



Lösungen für individuelle Prozesse

Umsetzung von BI-Lösungen mit Unterstützung einer Suchmaschine

6. Workshop Open Source Business Intelligence

05.03.2015

Tobias Kraft, exensio GmbH

Agenda

Suchmaschinen

Elasticsearch

BI-Stack mit Elasticsearch

Umsätze Pharma

Funktionen einer Suche

The screenshot shows a search engine interface with the following features highlighted by callouts:

- Unstrukturierte Suche**: Points to the search input field containing the text "exensio".
- Did you mean**: Points to the search input field.
- Synonyme**: Points to the search input field.
- Autocomplete**: Points to the search input field.
- Sortierung**: Points to the sorting options: "Sortieren nach: Titel Relevanz Datum".
- Strukturierte Suche**: Points to the tags of a search result: "Tags: Mitteilung, PPG, Aussendienst, Marketing".
- Facettierung**: Points to the left sidebar filter menu, specifically the "Marketing (18)" selection.
- Highlighting**: Points to the word "exensio" in the search result text.
- Blätterung**: Points to the pagination information: "Ergebnisse 1 bis 10 von 18 auf Seite 1".

The interface includes a left sidebar with filters for "TAXONOMIE", "ORGANISATIONSEINHEIT", "INFORMATIONSTYP", "REDAKTEUR", and "AUTOR". The main content area displays search results with titles, dates, authors, and snippets of text.

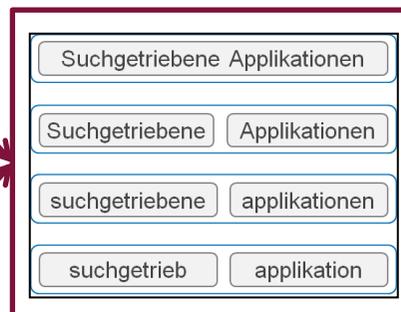
Speichern von Daten in einer Suchmaschine

Document

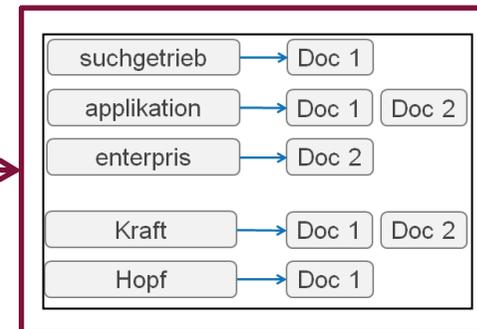
```
{
  "conference": "Berlin Expert Days",
  "title": "Suchgetriebene Applikationen",
  "speakers": [
    "Hopf, Florian",
    "Kraft, Tobias"
  ],
  "date": "04-03-2014",
  "location": "Berlin"
}
```

```
{
  "conference": "JUG SAXONY DAY",
  "title": "Grails für Enterprise-Applikationen?",
  "speakers": [
    "Kraft, Tobias"
  ],
  "date": "04-04-2014",
  "location": "Dresden"
}
```

Analyzing



Struktur



Aufbau

```
1 PUT /bedcon/talk/_mapping
2 {
3   "properties": {
4     "conference": {
5       "type": "string",
6       "format": "dateOptionalTime"
7     },
8     "date": {
9       "type": "date",
10      "format": "dateOptionalTime"
11    },
12    "location": {
13      "type": "string"
14    },
15    "speakers": {
16      "type": "multi_field",
17      "fields": {
18        "speakers": {
19          "type": "string"
20        },
21        "letter": {
22          "type": "string",
23          "analyzer": "single_char_analyzer"
24        }
25      }
26    },
27    "title": {
28      "type": "string",
29      "analyzer": "german"
30    }
31  }
32 }
```

Elasticsearch im Überblick



- Suchmaschine unter Apache 2 Open Source License
- Erstes Release 2010

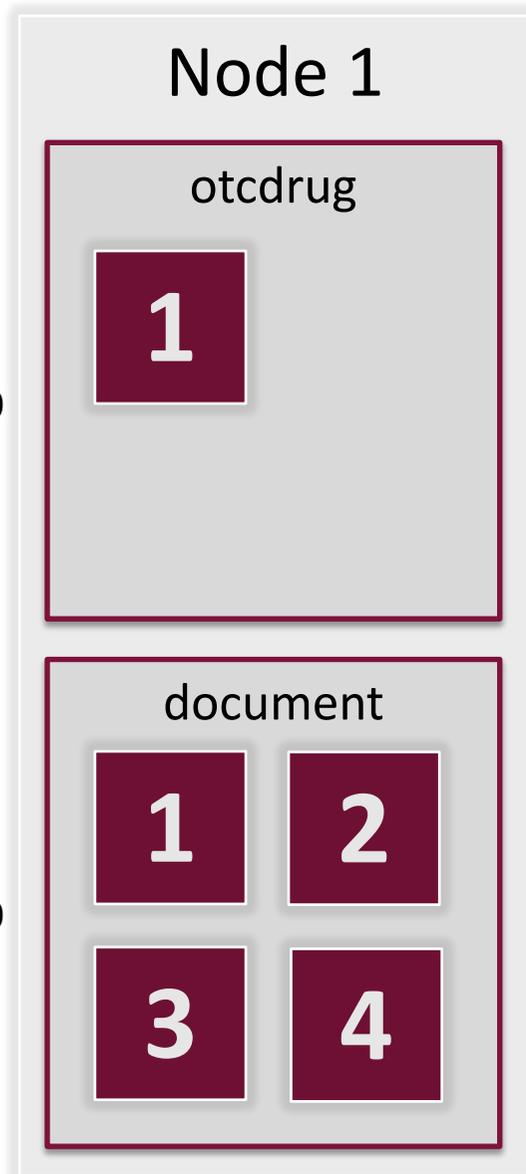
- Basiert auf Java
- Basiert auf Lucene

- JSON-API
- Schemalos
- Plugins

- Runterladen und loslegen
- Im Trend

```
1 GE {
2
3
4
5
6
7
8 }
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
"hits": {
  "total": 7124,
  "max_score": 0.66425383,
  "hits": [
    {
      "_index": "factmarket",
      "_type": "jdbc",
      "_id": "AUvW12UnPHK6qcGfww5-",
      "_score": 0.66425383,
      "_source": {
        "id": 42181,
        "version": 0,
        "aut_idem": null,
        "date_created": "2013-01-16T00:52",
        "gross_sales": null,
        "last_updated": "2013-01-16T00:52",
        "launch_revenues0": null,
        "launch_revenues1": null,
        "market_gen_launches0": null
```

