

## Solr Index erstellen

### 1. Solr Ordner in HDFS erstellen

- `sudo -u hdfs hadoop fs -mkdir /solr`
- `sudo -u hdfs hadoop fs -chown solr /solr`

### 2. Instance directory generieren

- `solrctl instancedir --generate $HOME/solr_configs`

### 3. Konfiguration in Zookeeper laden

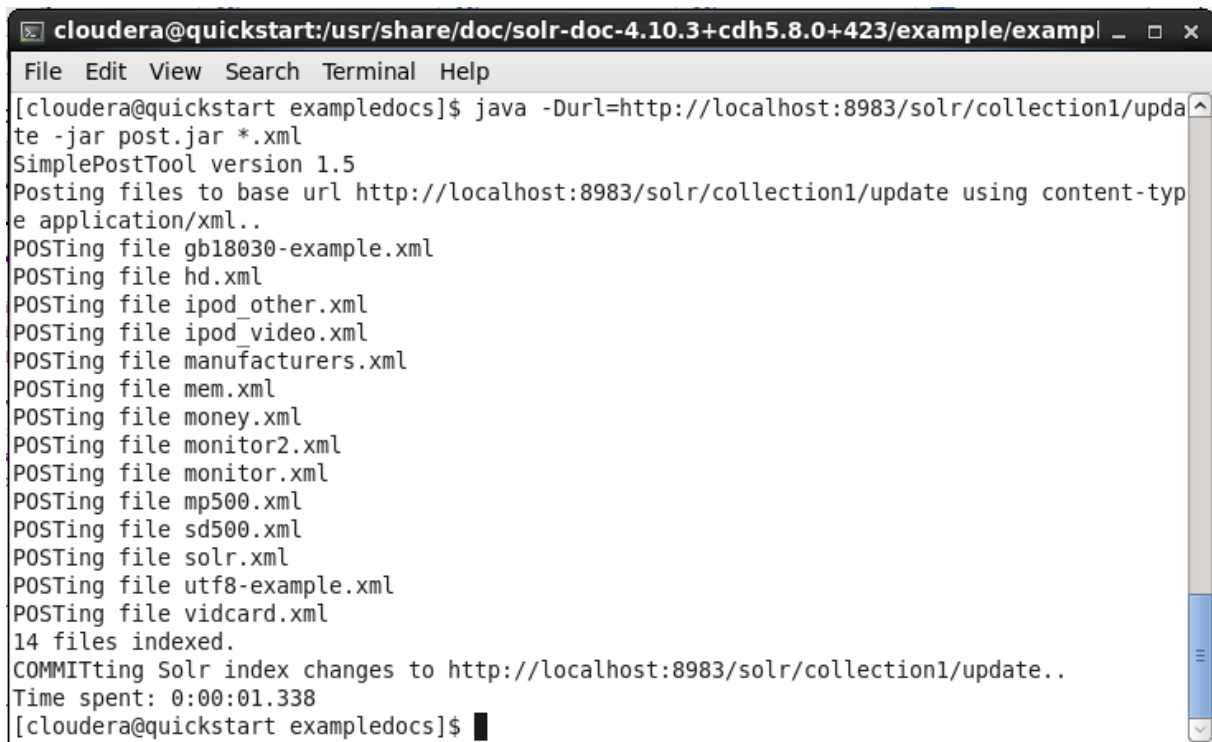
- `solrctl instancedir --create collection1 $HOME/solr_configs`

### 4. Solr Collection erstellen

- `solrctl collection --create collection1 -s 1`

### 5. Indizierung der Daten

- `cd /usr/share/doc/solr-doc*/example/exampledocs`
- `java -Durl=http://localhost:8983/solr/collection1/update -jar post.jar *.xml`

