Side by Side with Solr and Elasticsearch

Rafał Kuć

sematext

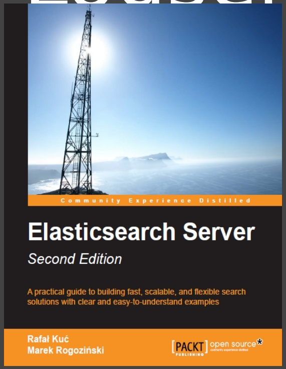
Radu Gheorghe

sematext

Rafał

sematext

Logsene


Community Experience Distilled

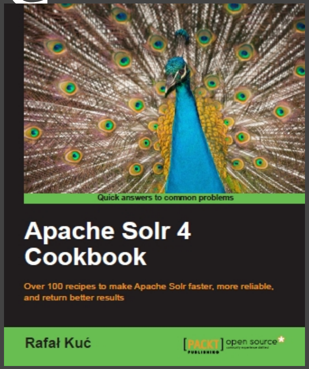
Elasticsearch Server

Second Edition

A practical guide to building fast, scalable, and flexible search solutions with clear and easy-to-understand examples

Rafał Kuć
Marek Rogoziński

[PACKT] open source*

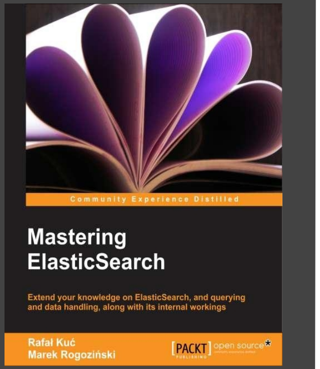

Quick answers to common problems

Apache Solr 4 Cookbook

Over 100 recipes to make Apache Solr faster, more reliable, and return better results

Rafał Kuć

[PACKT] open source*


Community Experience Distilled

Mastering ElasticSearch

Extend your knowledge on ElasticSearch, and querying and data handling, along with its internal workings

