Corralling logs with ELK

Open Source Log Analytics

Mark Walkom
@warkolm
mark.walkom@elasticsearch.com

elasticsearch.

xero

seek

XING

stackoverflow

Mc Graw Hill

SOUNDCLOUD

Fog Creek SOFTWARE

foursquare

loggly

NETWORKED INSIGHTS

wotif.com

What is a log?

- Time-based data

  String containing numbers and text

- This data is everywhere!

  Server logs
  Twitter stream
  Financial transactions
  Metric / monitoring data

- Log all things!!!!

- Format "Standards" is Format Frustration

Why collect & centralise logs?

- Access log files without system access
- Shell scripting: Too limited or slow
- Using unique ids for errors, aggregate it across your stack
- Reporting (everyone can create his/her own report)
- Bonus points: Unify your data to make it easily searchable

Elasticsearch in 10 seconds

- Schema-free, REST & JSON based document store
- Distributed and horizontally scalable
- Open Source: Apache License 2.0
- Zero configuration
- Written in Java, extensible
- APIs for everything

Basic terms

- Index

  Logical collection of data; might be time based
  Analogous to a database

- Shard(s)

  Split logical data (index) over several machines
  Write scalability
  Control data flows

- Replica(s)

  Read scalability
  Removing SPOF

elasticsearch.

Cluster management

- Single master at any point in time
  Responsible for cluster state (node entry, index creation)
- Multicast or unicast based discovery
- Configuration is required here
  Multicast - Tell each node the name of the cluster to join
  Unicast - use IP(s) of existing nodes to join
- Tip: Keep master-eligible node count uneven, helps to prevent split brain

Sizing a cluster or node

- Data and operation dependent

  How big are your documents? How many fields in them?
  What is your query rate?
  Do you do facets/aggregations, sorting, custom scoring?
  What is your write rate?
  Do you delete documents? Update them?
  Is the data time-based?

- Test on one node, one shard, no replicas

  Look at shard size, JVM heap usage and GC frequency, number of shards/node, docs per shard, CPU and disk utilisation

- Tip: No more than 31 GB heap

Ecosystem

- Plugins

  Many third party plugins available
  Languages, monitoring, attachments, transport, scripting
  Build your own!

- Clients for many languages

  Ruby, python, php, perl, javascript
  Scala, clojure, go, .NET coming soon

- Hadoop integration

Elasticsearch
Installation & first steps

2 minutes to live

```
$ wget https://download.elasticsearch.org/...
$ tar -xf elasticsearch-1.4.2.tar.gz
$ ./elasticsearch-1.4.2/bin/elasticsearch
...
[2014-01-19 14:53:11,508][INFO ][node] [Scanner] started
...
```

Also puppet/chef modules and RPM/DEB repos

Is it alive?

```
» curl localhost:9200
{
  "status" : 200,
  "name" : "Scanner",
  "version" : {
    "number" : “1.4.2",
    "build_hash" : "a70f3ccb52200f8f2c87e9c370c6597448eb3e45",
    "build_timestamp" : "2015-01-10T09:07:17Z",
    "build_snapshot" : false,
    "lucene_version" : “4.10.2"
  },
  "tagline" : "You Know, for Search"
}
```

Tools for Everyone!

REST-based management

- Elasticsearch is full of monitoring APIs
  Everything is returned as JSON
- Humans are not the world's best JSON parsers
- TIP: use ?pretty on end of curl requests
- But what if elasticsearch had an easy to use interface from the shell?

Which node is the master?

```
$ curl "localhost:9200/_cluster/state?pretty&filter_metadata=true&
filter_routing_table=true"
{
  "cluster_name" : "elasticsearch",
  "master_node" : "GNf0hEXlTfaBvQXKBF300A",
  "blocks" : { },
  "nodes" : {
    "ObdRqLHGQ6CMI5rOEstA5A" : {
      "name" : "Triton",
      "transport_address" : “inet[/10.0.1.11:9300]”,
      "attributes" : { }
    },
    "4C7pKbfhTvu0slcSy_G4_w" : {
      "name" : "Kid Colt",
      "transport_address" : "inet[/10.0.1.12:9300]",
      "attributes" : { }
    },
    "GNf0hEXlTfaBvQXKBF300A" : {
      "name" : "Lang, Steven",
      "transport_address" : "inet[/10.0.1.13:9300]",
      "attributes" : { }
    }
  }
}
```

elasticsearch.

Now who is the master?

```
$ curl localhost:9200/_cat/master
GNf0hEXlTfaBvQXKBF300A 10.0.1.13 Lang, Steven
```

_cat/* api

- /_cat/aliases
- /_cat/allocation
- /_cat/count
- /_cat/fielddata
- /_cat/health
- /_cat/indices
- /_cat/master
- /_cat/nodes
- /_cat/pending_tasks
- /_cat/plugins
- /_cat/recovery
- /_cat/shards
- /_cat/thread_pool

Elasticsearch Scaling

- Provision a new node
- Point it to existing node/cluster
- Shards will auto balance
- Query/insert via any node
- Survive node loss with replicas
- TIP: use noop scheduler on linux to maximise I/O

elasticsearch.

