Elasticsearch in Ecommerce

Alexander Reelsen
@spinscale
alexander.reelsen@elasticsearch.com

about

- Me

  Interested in metrics, ops and the web
  Likes the JVM
  Working with elasticsearch since 2011

- Elasticsearch, founded in 2012

  Products: Elasticsearch, Logstash, Kibana
  Professional services: Support & development subscriptions
  Trainings

Agenda

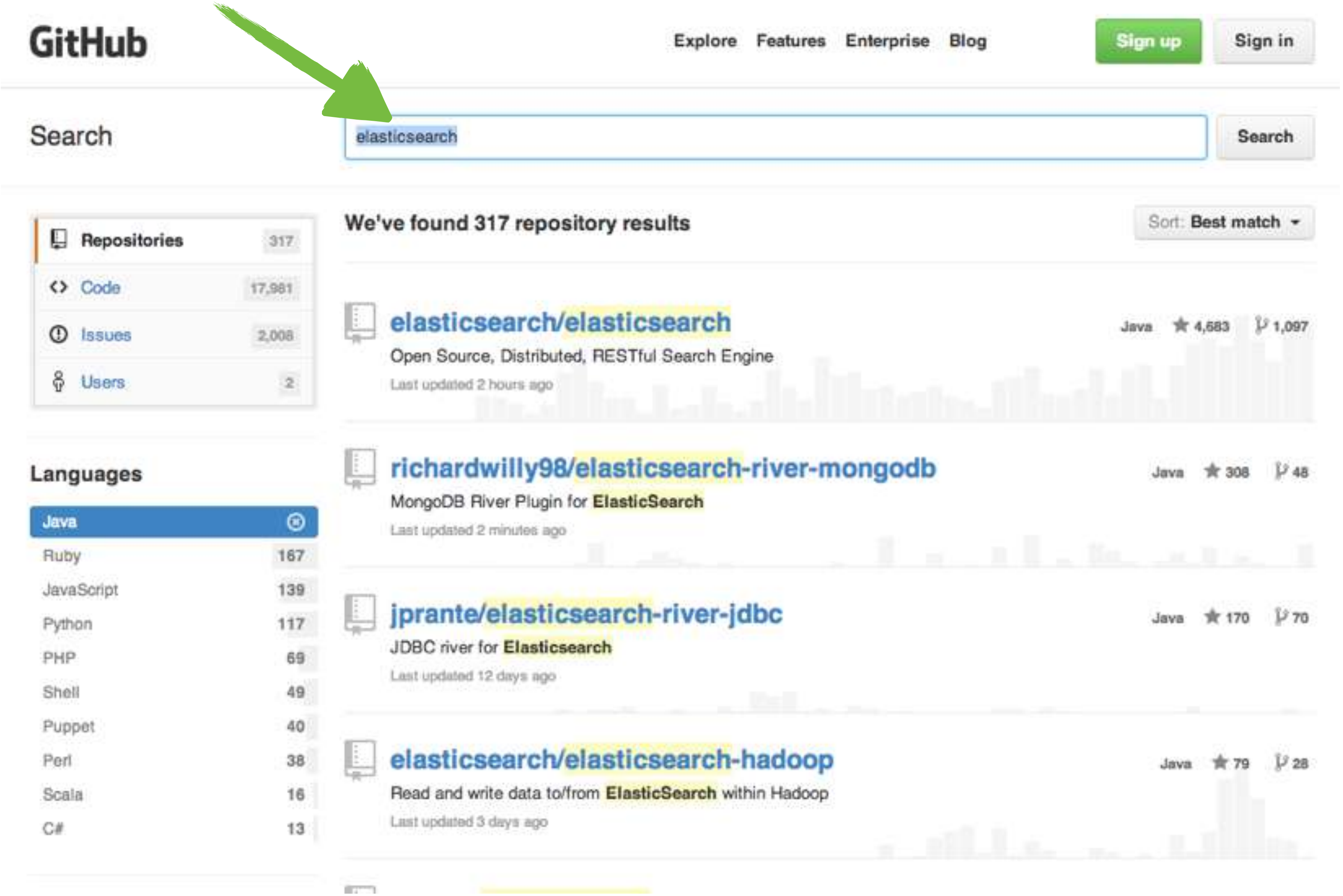
- Introduction
- Ecommerce Use-Cases

  Product/Full-text search
  Logfiles
  Analytics

- Elasticsearch 1.0
- Q & A

Introduction

Unstructured search



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

Languages

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

We've found 317 repository results

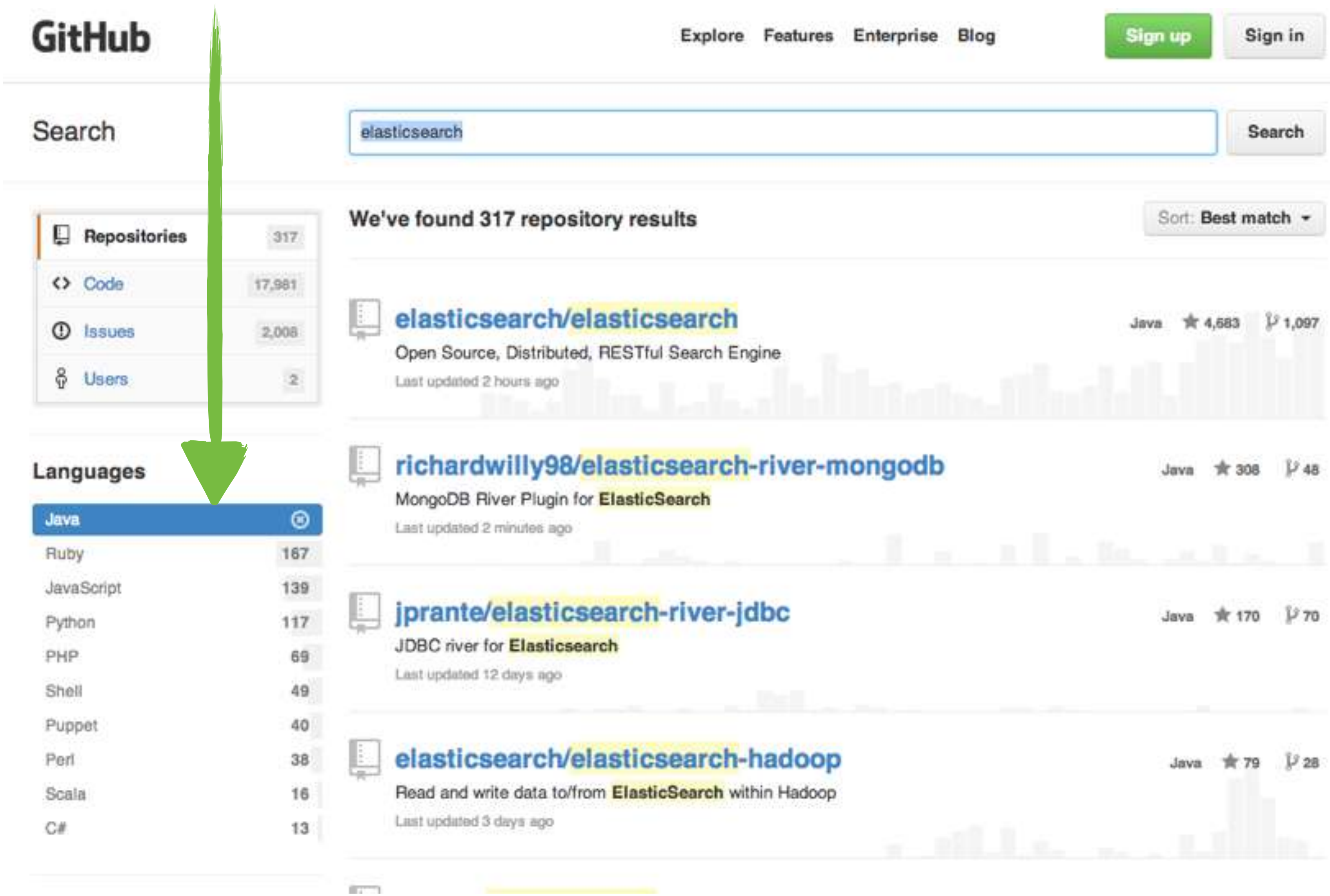
elasticsearch/elasticsearch Java ★ 4,683 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago

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

Structured search



elasticsearch

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

We've found 317 repository results

Languages

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

elasticsearch/elasticsearch — Java ★ 4,683 ⑂ 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago

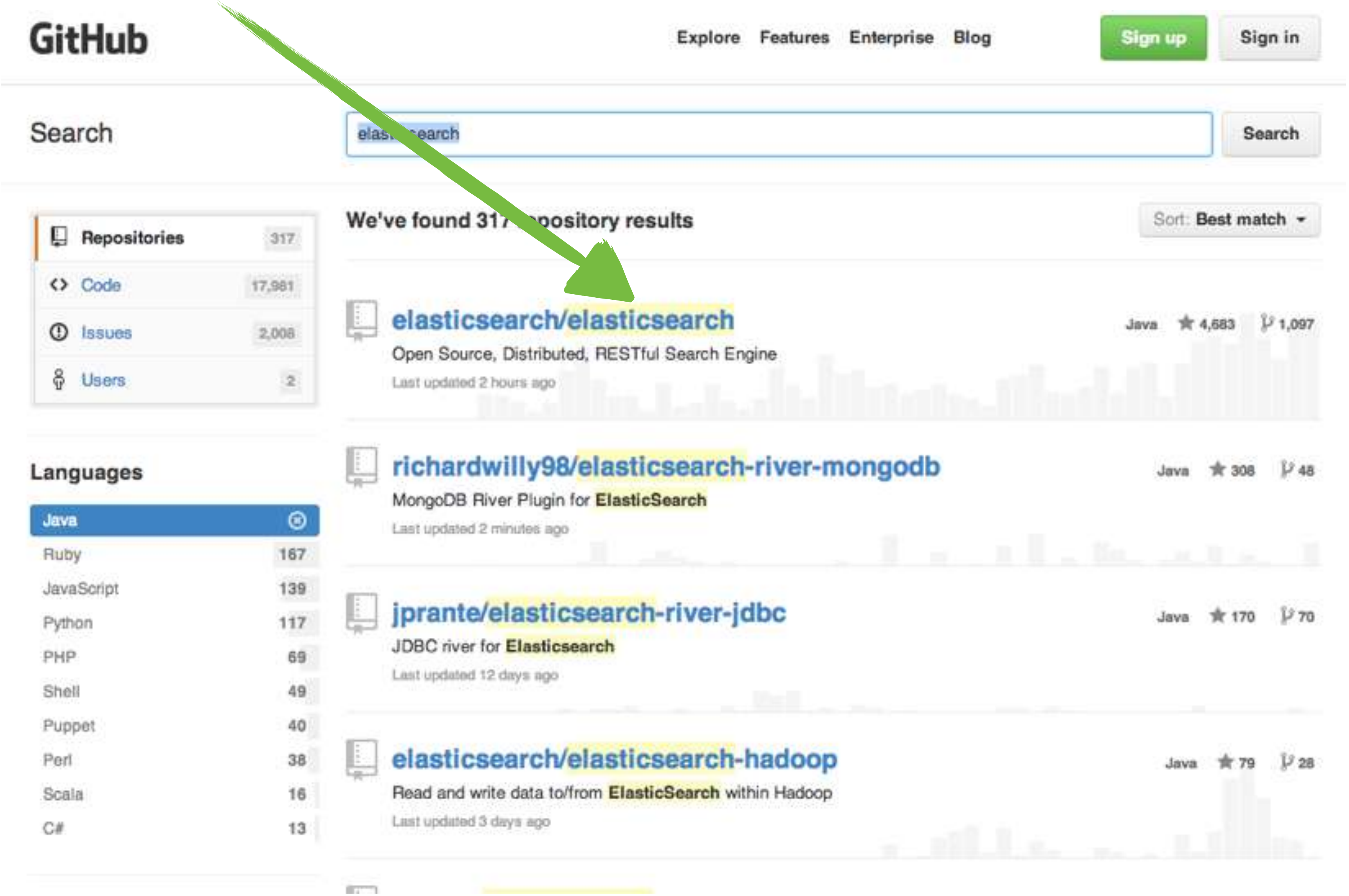
richardwilly98/elasticsearch-river-mongodb — Java ★ 308 ⑂ 48
MongoDB River Plugin for ElasticSearch
Last updated 2 minutes ago

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.

Enrichment



elasticsearch

We've found 317 repository results

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

Languages

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

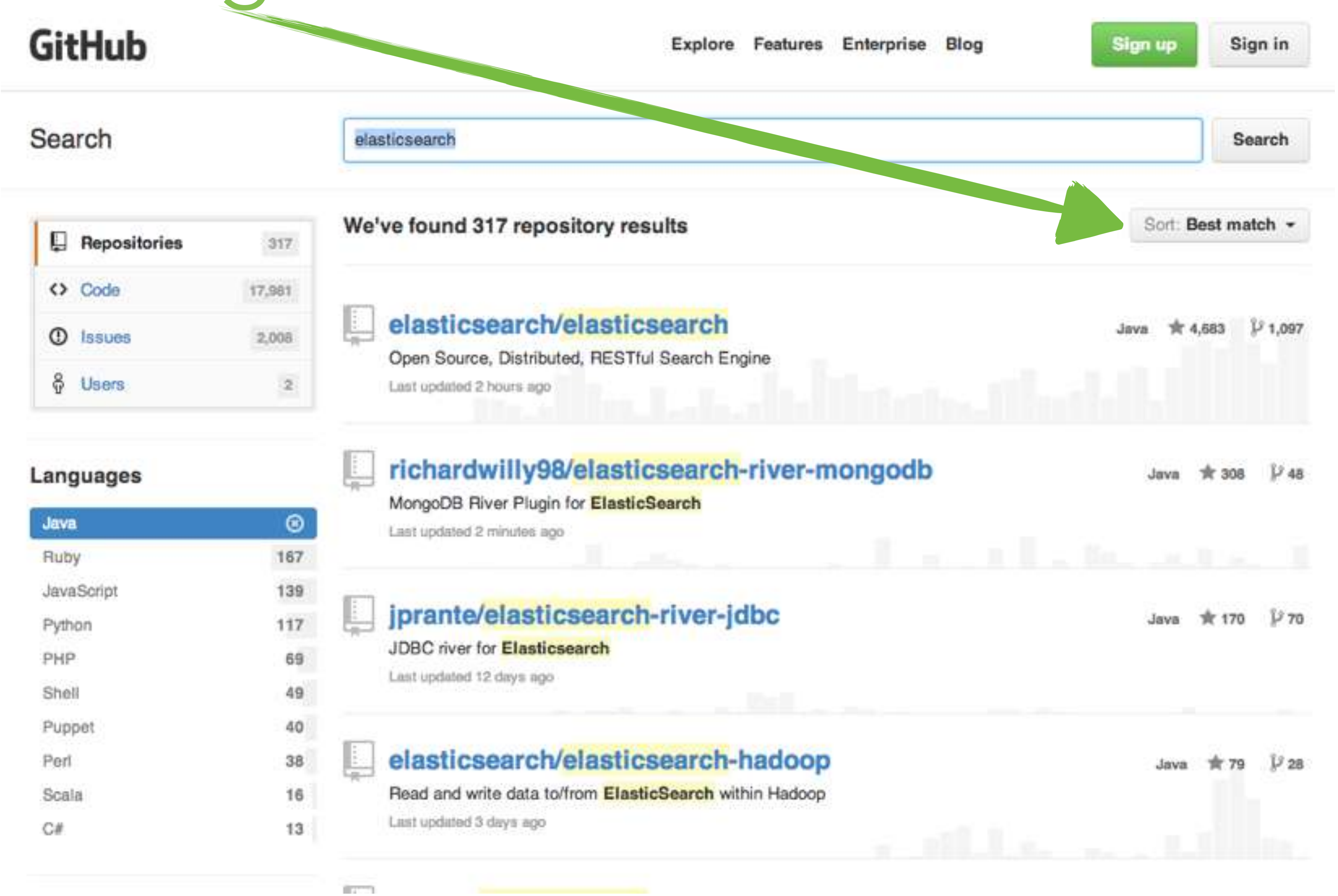
elasticsearch/elasticsearch — Java ★ 4,683 ⑂ 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago

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

Sorting



GitHub

Search

elasticsearch

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

Languages

Java	
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

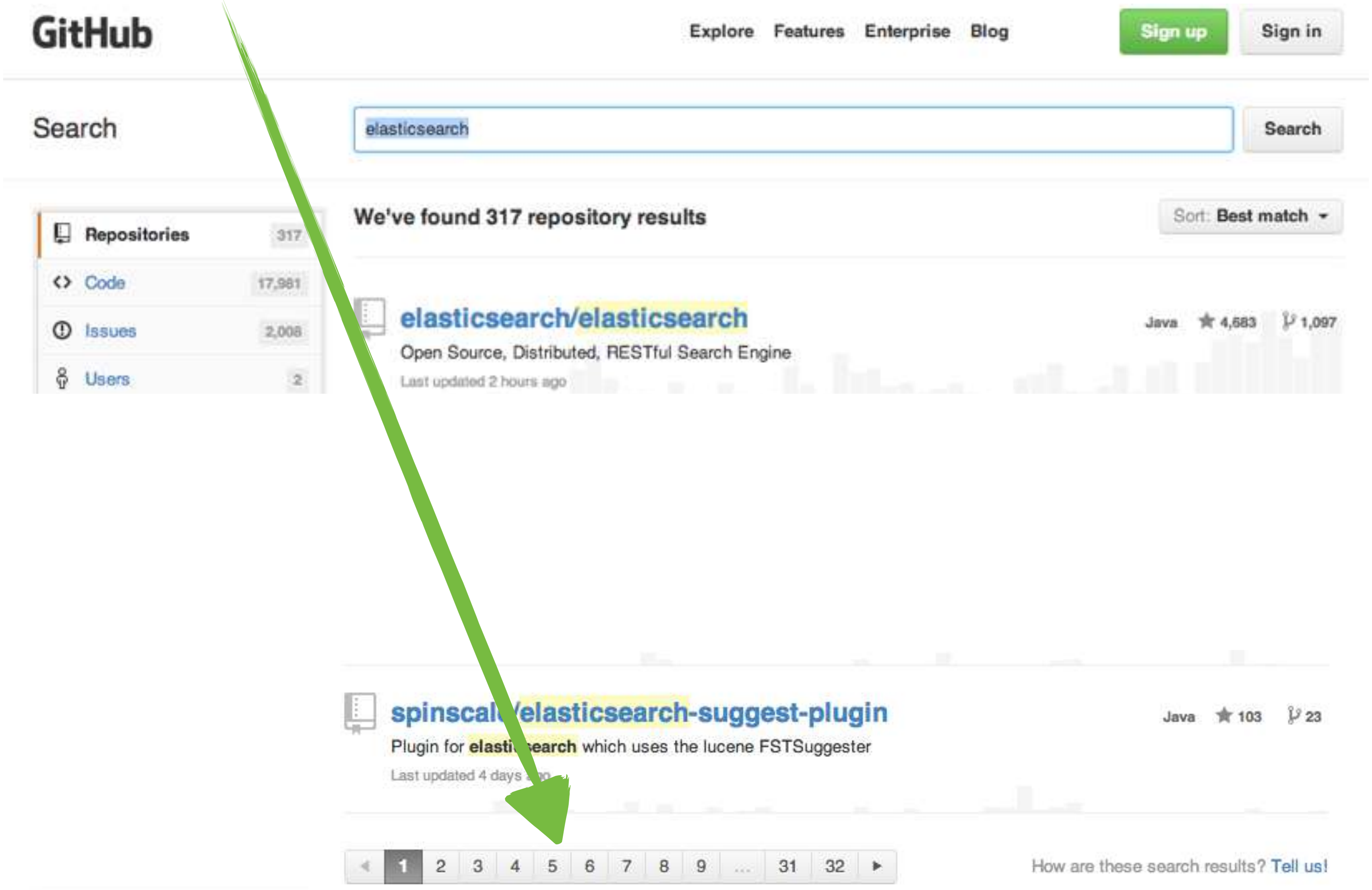
elasticsearch/elasticsearch — Java ★ 4,683 ⑂ 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago

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

Pagination



We've found 317 repository results

elasticsearch/elasticsearch — Java ★ 4,683 ⑂ 1,097

Open Source, Distributed, RESTful Search Engine

Last updated 2 hours ago

spinscale/elasticsearch-suggest-plugin — Java ★ 103 ⑂ 23

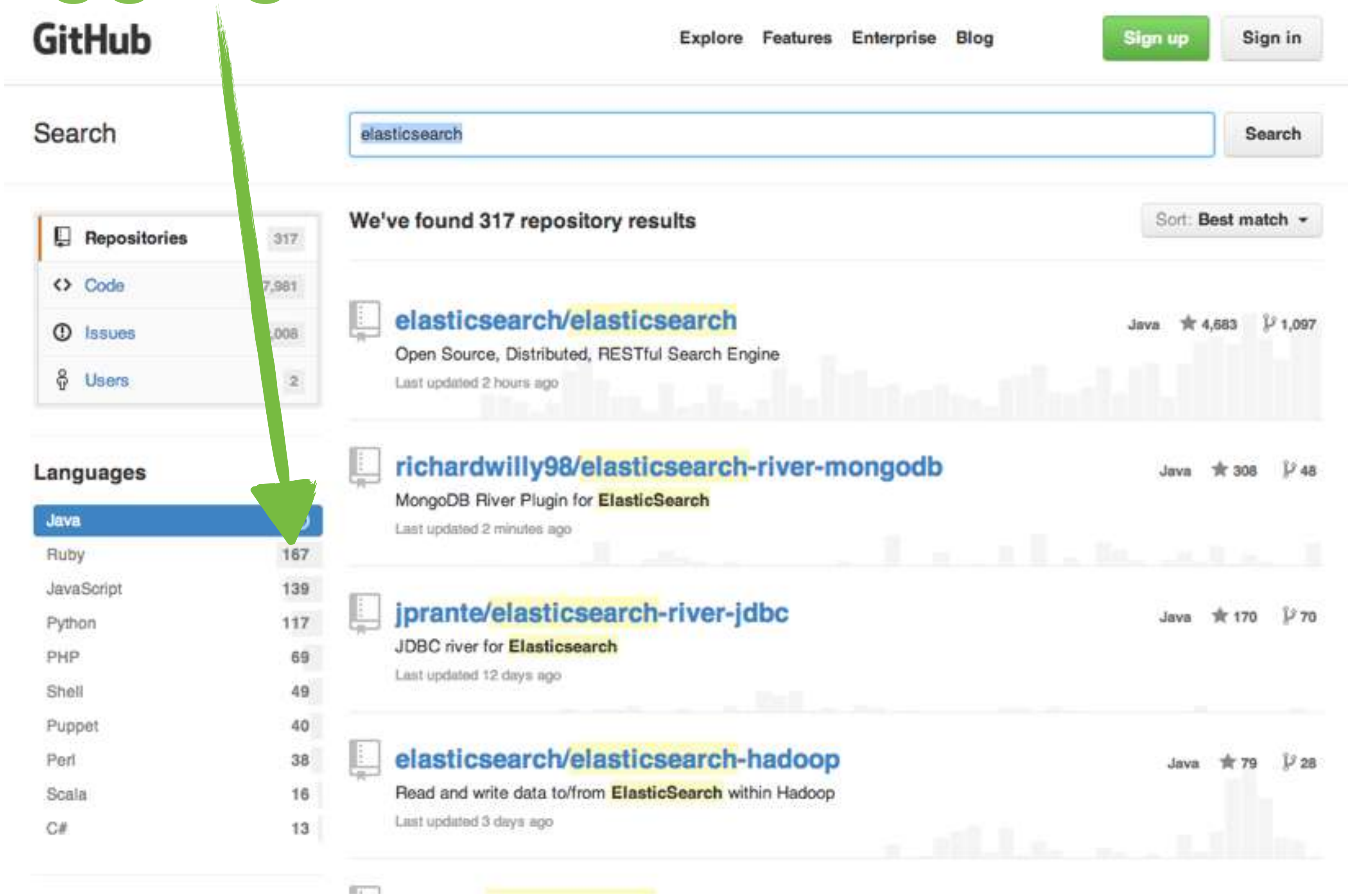
Plugin for elasticsearch which uses the lucene FSTSuggester

Last updated 4 days ago

How are these search results? Tell us!

elasticsearch.

Aggregation



GitHub

Explore Features Enterprise Blog

Sign up Sign in

Search

elasticsearch

Search

Repositories	317
Code	7,981
Issues	,008
Users	2

Languages

Java	
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

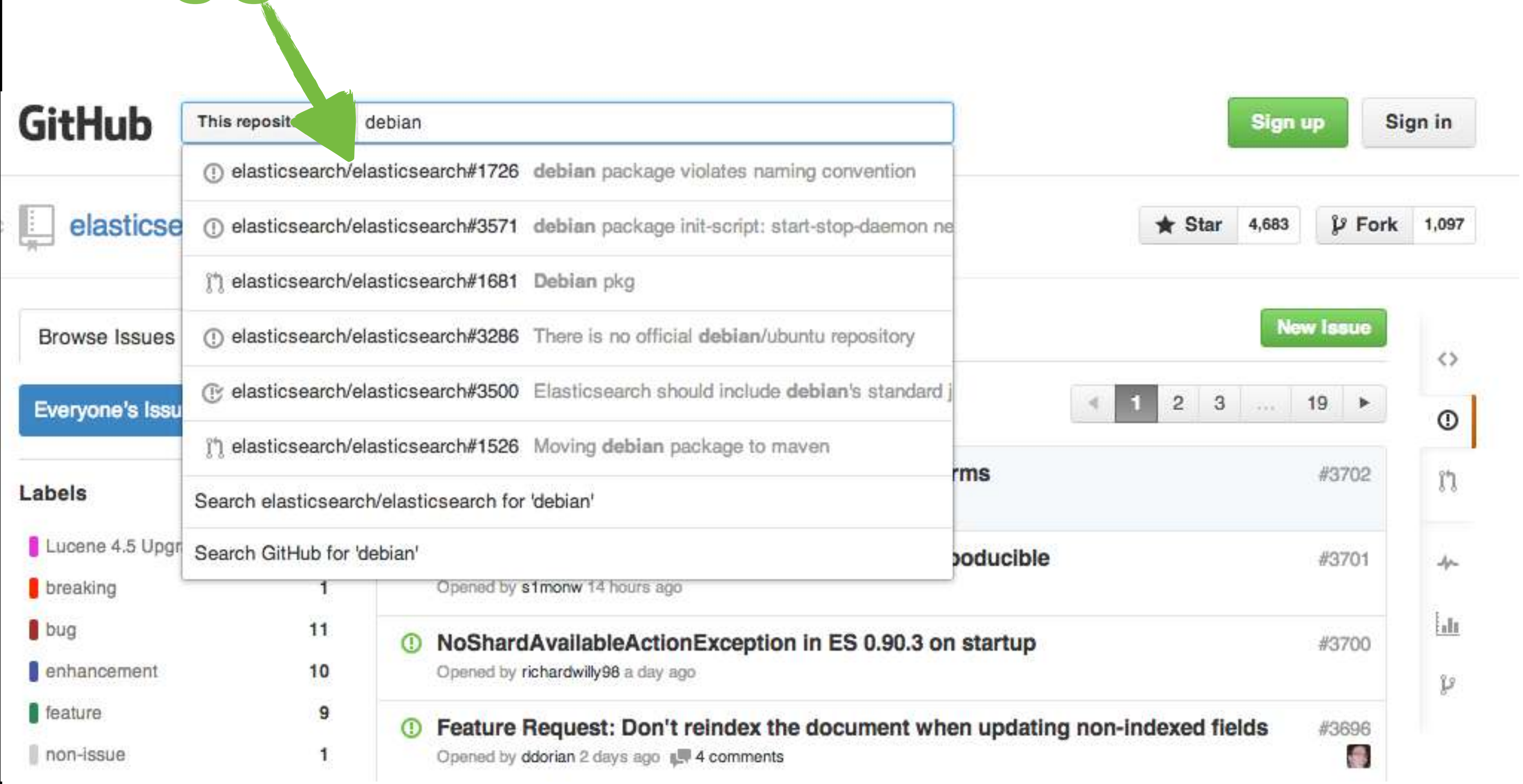
elasticsearch/elasticsearch Java ★ 4,683 1,097
Open Source, Distributed, RESTful Search Engine
Last updated 2 hours ago

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

Suggestions



elasticsearch/elasticsearch#1726 debian package violates naming convention

elasticsearch/elasticsearch#3571 debian package init-script: start-stop-daemon ne

elasticsearch/elasticsearch#1681 Debian pkg

elasticsearch/elasticsearch#3286 There is no official debian/ubuntu repository

elasticsearch/elasticsearch#3500 Elasticsearch should include debian's standard j

elasticsearch/elasticsearch#1526 Moving debian package to maven

Search elasticsearch/elasticsearch for 'debian'

Search GitHub for 'debian'

Lucene 4.5 Upgr	
breaking	1
bug	11
enhancement	10
feature	9
non-issue	1

rms #3702

oducible #3701
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

elasticsearch.

