field_len),
kv_field_len, 0)| rex mode=sed field=_raw "s/ //g" | rex mode=sed field=_raw "s/\n//g" | eval raw_len=len(_raw) | eval time_len =
if(isnull(timestamp), 19, 0) | fields - timestamp _time key_value_fields | stats count first(*) as * by sourcetype punct | eval
total_field_len = 0 | foreach * [eval total_field_len = if(isnotnull('<<FIELD>>'), len('<<FIELD>>') + total_field_len, 0 + total_field_len)]
| eval raw_len_len = len(raw_len) | eval st_len=len(sourcetype) | eval t_len = len(time_len) | eval punct_len = len(punct) | eval
count_len = len(count) | eval total_field_len = total_field_len - st_len - raw_len_len - punct_len - count_len - t_len + kv_field_len |
table sourcetype punct count raw_len total_field_len | eval perc_fields = round((total_field_len/raw_len)*100) | eval perc_fields =
case(perc_fields>100,"100", perc_fields<0, "0", 1=1, perc_fields) | eventstats sum(count) as total by sourcetype | eval
loaded_perc_fields = count*perc_fields | stats sum(loaded_perc_fields) as loaded_perc_fields by sourcetype total | eval
loaded_perc_fields = round(loaded_perc_fields/total,`field_extraction_percentage_round_int`) | table sourcetype
loaded_perc_fields
```

Splunk Admin Time

Prioritizing work

Data Import/Definition Pipeline

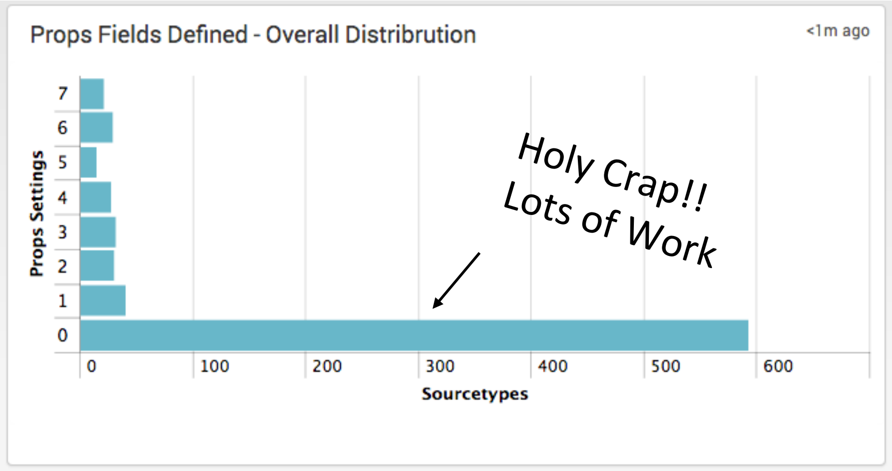
(Mark's View)



- Sourcetyping
- Line breaking
- Timestamp
- Host field
- etc

- Base level field extraction
- Normalized field names
- Field name alignment within Common Information Model (CIM)
- Knowledge Objects

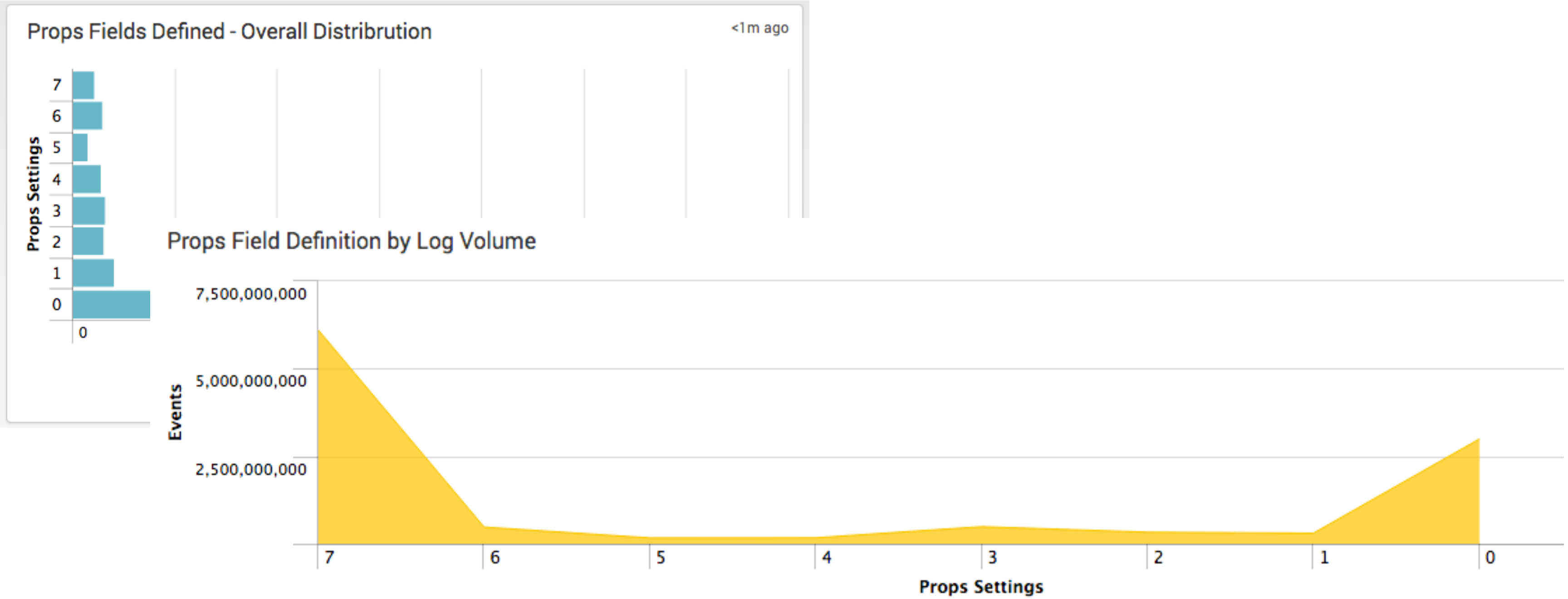
Props Score Breakdown



Holy Crap!!
