.conf2015

What You Think Is Real Is Not Real, Learn How Splunk Uncovered The Truth!

Jeff Kent

President m-mobo

Alex Gitelzon

System Administrator, APM

Dennis Morton

Splunk Expert m-mobo

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.

.conf2015

How Do You Know What Is Real?

splunk>

Video

Subconscious Test

5 seconds to read the sentence, only read it once and count the number of letter “F”

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS

# Of F's

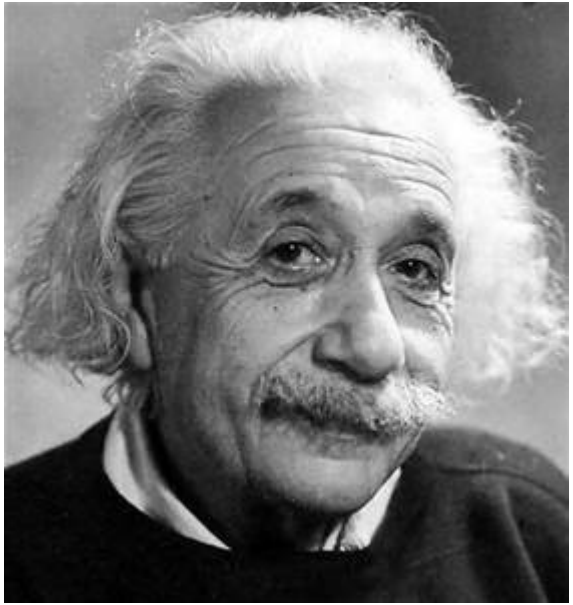
How many counted 3?

FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS

Agenda

- APM Business Model And Challenges
- Splunk Architecture And Initial Configurations
- Problems
- Solutions
- Demo
- Lessons Learned

M-mobo



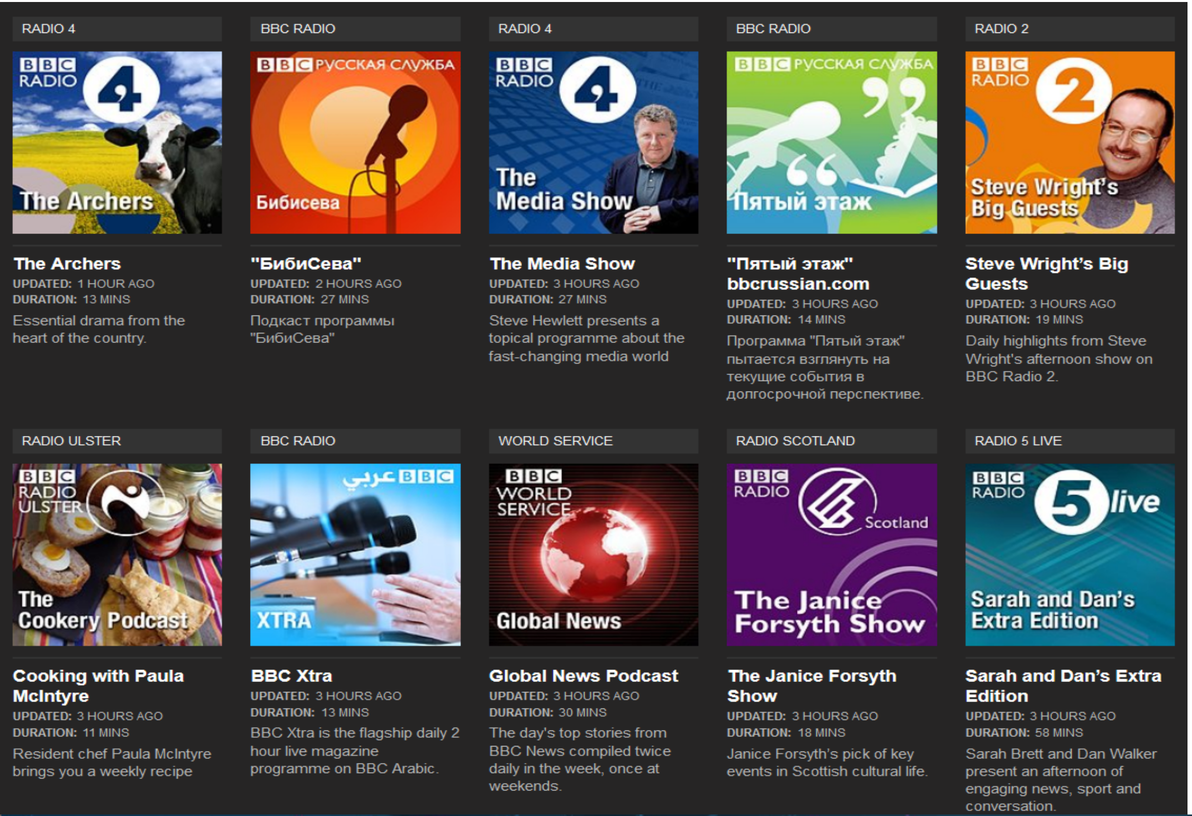
American Public Media (APM)

