The State Of Open Source Logging

Rashid Khan (@rashidkpc)

Shay Banon (@kimchy)

elasticsearch.

Rashid Khan

Developer @ elasticsearch

Operations guy

Logging Nerd

Kibana project

IRC/Twitter: rashidkpc

elasticsearch.

Logs suck.

3am What just happened?!

Local logs for x in $OMG; do ssh $x 'grep ...

Central Syslog sed/awk/grep spaghetti

My horrid scripts As if anyone used them

Logs told me bad things. **I hated logs.**

Log formats aren't any good either

38.4.41.91 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"

040908 10:00:00 MySql: Normal shutdown
MySql: ready for connections.
Version: '4.0.21-nt-log' socket: '' port: 3306 Source distribution

"RasBox","RAS",10/22/2006,09:13:09,1,"ACME\slimjim","ACME\slimjim",,,,,,"192.168.132.45",12,,"192.168.132.45",,,
0,"CONNECT 24000",1,2,4,,0,"311 1 192.168.132.45 07/31/2006 21:35:14 749",,,,,,,,,,,,,,,,,,,,,,,,"MSRASV5.00",
311,,,,

Feb 12 19:11:27 enigma su: dcid to root on /dev/ttyp0
SU 07/23 00:57 + ??? root-root
Jul 5 22:13:15 lili su[2614]: - pts/6 dcid-root

[2007-09-01 19:08:49.862 ADT] : LOG: connection received:
host=192.168.2.99 port=37142
[2007-09-01 19:08:49.869 ADT] 192.168.2.99: FATAL: password
authentication failed for user "ossec_user"

elasticsearch.

Commercial Log Analysis

Punishes you for using it.

Value vs Cost

Assumes every event has the same value, assigns it the same cost

Small data sets with high value events.

Java stack trace vs a single mail log line

Open Source

Limited by ingenuity, not by licensing

Component based

No vendor lock in

Extensible

What we need

1. A definition of logs
2. Something to collect logs
3. A way to mold them to our needs
4. Somewhere to put them
5. An interface for interacting with them
6. **Bonus** A way to add value to existing data

fluentd

Lightweight

Ruby

6 supported inputs (+ community)

12 supported outputs (+ community)

Speaks JSON

Input / Buffer / Output

Logstash

33 input sources

28 filters

46 output destinations

Ruby wrapped in JRuby

Actively developed

Engaged user community

Other Options

Graylog2

ELSA

Flume

Scribe

Logstash

An event pipe

As simple as you want: Single box

As complex as you can dream: Distributed and routed

Serious about Simplicity

Logstash Inputs

amqp, drupal_dblog, eventlog, exec, file, ganglia, gelf, gemfire, generator, heroku, irc, log4j, lumberjack, pipe, redis, relp, sqs, stdin, stomp, tcp, twitter udp, xmpp, zenoss, zeromq

Logstash Filters

alter, anonymize, checksum, csv, **date**, **dns**, environment, gelfify, **geoip**, **grep**, **grok**, grokdiscovery, json, kv, **metrics**, **multiline**, **mutate**, noop, split, syslog_pri, urldecode, xml, zeromq

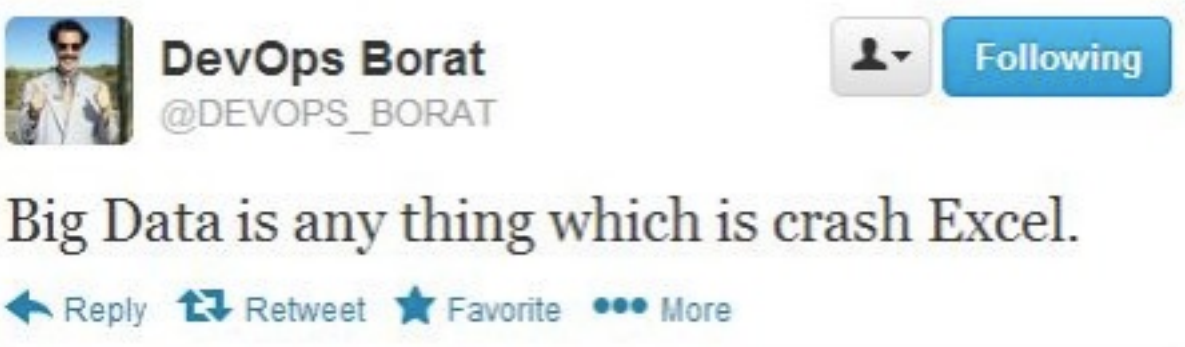
elasticsearch.

