

elasticsearch beyond full-text search

#gotoaar #elasticsearch

Alexander Reelsen

@spinscale

alexander.reelsen@elasticsearch.com



About me

- Elasticsearch core developer
Features, bug fixing, package maintenance, documentation, blog posts
- Development support
- Production support
- Trainings
- Conferences & talks

- Interests: Java, JavaScript, web apps

Beyond full-text search?



Unstructured search

GitHub

Explore Features Enterprise Blog

Sign up

Sign in

Search

elasticsearch

Search

📁 Repositories	317
🔗 Code	17,981
🕒 Issues	2,008
👤 Users	2

Languages

Java	167
Ruby	167
JavaScript	139
Python	117
PHP	69
Shell	49
Puppet	40
Perl	38
Scala	16
C#	13

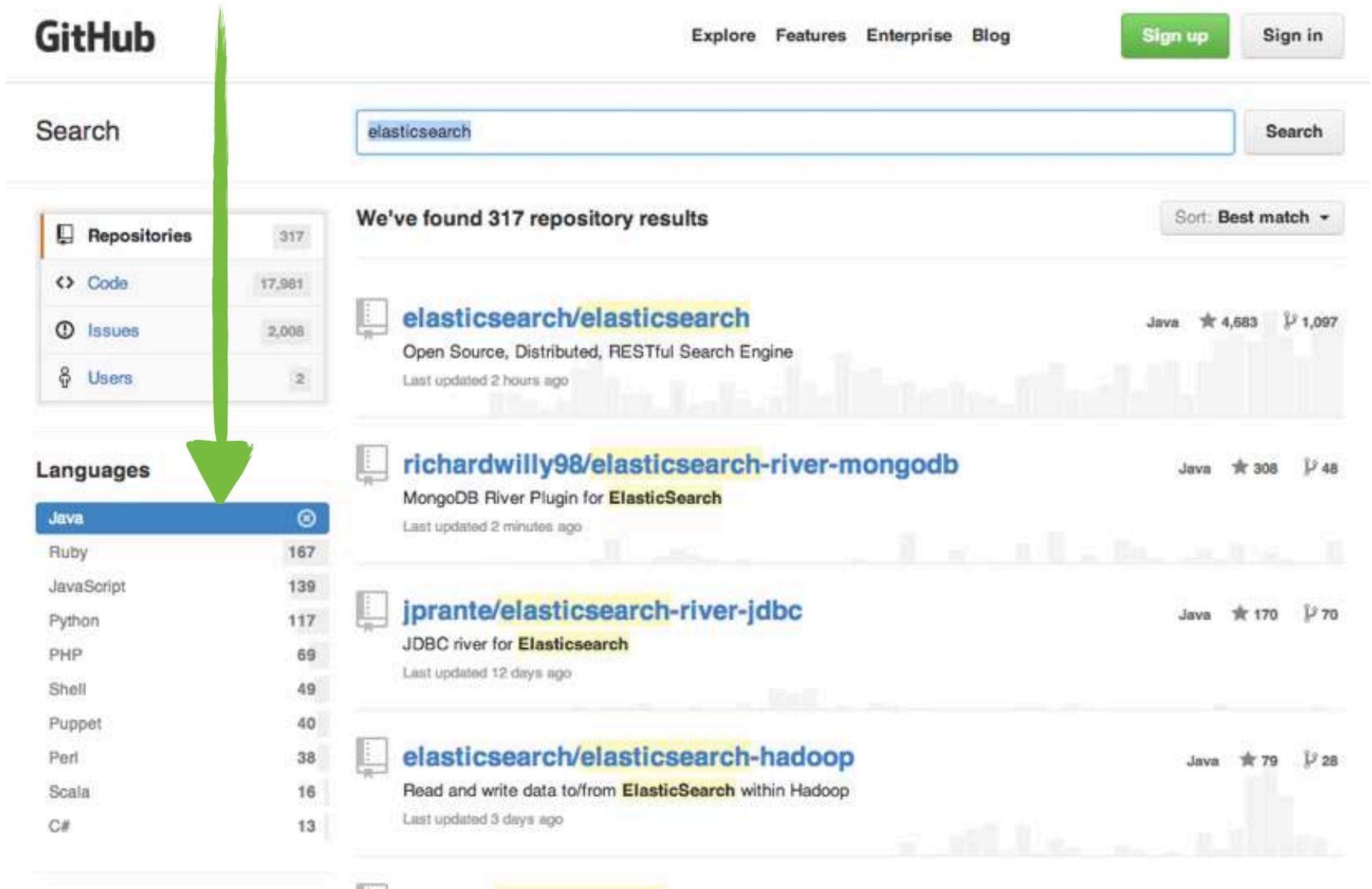
We've found 317 repository results

Sort: Best match ▾

-  **elasticsearch/elasticsearch** Java ★ 4,583 📄 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago
-  **richardwilly98/elasticsearch-river-mongodb** Java ★ 308 📄 48
MongoDB River Plugin for **ElasticSearch**
Last updated 2 minutes ago
-  **jprante/elasticsearch-river-jdbc** Java ★ 170 📄 70
JDBC river for **Elasticsearch**
Last updated 12 days ago
-  **elasticsearch/elasticsearch-hadoop** Java ★ 79 📄 28
Read and write data to/from **ElasticSearch** within Hadoop
Last updated 3 days ago

elasticsearch.

