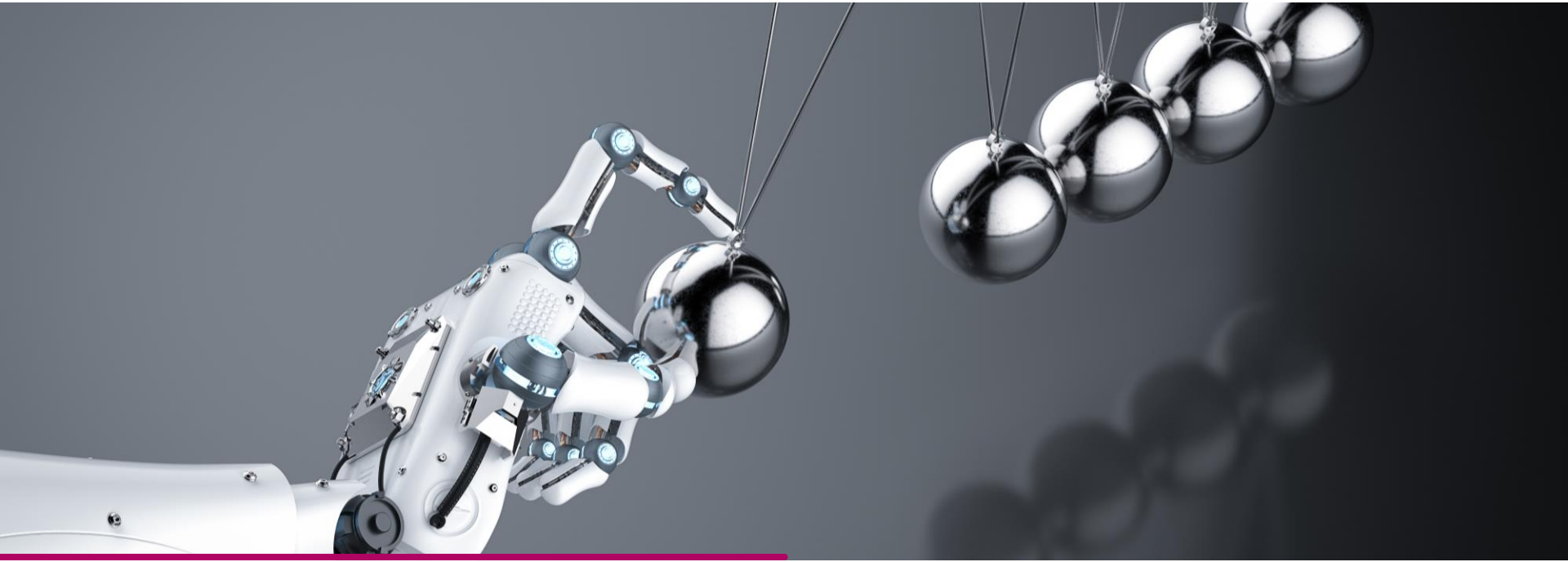
Transparenzregister Kapitalgesellschaften