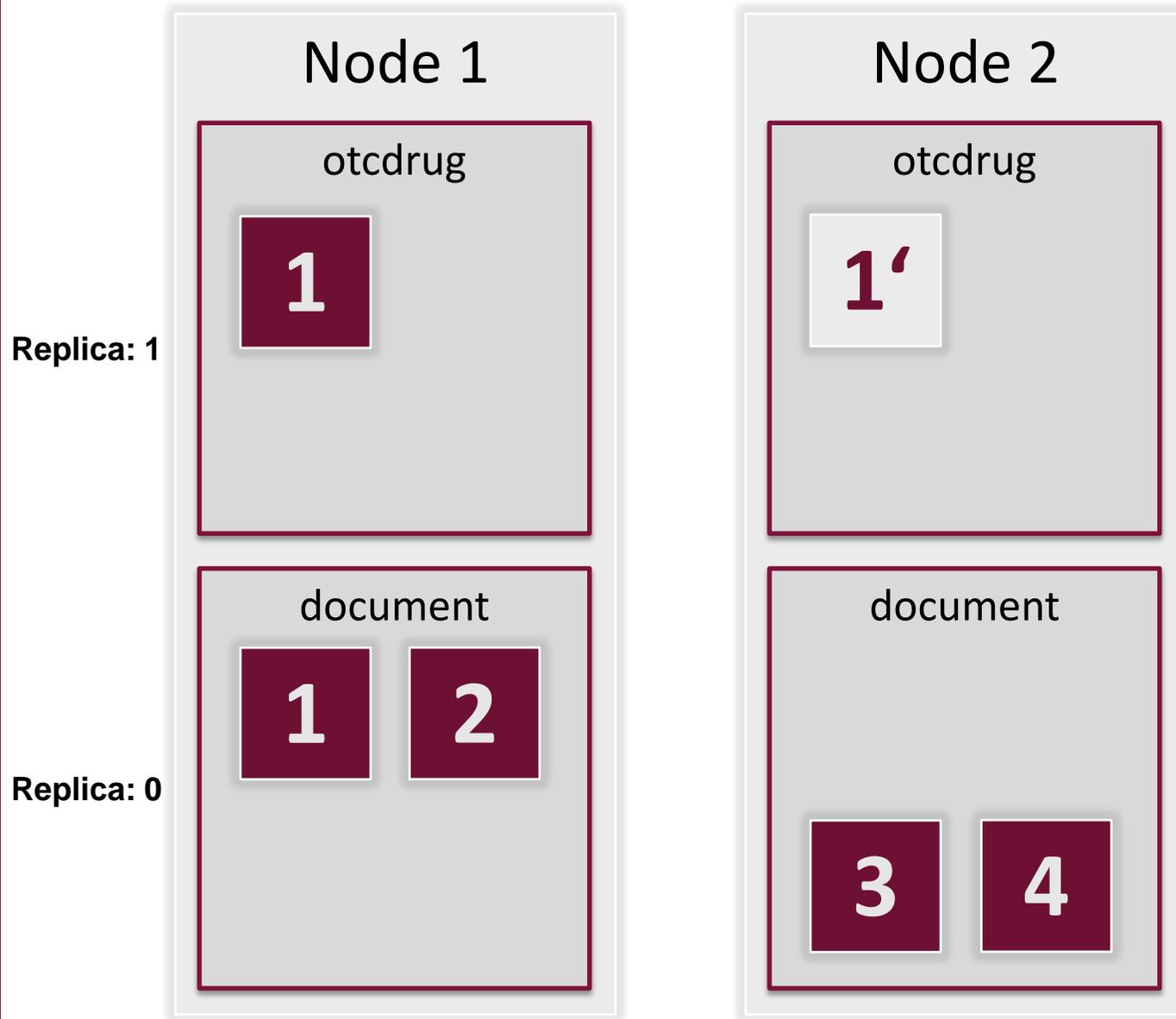
Große Datenmengen über Shards verwalten



Replica: 0

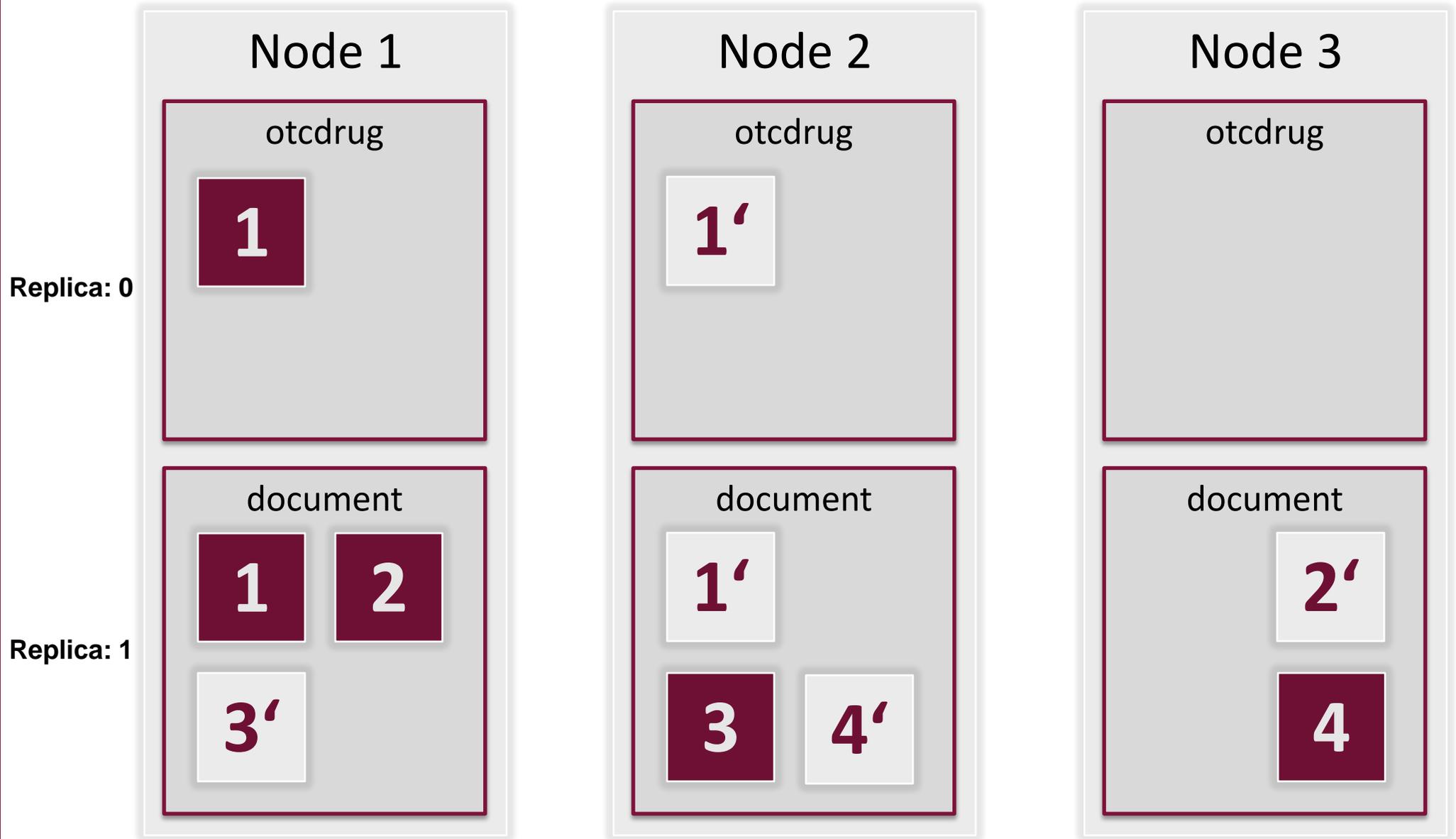
Replica: 0

Große Datenmengen über Shards verwalten



Neuer Knoten im Cluster

Große Datenmengen über Shards verwalten



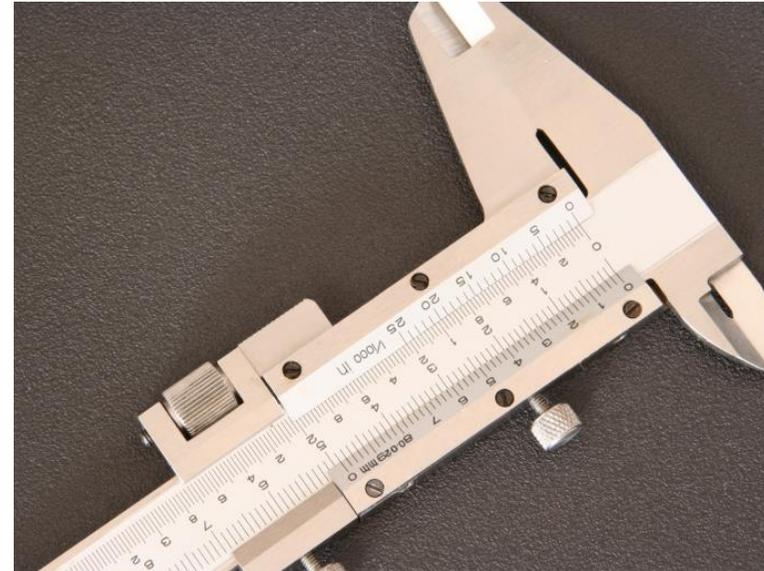
Aggregationen - Buckets und Metrics



<http://mrg.bz/IQNZFq>

Buckets

- Terme
- Ranges
- Histogramme
- Geo-Distanz



<http://mrg.bz/Nn57cJ>

Metrics

- Anzahl
- Summe
- Min / Max / Average
- Varianz
- Perzentile

Aggregationen für Analysen

```
1  POST /events/_search
2  {
3  "aggs" : {
4    "categories" : { "terms" : { "field" : "category" } },
5    "indications" : { "terms" : { "field" : "indication" } },
6    "locations" : { "terms" : { "field" : "location" } },
7    "countries" : { "terms" : { "field" : "country" } },
8    "created_range": {
9      "date_range": {
10       "field": "created",
11       "format": "yyyy-MM-dd'T'HH:mm:ss.SSSZ",
12       "ranges": [
13         { "from": "now-1M/M" },
14         { "from": "now-3M/M" },
15         { "from": "now-6M/M" }
16       ]
17     }
18   }
19 }
20 }