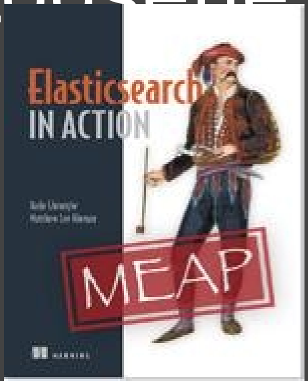
Rafał Kuć
Marek Rogoziński

[PACKT] open source*

Radu

sematext

Logsene

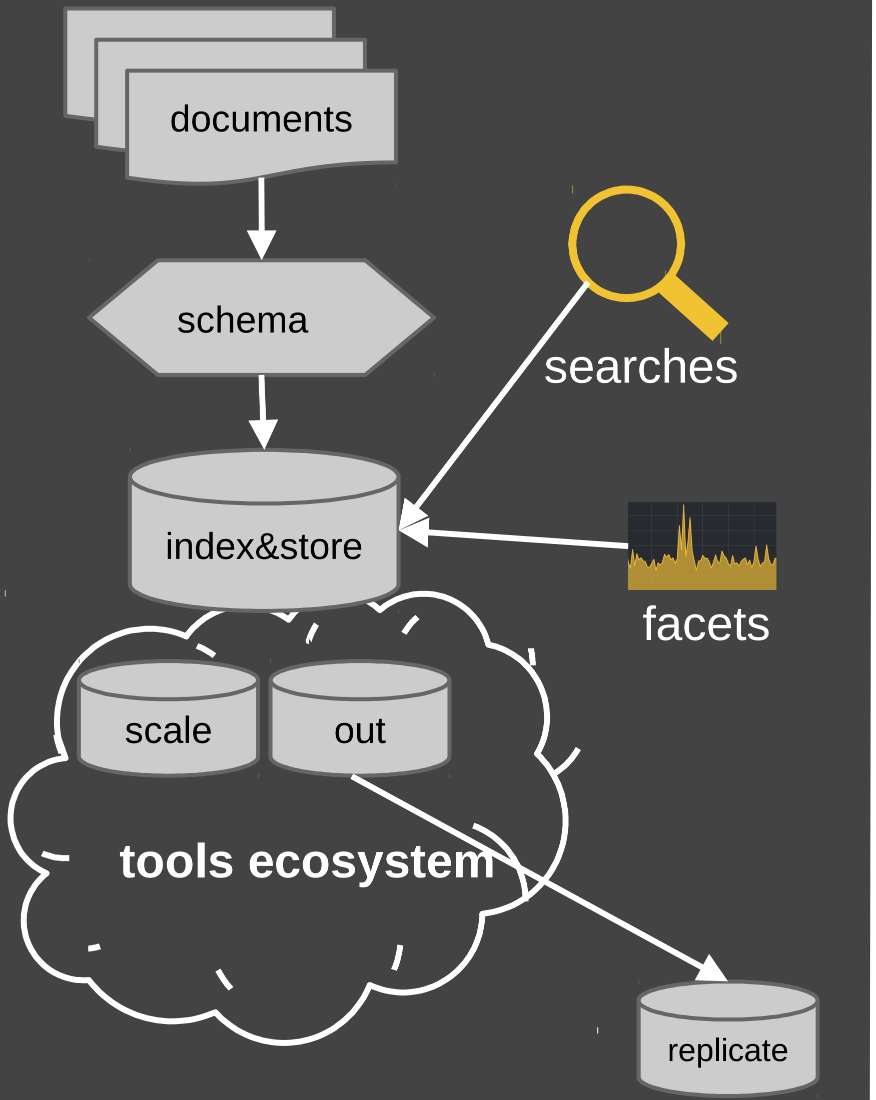

Elasticsearch IN ACTION

[illegible]
[illegible]