Automatische Daten Extraktion aus Web-Quellen



Wir geben Impulse

Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

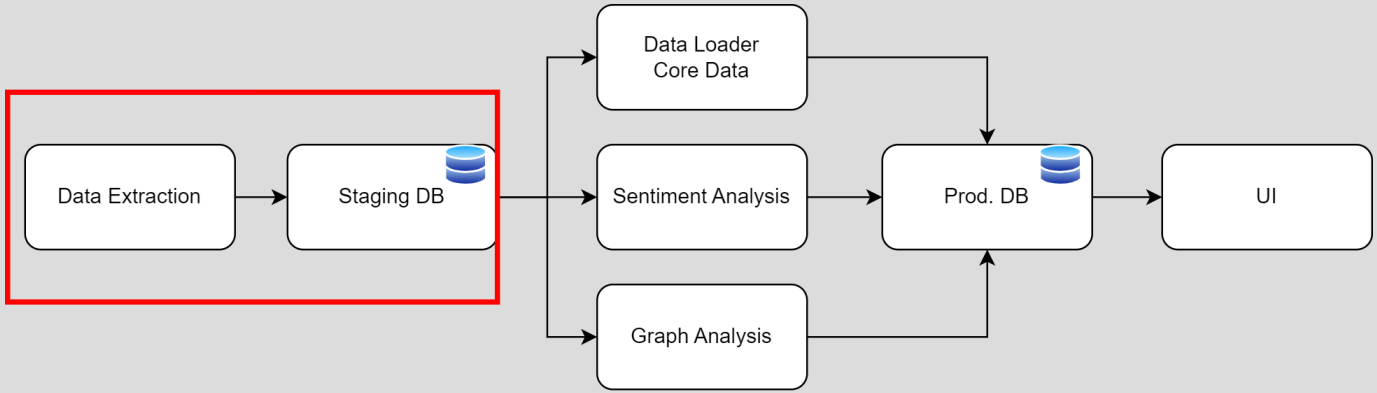
Agenda

- Zielsetzung
- Einblick: Datengrundlage
- Grundlagen – Web Mining & ETL
- Rechtliche Rahmenbedingungen
- Informationsgewinnung via
  - APIs
  - Web Scraping
  - Web Automation
- Orchestrierung
- Ausblick

Automatische Daten Extraktion aus Web-Quellen

Zielsetzung

- Datenbank anreichern mit:
  - Unternehmens-Stammdaten
  - Bewegungsdaten wie Aufsichtsrat, Jahresabschlüsse etc.
  - Dokumenten wie Nachrichten-Artikeln



Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

Datengrundlage



Düsseldorf, den 29. Juni 2022

Grant Thornton AG
Wirtschaftsprüfungsgesellschaft
Eckhard Lewe, Wirtschaftsprüfer
Renate Hermsdorf, Wirtschaftsprüferin

Bilanz zum 31. Dezember 2021

Aktiva

	31.12.2021	Vorjahr
A. Anlagevermögen		
I. Immaterielle Vermögensgeg		
Entgeltlich erworbene Softwar		
II. Sachanlagen		
1. Grundstücke und Bauten		
2. Technische Anlagen und Ma		
3. Andere Anlagen, Betriebs- u		
4. Geleistete Anzahlung und A		
III. Finanzanlagen		
Sonstige Ausleihungen		
B. Umlaufvermögen		
I. Vorräte		
Waren		
II. Forderungen und sonstige V		
1. Forderungen aus Lieferunge		
2. Forderungen gegen verbund		
3. Sonstige Vermögensgegens		

Handelsregister B des Amtsgerichts Hamm	Abteilung B Wiedergabe des aktuellen Registerinhalts Abruf vom 07.06.2023 19:37	Nummer der Firma: **HRB 5363**
Abdruck	Seite 1 von 2	

1. **Anzahl der bisherigen Eintragungen:**

   51

2. **a) Firma:**

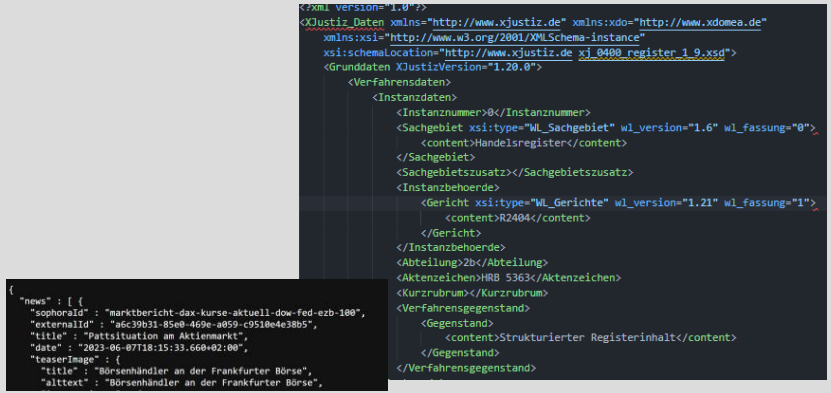
   GEA Farm Technologies GmbH

   b) Sitz, Niederlassung, inländische Geschäftsanschrift, empfangsberechtigte Person, Zweigniederlassungen:

   Bönen
   Geschäftsanschrift: Siemensstraße 25-27, 59199 Bönen

   c) Gegenstand des Unternehmens:

   Entwicklung, Herstellung und der Vertrieb von Landtechnik, insbesondere von Komponenten und Anlagen
   (a) zur Gewinnung, Kühlung, Behandlung und Lagerung von Milch;
   (b) für das Milchvieh-Herdenmanagement;
   (c) zur Tierhygiene und Sicherung der Milchqualität und
   (d) zur Aufstallung von Tieren;
   sowie die Herstellung und der Vertrieb von Anlagen und Fahrzeugen zur Aufbereitung und zum Transport von Gülle.



```
<?xml version="1.0"?>
<XJustiz_Daten xmlns="http://www.xjustiz.de" xmlns:xdo="http://www.xdomea.de"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.xjustiz.de xj_0400_register_1_9.xsd">
    <Grunddaten XJustizVersion="1.20.0">
        <Verfahrensdaten>
            <Instanzdaten>
                <Instanznummer>0</Instanznummer>
                <Sachgebiet xsi:type="WL_Sachgebiet" wl_version="1.6" wl_fassung="0">
                    <content>Handelsregister</content>
                </Sachgebiet>
                <Sachgebietszusatz></Sachgebietszusatz>
                <Instanzbehoerde>
                    <Gericht xsi:type="WL_Gerichte" wl_version="1.21" wl_fassung="1">
                        <content>R2404</content>
                    </Gericht>
                </Instanzbehoerde>
                <Abteilung>2b</Abteilung>
                <Aktenzeichen>HRB 5363</Aktenzeichen>
                <Kurzrubrum></Kurzrubrum>
                <Verfahrensgegenstand>
                    <Gegenstand>
                        <content>Strukturierter Registerinhalt</content>
                    </Gegenstand>
                </Verfahrensgegenstand>
```



```
{
  "news" : [ {
    "sophoraId" : "marktbericht-dax-kurse-aktuell-dow-fed-ezb-100",
    "externalId" : "a6c39b31-85e0-469e-a059-c9510e4e38b5",
    "title" : "Pattsituation am Aktienmarkt",
    "date" : "2023-06-07T18:15:33.660+02:00",
    "teaserImage" : {
      "title" : "Börsenhändler an der Frankfurter Börse",
      "alttext" : "Börsenhändler an der Frankfurter Börse",
```



```
<?xml version="1.0" encoding="UTF-8"?>
<rss version="2.0" xmlns:atom="http://www.w3.org/2005/Atom" xmlns:content="http://purl.org/rss/1.0/modules/content/" xmlns:media="http://search.yahoo.com/mrss/">
  <channel>
    <xhtml:meta xmlns:xhtml="http://www.w3.org/1999/xhtml" name="robots" content="noindex,follow" />
    <atom:link href="https://www.handelsblatt.com/contentexport/feed/wirtschaft" rel="self" type="application/rss+xml"/>
    <title>Handelsblatt Online - Wirtschaft - Politik, Unternehmen und Finanzen</title>
    <link>http://www.handelsblatt.com</link>
    <description>Das Handelsblatt ist die größte Wirtschafts- und Finanzzeitung in deutscher Sprache. Rund 200 Redakteure, Korrespondenten und ständige Mitarbeiter rund um den Globus sorgen für eine aktuelle,
umfassende und fundierte Berichterstattung.</description>
    <language>de-de</language>
    <copyright>Handelsblatt GmbH</copyright>
    <lastBuildDate>Wed, 07 Jun 2023 19:39:11 +0200</lastBuildDate>
    <ttl>113</ttl>
    <image>
      <url>https://www.handelsblatt.com/images/rss_logo_exklusiv/3595600/1-formatOriginal.jpg</url>
      <title>Handelsblatt Online - Wirtschaft - Politik, Unternehmen und Finanzen</title>
      <link>http://www.handelsblatt.com</link>
    </image>
```