```

Aggregationen

```

1 POST /event
2 {
3   "aggs" :
4     "categ
5     "indic
6     "locat
7     "count
8     "creat
9     "d
10
11
12
13
14
15
16
17 }
18 }
19 }
20 }
15   "aggregations": {
16     "locations": {
17       "buckets": [
18         {
19           "key": "Berlin",
20           "doc_count": 626
21         },
22         {
23           "key": "Basel",
24           "doc_count": 93
25         },
26         .....
27       ]
28     },
29     "countries": {
30       "buckets": [
31         {
32           "key": "Germany",
33           "doc_count": 671
34         },
35         {
36           "key": "Switzerland",
37           "doc_count": 107
38         },

```

BI mit Elasticsearch

ETL

Speicherung / Berechnungen

Analyse



Logstash

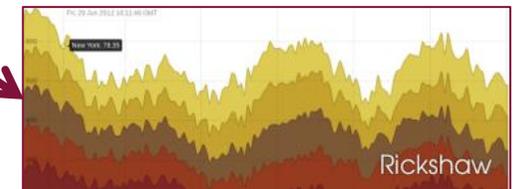
rivers

Implementierungen für

- JDBC
- CSV
- ...



Eigene Loader mit ES-Client (Bsp. SpringBoot)

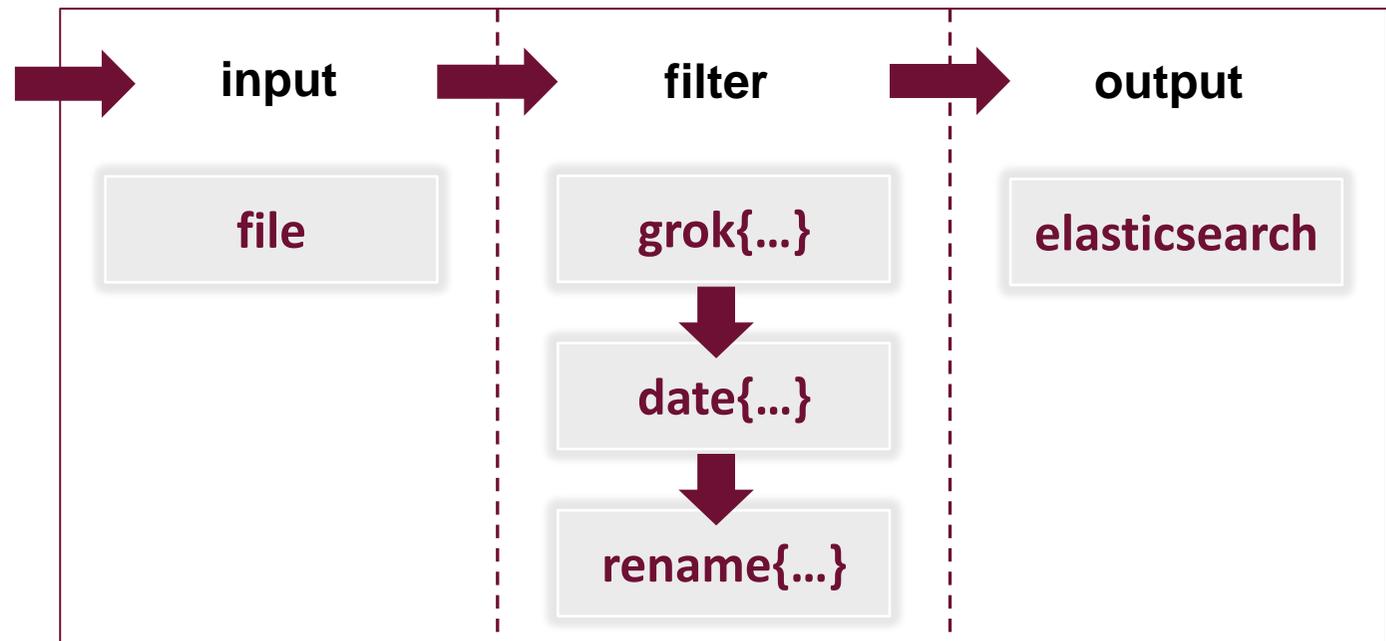


Eigene Visualisierung

Datentransport mit Logstash



- Event Processing Engine
- Optimiert für Log-Dateien
- Pipeline-Prinzip
 - Input (50+)
 - Filter (60+)
 - Output (75+)