Structured search



The screenshot shows the GitHub search interface. At the top left is the GitHub logo. To the right are links for 'Explore', 'Features', 'Enterprise', and 'Blog', along with 'Sign up' and 'Sign in' buttons. Below the logo is a 'Search' section with a search bar containing 'elasticsearch' and a 'Search' button. On the left side, there are navigation links for 'Repositories' (317), 'Code' (17,981), 'Issues' (2,008), and 'Users' (2). Below these is a 'Languages' section with a list of languages and their repository counts: Java (317), Ruby (167), JavaScript (139), Python (117), PHP (69), Shell (49), Puppet (40), Perl (38), Scala (16), and C# (13). A large green arrow points from the top of the page down to the 'Java' language filter, which is highlighted in blue. The main content area shows search results for 'elasticsearch'. The first result is 'elasticsearch/elasticsearch', an Open Source, Distributed, RESTful Search Engine, last updated 2 hours ago, with 4,583 stars and 1,097 forks. The second result is 'richardwilly98/elasticsearch-river-mongodb', a MongoDB River Plugin for ElasticSearch, last updated 2 minutes ago, with 308 stars and 48 forks. The third result is 'jprante/elasticsearch-river-jdbc', a JDBC river for Elasticsearch, last updated 12 days ago, with 170 stars and 70 forks. The fourth result is 'elasticsearch/elasticsearch-hadoop', for reading and writing data to/from ElasticSearch within Hadoop, last updated 5 days ago, with 79 stars and 28 forks. The search results are sorted by 'Best match'.

Enrichment

GitHub

Explore Features Enterprise Blog

Sign up

Sign in

Search

elasticsearch

Search

Repositories	317
Code	17,981
Issues	2,008
Users	2

Languages

Java	167
Ruby	167
JavaScript	139
Python	117
PHP	69
Shell	49
Puppet	40
Perl	38
Scala	16
C#	13

We've found 317 repository results

Sort: Best match

- elasticsearch/elasticsearch** Java ★ 4,583 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago
- richardwilly98/elasticsearch-river-mongodb** Java ★ 308 48
MongoDB River Plugin for ElasticSearch
Last updated 2 minutes ago
- jprante/elasticsearch-river-jdbc** Java ★ 170 70
JDBC river for Elasticsearch
Last updated 12 days ago
- elasticsearch/elasticsearch-hadoop** Java ★ 79 28
Read and write data to/from ElasticSearch within Hadoop
Last updated 5 days ago

elasticsearch.

Sorting

GitHub

Explore Features Enterprise Blog

Sign up

Sign in

Search

elasticsearch

Search

Sort: Best match ▾

Repositories	317
Code	17,981
Issues	2,008
Users	2

Languages

Java	167
Ruby	139
JavaScript	117
Python	69
PHP	49
Shell	40
Puppet	38
Perl	16
Scala	13
C#	

We've found 317 repository results

- elasticsearch/elasticsearch** Java ★ 4,583 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago
- richardwilly98/elasticsearch-river-mongodb** Java ★ 308 48
MongoDB River Plugin for ElasticSearch
Last updated 2 minutes ago
- jprante/elasticsearch-river-jdbc** Java ★ 170 70
JDBC river for Elasticsearch
Last updated 12 days ago
- elasticsearch/elasticsearch-hadoop** Java ★ 79 28
Read and write data to/from ElasticSearch within Hadoop
Last updated 3 days ago

elasticsearch.

Pagination

GitHub

Explore Features Enterprise Blog

Sign up

Sign in

Search

elasticsearch

Search

📁	Repositories	317
<>	Code	17,981
🕒	Issues	2,008
👤	Users	2

We've found 317 repository results

Sort: Best match ▾

 **elasticsearch/elasticsearch** Java ★ 4,583 🍴 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago

 **spinscal/elasticsearch-suggest-plugin** Java ★ 103 🍴 23
Plugin for **elasticsearch** which uses the lucene FST Suggester
Last updated 4 days ago

◀ 1 2 3 4 5 6 7 8 9 ... 31 32 ▶

How are these search results? [Tell us!](#)



Aggregation

GitHub

Explore Features Enterprise Blog

Sign up

Sign in

Search

elasticsearch

Search

Repositories 317

Code 7,981

Issues 1,008

Users 2

Languages

Java 167

Ruby 167

JavaScript 139

Python 117

PHP 69

Shell 49

Puppet 40

Perl 38

Scala 16

C# 13

We've found 317 repository results

Sort: Best match

elasticsearch/elasticsearch Java ★ 4,583 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago

richardwilly98/elasticsearch-river-mongodb Java ★ 308 48
MongoDB River Plugin for ElasticSearch
Last updated 2 minutes ago

jprante/elasticsearch-river-jdbc Java ★ 170 70
JDBC river for Elasticsearch
Last updated 12 days ago

elasticsearch/elasticsearch-hadoop Java ★ 79 28
Read and write data to/from ElasticSearch within Hadoop
Last updated 3 days ago

elasticsearch.

Suggestions



GitHub This repository:

[Sign up](#) [Sign in](#)

★ **Star** 4,683 [Fork](#) 1,097

[New Issue](#)

1 2 3 ... 19

Labels

- Lucene 4.5 Upgrade
- breaking
- bug
- enhancement
- feature
- non-issue

Issues

Issue Title	Count
elasticsearch/elasticsearch#1726 debian package violates naming convention	1
elasticsearch/elasticsearch#3571 debian package init-script: start-stop-daemon ne	11
elasticsearch/elasticsearch#1681 Debian pkg	10
elasticsearch/elasticsearch#3286 There is no official debian /ubuntu repository	9
elasticsearch/elasticsearch#3500 Elasticsearch should include debian 's standard j	9
elasticsearch/elasticsearch#1526 Moving debian package to maven	1

Search elasticsearch/elasticsearch for 'debian'

Search GitHub for 'debian'