Lots of Work

....but before you slit your wrists

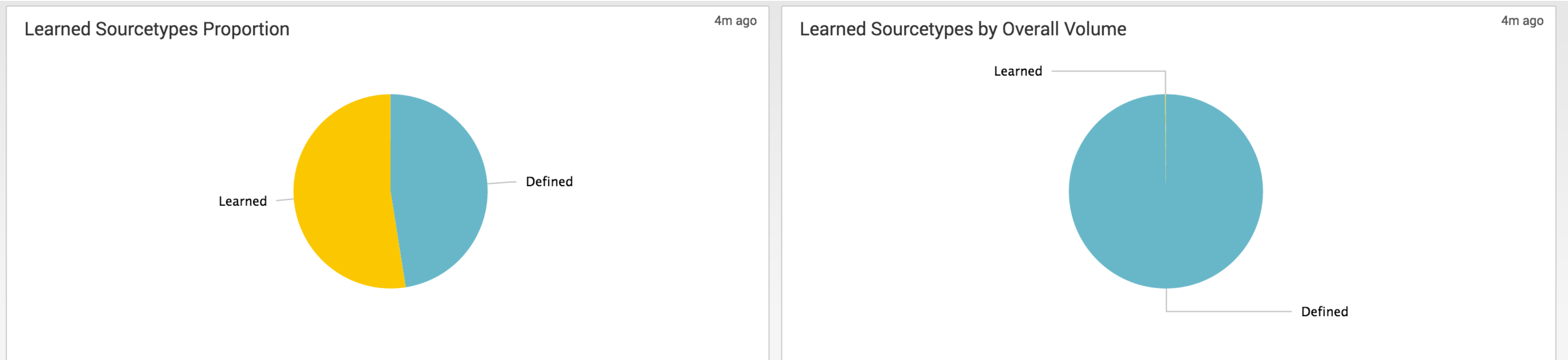
Props Score Breakdown



Props Fields Defined - Overall Distribrution

Props Field Definition by Log Volume

Learned Sourcetypes Quickview



Learned Sourcetypes Proportion

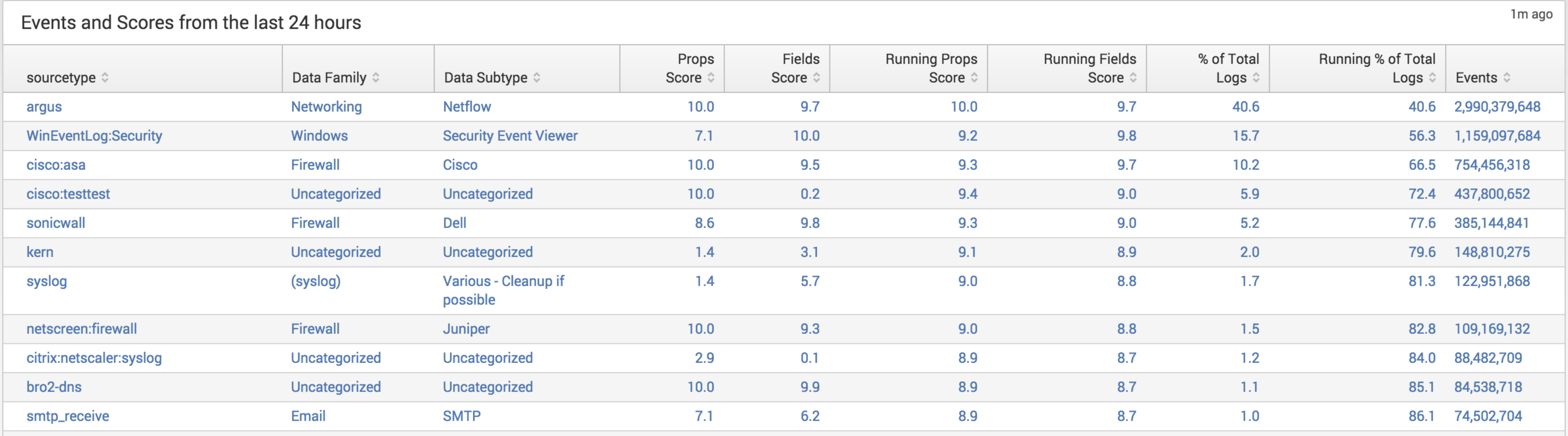
4m ago

Learned Sourcetypes by Overall Volume

4m ago

Learned = “too_small” OR -\d+$
Defined = not the above

Sourcetype Running Score List

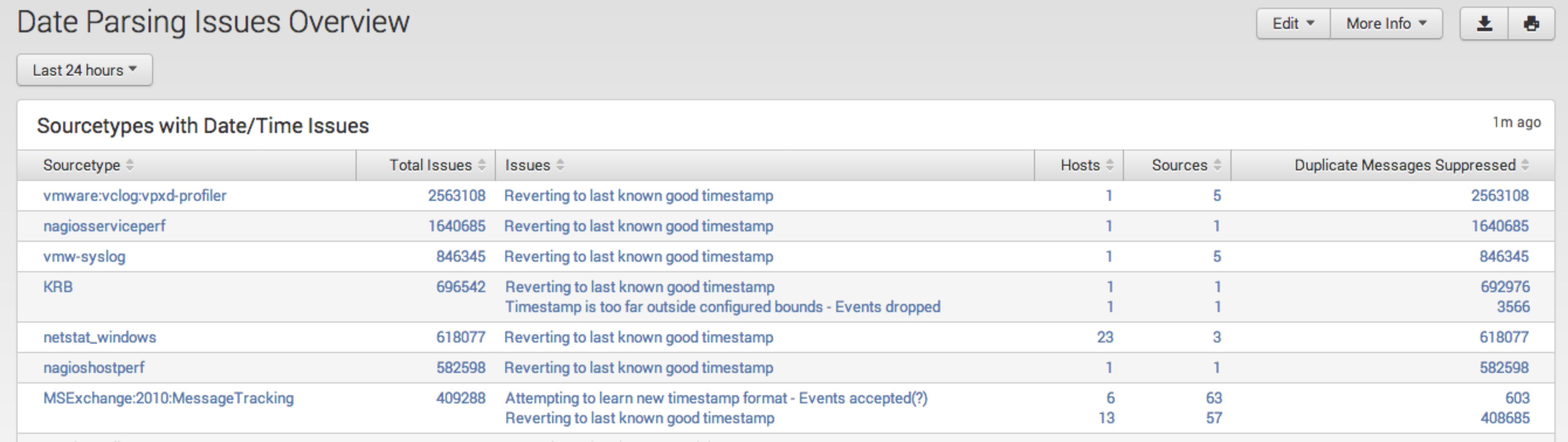


Events and Scores from the last 24 hours

sourcetype	Data Family	Data Subtype	Props Score	Fields Score	Running Props Score	Running Fields Score	% of Total Logs	Running % of Total Logs	Events
argus	Networking	Netflow	10.0	9.7	10.0	9.7	40.6	40.6	2,990,379,648
WinEventLog:Security	Windows	Security Event Viewer	7.1	10.0	9.2	9.8	15.7	56.3	1,159,097,684
cisco:asa	Firewall	Cisco	10.0	9.5	9.3	9.7	10.2	66.5	754,456,318
cisco:testtest	Uncategorized	Uncategorized	10.0	0.2	9.4	9.0	5.9	72.4	437,800,652
sonicwall	Firewall	Dell	8.6	9.8	9.3	9.0	5.2	77.6	385,144,841
kern	Uncategorized	Uncategorized	1.4	3.1	9.1	8.9	2.0	79.6	148,810,275
syslog	(syslog)	Various - Cleanup if possible	1.4	5.7	9.0	8.8	1.7	81.3	122,951,868
netscreen:firewall	Firewall	Juniper	10.0	9.3	9.0	8.8	1.5	82.8	109,169,132
citrix:netscaler:syslog	Uncategorized	Uncategorized	2.9	0.1	8.9	8.7	1.2	84.0	88,482,709
bro2-dns	Uncategorized	Uncategorized	10.0	9.9	8.9	8.7	1.1	85.1	84,538,718
smtp_receive	Email	SMTP	7.1	6.2	8.9	8.7	1.0	86.1	74,502,704

Good weekly/bi-weekly/monthly admin report

Identifying Date/Time Issues

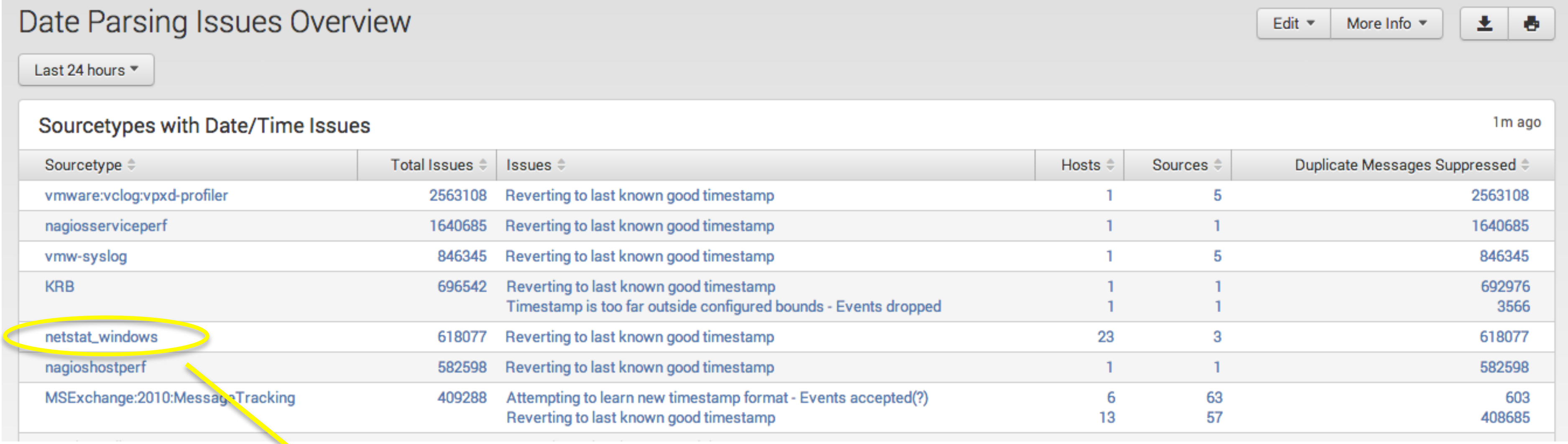


Date Parsing Issues Overview

Sourcetypes with Date/Time Issues