Elasticsearch in 10 seconds

- Schema-free, REST & JSON based distributed document store
- Open Source: Apache License 2.0
- Zero configuration
- Written in Java, extensible

XING
stackoverflow
SOUNDCLOUD
Mc Graw Hill
Fog Creek SOFTWARE
foursquare
LiveChat.com
loggly
NETWORKED INSIGHTS

Zero configuration

```
$ wget https://download.elasticsearch.org/...
$ tar -xf elasticsearch-1.0.0.RC1.tar.gz
$ ./elasticsearch-1.0.0.RC1/bin/elasticsearch -f
...
[2014-01-19 14:53:11,508][INFO ][node] [Scanner] started
...
```

Is it alive?

```
» curl localhost:9200
{
  "status" : 200,
  "name" : "Scanner",
  "version" : {
    "number" : "1.0.0",
    "build_hash" : "e018cda7e7a32643d59e0ac3cdb412ccc239af04",
    "build_timestamp" : "2014-01-17T15:11:47Z",
    "build_snapshot" : true,
    "lucene_version" : "4.6"
  },
  "tagline" : "You Know, for Search"
}
```

Create...

```
» curl -XPUT localhost:9200/books/book/1 -d '
{
  "title" : "Elasticsearch - The definitive guide",
  "authors" : "Clinton Gormley",
  "started" : "2013-02-04",
  "pages" : 230
}'
```

Update...

```
» curl -XPUT localhost:9200/books/book/1 -d '
{
  "title" : "Elasticsearch - The definitive guide",
  "authors" : [ "Clinton Gormley", "Zachary Tong" ],
  "started" : "2013-02-04",
  "pages" : 230
}'
```

Delete…

```
» curl –X DELETE localhost:9200/books/book/1
```

Realtime GET…

```
» curl –X GET localhost:9200/books/book/1
» curl –X GET localhost:9200/books/book/1/_source
```

Search

```
» curl -XGET localhost:9200/books/_search?q=elasticsearch
```

```
{
  "took" : 2, "timed_out" : false,
  "_shards" : { "total" : 5, "successful" : 5, "failed" : 0 },
  "hits" : {
    "total" : 1, "max_score" : 0.076713204,
    "hits" : [ {
      "_index" : “books", "_type" : “book", "_id" : "1",
      "_score" : 0.076713204, "_source" : {
        "title" : "Elasticsearch - The definitive guide",
        "authors" : [ "Clinton Gormley", "Zachary Tong" ],
        "started" : “2013-02-04", "pages" : 230
      }
    } ]
  }
}
```

Search - Query DSL

```
» curl -XGET ‘localhost:9200/books/book/_search' -d '{
    "query": {
        "filtered" : {
            "query" : {
                "match": {
                    "text" :  {
                        "query" : “To Be Or Not To Be",
                        "cutoff_frequency" : 0.01
                    }
                }
            },
            "filter" : {
                "range": {
                  "price": {
                    "gte": 20.0
                    "lte": 50.0
       ...
    }
}'
```

