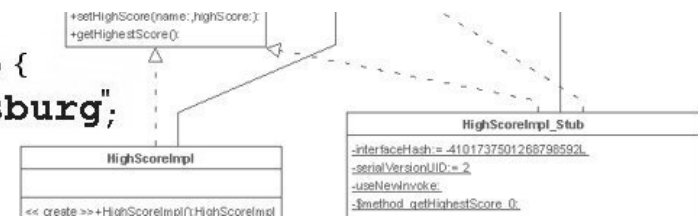


Bulk Web-Crawler mit Spring Batch



```
public class JavaUserGroup {  
    private String name = "Augsburg";  
}
```

Vorträge, Stammtische und vieles mehr...



Anforderung - funktional

Wir wollen automatisiert Überprüfung, ob bestimmte Produkte (Bücher) in einem Online-Shop gelistet sind.

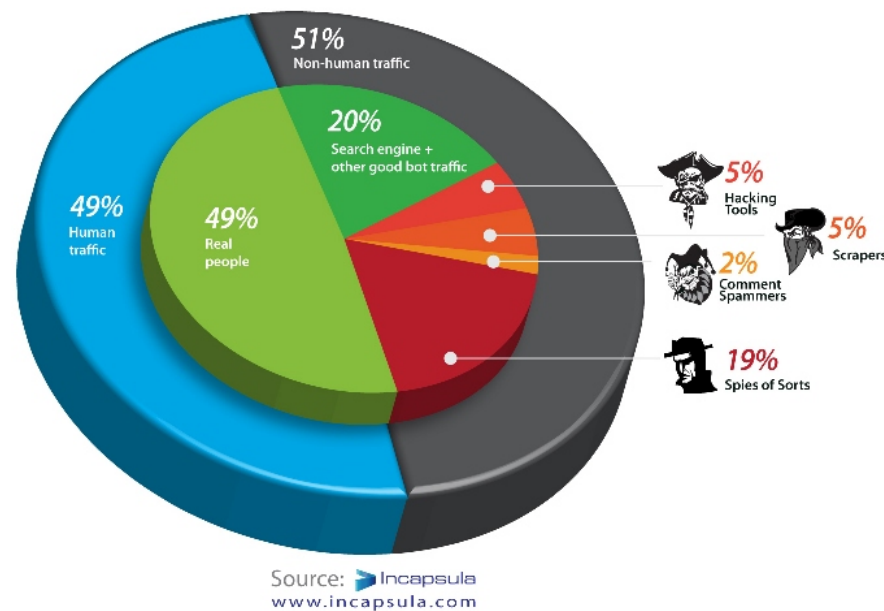
Site	Produkt	Status
AMAZON_DE	0815	FOUND
HIVE_UK	1234	NOT_FOUND
...