Opened by s1monw 14 hours ago

NoShardAvailableActionException in ES 0.90.3 on startup #3700
Opened by richardwilly98 a day ago

Feature Request: Don't reindex the document when updating non-indexed fields #3696
Opened by ddorian 2 days ago 4 comments

Introduction



Elasticsearch in 10 seconds

- Schema-free, REST & JSON based distributed document store
- Open source: Apache License 2.0
- Zero configuration

- Used by github, mozilla, soundcloud, stack overflow, foursquare, fog creek, stumbleupon

Zero configuration

```
$ wget https://download.elasticsearch.org/...  
$ tar -xf elasticsearch-0.90.5.tar.gz  
$ ./elasticsearch-0.90.5/bin/elasticsearch -f  
... [INFO ][node][Ghost Maker] {0.90.5}[5645]: initializing ...
```

Index & search data

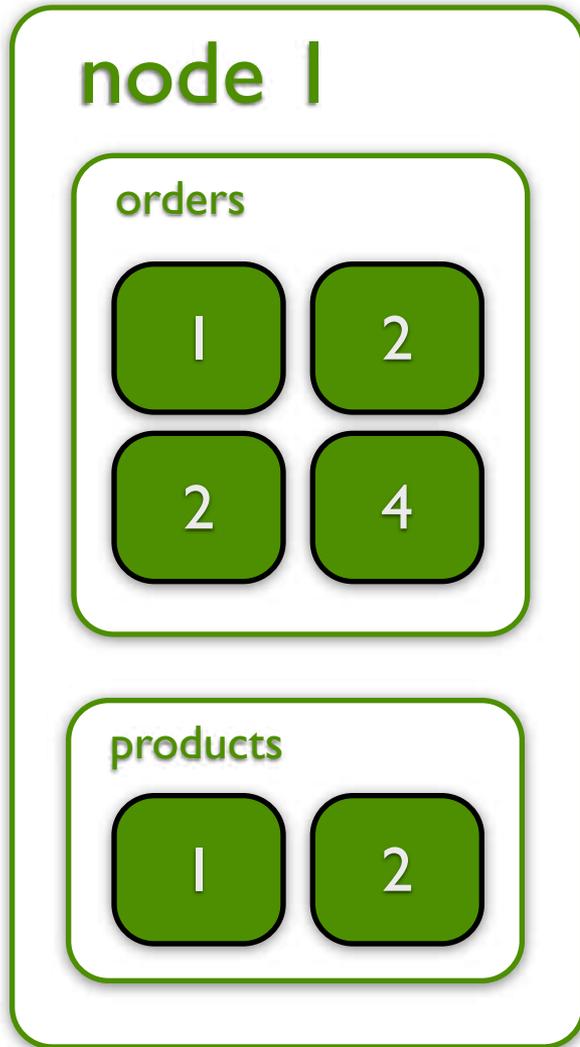
```
curl -X PUT localhost:9200/products/product/1 -d '{
  "created_at" : "2013/09/05 15:45:10",
  "name" : "Macbook Air",
  "price" : {
    "net" : 1699,
    "tax" : 322.81,
  }
}'
```

```
curl -X GET 'localhost:9200/products/product/_search?q=macbook'
```

Distributed

- **Replication: Data duplication**
 - Read scalability
 - Removing SPOF
- **Sharding: Data partitioning**
 - Split logical data over several machines
 - Write scalability
 - Control data flows