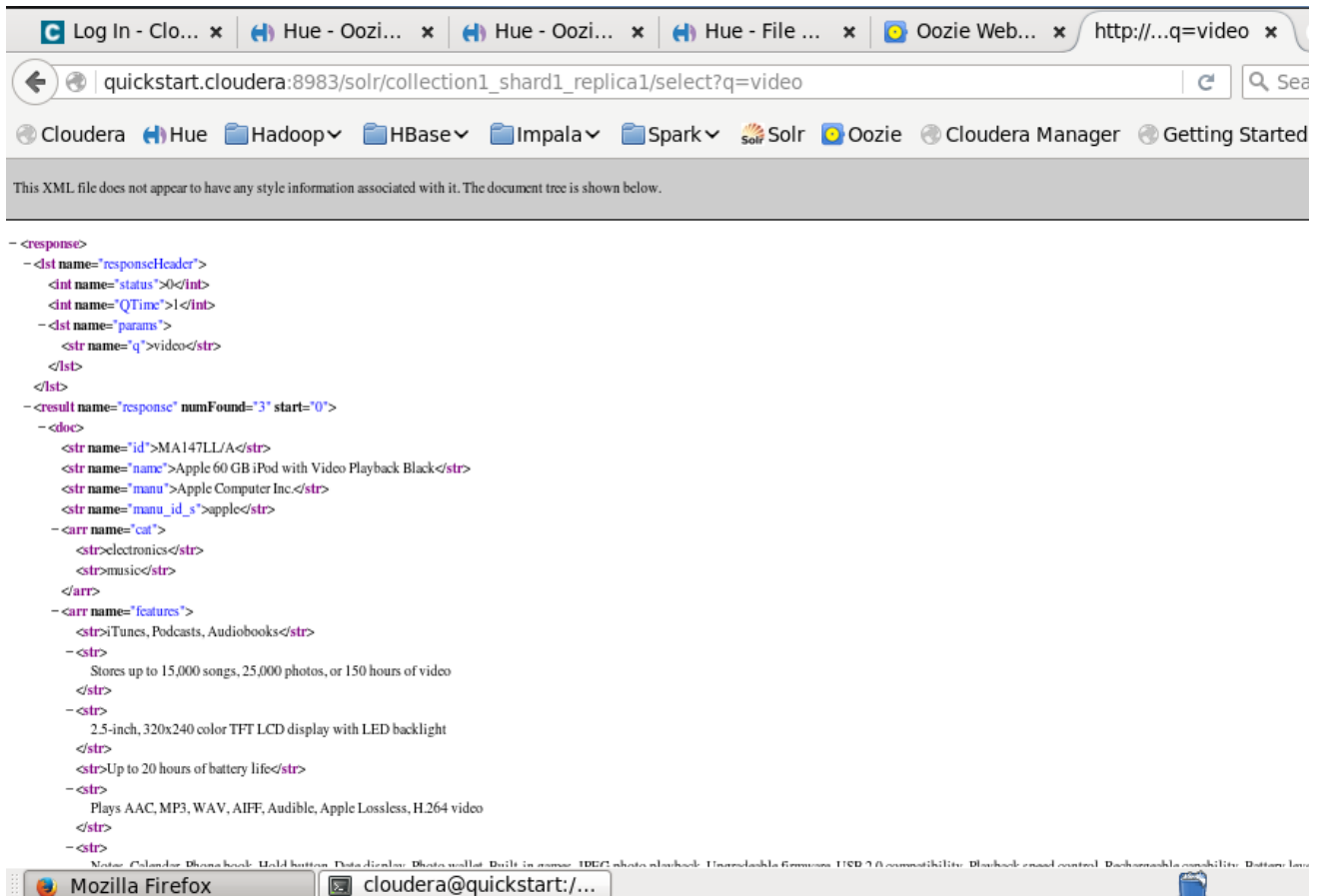


```
cloudera@quickstart:/usr/share/doc/solr-doc-4.10.3+cdh5.8.0+423/example/exampledocs$ java -Durl=http://localhost:8983/solr/collection1/update -jar post.jar *.xml
SimplePostTool version 1.5
Posting files to base url http://localhost:8983/solr/collection1/update using content-type application/xml..
POSTing file gb18030-example.xml
POSTing file hd.xml
POSTing file ipod_other.xml
POSTing file ipod_video.xml
POSTing file manufacturers.xml
POSTing file mem.xml
POSTing file money.xml
POSTing file monitor2.xml
POSTing file monitor.xml
POSTing file mp500.xml
POSTing file sd500.xml
POSTing file solr.xml
POSTing file utf8-example.xml
POSTing file vidcard.xml
14 files indexed.
COMMITting Solr index changes to http://localhost:8983/solr/collection1/update..
Time spent: 0:00:01.338
[cloudera@quickstart exampledocs]$
```