Anforderung - nicht-funktional

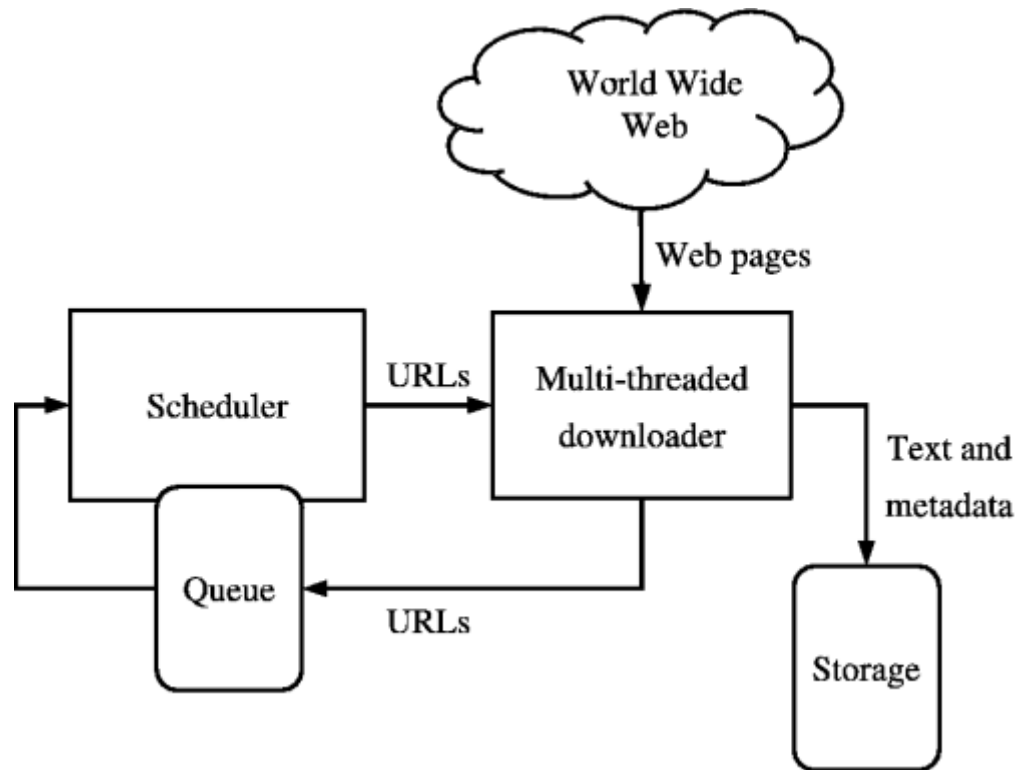
- ♥ Nicht-funktional:
 - REST-Prinzipien
 - optimiert ist für Hintergrundverarbeitung
 - Monitoring
 - batchfähig
 - Parallelisierung über Threads

Web-Crawler

- aka. internet bots, Spiders, Indexers, Scrapers
- Web-Crawler verursachen ca. 50 % des Internet Traffics weltweit!



Crawler: prinizieller Aufbau



Quelle: http://en.wikipedia.org/wiki/Web_crawler

Crawler: Richtlinien

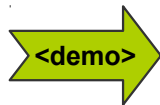
- selection policy
- re-visit policy
- politeness policy
- parallelization policy

Analyse der Zielseite

http://www.hive.co.uk/

ISBN 978-1-408-85567-6

„Harry Potter And The Prisoner Of Azkaban“



The screenshot shows the Hive website interface. At the top, there's a navigation bar with social media links (Facebook, Twitter, Pinterest, YouTube) and a search bar. The search bar contains the ISBN 978-1-408-85567-6. Below the search bar, there's a horizontal menu with categories: About Us, Books, Children's Books, eBooks, DVD & Blu-ray, Music, Stationery & Gifts, Blog, and Store locator. The main content area displays the search results for '978-1-408-85567-6'. It shows a list of products, with the first result being 'Sinfonias 1 - 6 - Benda' by Keith Anderson, Christian Benda, and Jan Kotzmann. The page also includes a 'Refine by...' section on the left with media type filters (Book, eBook, DVD, Music) and a 'Sort by' dropdown menu set to 'Relevancy'.

hive Shop locally online

Search All products 978-1-408-85567-6 Go

switch to advanced search

About Us Books Children's Books eBooks DVD & Blu-ray Music Stationery & Gifts Blog Store locator

Home Search results for '978-1-408-85567-6,'

Refine by...

Media Type

- Book (142,946)
- eBook (11,704)
- DVD (3,713)
- Music (3,023)

You searched for '978-1-408-85567-6,' and these are the results

Sort by: Relevancy Showing 1 - 20 of 161,386 products

Products per page: 20

1 2 3 4 5

Sinfonias 1 - 6 - Benda by Keith Anderson, Christian Benda and Jan Kotzmann

Request-Sequenz



URL	Status	Domain
POST search	302 Moved Temporarily	hive.co.uk

Headers	Post	HTML	Cache	Cookies
Response Headers view source				
Cache-Control	no-store, no-cache, must-revalidate, post-check=0, pre-check=0			
Connection	keep-alive			
Content-Length	0			
Content-Type	text/html			
Date	Mon, 09 Feb 2015 09:01:17 GMT			
Expires	Thu, 19 Nov 1981 08:52:00 GMT			
Location	/search/9781408855676/mediatype/all/			
Pragma	no-cache			
Server	nginx			