Logstash in 10 seconds

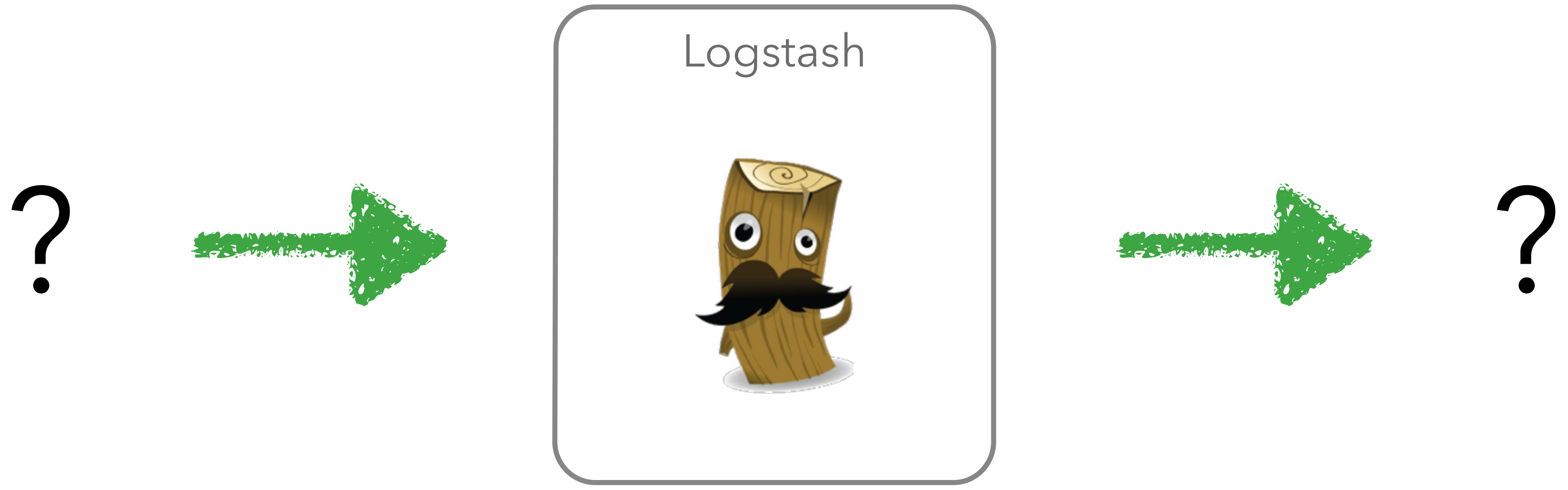
- Managing events and logs
- Collect, parse, enrich and store data
- Modular: many, many inputs and outputs
- Apache License 2.0
- Ruby app (JRuby)
- Part of Elasticsearch family

Logstash architecture

Input
collect and split

Filter
alter and enrich

Output
store and visualise



Inputs

- Monitoring: collectd, graphite, ganglia, snmptrap, zenoss
- Datastores: elasticsearch, redis, sqlite, s3
- Queues: rabbitmq, zeromq
- Logging: eventlog, lumberjack, gelf, log4j, relp, syslog, varnish log
- Platforms: drupal_dblog, gemfire, heroku, sqs, s3, twitter
- Local: exec, generator, file, stdin, pipe, unix
- Protocol: imap, irc, stomp, tcp, udp, websocket, wmi, xmpp

Filters

- alter, anonymize, checksum, csv, drop, multiline
- dns, date, extractnumbers, geoip, i18n, kv, noop, ruby, range
- json, urldecode, useragent
- metrics, sleep
- grok
- … many, many more …

Outputs

- Store: elasticsearch, gemfire, mongodb, redis, riak, rabbitmq
- Monitoring: ganglia, graphite, graphtastic, nagios, opentsdb, statsd, zabbix
- Notification: email, hipchat, irc, pagerduty, sns
- Protocol: gelf, http, lumberjack, metriccatcher, stomp, tcp, udp, websocket, xmpp
- External Monitoring: boundary, circonus, cloudwatch, datadog, librato
- External service: google big query, google cloud storage, jira, loggly, riemann, s3, sqs, syslog, zeromq
- Local: csv, exec, file, pipe, stdout, null

2 more minutes to live

```
$ wget https://download.elasticsearch.org/...
$ tar -xf logstash-1.4.2.tar.gz
$ ./logstash-1.4.2/bin/logstash -f sample.conf
```

Also puppet/chef modules and RPM/DEB repos

Simple example

- Download, create config and run

```
input {
    stdin {}
}

output {
    stdout { debug => true }
}
```

```
echo foo | logstash-1.4.2/bin/logstash -f sample.conf
{
       "message" => "foo",
      "@version" => "1",
    "@timestamp" => "2015-01-10T13:30:59.648Z",
          "host" => “kryptic.elasticsearch.org”
}
```

Simple filter with grok

```
input {
    stdin {}
}

filter {
  grok {
    match => [ "message", "%{WORD:firstname} %{WORD:lastname} %{NUMBER:age}" 
]
  }
}

output {
    stdout { debug => true }
}
```

sample.conf

Simple filter with grok

```
echo “Nick Fury 100" | logstash-1.4.2/bin/logstash -f
sample.conf
{
       "message" => “Nick Fury 100",
      "@version" => "1",
    "@timestamp" => "2014-01-10T16:56:02.502Z",
          "host" => "kryptic",
     "firstname" => "Nick",
      "lastname" => "Fury",
           "age" => "100"
}
```

Syslog example with grok

```
Jan 10 04:04:01 lvps109-104-93-171 postfix/smtpd[11105]:
connect from mail-we0-f196.google.com[74.125.82.196]
```

```
input { stdin {} }

filter {
  grok {
    match => { "message" => "%
{SYSLOGTIMESTAMP:syslog_timestamp} %
{SYSLOGHOST:syslog_hostname} %{DATA:syslog_program}(?:\[%
{POSINT:syslog_pid}\])?: %{GREEDYDATA:syslog_message}" }
  }
  date {
    match => [ "syslog_timestamp",
               "MMM  d HH:mm:ss", "MMM dd HH:mm:ss" ]
  }
}

output { stdout { debug => true } }
```

Syslog example with grok

```
cat sample-syslog.txt| logstash-1.4.2/bin/logstash -f
sample-syslog.conf
{
             "message" => "Jan 10 04:04:01
lvps109-104-93-171 postfix/smtpd[11105]: connect from
mail-we0-f196.google.com[74.125.82.196]",
            "@version" => "1",
          "@timestamp" => "2015-01-10T04:04:01.000+02:00",
                "host" => “kryptic.elasticsearch.org",
    "syslog_timestamp" => "Jun 10 04:04:01",
     "syslog_hostname" => "lvps109-104-93-171",
      "syslog_program" => "postfix/smtpd",
          "syslog_pid" => "11105",
      "syslog_message" => "connect from mail-we0-
f196.google.com[74.125.82.196]"
}
```

elasticsearch.

