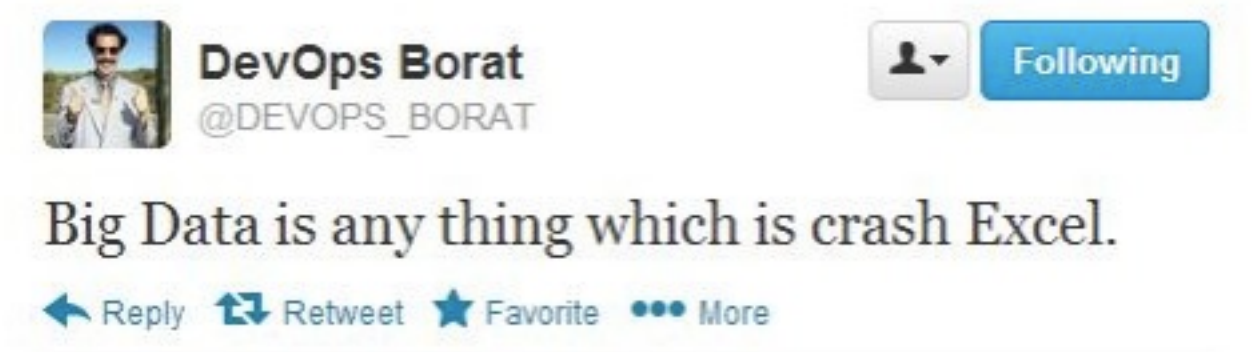
Data Zen

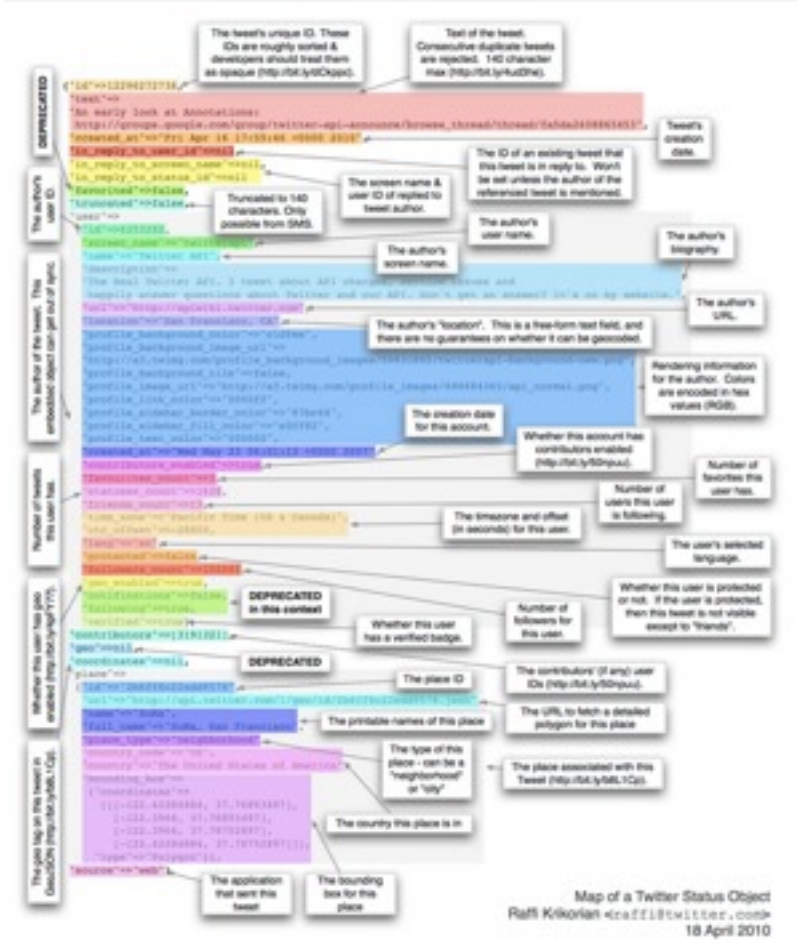
Tweet

just 140 characters?