URL	Status
POST search	302 Moved Temporarily
GET /search/9781408855676/mediatype/all/	302 Moved Temporarily
http://www.hive.co.uk/book/harry-potter-and-the-prisoner-of-azkaban/18287543/	
GET swc.js	200 OK
GET swc.css	200 OK
GET main-1123471070	200 OK

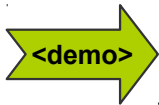


- POST search zur Such-Seite
- GET /search/9781408855676/mediatype/all/
- GET /book/harry-potter-and-the-prisoner-of-azkaban/18287543/

„Canonical Links“

- Identischer Seiten-Content wird unter verschiedenen URLs angeboten (Domain-intern, aber auch domainübergreifend)
- Beispiele:
 - <http://.../search/9781408855676/mediatype/all/>
 - <http://.../book/harry-potter-and-the-prisoner-of-azkaban/18287543/>
 -

Request-Seqzenz

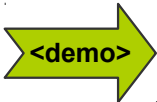


Status	Domain	Size
302 Moved Temporarily	hive.co.uk	0 B
302 Moved Temporarily	hive.co.uk	0 B
200 OK	hive.co.uk	52.9 KB

- POST, 0 bytes Content, nur HTTP redirect
- GET, 0 bytes Content, nur HTTP redirect
- GET, 52.9 KB Content
- Danach lädt der Browser nur noch CSS/JS Ressourcen vom Webserver.

Analyse URL-Aufbau

<http://www.hive.co.uk/book/harry-potter-and-the-prisoner-of-azkaban/18287543/>



Suche nach “**18287543**” im HTML-Code ...

```
div.mc_sku < div#mc_data < body.js-enabled < html
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1
<html xmlns:fb="http://www.facebook.com/2008/fbml" xmlns="http://www.w3.org/1999/xhtml">
  <head>
  <body class="js-enabled" data-twttr-rendered="true">
    <div id="mc_data" style="display: none;">
      <div class="mc_event">VIEW</div>
      <div class="mc_retailer">HIVE</div> <div class="mc_sku">18287543</div>
      <div class="mc_sku">18287543</div>
    </div>
    <script type="text/javascript">
```

Algorithmus 0.1

1. perform an HTTP GET request to
http://www.hive.co.uk/search/9781408855676/mediatype/all/
2. if HTTP response returns code **302**,
then product was **found**.

Der Algorithmus ist unvollständig, da er den Fall nicht berücksichtigt, wenn ein Produkt nicht gefunden wurde.

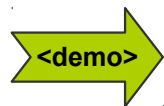


Algorithmus 0.2


1. perform an HTTP GET request to
<http://www.hive.co.uk/search/9781408855676/mediatype/all/>
2. if HTTP response returns code **302**,
then product was **found**.
3. if HTTP response returns code **200**
the product was **not found**.

Optimierung

Im Fehlerfall antwortet der Web-Server mit einer Fehlerseite:



Home Search results for '9781408855671'

 Sorry, we did not find any results for **9781408855671**.
Please try again with different search criteria.

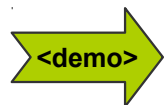
Search Tips

- Double-check your spelling
- Try using fewer or less specific keywords
- Try browsing categories

40.7 KB	87.83.25.62:80
48.8	
26.5	Response Body 40.7 KB (41,708 B)
87.9	Post Body 0 B
2.8	Total Received* 41.2 KB (42,149 B)
7	Total Sent* 577 B
2.0	* Including HTTP Headers

Dieser Request umfasst einen Download von ca. **40 KB** für jedes Produkt, das **nicht** gefunden wurde.

