



Effizientes Suchen mit Apache Solr

Java Forum Stuttgart 2023

Matthias Graf | 13. Juli 2023 | Stuttgart

searching "hot dog"

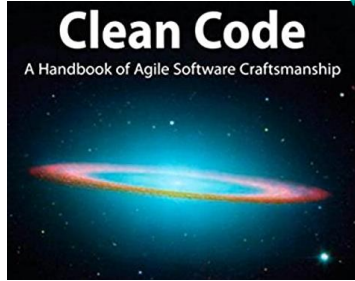
google:



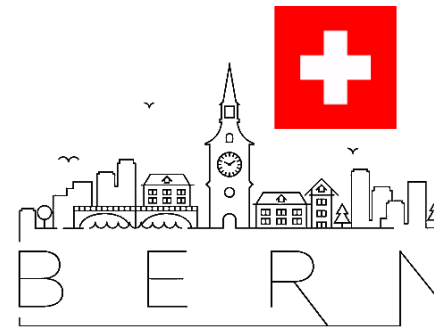
bing:



from https://www.reddit.com/r/dankmemes/comments/pi2aj3/why_is_my_search_engine_called_e621/



About Me



Übersicht

Um was geht es heute?

Inhalt

- Was ist “Suche”? Und wieso ist das wichtig?
- Was ist Solr und wie funktioniert es?
- Wie kann ich Solr in eine Applikation integrieren?

Out-Of-Scope

- Suche in Binärdateien
- Schema Optimierungen
- Performance und Load Balancing
- Query-Languages

Ziel

- Wissen wann Solr sinnvoll eingesetzt werden kann
- Wissen wie Solr Integriert werden kann

Was ist Suche?

Fiitern oder Suchen?

Solr Example

Search

All Games

ID	Title	Release Date	Team	Genres
0	Elden Ring	2022-02-25	Bandai Namco Entertainment, FromSoftware	Adventure, RPG
1	Hades	2019-12-10	Supergiant Games	Adventure, Brawler, Indie, RPG
2	The Legend of Zelda: Breath of the Wild	2017-03-03	Nintendo, Nintendo EPD Production Group No. 3	Adventure, RPG
3	Undertale	2015-09-15	tobyfox, 8-4	Adventure, Indie, RPG, Turn Based Strategy
4	Hollow Knight	2017-02-24	Team Cherry	Adventure, Indie, Platform
5	Minecraft	2011-11-18	Mojang Studios	Adventure, Simulator
6	Omori	2020-12-25	OMOCAT, PLAYISM	Adventure, Indie, RPG, Turn Based Strategy
7	Metroid Dread	2021-10-07	Nintendo, MercurySteam	Adventure, Platform
8	Among Us	2018-06-15	InnerSloth	Indie, Strategy
9	NieR: Automata	2017-02-23	PlatinumGames, Square Enix	Brawler, RPG

Showing 1 to 10 of 1512 rows

10

rows per page

<

1

2

3

4

5

...

152

>

Filtern

Suchen

Binary Data
Updates
Performance
Autocomplete
Imports
Highlighting
Features
Faceting
Clustering

Avoid Navigation
Simplicity
Full-Text
Intuitive
Suggestions
Usability
Fast
Mobile-Friendly
Stemming
Context
Wildcards
Relevance
Spelling
Similarity
Ordering
Phonetic Matching
Synonyms
Distance

Wie bauen wir eine gute Suchfunktion?

- Für eine gute Suche braucht **spezielle Tools**
 - Eine „LIKE '%\$1%'“ SQL-Query reicht nicht