Distributed & scalable

- Replication

  Read scalability

  Removing SPOF

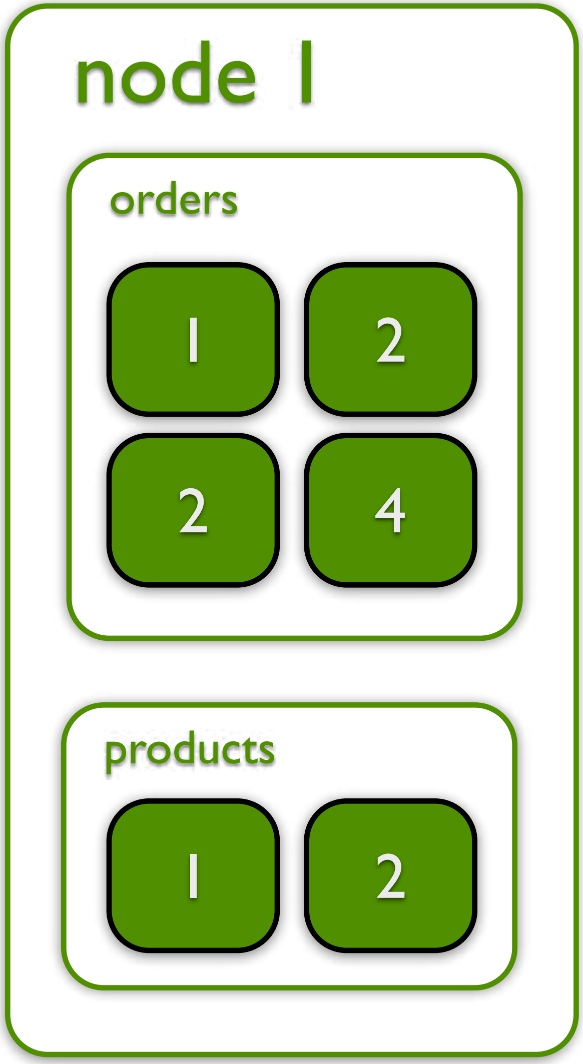
- Sharding

  Split logical data over several machines

  Write scalability

  Control data flows

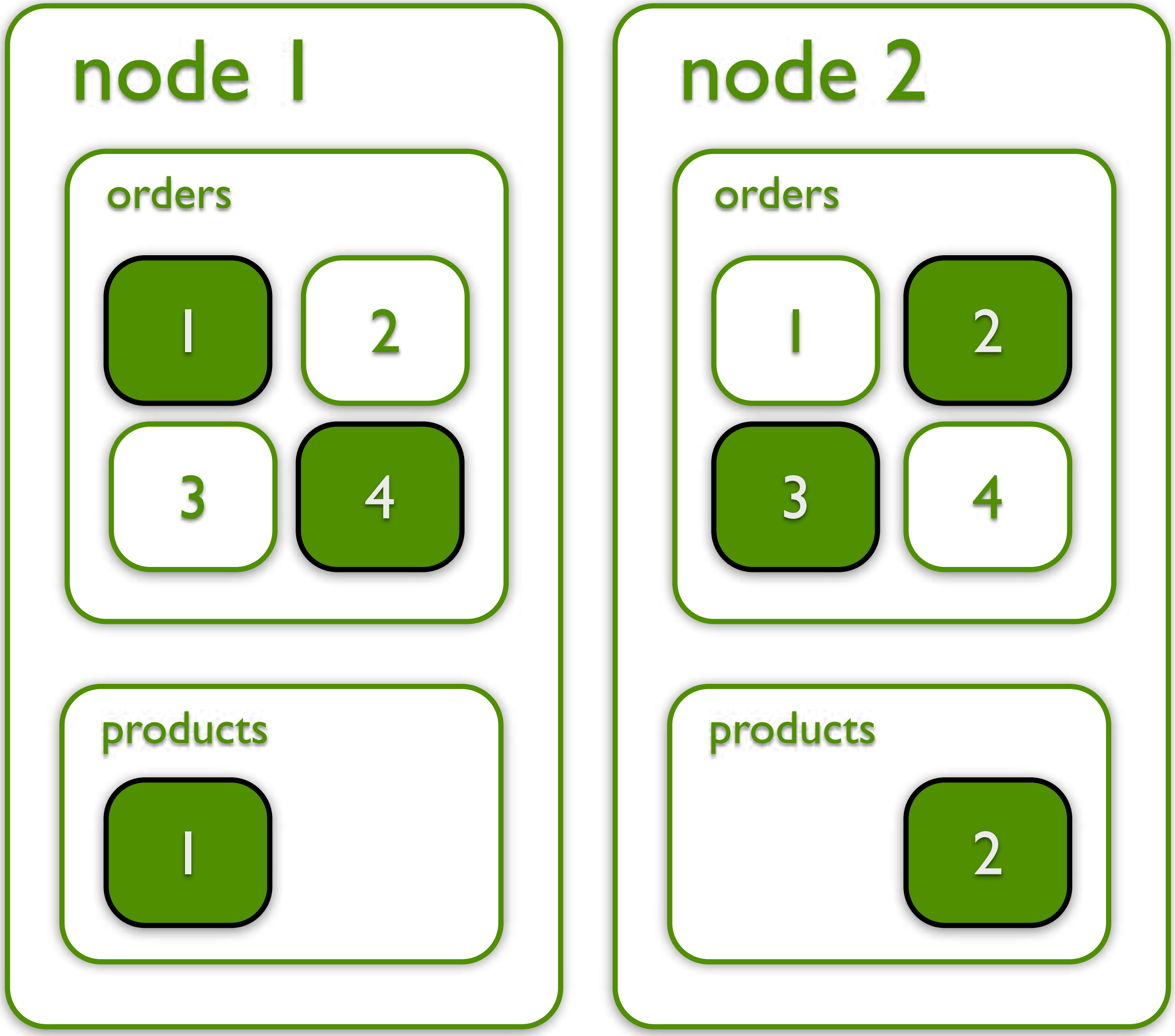
Distributed & scalable



```
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 and scalable



node 1

orders

1 2
3 4

products

1

node 2

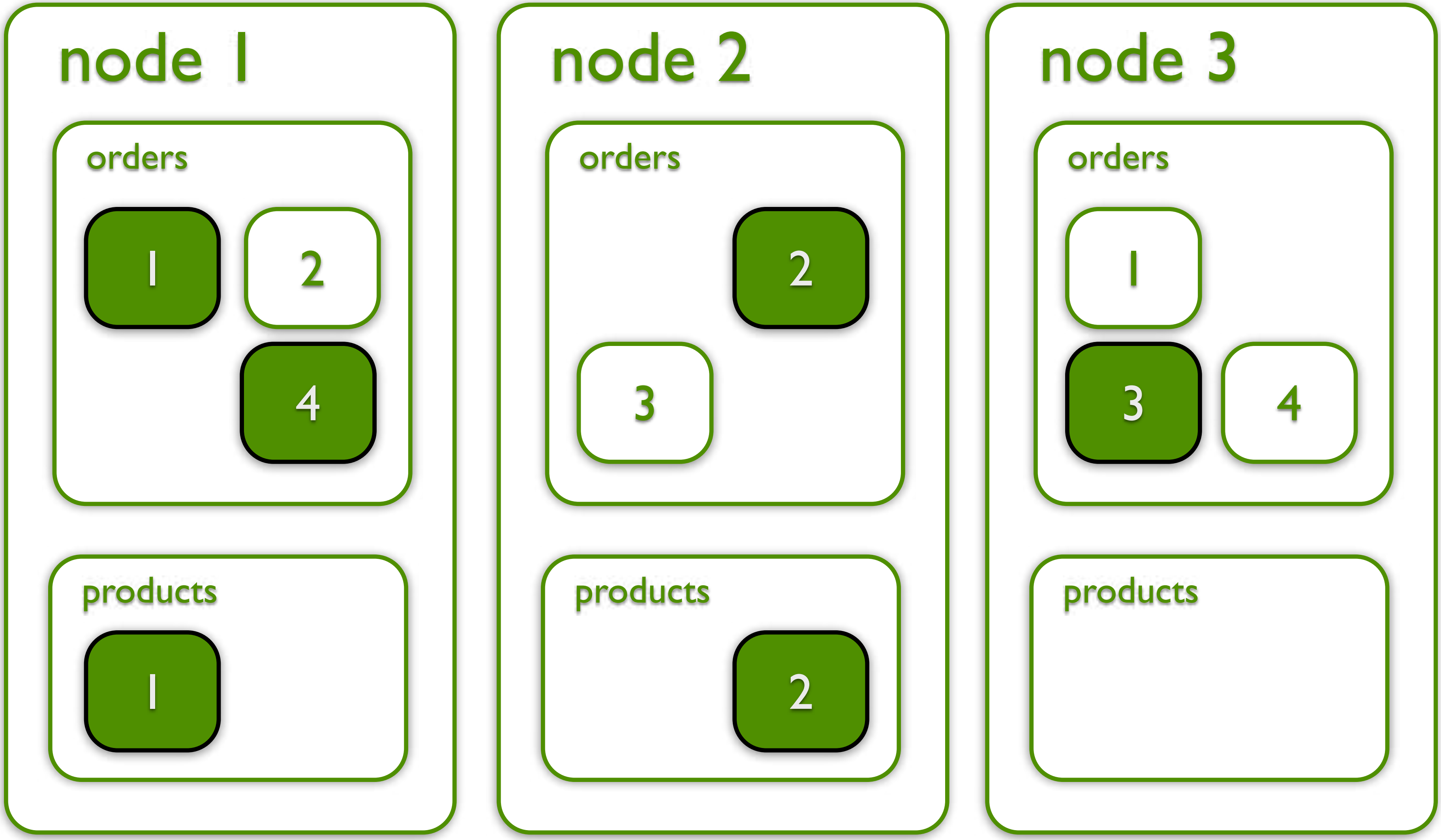
orders

1 2
3 4

products

2

Distributed & scalable

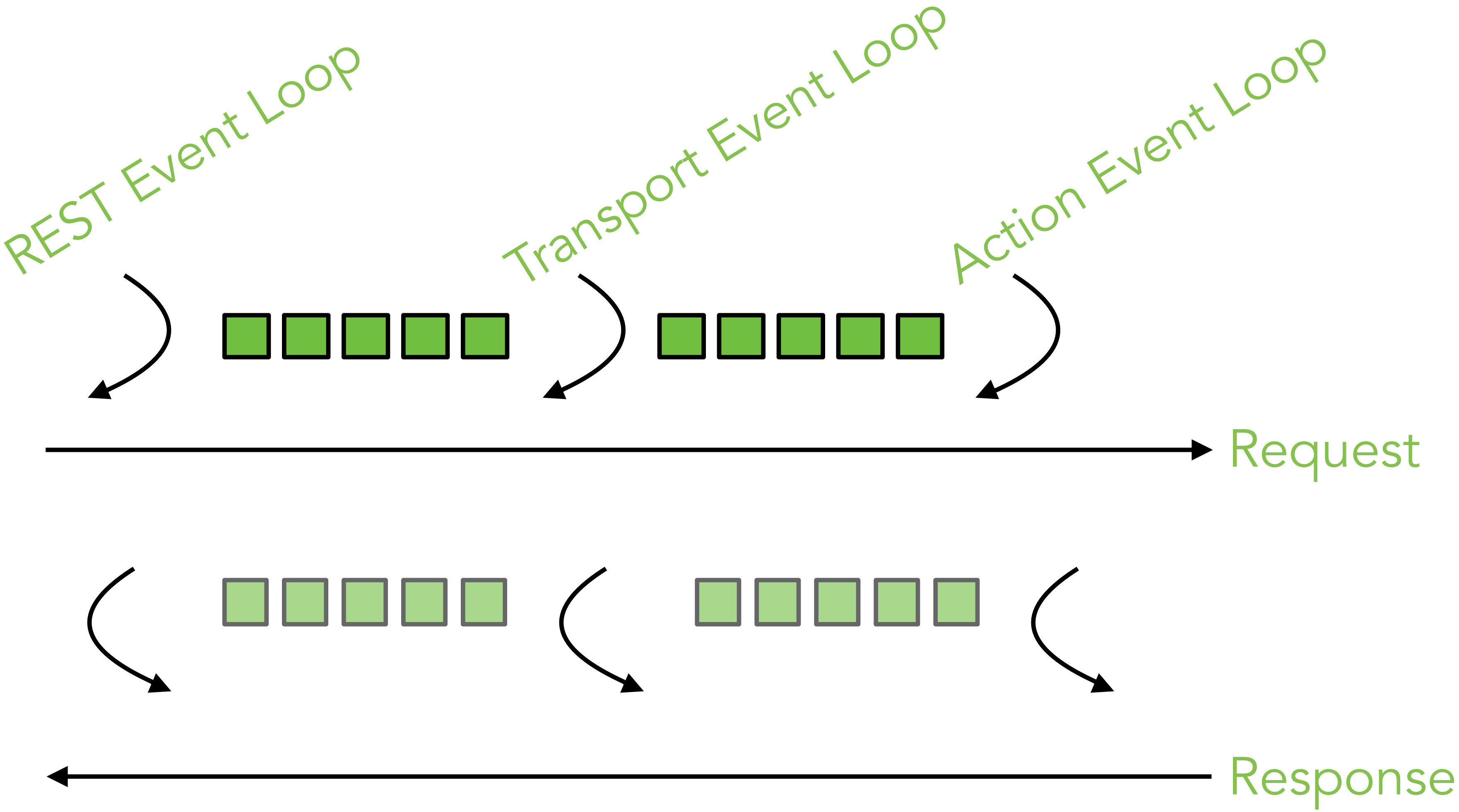


elasticsearch.

Distributed & scalable

- JVM (high level & high performance if done right)
- Netty (async networking on top of the JVM)
- Lucene (fulltext search library)
- HPPC (high performance primitive collections)
- Google Guice (for extension & dependencies)

A request under the hood



elasticsearch.

Think async!

- Enforces event driven architecture
- Support for non-blocking model
- Enforce loose coupling
- Prefers push over pull
- Callback based concurrency
- Helps to avoid contention on resources / threads

Ecosystem

- Plugins
- Clients for many languages

  Ruby, python, php, perl, javascript, (.NET coming)
  Scala, clojure, go
- Kibana
- Logstash
- Hadoop integration

Use-case:
Product search engine

Product search engine

- Just index all your products and be happy?

  Search is not that easy

- Gathered experience at an b2b ecommerce platform in the hotel and gastronomy sector

  First solution was self written using bobo/zoie turned out to be unmaintainable

  Switched to elasticsearch then

- Decompounding, Suggestions, Faceting, Custom scoring, Analytics, Price agents, Query optimization, beyond search

Domain specific knowledge

- Search term: Topf

  What is expected? Blumentopf? Kochtopf?
  Or: Tuch (Handtuch, Halstuch, Geschirrtuch)
  Or: Decke (Tischdecke, Löschdecke, Mitteldecke)

- Decompounding (compound word token filter)

  Blumentopf also needs to match Leuchtblumentopf

- Synonyms

  Portmonee/Portemonnaie/Geldbörse

Neutrality? Really?

- Is full-text search relevancy really your preferred scoring algorithm?

- Possible influential factors

  Age of the product, been ordered in last 24h
  On stock?
  Provision
  No shipping costs
  Special offer
  Rating (product or seller)

  http://www.elasticsearch.org/guide/en/elasticsearch/reference/current/query-dsl-function-score-query.html

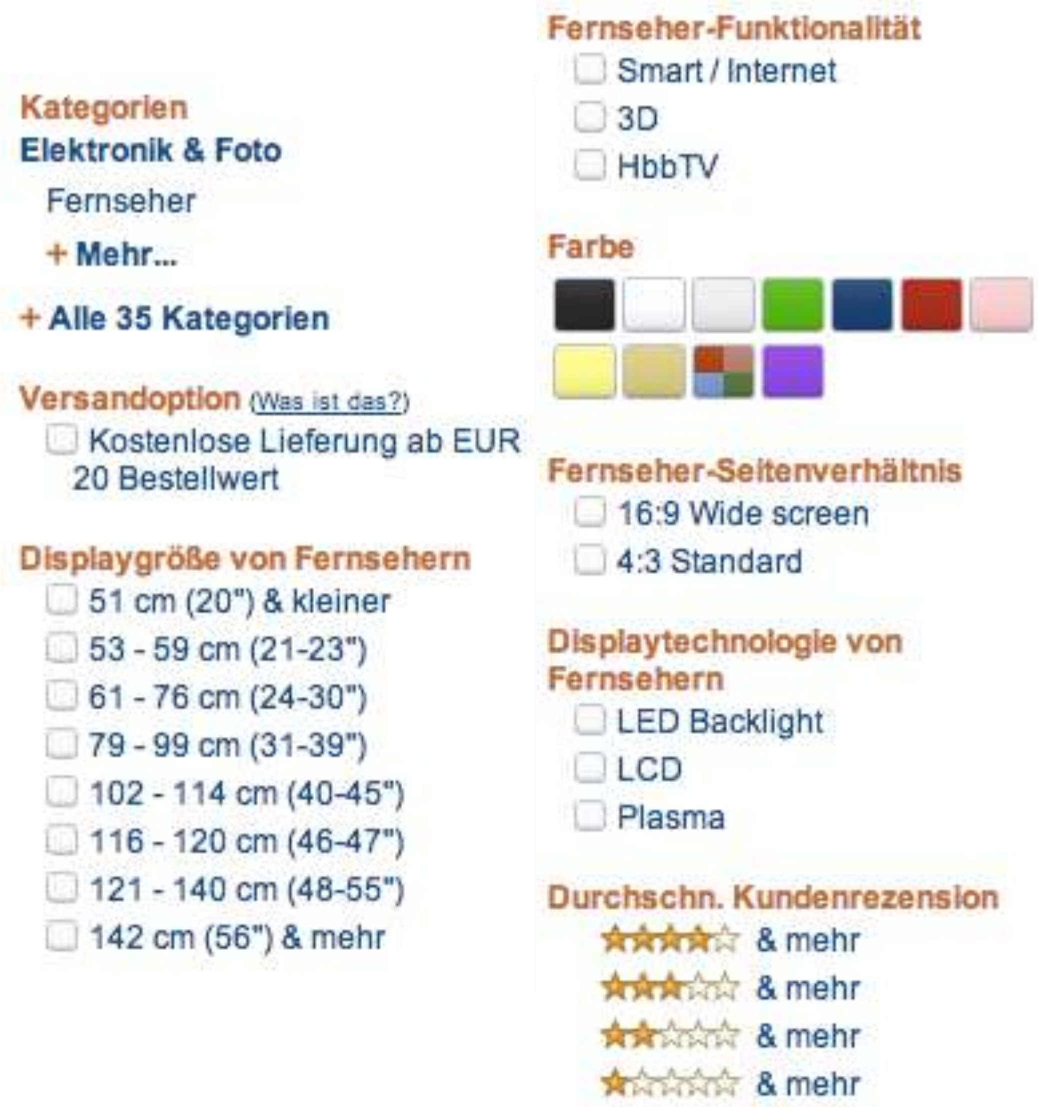
Faceting & Filtering

- Products grouped by

  Category
  Material
  Brand

- Allowing to filter

  All of the facets
  Price range
  Color
  Seller
  Ratings (hard!)



Kategorien
Elektronik & Foto
Fernseher
+ **Mehr...**
+ **Alle 35 Kategorien**

Versandoption (Was ist das?)
□ Kostenlose Lieferung ab EUR 20 Bestellwert

Displaygröße von Fernsehern
□ 51 cm (20") & kleiner
□ 53 - 59 cm (21-23")
□ 61 - 76 cm (24-30")
□ 79 - 99 cm (31-39")
□ 102 - 114 cm (40-45")
□ 116 - 120 cm (46-47")
□ 121 - 140 cm (48-55")
□ 142 cm (56") & mehr

Fernseher-Funktionalität
□ Smart / Internet
□ 3D
□ HbbTV

Farbe
⬛ ⬜ ⬜ 🟩 🟦 🟥 🟥
🟨 🟨 ▨ 🟪

Fernseher-Seitenverhältnis
□ 16:9 Wide screen
□ 4:3 Standard

Displaytechnologie von Fernsehern
□ LED Backlight
□ LCD
□ Plasma

Durchschn. Kundenrezension