- 90 Radio stations in the Midwest, California, Florida
- 20 Nationally distributed programs
- 900 Stations carry programming to 18 million listeners
- Winner of multiple Media Awards

Business Model

- Revenue
  - Generated through Ads
  - Listeners or subscribers
- Expenses
  - Content Delivery Networks (CDN)
- Important metrics
  - “What” Popular programming
  - “How” are they consuming the content
  - “When” are they listening
  - “Where” what advertising markets

Challenges



RADIO 4

The Archers
UPDATED: 1 HOUR AGO
DURATION: 13 MINS
Essential drama from the heart of the country.

BBC RADIO

"БибиСева"
UPDATED: 2 HOURS AGO
DURATION: 27 MINS
Подкаст программы "БибиСева"

RADIO 4

The Media Show
UPDATED: 3 HOURS AGO
DURATION: 27 MINS
Steve Hewlett presents a topical programme about the fast-changing media world

BBC RADIO

"Пятый этаж" bbcrussian.com
UPDATED: 3 HOURS AGO
DURATION: 14 MINS
Программа "Пятый этаж" пытается взглянуть на текущие события в долгосрочной перспективе.

RADIO 2

Steve Wright's Big Guests
UPDATED: 3 HOURS AGO
DURATION: 19 MINS
Daily highlights from Steve Wright's afternoon show on BBC Radio 2.

RADIO ULSTER

Cooking with Paula McIntyre
UPDATED: 3 HOURS AGO
DURATION: 11 MINS
Resident chef Paula McIntyre brings you a weekly recipe

BBC RADIO

BBC Xtra
UPDATED: 3 HOURS AGO
DURATION: 13 MINS
BBC Xtra is the flagship daily 2 hour live magazine programme on BBC Arabic.

WORLD SERVICE

Global News Podcast
UPDATED: 3 HOURS AGO
DURATION: 30 MINS
The day's top stories from BBC News compiled twice daily in the week, once at weekends.

RADIO SCOTLAND

The Janice Forsyth Show
UPDATED: 3 HOURS AGO
DURATION: 18 MINS
Janice Forsyth's pick of key events in Scottish cultural life.

RADIO 5 LIVE

Sarah and Dan's Extra Edition
UPDATED: 3 HOURS AGO
DURATION: 58 MINS
Sarah Brett and Dan Walker present an afternoon of engaging news, sport and conversation.

