© 2018 SPLUNK INC.

Splunk user segmentation for effective user enablement & adoption

Drinking the “Splunk Champagne” by applying Statistics & Machine Learning to Splunk’s internal logs

Anand Ladda | Senior Solutions Engineer

October 2018

© 2018 SPLUNK INC.

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. © 2018 Splunk Inc. All rights reserved.

splunk> .conf18

© 2018 SPLUNK INC.

Whoami

Anand Ladda

- Senior Solutions Engineer@Splunk supporting large business in the DC Metro Area
- 15+ years in Data & Analytics Space
- Currently working on a project at the intersection of my 3 passions – Cricket, Splunk & ML

© 2018 SPLUNK INC.

Am I in the right place?

Some familiarity with…

- Splunk Audit logs
- Search Processing Language (SPL)
- Basic Stats & ML concepts

splunk> .conf18

© 2018 SPLUNK INC.

Agenda

- Motivation - Machine Learning on user activity data
- What Splunk Logs about itself - “Monitor the monitors”
  - _internal, _audit, _introspection
  - What are the users up to?
- Demo – Where to find the data
- ML in Splunk – Quick Overview
- Applying ML to user activity data
- Demo - How it all comes together
- Real World Use Case Outcome
- Next Steps

splunk> .conf18

© 2018 SPLUNK INC.

Motivation - Machine Learning on user activity data

- As Splunk adoption increases, users progress at different rates. Segmentation of users is key to building targeted training programs
- Splunk logs information about user activity
- Splunk's for Analytics & Data Science course covered a use case around "Retail Market Segmentation" – Segmenting customers based on recency, frequency and monetary spend to build effective marketing campaign
- Extending that same concept to user activity data – Segment Splunk users by recency (of searches), frequency (of searches) and spend (search execution time)

© 2018 SPLUNK INC.

What Splunk logs about itself

Monitor the monitors

splunk> .conf18

© 2018 SPLUNK INC.

What Splunk Logs about itself

- _internal - Primary index capturing a variety of data about Splunk's internal state. Useful in troubleshooting issues within the Splunk "stack". Types of logs captured in this index
  - **splunkd.log** - The primary log written to by the Splunk server. Any stderr messages generated by scripted inputs, scripted search commands, and so on, are logged here.
  - metrics.log - Contains periodic snapshots of Splunk performance and system data, in particular around indexing queue performance, throughput, etc
- **_audit** - Information about user activity - about user log on failures/successes, modifying a setting, updating a lookup, search statistics by user, etc.
- _introspection –Introspection data about your Splunk instance and environment. Metrics like OS resource usage, disk I/O data, KV store performance data, to aid in reporting on system resource utilization and troubleshooting problems with your Splunk Enterprise deployment

What are the users up to?

Audit logs to the rescue

© 2018 SPLUNK INC.

How many searches is a user running, their runtimes and recency?

Search Activity by User (34)

	User	Search Count	Median Runtime	90th Percentile Runtime	Cumulative Runtime	Last Search
1		175	0.42s	0.69s	1min 20.89s	2018-09-05 21:41:45
2		31	0.63s	24.14s	4min 52.47s	2018-09-04 13:05:02
3		15	0.40s	0.64s	6.85s	2018-09-05 08:52:22
4		7	0.38s	0.68s	3.04s	2018-09-05 15:44:31
5		54	0.45s	1.52s	2min 21.60s	2018-09-01 06:14:55
6		26	0.39s	1.03s	47.50s	2018-09-05 07:19:21
7		11	0.50s	0.67s	5.47s	2018-09-05 09:30:07
8		78	0.35s	0.94s	1min 7.90s	2018-09-03 04:03:46
9		8	0.36s	0.61s	2.98s	2018-09-03 02:40:43
10		22	0.66s	1.27s	15.51s	2018-09-04 14:08:47

How efficient are the scheduled searches?