Sourcetype	Total Issues	Issues	Hosts	Sources	Duplicate Messages Suppressed
vmware:vclog:vpxd-profiler	2563108	Reverting to last known good timestamp	1	5	2563108
nagiosserviceperf	1640685	Reverting to last known good timestamp	1	1	1640685
vmw-syslog	846345	Reverting to last known good timestamp	1	5	846345
KRB	696542	Reverting to last known good timestamp Timestamp is too far outside configured bounds - Events dropped	1 1	1 1	692976 3566
netstat_windows	618077	Reverting to last known good timestamp	23	3	618077
nagioshostperf	582598	Reverting to last known good timestamp	1	1	582598
MSExchange:2010:MessageTracking	409288	Attempting to learn new timestamp format - Events accepted(?) Reverting to last known good timestamp	6 13	63 57	603 408685

Identifying Date/Time Issues



Date Parsing Issues Overview

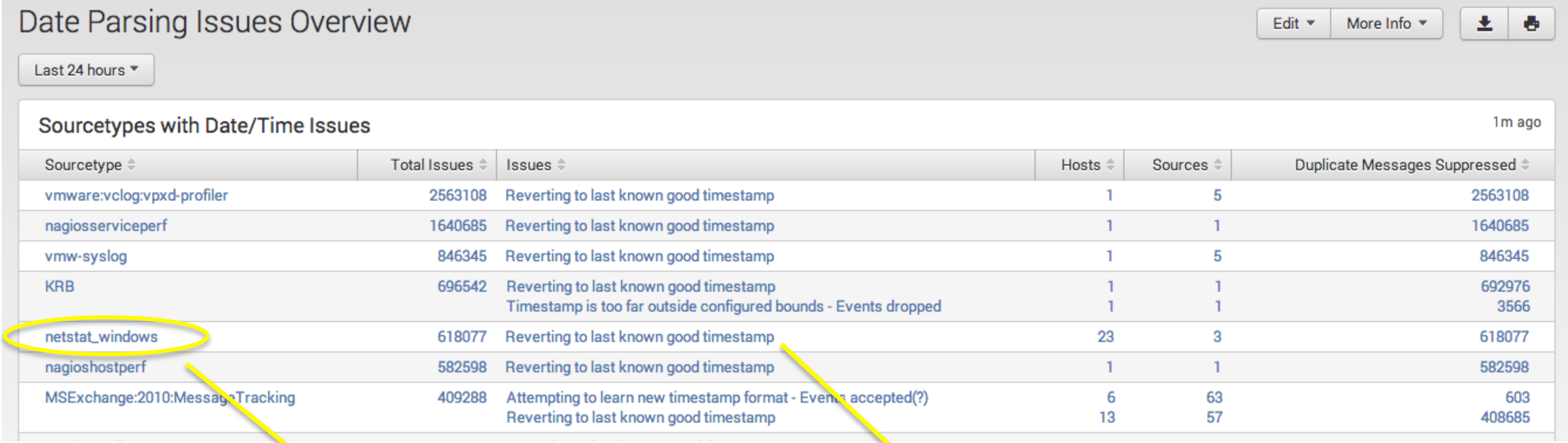
Sourcetypes with Date/Time Issues

1m ago

Sourcetype	Total Issues	Issues	Hosts	Sources	Duplicate Messages Suppressed
vmware:vclog:vpxd-profiler	2563108	Reverting to last known good timestamp	1	5	2563108
nagiosserviceperf	1640685	Reverting to last known good timestamp	1	1	1640685
vmw-syslog	846345	Reverting to last known good timestamp	1	5	846345
KRB	696542	Reverting to last known good timestamp Timestamp is too far outside configured bounds - Events dropped	1 1	1 1	692976 3566
netstat_windows	618077	Reverting to last known good timestamp	23	3	618077
nagioshostperf	582598	Reverting to last known good timestamp	1	1	582598
MSExchange:2010:MessageTracking	409288	Attempting to learn new timestamp format - Events accepted(?) Reverting to last known good timestamp	6 13	63 57	603 408685

These events don't have timestamps!