Initial Splunk Configurations

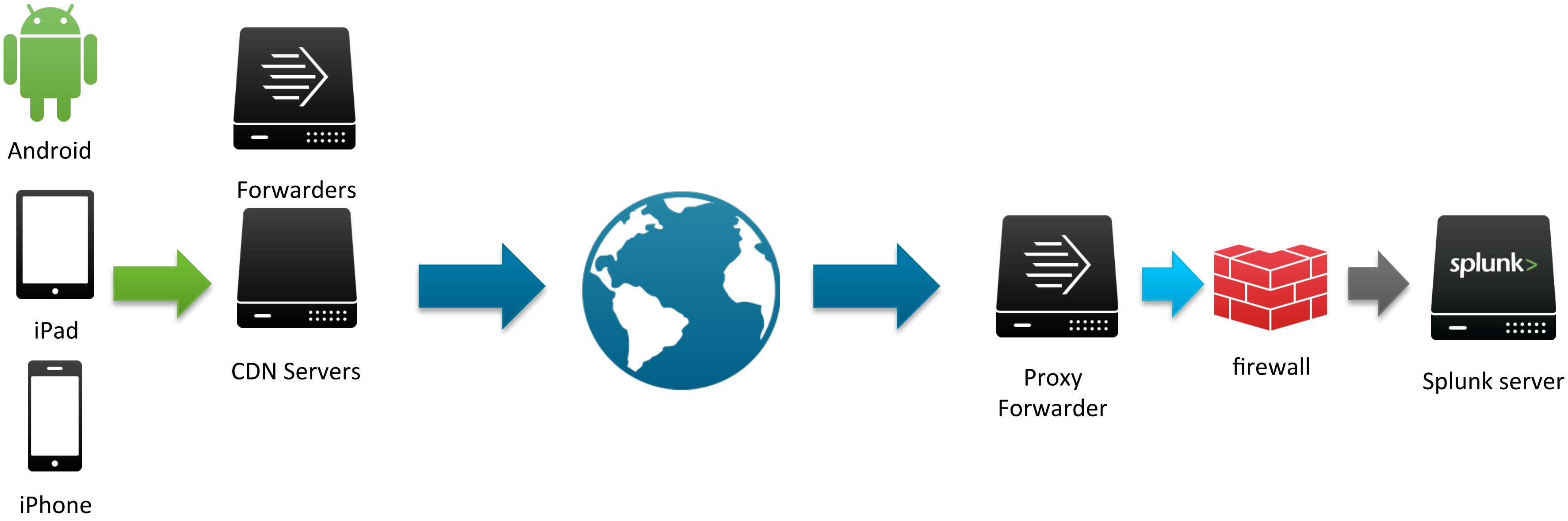
Early successes with Splunk:

- Initial data loading was relatively easy
- We very quickly identified an issue with download counts

Initial challenges with Splunk:

- Reports were slow, taking over an hour to run for a months worth of data
- Lookup table for agent user strings were inaccurate
- Could not get a consolidated view of all listener behavior across sources
- Vendor fields kept changing
- Issues with collecting the exact data from an application called Wowza

Splunk Architecture



Example log

```
203.0.113.110 "" 2015-07-01 04:59:59 "GET /podcast/marketplace/
segments/2015/06/30/mp_20150630_seg_21_64.mp3?
listeningSessionID=558878963c788f58_719653_N3egfk0g_0000000hRr5
HTTP/1.1" 200 1946331 "" "Marketplace/2014.04.24.1100 CFNetwork/
672.1.15 Darwin/14.0.0" 3
```

Marketplace App on iPhone

```
192.0.2.150 "" 2015-07-01 04:59:59 "GET /podcast/marketplace/pm/
2015/06/30/pm_20150630_pod_64.mp3 HTTP/1.1" 302 255 "" "Dalvik/2.1.0
(Linux; U; Android 5.0; SM-N900P Build/LRX21V)" 0
```

Android Phone

```
203.0.113.223 "" 2015-07-01 04:59:59 "GET /podcast/marketplace/
podcast_nu/2015/06/19/weekend_20150619_pod_nu_64.mp3 HTTP/1.1"
302 271 "" "Dalvik/1.4.0 (Linux; U; Android 2.3.4; Kindle Fire Build/
GINGERBREAD)" 31
```