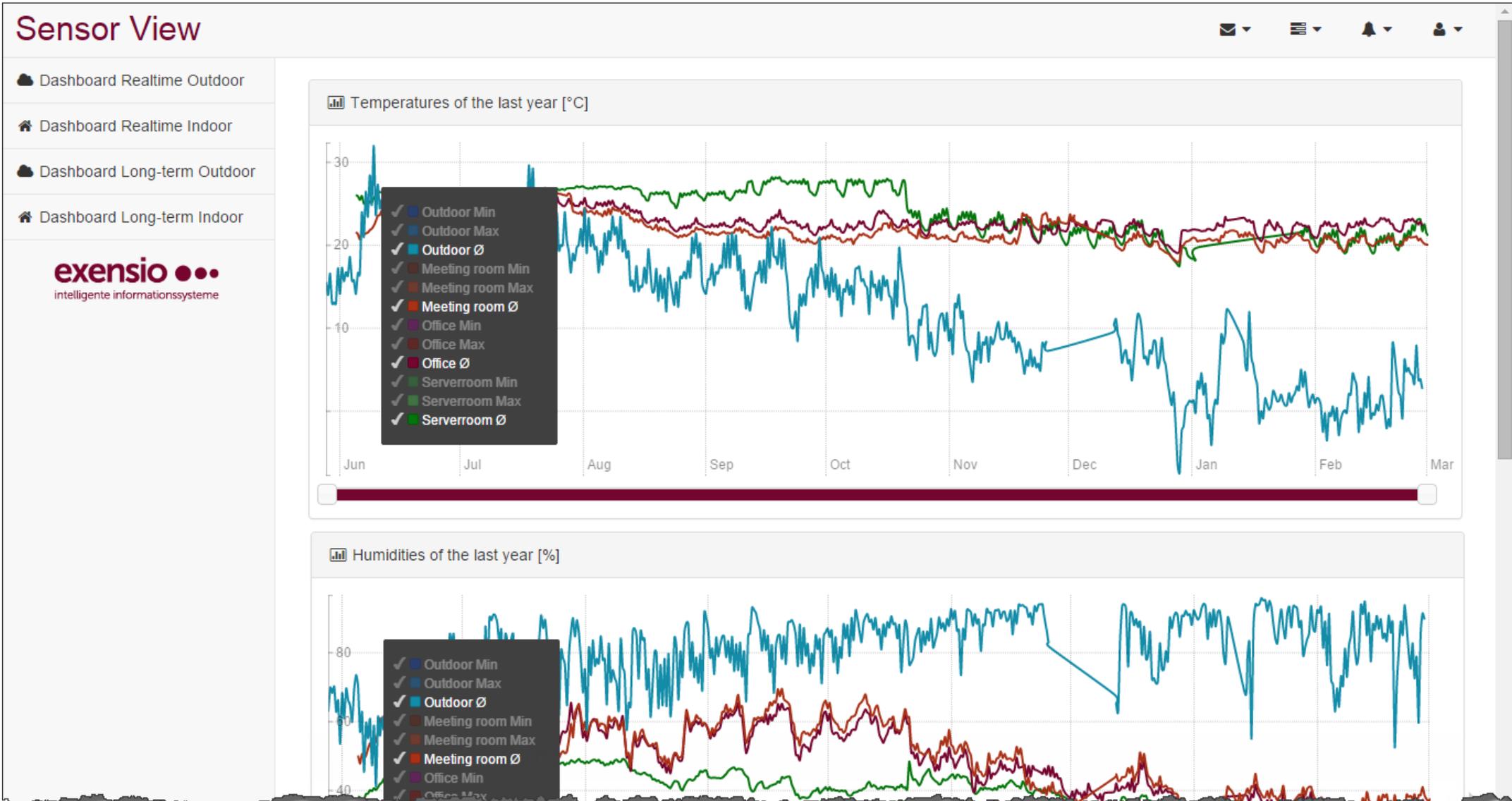
Visualisierungen mit Kibana



- Aktuelles Release: Kibana 4
- Browserbasierte Visualisierung von Daten
 - Abfragen über JSON an ES
- Aufbereitung über
 - Discover
 - Visualize
 - Dashboards