Saved Search Name	User	Efficiency	App	Host	Avg Runtime Secs	Weekly Count	Total Runtime Secs	Ran Every X Mins	Avg Runtime In Mins
My 93rd Alert		0.0011	search	ch-demo-dod	47076.7161	11193	52396385.027	0.9006	784.6119
Alert		0.0021	splunk_app_aws	ch-demo-aws41	25854.0952	11017	284834567.169	0.9149	430.9016
SendNotablesToUBA		80.1633	search	ch-demo-zeus	3.7554	2009	7544.599	5.0174	0.0626
Bad Status Codes > 10		100.5803	search	ch-demo-itsi.hod.cloud	0.5982	10052	6012.467	1.0028	0.0100
Product Sold by Brand		135.3485	webidemo	ch-demo-hunk	26.5980	168	4468.471	60.0000	0.4433
Customer Activity Map		183.8695	webidemo	ch-demo-hunk	19.5791	168	3289.290	60.0000	0.3263
Reopened incident is too many		213.3711	splunk_app_servicenow	ch-demo-snow.hod.cloud	1.4109	2009	2834.533	5.0174	0.0235
Kepware mcollect		219.2982	Windfarms	ch-demo-iot2	1.3680	2016	2757.881	5.0000	0.0228
Summary - Kepware Windfarm		304.6255	search	ch-demo-iot2	0.1971	10073	1985.390	1.0007	0.0033
Accelerated Search New		430.4729	webidemo	ch-demo-hunk	8.7265	161	1404.967	62.6087	0.1454

What are the common search commands?

	Command	Count	Average Runtime	Max Runtime
1	search	169	1min 44.44s	1h 1min 19.59s
2	typeahead	120	0.44s	1.98s
3	metadata	49	2min 6.69s	44min 52.46s
4	table	29	25.46s	1min 48.05s
5	fieldformat	24	28.29s	1min 48.05s
6	timechart	10	2min 16.35s	16min 43.76s
7	stats	7	11.15s	32.49s
8	spath	7	0.16s	0.17s
9	fields	5	35.91s	1min 3.00s
10	eval	5	12.35s	47.83s

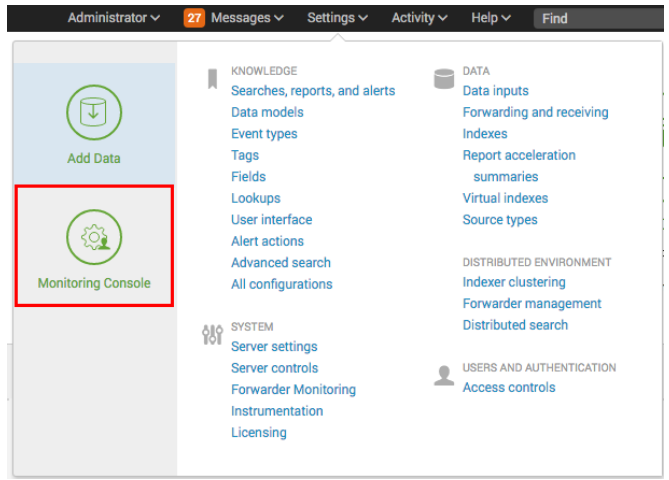
And many more…

splunk> .conf18

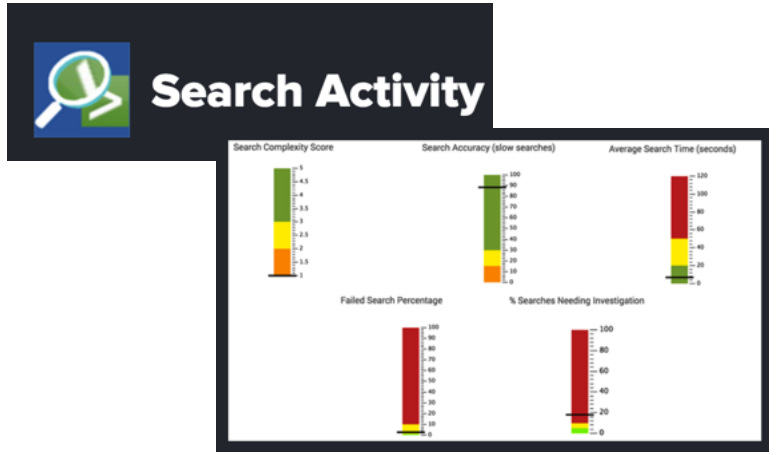
© 2018 SPLUNK INC.