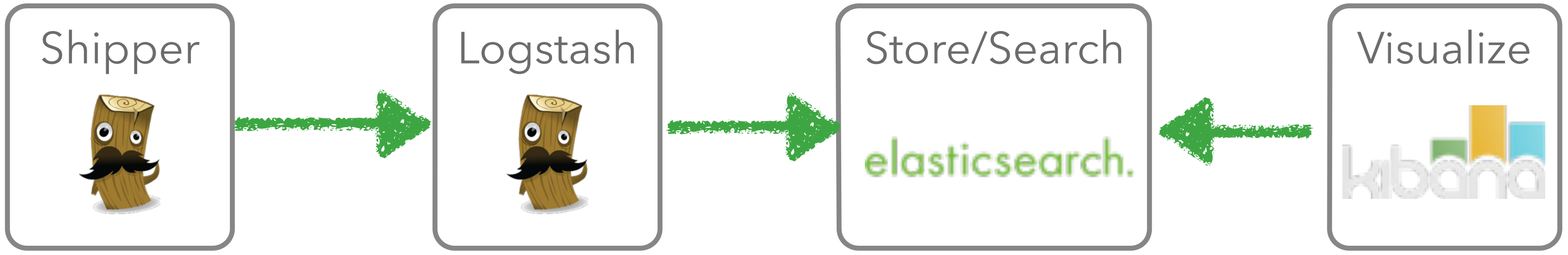
CLF log files

```
{
        "message" => "193.99.144.85 - - [23/Jan/2014:17:11:55 +0000]
\"GET / HTTP/1.1\" 200 140 \"-\" \"Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/535.19 (KHTML, like Gecko) Chrome/18.0.1025.5 Safari/
535.19\"",
       "@version" => "1",
     "@timestamp" => "2014-01-24T07:56:02.460Z",
           "host" => "kryptic.local",
       "clientip" => "193.99.144.85",
          "ident" => "-",
           "auth" => "-",
      "timestamp" => "23/Jan/2014:17:11:55 +0000",
           "verb" => "GET",
        "request" => "/",
    "httpversion" => "1.1",
       "response" => "200",
          "bytes" => "140",
       "referrer" => "\"-\"",
          "agent" => "\"Mozilla/5.0 (Windows NT 6.1; WOW64)
AppleWebKit/535.19 (KHTML, like Gecko) Chrome/18.0.1025.5 Safari/
535.19\""
}
```

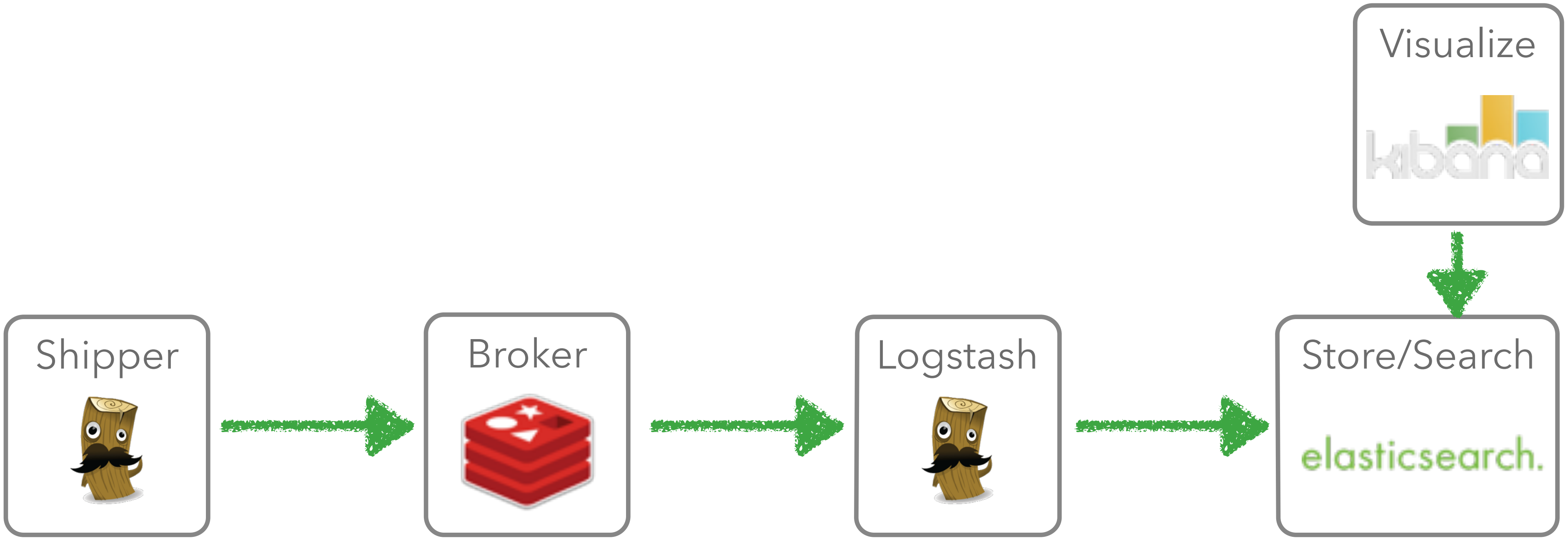
Write to elasticsearch

```
input { stdin {} }

filter {
  grok {
    match => [ message, "%{COMBINEDAPACHELOG}" ]
  }
}

output {
  elasticsearch {
    protocol => “http”
  }
}
```