Kindle Fire

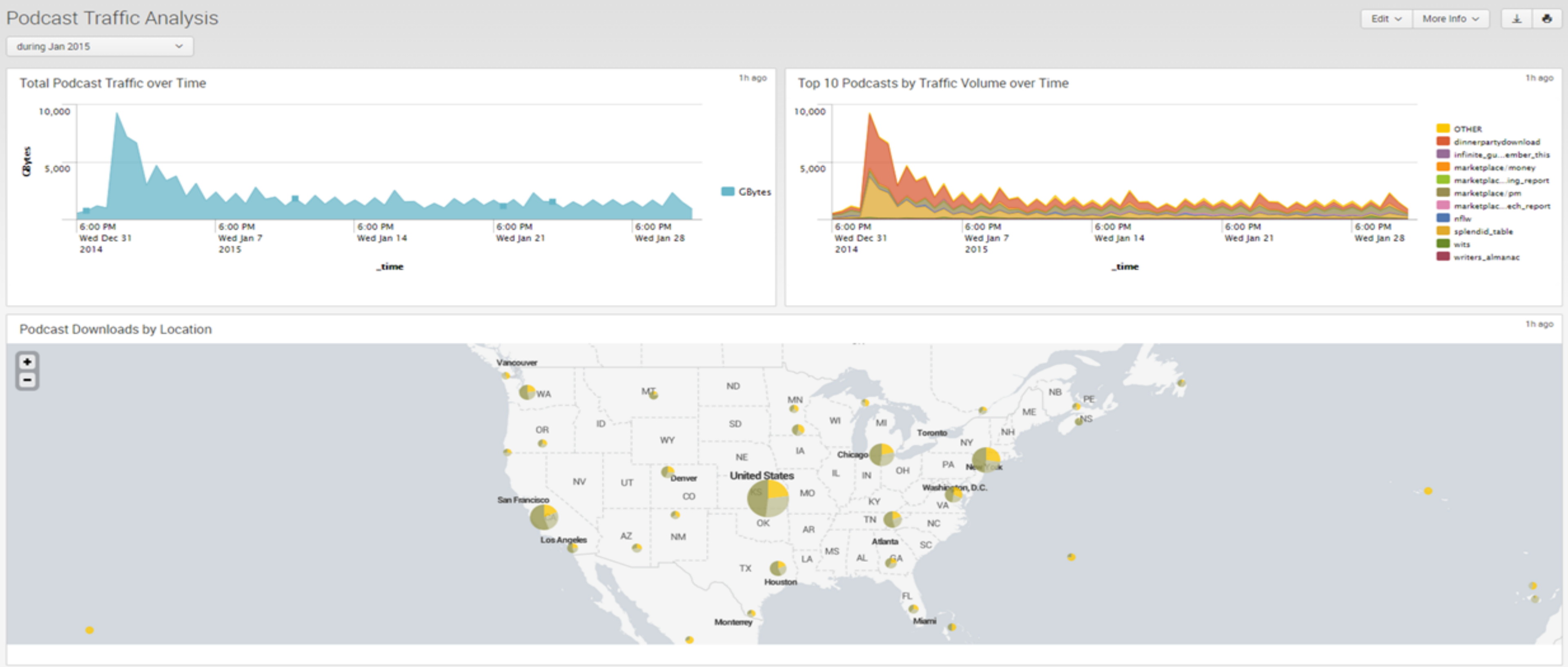
```
198.51.100.133 "" 2015-07-01 04:59:59 "GET /podcast/nflw/2015/06/27/
nflw_20150627_64.mp3?
listeningSessionID=5588552cf8d07783_728723_4UwUCKLg_0000000LfT4
HTTP/1.1" 206 297 "" "AppleCoreMedia/1.0.0.12F70 (iPhone; U; CPU OS 8_3
like Mac OS X; en_us)" 0
```

iPhone iOS 8.3

Code From Initial Configurations And Fixes?

- index=aod Podcast_Name=table OR Podcast_Name=splendid_table | dedup downloadId | timechart count by file limit=20
- index=aod status=302 | stats first(Podcast_Name) as Podcast_Name count(uri) as count by uri | sort Podcast_Name asc| fields Podcast_Name,uri,count
- How to make all searches start at 7 days instead of all-time?
- In /opt/splunk/etc/system/local/ui-prefs.conf
  - [search]
  - dispatch.earliest_time = -7d@d
  - dispatch.latest_time = now

Report Example



Issues with Existing Tool

Definition	Existing tool	Splunk
Sessions	30 min – 2 hours	Seconds to minutes
Active Listeners	30 min – 2 hours	Seconds to minutes
Implement report changes	Days to Months	Within Seconds
Outliers	Couldn’t identify	Identified
Geolocation	No	Yes

Robots

- One of the early reports that we created was to add up the bytes by IP address
- Splunk allows easy drill down into data
- That allowed us to see that we had a very large amount of traffic from one ip address

Fixing Consolidated View

- index=aod status=20* method=GET
- | eval listeningSessionID = if(isnull(listeningSessionID) OR listeningSessionID="", downloadId, listeningSessionID)
- | stats sum(bytes) as TotalSize by Podcast_Name listeningSessionID clientip uri_path date_hour
- | where TotalSize > 100000
- | chart count(Podcast_Name) as TotalCount over Podcast_Name
- | sort - TotalCount

- |`tstats` sum(Web.bytes) as TotalSize from datamodel=Web where Web.status=20* AND Web.http_method=GET AND Web.App=adswizz by Web.podcast_name, Web.client_id, Web.src, Web.uri_path, Web.date_hour
- | where TotalSize > 100000
- | `drop_dm_object_name("Web")`
- | chart count(podcast_name) as TotalCount over podcast_name
- | sort - TotalCount

Solution Approach

- Leverage the accelerated Web CIM
- Add additional fields where appropriate to the standard datamodel
- Rewrite existing reports using “tstats”
- Add additional panels with valuable information

Why Accelerated Data Models

- Three main methods for making searches faster:
  - Report Acceleration
  - Summary Indexing
  - Data Model Acceleration
- Each has their advantages/disadvantages
- Chose data model acceleration because...
  - Easy to "normalize" the various inputs - one report to cover all data sources!
  - Easy to support and extend vs other options
  - Freakin' fast...
  - Splunk uses it in their very popular Enterprise Security product!