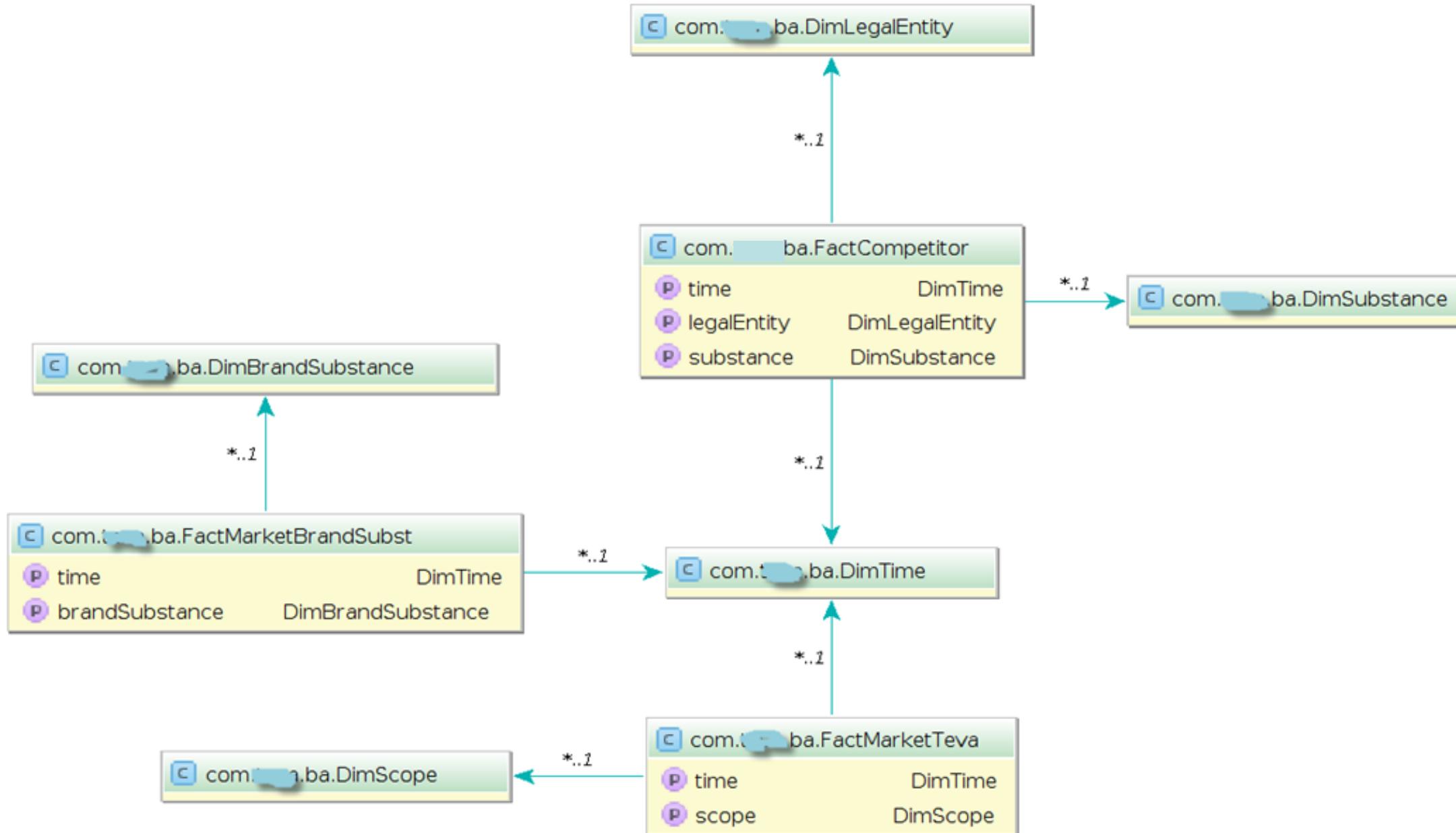


Eigene Analyse-Komponenten sind einfach zu erstellen



Darstellung der JSON-Rückgabewerte von ES auf Basis von JavaScript

Marktanalysen Pharmamarkt – Starschema



Marktanalysen Pharmamarkt – Suchmaschine

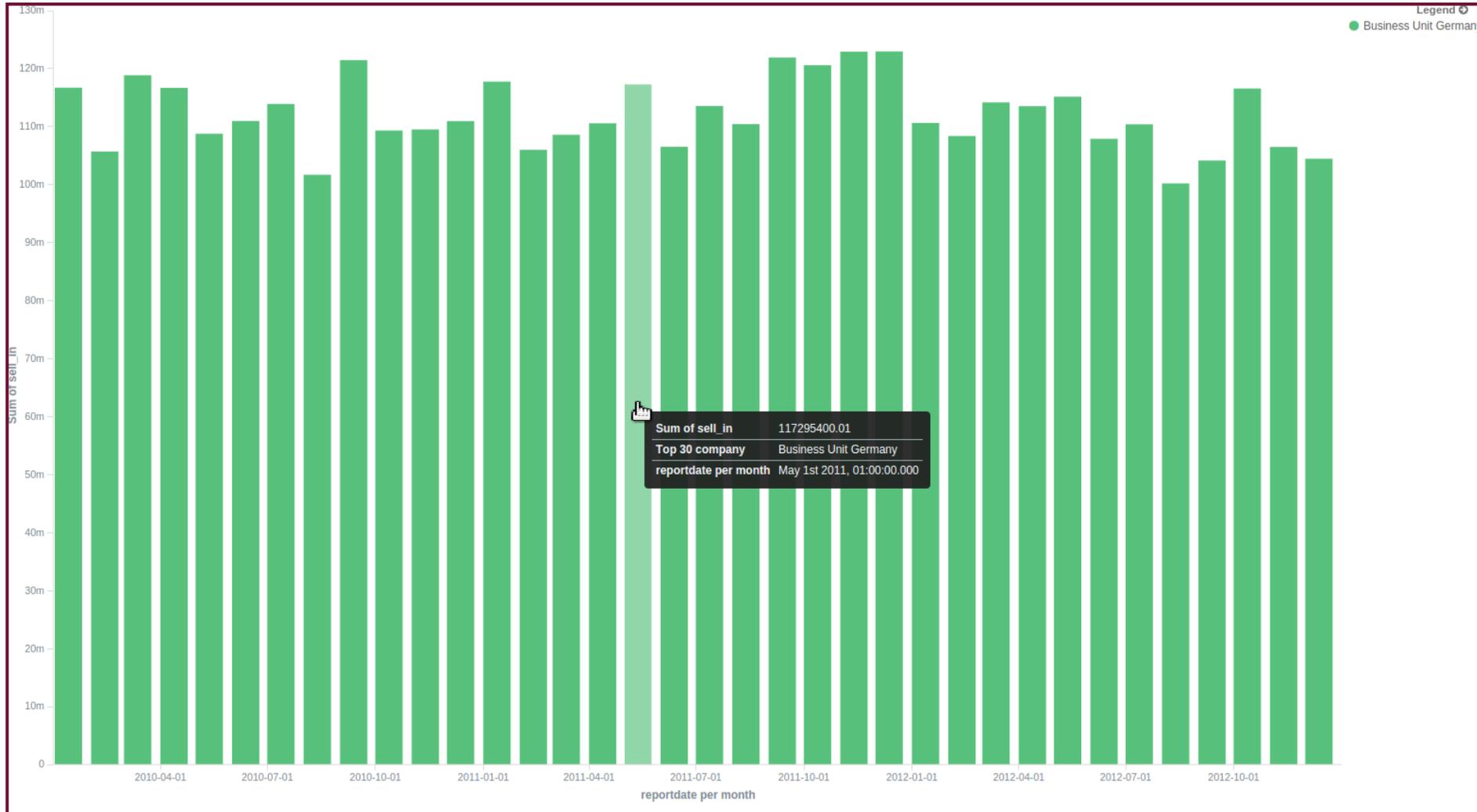
```
1 {  
2   "factmarket": {  
3     "mappings": {  
4       "market": {  
5         "properties": {  
6           "aut_idem": {  
7             "type": "double"  
8           },  
9           "bu": {  
10            "type": "string",  
11            "index": "not_analyzed"  
12          },  
13          "company": {  
14            "type": "string",  
15            "index": "not_analyzed"  
16          },  
17          "reportdate": {  
18            "type": "date",  
19            "format": "dateOptionalTime"  
20          },  
21          "id": {  
22            "type": "long"  
23          },  
24          "level": {  
25            "type": "long"  
26          },  
27          "line": {  
28            "type": "string",  
29            "index": "not_analyzed"  
30          },  
31          "market_gen_launches0": {  
32            "type": "double"  
33          },  
34          "market_gen_launches1": {  
35            "type": "double"  
36          },  
37          "market_launches0": {  
38            "type": "double"  
39          },  
40          "market_launches1": {  
41            "type": "double"  
42          },  
43          "market_sell_out": {  
44            "type": "double"  
45          },  
46          "market_vo_aut_idem": {
```