Lösung: GET-Request durch HEAD-Request ersetzen



HEAD <http://www.hive.co.uk/search/9781408855676/mediatype/all/>

● 302 Found 📶 0 bytes ⌚ 638 ms



Algorithmus 1.0

1. perform an HTTP HEAD request to
http://www.hive.co.uk/search/9781408855676/mediatype/all/
2. if HTTP response returns code **302**,
then product was **found**.
3. if HTTP response returns code **200**
the product was **not found**.

Einschub: Content Inspection

Der Algorithmus basiert nur auf HTTP-Ebene und berücksichtigt keine HTML-Seiteninhalte.

- Durch *Content Inspection* werden Informationen aus dem HTML extrahiert:
 - element by id
 - element by css style
 - Werte in Tags (z.B. ``)
 - hidden divs (`display:none`)
 - Ausgehende Links (``)

Implementierung - Teil 1

- Setup Java-Main mit Maven und Apache HTTPClient
- Absetzen des HEAD-Request
- Absetzen GET-Request und Download des HTML-Contents
- HTML Content Inspection mit *htmlcleaner*



Frameworks/Libraries

- Spring 4.1
- Spring Batch 3.0
- Apache HttpClient 4.3
- HtmlCleaner 2.8
- EclipseLink 2.5.2
- PostgreSQL 9.3
- Maven 3
- Java 8

htmlcleaner

- API ähnlich SAX (XML)
- Visitor-Pattern
 - ressourcenschonend
 - skalierbar
- Statushandling:
 - Boolean-Flag: found = true/false
- Single-Pass vs. Multi-Pass

HACKING...