Logstash Outputs

amqp, boundary, circonus, cloudwatch, datadog, elasticsearch, elasticsearch_http, elasticsearch_river, email, exec, file, ganglia, gelf, gemfire, graphite, graphtastic, http, internal, irc, juggernaut, librato, loggly, lumberjack, metriccatcher, mongodb, nagios, nagios_nsca, null, opentsdb, pagerduty, pipe, redis, riak, riemann, sns, sqs, statsd, stdout, stomp, syslog, tcp, websocket, xmpp, zabbix, zeromq

Tweet

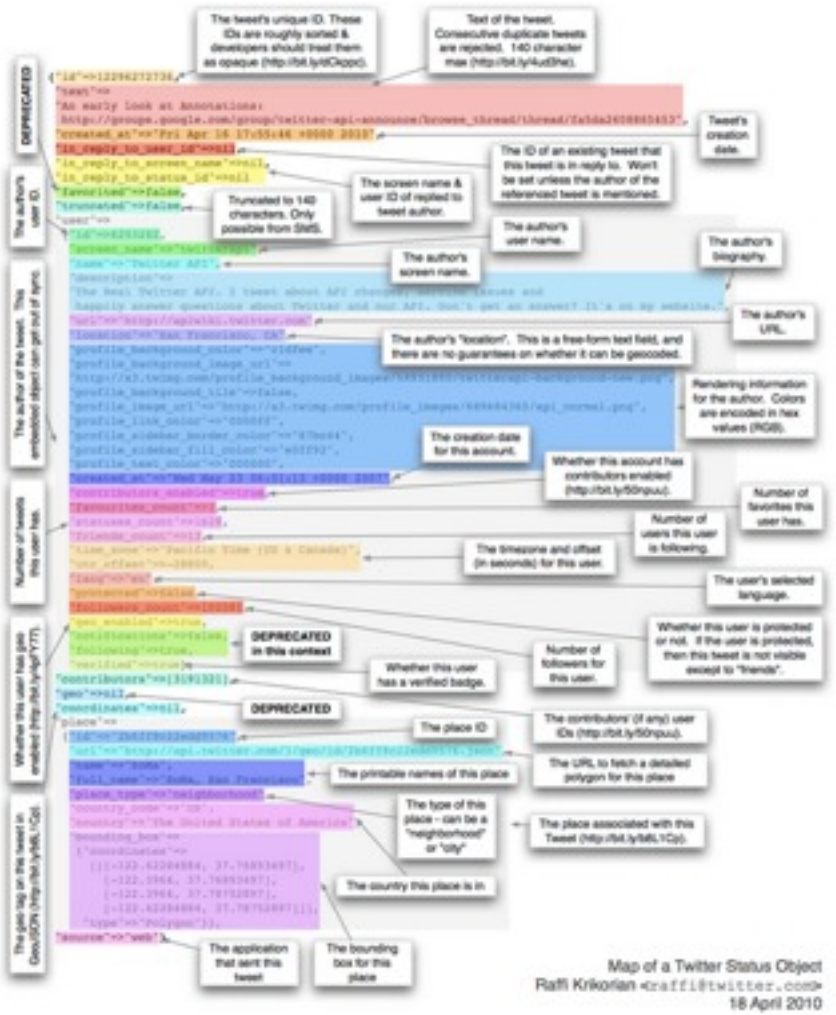
just 140 characters?



elasticsearch.

