

Solr Grundlagen

20.11.2015, Marcel Becker

© Copyright 2013 anderScore GmbH

1. Was ist Solr?

2. Grundlagen Suchen

3. Schema

4. Solr Features / Solr Cloud

5. Hands On

6. Fragen

7. Links



- Software Entwickler (Dipl.-Inf.) bei der anderScore GmbH, Köln
- Einige Jahre Erfahrung im Umfeld
 - Java
 - Web
 - Agile Methoden

idealo.de

IMMOBILIEN
SCOUT 24

MyHammer

PARSHIP
Deutschlands größte Partnervermittlung

trivago®

Google
Deutschland

YAHOO!



WIKIPEDIA
Die freie Enzyklopädie

10 Jahre
weg.de
Gut beraten, besser erholt.

mobile.de

ebay

bing

amazon

You Tube

- Wenig Struktur
- Natürliche Sprache
- Viele Treffer
- Große Datenmengen

High Performance

Index & Search



Apache Software License

Java Bibliothek

Stand-alone Server

HTTP / XML /
JSON / CSV



Apache Software License

Cluster / Cloud

Grundlagen

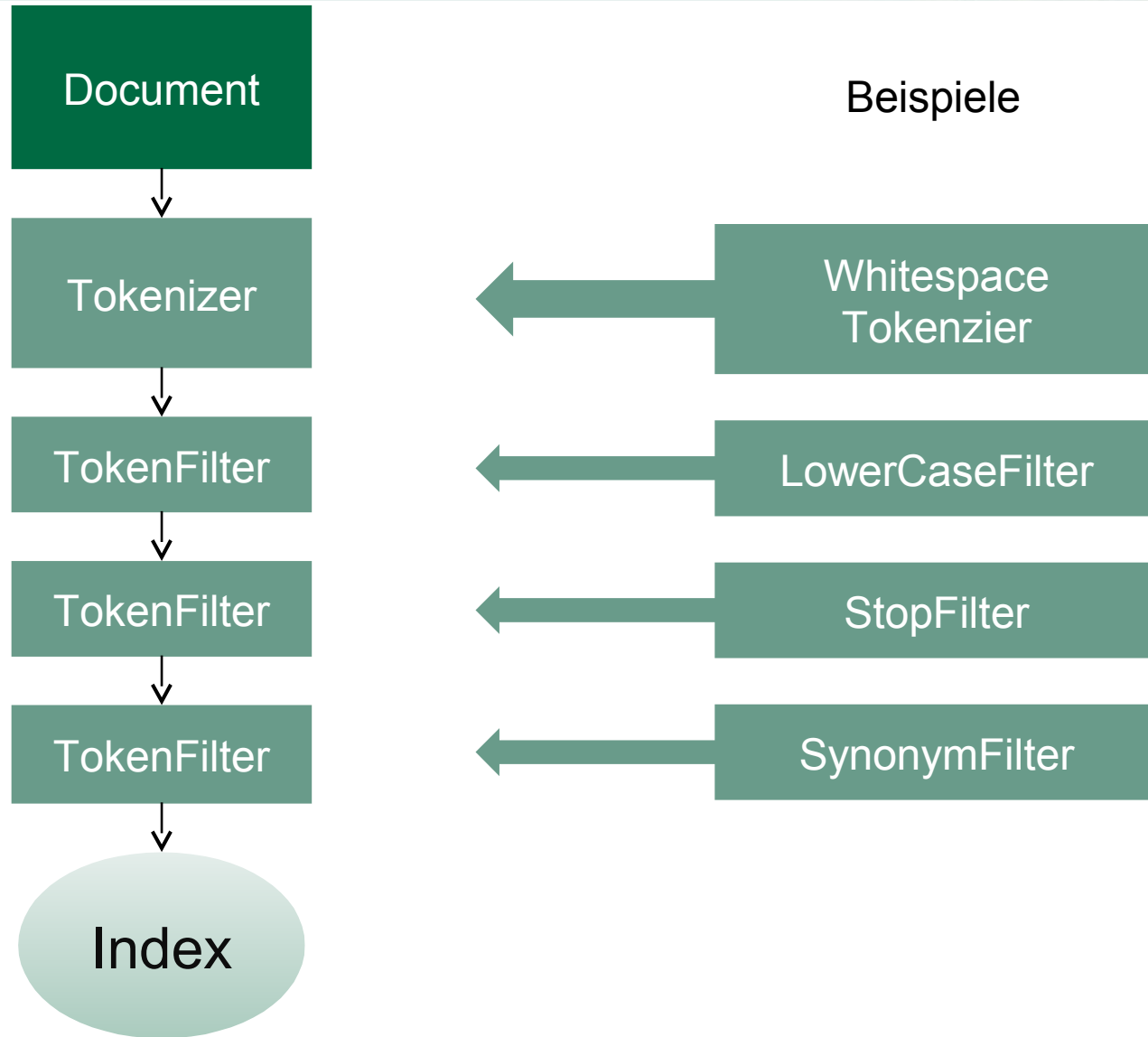
● Dokumente

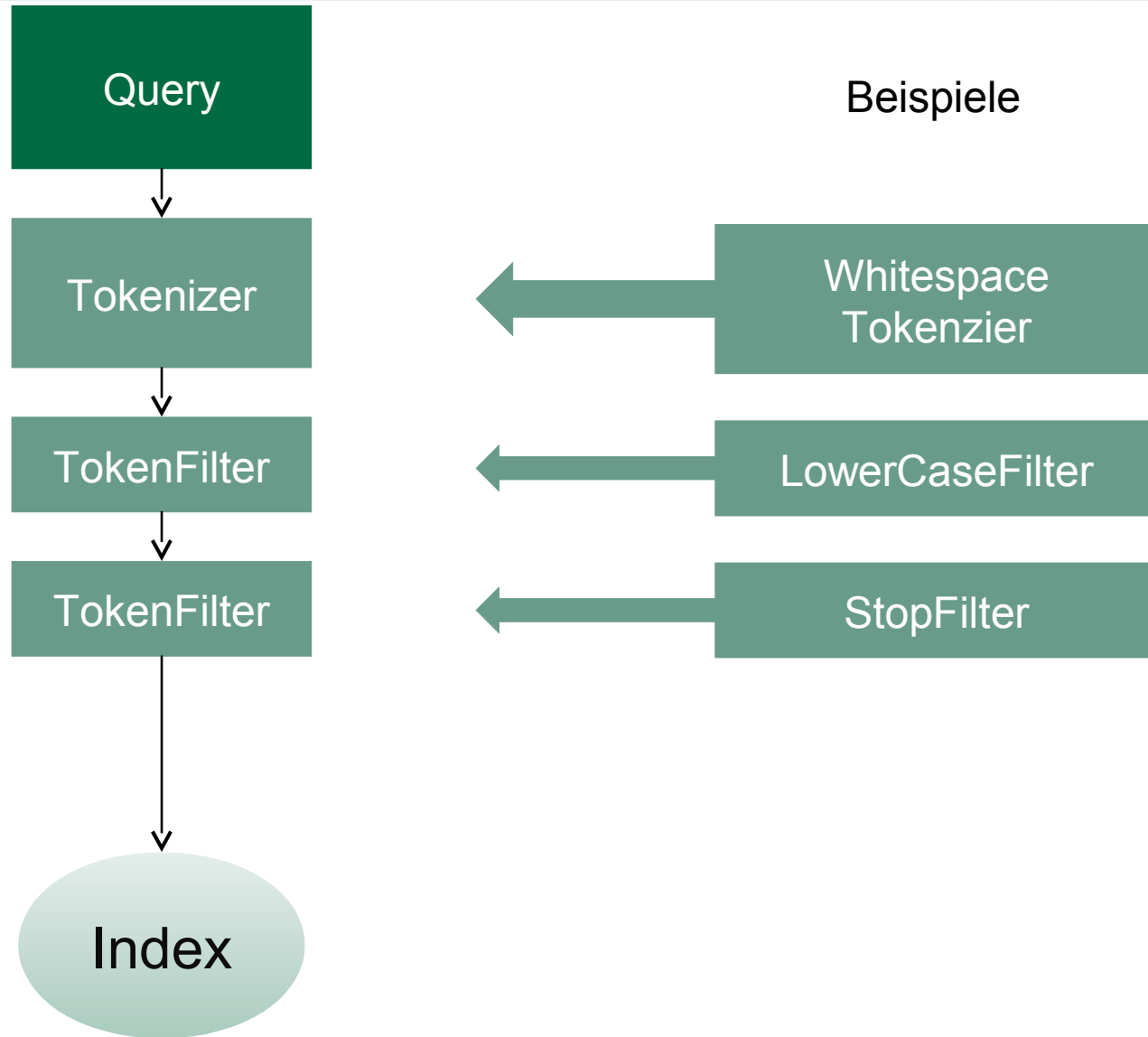
- Dokument 1: Die Hauptstadt von Frankreich heißt Paris.
- Dokument 2: Paris ist die Hauptstadt von Frankreich.
- Dokument 3: Die Hauptstädte von England bzw. Frankreich heißen London bzw. Paris.