Data Inputs Overview

- Three types of data:
  - Adswizz, Wowza, and Icecast
  - All fairly stock Apache-like logs
  - Radically different names for the same fields
- Adswizz example:

```
24.189.xxx.xxx "" 2015-08-20 20:34:03 "GET /podcast/marketplace/
podcast_nu/2015/08/20/hlppodcast2_64.mp3 HTTP/1.1" 302 260 "http://
www.marketplace.org/topics/wealth-poverty/york-fig/york-fig-where-
are-they-now" "Mozilla/5.0 (Windows NT 5.1) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/44.0.2403.107 Safari/537.36" 0
```

- Goal is to map each specific data sources fields into the Web CIM
- See here for Web CIM information:
  - http://docs.splunk.com/Documentation/CIM/latest/User/Web#Fields_for_Web_event_objects

Adswizz CIM Field Mapping Example

- EVAL-app = "adswizz"
- FIELDALIAS-adswizz-web-cim = **method AS http_method** clientip AS src timetaken AS duration **useragent AS http_user_agent** uri AS url referer AS http_referrer bytes AS bytes_out Podcast_Name AS podcast_name AS media_name root AS category
- EVAL-is_ondemand=if(match(uri_path,"^/podcast"),0,1)
- EVAL-is_podcast=if(match(uri_path,"^/podcast"),1,0)
- EVAL-**client_id** = if(isnull(listeningSessionID) OR listeningSessionID="", downloadId, listeningSessionID)

Adswizz CIM Field Mapping Example (cont.)

- Need to do more than just map fields to the CIM – must ensure it's tagged properly!
- Recommended method is to use an eventtype and tag off of that.
- Example:
  - Eventtypes.conf

```
[adswizz]
search = index=aod sourcetype=aod
#tags = web
```

- Tags.conf

```
[eventtype=adswizz]
web = enabled
```

Using Acelerated Data – tstats Basics

- Similar to stats, but different and a bit strange
- Example #1: table of all podcast download errors:

```
| tstats count from datamodel=Web where Web.is_podcast=1 AND
(Web.status=40* OR Web.status=50*) by Web.podcast_name,Web.status
```

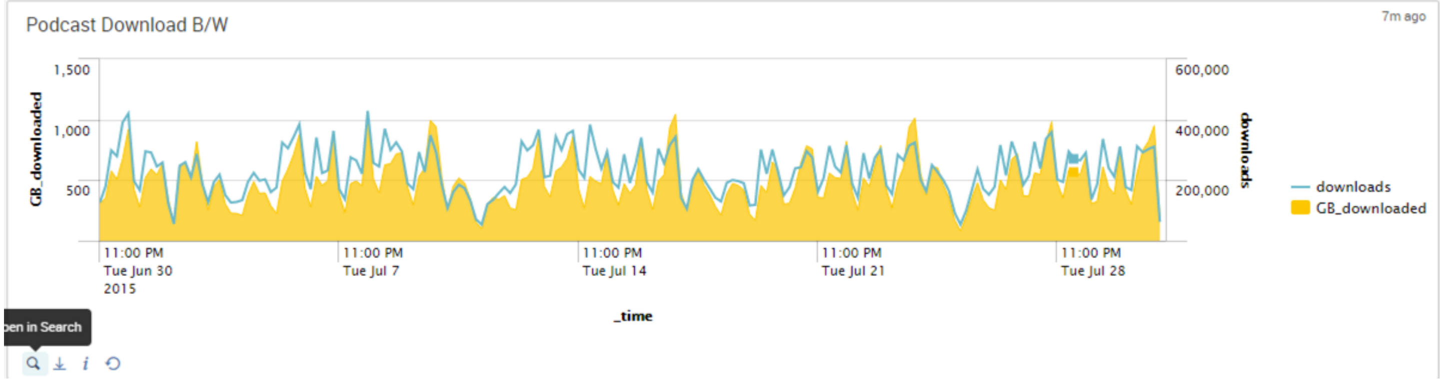
- Example #2: graph the previous over time:
  - Notice the use of “sum(count)” in the timechart command!
  - `drop_dm_object_name` macro removes the “Web” prefix from fieldnames

```
| `tstats` count from datamodel=Web where Web.is_podcast=1 AND
(Web.status=40* OR Web.status=50*) by Web.status,_time span=12h |
`drop_dm_object_name("Web")`| timechart sum(count) by status
```