Tweet

just 140 characters?



elasticsearch.

Log

just a message?

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

timestamp

hostname

process

code location

ip address

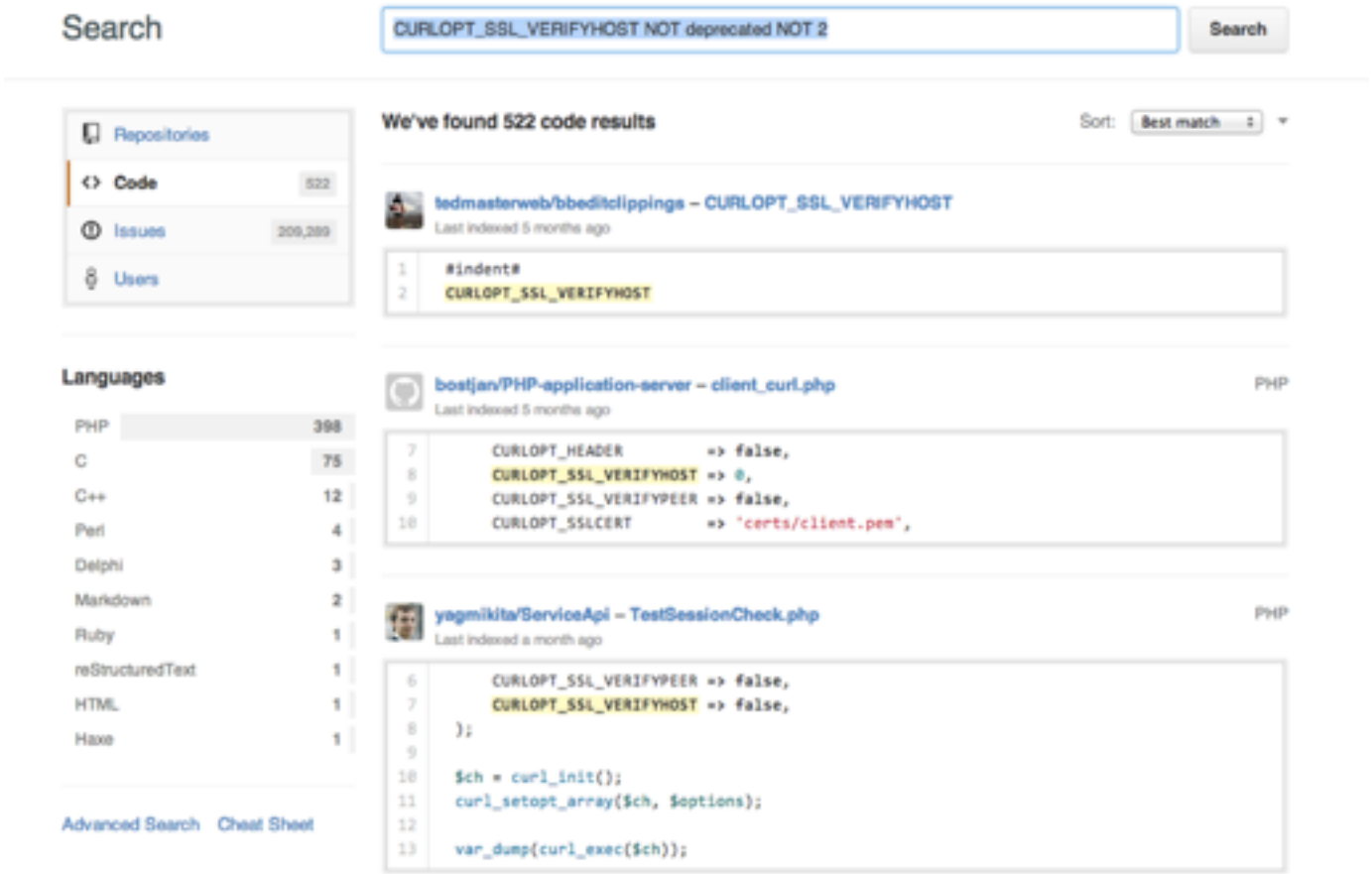
parameter values

request id

who committed this?!!!