Splunk Apps that provide insights on user activity

- **Monitoring Console** - Monitoring Console is the Splunk Enterprise monitoring tool. It lets you view detailed topology and performance information about your Splunk Enterprise deployment. Before Splunk Enterprise version 6.5.0, the Monitoring Console was called the Distributed Management Console



- **Search Activity App -** Search Activity helps Splunk champions monitor users, grow usage, and understand personas. It provides metrics on use, organizational information, and adoption



splunk> .conf18

© 2018 SPLUNK INC.

Where to find the data

I need it stat

© 2018 SPLUNK INC.

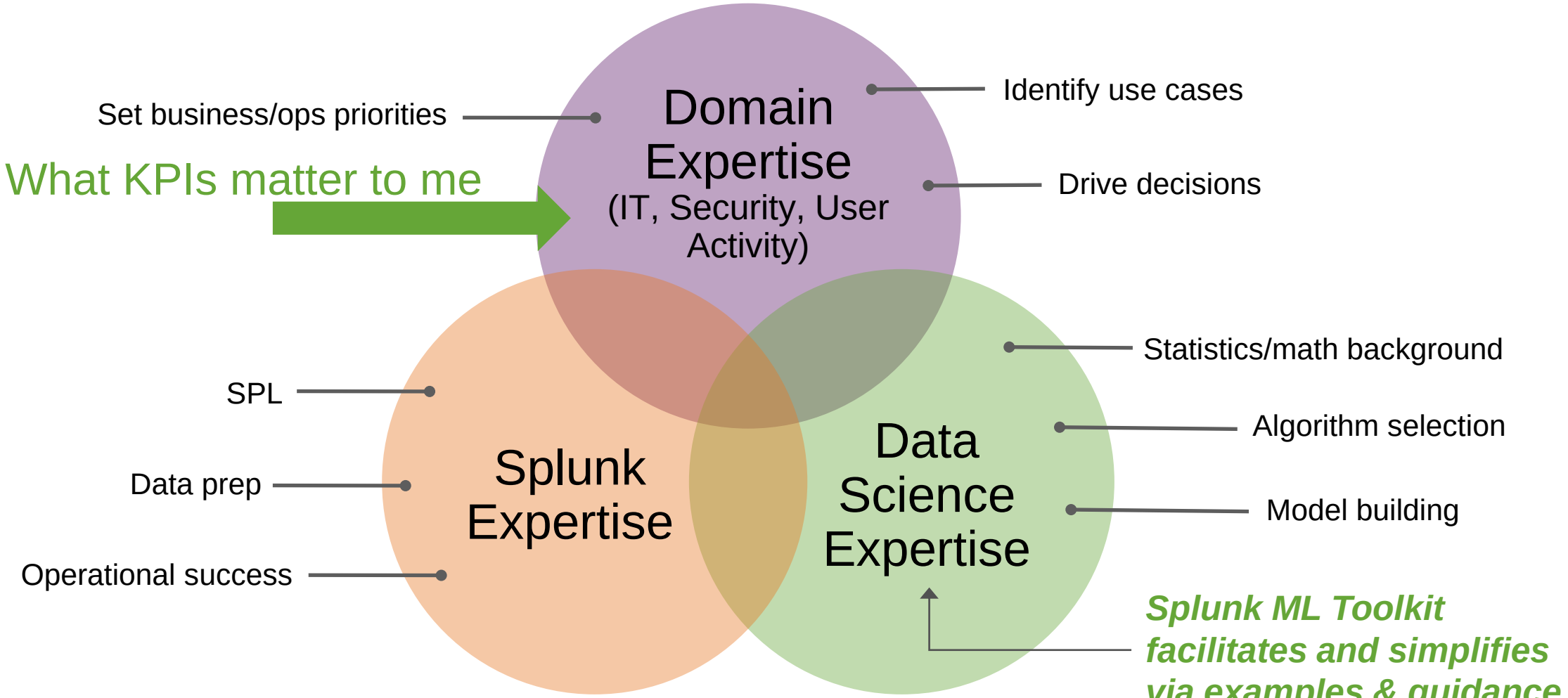
Machine Learning in Splunk

Quick Overview

splunk> .conf18

© 2018 SPLUNK INC.