MEAP

MANNING

Overview



documents

schema

searches

index&store

facets

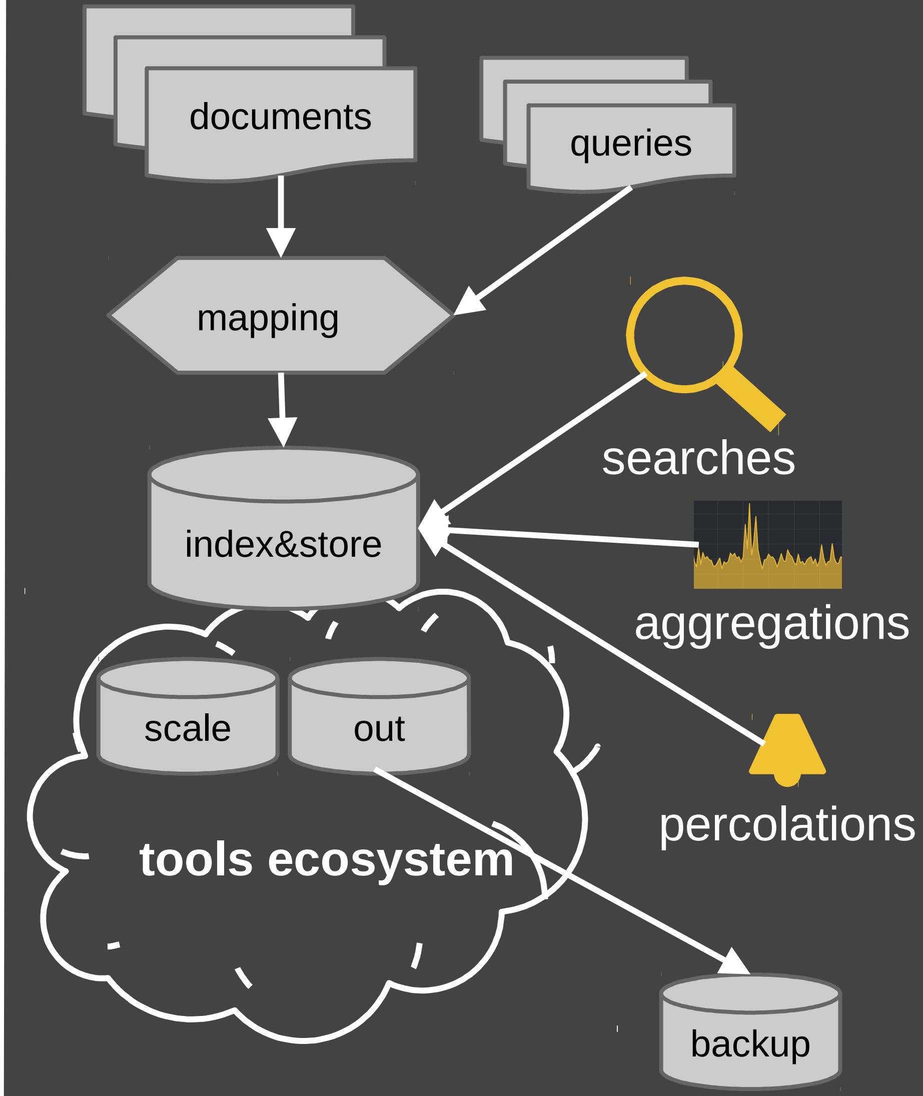
scale

out

tools ecosystem

replicate

Agenda



documents

queries

mapping

searches

index&store

aggregations

scale

out

percolations

tools ecosystem

backup

Let’s Index Videos

```
{
  "id": "4",
  "url": "https://www.youtube.com/watch?v=IutoHcJT61k",
  "title": "#bbuzz: Rafał Kuć: Battle of the Giants: Solr vs ElasticSearch, Round 2",
  "uploaded_by": "newthinking communications",
  "upload_date": "2013-06-19",
  "views": 380,
  "likes": 1,
  "tags": ["elasticsearch", "solr", "lucene", "comparison"]
}
```

Examples available at:
https://github.com/sematext/berlin-buzzwords-samples/

hkcarworld.com

Demo time: Start your engines!

Schema

schema.xml+... -> ZooKeeper

```
<schema
name="BerlinBuzzwords2014"
version="1.5">
 <fields>
   <field name="id" type="string"
          indexed="true" stored="true"
          required="true"
          multiValued="false" />
    ...
   <field name="tags" type="string"
         indexed="true" stored="true"
         multiValued="true"/>
 </fields>
...
</schema>
```

Mapping

PUT -> /bbuzz/videos/_mapping

```
{
  "videos": {
    "_id": {
      "path": "id"
    },
    "properties": {
...
      "tags": {
        "type": "string",
        "index": "not_analyzed"
      },
...
    }
  }
}
```

“q” Parameter

```
GET -> /solr/bbuzz/select
params -> q=elasticsearch
          fl=*,score
...
<result name="response"
       numFound="7" start="0">
 <doc>
  <float
name="score">0.44896343</float>
  <str name="id">2</str>
  <str name="url">
     /watch?v=6QX5hXf_e7c</str>
  <str name="title">Introduction
     to Elasticsearch by Radu</str>
  ...
 </doc>
...
```

URI Request

```
GET -> /bbuzz/videos/_search
params -> q=elasticsearch

...
"hits" : [ {
 "_index" : "bbuzz",
 "_type" : "videos",
 "_id" : "2",
 "_score" : 0.26516503,
 "_source" : {
   "url": "/watch?v=6QX5hXf_e7c",
   "title": "Introduction to
Elasticsearch by Radu",
...
```

Bool Query

```
GET -> /solr/bbuzz/select

q=title:elasticsearch OR tags:logs

q=title:elasticsearch tags:logs
q.op=OR
```

Bool Query

```
GET -> /bbuzz/videos/_search

{
    "query": {
        "bool": {
            "should": [
                {
                    "match": {
                        "title": "elasticsearch"
                    }
                },
                {
                    "term": {
                        "tags": "logs"
...
```

Grouping

```
GET -> /solr/bbuzz/select

q=elasticsearch
group=true
group.field=uploaded_by
```