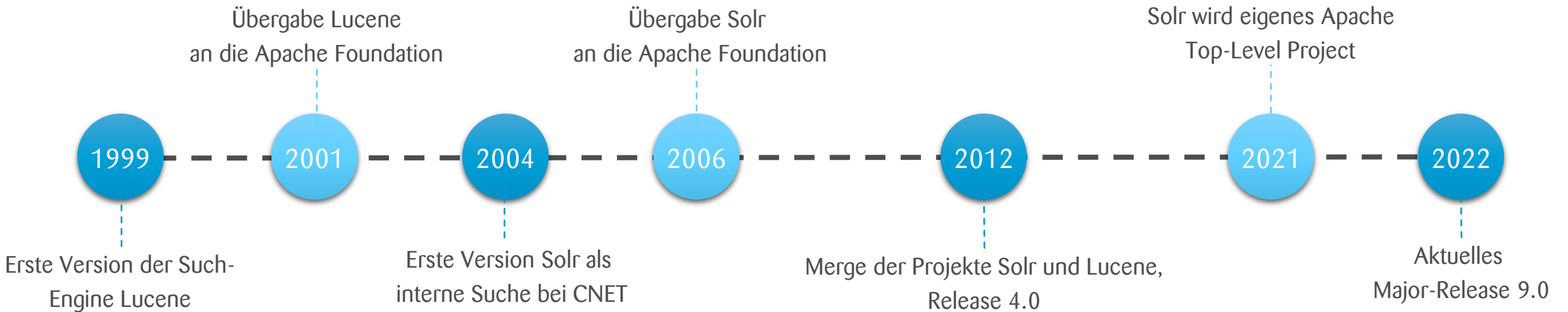
=> Solche Tool existieren!
- Eine wirklich gute Suche **braucht kontinuierlichen Aufwand**
 - Nach was sucht der Benutzer? Findet er die richtigen Resultate?
 - Wie benutzt der Benutzer die Suchfunktion?
 - Was für Daten werden durchsucht?

Was ist Solr?

Apache Solr

Facts and Figures

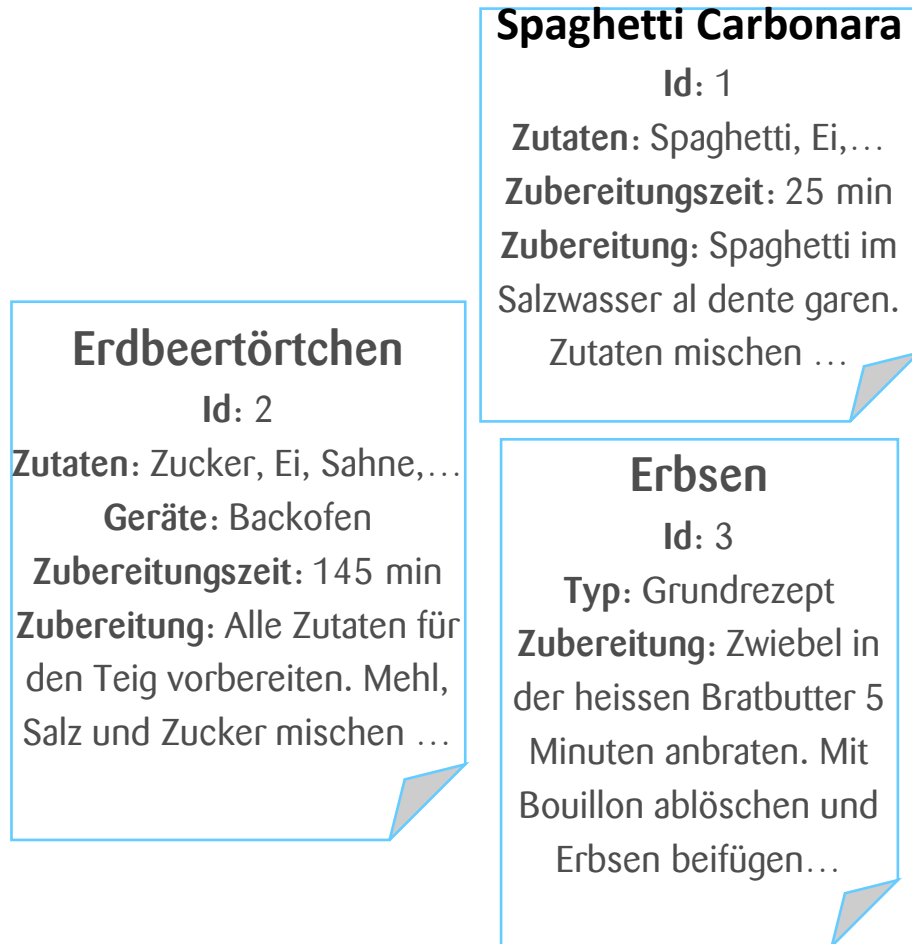
- Open-Source-Enterprise-Suchplattform
- Entwickelt in Java, basiert auf Apache Lucene
- Index-Suche
- Wichtigste Features: Volltextsuche, Highlighting, Faceting und Clustering, Replikation und Balancing, Binärdokumente etc.
- Weltweit in Verwendung unter anderem von DuckDuckGo, Adobe, Instagram, eBay, Netflix, Disney