Deploying ELK

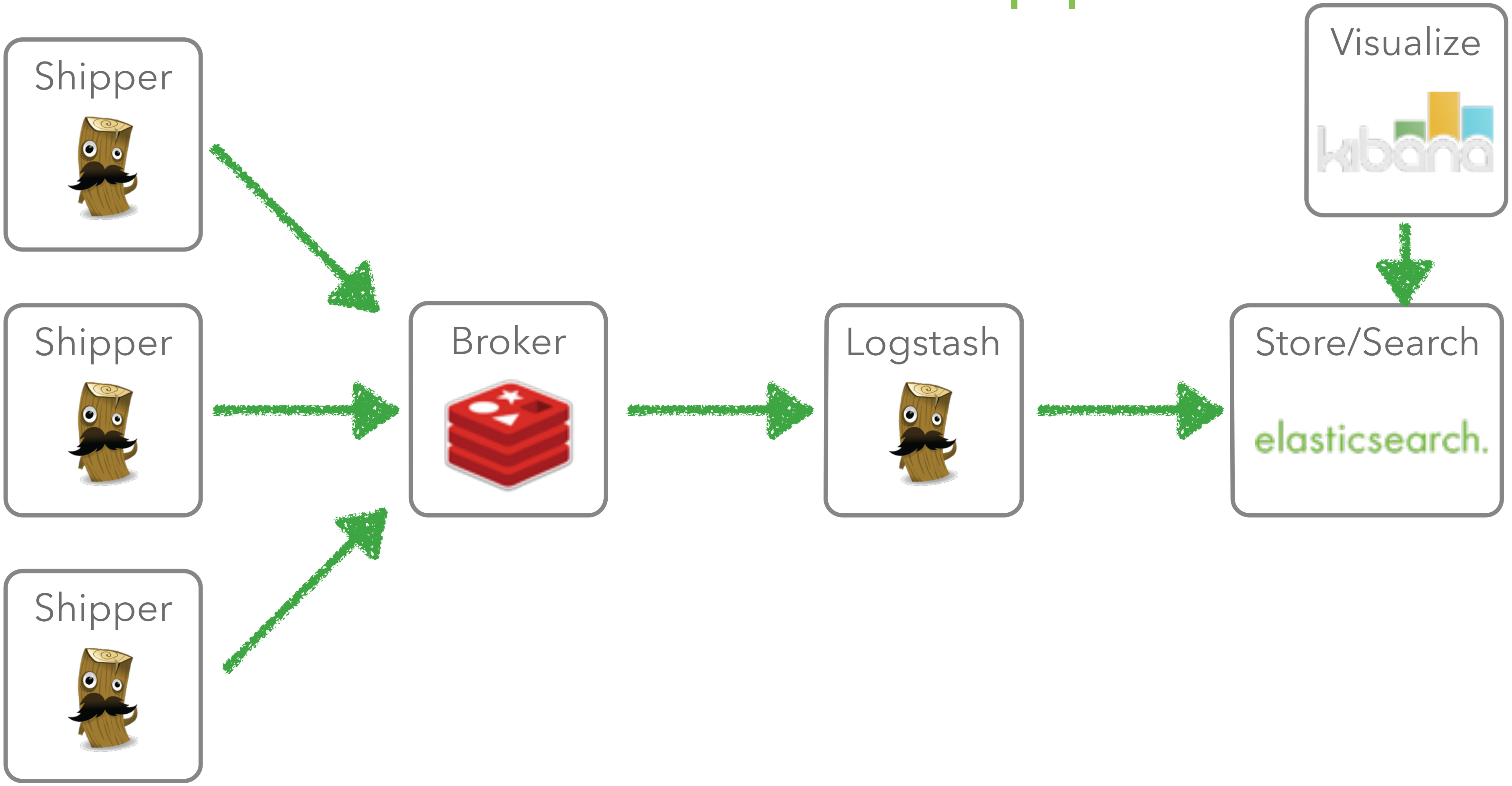


Add a broker



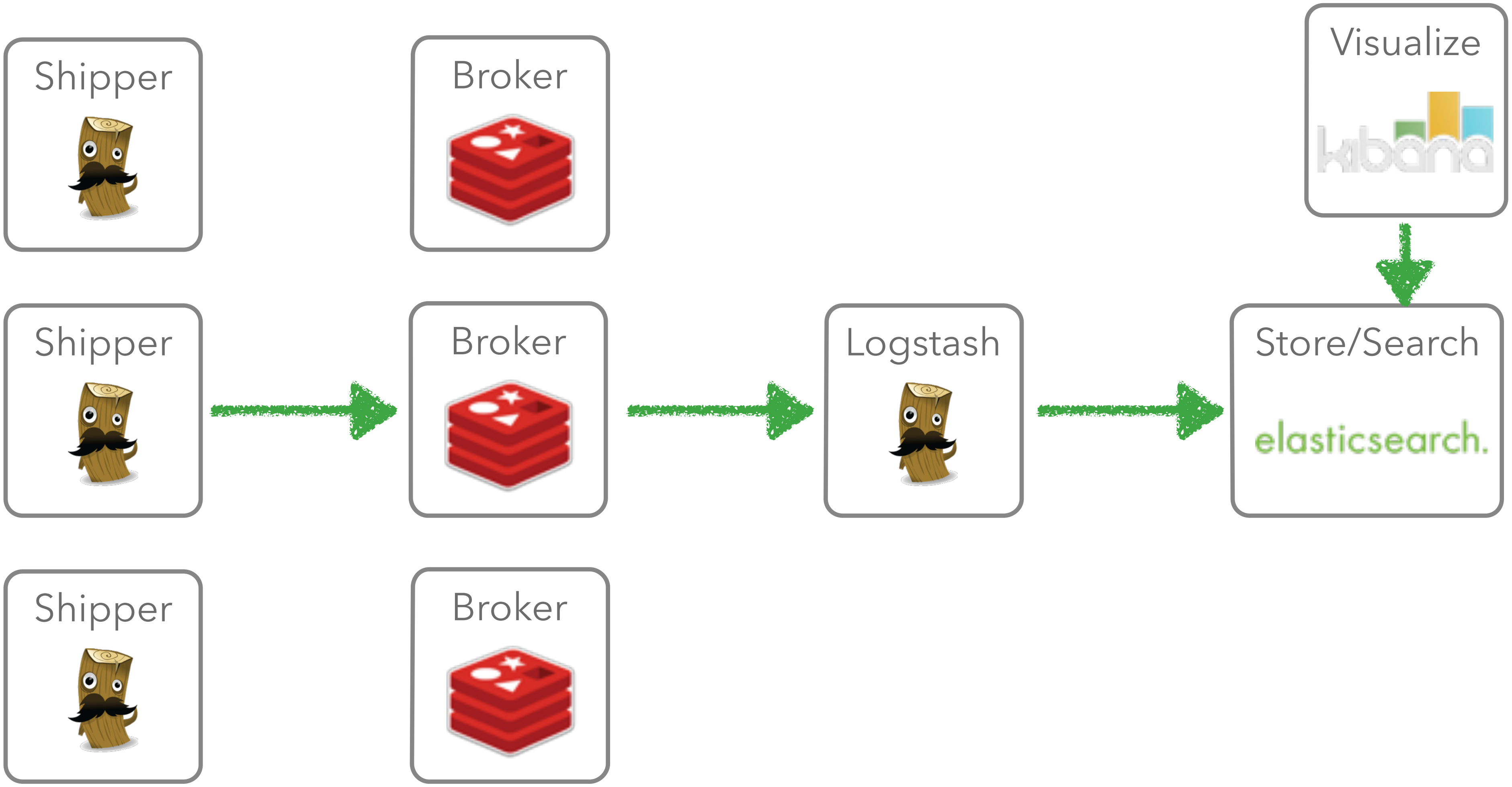
elasticsearch.