Code

just code?



elasticsearch.

Metric

just a number?

URL: 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

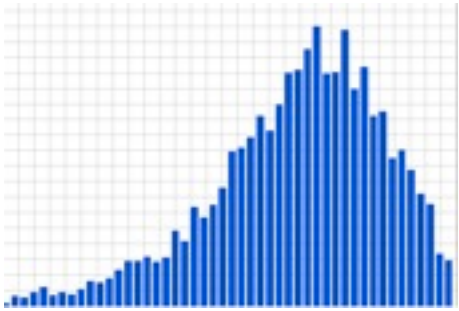
in the past month

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**

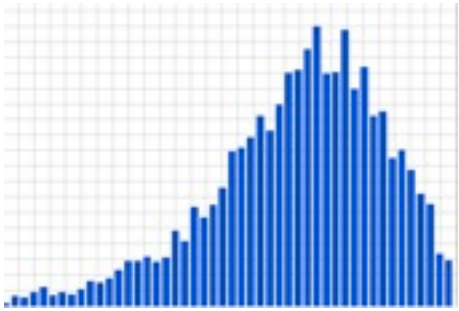
and you shall be answered

show me the tweets that mention **romney**

in ohio

in the past month

broken by day



analytics

elasticsearch.

Ask **Anything**

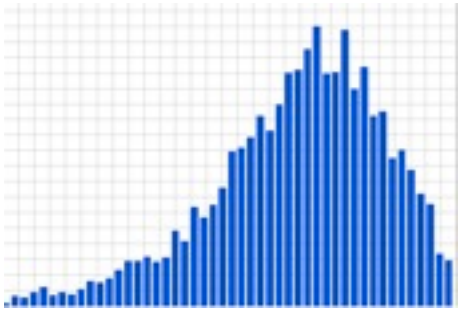
and you shall be answered

show me the tweets that mention **romney**

in **california**

in the past month

broken by day



analytics

elasticsearch.

Ask **Anything**

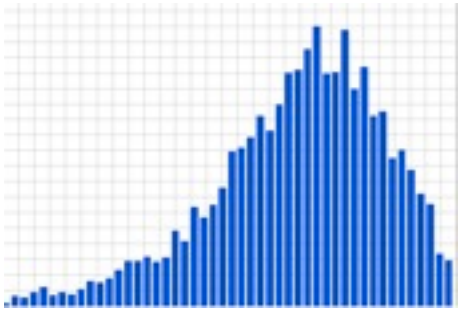
and you shall be answered

show me the tweets that mention **romney**

in **california**

in the past **year**

broken by day



analytics

elasticsearch.

Ask **Anything**

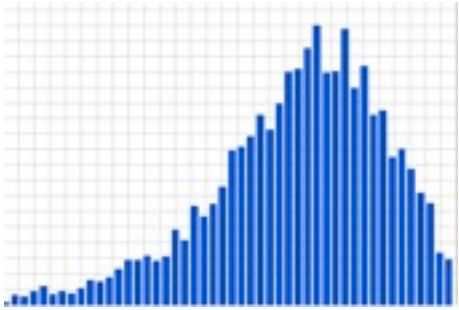
and you shall be answered

show me the tweets that mention **romney**

in **california**

in the past **year**

broken by **month**



analytics

elasticsearch.