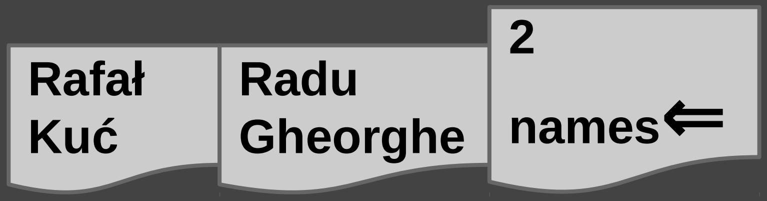
Percolator

```
PUT -> /bbuzz/.percolator/1

{
  "query" : {
    "term" : { "tags" : "elasticsearch" }
  }
}

GET -> /bbuzz/videos/_percolate

{
 "doc": {
  "title": "Scaling Massive ES Clusters",
  "tags": [ "elasticsearch", "scaling"]
 }
}
```

Hierarchies

```
names:
 -> first: Rafał, last: Kuć
 -> first: Radu, last: Gheorghe
```

nested (block join)

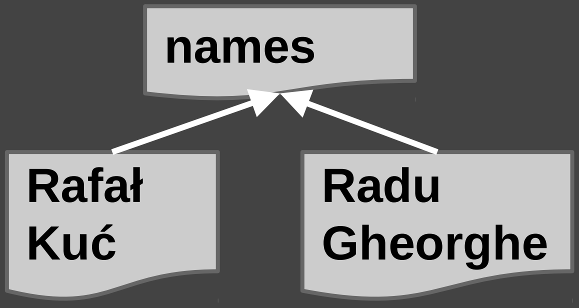


Rafał Kuć

Radu Gheorghe

2 names⇐

parent-child (query time join)



names

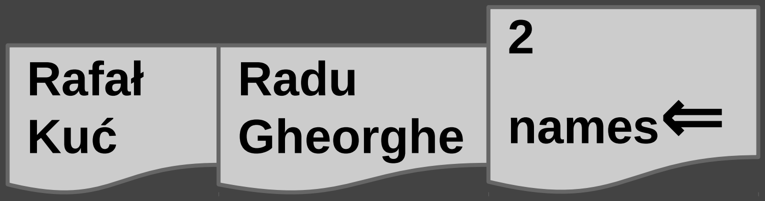
Rafał Kuć

Radu Gheorghe

Hierarchies

```
"names": [
   { "first": "Rafał", "last": "Kuć" },
   { "first": "Radu", "last": "Gheorghe" },
]
```

nested (block join)

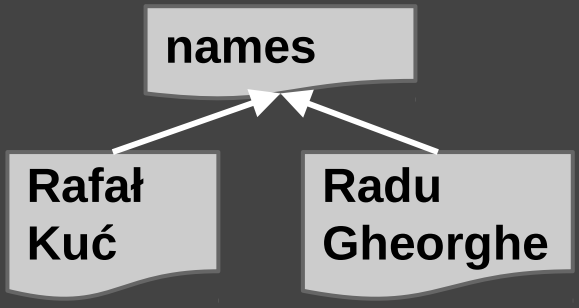


Rafał Kuć

Radu Gheorghe

2 names⇐

parent-child



names

Rafał Kuć

Radu Gheorghe

Facets

```
facet=true
facet.field=tags
```

```
facet=true
facet.query=uploaded_by:LuceneSolrRevolution
facet.query=uploaded_by:"newthinking communications"
```

Aggregations

```
"aggregations" : {
  "tags" : {
    "terms" : { "field" : "tags" }
  }
}
```

```
"aggregations": {
  "uploader_count": {
    "cardinality": {
      "field": "uploaded_by"
    }
  }
}
```

