Taming Your Data

Mark Runals
Sr Security Engineer
The Ohio State University
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 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.
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
- Props/field extraction score methodology
- Look at data curator app

FYI - Splunk Admin Focused Presentation
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 Splunk team

Incredible roll-on/adoption rate
Fast Forward a Year or 2 +/-

- 2TB Of data
- 1,800+ Splunk agents
- 10k Devices
- 12 Types of firewalls
- Multiple OS
- 90+ Teams with data in Splunk
- 700+ Sourcetypes – many ‘learned’
- 350+ People
Fast Forward a Year or 2 +/-

- 2TB Of data
- 1,800+ Splunk agents
- 10k Devices
- 12 Types of firewalls
- Multiple OS
- 90+ Teams with data in Splunk
- 700+ Sourcetypes – many ‘learned’
- 350+ People

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 and without specific data definition Splunk will generally ingest data correctly

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

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

Factors in play

- Hardware
- Ratio of indexers to total log volume
- Sourcetype velocity
- Data distribution (forwarders pre 5.0.4 will favor first indexer listed in autoLB outputs.conf)
- Weird date/time information in your logs
- Etc...
Data Import/Definition Pipeline

(Mark’s View)

DM = Index Time Processing
- Sourcetyping
- Line breaking
- Timestamp
- Host field
- etc

KM = Search Time Processing
- Base level field extraction
- Normalized field names
- Field name alignment within Common Information Model (CIM)
- Knowledge objects
The Plan

Data Management
Score based on ‘Getting Data in Correctly’ .conf 2012 preso

Knowledge Management
Score based on length of fields relative to _raw length
(conversation with Kevin Meeks)

Identify Common Issues
Munge through internal logs

Data Taxonomy
Create way to classify sourcetypes
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE =
LINE_BREAKER =
TRUNCATE =
TZ =
Data Management – 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
[mah_data_stanza]
TIME_PREFIX =
MAX_TIMESTAMP_LOOKAHEAD =
TIME_FORMAT =
SHOULD_LINEMERGE = False +1
LINEBREAKER =
TRUNCATE =
TZ =

....but what if my data should be merged?
Data Management – Props Score

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

AND

One of these is populated
BREAK_ONLY_BEFORE
MUST BREAK AFTER
MUST NOT BREAK BEFORE
MUST NOT BREAK AFTER

+1
Data Management – 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
[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
Data Management – 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 = 6
Data Management – 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
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.
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.
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

\[
\frac{\text{Length of Fields}}{\_\text{raw length}} = \% \text{ of Event has Fields Defined}
\]
Field Extraction Score Methodology

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

 rawData length

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

\[
\frac{\text{raw data length}}{\text{raw data length}} = \% \text{ of Event has Fields Defined}
\]
# Field Extraction Score Methodology

<table>
<thead>
<tr>
<th>Length of Fields</th>
<th>= 96</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Account for any autokv field names</td>
<td>2. Do convoluted search to get length of fields</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>_raw length</th>
<th>= 171</th>
</tr>
</thead>
</table>

\[
\frac{96}{171} = 56\% 
\]

% 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
Data Taxonomy

Version 1 – deprecated out of the box

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

Created a 2\textsuperscript{nd} 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 *
Data Taxonomy – Interactive Host Dashboard
Data Taxonomy – Interactive Host Dashboard

