08.04.2021

**Projektarbeit Meeting 1**

1. Festlegung von Zielen für die Projektarbeit:

* Eine Anwendung entwickeln die dazu in der Lage ist viele Daten zu sammeln und geordnet abzulegen.
* Eine Anwendung entwickeln die sich gut horizontal skalieren lässt und sich gut auf wandelnde Anforderungen anpassen lässt.
* Projektabläufe planen und in einer echten Projektsituation durchführen
* Tiefere Kenntnisse über Dataengineering erlangen
* Einarbeiten in ein Thema über das man wenig Kenntnisse hat und im Prozess weitere Ideen zur Wertschöpfung finden

1. Festlegung Technologien:

* GitHub für Ticketverwaltung und Versionsverwaltung <https://github.com/MuckFoe/DataKraken>
* Hadoop auf Linux
* Apache Ambari für Monitoring und Technologiemanagement
* Scraping über Pythonscript ( Generalisiert nutzbar, GUI zur Konfiguration, Auswahl des zu scapenden HTML-Elements evtl. über Inspector-tool)
* MongoDB/ OrientDB/ HBase/ ElasticSearch für die Datenablage
* Zepelin zur Auswertung der Daten

1. Festlegung der Zieldaten:

* Nachrichten( Metadaten + Artikel)
* Coronazahlen
* Wetter
* Coronamaßnahmen

Von 64 Städten ausgewählt nach den Kriterien Größe und Bundesland. Dabei sollen alle Landeshauptstädte und die größten 3 Städte jedes Bundeslandes abgedeckt werden. Die restlichen Plätze werden dann mit den größten oder besonders interessanten Städten aufgefüllt. Besonders interessant sind hierbei auch Städte in Grenznähe oder in spärlicher besiedelten Bundesländern gelegenen, um zu verhindern am Ende nur Daten aus NRW zu haben.

1. Festlegung eines Vorgehensmodells

* Scrum

1. Erste Epics definieren:
   1. Alle Technologien auf den Nodes deployt
   2. Cluster läuft
   3. Scraper für Coronazahlen
   4. Scraper für Nachrichten
   5. Scaper für Maßnahmen
   6. Scraper für Wetter
   7. Datenauswertungen
2. Festlegung eines wöchentlichen Termins:

* Montags 8:00Uhr