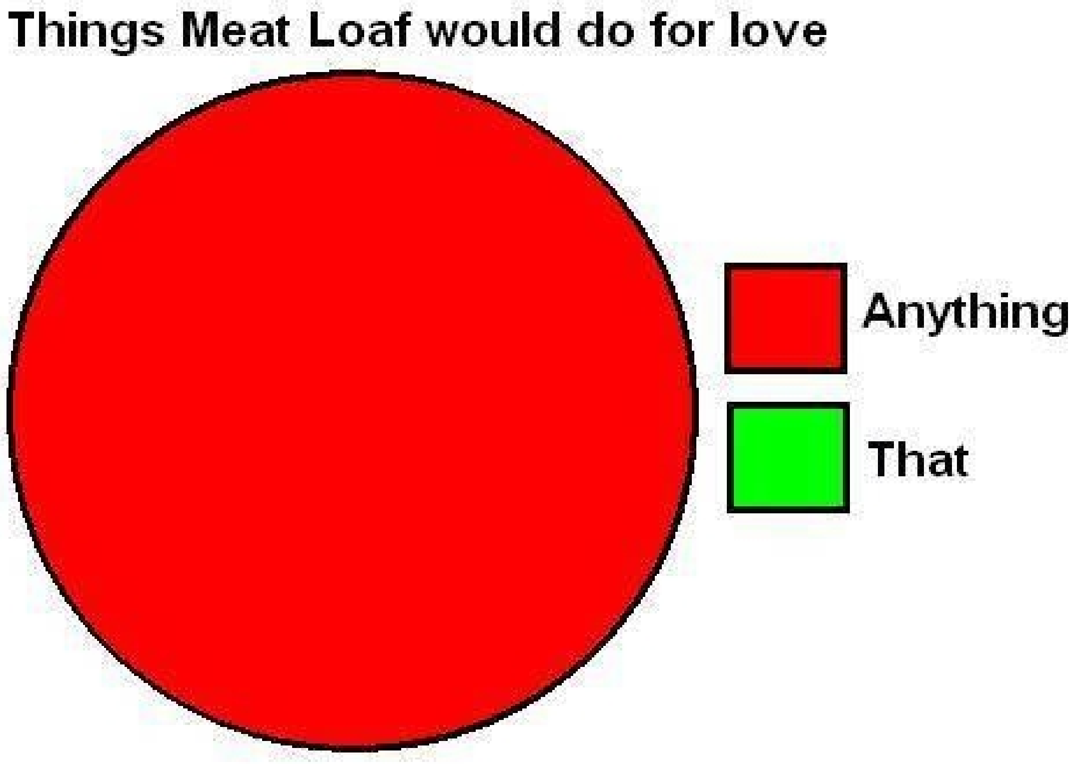
Pivot Facets

```
facet=true
facet.pivot=tags,views
```

Nesting Aggs

```
"aggregations" : {
    "tags" : {
        "terms" : { "field" : "tags" },
        "aggregations": {
            "dates": {
                "date_histogram": {
                    "field": "upload_date",
                    "interval": "month",
                    "format" : "yyyy-MM"
                }
            }
        }
    }
}
```

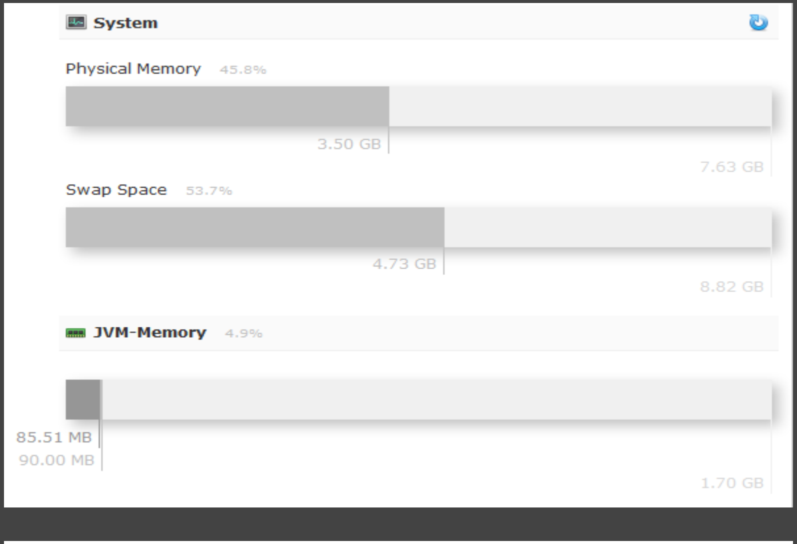
Demo time: Graph all the things!