Host B

Data Family Distribution

Detailed View of Data Taxonomy by Host

<table>
<thead>
<tr>
<th>data_family</th>
<th>data_subtype</th>
<th>source_type</th>
<th>index</th>
<th>events</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows</td>
<td>Security Event Viewer</td>
<td>WinEventLog\Security</td>
<td></td>
<td>10170</td>
</tr>
<tr>
<td>Uncategorized</td>
<td>Uncategorized</td>
<td>Windows\Process</td>
<td></td>
<td>8972</td>
</tr>
<tr>
<td>Networking</td>
<td>Netstat - Windows</td>
<td>netstat\windows</td>
<td></td>
<td>2847</td>
</tr>
<tr>
<td>Server Performance</td>
<td>Network Interface</td>
<td>Perfmon\LocalNetwork</td>
<td></td>
<td>888</td>
</tr>
<tr>
<td>Server Configuration</td>
<td>Arp</td>
<td>arp\windows</td>
<td></td>
<td>600</td>
</tr>
<tr>
<td>Server Information</td>
<td>CPU info</td>
<td>Windows\CPU</td>
<td></td>
<td>596</td>
</tr>
<tr>
<td>Server Performance</td>
<td>Memory Usage</td>
<td>Perfmon\Memory</td>
<td></td>
<td>447</td>
</tr>
<tr>
<td>Server Applications</td>
<td>Installed Applications (Windows)</td>
<td>Windows\QFE</td>
<td></td>
<td>302</td>
</tr>
<tr>
<td>Server Performance</td>
<td>CPU Time</td>
<td>Perfmon\CPU\Time</td>
<td></td>
<td>298</td>
</tr>
<tr>
<td>User Accounts</td>
<td>qwinsta</td>
<td>qwinsta</td>
<td></td>
<td>298</td>
</tr>
</tbody>
</table>
Data Curator App

Goals

• Flexible scoring scale
• Generate aggregate, system maturity scores
• Generate ~accurate individual maturity score
• Show what app/package contained props settings
• Show current props settings
• Highlight issues related to/solvable by props settings
  – Line breaking
  – Timestamp
  – Transforms issues

Take Note!

• Will NOT tell you what the settings should be
• Requires Splunk 6 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: 5,197,543,713
- Observed Sourcetypes: 787 (Including Learned)
- Defined Sourcetypes: 195 (Based on Props/Configs)
- Props Definition Score: 4.5 (Out of 10)
- Field Extraction Score: 8.6 (Out of 10)
- Logs Dropped due to TimeStamp Issues:
  - Wed Aug 13 2014: 5,000
  - Fri Aug 15: 15,000
  - Sun Aug 17: 10,000
  - Tue Aug 19: 15,000
Props Score Breakdown

Holy Crap!!
Lots of Work

....but before you slit your wrists
Props Score Breakdown
Learned Sourcetypes (-too_small OR -#)

Beware of diminishing returns on working the ‘long tail’
Sourcetype Deep Dive Dashboard

Avamar Logs

Props Definition Score
1 out of 10

Field Extraction Score
1.0 out of 10

Sourcetype Uniformity
13% (based on Punct field)

Props Configs - Common Fields of Interest

<table>
<thead>
<tr>
<th>setting</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATETIME_CONFIG</td>
<td>-</td>
</tr>
<tr>
<td>LINE_BREAKER</td>
<td>-</td>
</tr>
<tr>
<td>MAX_DAYS_ADD</td>
<td>2000</td>
</tr>
<tr>
<td>MAX_TIMESTAMP_LOOKAHEAD</td>
<td>128</td>
</tr>
<tr>
<td>SHOULD_LINEMERGE</td>
<td>-</td>
</tr>
<tr>
<td>TIME_FORMAT</td>
<td>True/Default</td>
</tr>
<tr>
<td>TIME_PREFIX</td>
<td>-</td>
</tr>
<tr>
<td>TRUNCATE</td>
<td>10000</td>
</tr>
<tr>
<td>TZ</td>
<td>-</td>
</tr>
</tbody>
</table>

Overall Field Extraction Percentage
10.3%
Based on raw length and volume

Perc Fields (Field Length) by Punct

<table>
<thead>
<tr>
<th>punct</th>
<th>Events</th>
<th>Raw field length</th>
<th>Combined Field Lengths</th>
<th>Field Extraction Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>708</td>
<td>89</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>-</td>
<td>707</td>
<td>55</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>-</td>
<td>789</td>
<td>89</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>-</td>
<td>123</td>
<td>100</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>
Sourcetype Deep Dive Dashboard

Avamar Logs

Props Definition Score
1 OUT OF 10

Field Extraction Score
1.0 OUT OF 10

Sourcetype Uniformity
13 % (BASED ON PUNCT FIELD)

