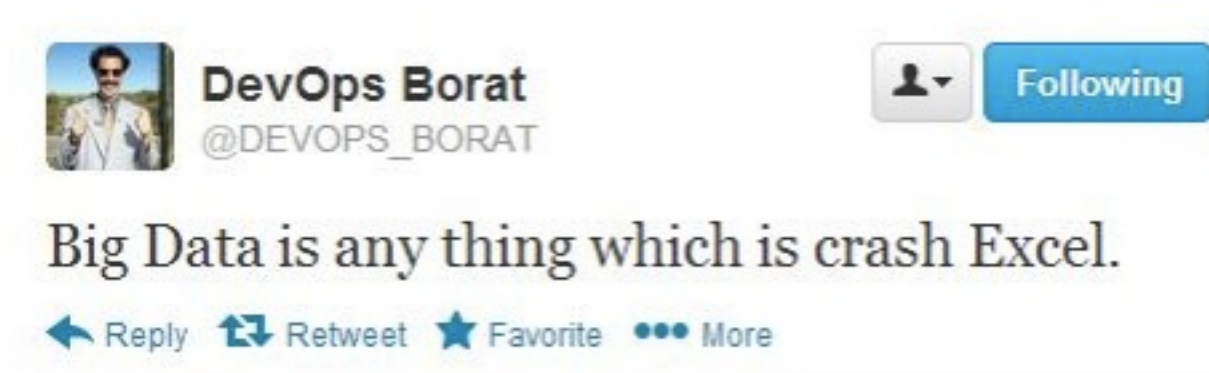


Data Zen



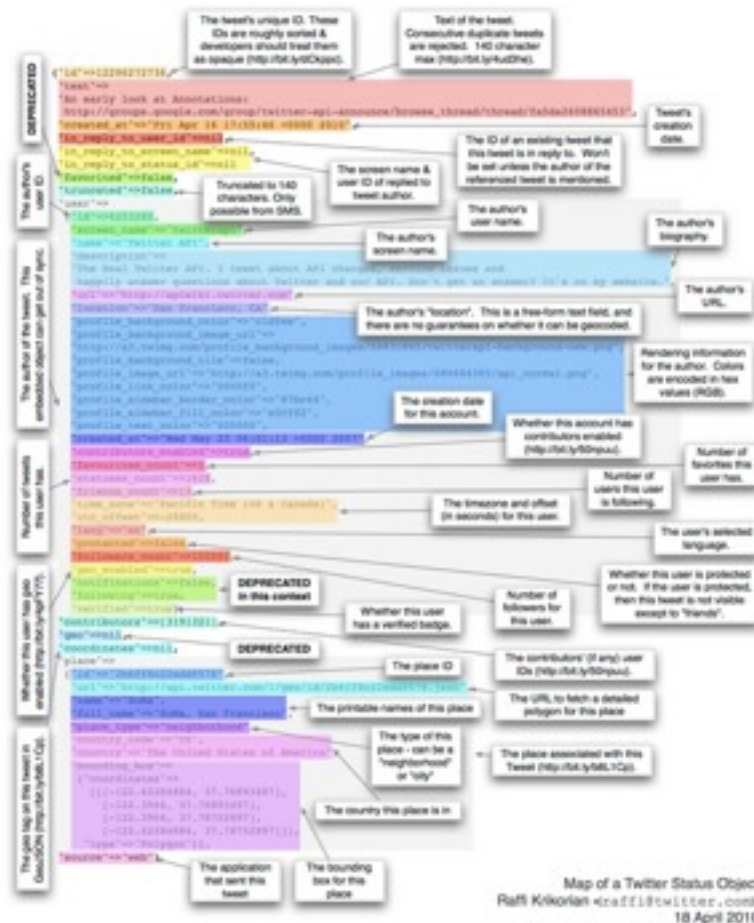
Tweet

just 140 characters?



Tweet

just 140 characters?



Log

just a message?

I'm broken. Please show this to someone who can fix can fix

timestamp hostname process
code location ip address
request id parameter values

who committed this?!!!