Implementierung – Teil 2

- Setup Spring Batch
- Konfiguration der Jobs
- Batch-Metadaten-Tabellen

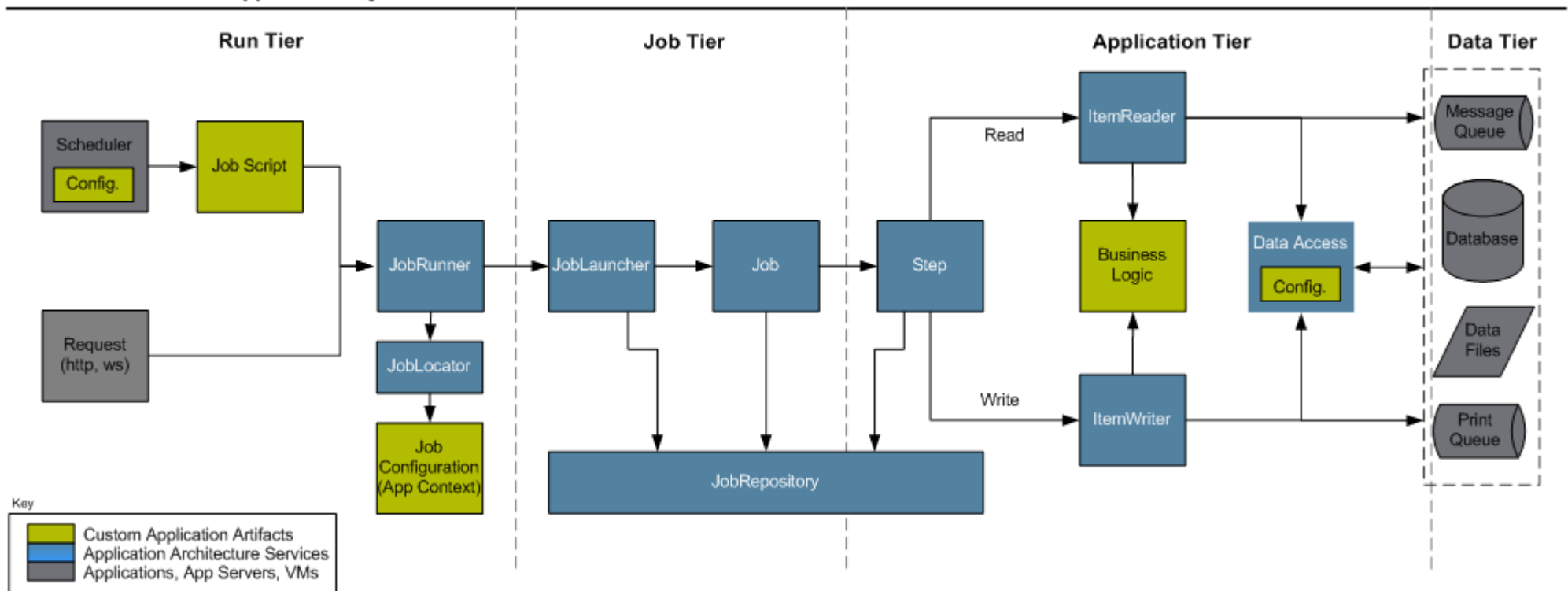
Spring Batch

„process large volumns of information without user interaction“

- ♥ Konzepte
 - Jobs
 - Steps
 - Tasklets
 - Chunk oriented: Reader-Processor-Writer
- ♥ AOP
 - Step/Chunk/Item Listeners
 - Retry Template
 - Transaction
 - Logging
- ♥ Spring Batch ist KEIN Scheduler

Spring Batch

Batch Application Style – Interactions and Services



Quelle: spring.io

„bulkCrawlerJob“

- Reader
 - CSV File
 - FlatFile Reader
- Processor
 - Crawler Implementierung
- Writer
 - Status loggen
 - später auch in die Datenbank

HACKING...






Spring Batch Admin UI

← → ↻ 🏠 📄 localhost:8080/sbent-sample/jobs/executions ☆ * 🔍

Spring Batch Admin



Home Jobs Executions Files SpringSource Spring Batch

Recent and Current Job Executions

Stop All

ID	Instance	Name	Date	Start	Duration	Status	ExitCode
<u>5</u>	5	report	2012-09-25	17:07:55	00:00:00	COMPLETED	COMPLETED
<u>4</u>	4	processReturns	2012-09-25	17:07:05	00:00:34	COMPLETED	COMPLETED
<u>3</u>	3	loadReturn	2012-09-25	17:06:32	00:00:22	COMPLETED	COMPLETED
<u>2</u>	2	loadFromQueue	2012-09-25	17:06:21	00:00:00	COMPLETED	COMPLETED
<u>1</u>	1	createQueueTestData	2012-09-25	17:06:09	00:00:02	COMPLETED	COMPLETED

Rows: 1-5 of 5 Page Size: 20

© Copyright 2009-2010 SpringSource. All Rights Reserved. [Contact SpringSource](#)