Custom Machine Learning – Success Formula



splunk> .conf18

© 2018 SPLUNK INC.

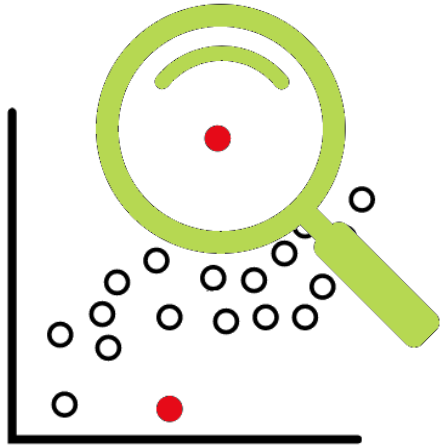
Overview of ML at Splunk

splunk> .conf18

© 2018 SPLUNK INC.

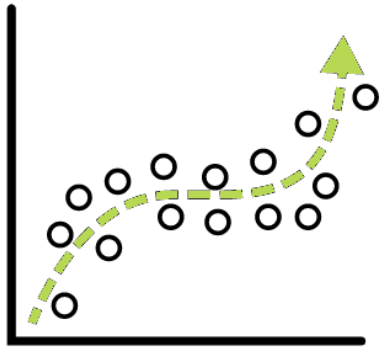
Splunk Customers Have ML Problems

Anomaly detection



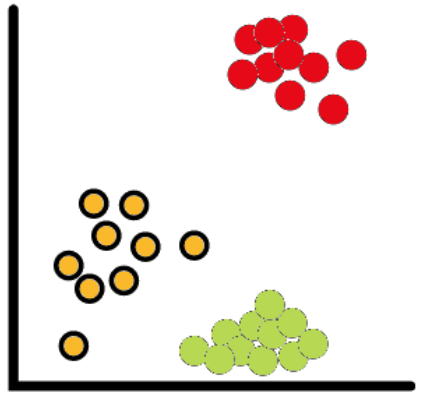
Deviation from past behavior
Deviation from peers
(aka Multivariate AD or Cohesive AD)
Unusual change in features
ITSI Anomaly Detection

Predictive Analytics



Predict Service Health Score
Predicting Churn
Predicting Events
Trend Forecasting
Early warning of failure – predictive maintenance

Clustering



Identify peer groups
Event Correlation
Reduce alert noise
ITSI Event Analytics

splunk> .conf18

© 2018 SPLUNK INC.

Want to learn more about ML in Splunk

Go see these talks!

FN1398 - Splunk and the Machine Learning Toolkit in Action: Customer Use Cases

- Wednesday, Oct 03, 12:45 p.m. - 1:30 p.m.

FN1418 - Getting Your Data Ready for Machine Learning

- Wednesday, Oct 03, 12:45 p.m. - 1:30 p.m.

FN1424 - Overview of AI/ML Capabilities Across Our Portfolio

- Tuesday, Oct 02, 2:15 p.m. - 3:00 p.m.

splunk> .conf18

© 2018 SPLUNK INC.