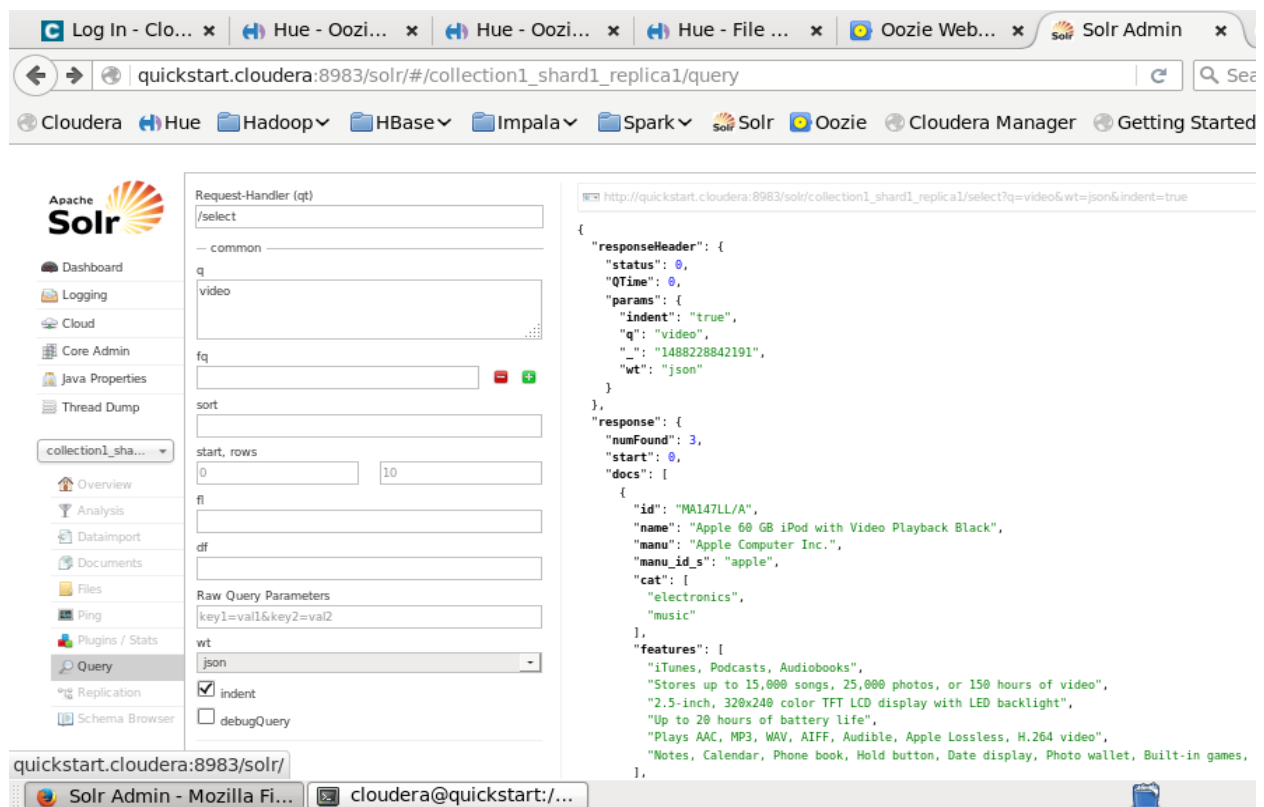
- Web-Ui erreichbar unter: <http://quickstart.cloudera:8983/solr/>

## 6. Querys

- URL: [http://quickstart.cloudera:8983/solr/collection1\\_shard1\\_replica1/select?q=video](http://quickstart.cloudera:8983/solr/collection1_shard1_replica1/select?q=video)



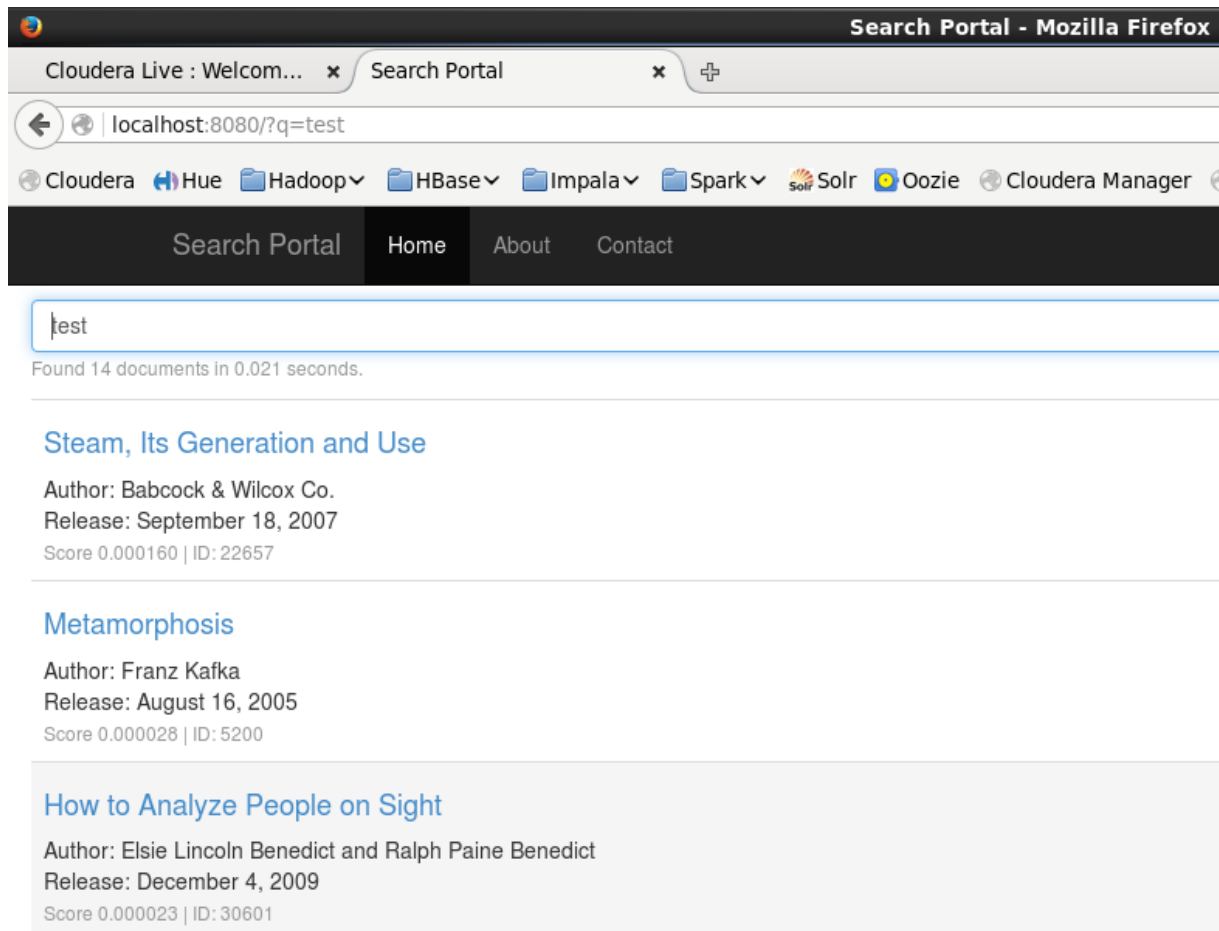
- Web-UI: [http://quickstart.cloudera:8983/solr/#/collection1\\_shard1\\_replica1/query](http://quickstart.cloudera:8983/solr/#/collection1_shard1_replica1/query)