Scale out the shipper

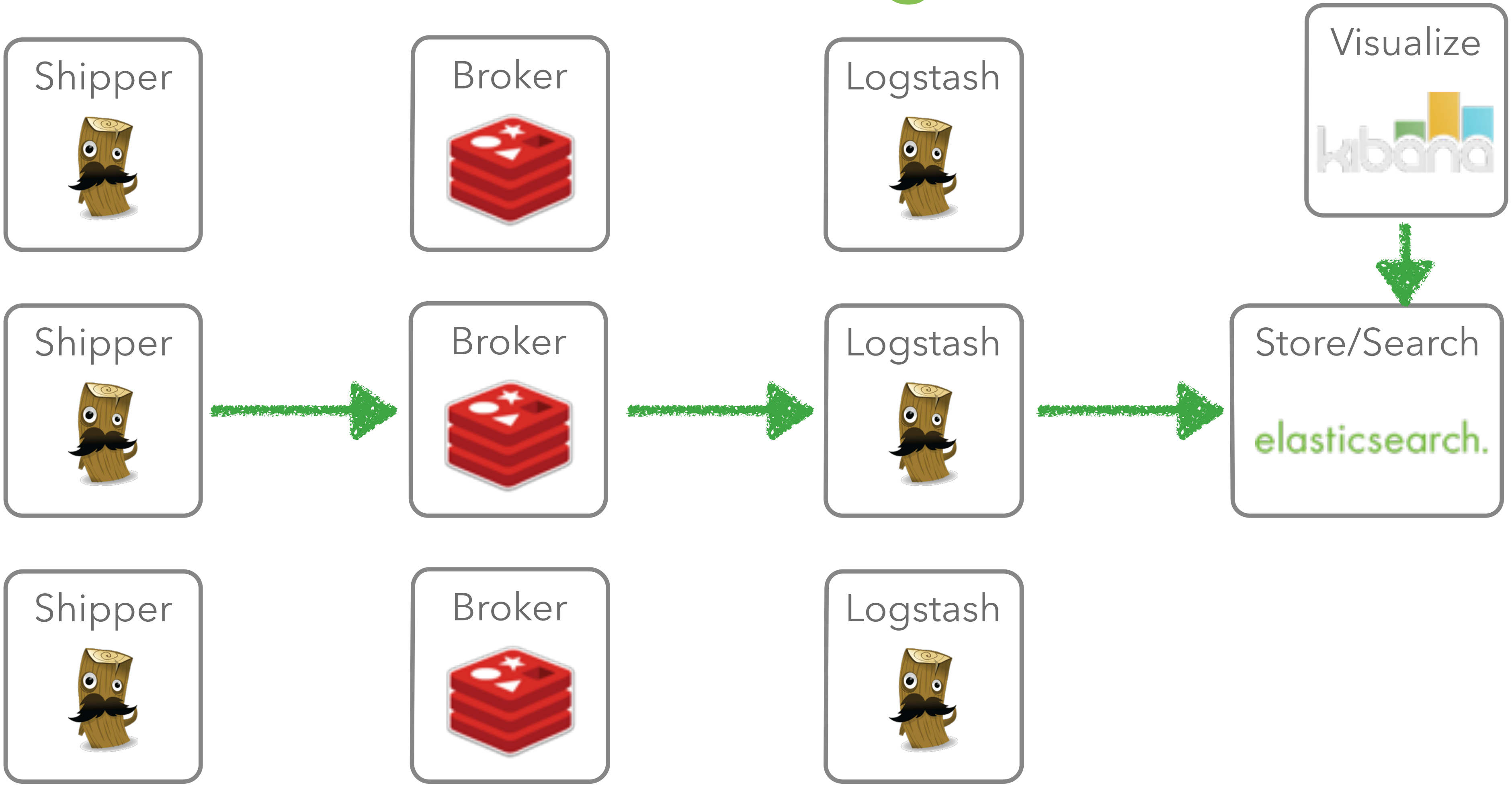


elasticsearch.