Tweet

just 140 characters?



elasticsearch.

Log
just a message?

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

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?!!!

Metric

just a number?

URL: https://download.elasticsearch.org/elasticsearch/elasticsearch-0.90.1.zip

Metric

just a number?

URL: https://download.elasticsearch.org/elasticsearch/elasticsearch-0.90.1.zip

timestamp

geo location

package

remote ip

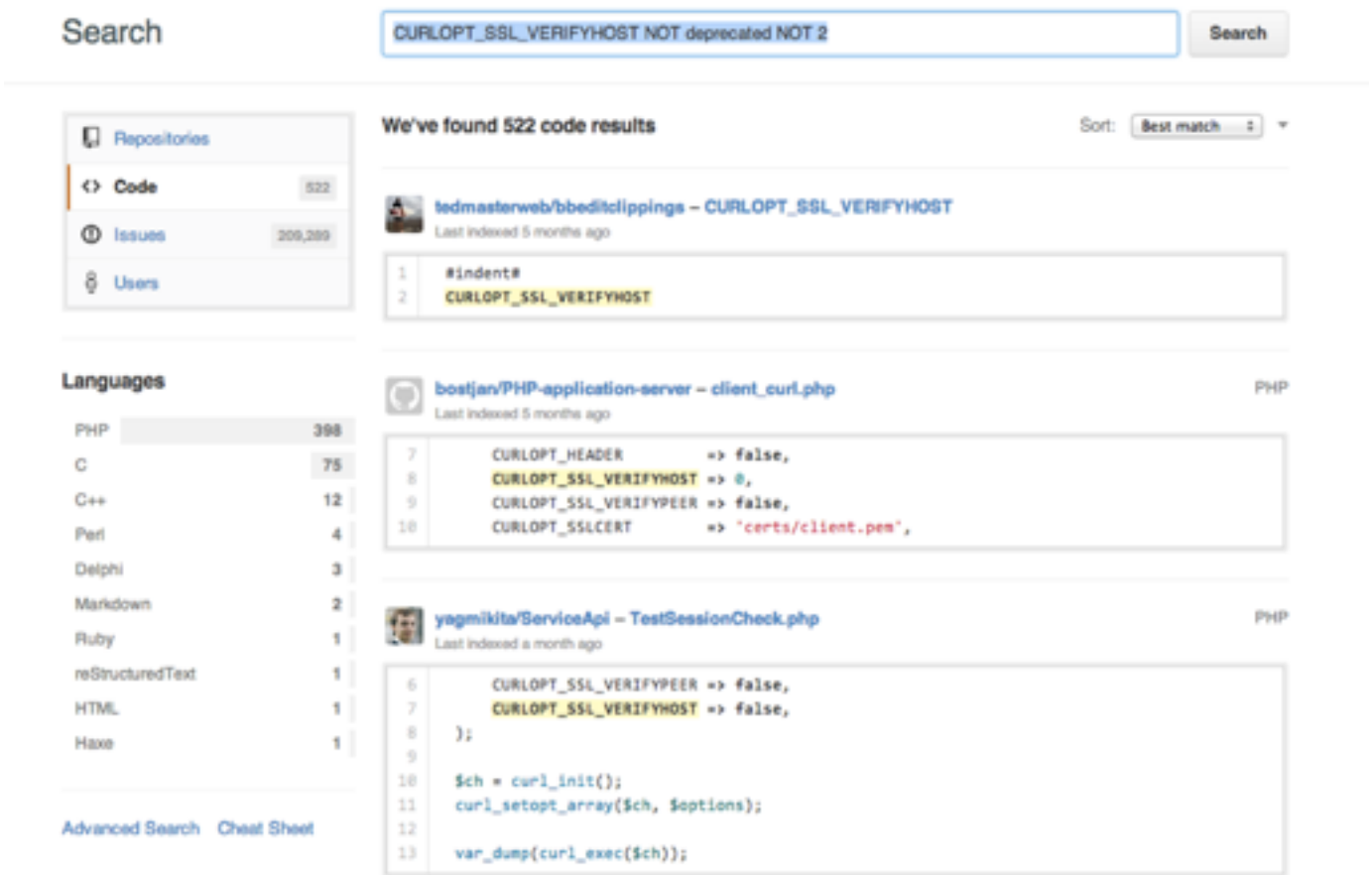
host name

format

product

Code

just code?



elasticsearch.