Identifying Date/Time Issues



Date Parsing Issues Overview

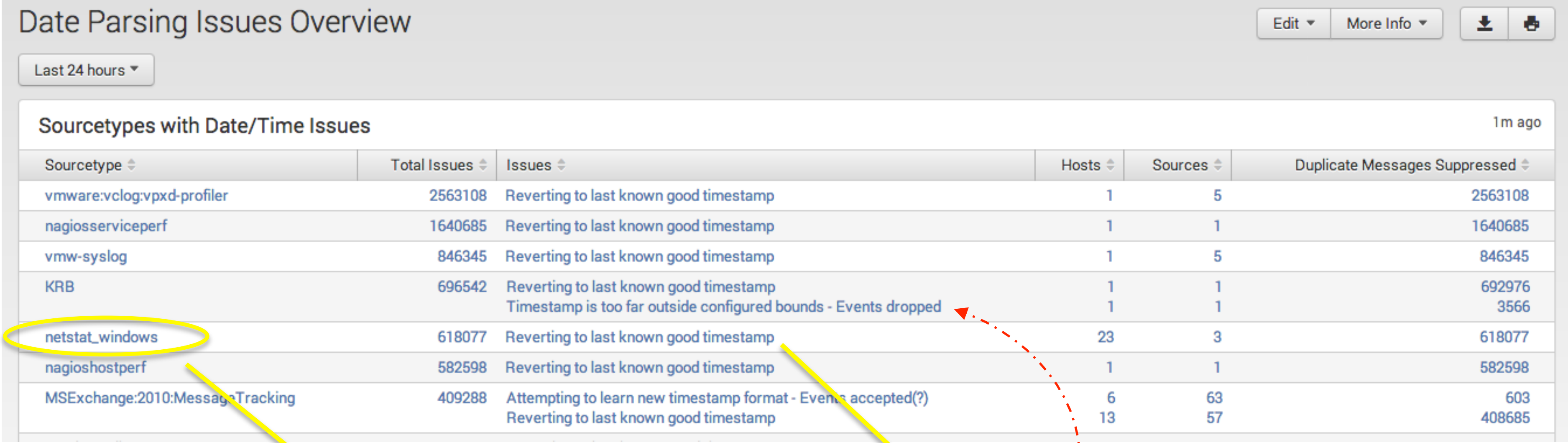
Sourcetypes with Date/Time Issues

Sourcetype	Total Issues	Issues	Hosts	Sources	Duplicate Messages Suppressed
vmware:vclog:vpxd-profiler	2563108	Reverting to last known good timestamp	1	5	2563108
nagiosserviceperf	1640685	Reverting to last known good timestamp	1	1	1640685
vmw-syslog	846345	Reverting to last known good timestamp	1	5	846345
KRB	696542	Reverting to last known good timestamp Timestamp is too far outside configured bounds - Events dropped	1 1	1 1	692976 3566
netstat_windows	618077	Reverting to last known good timestamp	23	3	618077
nagioshostperf	582598	Reverting to last known good timestamp	1	1	582598
MSExchange:2010:MessageTracking	409288	Attempting to learn new timestamp format - Events accepted(?) Reverting to last known good timestamp	6 13	63 57	603 408685

These events don't have timestamps!

What if Splunk thinks the last known good timestamp was 6 years ago?

Identifying Date/Time Issues



Date Parsing Issues Overview

Sourcetypes with Date/Time Issues

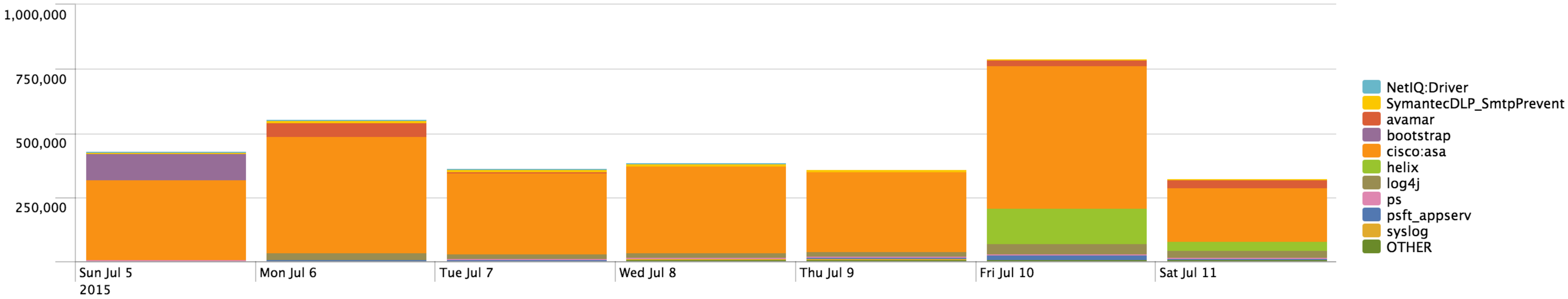
1m ago

Sourcetype	Total Issues	Issues	Hosts	Sources	Duplicate Messages Suppressed
vmware:vclog:vpxd-profiler	2563108	Reverting to last known good timestamp	1	5	2563108
nagiosserviceperf	1640685	Reverting to last known good timestamp	1	1	1640685
vmw-syslog	846345	Reverting to last known good timestamp	1	5	846345
KRB	696542	Reverting to last known good timestamp Timestamp is too far outside configured bounds - Events dropped	1 1	1 1	692976 3566
netstat_windows	618077	Reverting to last known good timestamp	23	3	618077
nagioshostperf	582598	Reverting to last known good timestamp	1	1	582598
MSExchange:2010:MessageTracking	409288	Attempting to learn new timestamp format - Events accepted(?) Reverting to last known good timestamp	6 13	63 57	603 408685

These events don't have timestamps!