● Invertierter Index:

- | | |
|-------------------------|--------------------|
| ● bzw. -> 3 | ● heißen -> 3 |
| ● die -> 1, 2, 3 | ● heißt -> 1, 2 |
| ● England -> 3 | ● ist -> 2 |
| ● Frankreich -> 1, 2, 3 | ● London -> 3 |
| ● Hauptstadt -> 1, 2 | ● Paris -> 1, 2, 3 |
| ● Hauptstädte -> 3 | ● von -> 1, 2, 3 |

Quelle: http://wikis.gm.fh-koeln.de/wiki_ir/InformationRetrieval/Invertierte-Liste







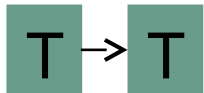
Dokument: Die Froscon findet in der Hochschule Bonn-Rhein-Sieg statt.



Tokens: froscon, findet, hochschule, fh, bonn-rhein-sieg, brs, statt



Query: FH BRS FrosCon 2015



Tokens: fh, brs, froscon, 2015



Match: fh, brs, froscon

- **tf** - Term Frequency
- **idf** - Inverse Document Frequency
- **coord** - Coordination Factor
- **fieldNorm** - Field length
- **distance** - bei Fuzzy Search
- **boost**
 - **Index**
 - **Query:** `title:"foo bar"^10`

```
<fieldType name="text_general" class="solr.TextField">
  <analyzer type="index">
    <charFilter class="solr.MappingCharFilterFactory" mapping="ascii.txt"/>
    <tokenizer class="solr.WhitespaceTokenizerFactory" />
    <filter class="solr.LowerCaseFilterFactory" />
    <filter class="solr.StopFilterFactory" words="stopwords.txt" />
    <filter class="solr.SynonymFilterFactory" synonyms="synonyms.txt" />
  </analyzer>
  <analyzer type="query">
    <charFilter class="solr.MappingCharFilterFactory" mapping="ascii.txt"/>
    <tokenizer class="solr.WhitespaceTokenizerFactory" />
    <filter class="solr.LowerCaseFilterFactory" />
    <filter class="solr.StopFilterFactory" words="stopwords.txt" />
  </analyzer>
</fieldType>

<field name="text" type="text_general" indexed="true" stored="true" />

<dynamicField name="*_txt" type="text_general" indexed="true" stored="true" />
```

● Java API

```
String collection = "anderscore";

Book book = new Book();
book.id = "1";
book.author_txt = "Eric Ries";
book.title_txt = "The Lean Startup";

Address address = new Address();
address.id = "2";
address.addressLine_txt = "Grantham-Allee 20";
address.postalCode_txt = "53757";
address.city_txt = "Sankt Augustin";
address.country_txt = "Germany";

CloudSolrClient c = new CloudSolrClient("localhost:2001");
c.addBean(collection, book);
c.addBean(collection, address);

c.close();|
```

● Admin UI

Solr

- Dashboard
- Logging
- Cloud
- Core Admin
- Java Properties
- Thread Dump
- anderscore_sh...
- Overview
- Analysis
- Dataimport
- Documents
- Files
- Ping
- Plugins / Stats
- Query

Request-Handler (qt)
/select

— common —

q
:

fq
[]

sort
[]

start, rows
0 10

fl
[]

df
[]

Raw Query Parameters
key1=val1&key2=val2

wt
json

☒ indent

http://localhost:1001/solr/anderscore_shard1_replica1/select?q=*&3A

```
{
  "responseHeader": {
    "status": 0,
    "QTime": 87,
    "params": {
      "q": ":*:*",
      "indent": "true",
      "wt": "json",
      "_": "1439148187292"
    }
  },
  "response": {
    "numFound": 0,
    "start": 0,
    "docs": []
  }
}
```


- Geospatial Support



Quelle: Google Maps

● Facets:

The screenshot displays a web interface for a search application. On the left, a facet menu titled 'Kategorie' is highlighted with a green border. It includes a link '» alle anzeigen' and three radio buttons for 'Bücher (29)', 'Filme (7)', and 'Musik (2)'. The main content area shows a navigation bar with 'Zurück' and 'Home' links. Below this, a search result summary states: 'Ihre Suche nach Per Anhalter durch die Galaxis ergab 38 Treffer.' At the bottom of the main area, there is a dropdown menu set to 'Beliebtheit' and a checkbox labeled 'nur verfügbare Artikel' which is currently unchecked.