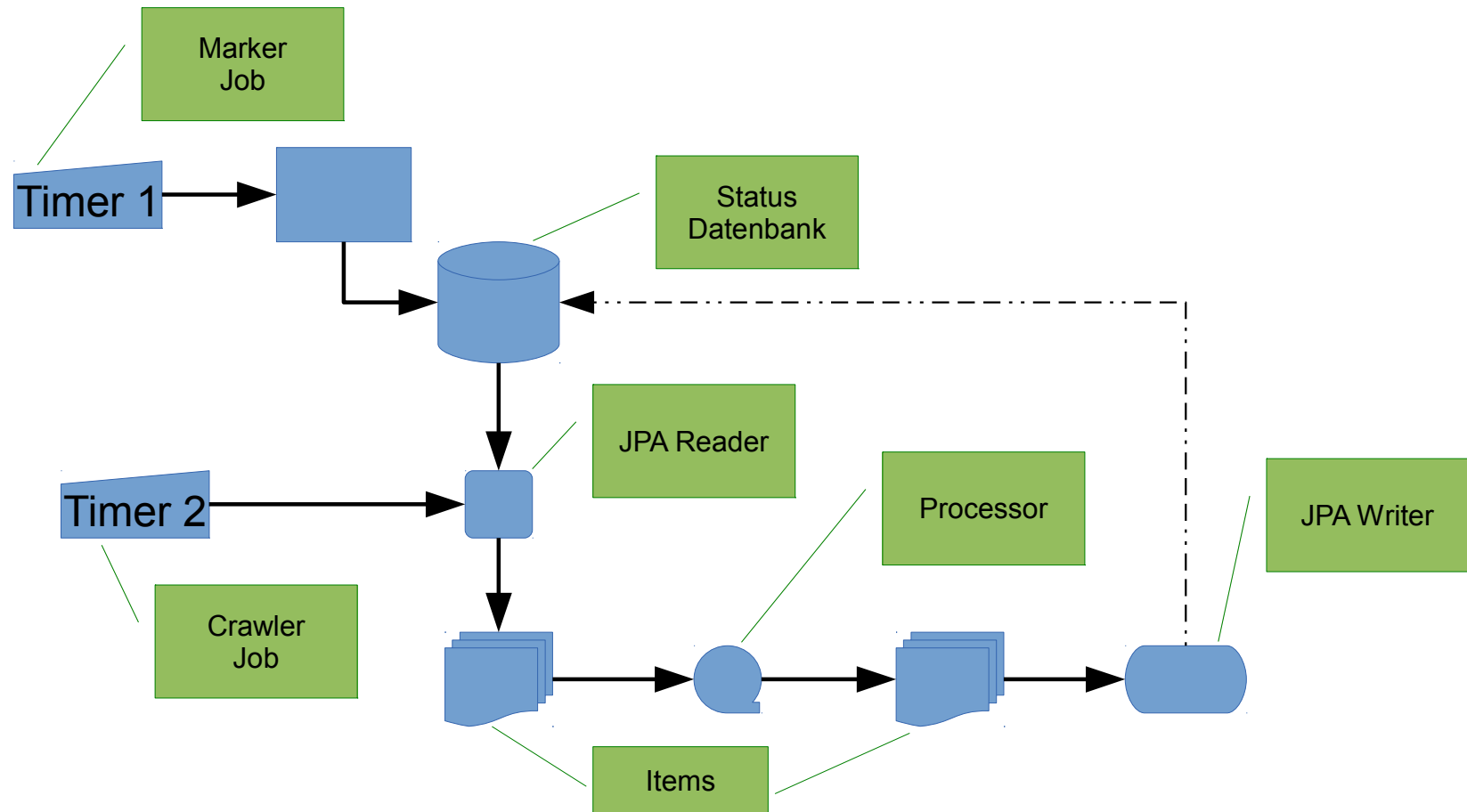
Implementierung – Teil 3

- ♥ Datenmodell
 - CrawlerStatus
 - Produkt
- ♥ Statushandling
 - Modellierung Tabelle und Statusnetz
 - Status-Datenbank anbinden
 - Scheduling
- ♥ Partitionierung mit Spring Batch
- ♥ Konfiguration der Pooling (Thread-Pool, DBCP)

HACKING...



Job Konfiguration - final



„bulkCrawlerJob“ - mit DB

- ♥ JPA Reader
 - ersetzt den CSV FlatFile reader
- ♥ JPA Writer
 - ersetzt den *LoggingItemWriter*
 - Status in Statusdatenbank via JPA
- ♥ zusätzlicher Job *bulkCrawlerMarkerJob*

Status-Modell

📍 Statusmodell (s. Tabelle)

📍 Statusnetz

- FOUND
- NOT_FOUND
- CHECK_FAILED

site_ident	isbn13	priority	crawler_result	since_timestamp	latest_timestamp	count
HIVE_UK	978-1-408-85567-6	90	FOUND	09:15:00	19:35:05	4
HIVE_UK	978-1-408-12122-3	70	FOUND	11:35:48	19:34:46	3
...

Scheduling-Logik

	Gestoppt		Freigegeben	
	NOT_FOUND		Prio 60 pub>7d last>48h	Prio 5 last>60d
	FOUND	Prio 25 last>48h		Prio 7 last>30d
	null	Prio 70	Prio 85	

HACKING...