Code just code?

Search

Repositories

Code 522

Issues 209,289

Users

Languages

PHP	398
C	75
C++	12
Perl	4
Delphi	3
Markdown	2
Ruby	1
reStructuredText	1
HTML	1
Haxe	1

[Advanced Search](#) [Cheat Sheet](#)

We've found 522 code results Sort: **Best match** ▾

tedmasterweb/bbeditclippings – CURLOPT_SSL_VERIFYHOST
Last indexed 5 months ago

```
1 #indent#
2 CURLOPT_SSL_VERIFYHOST
```

bostjan/PHP-application-server – client_curl.php
Last indexed 5 months ago

PHP

```
7  CURLOPT_HEADER      => false,
8  CURLOPT_SSL_VERIFYHOST => @,
9  CURLOPT_SSL_VERIFYPEER => false,
10 CURLOPT_SSLCERT      => 'certs/client.pem',
```

yagmikita/ServiceApi – TestSessionCheck.php
Last indexed a month ago

PHP

```
6  CURLOPT_SSL_VERIFYPEER => false,
7  CURLOPT_SSL_VERIFYHOST => false,
8  );
9
10 $ch = curl_init();
11 curl_setopt_array($ch, $options);
12
13 var_dump(curl_exec($ch));
```

Metric

just a number?

URL: [https://download.elasticsearch.org/
elasticsearch/elasticsearch-0.90.1.zip](https://download.elasticsearch.org/elasticsearch/elasticsearch-0.90.1.zip)

timestamp geo location package
remote ip host name
format product

Ask

and you shall be answered

show me the tweets that mention obama

unstructured

in ohio

structure

in the past month

moar structure

Ask

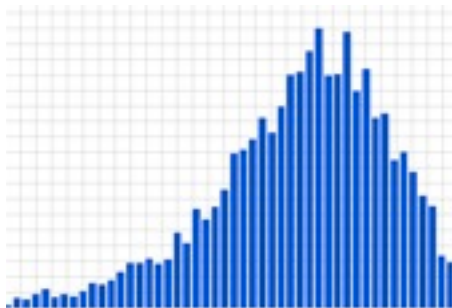
and you shall be answered

show me the tweets that mention obama
in ohio
in the past month

total: 255010294

Ask
and you shall be answered

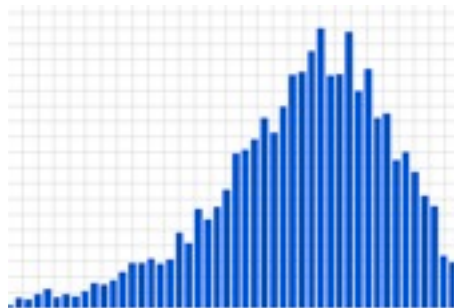
show me the tweets that mention obama
in ohio
in the past month
broken by day



analytics

Ask **Anything** and you shall be answered

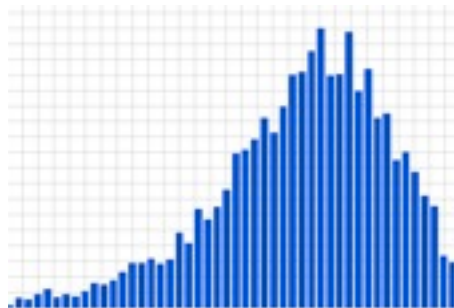
show me the tweets that mention **romney**
in ohio
in the past month
broken by day



analytics

Ask **Anything** and you shall be answered

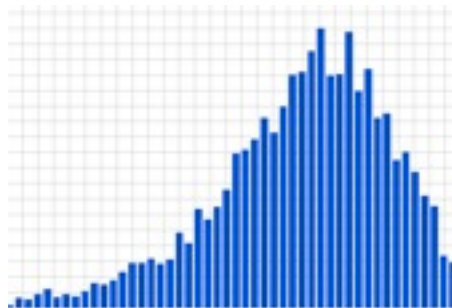
show me the tweets that mention **romney**
in **california**
in the past month
broken by day



