Elasticsearch - Eine Einführung

Till Hildebrandt, Frederik Schnoege, Tobias Jansing

Elasticsearch Allgemein

Was ist Elasticsearch?

- Verteilte Suchmaschine und Analytics-Engine
- Speicherung von Dokumenten im JSON-Format
- REST-API
- Xing, GitHub, Stackoverflow...

Was ist Elasticsearch?

Besonderheiten

Datenverteilung durch Sharding auf Clustern

Automat. Replikation

Verteilte Suchfunktion

Analytics (Kibana, Logstash)

Ziele

Verfügbarkeit

Performanz

Skalierbarkeit

Ausfallsicherheit

Lastverteilung

Vorteile

Nahezu-Echtzeit-Suche

Analytics

Geodaten + Score-Werte

Open Source

Viele Sprachen verfügbar

Kibana

Apache - Total Visitors

4,931,584

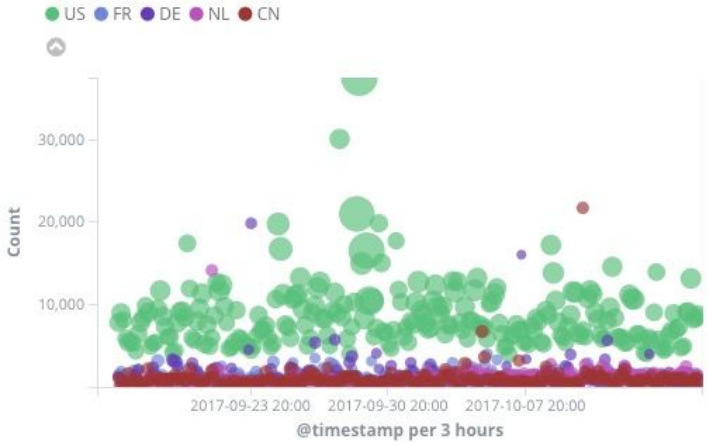
Apache - Unique Visitors

29,740

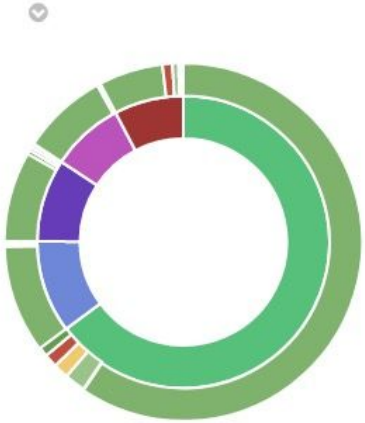
Apache - Unique Visitors ...

City	Unique Visitors
Beijing	562
Redmond	445
Ashburn	400
Chicago	373
Los Angeles	245
Seattle	233
San Jose	232
Singapore	208

