The Power of Data Normalization

A look at the Common Information Model

Mark Bonsack, CISSP
Vladimir Skoryk, CISSP, CCFE, CHFI, CISA, CISM, RGTT

September 27, 2017 | Washington, DC

| Staff Sales Engineer, Splunk
| PS Supreme Architect, Splunk
Introduction
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, Splunk>, Listen to Your Data, The Engine for Machine Data, Splunk Cloud, Splunk Light and SPL are trademarks and registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners. © 2017 Splunk Inc. All rights reserved.
Who are we?

Mark: Staff Systems Engineer, Southwest Majors
• 6 years @ Splunk
• Focus: Security, Networking, IT Operations space

Vladimir: PS Supreme Architect, Professionally homeless
• 4 years @ Splunk
• Focus: Security
We will discuss:
This is where the subtitle goes

1. Why data normalization is important
2. The Splunk Common Information Model
3. Tools and Resources on the road to CIM compliance
Quick poll

► Have you heard of the Splunk Common Information model (CIM)?

► Have you worked on normalizing data using the Splunk CIM?
Data Normalization

Making the Most of Schema on the Fly
Late Binding Schema Rewards Time Invested

SPLUNK Late-Binding Schema

ETL into an RDBMS

<table>
<thead>
<tr>
<th>Time Invested</th>
<th>Value Returned</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>
Late Binding Schema Rewards Time Invested

CHANGE leads to Zeno’s Paradox…
always halfway to done, never done!
All Data is Relevant = Big Data

- Databases
- Email
- Web
- Desktops
- Servers
- DHCP/ DNS
- Network Flows
- Hypervisor
- Badges
- Firewall
- Authentication
- Vulnerability Scans
- Custom Apps
- Service Desk
- Storage
- Mobile
- Intrusion Detection
- Data Loss Prevention
- Anti-Malware
- Industrial Control
- Call Records
All Data is Relevant = Big Data

I don’t know how to ask four hundred systems if something changed!
The Value of Normalization

Makes things easier for a search user
Simple apps can play nicely together
Complex apps become far more useful
Normalization: Not just a dirty word

Normalizing at index time is pretty lame
Normalizing the data before it’s stored is VERY lame
Normalizing with tags and fields at search time is very AWESOME
The Splunk Common Information Model

Unlocking the Power of Data Normalization
Different products, same view!
Splunk did not invent the CIM...
…but is the only platform where it doesn’t suck

- **Splunk Common Information Model**
  - Makes things far easier for search users
  - Makes standalone apps more powerful
  - Makes enterprise apps possible

- **Splunk Technology Add-ons**
  - First “go-to” for CIM compliance
    - Look for “CIM Compliance” in Splunkbase
  - Translate data to the CIM
  - Get gnarly data into Splunk
Splunk CIM Architecture

Machines -> Data -> Information -> Users

Bits & Bytes → Extractions & Tags → Schema & Acceleration → Reports & Alerts

RAW DATA → TECHNOLOGY ADD-ONS → DATA MODELS → SEARCH
The Splunk CIM Add-On
CIM != Data Models, but it is part of the Implementation

- Set of Data Models representing least common denominator of domain

- Built upon normalized tags and fields generated at search time
- Enables correlation across data sources
- Simplifies app development
- Includes 22 preconfigured data models
CIM Powers Splunk Ecosystem

Premium Apps
- Splunk provided: ES, ITSI
- Partner apps can leverage too

Partner and Splunk Add-ons
- Bring data in
- Make that data easy to extract value from

Searching with Splunk across many data sources
How to Get Started

The path to CIM compliance
Start with Technology Add-ons!
• All work is done for you!
• Implements Field Extractions and Tags

Look for CIM Versions supported by the Add-on in Splunkbase
The CIM Data Models

- **Splunk_SA_CIM**
  - Packaged in premium apps
  - Domain specific datamodel definitions
    - Search>Pivot
    - Dataset structure
    - Acceleration
  - Provides CIM “dictionary”

- [https://splunkbase.splunk.com/app/1621/](https://splunkbase.splunk.com/app/1621/) (Splunk App)
CIM Data Model Details

- Alerts
- Application State
- Authentication
- Certificates
- Change Analysis
- Databases
- Data Loss Prevention
- Email
- Interprocess Messaging
- Intrusion Detection
- Inventory

- Java Virtual Machines
- Malware
- Network Resolution
- Network Sessions
- Network Traffic
- Performance
- Splunk Audit Logs
- Ticket Management
- Updates
- Vulnerabilities
- Web

ITSI Data models (Trial for CIM)

- Operating System
- Database
- Virtualization

- Load Balancer
- Application Server
- Web Server

New!
Why CIM?

