Workflow

In der Datei workfow.py ist der Workflow festgelegt. Zuerst werden „xkcd“ sowohl auf der hdfs als auch lokal gelöscht. Nachdem das passiert ist, werden die Ordner „xkcd“ und „raw“ sowohl auf der hdfs als auch lokal erstellt. Danach fängt der Download an. Dieser ist durch eine Funktion definiert, welche mit einer for-Schleife die einzelnen JSON Objekte speichert, nach Jahr sortiert und in die jeweiligen Ordner speichert. Der Ordner xkcd wird dann auf die HDFS gepusht. Zuletzt gibt es eine Pyspark Datei (data_to_database.py), welche im letzten Schritt ausgeführt wird.

In der Datei data_to_database.py wird zuallererst eine Spark Session eröffnet. Danach werden alle JSON Dateien von der HDFS in einem Dataframe gespeichert. Das Dataframe wird um alle Spalten bereinigt, die nicht gebraucht werden. Die Daten werden im csv Format in dem Ordner „final“ abgespeichert und in die Datenbank geladen.

Datenbank

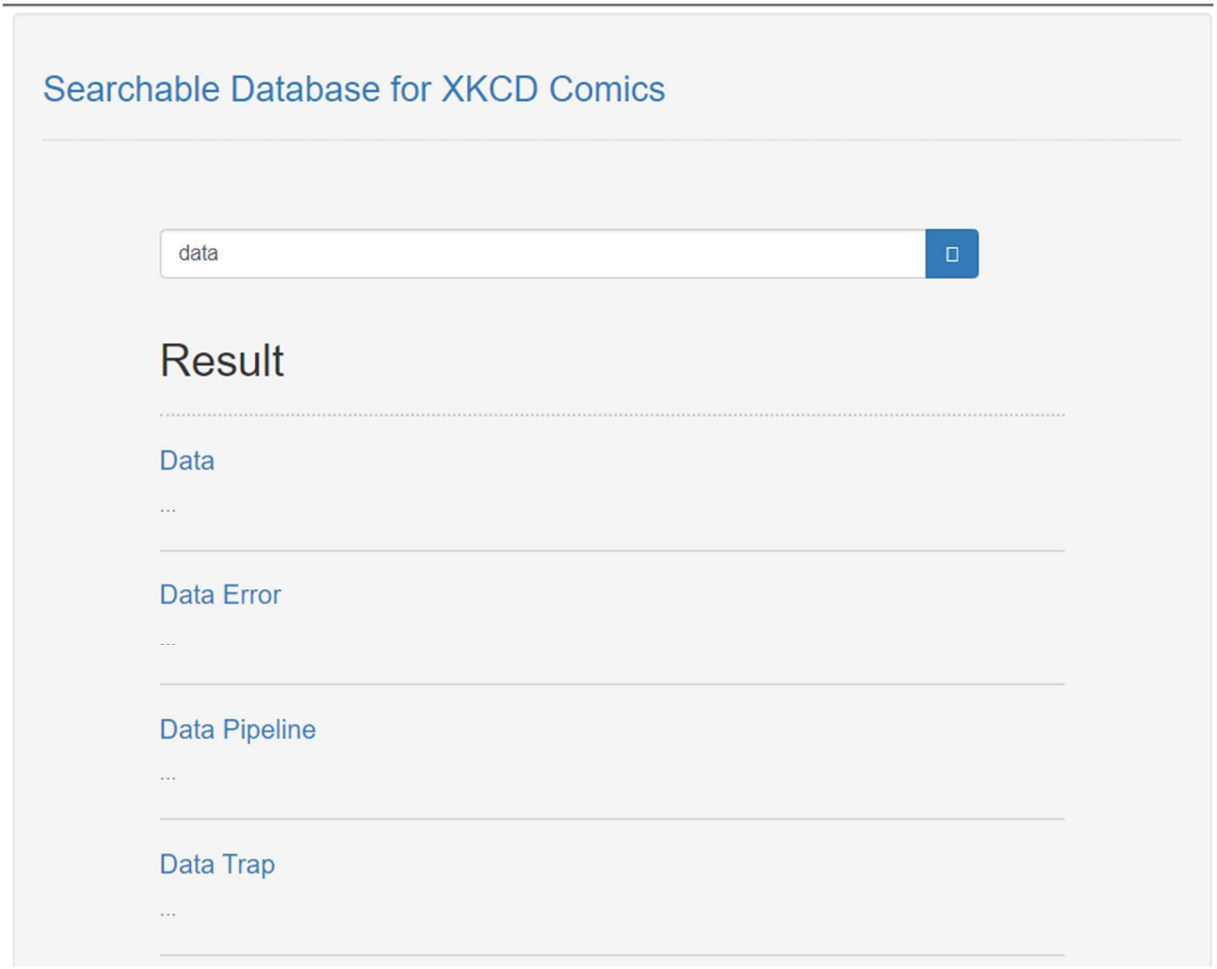
Für die Datenbank wurde ein neuer Docker Container aufgesetzt. Die Datenbank und die Tabelle wurde schon im Voraus erstellt.

Im Airflow Container muss noch der JDBC Driver heruntergeladen werden und in spark/jars kopiert werden.

Wget https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.0.31.tar.gz

Website

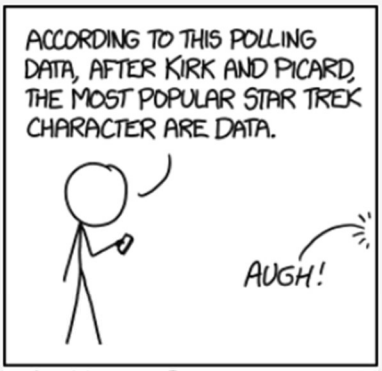
Die Website wurde lokal mit PHPStorm erstellt. Damit die Website läuft muss der Port für die Datenbank in der Firewall freigeschalten werden (3306). Zusätzlich muss die IP-Adresse der VM in den Dateien search.php und get_comics.php verändert werden.



Comics Database

Back

Data



https://imgs.xkcd.com/comics/data.png

Release Date: 3.10.2014

Zum Zeitpunkt der Abgabe hat alles einwandfrei funktioniert.

Bei Fragen oder Problemen: goette.elisabeth@gmail.com