Scale out the broker

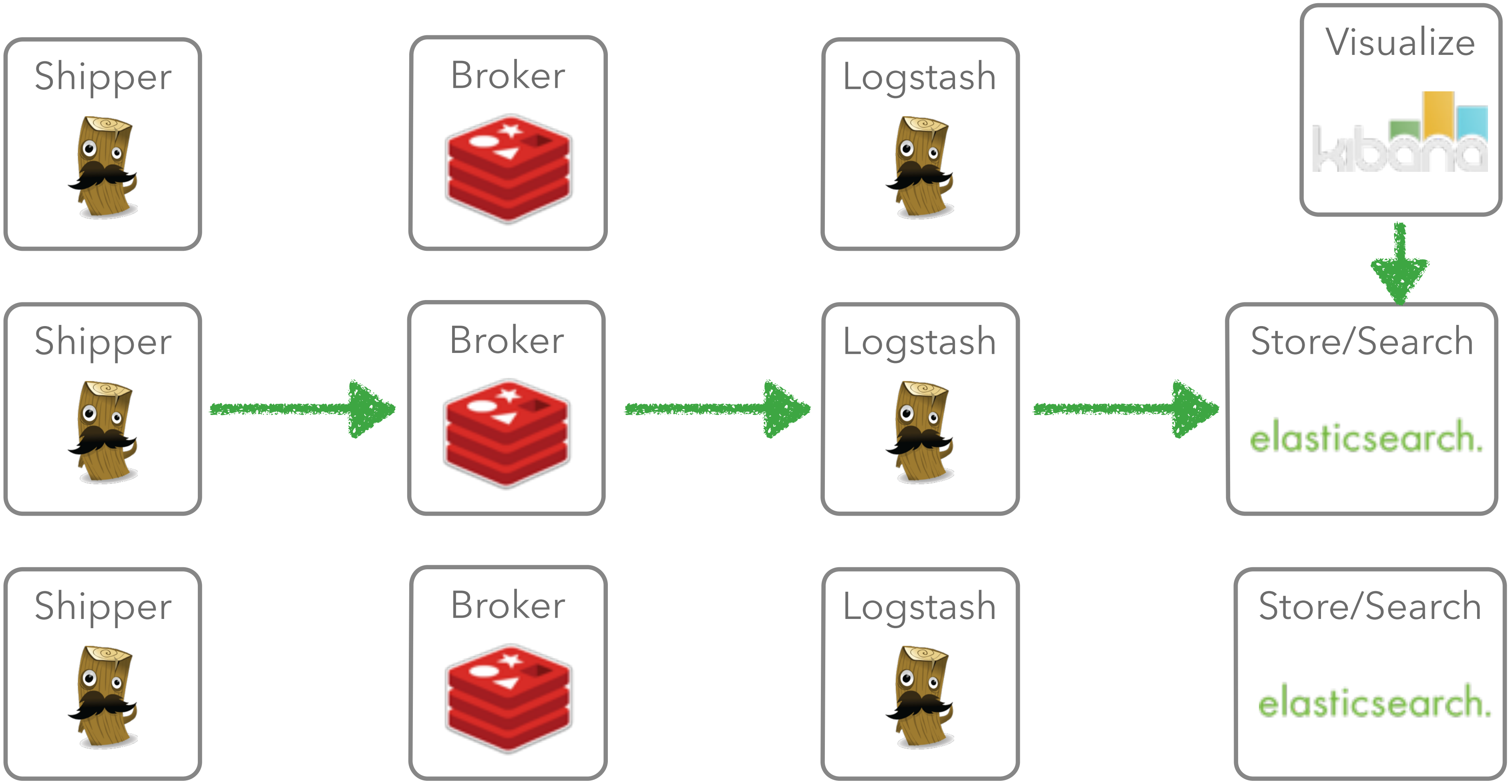


Scale out Logstash



elasticsearch.

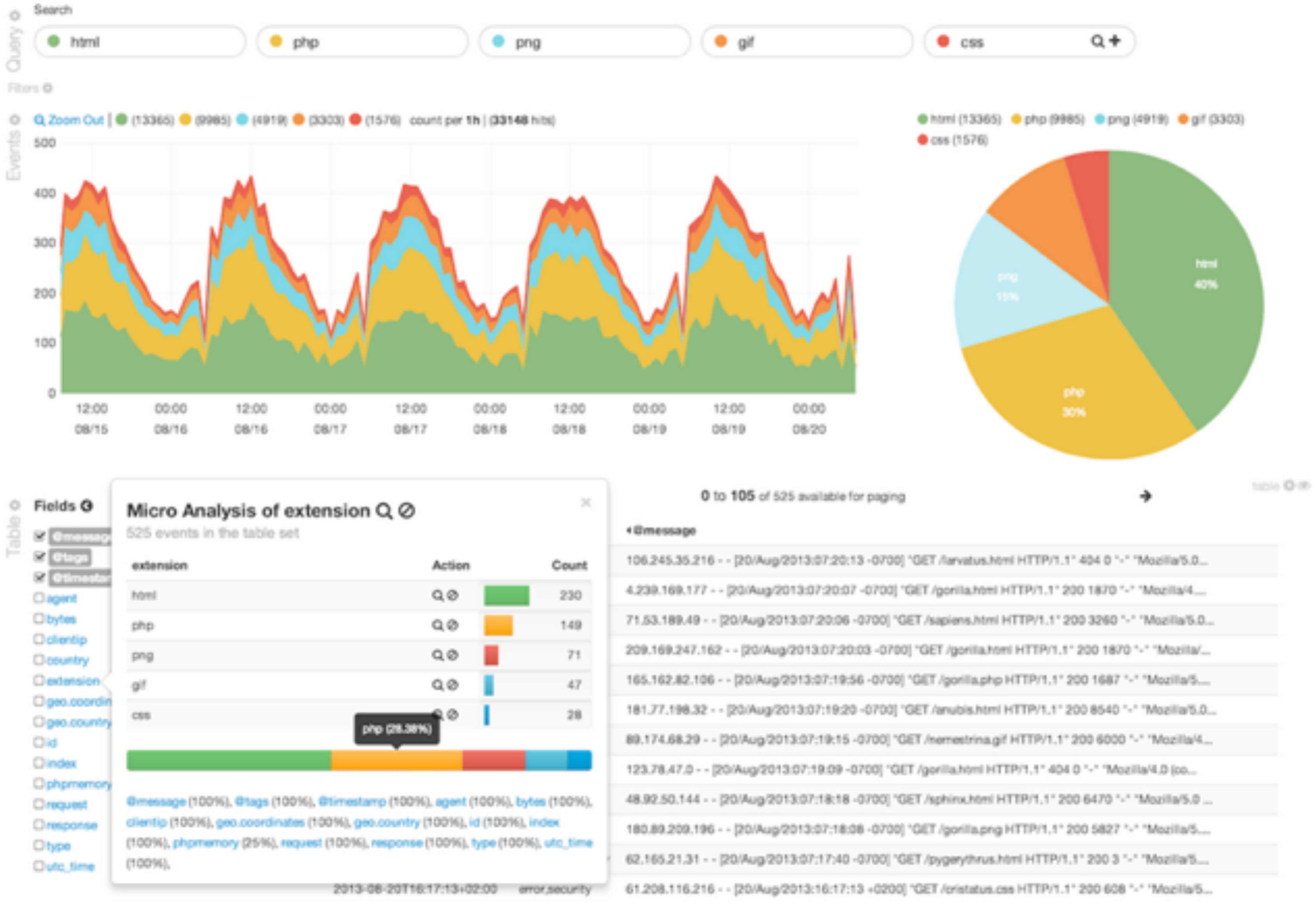
Scale out Elasticsearch



Visualise with Kibana

- jss/css
- Host under your favourite web server
  apache, nginx, IIS
- v4 on the way, currently in beta
- Lots of shiny!

