Deep dive into analytics using Aggregation

Boaz Leskes
@bleskes
Elasticsearch

an end-to-end search and analytics platform.
• full text search
• highlighted search snippets
• search-as-you-type
• did-you-mean suggestions
• combines full text search with geolocation
• uses more-like-this to find related questions and answers
GitHub

- search repositories, users, issues, pull requests
- search 130 billion lines of code
- track all alerts, events, logs
• index and analyse
5TB of log data every day
• combine visitor logs with social network data
• real-time feedback to editors
Feature summary

• Fully-featured search
  Relevance-ranked text search
  Scalable search
  High-performance geo, temporal, numeric range and key lookup
  Highlighting
  Support for complex document types (nested structures)
  Spelling suggestions
  Powerful query DSL
  “Standing” queries
  Real-time results
  Extensible via plugins

• Management
  Simple and robust deployments
  REST APIs for handling all aspects of administration/monitoring
  “Marvel” console for monitoring and administering clusters
  Special features to manage the life cycle of content

• Integration
  Hadoop (MapRed, Hive, Pig, Cascading..)
  Client libraries (Python, Java, Ruby, javascript..)
  Data connectors (Twitter, JMS..)
  Logstash ETL framework

• Powerful faceting/analysis
  Summarise large sets by any combinations of time, geo, category and more.
  “Kibana” visualisation tool

• Support
  Development and Production support with tiered levels
  Support staff are the core developers of the product

* Features we see as differentiators
Let’s talk data

data === json
Prepping up for my talk tomorrow at #nosql14 Cologne, where I’ll spend the coming two days. Drop by for everything #elasticsearch related.

Coder at Elasticsearch

Amsterdam
Data can be anything

- Questions
- Code
- Logs
- Credit card transactions
- Click logs
- …
aggregations == 50km view == patterns
aggregations == 50km view == patterns
a simple UI element...
... or more complex ...
.. even more complex?
Underpinnings

back to search
GET tweets/tweet/_search
{
  "query": {
    "filtered": {
      "query": {
        "match": {
          "text": "jumping"
        }
      },
      "filter": {
        "range": {
          "created_at": {
            "from": "2014-01-28T05:16:29+00:00",
            "to": "now"
          }
        }
      }
    }
  }
}
Inverted index

- nosql: 2 6 8 48 112 379
- 128: 6 9 10 48
- New York: 11 13 14 134 207
- lat=6.9 lon=50: 6 9
- F: 2 4 9 36 103 310
Our goal

Cameras and Optics
- Digital Camera Accessories (389)
- Photo accessories (77)
- Cameras, other (19)
- Optics (17)
- Digital Cameras (2)

Results

- 6
- 8
- 11
- 38
- 153
Counting results

| 6 | 8 | 11 | 38 | 153 |

- Cameras
- Optics
- Accessories
- Lenses
Counting

results

<table>
<thead>
<tr>
<th>6</th>
<th>8</th>
<th>11</th>
<th>38</th>
<th>153</th>
</tr>
</thead>
</table>

3 | 1 | 2 | 0 |
Field data

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>134</th>
<th>207</th>
<th>379</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>nosql</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>128</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>New York</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>lat=6.9</td>
<td>lon=50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copyright Elasticsearch 2013. Copying, publishing and/or distributing without written permission is strictly prohibited
Introducing aggregations

analysis lego
Before there were facets

• facets are awesome
• the serve well and long
• but… they are not scalable from a functionality perspective
Analysis lego
Buckets and metrics

<table>
<thead>
<tr>
<th>results</th>
<th>buckets</th>
<th>metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Cameras</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>Optics</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>Accessories</td>
<td>2</td>
</tr>
<tr>
<td>38</td>
<td>Lenses</td>
<td>0</td>
</tr>
</tbody>
</table>
Buckets and metrics

results

| 6  | 8  | 11 | 38 | 153 |

buckets

Cameras
Optics
Accessories
Lenses

metrics

2013
2012
2013
2012
3
1
2
0
GET localhost:9200/_search
{
  "aggs": {
    "countries": {
      "terms": {
        "field": "country"
      }
    },
    "subjects": {
      "terms": {
        "field": "subject"
      }
    },
    "aggs": {
      "avg_score": {
        "avg": {
          "field": "score"
        }
      }
    }
  }
}