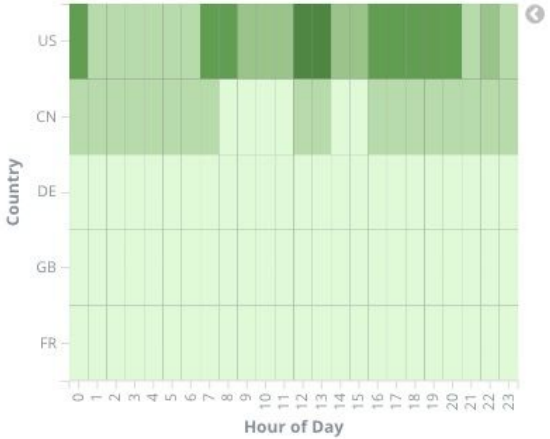
Apache - Bytes and Count



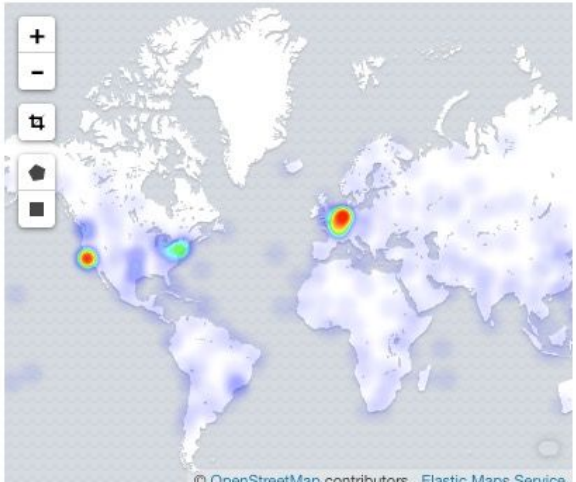
Apache - Country and Status



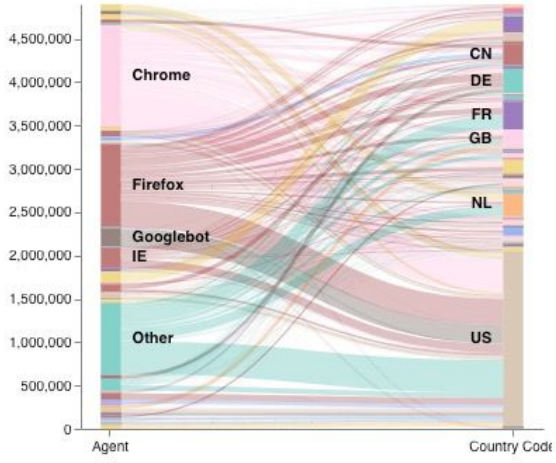
Apache - Country traffic by hour

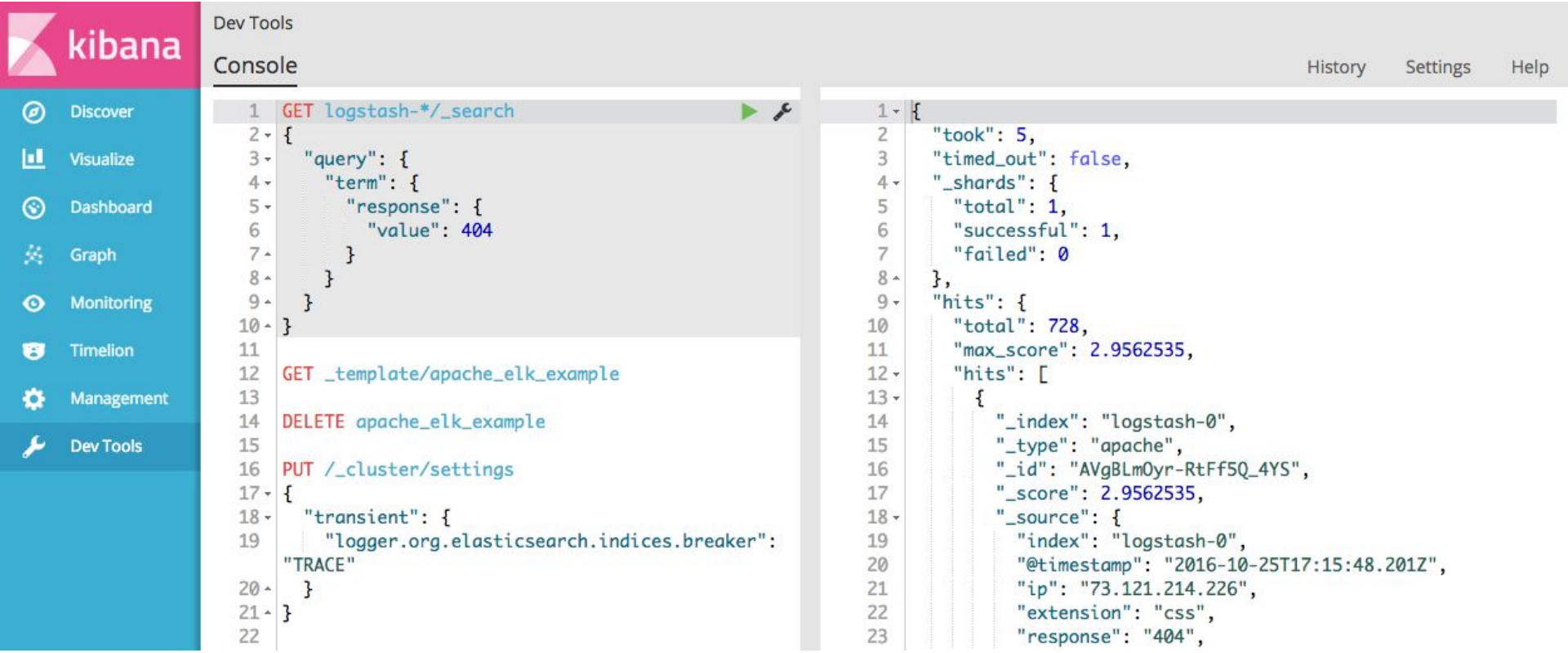


Apache - Visitor Map (geocentroid)