Props Configs - Common Fields of Interest

Not all items factor into score

Overall Field Extraction Percentage
10.3 %
BASED ON RAW LENGTH AND VOLUME

Perc Fields (Field Length) by Punct

<table>
<thead>
<tr>
<th>Punct</th>
<th>Events</th>
<th>Raw Field Length</th>
<th>Combined Field Length</th>
<th>Field Extraction Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>-</em></td>
<td>708</td>
<td>69</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td><em>-</em>__</td>
<td>707</td>
<td>55</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td><em>-</em>__-</td>
<td>789</td>
<td>89</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td><em>-</em>__-</td>
<td>123</td>
<td>100</td>
<td>7</td>
<td>7</td>
</tr>
</tbody>
</table>
**Sourcetype Deep Dive Dashboard**

**Avamar Logs**

<table>
<thead>
<tr>
<th>Props Definition Score</th>
<th>Field Extraction Score</th>
<th>Sourcetype Uniformity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 OUT OF 10</td>
<td>1.0 OUT OF 10</td>
<td>13 % (BASED ON PUNCT FIELD)</td>
</tr>
</tbody>
</table>

**Props Configs - Common Fields of Interest**

<table>
<thead>
<tr>
<th>field</th>
<th>setting</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATETIME_CONF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>LINEBREAKER</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MAX_DAYS_AGO</td>
<td>2006</td>
<td>128</td>
</tr>
<tr>
<td>MAX_TIMES_AGO_LOCKHEAD</td>
<td>True / Default</td>
<td>-</td>
</tr>
<tr>
<td>SHOULD_TRUNCATE</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TIME_FORMAT</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TIME_PREFIX</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TRUNCATE</td>
<td>10000</td>
<td>-</td>
</tr>
<tr>
<td>TZ</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Overall Field Extraction Percentage**

10.3 %

Based on raw length and volume

---

*Loaded score based on volume of events per punct. Score created on the fly*
Sourcetype Deep Dive Dashboard

Avamar Logs

- Props Definition Score: 1 out of 10
- Field Extraction Score: 1.0 out of 10
- Sourcetype Uniformity: 13% (based on punct field)

Props Configs - Common Fields of Interest

- DATETIME_CONFIG
- LINE_BREAKER
- MAX_DAYS_ADD
- MAX_TIMESTAMP_LOOKAHEAD
- SHOULD_LINEMERGE
- TIME_FORMAT
- TIME_PREFIX
- TRUNCATE
- TZ

Overall Field Extraction Percentage: 10.3% (based on raw length and volume)

Perc Fields (Field Length) by Punct

- Had 75 unique punct

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

ABDCB (learned)

<table>
<thead>
<tr>
<th>Props Definition Score</th>
<th>Field Extraction Score</th>
<th>Sourcetype Uniformity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.5</td>
<td>99%</td>
</tr>
<tr>
<td>OUT OF 10</td>
<td>OUT OF 10</td>
<td>(BASED ON PUNCT FIELD)</td>
</tr>
</tbody>
</table>

Props Configs - Common Fields of Interest

No results found.

Overall Field Extraction Percentage

14.9%  
BASED ON RAW LENGTH AND VOLUME

Percent Fields (Field Length) by Punct

<table>
<thead>
<tr>
<th>punct</th>
<th>Events</th>
<th>Raw field length</th>
<th>Combined Field Lengths</th>
<th>Field Extraction Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>[&lt;&gt;: ]:</td>
<td>196</td>
<td>141</td>
<td>21</td>
<td>15</td>
</tr>
<tr>
<td>[&lt;&gt;: ]:####</td>
<td>1</td>
<td>60</td>
<td>3</td>
<td>5</td>
</tr>
</tbody>
</table>