Ask

and you shall be answered

show me the tweets that mention obama

Ask

and you shall be answered

show me the tweets that mention obama

unstructured

Ask

and you shall be answered

show me the tweets that mention obama

unstructured

in ohio

Ask

and you shall be answered

show me the tweets that mention obama

unstructured

in ohio

structure

elasticsearch.

Ask

and you shall be answered

show me the tweets that mention obama

unstructured

in ohio

structure

in the past month

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

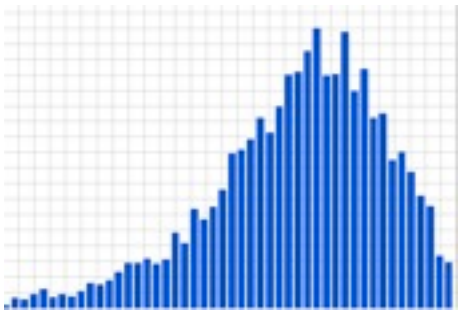
and you shall be answered

show me the tweets that mention obama

in ohio

in the past month

broken by day



analytics

elasticsearch.

Ask **Anything**

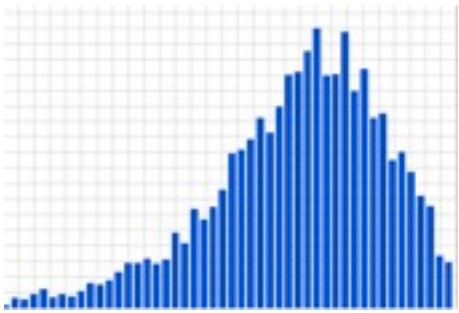
and you shall be answered

show me the tweets that mention **romney**

in ohio

in the past month

broken by day



analytics

elasticsearch.

Ask **Anything**

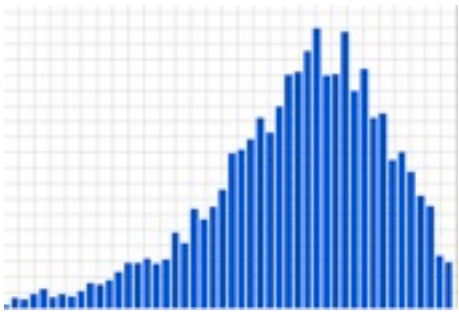
and you shall be answered

show me the tweets that mention **romney**

in **california**

in the past month

broken by day



analytics

elasticsearch.

Ask **Anything**

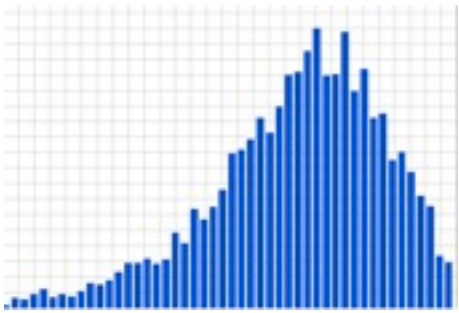
and you shall be answered

show me the tweets that mention **romney**

in **california**

in the past **year**

broken by day



analytics

elasticsearch.

Ask **Anything**

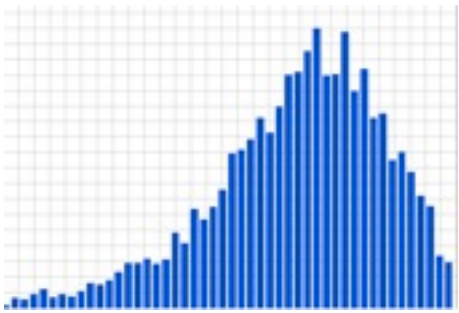
and you shall be answered

show me the tweets that mention **romney**

in **california**

in the past **year**

broken by **month**



analytics

elasticsearch.