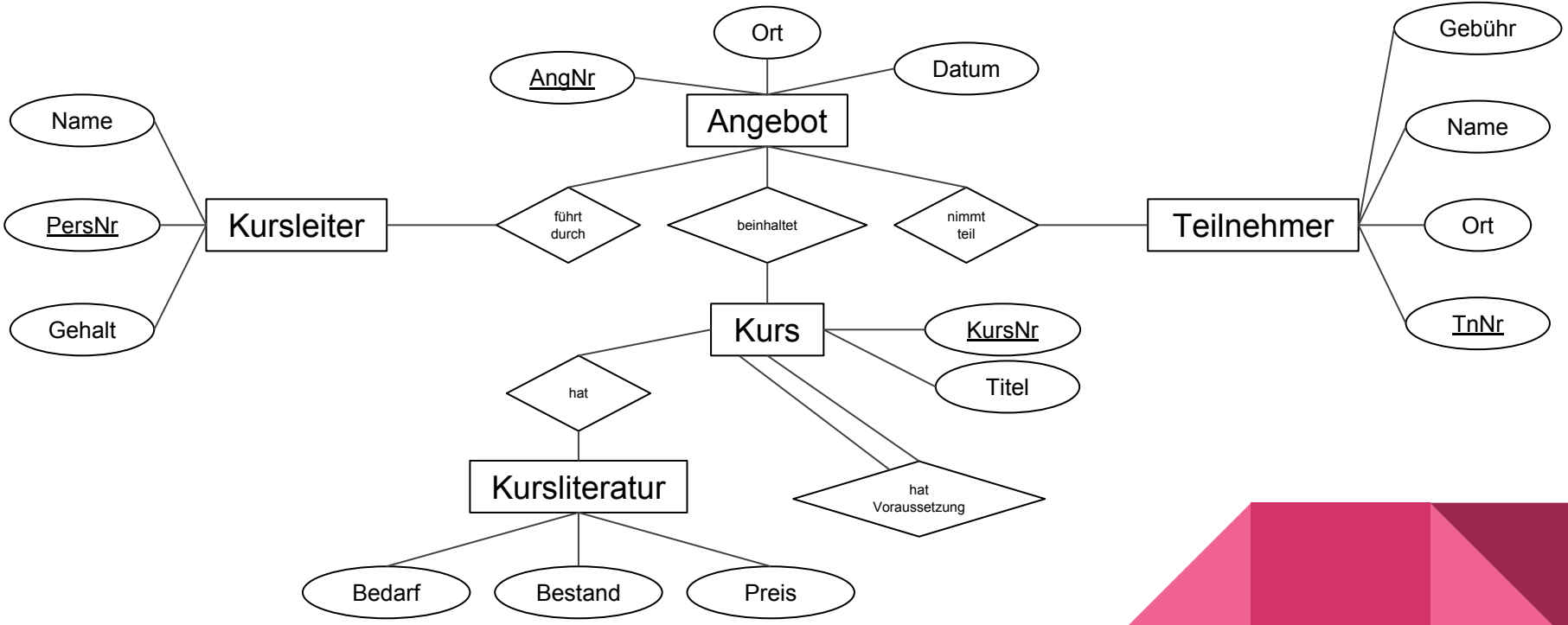
Apache - Browser to Country (vega)




kibana
Discover
Visualize
Dashboard
Graph
Monitoring
Timelion
Management
Dev Tools
Dev Tools
Console
History
Settings
Help
GET logstash-*/_search
{
"query": {
"term": {
"response": {
"value": 404
}
}
}
}
GET _template/apache_elk_example
DELETE apache_elk_example
PUT /_cluster/settings
{
"transient": {
"logger.org.elasticsearch.indices.breaker": "TRACE"
}
}
{
"took": 5,
"timed_out": false,
"_shards": {
"total": 1,
"successful": 1,
"failed": 0
},
"hits": {
"total": 728,
"max_score": 2.9562535,
"hits": [
{
"_index": "logstash-0",
"_type": "apache",
"_id": "AVgBLmOyr-RtFf5Q_4YS",
"_score": 2.9562535,
"_source": {
"index": "logstash-0",
"@timestamp": "2016-10-25T17:15:48.201Z",
"ip": "73.121.214.226",
"extension": "css",
"response": "404",