What if Splunk thinks the last known good timestamp was 6 years ago?

Identifying Date/Time Issues



Cisco:ASA Logs

45 Firewalls

1 couldn't reach NTP servers > 2 month time skew

Some Other Features

Data Taxonomy

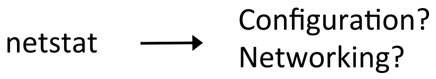
Version 1 – deprecated out of the box

Designed to answer “What type of data is in Splunk?”

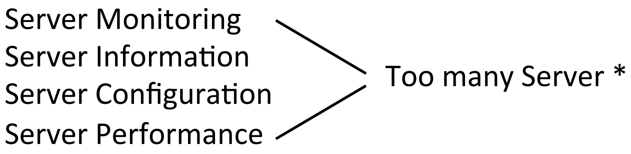
Created a 2 field classification csv for several hundred sourcetypes

- Data Family
- Data Subtype

Very useful but too many one-to-many relationships based on data use



netstat → Configuration? Networking?

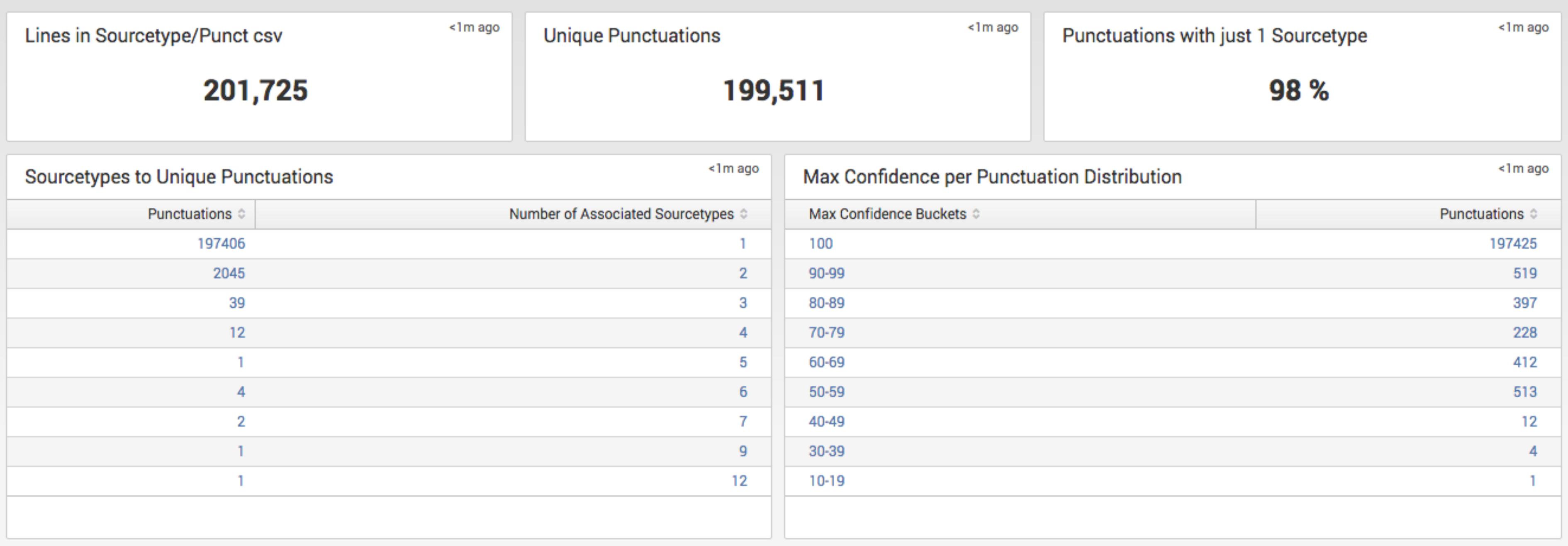


Server Monitoring
Server Information
Server Configuration
Server Performance

Too many Server *

Sourcetype Punctuation Overview

30 Minute Sampled Data



Lines in Sourcetype/Punct csv <1m ago

201,725

Unique Punctuations <1m ago

199,511

Punctuations with just 1 Sourcetype <1m ago

98 %

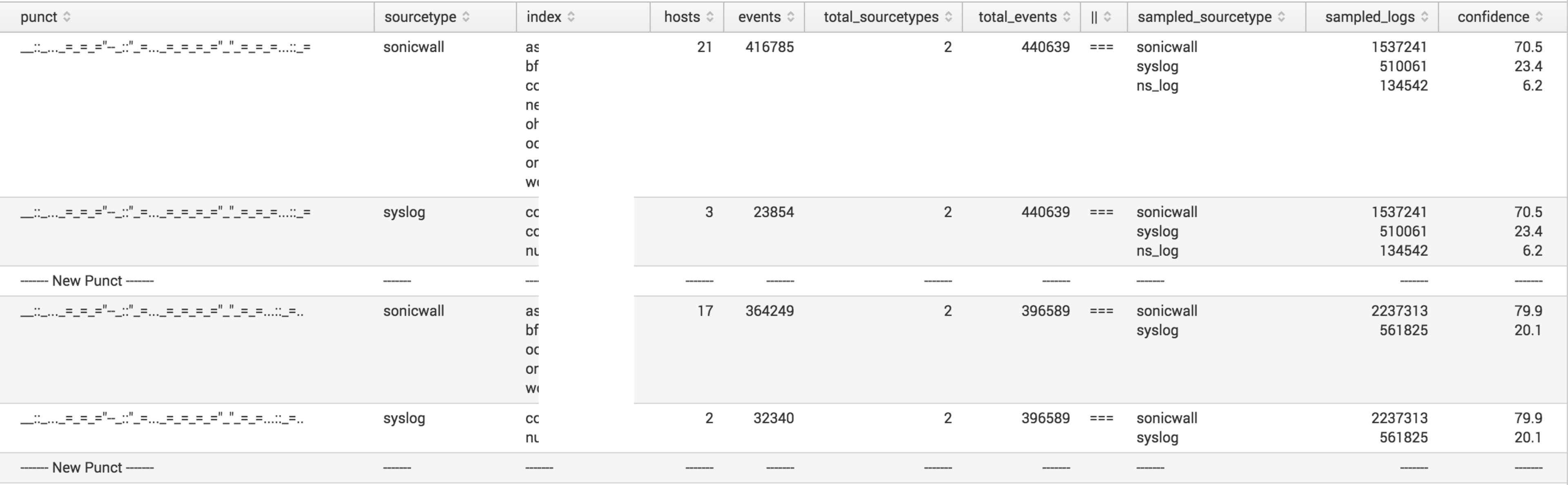
Sourcetypes to Unique Punctuations <1m ago

Punctuations ⇕	Number of Associated Sourcetypes ⇕
197406	1
2045	2
39	3
12	4
1	5
4	6
2	7
1	9
1	12

Max Confidence per Punctuation Distribution <1m ago

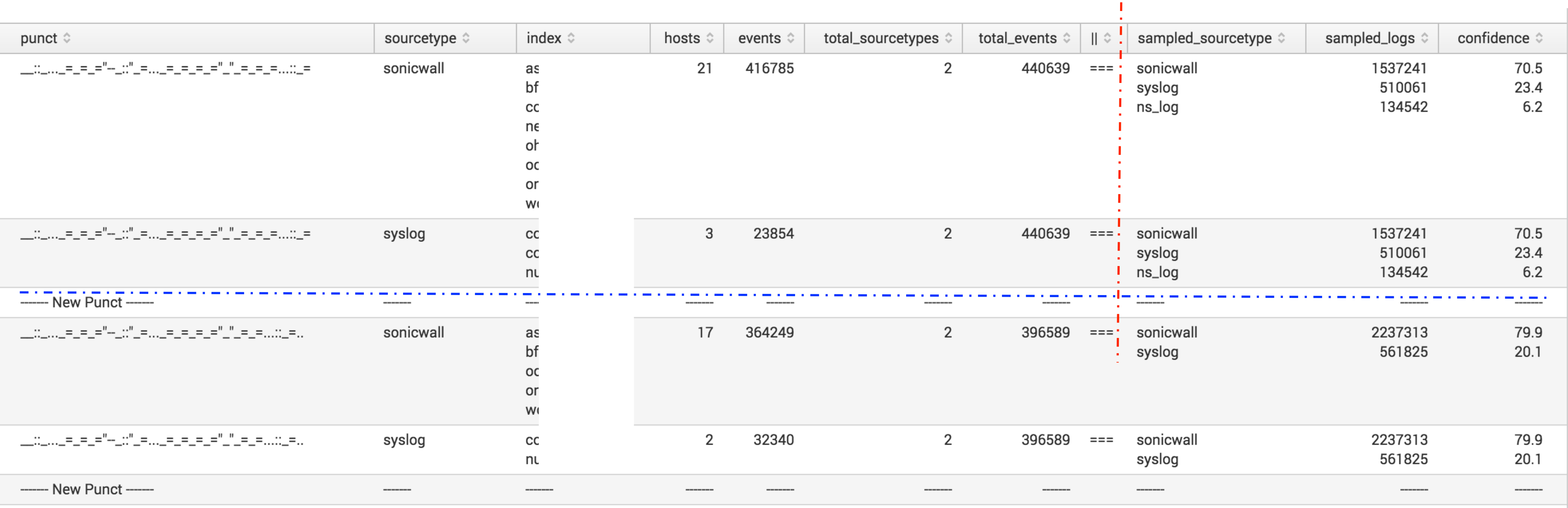
Max Confidence Buckets ⇕	Punctuations ⇕
100	197425
90-99	519
80-89	397
70-79	228
60-69	412
50-59	513
40-49	12
30-39	4
10-19	1

Sourcetype Punctuation Overview



<table>
<tr><th>punct</th><th>sourcetype</th><th>index</th><th>hosts</th><th>events</th><th>total_sourcetypes</th><th>total_events</th><th>||</th><th>sampled_sourcetype</th><th>sampled_logs</th><th>confidence</th></tr>