Status/Visualisierung

tango

Application ▾ Dashboard ▾ Daten ▾ User ▾

Suche

Q

Crawler Overview

Sites overview


	F	F(!)	¬F	¬F(!)	S	
AMAZON_COM_v1	64294	2181	2827	2104	67121	2015-01-28 09:06:24
AMAZON_CO_UK_v1	66678	2468	443	7	67121	2015-01-28 09:11:07
AMAZON_COM_v1	67013	2803	108	7	67121	2015-01-28 09:26:45
AMAZON_FR_v1	65735	1541	1386	23	67121	2015-01-28 05:36:27
BOOKBUTLER_DE_v1	64574	1370	5547	4013	67121	2015-01-28 09:46:26
					67121	2015-01-28 09:46:10
					67121	2015-01-28 09:43:53
					67121	2015-01-28 09:45:33
					67129	2015-01-28 09:31:53
BUECHER_DE_v1	66423	2312	698	106	67121	2015-01-28 09:35:06
CALVENDO_DE_v1	54298	184	12831	10103	67129	2015-01-28 09:43:59
DNB_v1	59831	2570	7290	6956	67121	2015-01-28 09:35:31

Recently Crawled

				cnt
978-3-660-89044-0	WELTBILD_DE_v1	FOUND	2015-01-28 09:47:00	4
978-3-660-95377-0	BOOKBUTLER_DE_v1	FOUND	2015-01-28 09:46:26	4
978-3-660-07918-0	BUCH24_DE_v1	FOUND	2015-01-28 09:46:10	4
978-3-660-81644-0	WELTBILD_DE_v1	FOUND	2015-01-28 09:45:58	4
978-3-660-07918-0	BUCHKATALOG_DE_v1	FOUND	2015-01-28 09:45:33	4
978-3-660-30120-5	BUCHKATALOG_DE_v1	FOUND	2015-01-28 09:44:26	4
978-3-660-41086-0	EBAY_DE_v1	FOUND	2015-01-28 09:44:26	5
978-3-660-89044-0	THALIA_DE_v1	FOUND	2015-01-28 09:44:13	4
978-3-664-06503-5	BUCHKATALOG_DE_v1	FOUND	2015-01-28 09:44:13	6
978-3-660-71017-5	CALVENDO_DE_v1	FOUND	2015-01-28 09:43:59	3
978-3-660-95377-0	BUCHHANDEL_DE_v1	FOUND	2015-01-28 09:43:53	4
978-3-660-76207-5	BUCH24_DE_v1	FOUND	2015-01-28 09:42:20	4



Status

Verfügbarkeitsmonitor 			
Seite	Status	Link	Zuletzt überprüft
Amazon COM	✓		2014-11-22 13:51:48
Amazon UK	✓		2014-11-22 13:51:59
Amazon DE	✓		2014-11-22 13:52:02
Amazon FR	✓		2014-12-19 01:37:36
bookbutler.de	✗		2015-01-26 03:21:21
buch24.de	✓		2014-11-19 08:56:09
buchhandel.de	✓		2014-11-22 13:22:36
buchkatalog.de	✓		2014-11-22 13:52:08
buch.de	✓		2014-11-22 13:22:15
buecher.de	✓		2014-11-22 13:51:08
CALVENDO.de	✓		2014-11-22 13:22:32
Deutsche National Bibliothek	✓		2014-11-22 13:52:15
ebay.de	✓		2014-11-22 13:51:23
ebook.de	✓		2014-11-22 13:51:13
hive.co.uk	✗		2015-01-26 03:21:22
kalenderhaus.de	✓		2014-11-22 13:22:26
thalia.de	✓		2014-11-22 13:51:27
weltbild.de	✗		2015-01-26 03:21:25



Links

- <http://www.michaelnielsen.org/ddi/how-to-crawl-a-quarter-billion-webpages-in-40-hours/>
- <http://www.baeldung.com/httpclient4>
- <http://www.servage.net/blog/2013/04/08/rest-principles-explained>
- http://en.wikipedia.org/wiki/Web_crawler
- Oracle Java API Docs: **ThreadPoolExecutor**

Das war's...

Zumindest von meiner Seite...

Danke für die Aufmerksamkeit!