★★★★☆ & mehr
★★★☆☆ & mehr
★★☆☆☆ & mehr
★☆☆☆☆ & mehr

Product variants?

- How to handle product variants?

  Same product by the same merchant
  Same product in different sizes, colours (clothing)

- Solution: Patched elasticsearch with grouping support, which was done by creating an image hash from the image and grouping on it

- Unsolved: Same product by different merchant

  Unless the exact same image is used, unlikely

- Better solution: Parent/child support

Notification with Percolation

- Customer: If a product matches name *X* and costs below price *Y*, is color *Z*, then I want to get a mail

  More likely: Notify customer, when it is back on stock

- Enter percolation!

  Not: Index a document and fire a query

  But: Index a query and check a document against if it matches

https://speakerdeck.com/javanna/whats-new-in-percolator

More than pure search

- Users (ab)use the search bar for everything

  Imprint, Careers, Jobs, special offer
  Requires a special component between web app and search which redirects for special search terms to landing pages

- Analytics

  Save all your queries, and analyze
  Most searched terms
  Most searched terms with zero results
  Searched terms, which lead to an add-to-cart action
  Searched terms, which lead to complete abort

Beware: Data quality

- Data quality can kill all your search improvements in no time

  Tough bet, if you rely on external products

  Will require you to have an own ETL pipeline, before the data goes into search or your platform (hard!)

- Less products, but more enriched results in more relevant searches

- Tough in a multi merchant environment in a non IT driven industry with lots of small businesses

Use-case:
Log file analysis

Enter logstash

- Managing events and logs
- Collect data
- Parse data
- Enrich data
- Store data (search and visualizing)

Enter logstash

- Managing events and logs
- Collect data

} Input

- Parse data
- Enrich data

} Filter

- Store data (search and visualizing)

} Output

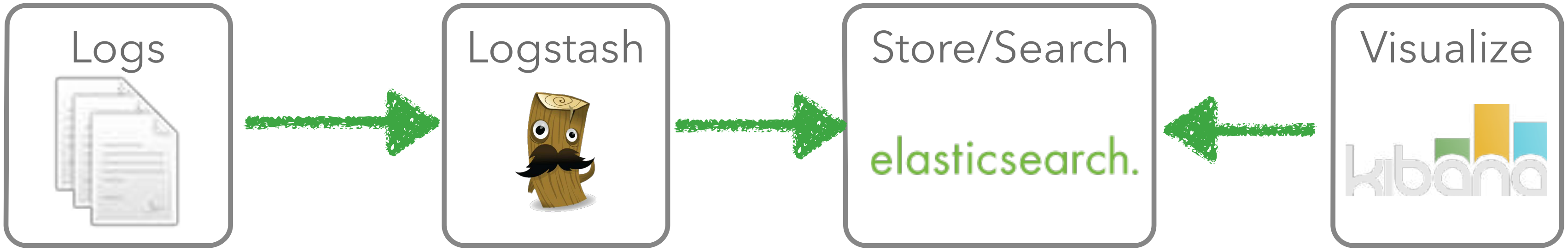
Data pipeline

- Use a shipper to get your logfiles from all hosts to logstash or a broker (redis, rabbitmq, flume)
- Run data through logstash data pipeline for enrichment
- Store data in elasticsearch
- Use kibana for dashboards and visualisation

Parsing and enrichment?

- Add geo information about an IP
- Parse multi-line exceptions from a java application
- Use grok to have tons of predefined regexes
- Metrics for event throughput information
- HTTP User-Agent extraction
- Enrichment by range values of a field

Use case: Log files



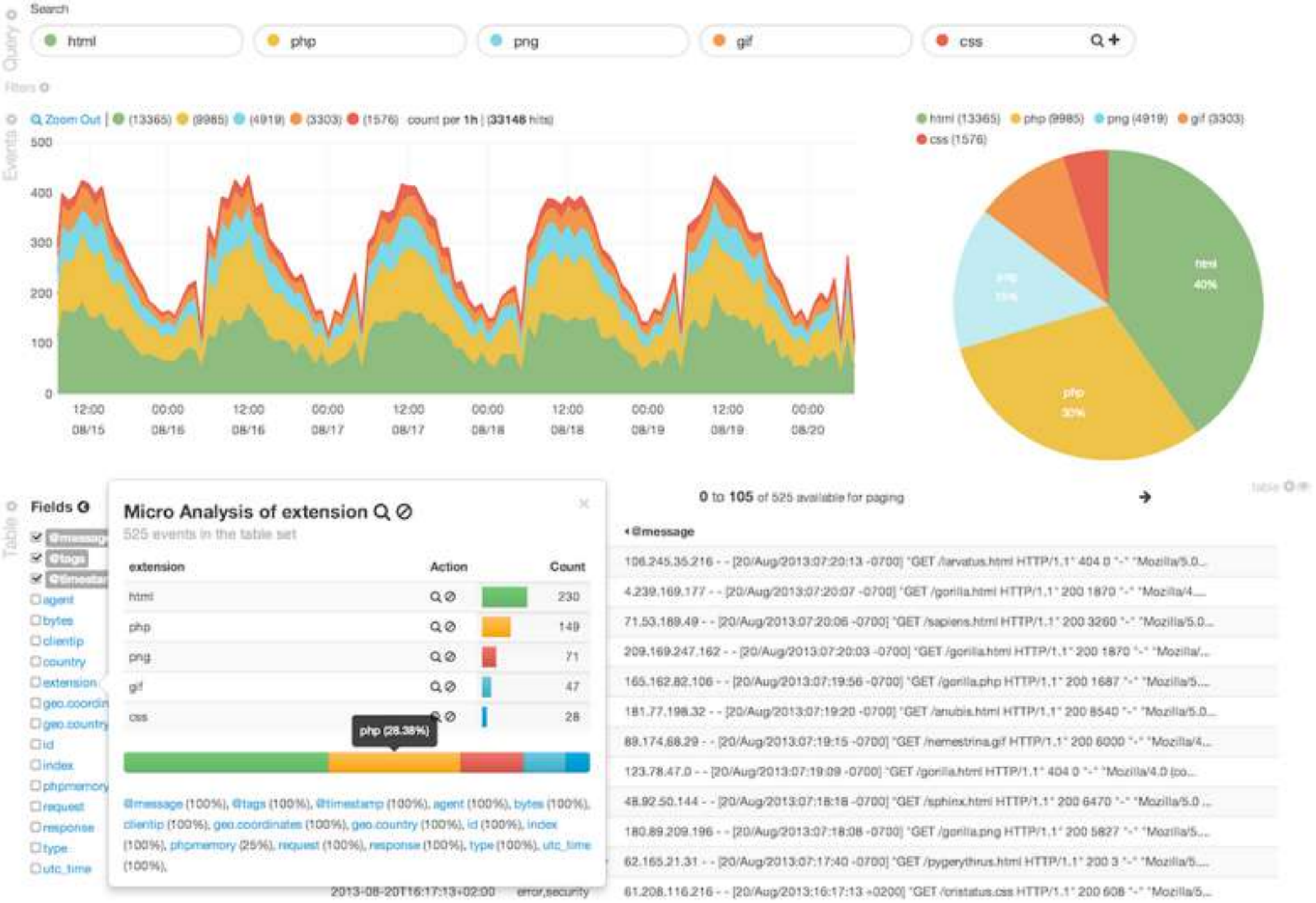
Logs

Logstash

Store/Search

Visualize

Kibana



Micro Analysis of extension

525 events in the table set

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

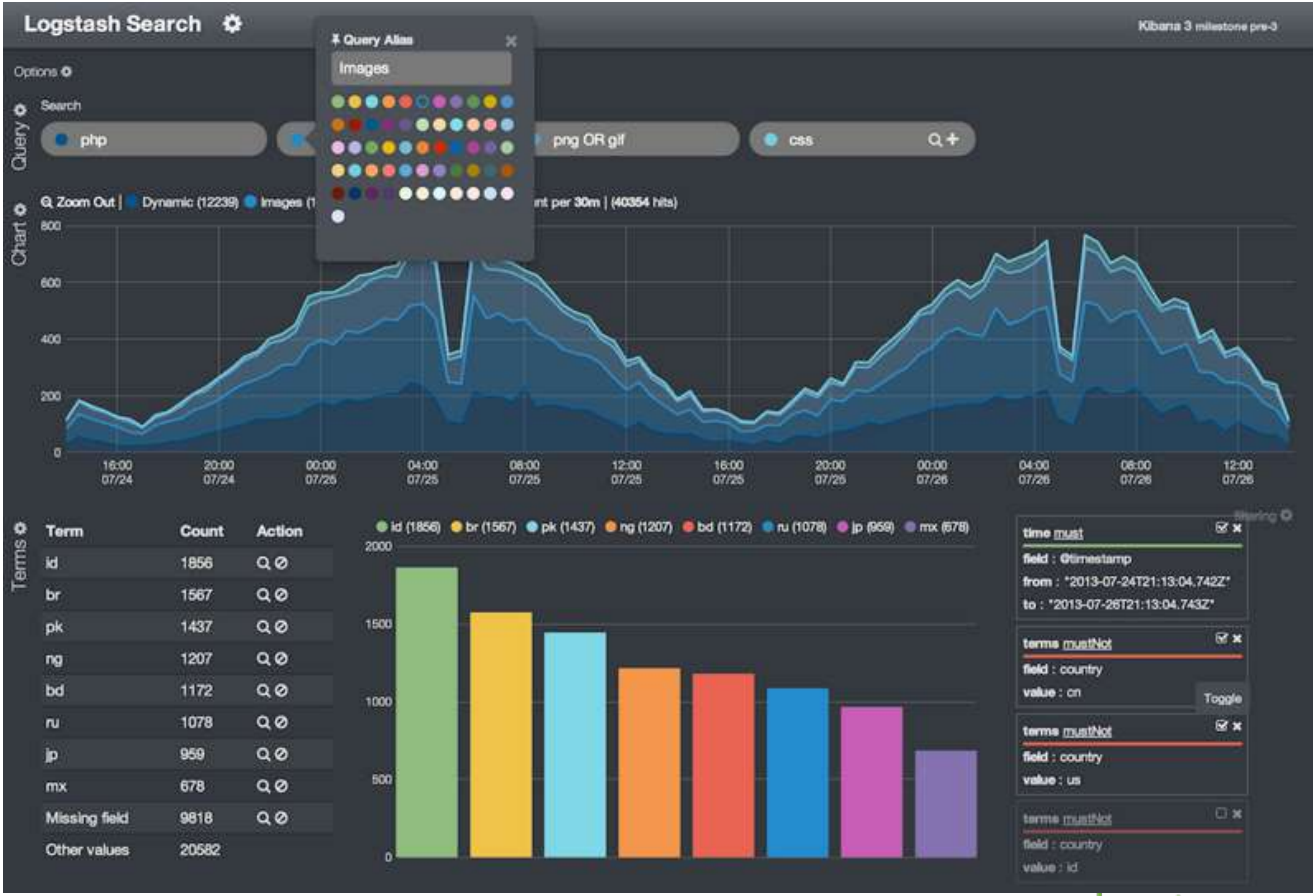
@message (100%), @tags (100%), @timestamp (100%), agent (100%), bytes (100%), clientip (100%), geo.coordinates (100%), geo.country (100%), id (100%), index (100%), phpmemory (25%), request (100%), response (100%), type (100%), utc_time (100%),

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 /arvatus.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...
```