ER-Modell

ER-Modell


Ort
AngNr
Datum
Angebot
Gebühr
Name
Name
PersNr
Kursleiter
führt durch
beinhaltet
nimmt teil
Teilnehmer
Ort
Gehalt
Kurs
KursNr
TnNr
Titel
hat
Kursliteratur
hat Voraussetzung
Bedarf
Bestand
Preis

```
PUT angebot
{
  "settings": {
    "number_of_shards": 1
  },
  "mappings": {
    "_doc": {
      "properties": {
        "AngNr": {
          "type": "keyword"
        },
        "Kurs": {
          "type": "nested",
          "properties": {
            "KursNr": {
              "type": "keyword"
            },
            "Titel": {
              "type": "keyword"
            },
            "KursLit": {
              "type": "nested",
              "properties": {
                "Bestand": {
                  "type": "integer"
                },
                "Bedarf": {
                  "type": "integer"
                },
                "Preis": {
                  "type": "double"
                }
              }
            },
            "Voraus": {
              "type": "nested",
              "properties": {
                "VorNr": {
                  "type": "keyword"
                }
              }
            }
          }
        },
        "Datum": {
          "type": "date",
          "format": "dd.MM.yyyy"
        },
        "Ort": {
          "type": "keyword"
        },
        "Kursleiter": {
          "type": "nested",
          "properties": {
            "PersNr": {
              "type": "keyword"
            },
            "Name": {
              "type": "keyword"
            },
            "Gehalt": {
              "type": "double"
            }
          }
        },
        "Teilnehmer": {
          "type": "nested",
          "properties": {
            "TnNr": {
              "type": "keyword"
            },
            "Name": {
              "type": "keyword"
            },
            "Ort": {
              "type": "keyword"
            },
            "Gebuehr": {
              "type": "double",
              "null_value": 0
            }
          }
        }
      }
    }
  }
}
```

Queries

```
GET /angebot/_search?filter_path=aggregations.Orte.buckets.key
{
  "aggs" : {
    "Orte" : {
      "terms" : {
        "field" : "Ort"
      }
    }
  }
}
```

```
POST angebot/_update_by_query
{
  "query": {
    "bool": {
      "must": {
        "match": {
          "Ort": "Kiel"
        }
      }
    }
  },
  "script": "ctx._source.Ort = 'Lübeck'"
}
```

```
GET angebot/_search?filter_path=aggregations.*.*.buckets.key,aggregations.*.*.buckets.Vorauscount.value
{
    "aggs": {
        "Kurs.Titel": {
            "nested": {
                "path": "Kurs"
            },
            "aggs": {
                "Kurs.Titel": {
                    "terms": {
                        "field": "Kurs.Titel"
                    },
                    "aggs": {
                        "Vorauscount": {
                            "bucket_script": {
                                "buckets_path": {
                                    "vCnt": "Kurs.Voraus._count",
                                    "kCnt": "_count"
                                },
                                "script": "params.vCnt / params.kCnt"
                            }
                        },
                        "Kurs.Voraus": {
                            "nested": {
                                "path": "Kurs.Voraus"
                            }
                        },
                        "voraus_bucket_filter": {
                            "bucket_selector": {
                                "buckets_path": {
                                    "vCnt": "Kurs.Voraus._count",
                                    "kCnt": "_count"
                                },
                                "script": "params.vCnt / params.kCnt >= 2"
                            }}}}}}}
```

Fazit, Erfahrungen und Probleme

Erfahrungen und Probleme

- Fehlende SQL-Standardfeatures
  - Fehlende Joins → verkompliziert Datenstruktur
  - Daten manipulieren, formatierte Ausgaben etc. schwierig
  - Direkte Umwandlung von SQL-Befehlen in Elastic-Befehle nicht einfach möglich
  - Keine Subqueries, Speicherung der Resultate einzelner Queries nötig
- Doppelte Datenhaltung notwendig, obwohl nicht angestrebt
- Keine Tabellen sondern Objekte → völlig andere Verwendung

Fazit

- Daten ungeeignet für Elasticsearch
  - Konzipiert für unstrukturierte Daten, keine Normalisierung der Daten
  - Fehlende SQL-Features bzw. andere Architektur verkomplizieren Übergang
  - Filter, Aggregationen etc. teils sehr kompliziert und unübersichtlich
- Geeignet eher für das Speichern und Abfragen von Dokumenten
  - Elastic-Prinzip: Wegschreiben und bei Bedarf alles Abfragen
  - Verknüpfung von Informationen aus unterschiedlichen Dokumenten schwierig
- SQL-Denkweise muss angepasst werden
- Elastic ist schnell

Noch Fragen?