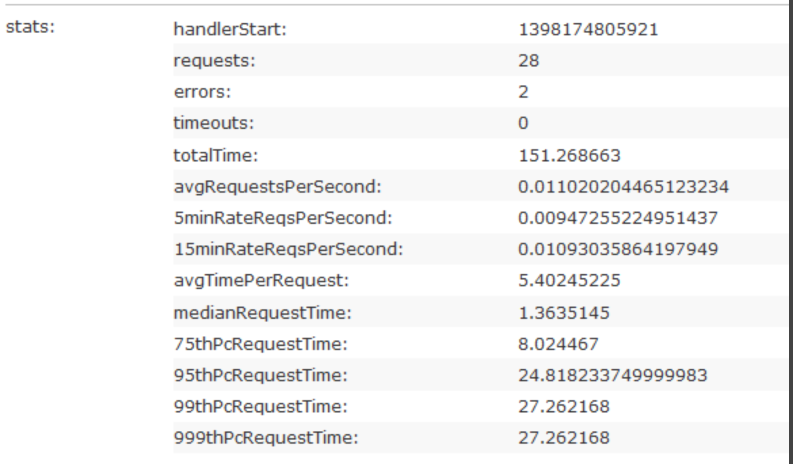


http://f1.thejournal.ie/media/2013/05/meatloaf-2.jpg

Stats

JMX / Solr admin / clusterstate





stats:		
	handlerStart:	1398174805921
	requests:	28
	errors:	2
	timeouts:	0
	totalTime:	151.268663
	avgRequestsPerSecond:	0.011020204465123234
	5minRateReqsPerSecond:	0.00947255224951437
	15minRateReqsPerSecond:	0.01093035864197949
	avgTimePerRequest:	5.40245225
	medianRequestTime:	1.3635145
	75thPcRequestTime:	8.024467
	95thPcRequestTime:	24.818233749999983
	99thPcRequestTime:	27.262168
	999thPcRequestTime:	27.262168

Stats APIs

GET -> /_stats

```
    "index_total" : 15118403,
    "index_time" : "4.2h",
...
    "query_total" : 41092,
     "query_time" : "57.2m",
```

GET -> /_cluster/stats

```
    "heap_used_in_bytes" : 83960392,
...
```

Backup

```
PUT -> /_snapshot/bbuzz
{
    "type": "fs",
    "settings": {
        "location": "/mnt/bbuzz_backup"
    }
}'
```

```
PUT -> /_snapshot/bbuzz/1
{
    "indices": "bbuzz"
}'
```

```
POST -> /_snapshot/bbuzz/1/_restore"
```

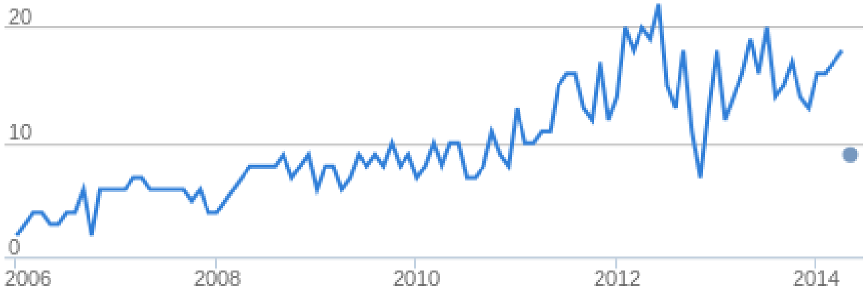
Demo time: Scaling out

© RTF 99 www.toddtoons.com

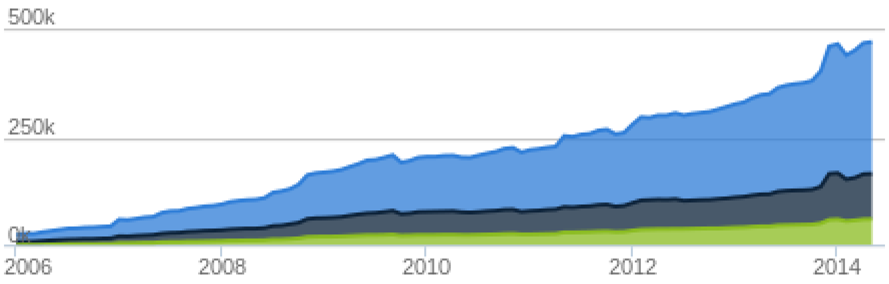
"I think it's called a scale,
but mom calls it a @#$&% liar!"