analytics

Ask **Anything** and you shall be answered

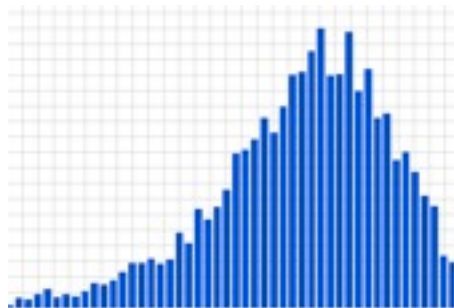
show me the tweets that mention **romney**
in **california**
in the past **year**
broken by day



analytics

Ask **Anything** and you shall be answered

show me the tweets that mention **romney**
in **california**
in the past **year**
broken by **month**



analytics

Ask **Anything** and you shall be answered

with as little (or no) data munging as possible!

Data Triangulation



Data Triangulation

text



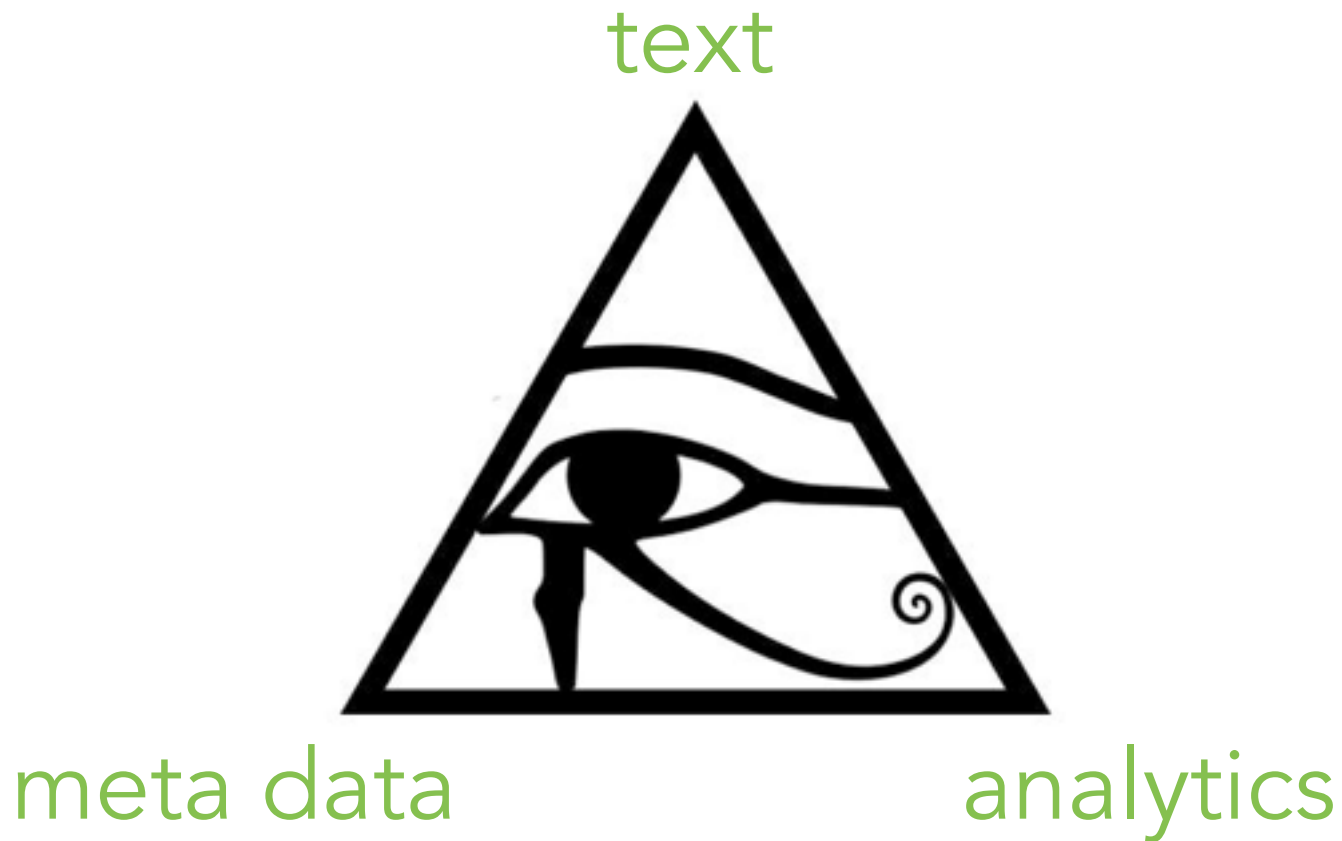
Data Triangulation

text



meta data

Data Triangulation



Data Triangulation



Data Triangulation

unstructured



meta data

analytics

[elasticsearch.](https://www.elasticsearch.com/)

Data Triangulation

unstructured



structure

analytics

[elasticsearch.](https://www.elasticsearch.com)

Data Triangulation

unstructured



structure

aggregation

elasticsearch.

Fresh!

realtime is the only time



Fresh!
what is realtime?

how quickly can we get results?

milliseconds!

Fresh!
what is realtime?

how quickly can we see new data?

milliseconds!

Fresh!

what is realtime?

how big is the data?

irrelevant

(but make sure enough HW)