1m ago
Sourcetype Deep Dive Dashboard

Argus
Identifying Date/Time Issues

Date Parsing Issues Overview

Source types with Date/Time Issues

<table>
<thead>
<tr>
<th>Source Type</th>
<th>Total Issues</th>
<th>Issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>vmware:volg:vxd-profile</td>
<td>2553108</td>
<td>Reverting to last known good timestamp</td>
</tr>
<tr>
<td>nagios:serviceperf</td>
<td>1640685</td>
<td>Reverting to last known good timestamp</td>
</tr>
<tr>
<td>vmw-syslog</td>
<td>846345</td>
<td>Reverting to last known good timestamp</td>
</tr>
<tr>
<td>KRB</td>
<td>696542</td>
<td>Reverting to last known good timestamp</td>
</tr>
<tr>
<td>netstat:windows</td>
<td>618077</td>
<td>Reverting to last known good timestamp</td>
</tr>
<tr>
<td>nagios:hostperf</td>
<td>582598</td>
<td>Reverting to last known good timestamp</td>
</tr>
<tr>
<td>MSExchange:2010:MessageTracking</td>
<td>409288</td>
<td>Attempting to learn new timestamp format - Events accepted (??)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Reverting to last known good timestamp</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hosts</th>
<th>Sources</th>
<th>Duplicate Messages Suppressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>2553108</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1640685</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>846345</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>692976</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>3566</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>618077</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>582598</td>
</tr>
<tr>
<td>6</td>
<td>63</td>
<td>409288</td>
</tr>
</tbody>
</table>
Identifying Date/Time Issues

These events don’t have timestamps!
Identifying Date/Time Issues

Date Parsing Issues Overview

These events don’t have timestamps!

What if Splunk thinks the last known good timestamp was 6 years ago?
Identifying Date/Time Issues

These events don’t have timestamps!

What if Splunk thinks the last known good timestamp was 6 years ago?
Date/Time Workspace Dashboard

Pre-populated with sourcetypes having issues

Additional Dashboard Elements
- Clustered internal logs giving you a level of visibility
- 100 most recent events

(No time information set)

(DATETIME_CONFIG added to view after screenshot)
Line Breaking/Truncate Workspace Dashboard

### Props Config

<table>
<thead>
<tr>
<th>app</th>
<th>setting</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-oracle_audit</td>
<td>DATETIME_CONFIG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>LINEBREAKER</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAX_DAYS_AGO</td>
<td></td>
</tr>
<tr>
<td></td>
<td>MAX_TIMESTAMP_LOOKAHEAD</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SHOULD_LINE器GERE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIME_FORMAT</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIME_PREFIX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRUNCATE</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TZ</td>
<td></td>
</tr>
</tbody>
</table>

### Issues Identified

<table>
<thead>
<tr>
<th>data_source</th>
<th>total_events</th>
<th>events</th>
<th>issue</th>
<th>limit</th>
<th>sources</th>
<th>hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>oracle_audit</td>
<td>40</td>
<td>40</td>
<td>Lines too long - adjust truncate setting (keep eye toward line breaking)</td>
<td>10000</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

### Clustered Internal Error Messages

<table>
<thead>
<tr>
<th>cluster_count</th>
<th>raw</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>08-29-2014 17:04:52.726-0400 WARN LineBreakingProcessor - Truncating line because limit of 10000 has been exceeded with a line length &gt;= 10136 - data_source=... data_host=... data_source_type=oracle_audit</td>
</tr>
</tbody>
</table>

### Longest Events (top 20)

<table>
<thead>
<tr>
<th>i</th>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Line Breaking/Truncate Workspace Dashboard

### Props Config

<table>
<thead>
<tr>
<th>app</th>
<th>setting</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA-oracle_audit</td>
<td>DATETIME_CONFIG</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>LINE_BREAKER</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MAX_DAYS_AGO</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>MAX_TIMESTAMP_LOOKAHEAD</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>SHOULD_LINEMERGE</td>
<td>True / Default</td>
</tr>
<tr>
<td></td>
<td>TIME_FORMAT</td>
<td>%Y-%m-%d %H:%M:%S %Z</td>
</tr>
<tr>
<td></td>
<td>TIME_PFX</td>
<td>TRUNCATE</td>
</tr>
<tr>
<td></td>
<td>TZ</td>
<td>EST/Eastern</td>
</tr>
</tbody>
</table>