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

text

meta data

analytics

Data Triangulation

text

meta data

analytics

Data Triangulation

unstructured

meta data

analytics

Data Triangulation

unstructured

structure

analytics

Data Triangulation

unstructured

structure aggregation

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

Data Triangulation

unstructured

structure

aggregation

SIMPLE!

all talk, no game?

logstash

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

logstash

the de-facto OS log management platform

diverse set of inputs, outputs & filters

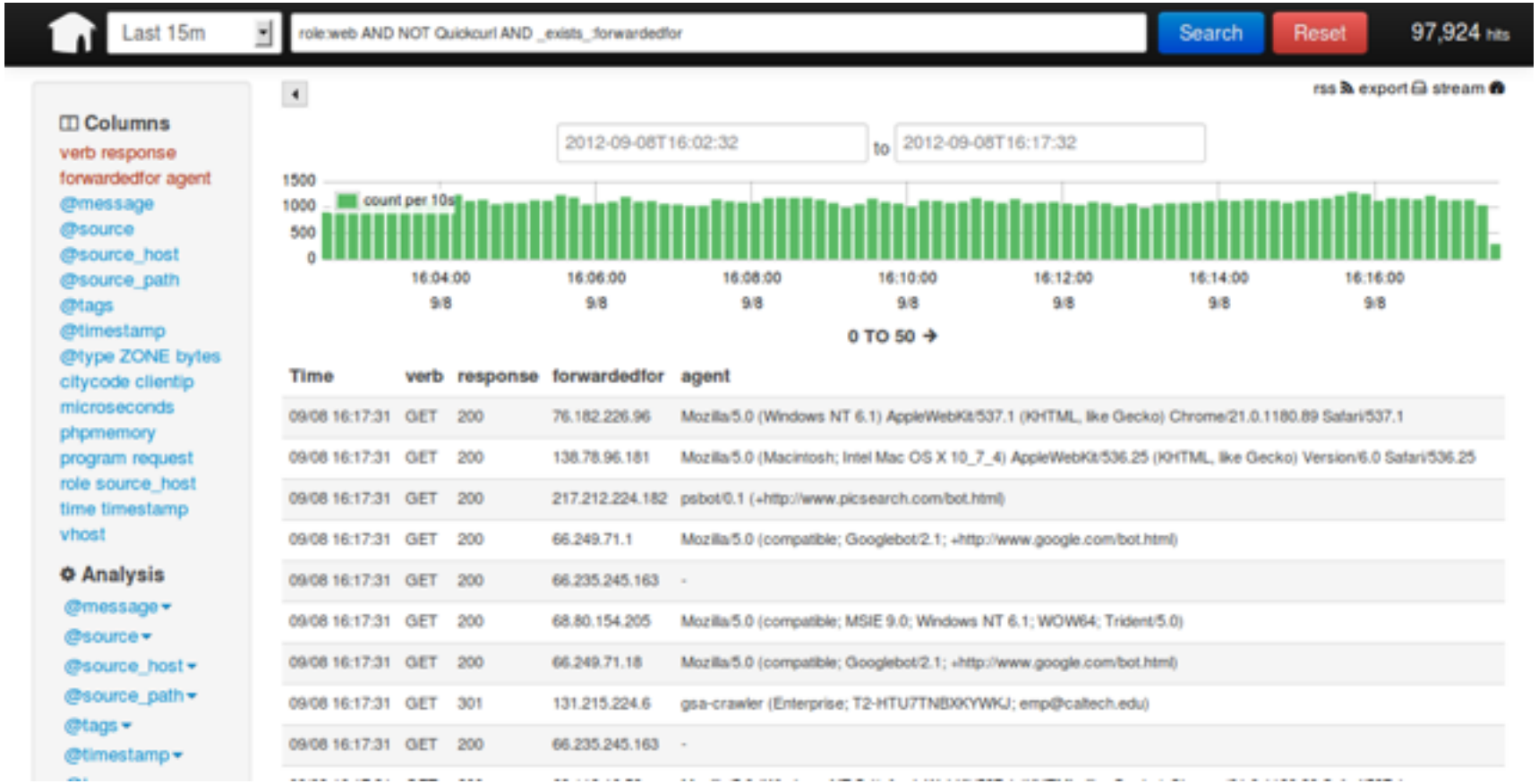
storing the log data

searching the data

exploring the data

logstash

kibana



elasticsearch.

