SEARCH **ANALYTICS** BIG DATA

Consulting • Software • Development • Training

CUSTOM SOLR TOKENIZER
FLEXIBLE TOKENIZER WITH JFLEX

2014 BerlinBuzzword

Agenda

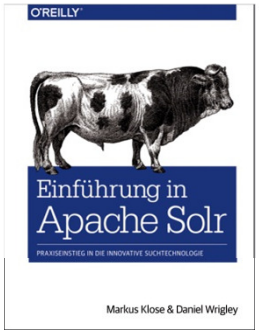
- ME & SHI
- JFLEX Tokenizer
  - Motivation
  - JFlex ?!
  - Solr implementation
  - Demo
- Q & A

ME & SHI

Markus Klose – Search Consultant

- Expertise in Solr, Lucene, Elasticsearch, Fast ESP
- Certified Apache Solr Trainer

- Speaker, Blogger, Coder
- Author “Einführung in Apache Solr”



- @markus_klose

SHI GmbH & Co KG

2014 Delivering mission-critical data-driven solution for multiple industries.

2013 Partnering with Smartlogic™ | THE CONTENT INTELLIGENCE COMPANY

2011 Partnering with LucidWorks LucidWorks™

2000 Embracing Open Source.

1994 Foundation. Development of home-grown information retrieval platform.

OUR MISSION

Vendor-independent **IT Consulting and Software Engineering** company.

Dedicated to deliver next generation **Semantic Search, Big Data and Exploratory Data Analytics** solutions.

Using **Enterprise Data Hub** approach for 360° data integration.

And helping customers to **Accelerate (e)Business** through better technology adoption and data utilization.

WHAT WE DO

WITH SERVICE OF EXCELLENCE

- Strategy Consulting
- Technical Consulting
- Architecture Review
- Development Support
- Team Enablement through Workshops and Trainings
- Technology Comparison
- Tuning & Troubleshooting
- Migration Services
- Experts to Hire
- Service Level Agreements

AND WELL-ESTABLISHED PRODUCTS

- Apache Solr/Lucene
- Elasticsearch
- Kibana
- Logstash
- Apache Mahout
- Apache Hadoop, Pig, Hive
- LucidWorks Search
- LucidWorks Search Big Data

SKILLS AND KNOW-HOW APPLIED

- Software Architecture
- Coding Services for Java, C++/C, .NET, PHP for multiple OSs.
- Continuous Integration and Test Driven Development
- Managing Software Project Lifecycle

HELPING CLIENTS TO DEVELOP COMPETITIVE DATA-DRIVEN SOLUTIONS

- Explorative Data Analytics
- Commerce Search
- Identity Search
- Knowledge Management
- Intranet Portal Search
- Call Center Search
- Cyber Security
- Website Search
- Fraud Detection
- Governance and Compliance

CONTACT

SHI GmbH & Co KG
Curt-Frenzel-Str. 12
86167 Augsburg
Germany
info@shi-gmbh.com
+49.821.74 82 633 0