Applying ML to User Activity Data

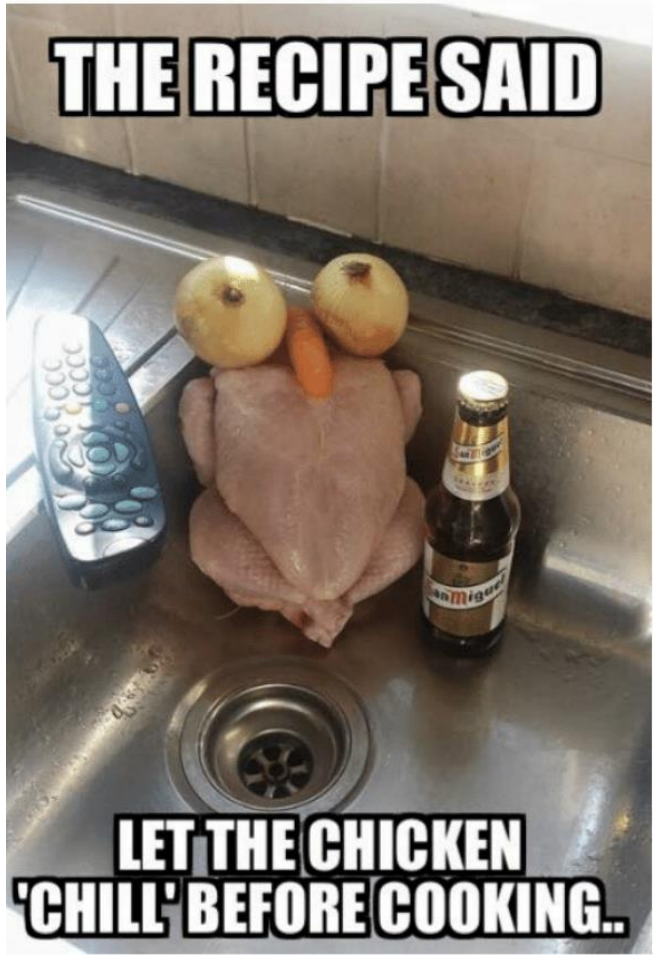
Recipe

splunk> .conf18

© 2018 SPLUNK INC.

Recipe

1. Use **stats** to get the user activity data which contains runtimes, search count and recency.
2. **Review** data using fieldsummary
3. **Clean** the data by removing outliers using **eventstats**
4. **Normalize** “features” using **StandardScaler**
5. **Fit** into **clusters** using **kmeans**



splunk> .conf18

© 2018 SPLUNK INC.

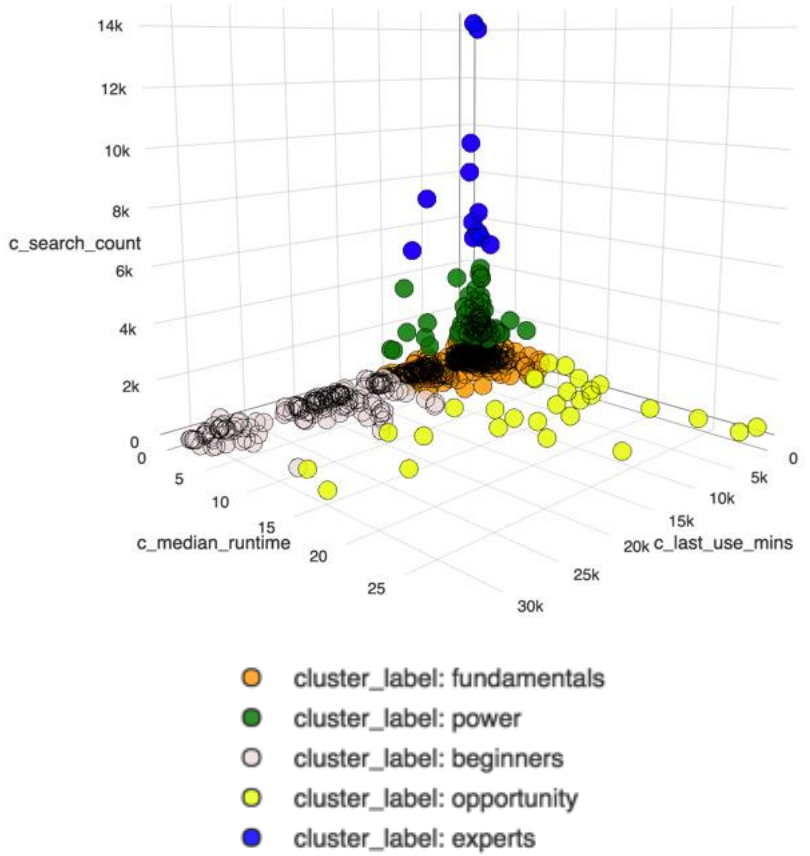
How it all comes together - Demo

Proof is in the pudding

© 2018 SPLUNK INC.