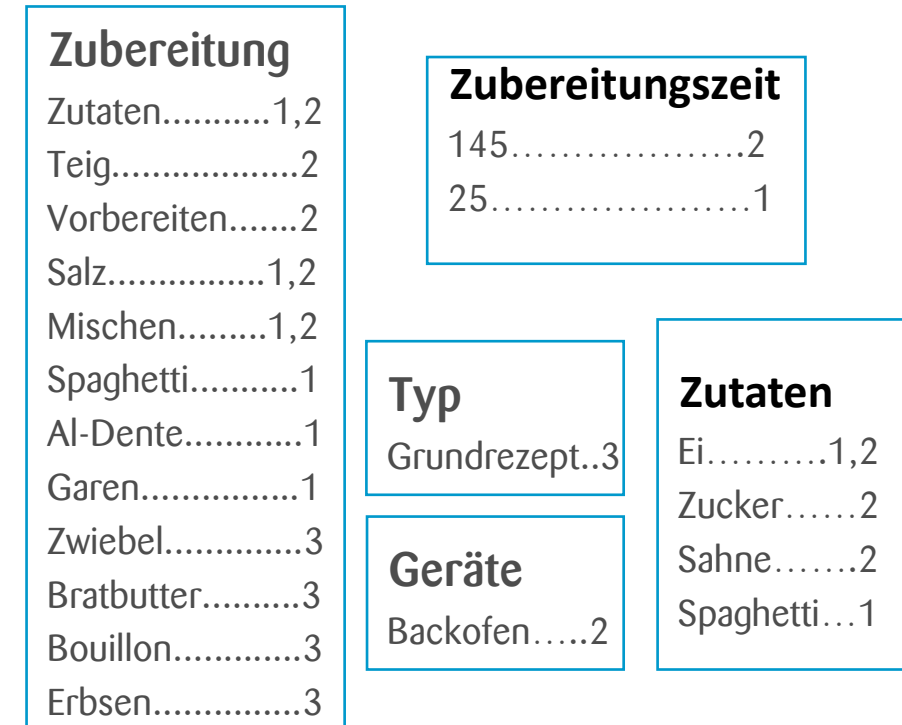
Begriffe

Document	Eine Menge von Daten, die etwas beschreiben. Grundlegende Informationseinheit von Solr. Im Dokumente setzen sich aus Feldern zusammen (Semi-Strukturiert). Sollten über eine ID verfügen.
Field	Teil eines Document, können verschiedene Arten von Daten enthalten, z.B. Text, Fließkommazahlen, Listen etc.
Collection	Dokumente, die in einem einzigen logischen Index mit einer einzigen Konfiguration gruppiert sind.
Core	Eine einzelne Solr-Instanz / logischer Knoten. Mehrere Cores können auf einem einzigen physikalischen Knoten laufen.
Shard	Dokumente werden Shards zugewiesen. So kann eine Collection wenn nötig auf mehrere Cores verteilt werden.

Was ist ein Index?