Fun With tstats

How about a graph of total podcast downloads and bandwidth?

```
|`tstats` count,sum(Web.bytes) as total_size from datamodel=Web
by Web.podcast_name,_time span=4h | where total_size>10000 | eval
total_GB=total_size/(1024*1024*1024) | timechart span=4h
sum(count) as downloads, sum(total_GB) as GB_downloaded
```



More tstats

- “summariesonly” option
  - In general, set this to “true” to ensure that all dashboards run at maximum speed
  - Downside is that if you are querying a Data Model that is not 100% accelerated you might miss data (though this is unlikely in regular practice)
- “allow_old_summaries” option
  - Useful during data model development
  - Set to true to prevent having to wait until the acceleration is up-to-date
- “prestats” Option
- Using macros – example:

Useful Macros for tstats

- The following macros make using tstats simpler!
- [tstats]
- definition = tstats `summariesonly` `allow_old_summaries`
- iseval = 0

- [allow_old_summaries]
- definition = allow_old_summaries=t
- iseval = 0

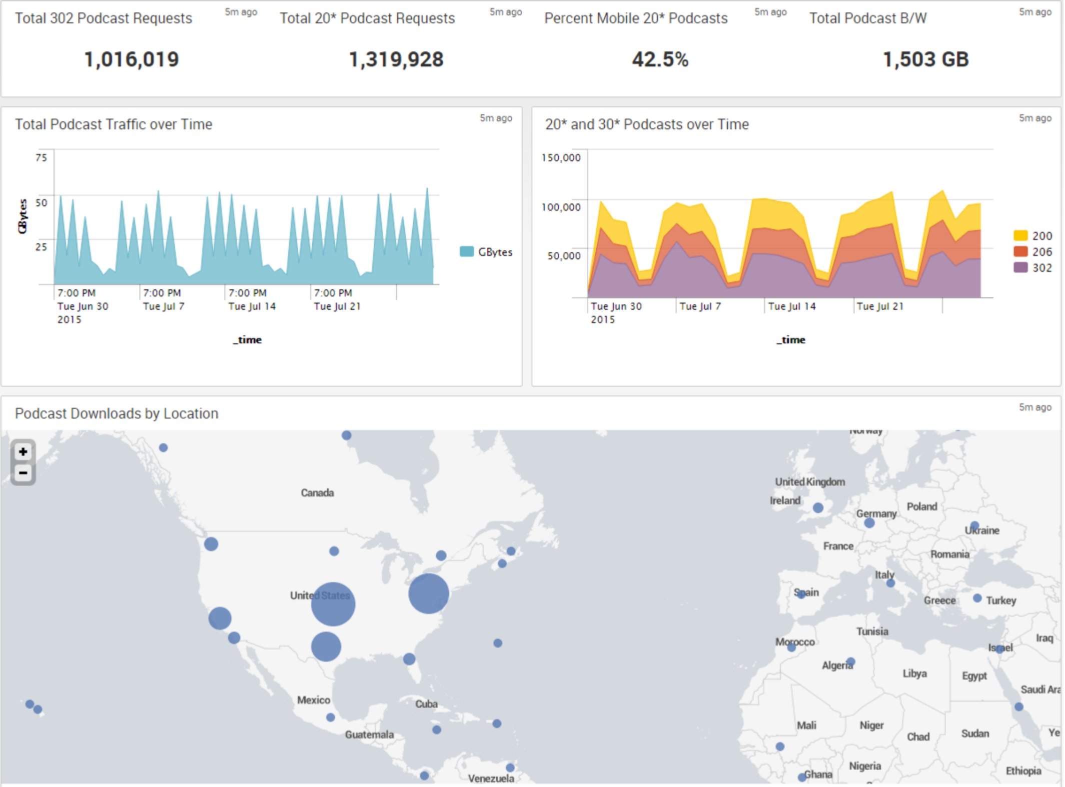
- [summariesonly]
- definition = summariesonly=t
- iseval = 0