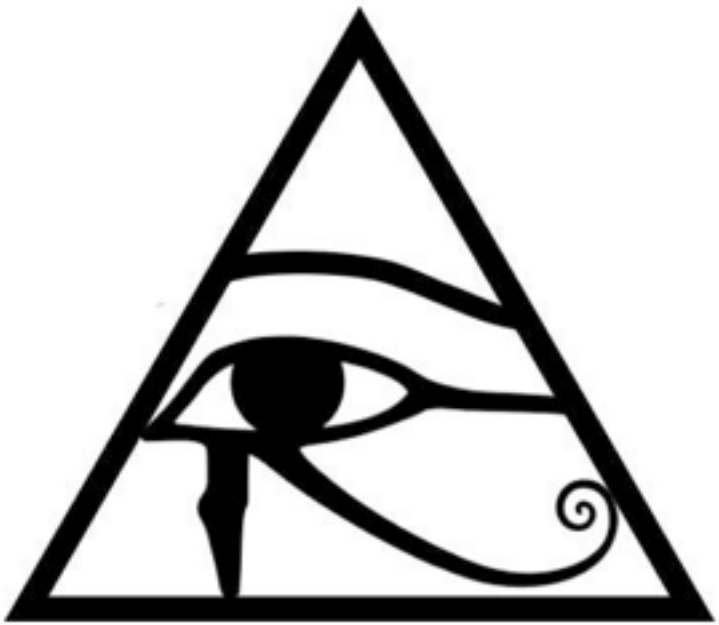
Ask **Anything**

and you shall be answered

with as little data munging as possible!

Data Triangulation

text

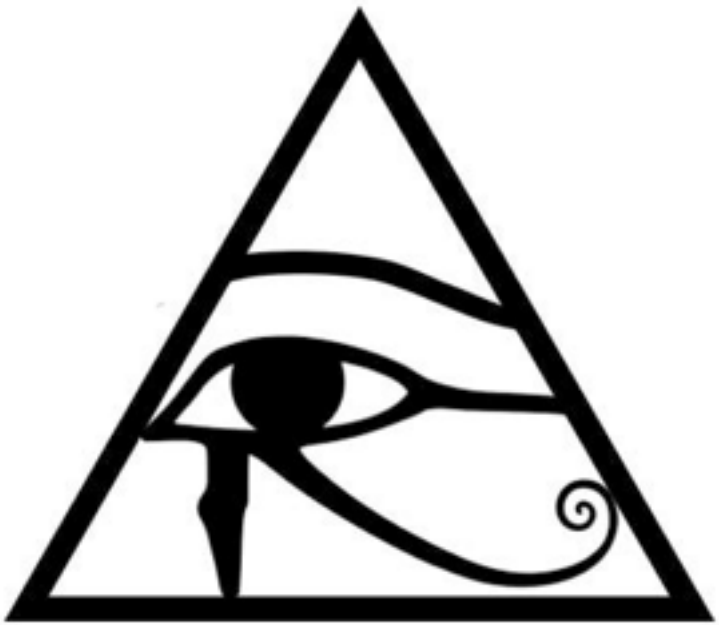


meta data

analytics

Data Triangulation

unstructured



structure aggregation

Fresh!

realtime is the only time

Fresh!

what is realtime?

how quickly can we get results?

Fresh!

what is realtime?

how quickly can we get results?

took: 70ms

Fresh!

what is realtime?

how quickly can we see new data?

Fresh!

what is realtime?

how quickly can we see new data?

milliseconds!

Fresh!

what is realtime?

how big is the data?

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 systems

collocation

Data Fight Club

collocation

the **second** rule of distributed systems

collocation

Data Fight Club

collocation

in order to achieve data triangulation

a system should provide all of them

Final Words

elasticsearch!