Real World Use Case Outcome

- Results from analysis of 30 days of search activity data from the “real world”
- Clusters plotted along the original axes for ease of understanding
  - **Beginners** Possibly new to Splunk users with low search counts and infrequent searches. Target them with training sessions, user group meetings, L n Ls – Bring them into the fold!
  - **Fundamentals** Got the basics right. Continue to build them up keeping them informed of new features and capabilities
  - **Power** are burgeoning experts. Headed in the right direction and we want to keep it that way with advanced training and targeted LnL sessions
  - **Experts** Heavy hitters! Help them Pay it Forward
  - **Opportunity** Group that can benefit from some basics in Splunk training



splunk> .conf18

© 2018 SPLUNK INC.

What's Next

- Overlay Splunk training transcripts data to correlate user activity with skill level
- Use the cohorts to build targeted enablement programs – Splunk training, lunch & learns, wiki updates, etc
- User other aspects of the audit logs to build a better understanding of Splunk usage within your customer base
- Try the monitoring console & search activity apps to assist in this process

© 2018 SPLUNK INC.

Thank You & Questions

Don't forget to rate this session in the .conf18 mobile app

© 2018 SPLUNK INC.

Appendix – SPL for examples shown

Ah Here you are!

© 2018 SPLUNK INC.

How many searches is a user running, their runtimes and recency?

Search from one of the panels in the Monitoring Console

Search Activity by User (34)

	User	Search Count	Median Runtime	90th Percentile Runtime	Cumulative Runtime	Last Search
1	abearsley	175	0.42s	0.69s	1min 20.89s	2018-09-05 21:41:45
2	agrant	31	0.63s	24.14s	4min 52.47s	2018-09-04 13:05:02
3	agutknecht	15	0.40s	0.64s	6.85s	2018-09-05 08:52:22
4	bting	7	0.38s	0.68s	3.04s	2018-09-05 15:44:31
5	chwang	54	0.45s	1.52s	2min 21.60s	2018-09-01 06:14:55
6	clin	26	0.39s	1.03s	47.50s	2018-09-05 07:19:21
7	cputnam	11	0.50s	0.67s	5.47s	2018-09-05 09:30:07
8	cwinata	78	0.35s	0.94s	1min 7.90s	2018-09-03 04:03:46
9	dlambrou	8	0.36s	0.61s	2.98s	2018-09-03 02:40:43
10	dwaters	22	0.66s	1.27s	15.51s	2018-09-04 14:08:47

```
`sim_set_search_head` `sim_audit_get_searches_for_groups(*)`
| stats min(_time) as _time, values(user) as user, max(total_run_time) as total_run_time, first(search) as search, first(search_type) as search_type, first(apiStartTime) as apiStartTime, first(apiEndTime) as apiEndTime by search_id, host
| where isnotnull(search) AND search_type="ad hoc"
| search host=* `sim_set_search_head`
| stats dc(host) as count_host, median(total_run_time) as median_runtime, sum(total_run_time) as cum_runtime, count(search) as count, max(_time) as last_use by user
| eval median_runtime = if(isnotnull(median_runtime), median_runtime, "-")
| eval cum_runtime = if(isnotnull(cum_runtime), cum_runtime, "-")
| `sim_time_format(last_use)`
| fields user, count, count_host, median_runtime, cum_runtime, last_use
| sort - count
| rename user as User, count_host as "Search Head Count", count as "Search Count", median_runtime as "Median Runtime", cum_runtime as "Cumulative Runtime", last_use as "Last Search"
| fieldformat "Median Runtime" = `sim_convert_runtime('Median Runtime')`
| fieldformat "Cumulative Runtime" = `sim_convert_runtime('Cumulative Runtime')`
```

splunk> .conf18

© 2018 SPLUNK INC.