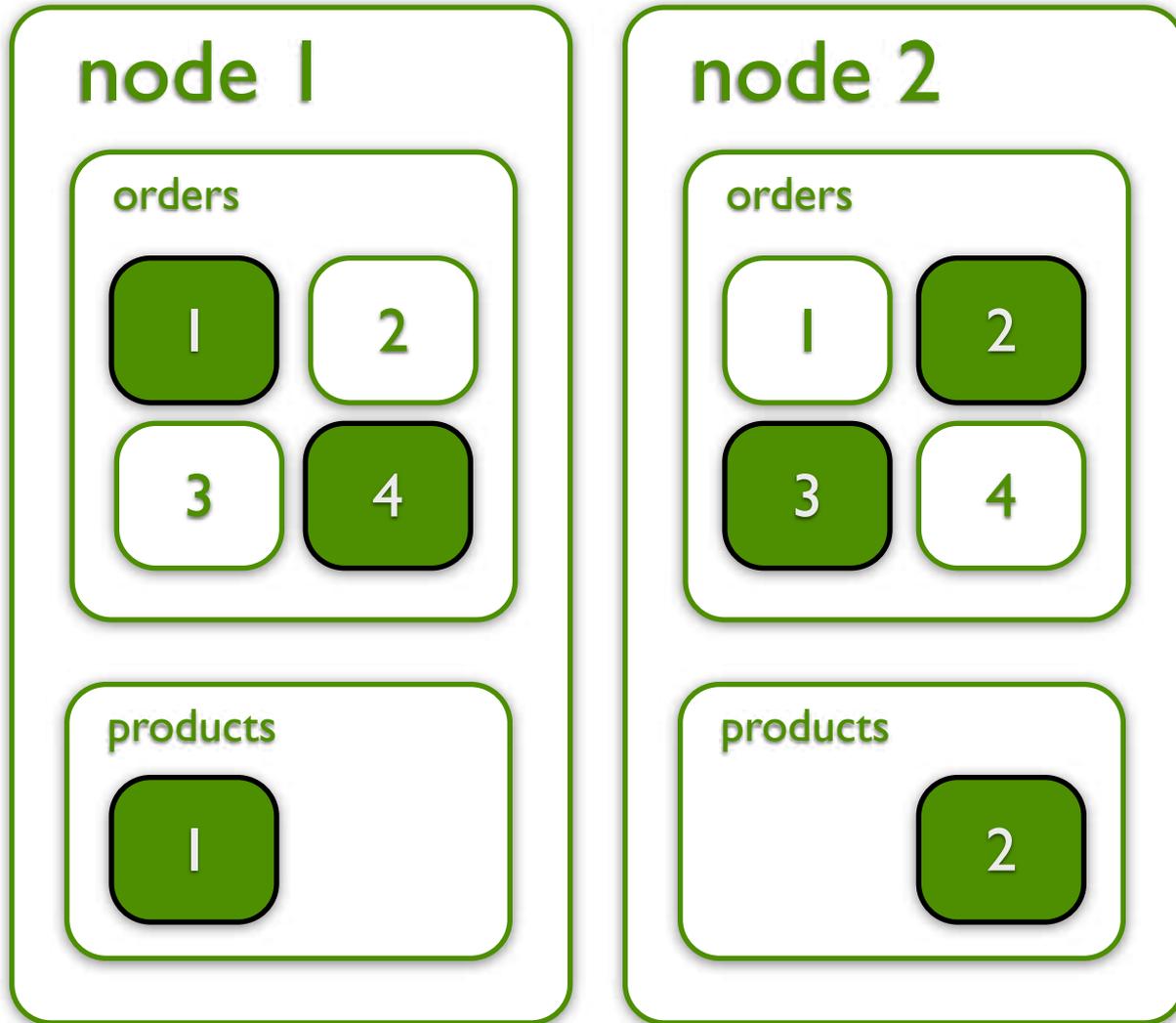
Distributed



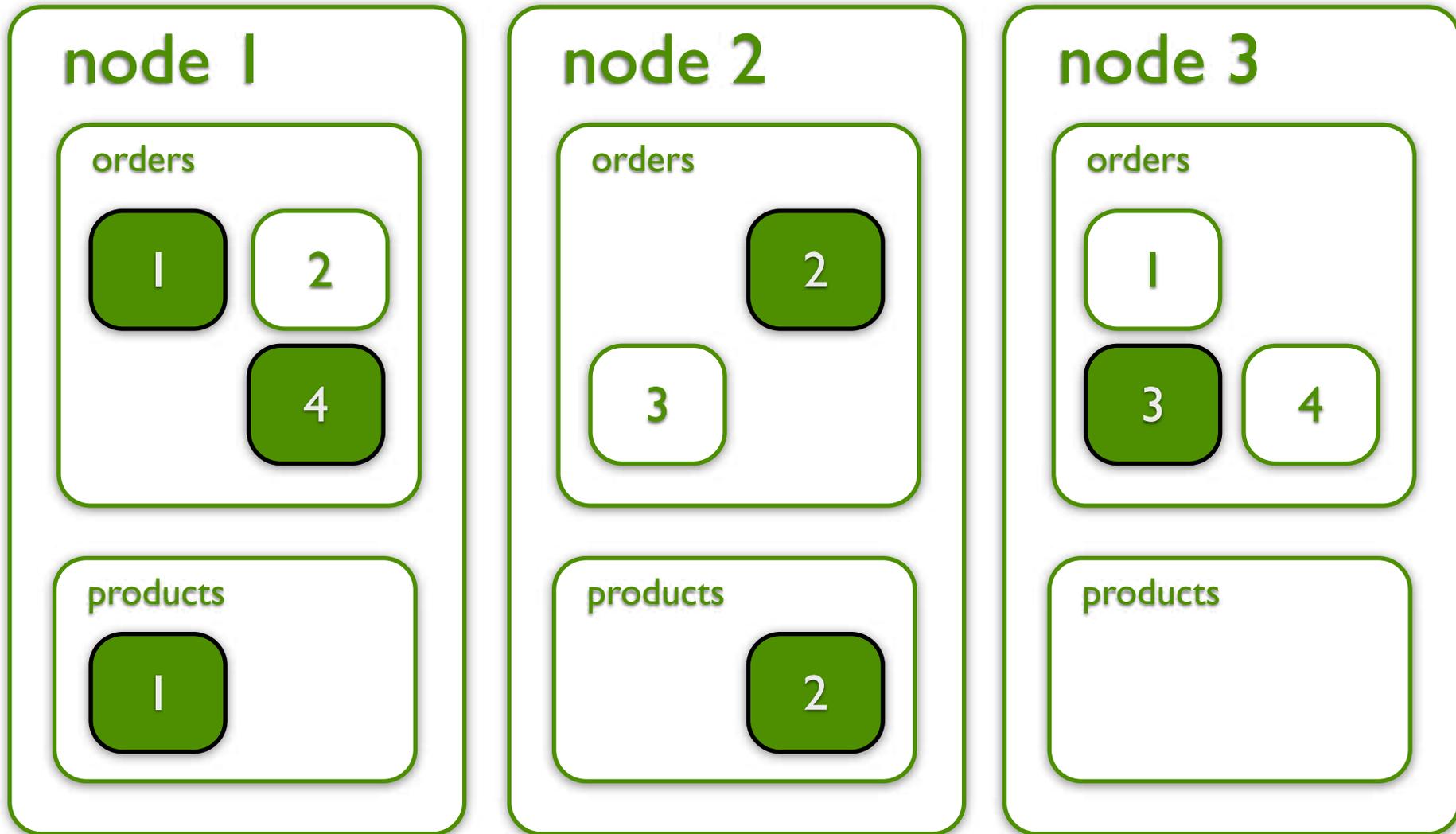
```
curl -X PUT localhost:9200/orders -d '{  
  "settings.index.number_of_shards" : 4  
  "settings.index.number_of_replicas" : 1  
}'
```

```
curl -X PUT localhost:9200/products -d '{  
  "settings.index.number_of_shards" : 2  
  "settings.index.number_of_replicas" : 0  
}'
```

Distributed



Distributed



Ecosystem

- Plugins
- Clients for many languages
Ruby, Python, PHP, Perl
Javascript, Scala, Clojure
- Kibana & Logstash
- Hadoop integration

**From data
to information**



What is data?