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	

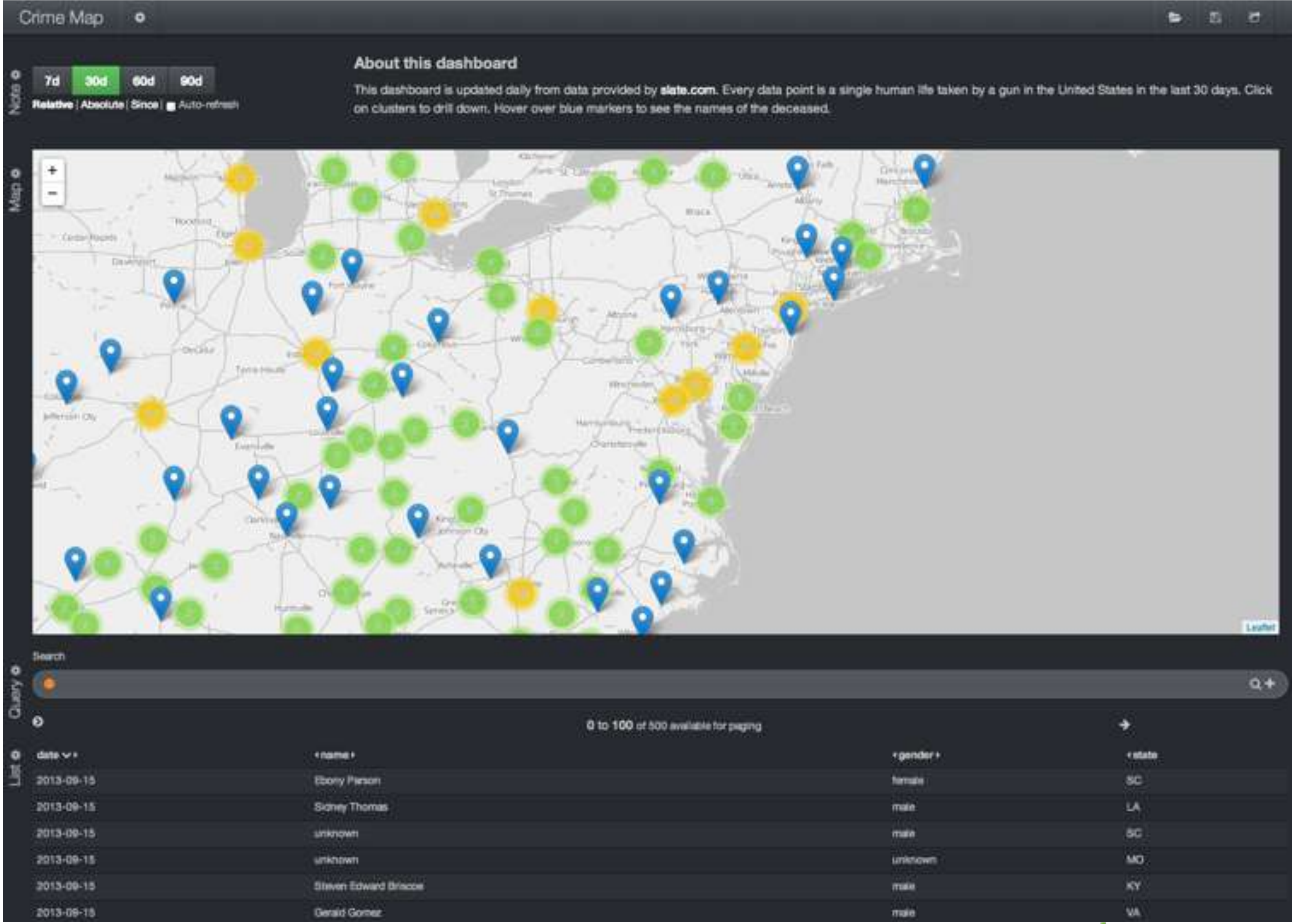
time must ☑ ✖
field : @timestamp
from : "2013-07-24T21:13:04.742Z"
to : "2013-07-26T21:13:04.743Z"