How efficient are the scheduled searches?

Courtesy Splunk Professional Services

- Description: The efficiency panel is a ranking of searches based on how efficient the searches are. The value represents a function of how often the search runs and how long it takes to run. A search running often and takes a long time will have a low efficiency value. Searches that run in less time raise efficiency value. Higher efficiency values, relative to each other, are better. Anything below 10 should be considered for improvement in SPL, time range, or change in frequency of scheduling.

Saved Search Name	User	Efficiency	App	Host	Avg Runtime Secs	Weekly Count	Total Runtime Secs	Ran Every X Mins	Avg Runtime In Mins
My 93rd Alert		0.0011	search	ch-demo-dod	47076.7161	11193	52396385.027	0.9006	784.6119
Alert		0.0021	splunk_app_aws	ch-demo-aws41	25854.0952	11017	284834567.169	0.9149	430.9016
SendNotablesToUBA		80.1633	search	ch-demo-zeus	3.7554	2009	7544.599	5.0174	0.0626
Bad Status Codes > 10		100.5803	search	ch-demo-itsi.hod.cloud	0.5982	10052	6012.467	1.0028	0.0100
Product Sold by Brand		135.3485	webidemo	ch-demo-hunk	26.5980	168	4468.471	60.0000	0.4433
Customer Activity Map		183.8695	webidemo	ch-demo-hunk	19.5791	168	3289.290	60.0000	0.3263
Reopened incident is too many		213.3711	splunk_app_servicenow	ch-demo-snow.hod.cloud	1.4109	2009	2834.533	5.0174	0.0235
Kepware mcollect		219.2982	Windfarms	ch-demo-iot2	1.3680	2016	2757.881	5.0000	0.0228
Summary - Kepware Windfarm		304.6255	search	ch-demo-iot2	0.1971	10073	1985.390	1.0007	0.0033
Accelerated Search New		430.4729	webidemo	ch-demo-hunk	8.7265	161	1404.967	62.6087	0.1454

```
index=_internal sourcetype=scheduler source=*scheduler.log (user=*)
| stats avg(run_time) as average_runtime_in_sec count(savedsearch_name) as weekly_count sum(run_time) as total_runtime_sec by savedsearch_name user app host
| eval Ran_every_x_Minutes=round(60/(weekly_count/168), 4)
| eval average_runtime_in_minutes=round((average_runtime_in_sec/60), 4)
| eval average_runtime_in_sec=round(average_runtime_in_sec, 4)
| eval efficiency=round(((60/(weekly_count/168))/(average_runtime_in_sec/60)), 4)
| sort efficiency
| rename savedsearch_name AS "Saved Search Name", user AS "User", efficiency AS "Efficiency", app AS "App", host AS "Host", average_runtime_in_sec AS "Avg Runtime Secs", weekly_count AS "Weekly Count", total_runtime_sec AS "Total Runtime Secs", Ran_every_x_Minutes AS "Ran Every X Mins", average_runtime_in_minutes AS "Avg Runtime In Mins"
| table "Saved Search Name","User", "Efficiency", "App", "Host", "Avg Runtime Secs", "Weekly Count", "Total Runtime Secs", "Ran Every X Mins", "Avg Runtime In Mins"mulative Runtime" = `sim_convert_runtime('Cumulative Runtime')`
```

splunk> .conf18

© 2018 SPLUNK INC.