Apache Software Foundation

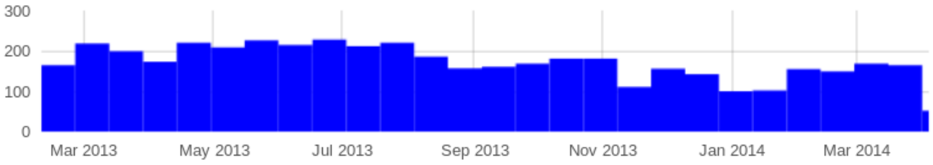
Contributors



Code

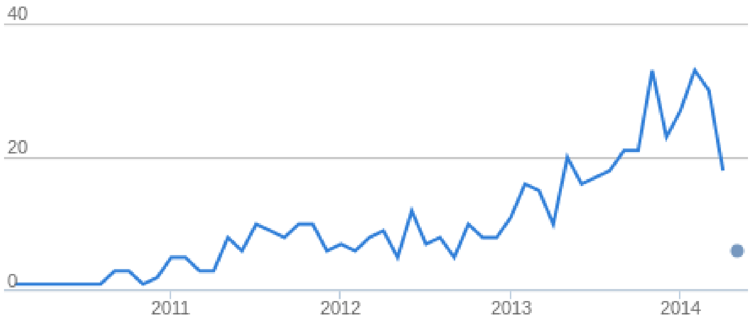


Mailing list

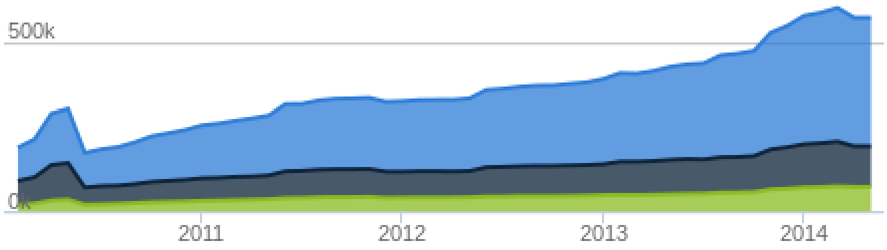


Elasticsearch

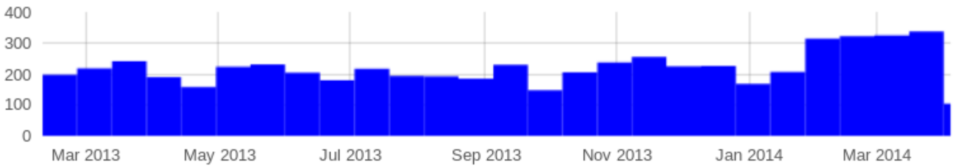
Contributors



Code



Mailing list



New juicy things to come

facet by function
https://issues.apache.org/jira/browse/SOLR-1581

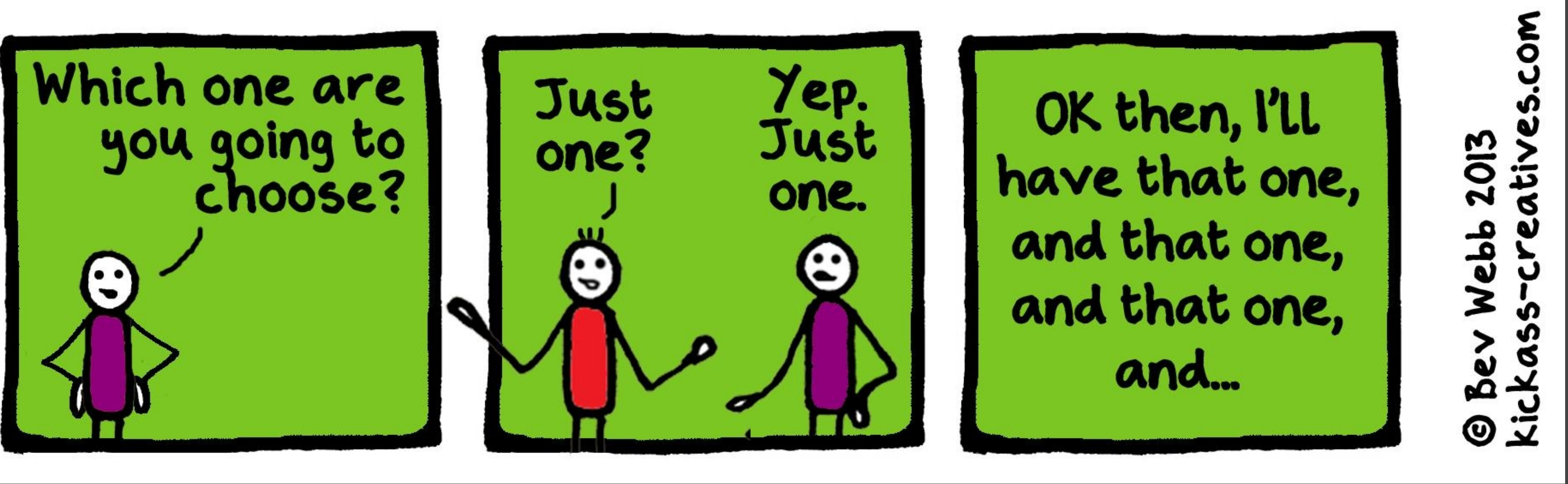
analytics component
https://issues.apache.org/jira/browse/SOLR-5302

Solr as standalone application
5.0 - no general issue yet

top_hits aggregation
https://github.com/elasticsearch/elasticse

minumum_should_match on has_child