soundcloud

need for realtime

data structure changes/updates

loose coupling with application

ha & scalability

extensibility

maintainability

soundcloud

need for realtime

data structure changes/updates

loose coupling with application

ha & scalability

extensibility

maintainability

soundcloud

default refresh time is 1sec.

data structure changes/updates

loose coupling with application

ha & scalability

extensibility

maintainability

soundcloud

default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

loose coupling with application

ha & scalability

extensibility

maintainability

soundcloud

default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

ha & scalability

extensibility

maintainability

soundcloud

default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

built in (shards & replicas)

extensibility

maintainability

soundcloud

default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

built in (shards & replicas)

plugin mechanism (custom scorer)

maintainability

soundcloud

default refresh time is 1sec.

fast reindexing (45min vs 24hrs)

index aliases

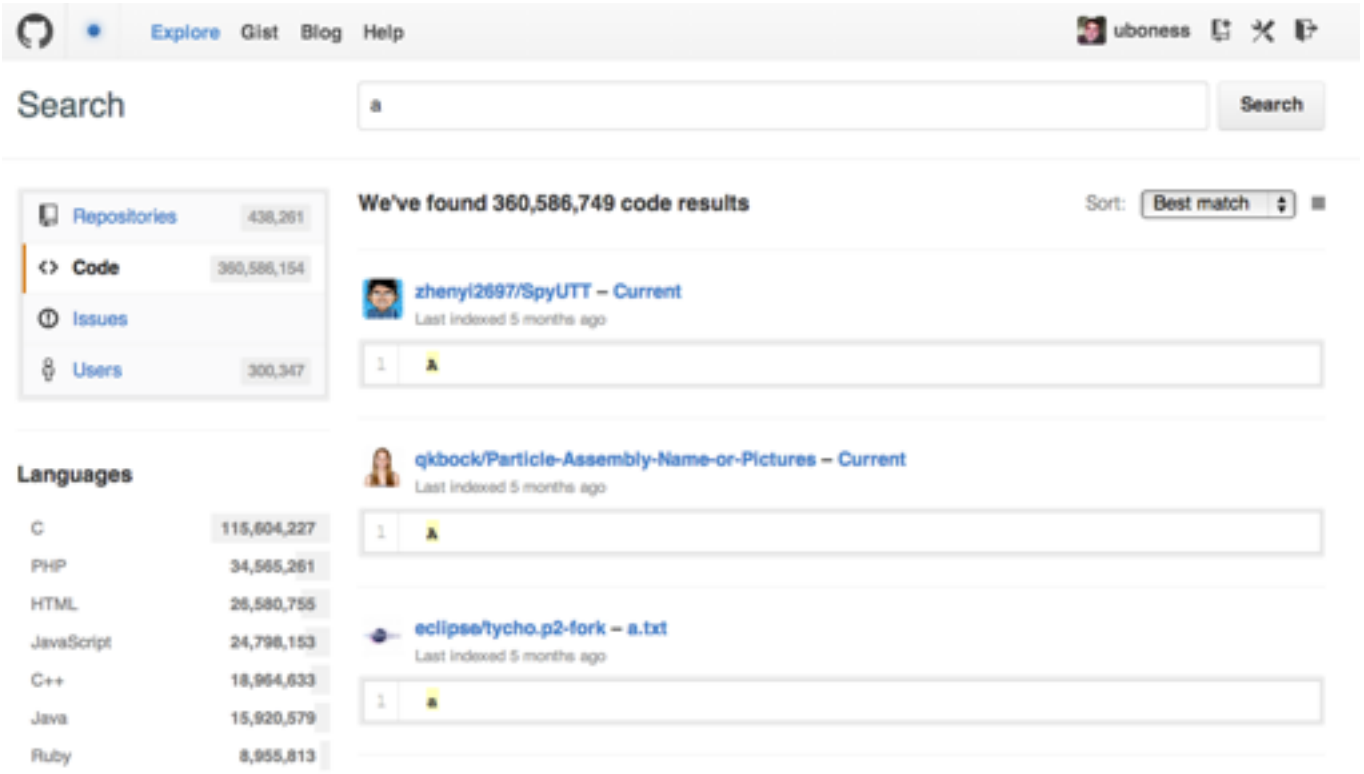
built in (shards & replicas)