Marktanalysen Pharmamarkt – Suchmaschine

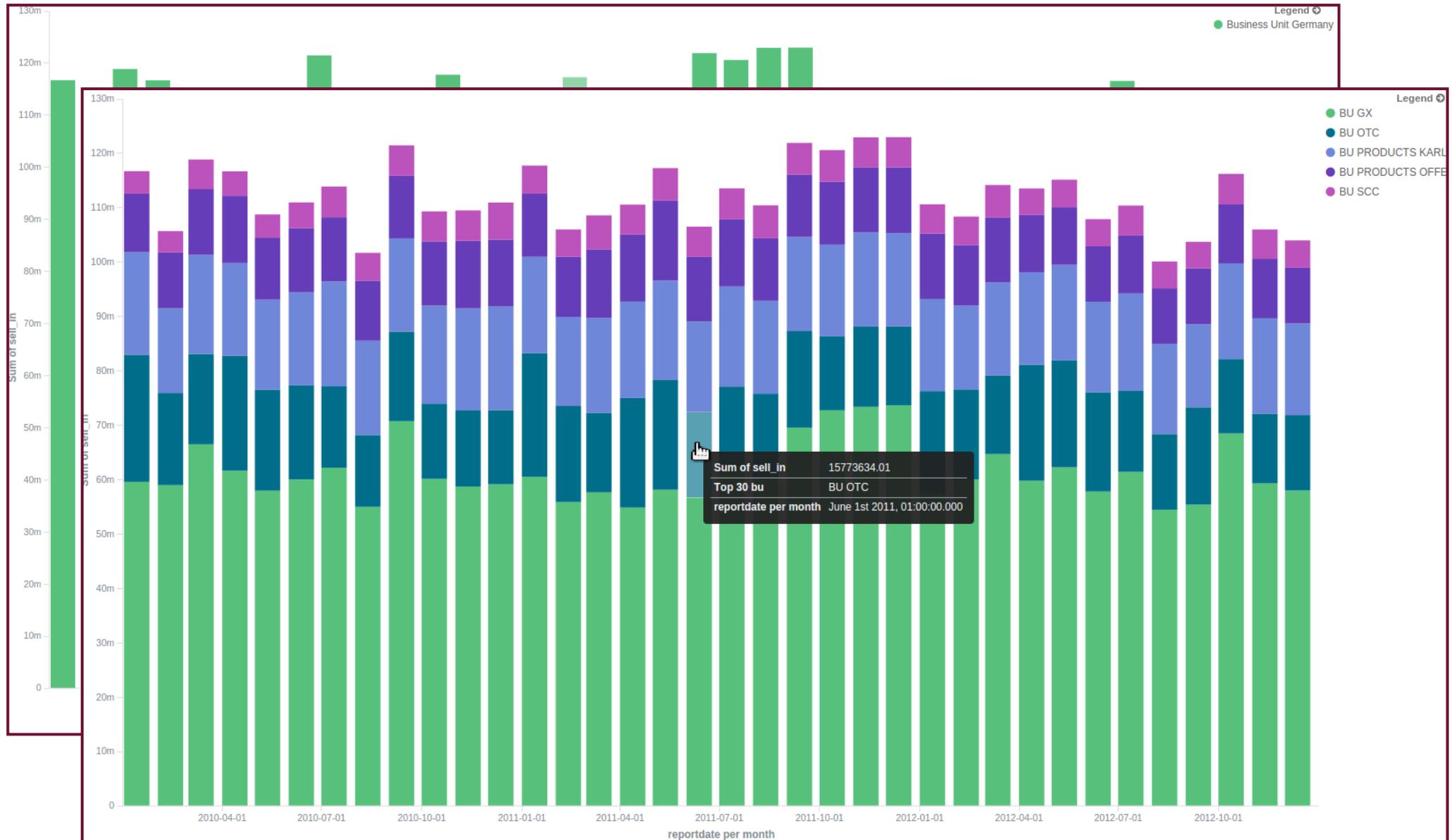


```
"parallel_imports": {  
  "type": "double"  
},  
"productih": {  
  "type": "string",  
  "index": "not_analyzed"  
},  
"productims": {  
  "type": "string",  
  "index": "not_analyzed"  
},  
"reportdate": {  
  "type": "date",  
  "format": "dateOptionalTime"  
},  
"revenue_discounted": {  
  "type": "double"  
},
```

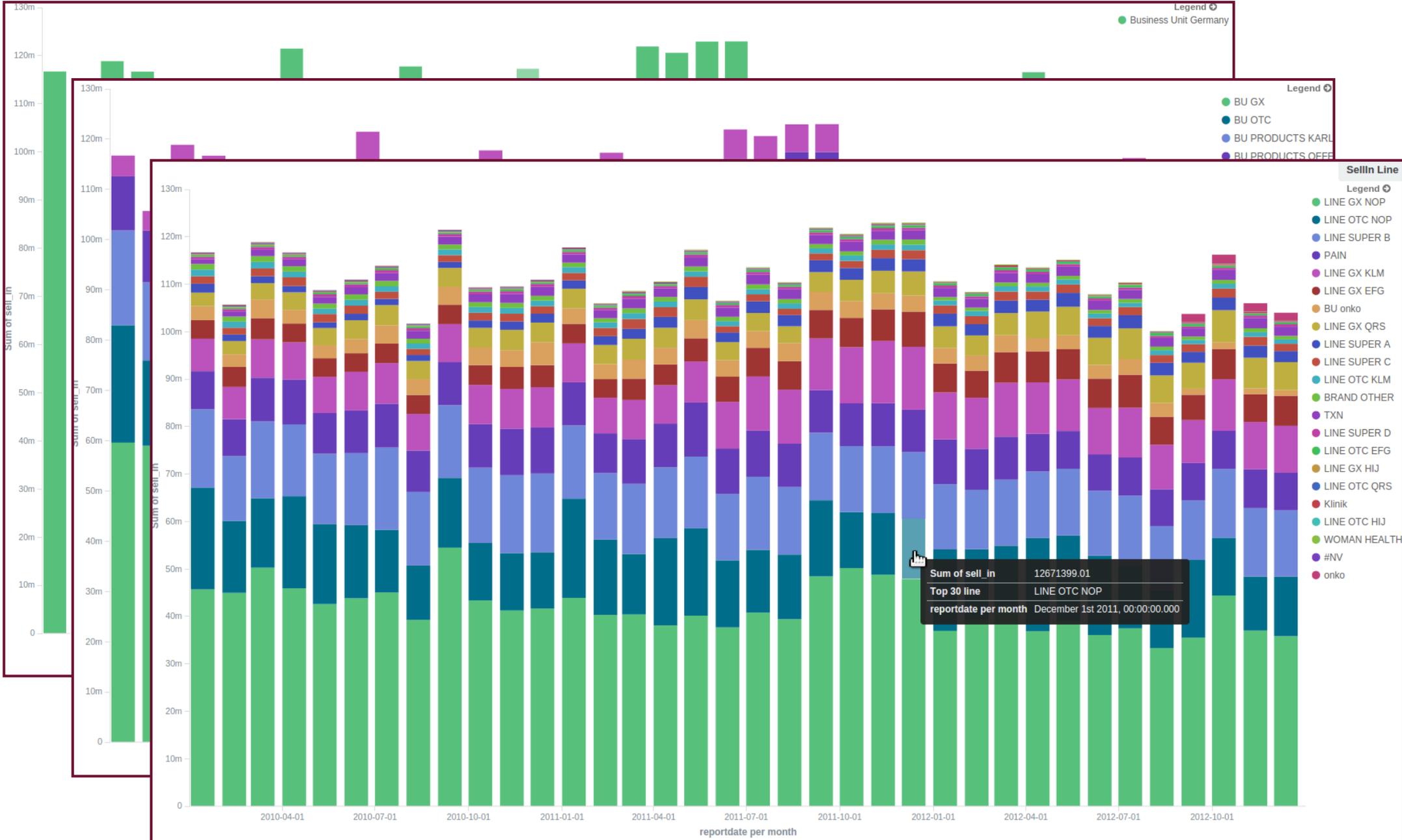
DrillDown des SellIn von Firma, Business Unit und Linie