- Whatever provides value for your business
- Domain data
 - Internal: Orders, products
 - External: Social media streams, email
- Application data
 - Log files
 - Metrics

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?
- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

Order as JSON

```
curl -X PUT localhost:9200/orders/order/1 -d '{
  "created_at" : "2013/09/05 15:45:10",
  "items" : [
    ...
  ]
  "total" : 245.37
}'
```

Asking questions to your data

- How many orders were created? **count**
- How many orders were created in the last month?



```
curl -X GET http://localhost:9200/orders/order/_count
```

- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?

filter & count



```
curl -X GET http://localhost:9200/orders/order/_count -d '{
  "range": {
    "created_at": {
      "gte": "2013/09/01",
      "lt": "2013/10/01"
    }
  }
}'
```

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?  **filter**
- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour? **count/day**

Asking questions to your data

```
curl -X GET http://localhost:9200/orders/order/_search -d '{
  "facets": {
    "created": {
      "date_histogram" : {
        "field" : "created_at",
        "interval" : "1d"
      },
      "facet_filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      }
    }
  }
}'
```

count/day

filter

Asking questions to your data

- How many orders were created?
- How many orders were created in the last month?
- How many orders were created every day in the last month?
- What is the average revenue per shopping cart?
- What is the average shopping cart size per order (EUR or #items)? Per hour?

filter

scripting

stats



Asking questions to your data

```
curl -X GET http://localhost:9200/orders/order/_search -d '{
  "facets": {
    "avg_revenue": {
      "facet_filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      }
    },
    "statistical" : {
      "script" : "doc[\u0027total\u0027].value * 0.1 + 2"
    }
  }
}'
```

filter

scripting

stats

Asking questions to your data

- How many orders were created?
filter
 - How many orders were created in the last month?
filter
 - How many orders were created every day in the last month?
scripting
 - What is the average revenue per shopping cart?
stats
 - What is the average shopping cart size per order (EUR or #items)? Per hour?
per hour
-

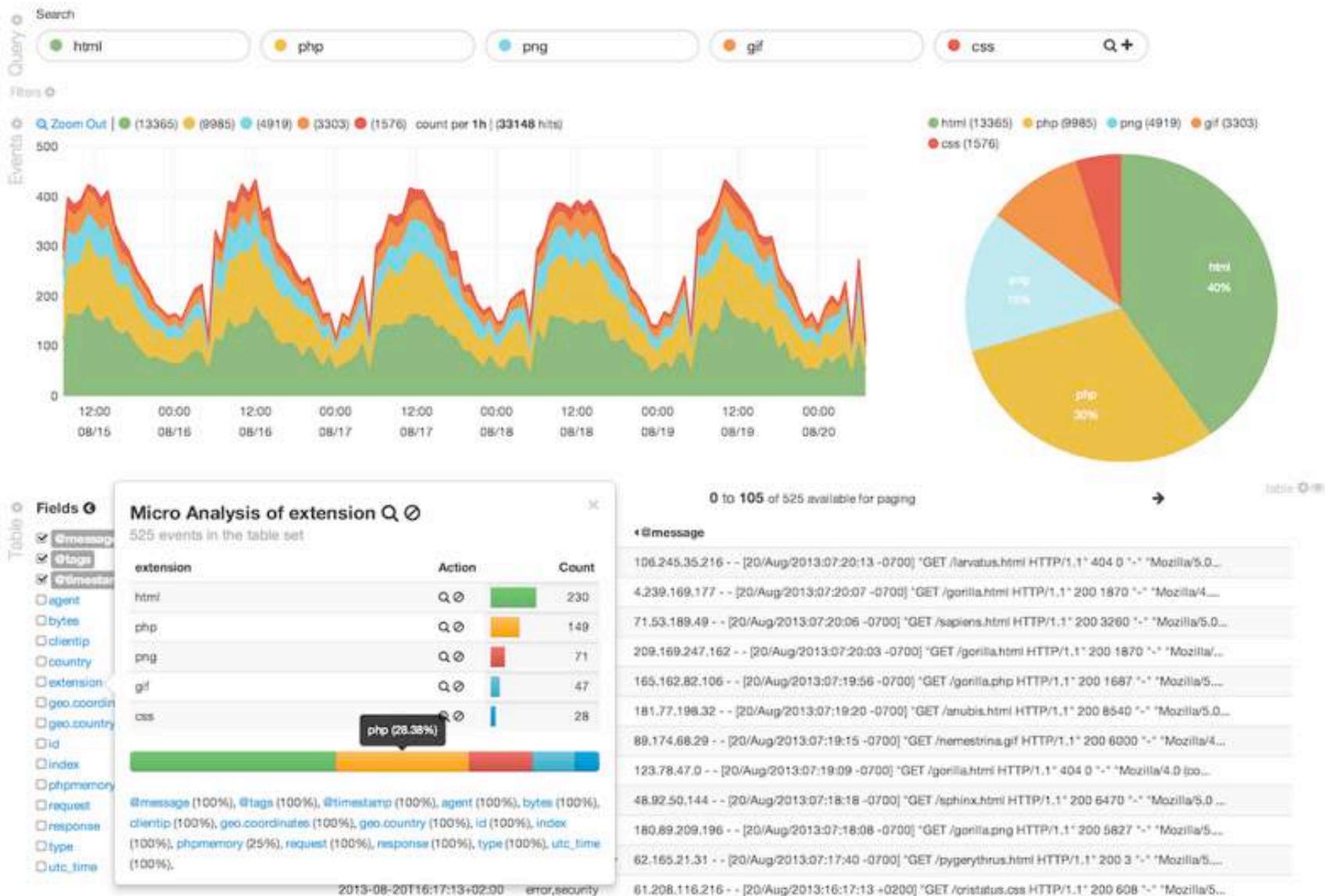
**From data
to visualization**



From numbers to simplicity

- JSON is not a management compatible notation
- Writing your own visualization app for all the different data is tedious
- Enter Kibana!