### Issues Identified

<table>
<thead>
<tr>
<th>data_sourcetype</th>
<th>total_events</th>
<th>events</th>
<th>issue</th>
<th>limit</th>
<th>sources</th>
<th>hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>oracle_audit</td>
<td>40</td>
<td>40</td>
<td>Lines too long - adjust truncate setting (keep eye toward line breaking)</td>
<td>10000</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

### Clustered Internal Error Messages

```
06-29-2014 17:04:52.726 -0400 WARN LineBreakingProcessor - Truncating line because limit of 10000 has been exceeded with a line length >= 10136.
```

### Longest Events (top 20)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Event</td>
<td></td>
</tr>
</tbody>
</table>
Line Breaking Sanity Check Dashboard

Source types have line breaking set but have multiple line counts in recent events

<table>
<thead>
<tr>
<th>sourcetype</th>
<th>Events</th>
<th>Unique Linecounts</th>
<th>App</th>
<th>LINEBREAKER</th>
<th>SHOULD_LINEMERGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>oracle_audit</td>
<td>23176</td>
<td>54</td>
<td>TA-oracle_audit</td>
<td>$(\n\n)+{(?:w(3)s+\d+\s+\d(2)\s(2)w(2)s)\d(2)w(2)\s)}</td>
<td>True / Default</td>
</tr>
<tr>
<td>apache_error</td>
<td>35496</td>
<td>33</td>
<td>osu netscaler_props</td>
<td>$(\n\n)+{(?:w(3)s+\d(2)\s+\d(2)\s)+}</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>osu tomcat_apache_props</td>
<td>$(\n\n)+{(?:w(3)s+\d(2)\s+\d(2)\s)+}</td>
<td>False</td>
</tr>
<tr>
<td>log4j_fineq</td>
<td>10660</td>
<td>8</td>
<td>osu tomcat_apache Props</td>
<td>$(\n\n)+{(?:w(3)s+\d(2)\s+\d(2)\s)+}</td>
<td>False</td>
</tr>
<tr>
<td>ipconfig</td>
<td>14</td>
<td>7</td>
<td>osu windows_event_manipulation Props</td>
<td>$(\n\n)+{(?:w(3)s+\d(2)\s+\d(2)\s)+}</td>
<td>True / Default</td>
</tr>
<tr>
<td>log4j_hraction</td>
<td>8823</td>
<td>7</td>
<td>osu tomcat_apache Props</td>
<td>$(\n\n)+{(?:w(3)s+\d(2)\s+\d(2)\s)+}</td>
<td>False</td>
</tr>
</tbody>
</table>
Line Breaking Sanity Check

Sourcetypes have line breaking set but have multiple line counts in recent events

<table>
<thead>
<tr>
<th>sourctype</th>
<th>Events</th>
<th>Unique Linecounts</th>
<th>App</th>
<th>LINE_BREAKER</th>
<th>SHOULD_LINEMERGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>oracle_audit</td>
<td>23176</td>
<td>54</td>
<td>TA-oracle_audit</td>
<td>$(\n\n)+$</td>
<td>True / Default</td>
</tr>
<tr>
<td>apache_error</td>
<td>35496</td>
<td>33</td>
<td>osu_netScaler_props</td>
<td>$(\n\n)+$</td>
<td>False</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>osu_tomcat_apache_props</td>
<td>$(\n\n)+$</td>
<td>False</td>
</tr>
<tr>
<td>log4j_finreq</td>
<td>10668</td>
<td>8</td>
<td>osu_tomcat_apache_props</td>
<td>$(\n\n)+$</td>
<td>False</td>
</tr>
<tr>
<td>ipconfig</td>
<td>14</td>
<td>7</td>
<td>osu_windows_event_manip搡tion_props</td>
<td>$(\n\n)+$</td>
<td>True / Default</td>
</tr>
<tr>
<td>log4j_hraction</td>
<td>8823</td>
<td>7</td>
<td>osu_tomcat_apache_props</td>
<td>$(\n\n)+$</td>
<td>False</td>
</tr>
</tbody>
</table>

Set in multiple apps; potential problem down the road?
Query Troubleshooting

Two main scheduled searches that are somewhat computationally expensive. Dashboard allows admin to compare run length & frequency to coverage.