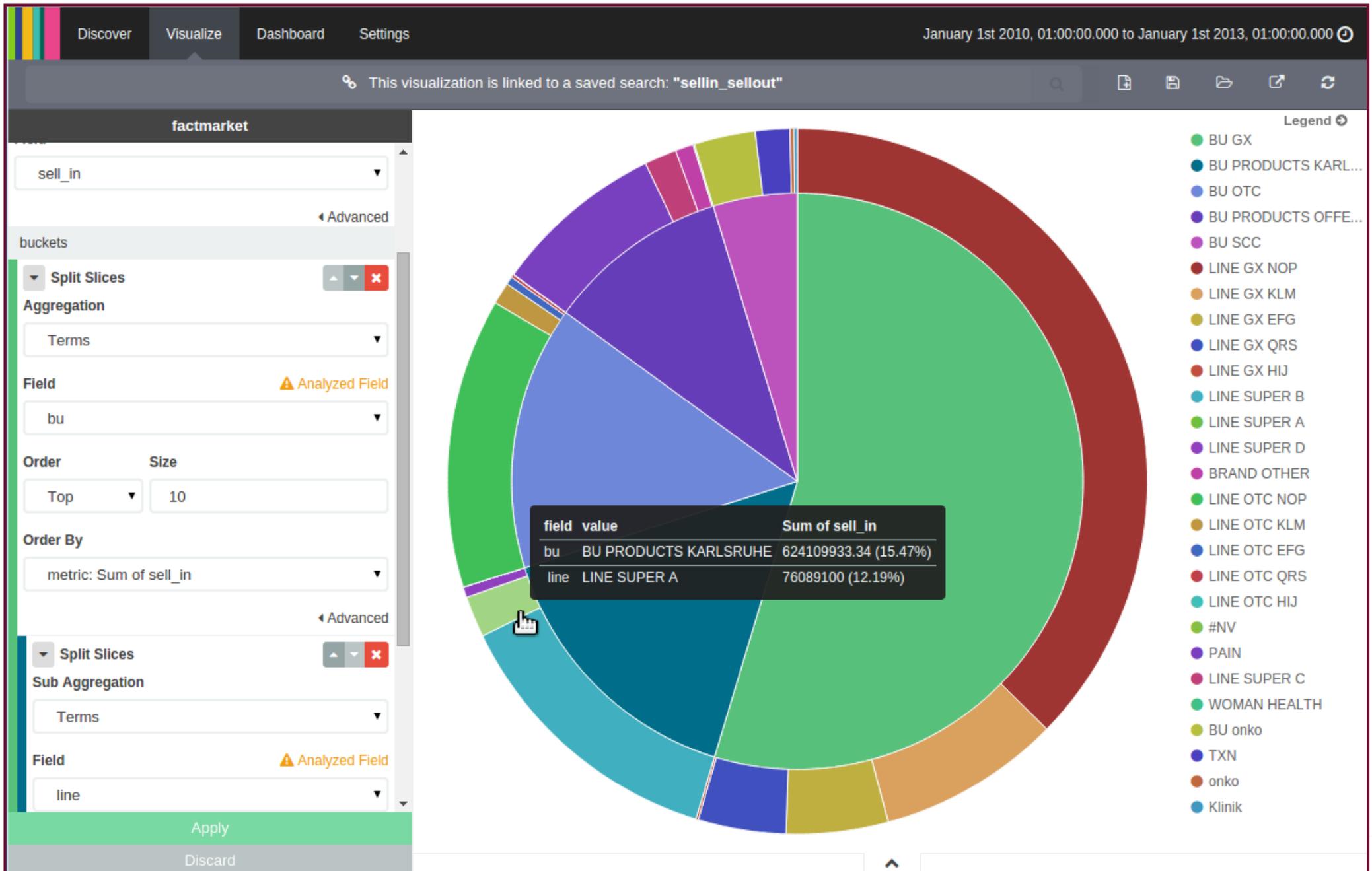
DrillDown des SellIn von Firma, Business Unit und Linie



DrillDown des SellIn von Firma, Business Unit und Linie



Donat-Diagramm für SellIn auf BU-und Linien-Ebene



Visualisierung von Aggregationen mit einer Tabelle

The screenshot shows the Kibana interface with a table visualization. The left sidebar contains the 'factmarket' visualization configuration, and the main area displays a table of aggregated data.

factmarket

metrics

- Metric: Sum of sell_in
- Metric: Average sell_in
- Metric: Sum of sell_out
- Metric: Average sell_out
- Metric: Sum of aut_idem
- Metric: Sum of market_vo_discounted

+ Add Aggregation

buckets

- Split Rows: reportdate per month

⌵ Add Sub Aggregation

view options ▶

This visualization is linked to a saved search: "sellin_sellout"

January 1st 2010, 01:00:00.000 to January 1st 2010, 01:00:00.000

reportdate per month	Sum of sell_in	Average sell_in	Sum of sell_out	Average sell_out	Sum of aut_idem	Sum of market_vo_discounted
January 1st 2010, 00:00:00.000	116734699.99	105071.737	118034600	106146.223	17808234	84383708
February 1st 2010, 00:00:00.000	105719299.99	95156.886	119628199.99	107579.317	15828266	83110111
March 1st 2010, 00:00:00.000	118879700	107002.43	135591199.99	121934.532	18449105	93359358
April 1st 2010, 01:00:00.000	116703599.99	105043.744	122794100	110426.349	18020430	84413940
May 1st 2010, 01:00:00.000	108780500.01	97912.241	117347800	105528.597	16855364	80295725
June 1st 2010, 01:00:00.000	110990000	99900.99	122783200.01	110416.547	17202901	84898494
July 1st 2010, 01:00:00.000	113909600.01	102528.893	129332899.99	116306.565	18646998	90992232
August 1st 2010, 01:00:00.000	101708600	91546.895	119142699.99	107142.716	16687787	82038710
September 1st 2010, 01:00:00.000	121488200.01	109350.315	129286200.01	116264.568	16767659	90829646
October 1st 2010, 01:00:00.000	109345699.99	98420.972	126259700.01	113542.896	17630365	88359374