Kibana



Micro Analysis of extension

525 events in the table set

extension	Action	Count
html		230
php		149
png		71
gif		47
css		28

php (28.38%)

@message (100%), @tags (100%), @timestamp (100%), agent (100%), bytes (100%), clientip (100%), geo.coordinates (100%), geo.country (100%), id (100%), index (100%), phpmemory (25%), request (100%), response (100%), type (100%), utc_time (100%).

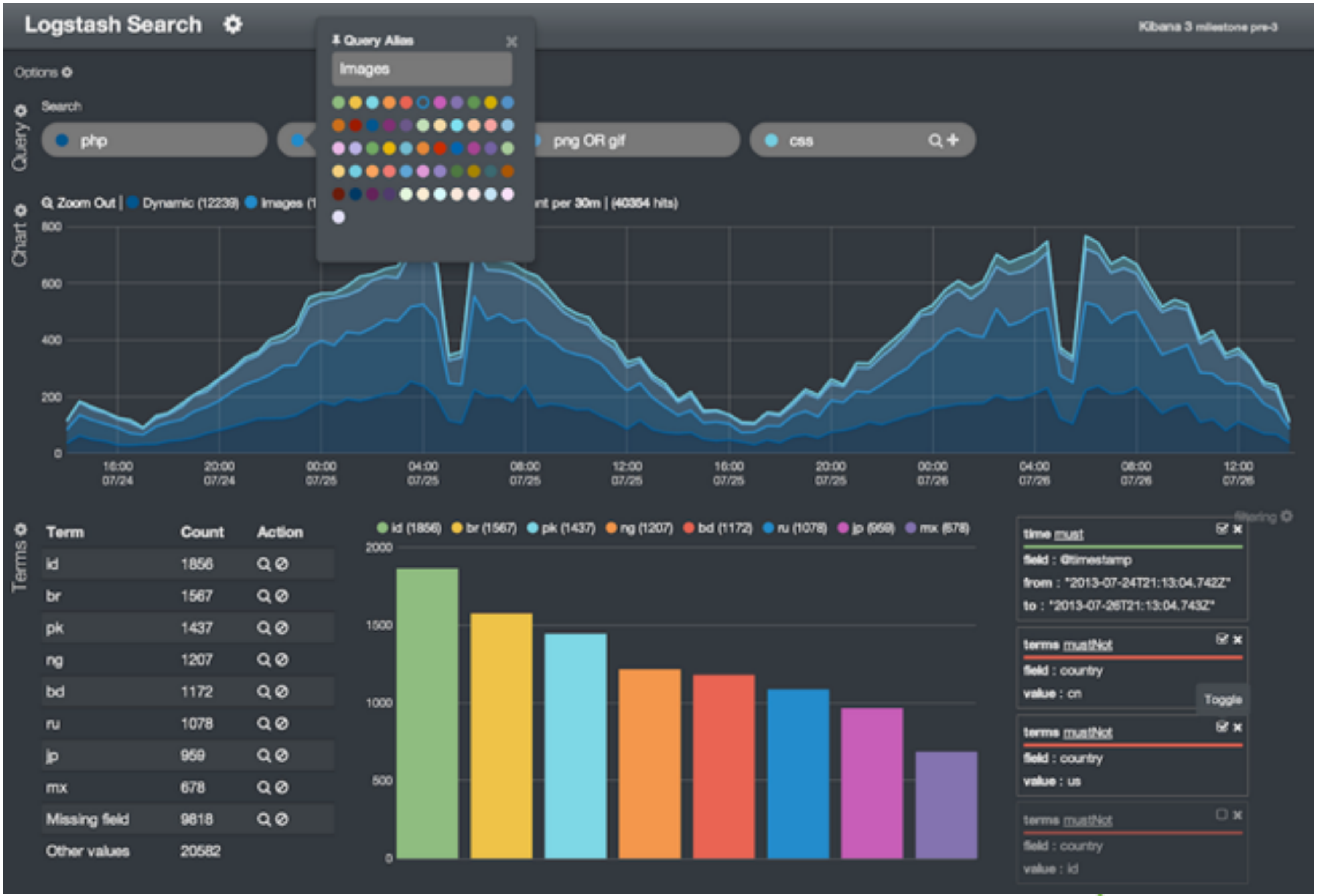
2013-08-20T16:17:13+02:00 error,security

0 to 105 of 525 available for paging

```
@message
106.245.35.216 - - [20/Aug/2013:07:20:13 -0700] "GET /larvatus.html HTTP/1.1" 404 0 "-" "Mozilla/5.0...
4.239.169.177 - - [20/Aug/2013:07:20:07 -0700] "GET /gorilla.html HTTP/1.1" 200 1870 "-" "Mozilla/4....
71.53.189.49 - - [20/Aug/2013:07:20:06 -0700] "GET /sapiens.html HTTP/1.1" 200 3260 "-" "Mozilla/5.0...
209.169.247.162 - - [20/Aug/2013:07:20:03 -0700] "GET /gorilla.html HTTP/1.1" 200 1870 "-" "Mozilla/...
165.162.82.106 - - [20/Aug/2013:07:19:56 -0700] "GET /gorilla.php HTTP/1.1" 200 1687 "-" "Mozilla/5....
181.77.198.32 - - [20/Aug/2013:07:19:20 -0700] "GET /anubis.html HTTP/1.1" 200 8540 "-" "Mozilla/5.0...
89.174.68.29 - - [20/Aug/2013:07:19:15 -0700] "GET /nemestrina.gif HTTP/1.1" 200 6000 "-" "Mozilla/4...
123.78.47.0 - - [20/Aug/2013:07:19:09 -0700] "GET /gorilla.html HTTP/1.1" 404 0 "-" "Mozilla/4.0 (co...
48.92.50.144 - - [20/Aug/2013:07:18:18 -0700] "GET /sphinx.html HTTP/1.1" 200 6470 "-" "Mozilla/5.0 ...
180.89.209.196 - - [20/Aug/2013:07:18:08 -0700] "GET /gorilla.png HTTP/1.1" 200 5827 "-" "Mozilla/5....
62.165.21.31 - - [20/Aug/2013:07:17:40 -0700] "GET /pygerythrus.html HTTP/1.1" 200 3 "-" "Mozilla/5....
61.208.116.216 - - [20/Aug/2013:16:17:13 +0200] "GET /cristatus.css HTTP/1.1" 200 608 "-" "Mozilla/5...
```