Kategorie

- » alle anzeigen
- ☐ Bücher (29)
- ☐ Filme (7)
- ☐ Musik (2)

Navigation: [Zurück](#) | [Home](#)

Ihre Suche nach **Per Anhalter durch die Galaxis** ergab 38 Treffer.

Beliebtheit ▼ ☐ nur verfügbare Artikel



Advanced Full-Text Search Capabilities

Powered by Lucene™, Solr enables powerful matching capabilities including phrases, wildcards, joins, grouping and much more across any data type.



Optimized for High Volume Traffic

Solr is proven at extremely large scales the world over



Standards Based Open Interfaces - XML, JSON and HTTP

Solr uses the tools you use to make application building a snap



Comprehensive Administration Interfaces

Solr ships with a built-in, responsive administrative user interface to make it easy to control your Solr instances



Need more insight into your instances? Solr publishes loads of metric data via JMX



High Scalability and Fault Tolerant

Built on the battle-tested Apache Zookeeper, Solr makes it easy to scale up and down. Solr bakes in replication, distribution, rebalancing and fault tolerance out of the box.



Flexible and Adaptable with easy configuration



Near Real-Time Indexing

Want to see your updates now? Solr takes



Extensible Plugin Architecture

Quelle: <http://lucene.apache.org/solr/features.html>

- Menge von Features, die helfen einen Index horizontal zu skalieren (verteilen):
 - Sharding
 - Replikation
- Ziele:
 - Skalierbarkeit
 - Performance
 - Hoch-Verfügbarkeit
 - Einfachheit / Komfort

- ZooKeeper
 - Zentralisierte Konfiguration
 - Cluster-State Management
 - Leader Election
- Node
- Core
- Collection
 - Shard
 - Replication Factor
 - Hash Range
 - Leader
- Replica
- Leader

Quelle: <http://de.slideshare.net/thelabdude/solr-exchange-introtosolrcloud>

Hands On



Danke

- <http://lucene.apache.org/solr/>
- <http://lucene.apache.org/index.html>
- http://wikis.gm.fh-koeln.de/wiki_ir/InformationRetrieval/Invertierte-Liste
- <http://de.slideshare.net/thelabdude/solr-exchange-introtoSolrcloud>

Anhang

- Classic Solr

Solr Core

Index Files

Core
Configuration
Files

- Solr Cloud (Logical View)

Solr Collection

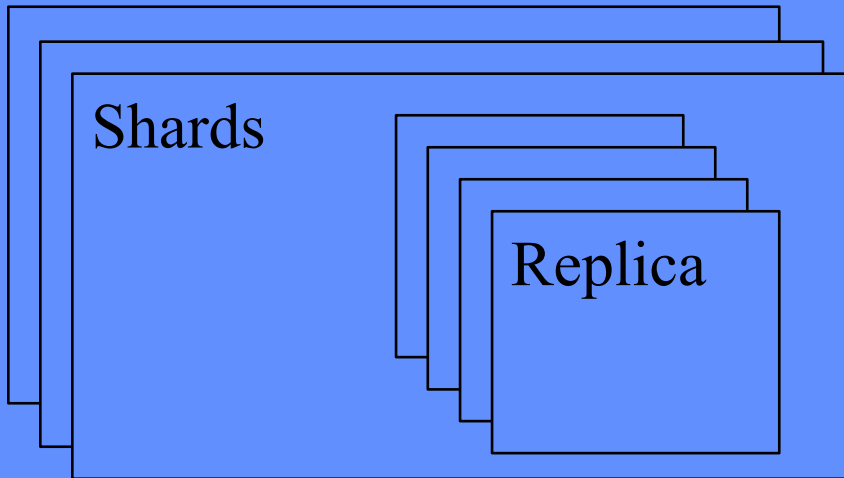
Shards

Replica

Config Set

- Solr Cloud (Physical View)

Solr Collection



Cluster

