

Tübix 2019

Webcrawling mit scrapy und PostgreSQL backend

06. Juli 2019

Janek Schoffit

Lightning talk

ZIELE

Ziele

- Open source stack
- Einfach skalierbar
- Erweiterbarkeit

SCRAPY

Scrapy

- Open source web spider framework
- Erweiterbar durch Plugins und Python libs
- Mit scrapyd als daemon deploybar

Scrapy

- Einfache Link extraction per regex
- Mächtiger Parser
- scrapy-rotating-proxy plugin

API

API

- REST API
 - Gin Gonic web framework
- Upload der crawler Daten
- Suchanfragen der Website

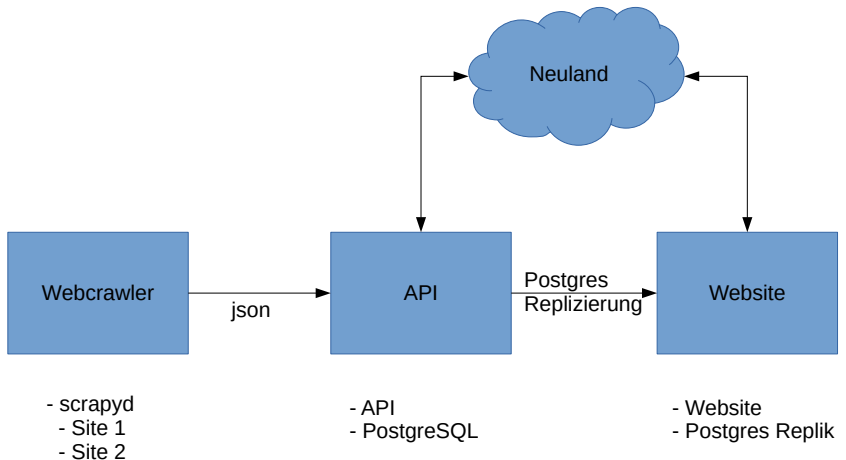
POSTGRESQL

PostgreSQL

- Relationale Datenbank
- JSON Felder mit Indexierung
- Fulltext search support
 - Vektor Spalte für schnelles durchsuchen

AUFBAU

Aufbau



PEACE OUT!