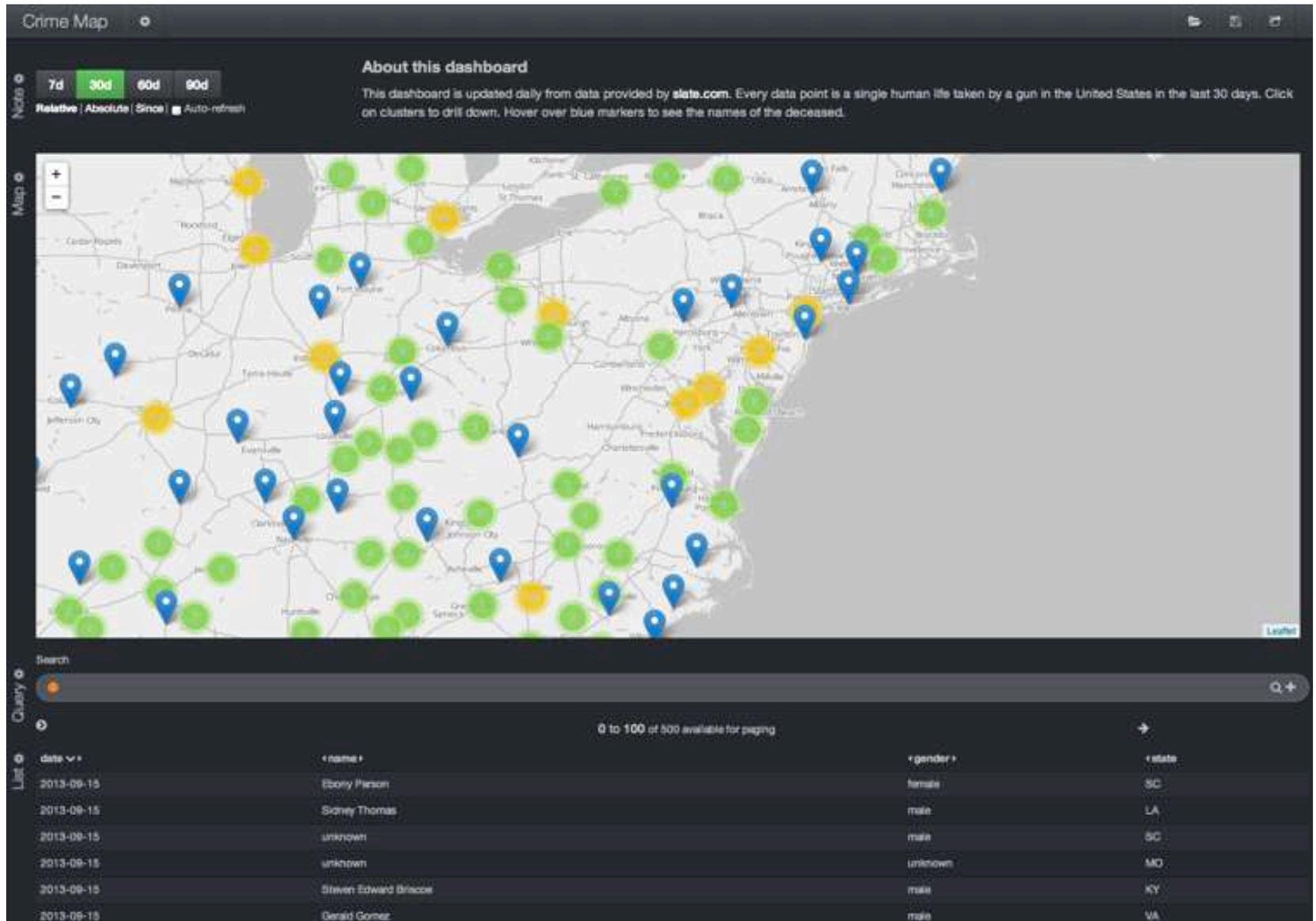
Kibana



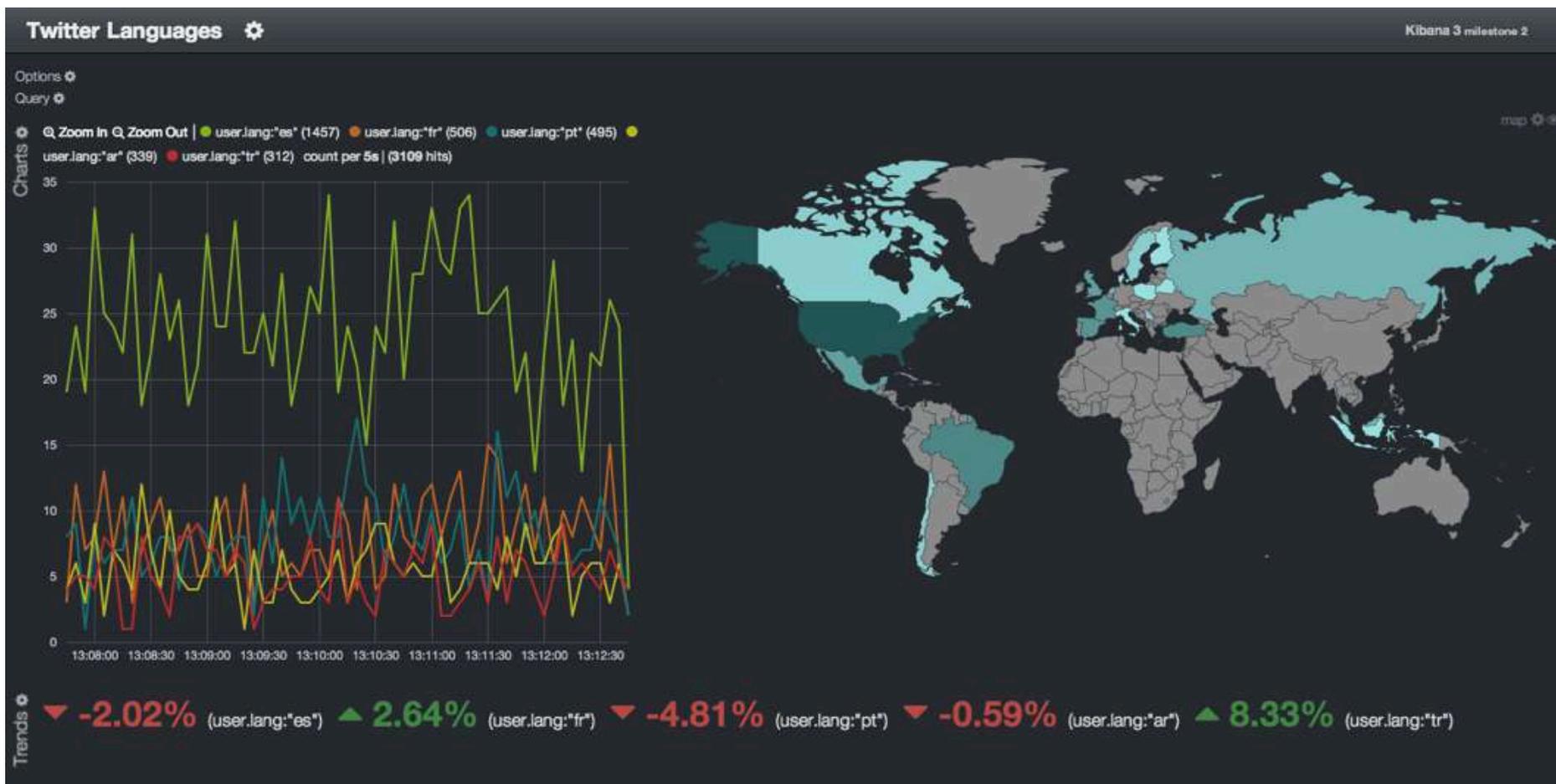
Kibana



Kibana



Kibana



**From data
to notification**



Houston, we have a problem!

- The average response time of your payment API just increased over 2 seconds over the last 15 minutes
- A credit card fraud detection kicks in
- Visits are exploding after the television commercial
- The “win-a-car” voucher has reached its usage limit
- Memory usage exceeds physical memory

Meet the metrics library!

- Measure inside your application
- Gauges, Timers, Counters, Meters, Histograms
- Healthchecks
- Report to elasticsearch



Meet the metrics library!

```
MetricRegistry metrics = new MetricRegistry();
```

```
Meter requestsMeter = metrics.meter("incoming-http-requests");
```

```
// in your app code  
requestsMeter.mark(1);
```

```
Timer responses = metrics.timer("responses");
```

```
Timer.Context context = responses.time();  
try {  
    // etc;  
    return "OK";  
} finally {  
    context.stop();  
}
```

Metrics elasticsearch reporter

- Reports from your application into elasticsearch
- Uses HTTP, no elasticsearch dependency
- Realtime notification via percolation
Sent an email, a pager alert or a MQ message

Percolation

- Normal: Index documents, run queries
- Percolator: Register queries, run against documents
- Use-case: Price agent, contextual ads, classification before indexing (geo, tag, categorization), metrics

Percolation support

```
ElasticsearchReporter reporter =  
    ElasticsearchReporter.forRegistry(registry)  
        .percolateNotifier(new PagerNotifier())  
        .percolateMetrics(".*")  
        .build();  
reporter.start(60, TimeUnit.SECONDS);
```

```
public class PagerNotifier implements Notifier {  
  
    @Override  
    public void notify(JsonMetric metric, String id) {  
        // send pager duty here  
    }  
}
```