<tr><td>__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=_=...::_=</td><td>sonicwall</td><td>as
bf
cc
ne
oh
oc
or
w</td><td>21</td><td>416785</td><td>2</td><td>440639</td><td>===</td><td>sonicwall
syslog
ns_log</td><td>1537241
510061
134542</td><td>70.5
23.4
6.2</td></tr>
<tr><td>__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=_=...::_=</td><td>syslog</td><td>cc
cc
nu</td><td>3</td><td>23854</td><td>2</td><td>440639</td><td>===</td><td>sonicwall
syslog
ns_log</td><td>1537241
510061
134542</td><td>70.5
23.4
6.2</td></tr>
<tr><td>------- New Punct -------</td><td>-------</td><td>---</td><td>-------</td><td>-------</td><td>-------</td><td>-------</td><td></td><td>-------</td><td>-------</td><td>-------</td></tr>
<tr><td>__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=...::_=..</td><td>sonicwall</td><td>as
bf
oc
or
w</td><td>17</td><td>364249</td><td>2</td><td>396589</td><td>===</td><td>sonicwall
syslog</td><td>2237313
561825</td><td>79.9
20.1</td></tr>
<tr><td>__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=...::_=..</td><td>syslog</td><td>cc
nu</td><td>2</td><td>32340</td><td>2</td><td>396589</td><td>===</td><td>sonicwall
syslog</td><td>2237313
561825</td><td>79.9
20.1</td></tr>
<tr><td>------- New Punct -------</td><td>-------</td><td>-------</td><td>-------</td><td>-------</td><td>-------</td><td>-------</td><td></td><td>-------</td><td>-------</td><td>-------</td></tr>
</table>

Sourcetype Punctuation Overview