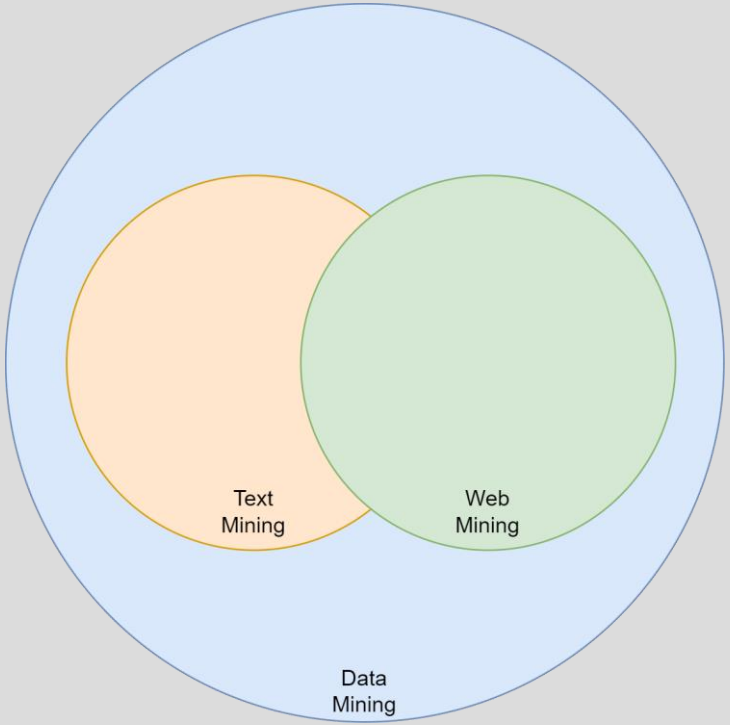

Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

Grundlagen: Web Mining

- Als Teildisziplin des Data Minings zu betrachten
- Web Mining – Informationsgewinn aus:
  - Audio
  - Video
  - Bilder
  - Text

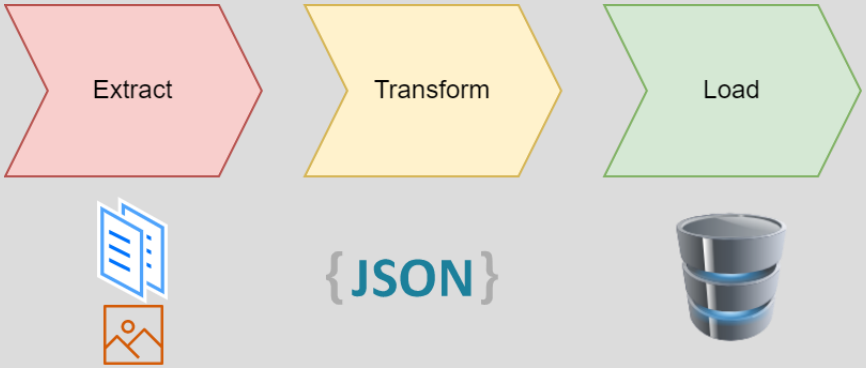
  Im Web
- ETL (**E**xtract, **T**ransform, **L**oad) Prozesse benötigt



Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

Grundlagen: ETL-Prozess



- Rohdaten beziehen
- Daten aufbereiten z.B. via Text Mining
- Ablage im Zielsystem z.B. NoSQL DB

Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

Grundlagen: Rechtliche Rahmenbedingungen

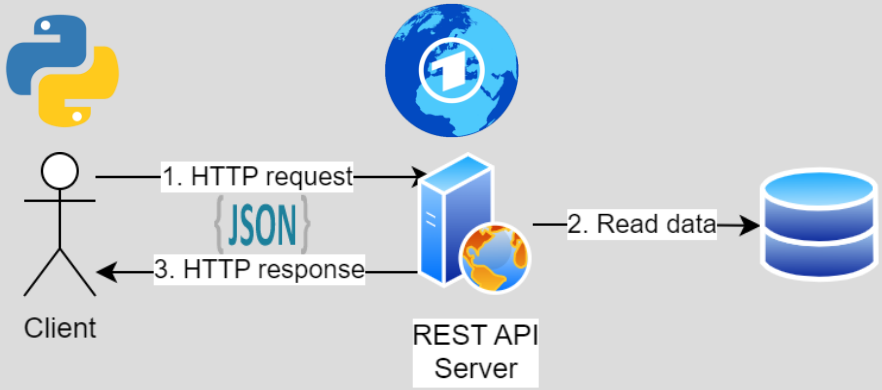
- Legal, wenn die Daten **frei zugänglich** im Internet zur Verfügung stehen
- Urheberrecht bei Verarbeitung und Weitergabe dennoch zu berücksichtigen
- Technische Maßnahmen, die das Web Scraping verhindern (z.B. Captchas), dürfen **nicht** umgangen werden
- Gemäß urheberrechtlichen Wissenschaftsschranke dürfen bis zu 75% geschützter Werke verwendet werden, wenn es sich um eine nicht-kommerzielle, wissenschaftliche Auswertung handelt

Quelle: Uni Hamburg

Automatische Daten Extraktion aus Web-Quellen

Informationsgewinnung via APIs

- i.d.R. RESTful APIs
- Anfrage durch Client mittels HTTP
  - Get
  - Put
  - Post
  - Delete
- Stateless
- Ermöglicht Trennung von Aufgaben, jedoch leichte Anbindung via Standard



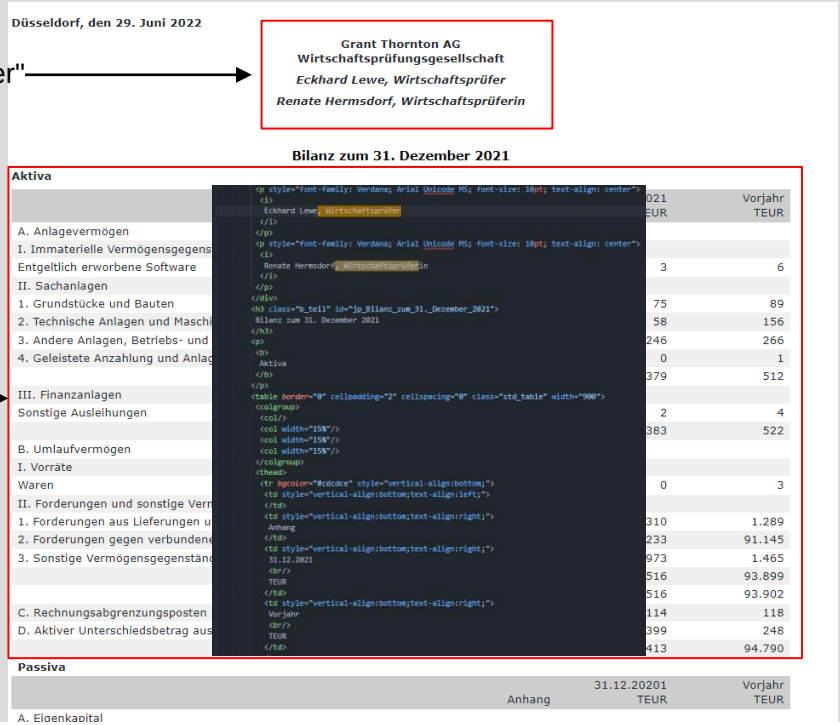
Automatische Daten Extraktion aus Web-Quellen

Informationsgewinnung via Web Scraping

- Traversieren des HTML-Dokumentes
- Auslesen von Tags (z.B. div, p, h1)
- Erweiterung: Web Crawling = Verfolgen der Links

r''[a-z A-ZöÖüÜäÄß,.'-]+, Wirtschaftsprüfer''

<b>.content = “Aktiva”;
<table>.class = “std_table”