@SHIEngineers

mma@shi-gmbh.com

mk@shi-gmh.com

@markus_klose

dwr@sgi-gmbh.com

@wrigley_dan

SEARCH **ANALYTICS** BIG DATA

Consulting • Software • Development • Training

CUSTOM TOKENIZER WITH JFLEX

JFlex based tokenizer - the idea is not new, but great

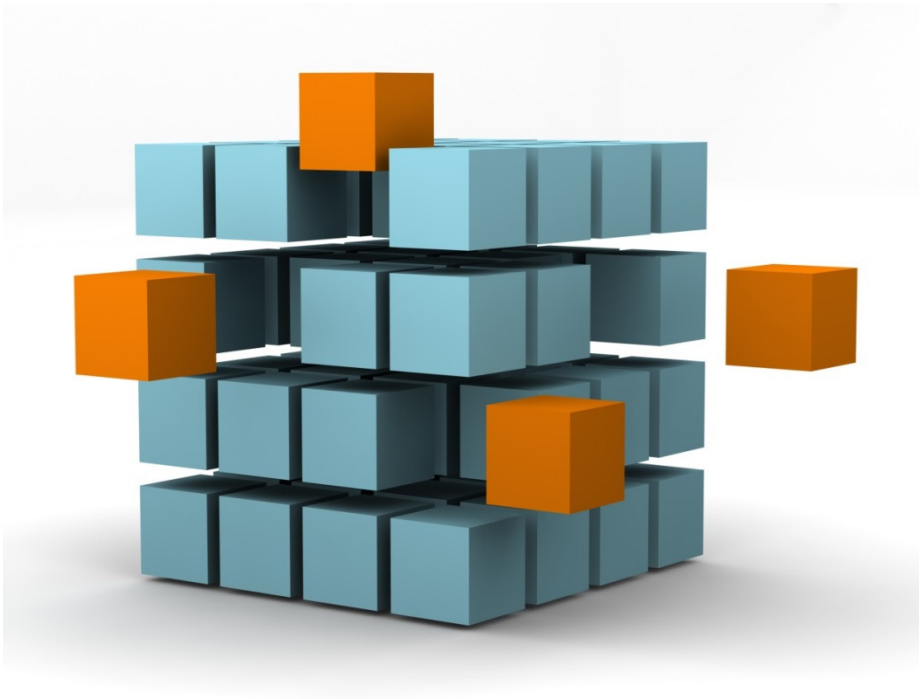
Motivation 1

- In customer projects we have to deal very often with custom „meta“ data
  - IDs
  - Type designation
  - Product description

- How to face that problem? PatternTokenizer?

Motivation 2

- Use and combine existing tools to be more flexible
- Configuration over Coding
- JFlex allready used in ClassicTokenizer / StandardTokenizer



UseCase – Type designation

- Product Data
  - nymj3x1,5 / nym-j 3x1,5 / nymj 3x1,5 / nym-j 3 x 1,5
- Search Input
  - nymj 3 1,5 / nym-j 3x1,5
- Index
  - nymj315 / nymj / nym / j / 315 / 3 / 15

JFlex - The Fast Scanner Generator

- JFlex is a lexical analyzer generator (aka scanner generator)

- Current version 1.5.1
- Download - http://jflex.de/download.html
- Mailing Lists
- BSD-style license
- CLI API & GUI

JFlex - The Fast Scanner Generator

- Berlin Buzzword 26.05.2014

- *LETTERS* -> „Berlin“, „Buzzword“
- *LETTERS* and SPACE -> „Berlin Buzzword“
- *DIGITS* -> „26“, „05“, „2014“
- *DIGITS* and . -> „26.05.2014“

- *LETTERS* and *SPACE* ***or*** *DIGITS* and .

  -> „Berlin Buzzword“ , „26.05.2014“

Custom Tokenizer – Project Setup

- JAVA - TokenizerFactory

  –> typical factory, tokenizer configuration

- JAVA - Tokenizer

  -> base class, token manipulation

- JFLEX – Scanner

  -> description of token patterns

- (JAVA – Scanner)

  -> Generated scanner

Demo

ISBN Tokenizer / URL Tokenizer

https://github.com/scherziglu

Resources

- JFlex Tokenizer
  - GitHub (https://github.com/scherziglu)
  - Solr Source Code (e.g. ClassicTokenizer)
  - @markus-klose / @SHIEngineers

- JFlex Websites

  http://jflex.de/

- Q & A

CONTACT

SHI GmbH & Co KG
Curt-Frenzel-Str. 12
86167 Augsburg
Germany
info@shi-gmbh.com
+49.821.74 82 633 0

@SHIEngineers

mma@shi-gmbh.com

mk@shi-gmh.com

@markus_klose

dwr@sgi-gmbh.com

@wrigley_dan