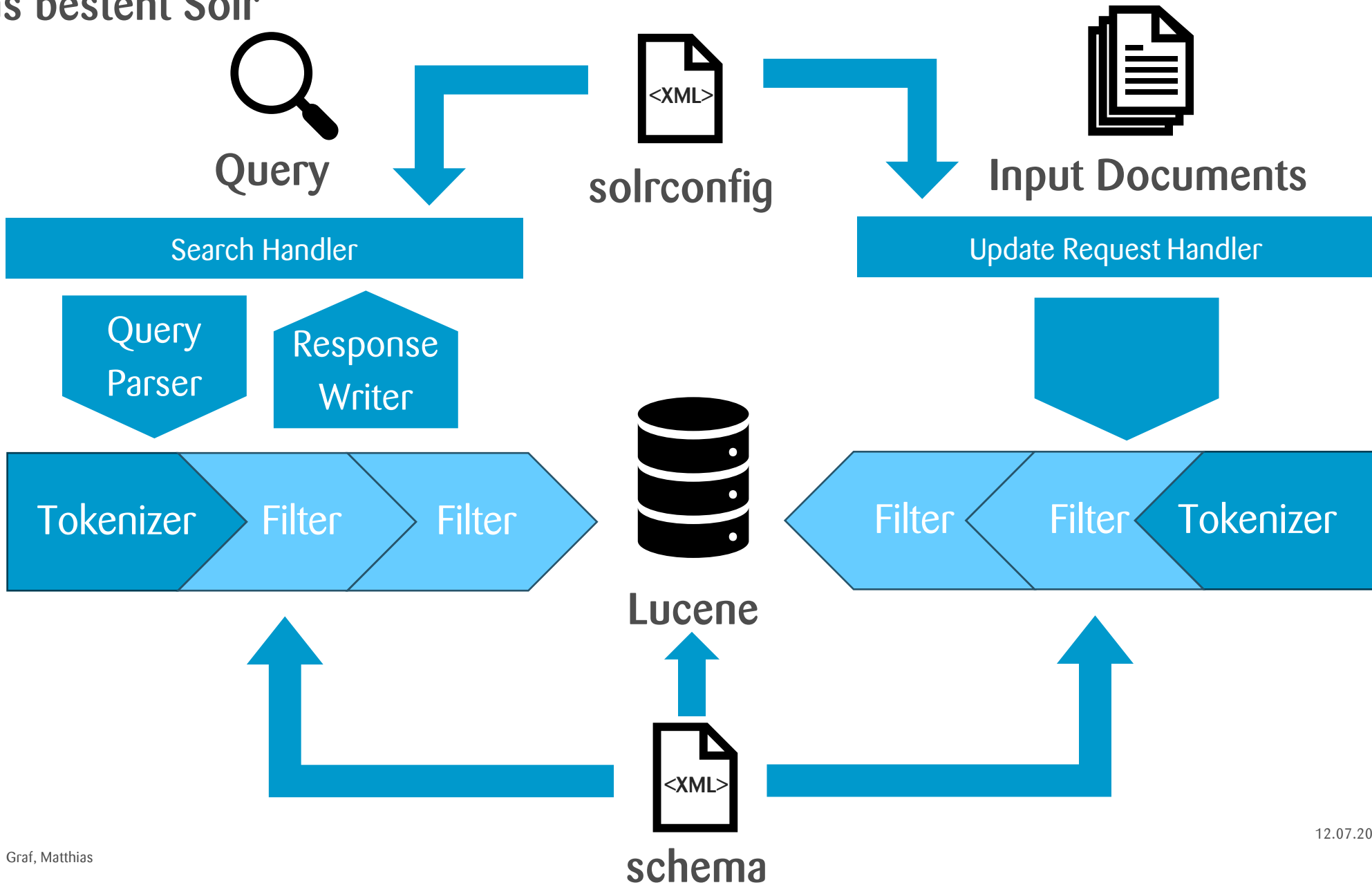


Semistrukturierte Daten (Dokumente)



Inverted Index

Aus was besteht Solr



Wie kann ich Solr benutzen?

Docker-Setup

Solr kann als Docker-Container gestartet werden

```
docker run -p 8983:8983 -t solr
```

Erweitertes Docker-Compose-Setup

```
solr:
  image: 'solr:9.2.0'
  ports:
    - "8983:8983"
  command:
    - solr-precreate
    - COLLECTION_NAME
  volumes:
    - ./src/main/solr:/opt/solr/server/solr/configsets/_default/conf
    - solr-data:/var/solr
```

Live Demo

Zusammenfassung Live Demo

- Solr-Server starten (Container oder Binary)
- Eigene Collection erstellen (precreate-script oder über UI)
- ManagementUI localhost:8983/solr ausprobieren
- Integration in eigene Applikation mittels Client-Library (Solrj) oder Rest-API
- Dokumente hinzufügen (Indexing)
- Suchen ausführen (Querying)

Tipps und Tricks

- Solr ist einfach zu integrieren, kann aber schwer zu meistern sein. Start als MVP und langsames Aufbauen
- Suche ist immer Kontext-Sensitiv: Was und wie sucht der Benutzer?
 - Eigene Abfragen vs. Solr-Queries?
 - Schema-Optimierungen
 - Faceting / Clustering
 - Konstantes Monitoring
- Best-Practice
 - Eine Collection für unterschiedliche Daten
 - Möglichst alle Daten aufnehmen, Schema-Änderungen und Reindex ist teuer
- Pain-Points
 - Berechtigungsprüfungen aus den Suchresultaten?
 - Wann Index updaten? Hohe Last oder veraltete Daten?

Takeaways

- Eine Suchfunktion ist wichtig, eine gute Suchfunktion zu implementieren ist aber schwer
- Apache Solr ist eine Open-Source Lösung mit guter Performance, vielen Features, einfach zu integrieren aber schwer zu meistern
- Start Slow and Improve

Fragen?



github.com/lizzyTheLizard/solr-jfs2023

Folien, Code-Beispiele und Referenzen

Weitere Informationen

solr.apache.org/guide

Offizielle Solr Dokumentation; Tutorials, Guides für Deployment und Konfiguration, Dokumentation der Query-Parser und vieles mehr

hub.docker.com/_/solr

Offizielles Docker-Image von Solr

baeldung.com/apache-solrj

SolrJ-Tutorial