Data Fight Club

collocation



Data Fight Club

collocation



the **first** rule of distributed system
collocation

the **second** rule of distributed system
collocation

Data Fight Club

collocation



in order to achieve data triangulation

a system should provide all of them

Data Triangulation

unstructured

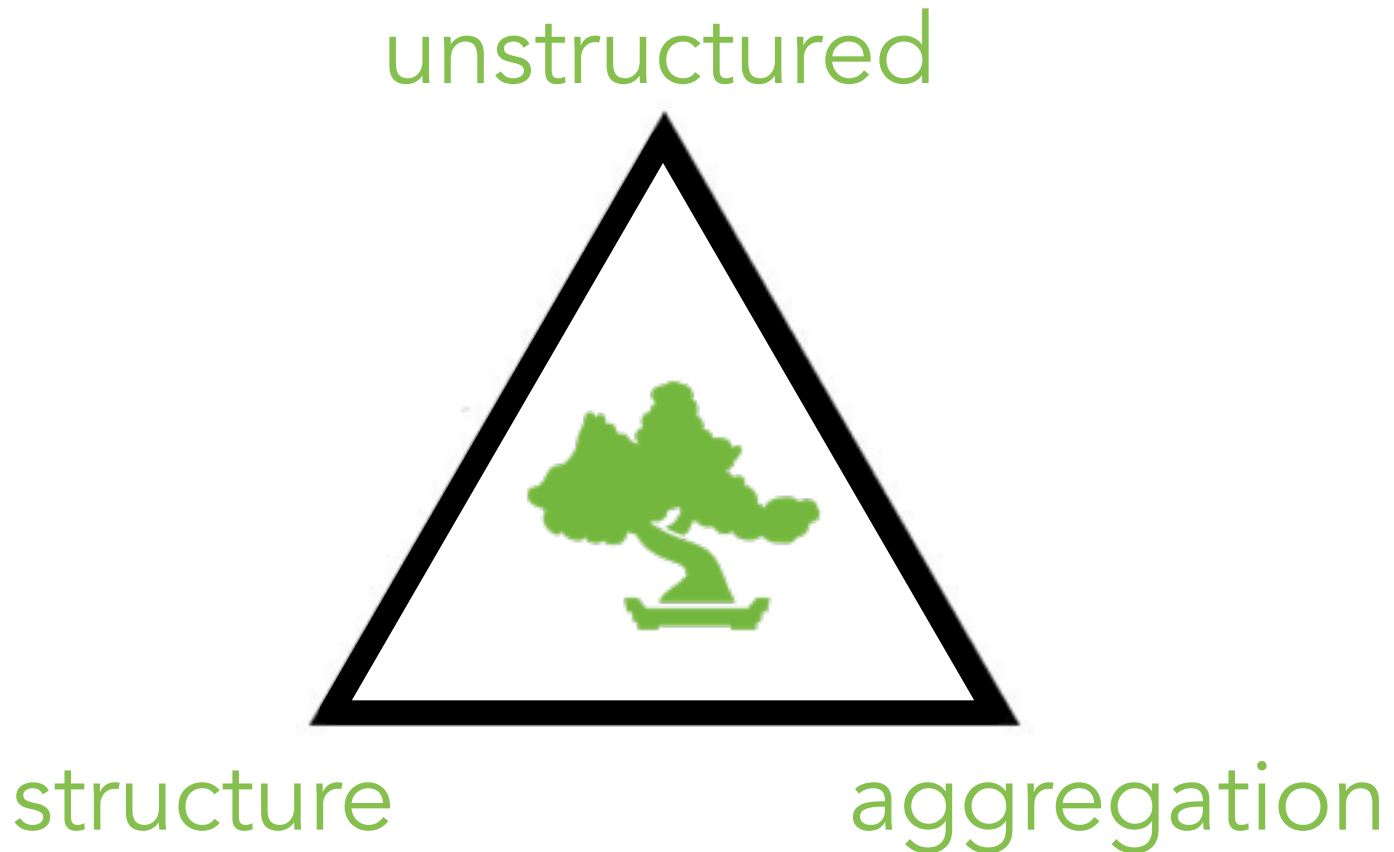


structure

aggregation

elasticsearch.

Data Triangulation



SIMPLE!

all talk, no game?



the de-facto OS log management platform
limited by ingenuity, not by licensing
diverse set of inputs, outputs & filters



logstash

detour

what's your favorite date format?