elasticsearch.

Kibana



Term	Count	Action
id	1856	
br	1567	
pk	1437	
ng	1207	
bd	1172	
ru	1078	
jp	959	
mx	678	
Missing field	9818	
Other values	20582	

time must
field : @timestamp
from : "2013-07-24T21:13:04.742Z"
to : "2013-07-26T21:13:04.743Z"

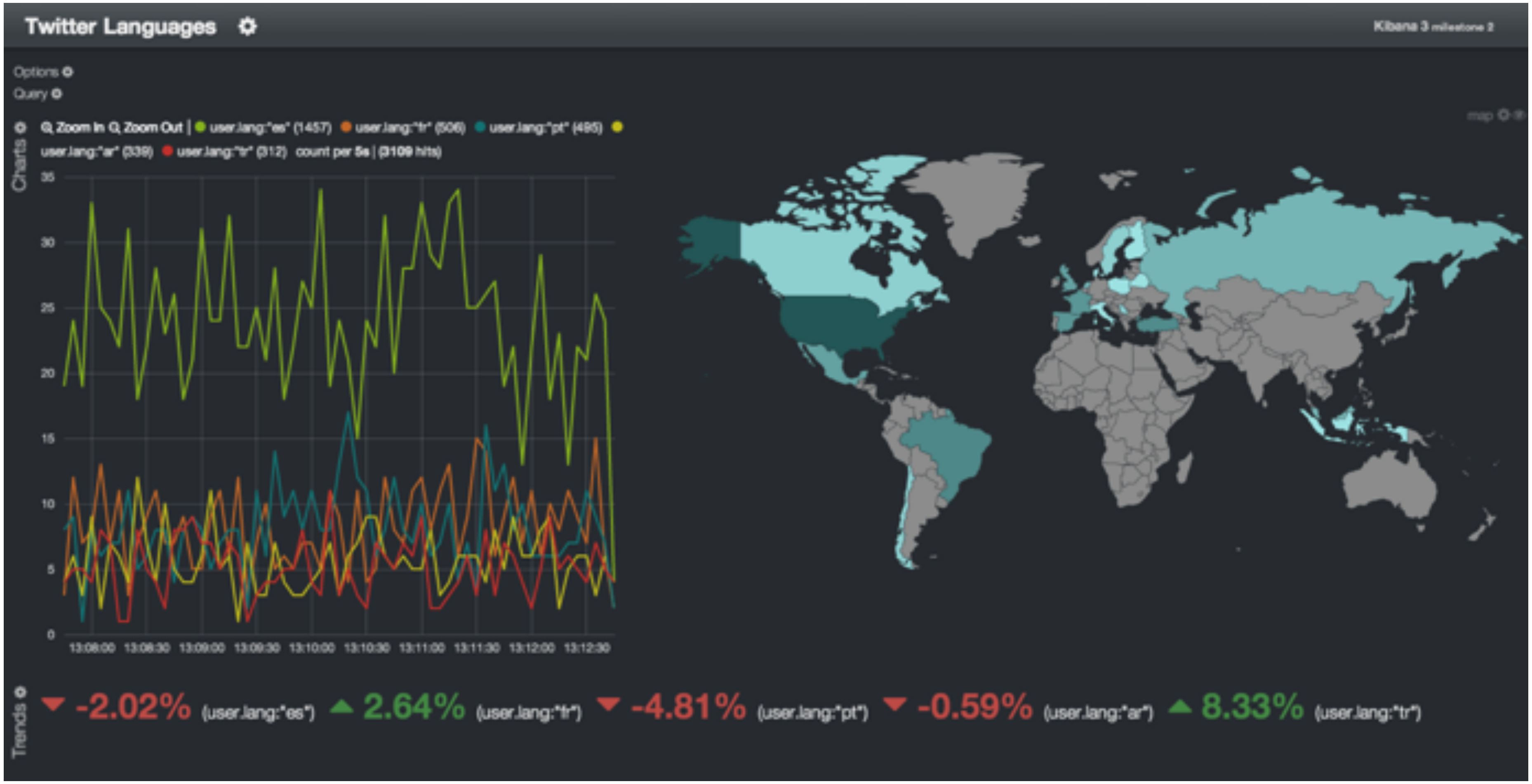
terms mustNot
field : country
value : cn

terms mustNot
field : country
value : us

terms mustNot
field : country
value : id

elasticsearch.

Kibana



elasticsearch.

Useful helpers

- Curator: index management
  http://www.elasticsearch.org/blog/curator-tending-your-time-series-indices/

- Puppet & Chef modules
  https://forge.puppetlabs.com/elasticsearch
  https://github.com/elasticsearch/cookbook-elasticsearch/

- logstash forwarder: low overhead collector
  https://github.com/elasticsearch/logstash-forwarder

- grokdebugger: log pattern matching
  http://grokdebug.herokuapp.com/

elasticsearch.

More info

- Github: https://github.com/elasticsearch
- Docs: http://www.elasticsearch.org/guide/
  elasticsearch and clients, logstash, kibana and more
- Google groups: elasticsearch and logstash-users
- IRC channels
  #elasticsearch, #logstash and #kibana on freenode
- We're hiring!
  jobs@elasticsearch.com