https://github.com/elasticsearch/elasticse

filters aggregation
https://github.com/elasticsearch/elasticse

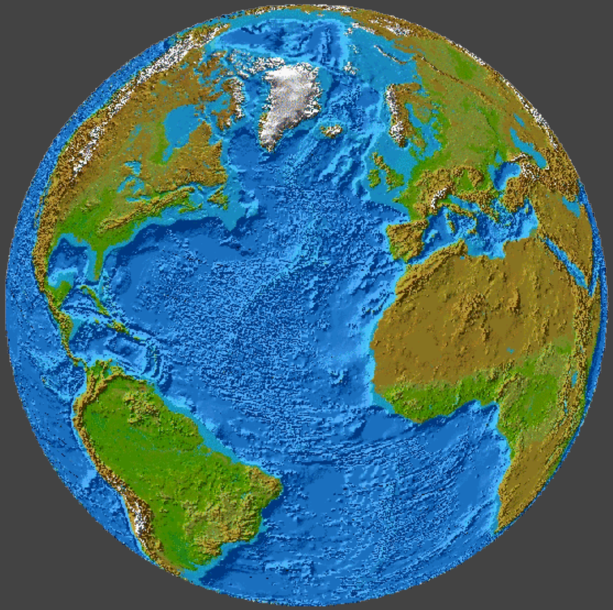

Which one are you going to choose?
Just one?
Yep. Just one.
OK then, I'll have that one, and that one, and that one, and...
© Bev Webb 2013
kickass-creatives.com

most projects work well with either

many small differences, few show-stoppers

choose the best. for your use-case.

Want to work with both? We’re hiring!



Worldwide

http://www.staff.amu.edu.pl/~zbzw/glob/glob.gif

Thank you!

Rafał Kuć
@kucrafal

Radu Gheorghe
@radu0gheorghe

Examples available at:
https://github.com/sematext/berlin-buzzwords-samples/

sematext
@sematext