040908



the de-facto OS log management platform

diverse set of inputs, outputs & filters

storing the log data

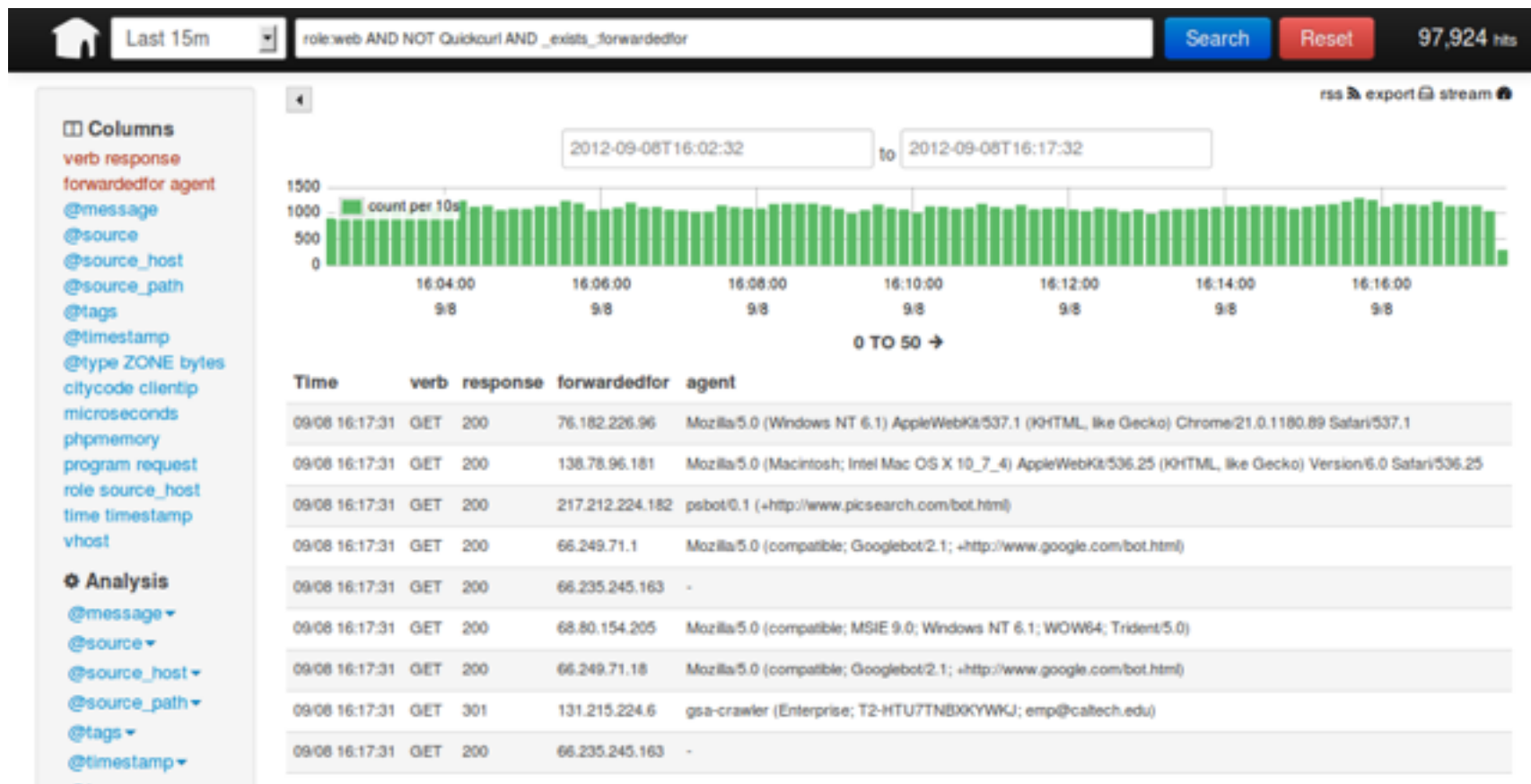
searching the data

exploring the data



logstash

kibana





need for realtime
data structure changes/updates
loose coupling with application
ha & scalability
extensibility
maintainability



need for realtime
data structure changes/updates
loose coupling with application
ha & scalability
extensibility
maintainability



default refresh time is 1sec.

data structure changes/updates

loose coupling with application

ha & scalability

extensibility

maintainability



default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

loose coupling with application

ha & scalability

extensibility

maintainability



default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

ha & scalability

extensibility

maintainability



default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

built in (shards & replicas)

extensibility

maintainability



default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

built in (shards & replicas)

plugin mechanism (custom scorer)

maintainability



default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

built in (shards & replicas)

plugin mechanism (custom scorer)

api centric



github

searches 20TB of data, including 1.3 billion files
and 130 billion lines of code.

The screenshot shows the GitHub search interface. At the top, there's a navigation bar with links to Explore, Gist, Blog, and Help. A search bar contains the letter 'a', and a 'Search' button is to its right. Below the search bar, the results are categorized into Repositories (438,261), Code (360,586,154), Issues, and Users (300,347). A 'Languages' section lists various programming languages with their respective file counts. The main results area shows 'We've found 360,586,749 code results' and lists three search results, each with a repository name, a user profile picture, and a 'Last indexed' date.

Category	Count
Repositories	438,261
Code	360,586,154
Issues	
Users	300,347

Language	Count
C	115,604,227
PHP	34,565,261
HTML	26,580,755
JavaScript	24,796,153
C++	18,964,633
Java	15,920,579
Ruby	8,955,813

