Real-time Data Analytics mit Elasticsearch

Bernhard Pflugfelder
inovex GmbH
Bernhard Pflugfelder

- Big Data Engineer @ inovex
- Fields of interest:
  - search
  - analytics
  - big data
  - bi
- Working with:
  - Lucene
  - Solr
  - Elasticsearch
  - Hadoop Ecosystem
- bpflugfelder@inovex.de
Agenda

- elasticsearch intro
- import your data
- analyze your data
- visualize your data
You know, ... don’t you?
data analysis landscape
the big picture
elasticsearch intro

- scalable
- document-oriented
- REST & JSON
- Lucene under the hood
- plugin architecture
- Apache 2 license
elasticsearch intro
architecture

node discovery
node types
fault tolerant
high availability
elasticsearch intro

document-oriented & flat data model
elasticsearch intro
core types, mapping, manipulation

- core types
- mapping
- insert, update, delete
- real-time get
- search query types
- snapshot & backup
getting data into elasticsearch …
getting data into elasticsearch

index api

http bindings

spring-data-elasticsearch

logstash

rivers

flume fluentd
getting data into elasticsearch

logstash

- log collection and management tool
- collects, parses and stores log events
- became part of the ELK stack
- seamless integration with elasticsearch
- plugin architecture:
  - inputs (syslog, ganglia, log4j and more)
  - codec (json, line, multiline, …)
  - filters (csv, json, date, grep, …)
  - outputs (elasticsearch, …)

- expect that logstash will be promoted to a more general ingestion pipeline
getting data into elasticsearch

rivers

- works as an elasticsearch plugin
- service for pulling data into cluster
- examples:
  - couchdb river
  - rabbitmq river
  - csv river
  - jdbc river
  - twitter river
  - wikipedia river
- runs on a single node
- automatic allocation
- shall be deprecated sooner or later
getting data into elasticsearch

elasticsearch and hadoop

- former hadoop-elasticsearch became official integration of elasticsearch and hadoop
- makes elasticsearch accessible from hive, pig, cascading and map/reduce
- automatic mapping between elasticsearch’s json and hadoop file formats
- every query to elasticsearch is performed by m/r jobs as follows:
  - one mapper task per shard
  - final aggregation by reducer
- elasticsearch works as a separate data store, index files are not stored in hdfs

```
CREATE EXTERNAL TABLE artists (  
  id BIGINT, name STRING, links STRUCT)  
STORED BY 'org.elasticsearch.hadoop.hive.ESSerializationHandler'  
TBLPROPERTIES('es.resource' = 'radio/artists/_search?q=*me*');

-- stream data from Elasticsearch
SELECT * FROM artists;
```

```
-- load data from HDFS into Pig using a schema
A = LOAD 'in/artists.dat' USING PigStorage()  
  AS (id:long, name, url:chararray, picture: chararray);  

-- ETL the data to your heart content
B = FOREACH A GENERATE name, TOTUPLE(url, picture) AS links;

-- save the result to Elasticsearch
STORE B INTO 'radio/artists' USING org.elasticsearch.hadoop.pig.ESSerialization();
```

```
JobConf conf = new JobConf();  
// index used for storing data
conf.set("es.resource", "radio/artists");  
// use dedicated output format
conf.setOutputFormatClass(ESOutputFormat.class);
...
JobClient.runJob(conf);
```

from http://www.elasticsearch.org/blog/elasticsearch-and-hadoop/
analyze your data ...
analyze your data
you know about facets, I am sure
analyze your data
same analysis methodology, other visualization == kibana panels
analyze your data
next generation of facets

facets

limited analysis functionality

aggregations

enabling custom analysis
analyze your data
aggregations (aggs)

- gives insight into data space by
  - slicing along dimensions
  - drill down
  - interactive
  - quick by using field data

- two types of aggregations
- many types of aggregators
- customize with scripting
- use over search api
- json in / json out
analyze your data

two aggregation types

Bucket aggs
Aggregations that split the original set of documents into separate buckets.

Metric aggs
Aggregations that compute a specific metrics over a set of documents by aggregating of all documents per bucket.
analyze your data
aggregation example

_id: 1
ref: seo

_id: 2
ref: direct

_id: 3
ref: seo

_id: 4
ref: other

_id: 5
ref: direct

_id: 6
ref: seo

bucket
ref: seo
id: 1, 3, 6
ref: direct
id: 2, 5
ref: other
id: 4

metrics
3
2
1
analyze your data
aggregation example

_id: 1
ref: seo

_id: 2
ref: direct

_id: 3
ref: seo

_id: 4
ref: other

_id: 5
ref: direct

_id: 6
ref: seo

buckets

desktop
id: 1

mobile
id: 3, 6

seo
id: 1, 3, 6

direct
id: 2, 5

other
id: 4

desktop
id: 1, 3, 6

metrics

1

2

1

1

1
analyze your data
customize your analysis with nested aggregators

```
"aggregations": {
  "<aggregation_name>": {
    "<aggregation_type>": {
      <aggregation_body>
    },
    ["aggregations": { [<sub_aggregation>]* }]}
  [,"<aggregation_name_2>": { ... }]*
} }
```

my_aggregation:

- bucket 1
- bucket 2
- ...
- bucket n
- metrics
analyze your data
many types of aggregators

- terms
- range
- date range
- histogram
- date histogram
- geo distance
- geohash grid
- ...

- min
- max
- sum
- avg
- value count
- percentiles
- cardinality
- ...

25
visualize your data …
visualize your data

kibana

light-weighted web frontend

visualize time-stamped data

fancy visualization

panels to visualize

creating dashboards

sharing dashboards
elasticsearch is not only a search technology

elasticsearch also provides powerful capabilities for data analytics

- aggregations framework
- real-time analytics
- plus: elasticsearch enables you to analyze unstructured along with structured data in one place

data analytics ecosystem of elasticsearch:

- ELK stack (ingestion + analysis + visualization)
- deep hadoop integration to avoid separate data silos and make use of the advantages of both words

... and thanks for your attention!
Thank you very much for your attention

Contact

Bernhard Pflugfelder
*Big Data Engineer*

Cell: 0173 3181088
Mail: bernhard.pflugfelder@inovex.de