{
  "hits": { ... },
  "aggregations": {
    "countries": {
      "buckets": [
        {
          "key": "USA",
          "doc_count": 5,
          "aggregations": {
            "subjects": {
              "buckets": [
                {
                  "key": "Mathematics",
                  "doc_count": 3,
                  "aggregations": {
                    "avg_score": {
                      "avg": {
                        "field": "score"
                      }
                    }
                  }
                }
              ]
            }
          }
        }
      ]
    }
  }
}
measure all the things

- doc count (free!)
- avg
- min
- max
- sum
- count
- stats
- extended stats
Buckets

- global
- filter
- missing
- terms
- range
- date range
- ip range
- histogram
- date histogram
- geo distance
- nested
- geohash grid
- cardinality
- percentiles
- significant terms
Example - range bucket

GET localhost:9200/grades/grade/_search
{
    "aggs" : {
        "age_groups" : {
            "range" : {
                "field" : "age",
                "ranges" : [
                    { "from" : 5, "to" : 10 },
                    { "from" : 10 }
                ],
            },
            "aggs" : {
                "avg_grade" : {
                    "avg" : {
                        "field" : "grade"
                    }
                }
            }
        }
    }
}

"age_groups": {
    "buckets": [
        {
            "from": 5,
            "to": 10,
            "doc_count": 911,
            "avg_grade": {
                "value": 81.603
            }
        },
        {
            "from": 10,
            "doc_count": 2276,
            "avg_grade": {
                "value": 82.357
            }
        }
    ]
}
Example - histogram

GET grades/grade/_search
{
   "aggs": {
      "grades_distribution": {
         "histogram": {
            "field": "grade",
            "interval": 10
         }
      }
   }
}

Example - histogram

GET grades/grade/_search
{
  "aggregations": {
    "grades_distribution": {
      "buckets": [
        {
          "key": 60,
          "doc_count": 467
        },
        {
          "key": 70,
          "doc_count": 873
        },
        {
          "key": 80,
          "doc_count": 930
        },
        {
          "key": 90,
          "doc_count": 915
        }
      ]
    }
  }
}
Significant terms

analytics as search
Common crimes

```
GET ukcrimes/_search
{
    "query": { }
    "aggregations": {
        "map": {
            "geohash_grid": {
                "field": "location",
                "precision": 5,
            },
            "aggregations": {
                "most_popular_crime_type": {
                    "terms": {
                        "field": "crime_type", "size": 1
                    }
                }
            }
        }
    }
}
```
Geo-what?
Geo-what?
Geo-what?
The common terms problem

- Our sites users
- Subscribers to Martha Stewart Living
- Consumers of furry pornography

Pet peeve #208: Geographic profile maps which are basically just population maps.
Uncommonly common

GET ukcrimes/_search
{
    "query": { }
    "aggregations": {
        "map": {
            "geohash_grid": {
                "field": "location",
                "precision": 5,
            },
            "aggregations": {
                "most_popular_crime_type": {
                    "significant_terms": {
                        "field": "crime_type",
                        "size": 1
                    }
                }
            }
        }
    }
}
Uncommonly common

989 crimes
Has 161 incidents of Possession of weapons vs expected 2

<table>
<thead>
<tr>
<th>Month</th>
<th>Num Incidents</th>
</tr>
</thead>
<tbody>
<tr>
<td>201301</td>
<td>72</td>
</tr>
<tr>
<td>201303</td>
<td>92</td>
</tr>
<tr>
<td>201304</td>
<td>89</td>
</tr>
<tr>
<td>201305</td>
<td>120</td>
</tr>
<tr>
<td>201306</td>
<td>115</td>
</tr>
<tr>
<td>201307</td>
<td>85</td>
</tr>
<tr>
<td>201308</td>
<td>114</td>
</tr>
<tr>
<td>201309</td>
<td>113</td>
</tr>
<tr>
<td>201310</td>
<td>96</td>
</tr>
<tr>
<td>201311</td>
<td>93</td>
</tr>
</tbody>
</table>
Demo!
thank you!

@elasticsearch , @bleskes

http://elasticsearch.org/resources
http://elasticsearch.com/support