terms mustNot ☑ ✖
field : country
value : cn

terms mustNot ☑ ✖
field : country
value : us

terms mustNot □ ✖
field : country
value : id

Kibana



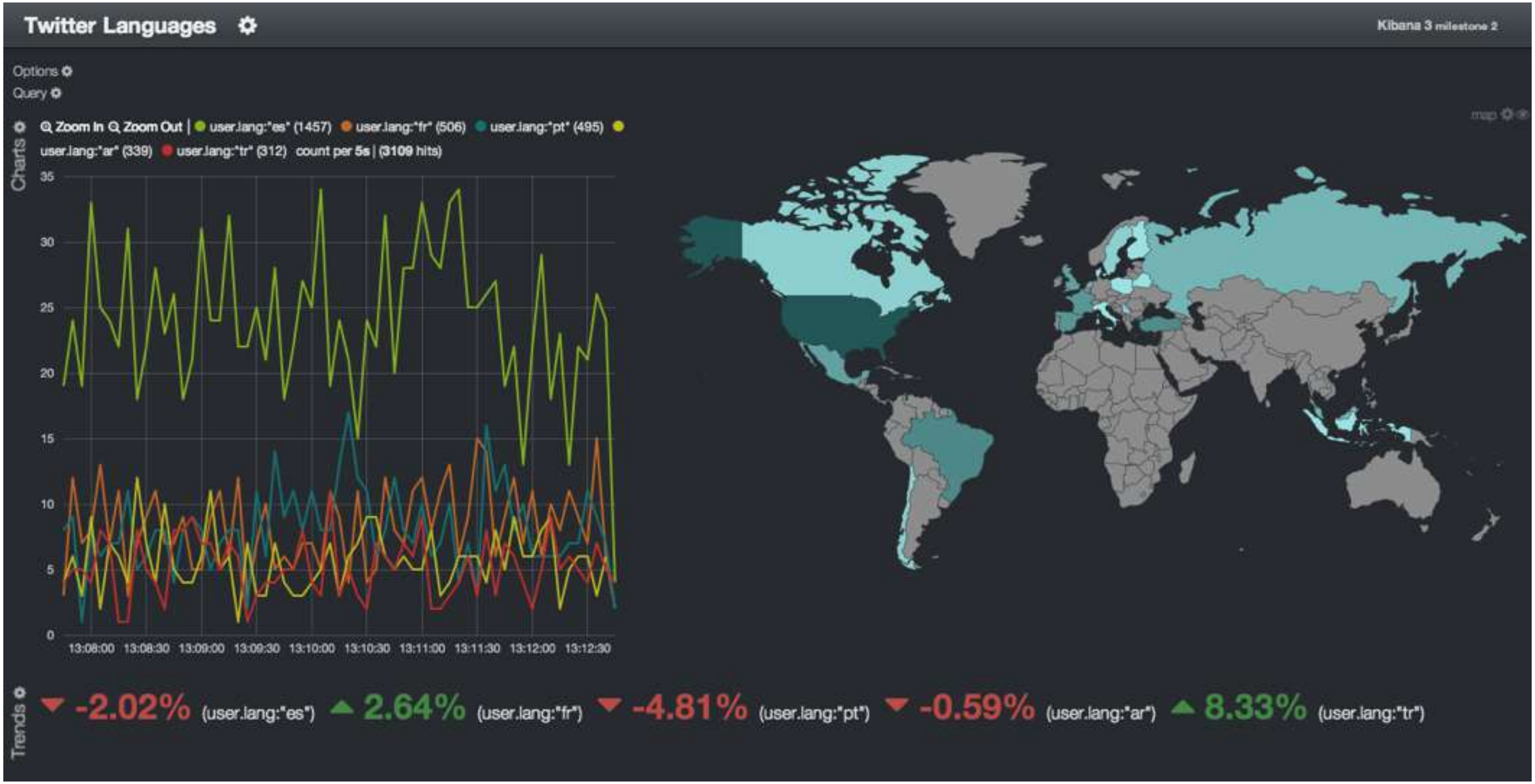
About this dashboard

This dashboard is updated daily from data provided by **slate.com**. Every data point is a single human life taken by a gun in the United States in the last 30 days. Click on clusters to drill down. Hover over blue markers to see the names of the deceased.

0 to 100 of 500 available for paging

date	name	gender	state
2013-09-15	Ebony Parson	female	SC
2013-09-15	Sidney Thomas	male	LA
2013-09-15	unknown	male	SC
2013-09-15	unknown	unknown	MO
2013-09-15	Steven Edward Briscoe	male	KY
2013-09-15	Gerald Gomez	male	VA

Kibana



● user.lang:"es" (1457) ● user.lang:"fr" (506) ● user.lang:"pt" (495) ● user.lang:"ar" (339) ● user.lang:"tr" (312) count per 5s | (3109 hits)

▼ -2.02% (user.lang:"es") ▲ 2.64% (user.lang:"fr") ▼ -4.81% (user.lang:"pt") ▼ -0.59% (user.lang:"ar") ▲ 8.33% (user.lang:"tr")

Not-only log files

- Analyse web streams in realtime

  meetup.com RSVP stream

  us gov page visits

- Billing data (payment morale?)

- IRC

  wikipedia changes

Use-case:
Analytics

Analytics

- Aggregation of information
- Facets are one dimensional

  Categories/brands/material of all results of this query
- Questions are multidimensional

  Average revenue per category id per day
- Enter Aggregations!

Create knowledge from data

- Orders

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

- Money

  What is the average revenue per shopping cart?
  What is the average shopping cart size per order per hour?

- Product portfolio

  Take the location of people into account for special offers?
  Analyse page views: Premium or low budget ecommerce site?

Aggregations

```
» curl -X POST 'localhost:9200/orders/order/_search' -d '
{
    "aggs" : {
        "average_order_size" : {
         "avg" : { "field" : "total" }
        }
    }
}
'
```

```
...
  "aggregations" : {
    "average_order_size" : {
      "value" : 658.369
    }
  }
...
```

Aggregations - Filters

```
{
  "aggs" : {
    "average_order_size_january" : {
      "filter" : {
        "range" : { "created_at" : { "gte" : "2014-01-01", "lt":
"2014-02-01" } } },
      "aggs" : {
        "avg" : { "avg" : { "field" : "total" } }
      }
    }
  }
}
```

```
...
  "aggregations" : {
    "average_order_size_january" : {
      "doc_count" : 8,
      "avg" : { "value" : 540.89375 }
    }
...
```

Aggregations - per day

```
{
   "aggs": {
     "by_day": {
       "filter": {
         "range": {
           "created_at": {
             "gte": “2014-01-01", "lt": "2014-02-01"
           }
         }
       },
       "aggs": {
         "monthly_filter": {
           "date_histogram": {
             "field": "created_at",
             "interval": "day",
             "format": "yyyy-MM-dd"
           },
           "aggs": {
             "average_order_size": { "avg": { "field": “total" } }
           }
} } } } } }
```

Aggregations - per day

```
...
  "aggregations" : {
    "by_day" : {
      "doc_count" : 8,
      "monthly_filter" : [ {
        "key_as_string" : "2014-01-01",
        "key" : 1388534400000,
        "doc_count" : 136,
        "average_order_size" : {
          "value" : 380.0
        }
      }, {
        "key_as_string" : "2014-01-06",
        "key" : 1388966400000,
        "doc_count" : 256,
        "average_order_size" : {
          "value" : 502.575
        }
      }, {
...
```

Aggregations - per hour

```
{
 "aggs": {
  "by_day": {
   "filter": {
    "range": {
     "created_at": { "gte": “2014-01-01", "lt": “2014-02-01" }
    }
   },
   "aggs": {
    "hourly_filter": {
     "histogram": {
      "interval": 1,
      "script": "doc[\u0027created_at\u0027].date.hourOfDay"
     },
     "aggs": {
      "average_order_size": {
       "avg": { "field": “total" }
      }
     }
    }
} } } }
```

Aggregations - per hour

```
...
  "aggregations" : {
    "by_day" : {
      "doc_count" : 8,
      "hourly_filter" : [ {
        "key" : 11,
        "doc_count" : 1,
        "average_order_size" : {
          "value" : 380.0
        }
      }, {
        "key" : 13,
        "doc_count" : 1,
        "average_order_size" : {
          "value" : 450.15
        }
      }
...
```

Elasticsearch 1.0

Elasticsearch 1.0

- Aggregations
- Snapshot/Restore
- Distributed/scalable percolator
- Cat API

  http://www.elasticsearch.org/blog/introducing-cat-api/
- Federated search: Tribe node

Thanks for listening!

Q & A

P.S. We’re hiring
http://elasticsearch.com/about/jobs
http://elasticsearch.com/support

Alexander Reelsen
@spinscale
alexander.reelsen@elasticsearch.com

elasticsearch.