## TF-IDF Suche

commands von Lesson 4 ausgeführt.

Ergebnis:



Search Portal - Mozilla Firefox

Cloudera Live : Welcom... x Search Portal x

localhost:8080/?q=test

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager

Search Portal Home About Contact

test

Found 14 documents in 0.021 seconds.

**Steam, Its Generation and Use**

Author: Babcock & Wilcox Co.  
Release: September 18, 2007  
Score 0.000160 | ID: 22657

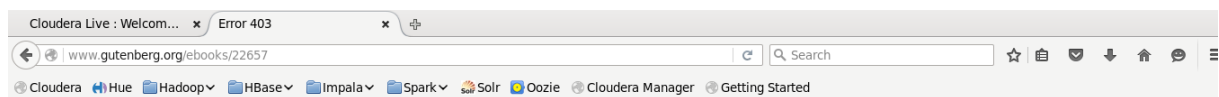
**Metamorphosis**

Author: Franz Kafka  
Release: August 16, 2005  
Score 0.000028 | ID: 5200

**How to Analyze People on Sight**

Author: Elsie Lincoln Benedict and Ralph Paine Benedict  
Release: December 4, 2009  
Score 0.000023 | ID: 30601

Sobald ich ein Ergebnis auswähle, erhalte ich den Error 403



Cloudera Live : Welcom... x Error 403 x

www.gutenberg.org/ebooks/22657

Search

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

### Error 403

Maybe you have just a wrong url. Go to <http://www.gutenberg.org/ebooks/> first to see if the error persists.

If you get the error again check that you:

- Don't use anonymizers, open proxies, VPNs, or TOR to access Project Gutenberg. This includes the Google proxies that are used by Chrome.
- Don't access Project Gutenberg from hosted servers.
- Don't use automated software to download lots of books. We have a limit on how fast you can go while using this site. If you surpass this limit you get blocked for 24h.
- We have a daily limit on how many books you can download. If you exceeded this limit you get blocked for 24h.
- If you use the RSS feed, set your update interval to 24 hours.

If you are sure that none of the above applies to you, and wish us to investigate the problem, we need to know your IP address. Go to [this site](#), don't sign up, just copy the IP address (it looks like: 12.34.56.78 but your numbers will be different) and [mail it to us](#). If that page also shows a proxy address, we need that one too.

## Vergleich

### **Solr**

- Rückgabe als JSON – somit detaillierte Informationen über Aufbau der Rückgabeinformationen
- Für ungeübte jedoch etwas verwirrend
- Über URL können weitere Parameter mitgegeben werden

### **TF-IDF**

- Gut designte Web-UI (Bootstrap?)
- Einfach zu verstehen
- Schnell (Suche nach „test“ in 0,021 Sekunden abgeschlossen)
- Ausgabe eines Score Wertes, der Informationen über die Häufigkeit (und dessen Relevanz?) des Suchwortes in einem Dokument gibt