- CIM != Data Models
- Universal way to refer to an object
- Consider
  - `destination_ip`
  - `d-ip`
  - `dstip`
  - `dest_ip`
  - `dst_ip`
  - `bob`
- CIM solves this, `dest_ip`
Splunk_SA_CIM

Check your compliance to CIM Data Models

- CIM Validation (S.o.S) datamodel
  - Basic tools to spot untagged or partially parsed data
Add-on Builder

Build CIM-compliant Add-ons and publish to Splunkbase!

Helpful when developing new content

- Point and click
- TA creation
- Field extractor
- CIM mapper
- Branding
- Best practice validator

Ready for certification

Congratulations, your add-on is ready to submit to Splunk App Certification.
SA-cim_vladiator
Comprehensive CIM validation

- Data preparation tool for data models
- Review datasets against particular data models
- Rapid verification and prototyping for TA’s
- Particularly helpful with premium apps
- Extendable to custom content
### SA-cim_vladiator

#### Data Model: Network_Traffic (and sub models) uses these fields:

<table>
<thead>
<tr>
<th>field</th>
<th>total_events</th>
<th>distinct_value_count</th>
<th>percent_coverage</th>
<th>field_values</th>
<th>is_cim_valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>action</td>
<td>2268</td>
<td>5</td>
<td>100.00</td>
<td>48.15% allowed&lt;br&gt;33.33% NONE&lt;br&gt;11.11% DROP&lt;br&gt;3.7% IGNORE&lt;br&gt;3.7% TRAFFIC_PACTION_NOTIFY</td>
<td>! found 4 unexpected values (NONE, DROP, IGNORE, TRAFFIC_PACTION_NOTIFY)</td>
</tr>
<tr>
<td>app</td>
<td>2268</td>
<td>5</td>
<td>70.37</td>
<td>29.65% NONE&lt;br&gt;18.52% NULL&lt;br&gt;11.11% GSL&lt;br&gt;7.41% HTTP</td>
<td>! event coverage less then 90%</td>
</tr>
<tr>
<td>bytes</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td>! no extracted values found</td>
</tr>
<tr>
<td>bytes_in</td>
<td>2268</td>
<td>1</td>
<td>18.52</td>
<td>0.18% 0</td>
<td>! event coverage less then 90%</td>
</tr>
<tr>
<td>channel</td>
<td>0</td>
<td>0</td>
<td></td>
<td>! no extracted values found</td>
<td>looking good!</td>
</tr>
<tr>
<td>dest</td>
<td>2268</td>
<td>75</td>
<td>100.00</td>
<td>0.79% 10.11.36.43&lt;br&gt;0.67% 10.11.36.49&lt;br&gt;0.57% 10.11.36.11</td>
<td>looking good!</td>
</tr>
<tr>
<td>dest_ip</td>
<td>2268</td>
<td>75</td>
<td>100.00</td>
<td>0.79% 10.11.36.43&lt;br&gt;0.67% 10.11.36.49&lt;br&gt;0.57% 10.11.36.11</td>
<td>looking good!</td>
</tr>
</tbody>
</table>

- **Check for unexpected values**: ! found 4 unexpected values (NONE, DROP, IGNORE, TRAFFIC_PACTION_NOTIFY)
- **Check for extraction coverage**: ! event coverage less then 90%
- **Check for missing extractions**: ! no extracted values found
- **Eat a cookie!**
SA-cim_vladiator: What else?

- Ability to assign score to dataset
- Ability to monitor score over time
- Ability to detect data format changes
  - Oh, remember that code upgrade last month? Log format changed…
  - Alert me!

<table>
<thead>
<tr>
<th>Total fields</th>
<th>Issue fields</th>
<th>% CIM Compliance</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>40</td>
<td>17%</td>
</tr>
</tbody>
</table>
Key Takeaways

1. CIM sets you up for success as your Splunk environment grows in size and sophistication

2. CIM != Data Models

3. Use Add-ons and the available tools to make your life easier
UX “Getting Data In” Survey

http://tinyurl.com/GDI-conf2017
Don't forget to rate this session in the .conf2017 mobile app