We've found 360,586,749 code results Sort: **Best match**

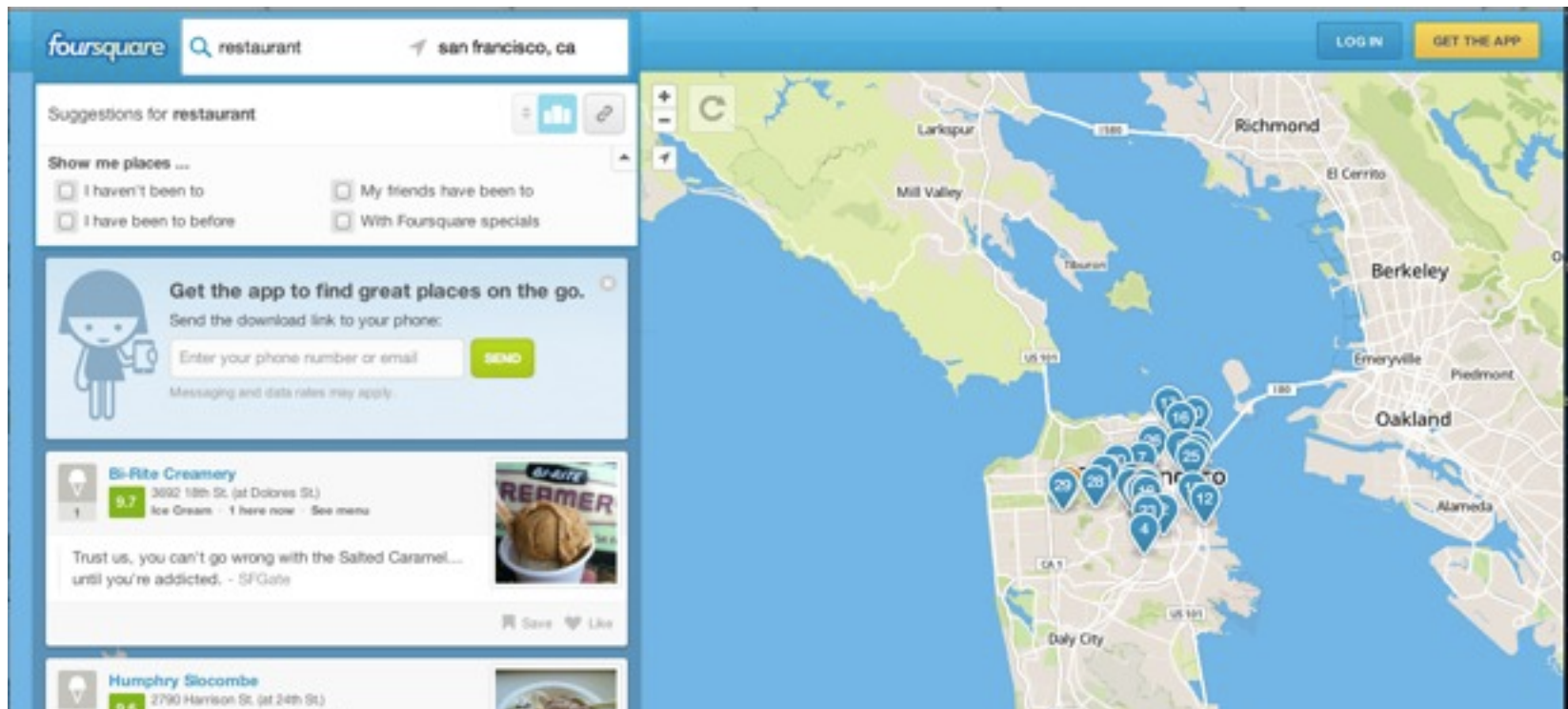
- zhenyi2697/SpyUTT – Current**
Last indexed 5 months ago
- qkbock/Particle-Assembly-Name-or-Pictures – Current**
Last indexed 5 months ago
- eclipse/tycho.p2-fork – a.txt**
Last indexed 5 months ago

elasticsearch.



foursquare

searching 50 million venues in real-time every day



“What's past is prologue.”
W.S.

where we're at?