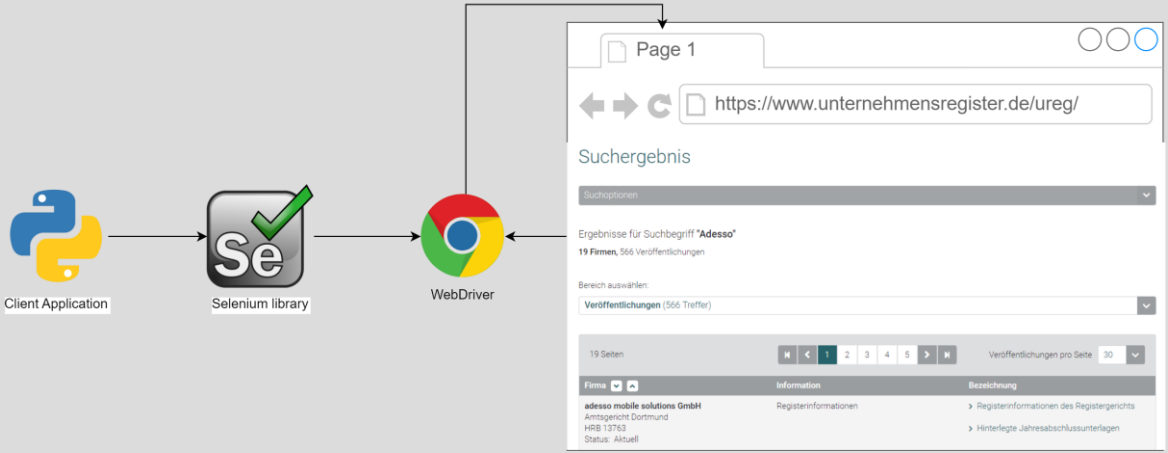


Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

Informationsgewinnung via Web Automation

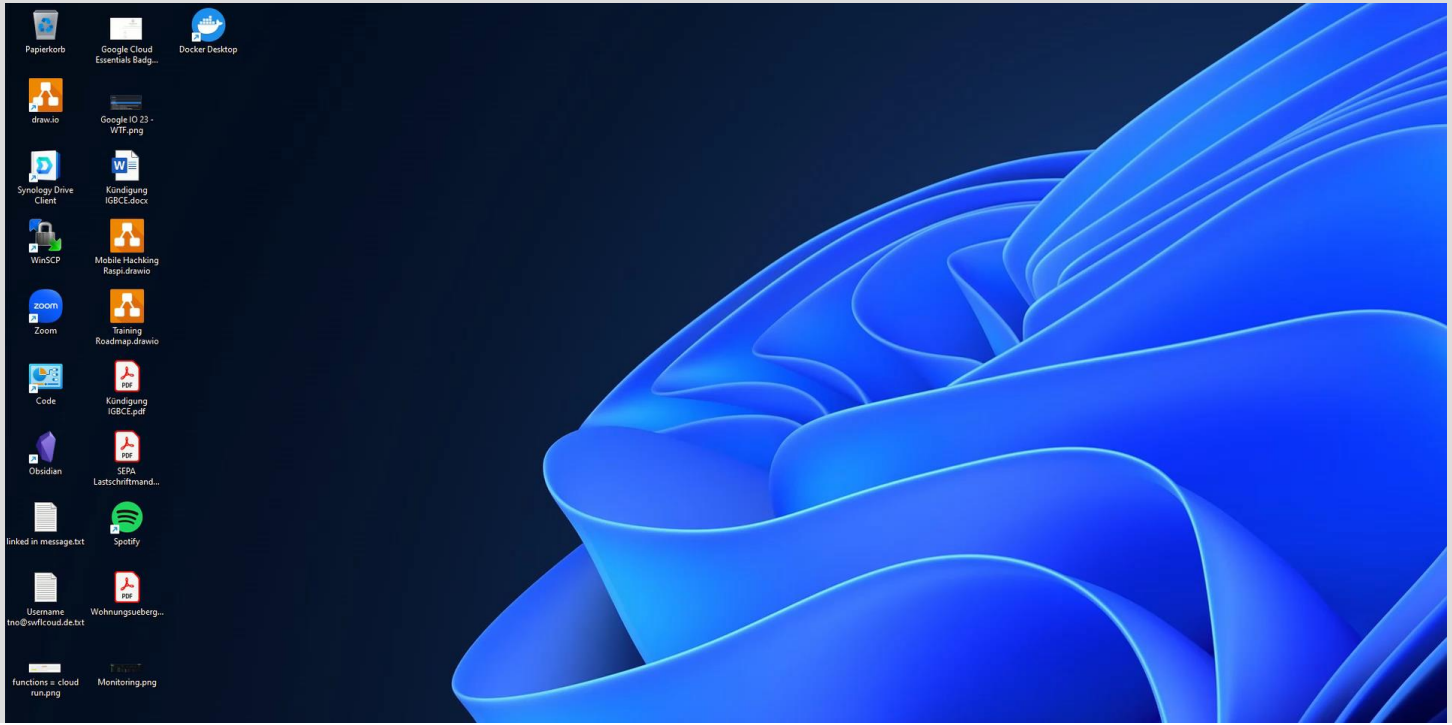
- Simuliert Interaktion des Nutzers mit der Applikation
- Lösung, wenn keine API o.ä.
- Ermöglicht durch Selenium und Webdriver



Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

Informationsgewinnung via Web Automation



Fachhochschule Südwestfalen
University of Applied Sciences

Automatische Daten Extraktion aus Web-Quellen

Informationsgewinnung via Web Automation



~ 50 Stunden pro Wildcard Suche * (26 Buchtaben + 10 Ziffern)

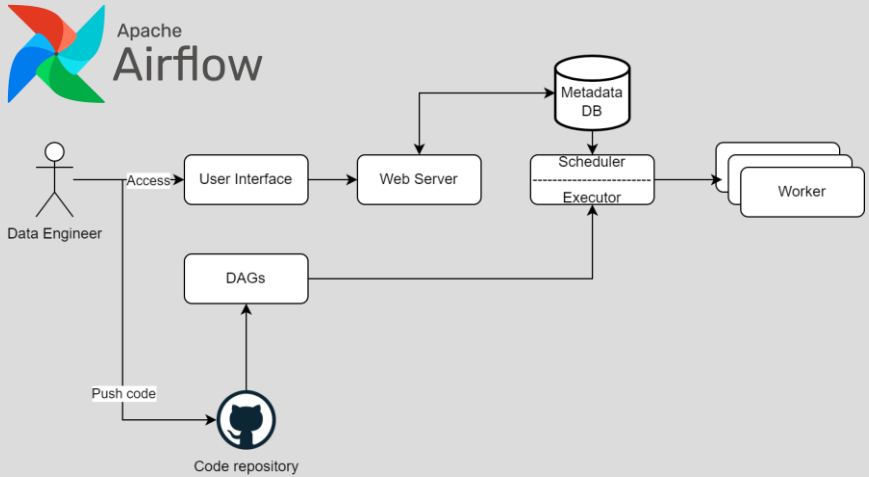
= 1.800 Stunden Gesamtlaufzeit

= 75 Tage Gesamtlaufzeit

Automatische Daten Extraktion aus Web-Quellen

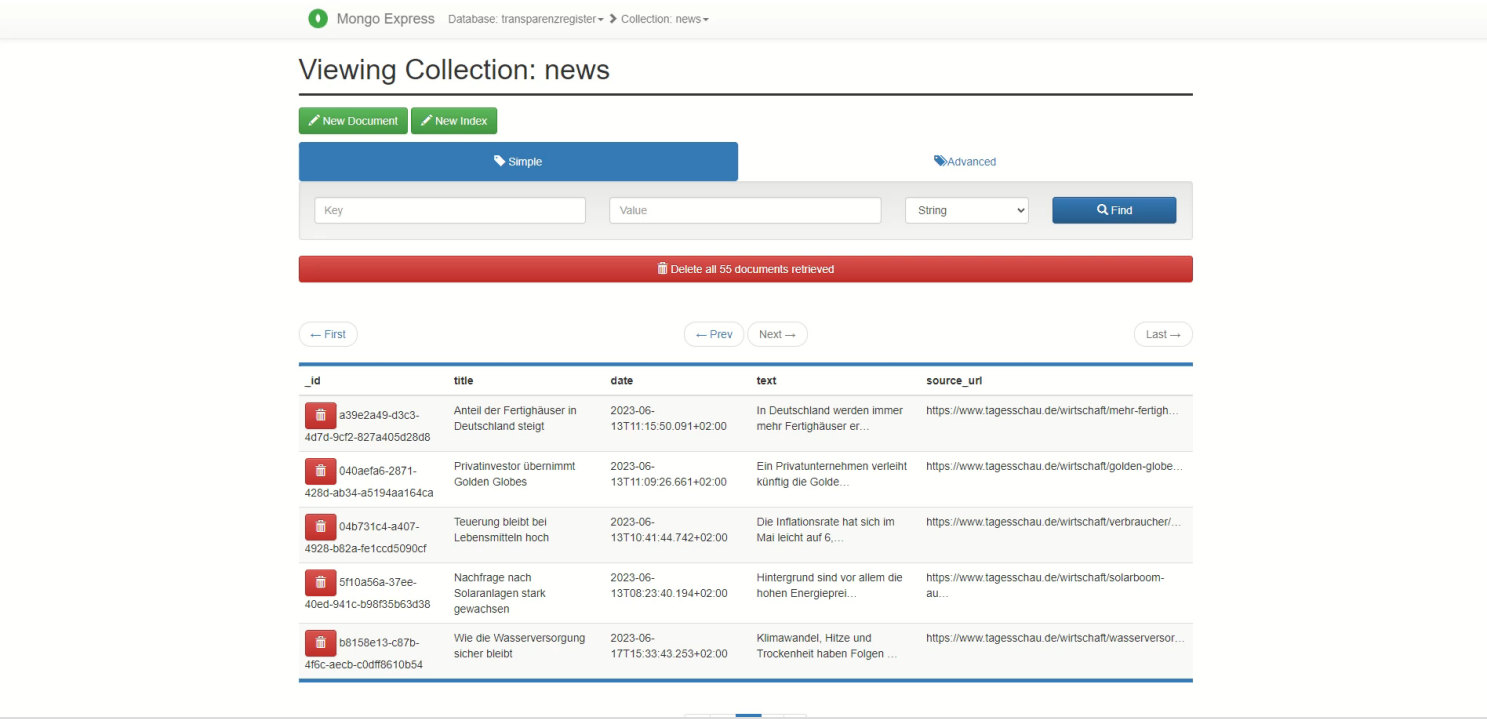
Orchestrierung

- Apache Airflow
- „Workflow-Automatisierung“
- Zentrales Element – DAG:
  - **D**irected
  - **A**cyclic
  - **G**raph
- Funktionen:
  - Ablaufplanung
  - Durchführung
  - Überwachung



Automatische Daten Extraktion aus Web-Quellen

Orchestrierung



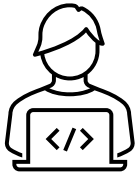
Automatische Daten Extraktion aus Web-Quellen

Ausblick

- „Puzzle-Stücke“ kombinieren
  - Unternehmensregister (Stammdaten) → Bundesanzeiger (Bewegungsdaten)
- Erweiterung um RSS-Feeds anderer Anbieter (z.B. Handelsblatt)
- Abgleich mit dem Team
  - Welche Daten werden benötigt?
  - Erstellung zentraler Model-Klassen
  - Wie organisieren wir unser Mono-Repo?
- Verbesserungen:
  - Code-Qualität
  - Tests
  - Multi-Threading der Daten-Extraktion und Auswertung

Automatische Daten Extraktion aus Web-Quellen

Danke fürs Zuhören



Code