plugin mechanism (custom scorer)

api centric

github

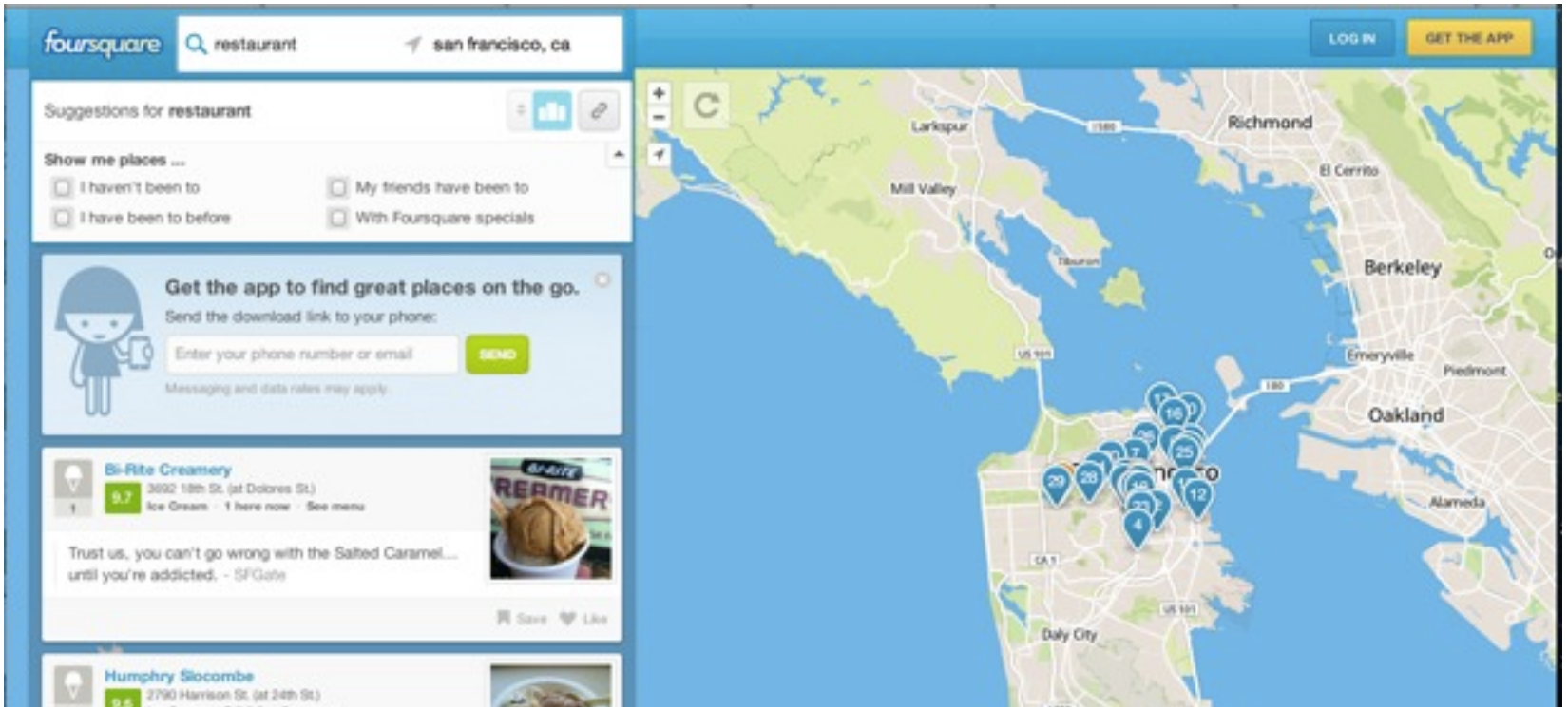
searches 20TB of data, including 1.3 billion files and 130 billion lines of code.