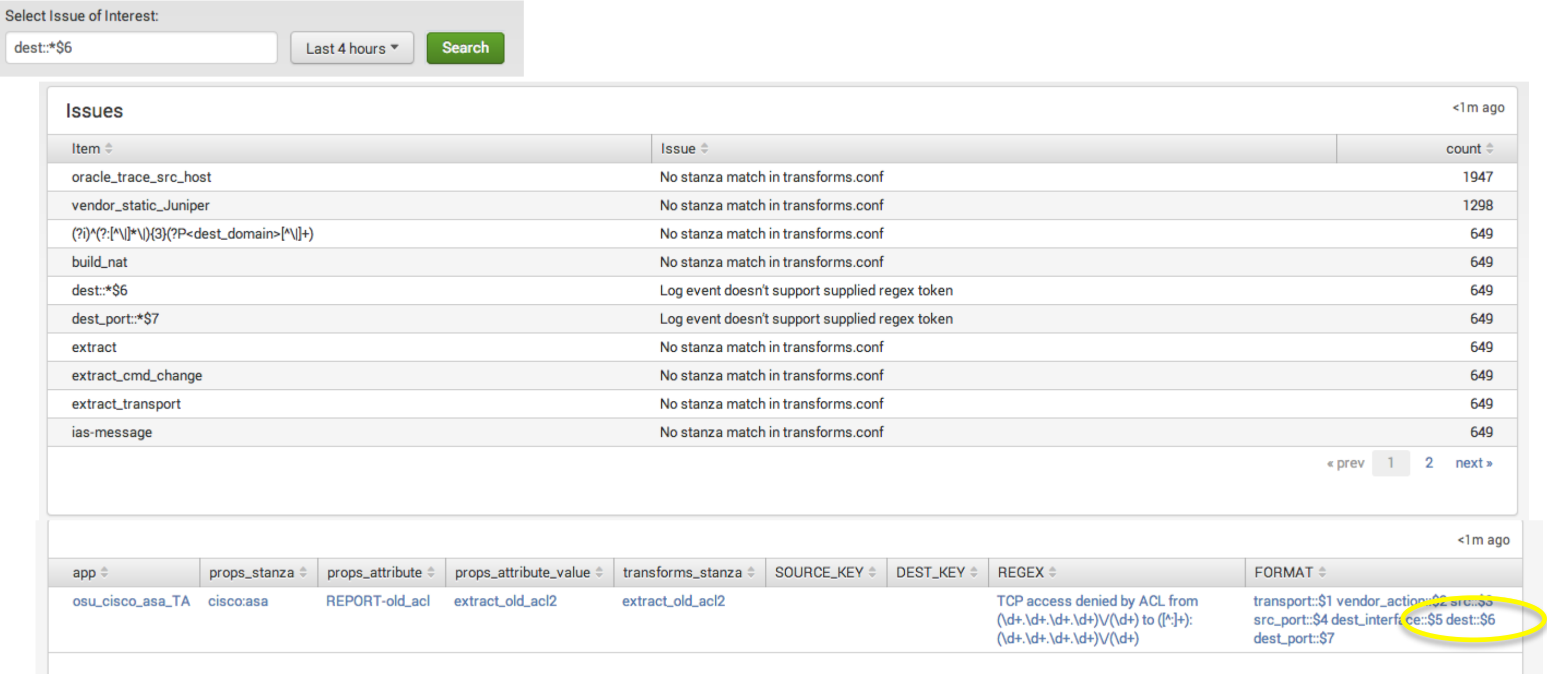
punct	sourcetype	index	hosts	events	total_sourcetypes	total_events	||	sampled_sourcetype	sampled_logs	confidence
__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=_=...::_=	sonicwall	as bf cc ne oh oc or w	21	416785	2	440639	===	sonicwall syslog ns_log	1537241 510061 134542	70.5 23.4 6.2
__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=_=...::_=	syslog	cc cc nu	3	23854	2	440639	===	sonicwall syslog ns_log	1537241 510061 134542	70.5 23.4 6.2
------- New Punct -------	-------	---	-------	-------	-------	-------		-------	-------	-------
__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=...::_=..	sonicwall	as bf oc or w	17	364249	2	396589	===	sonicwall syslog	2237313 561825	79.9 20.1
__::_..._=_=_="--_::"_=..._=_=_=_="_"_=_=...::_=..	syslog	cc nu	2	32340	2	396589	===	sonicwall syslog	2237313 561825	79.9 20.1
------- New Punct -------	-------	-------	-------	-------	-------	-------		-------	-------	-------

Sourcetype Punctuation Overview

Anecdotal Uses

- We have lots of data coming in via syslog receivers with sourcetype of “syslog”. Able to pull out cases where that data is actually set correctly elsewhere.
- Juniper firewall data collected by syslog receiver – sourcetype set on inputs. Someone deployed a Dell Sonicwall and pointed it to the Juniper syslog destination since “it would automatically come into Splunk”
- Unit standardized on iptables data being logged along a specific path. Quickly able to spot 3 systems that were still logging the data in /var/log/messages.