Cockpit - Sample App

Cockpit



Add percolation

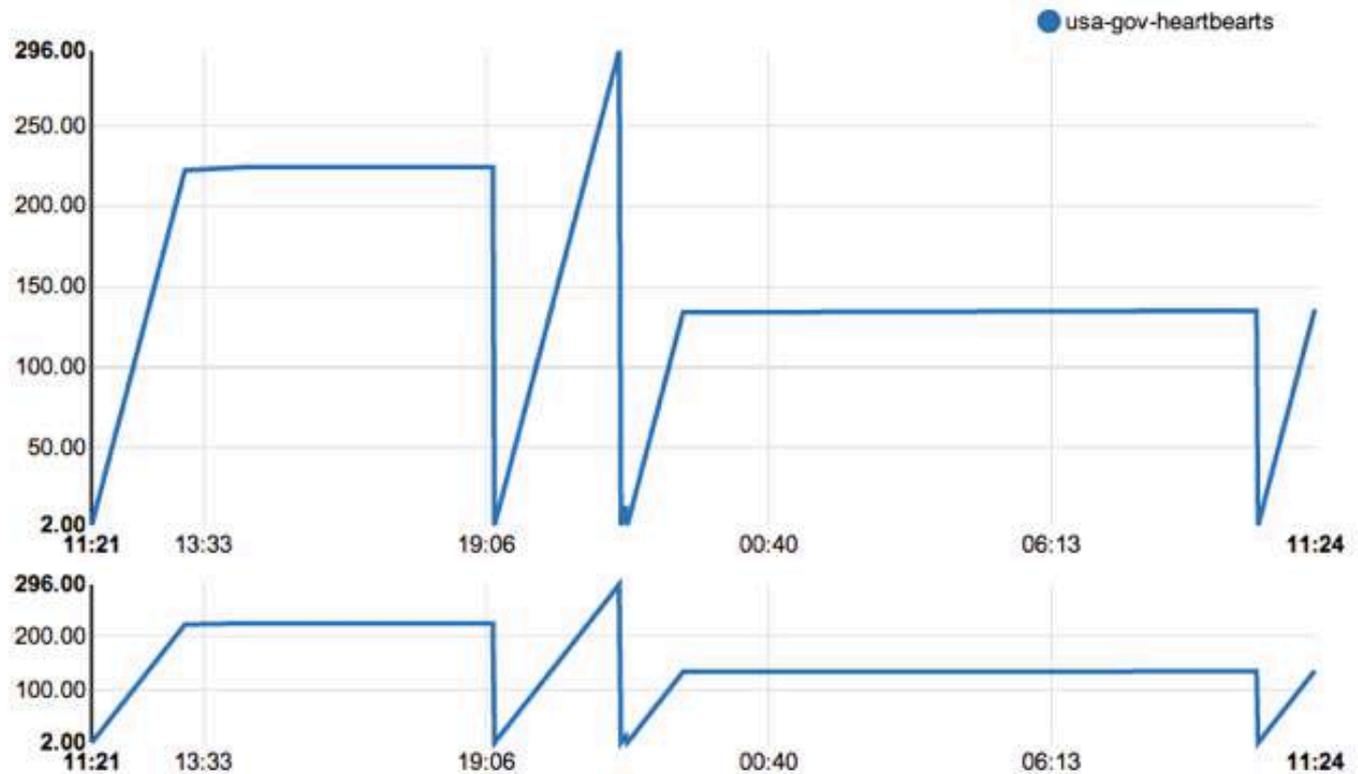
usagov-incoming-requests

m1_rate



Draw

usa-gov-
heartbeats



From **data**
to **insight**



Know it all!

- Long term data required (index everything!)
- Visualization is a great start
- Deep insight into your data required

Know your data

Know your data format

Concrete questions with lots of dimensions

Aggregations

- aka: composable facets
- Take the output of a facet operation
- Use it as an input of another facet operation

- Remember: What is the average shopping cart value per order per hour?

Aggregations

```
curl -X GET 'http://localhost:9200/orders/order/_search' -d '{
  "aggs" : {
    "avg_shopping_cart_per_hour" : {
      "filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      }
    },
    "date_histogram" : {
      "field" : "created_at",
      "interval" : "1h"
    },
    "aggregations" : {
      "avg" : { "avg" : { "field" : "total" } }
    }
  }
}'
```

Aggregations

```
curl -X GET 'http://localhost:9200/orders/order/_search' -d '{
  "aggs" : {
    "avg_shopping_cart_per_hour" : {
      "filter" : {
        "range": {
          "created_at": {
            "gte": "2013/09/01",
            "lt" : "2013/10/01"
          }
        }
      },
      "histogram" : {
        "script" : "doc[\u0027created_at\u0027].date.hourOfDay",
      },
      "aggregations" : {
        "avg" : { "avg" : { "field" : "total" } }
      }
    }
  }
}'
```

Ask complex questions

- Product pageviews

Sum of page views per price range including price statistics (min/max/avg/sum/count)

- Geo location

Physical store: Home of buyers per weekday combined with money spent

- Protip: Reduce memory consumption using probabilistic data structures, losing precision

roundup



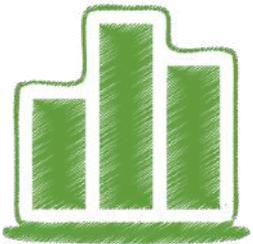
Roundup



Insight



elasticsearch.



Visualization



Notification

elasticsearch.

Thanks for listening!

We're hiring

<http://www.elasticsearch.com/about/jobs>

#gotoaar #elasticsearch

Alexander Reelsen

@spinscale

alexander.reelsen@elasticsearch.com



roadmap



Roadmap

- Elasticsearch 1.0

Distributed percolator (already in master)

Aggregations

Snapshot/Restore

links



Links

- Elasticsearch

<http://www.elasticsearch.org>

- Logstash

<http://logstash.net>

- Kibana

<http://three.kibana.org>

- elasticsearch-metrics-reporter

<https://github.com/elasticsearch/metrics-elasticsearch-reporter-java>

Links

- Clients

<http://www.elasticsearch.org/blog/unleash-the-clients-ruby-python-php-perl/>

- Metrics

<http://metrics.codahale.com/>

- Aggregations

<https://github.com/elasticsearch/elasticsearch/issues/3300>

- Elasticsearch Hadoop integration

<https://github.com/elasticsearch/elasticsearch-hadoop>

Links

- Talk on probabilistic data structures

<http://www.infoq.com/presentations/scalability-data-mining>

- Icons

<http://www.doublejdesign.co.uk/>

<http://www.iconarchive.com/>