elasticsearch

QueryDSL

Percolator

lucene 4 goodness

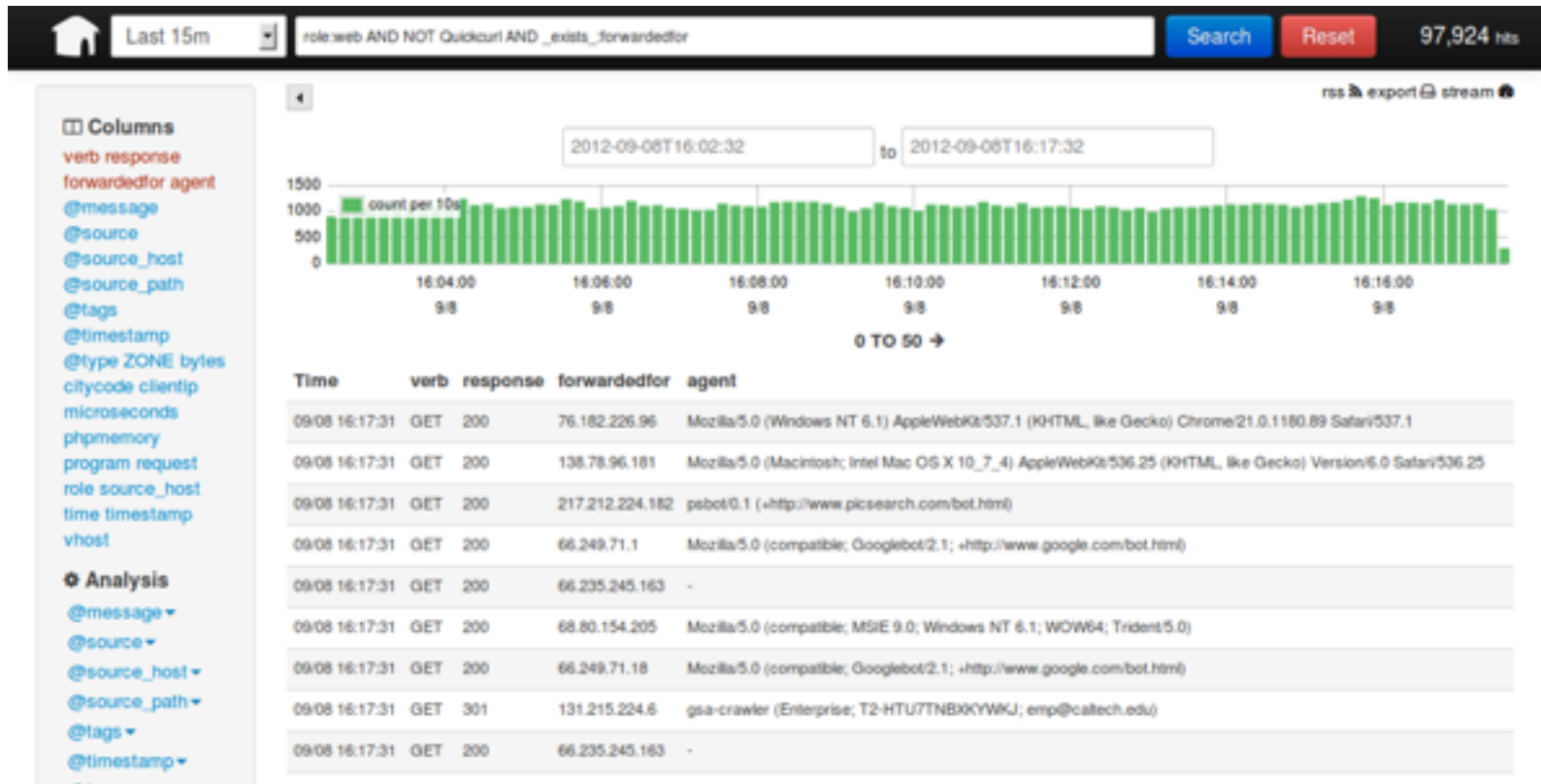
new & improved api's (suggest, parent/child, etc...)

smart shard allocator

substantial reduction of mem. footprint

where we're at?

kibana



<http://demo.kibana.org/#/dashboard/file/newtown>

elasticsearch.

where we're at?

kibana

complete rewrite

pure javascript

build, analyze, share

your data, your dashboard

on any data

at realtime

where we're at?

elasticsearch-hadoop

load data in hdfs into elasticsearch

index data directly to elasticsearch

access elasticsearch in your map/reduce jobs

still run long batch jobs, next to realtime access

did we already mention colocation?

where we're at?

elasticsearch-hadoop

pig

hive

cascading



the road ahead

towards 1.0

snapshot/restore api

aggregations

clients - ruby, python, php, perl, and more...

Final Words

elasticsearch!