What are the common Splunk commands

Panel from Splunk's Monitoring Console

	Command	Count	Average Runtime	Max Runtime
1	search	169	1min 44.44s	1h 1min 19.59s
2	typeahead	120	0.44s	1.98s
3	metadata	49	2min 6.69s	44min 52.46s
4	table	29	25.46s	1min 48.05s
5	fieldformat	24	28.29s	1min 48.05s
6	timechart	10	2min 16.35s	16min 43.76s
7	stats	7	11.15s	32.49s
8	spath	7	0.16s	0.17s
9	fields	5	35.91s	1min 3.00s
10	eval	5	12.35s	47.83s

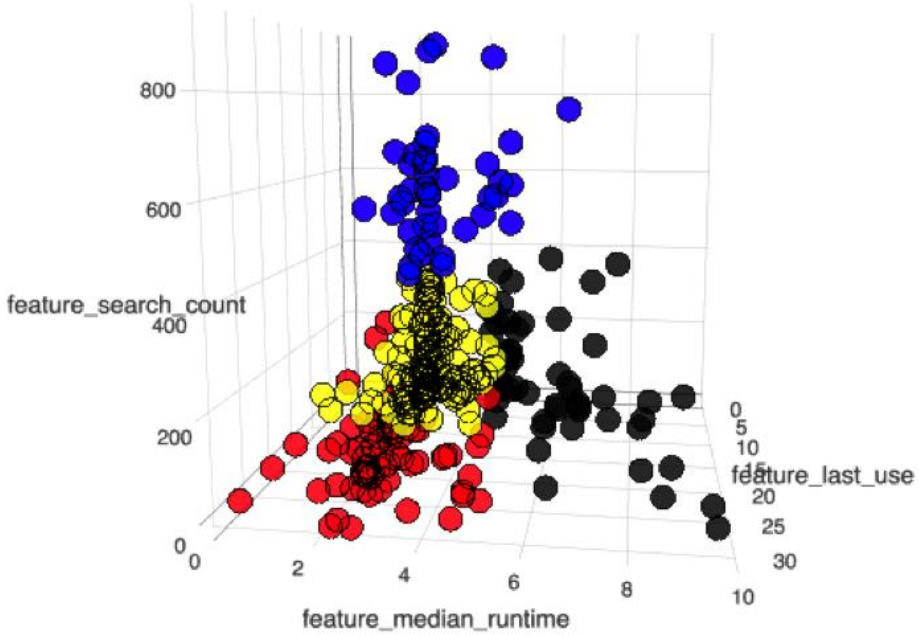
```
`dmc_audit_get_searches(CHSH01)`
    | stats min(_time) as _time, values(user) as user, max(total_run_time) as total_run_time, first(search) as search, first(search_type) as search_type, first(apiStartTime) as apiStartTime, first(apiEndTime) as apiEndTime by search_id
    | where isnotnull(search) AND search_type="ad hoc" | eval commands = commands(search)
| streamstats window=1 values(commands) as commands
| stats count avg(total_run_time) as avg_runtime max(total_run_time) as max_runtime by commands
| eval avg_runtime = round(avg_runtime, 2)
| eval max_runtime = round(max_runtime, 2)
| sort - count, - max_runtime, - avg_runtime
| rename commands as Command, avg_runtime as "Average Runtime", max_runtime as "Max Runtime", count as "Count"
| eval "Average Runtime" = `dmc_convert_runtime('Average Runtime')`
| eval "Max Runtime" = `dmc_convert_runtime('Max Runtime')`
```

splunk> .conf18

© 2018 SPLUNK INC.

Clustering Users based on Search Activity

- Finished SPL Product




```
`sim_set_search_head` `sim_audit_get_searches_for_groups(*)`
| stats min(_time) as _time, values(user) as user, max(total_run_time) as total_run_time, first(search) as search, first(search_type) as search_type,
first(apiStartTime) as apiStartTime, first(apiEndTime) as apiEndTime by search_id, host
| where isnotnull(search)
| stats median(total_run_time) as c_median_runtime, count(search) as c_search_count, max(_time) as last_use by user
| eval c_last_use_secs = now() - last_use
| fields - last_use
| eventstats p99(c_median_runtime) as p99_r p99(c_search_count) as p99_s p99(c_last_use_secs) as p99_l
| where c_median_runtime < p99_r AND c_search_count < p99_s AND c_last_use_secs < p99_l
| fit StandardScaler c_*
| fit KMeans k=5 SS* into clusterer
| table cluster SS_*
```