Speed Comparison

- Over the same 7 day period - A little under 75% faster
- First Search - This search has completed and has returned 115 results by scanning 8,716,481 events in 336 seconds
- Second Search - This search has completed and has returned 115 results by scanning 22,017,061 events in 122 seconds

- Over 30 days of data, 1st takes 23.2 minutes and 2nd search takes 9.4 minutes, about 60% faster

Live Demo



Total 302 Podcast Requests 5m ago

1,016,019

Total 20* Podcast Requests 5m ago

1,319,928

Percent Mobile 20* Podcasts 5m ago

42.5%

Total Podcast B/W 5m ago

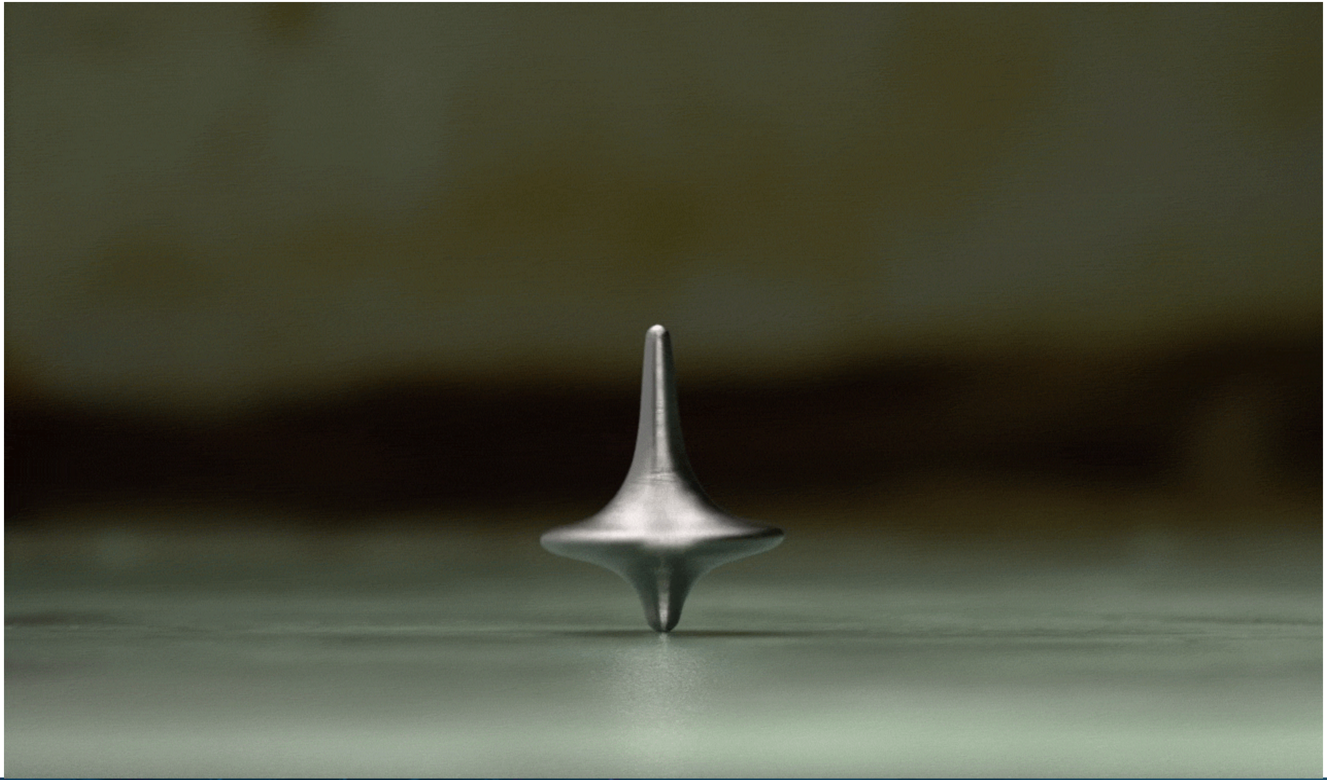
1,503 GB

Total Podcast Traffic over Time 5m ago

20* and 30* Podcasts over Time 5m ago

Podcast Downloads by Location 5m ago

Lessons Learned



.conf2015

THANK YOU

splunk>