Extract / Report / Transforms Issues



Select Issue of Interest:

dest::*$6

Issues

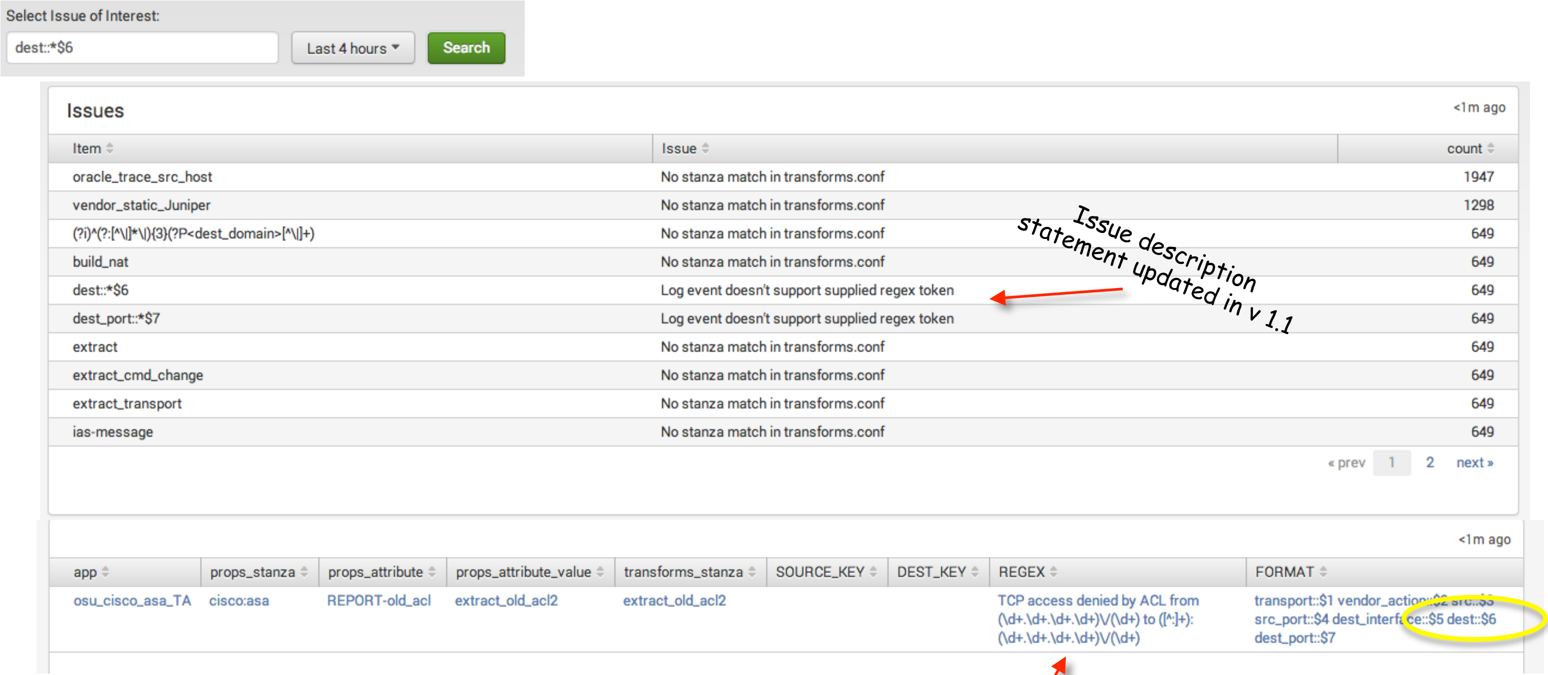
<1m ago

<table>
<tr><th>Item</th><th>Issue</th><th>count</th></tr>
<tr><td>oracle_trace_src_host</td><td>No stanza match in transforms.conf</td><td>1947</td></tr>
<tr><td>vendor_static_Juniper</td><td>No stanza match in transforms.conf</td><td>1298</td></tr>
<tr><td>(?i)^(?:[^\|]*\|){3}(?P<dest_domain>[^\|]+)</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>build_nat</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>dest::*$6</td><td>Log event doesn't support supplied regex token</td><td>649</td></tr>
<tr><td>dest_port::*$7</td><td>Log event doesn't support supplied regex token</td><td>649</td></tr>
<tr><td>extract</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>extract_cmd_change</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>extract_transport</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>ias-message</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
</table>

<1m ago

app	props_stanza	props_attribute	props_attribute_value	transforms_stanza	SOURCE_KEY	DEST_KEY	REGEX	FORMAT
osu_cisco_asa_TA	cisco:asa	REPORT-old_acl	extract_old_acl2	extract_old_acl2			TCP access denied by ACL from (\d+.\d+.\d+.\d+)\/(\d+) to ([^:]+):(\d+.\d+.\d+.\d+)\/(\d+)	transport::$1 vendor_action::$2 src::$3 src_port::$4 dest_interface::$5 dest::$6 dest_port::$7

Extract / Report / Transforms Issues



Select Issue of Interest:

dest::*$6

Issues

<1m ago

<table>
<tr><th>Item</th><th>Issue</th><th>count</th></tr>
<tr><td>oracle_trace_src_host</td><td>No stanza match in transforms.conf</td><td>1947</td></tr>
<tr><td>vendor_static_Juniper</td><td>No stanza match in transforms.conf</td><td>1298</td></tr>
<tr><td>(?i)^(?:[^\|]*\|){3}(?P<dest_domain>[^\|]+)</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>build_nat</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>dest::*$6</td><td>Log event doesn't support supplied regex token</td><td>649</td></tr>
<tr><td>dest_port::*$7</td><td>Log event doesn't support supplied regex token</td><td>649</td></tr>
<tr><td>extract</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>extract_cmd_change</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>extract_transport</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
<tr><td>ias-message</td><td>No stanza match in transforms.conf</td><td>649</td></tr>
</table>

Issue description statement updated in v 1.1

<1m ago

app	props_stanza	props_attribute	props_attribute_value	transforms_stanza	SOURCE_KEY	DEST_KEY	REGEX	FORMAT
osu_cisco_asa_TA	cisco:asa	REPORT-old_acl	extract_old_acl2	extract_old_acl2			TCP access denied by ACL from (\d+.\d+.\d+.\d+)\/(\d+) to ([^:]+): (\d+.\d+.\d+.\d+)\/(\d+)	transport::$1 vendor_action::$2 src::$3 src_port::$4 dest_interface::$5 dest::$6 dest_port::$7

Only 5 tokens

Other Focus Areas / Dashboards

Data Management

- Line Breaking
- Date Parsing
- Time zone issues
- Learned Sourcetypes

Knowledge Management

- Field Extraction
- Field Lookup (what sourcetypes have particular fields)
- Compare fields across multiple sourcetypes
- Extract, Transforms, Report

App Roadmap

Now

- Props maturity scores
- Field extraction scores
- Issues workspaces
- Mis-sourcetyped data
- Data Taxonomy
  Relatively non scaling

Next

- Dashboard optimization
  (ie searchTemplate)
- Tag based Data Taxonomy
- Any initial app bug fixes

After Next

- Tie in Data Model fields
- Field value?
- Expand issue troubleshooting
  Based on community feedback

?

.conf 14 *Getting Data in Correctly* presentation– Andrew Duca

Blog: runals.blogspot.com

Check out the Forwarder Health & Splunk Internal Change Mgmt app in Splunkbase

.conf2015

THANK YOU

splunk>