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

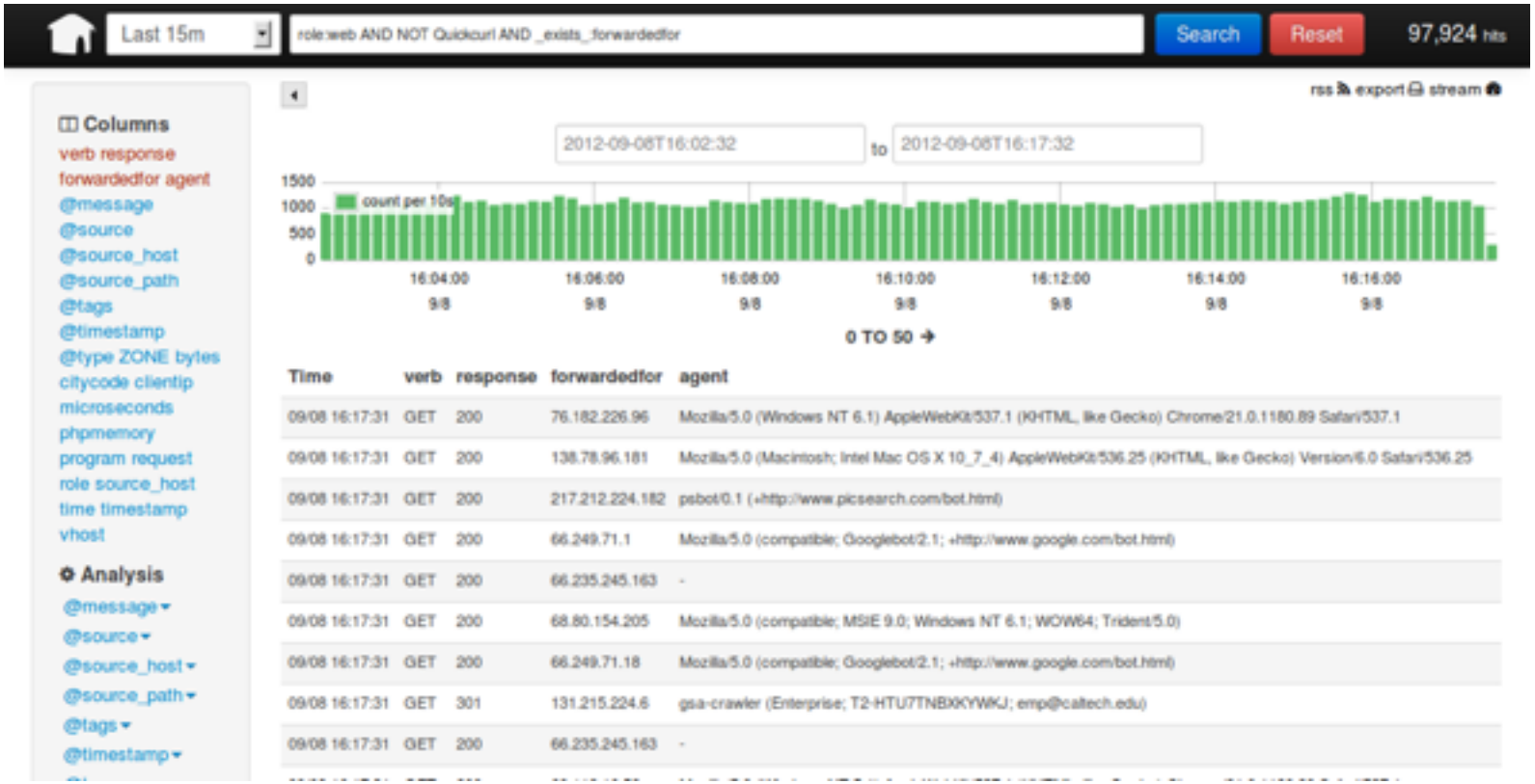
new & improved api's (suggest, parent/child, etc...)

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!