Sourcetype field length percentage query

- Current Cron Schedule: 51 ****
- Search Length (in Query): -60s
- Last Run: 220.80 SECONDS
- Avg Run Len: 139.81 SECONDS

Percentage of Non Learned Sourcetype Coverage

Fields Lengths Not Collected

Fields Lengths Collected
Extract/Report/Transforms Issues

Example Internal Warning Logs

08-21-2014 08:55:46.348 -0400 WARN SearchOperator:kv - IndexOutOfBoundsException invalid The FORMAT capturing group id: id=7, transform_name='Message'

08-21-2014 08:59:02.854 -0400 WARN SearchOperator:kv - Invalid key-value parser, ignoring it, transform_name='extract_cmd_change'

08-21-2014 08:59:03.345 -0400 WARN SearchOperator:kv - Invalid key-value parser, ignoring it, transform_name='(?i)^(?:\[\|\]*\|){3}(?P<dest_domain>\[\|\]+)' Wut?

Which app?

In props or transforms?

Solution: grep -r through 520+ packages in deployment-apps directory for ‘Message’?
Extract/Report/Transforms Issues
Extract/Report/Transforms Issues

Issue description statement updated in v 1.1

Only 5 tokens
Anyone know what the issue is?
Extract/Report/Transforms Issues

<table>
<thead>
<tr>
<th>Item</th>
<th>Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>oracle_trace_src_host</td>
<td>No stanza match in transforms.conf</td>
</tr>
<tr>
<td>vendor_static_kuniper</td>
<td>No stanza match in transforms.conf</td>
</tr>
<tr>
<td>(?)(?)(?)(?)(?)=dest_domain=([^)]+)</td>
<td>No stanza match in transforms.conf</td>
</tr>
<tr>
<td>build_net</td>
<td>No stanza match in transforms.conf</td>
</tr>
<tr>
<td>dest.*86</td>
<td>Log event doesn’t support supplied regex token</td>
</tr>
<tr>
<td>dest_port=87</td>
<td>Log event doesn’t support supplied regex token</td>
</tr>
<tr>
<td>extract</td>
<td>No stanza match in transforms.conf</td>
</tr>
<tr>
<td>extract_cmd_change</td>
<td>No stanza match in transforms.conf</td>
</tr>
<tr>
<td>extract_transport</td>
<td>No stanza match in transforms.conf</td>
</tr>
<tr>
<td>ias-message</td>
<td>No stanza match in transforms.conf</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>app</th>
<th>props_stanza</th>
<th>props_attribute</th>
<th>props_attribute_value</th>
<th>transforms_stanza</th>
<th>SOURCE_KEY</th>
<th>DEST_KEY</th>
<th>REGEX</th>
<th>FORMAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>osu_shibboleth Props</td>
<td>shibboleth</td>
<td>REPORT_shib_dest_domain</td>
<td>(?)(?)(?)(?)(?)=dest_domain=([^)]+)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Should be an EXTRACT
KM – Sourcetype Fields Comparison

Bottom of explanatory text. There is a freeform text search box at top of dashboard

<table>
<thead>
<tr>
<th>Field</th>
<th>WinEventLog:Security</th>
<th>cisco:asa</th>
<th>Source types with field in common</th>
</tr>
</thead>
<tbody>
<tr>
<td>action</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>app</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>dest</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>dest_ip</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>dest_port</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>ip</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>src</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>src_ip</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>src_port</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>vendor</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Account_Domain</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Account_Name</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Application_Name</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
App Roadmap

Now
• Props maturity scores
• Field extraction scores
• Issues workspaces
• 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 updated *Getting Data in Correctly* presentation— Andrew Duca

Blog: runals.blogspot.com

Check out the Forwarder Health app in Splunkbase
THANK YOU