Export: [Raw](#) [Formatted](#)

Competitive Intelligence als weiteres Einsatzszenario

Beispiel Shop/Handel: Mapping bspw. über den Weinnamen, falls keine eindeutige ID wie EAN Code vorhanden ist

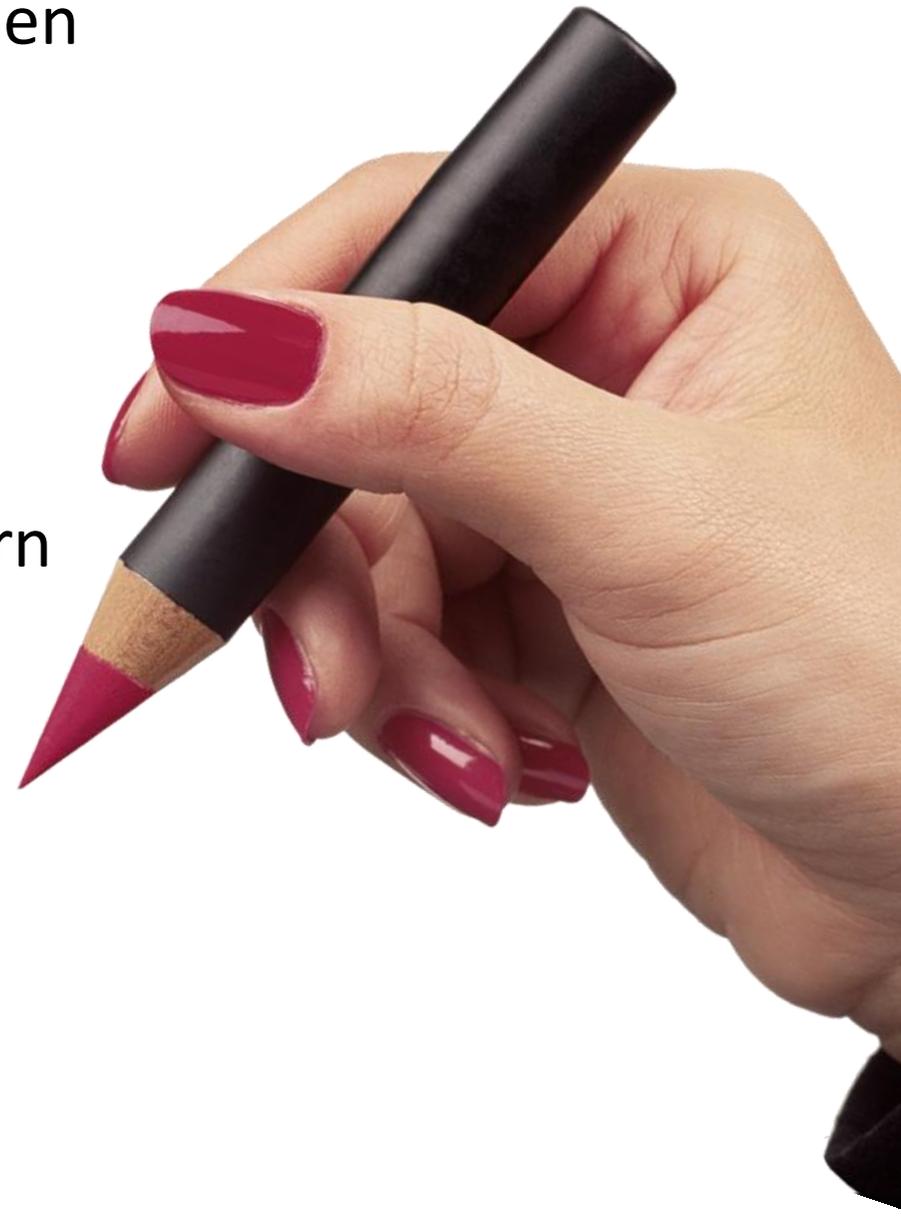
The screenshot shows a web interface for a wine shop. On the left, there are several filter sections: 'Händler' (Koelner-weinkeller.de (39)), 'Kategorie' (Rotwein (31), Weisswein (8)), 'Land' (Frankreich (39)), 'Anbauregion' (Rhône / Châteauneuf-du-Pape (39)), 'Weingut' (Château de Beaucastel (17), Chante Cigale (5), Château de Vaudieu (5), Vieux Telegraphe (5), Domaine Giraud (3), Domaine de la Janasse (2), Château La Nerthe (1), E. Guigal (1)), and 'Preise' (bis 10 Euro (0), 10-20 Euro (0), 20-50 Euro (13), 50-100 Euro (18)).

The main content area is divided into 'Suche' (Search) and 'Ergebnisse' (Results). The search section has two input fields: 'Exakte Suche' and 'Fuzzy Suche'. The results section shows a single product card for 'Châteauneuf-du-Pape blanc Chante Cigale - 2013'. The product card includes a bottle image, a price of 23.00 €, and a 'sofort lieferbar' status. The product details are as follows:

Etikette		Artikel	Châteauneuf-du-Pape blanc Chante Cigale - 2013
Kategorie	Weisswein	Land	Frankreich
Anbauregion	Rhône / Châteauneuf-du-Pape	Weingut	Chante Cigale
Geschmack	trocken	Stil	
Jahrgang	2013	Alkohol	14 %
Rebsorte	Grenache blanc, Rousanne, Bourboulenc, Clairette		
Preis	23.00 €	Artikelnummer	20024-13
Lieferstatus	sofort lieferbar	Lagerbestand	
Shop Name	REWE / Kölner Weinkeller	ArtikelURL	Link zur Artikelseite
Beschreibung	Wunderbare Weine sind die Weissen von der südlichen Rhône. Nur in Deutschland leider fast unbekannt. Mit wenig Säure und ihren nussigen und reifen Fruchtnoten erinnern sie ein wenig an große Burgunder. Doch zeigen sie auch eine spannende Kräuterwürze und feine Früchte, die nach einigen Jahren Reife perfekt balanciert sind.		
Shop URL		Shop URL	koelner-weinkeller.de

Fazit

- Aggregationen ersetzen Dimensionen im Star-Schema
- Near Realtime
- Leichtgewichtig
- Kostengünstige Alternative
- Einfach zu integrieren und erweitern
- Kombinierbar mit Vorteilen von Suchmaschinen
 - Matching
 - Unstrukturierte Daten





Lösungen für individuelle Prozesse

Fragen?
Vielen Dank!

 <http://blog.exensio.de>

 @tokraft

Partner:

 elasticsearch.

ORACLE
PARTNERNETWORK