Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt
Fakultät Informatik und Wirtschaftsinformatik

Projektarbeit

Entwicklung einer Data Mining Plattform für Corona Daten

vorgelegt an der Hochschule für angewandte Wissenschaften Würzburg-Schweinfurt in der Fakultät Informatik und Wirtschaftsinformatik zum Abschluss eines Studiums im Studiengang Informatik

N. N.

Eingereicht am: 21. August 2021

Erstprüfer: Prof. Rott
Zweitprüfer: Prof. Fertig

Zusammenfassung

TODO

Abstract

TODO

Danksagung

Inhaltsverzeichnis

1 Einführung

2 Entwicklungsprozess

Beim Entwicklungsprozess wurde sich für eine Scrum artige Lösung entschieden. So werden klassische Artefakte und Tools wie etwa das Daily oder der Backlog entsprechend angepasst angewandt. Anpassung meist im Sinne der zeitlichen Komponente. Da das Projekt parallel während des Semesters abläuft wurde hier das 'Daily', von einem täglichen auf einen wöchentlichen Rhythmus umgestellt. Das Backlog wird durch das Team selbst befüllt da es keinen 'Product Owner' gibt. Genauso wurden die Sprints abgewandelt, durch die zeitliche Entzerrung wird hier direkt auf eine dynamische Sprintlänge gesetzt die jeweils zum Sprint-Anfang abgestimmt worden ist.
Gänzlich verzichtet wurde auf die Sprint Retrospektive, zwar hat sich im Laufe des Projekts der Prozess dynamisch angepasst, allerdings wurde hier nicht explizit ein Artefakt bzw. Termin durchgeführt. Änderungen am Prozess wurden kurzfristig und in direkter Absprache mit allen Teammitgliedern beschlossen.

3 Datenablagekonzept

Für die Datenablage wurde zunächst ein entsprechendes Konzept entwickelt. Dieses Konzept basiert auf den Daten die abgelegt werden sollen. Hierzu wird zunächst ein Überblick über die zu ablegenden Daten gegeben:

- Corona Nachrichten/Artikel
- Corona RKI Daten
- Corona Maßnahmen
- Wetterdaten

Jede einzelne Datenquelle wird nun genauer beleuchtet und ein entsprechendes Datenablagekonzept erarbeitet. Da die Konzepte sich aber sehr ähneln wird der Hauptteil der Erklärung bei der grössten Komponente den Corona-Artikeln zu finden sein. Grundsätzlich werden alle Daten sowohl in Elasticsearch als auch im HDFS abgelegt. Wobei das HDFS hauptsächlich für Data Governance Zwecke genutzt wird. Bedeutet das hier die Rohdaten unverändert gespeichert werden sollen um später immer noch Zugriff auf die Original Daten zu haben, was eine Art 'sanity-check' der Daten ermöglicht.

3.1 Corona Nachrichten/Artikel

Die Corona Nachrichten bzw. Artikel werden in zwei unterschiedlichen Technologien abgelegt. Zunächst betrachten wir die Rohdaten. Wie bereits erklärt wird jeder Artikel in Rohformat gespeichert. Diese Daten sind aber für die Auswertung und zur Übersicht der Datensammlung erst mal unwichtig. Dies bedeutet das diese Daten nicht in einer Datenbank indiziert, sondern nur im HDFS abgelegt werden. Die Geschwindigkeit des HDFS für den Datenzugriff ist ausreichend um die Daten bei einer genauen Auswertung ad-hoc zu lesen. Streng genommen könnte man sogar argumentieren dass das HDFS nichts anderes als ein 'key-value store' ist. So wird im Hintergrund eine Datei die unter einem gewissen Pfad abgelegt worden ist für den User als klassischer Dateipfad angezeigt (Ordner durch 'forward-slashes' getrennt und am Ende des Pfades

ein Dateiname), intern aber der Pfad einen key darstellt. Dies ist nötig um die Verteilung und Ausfallsicherheit der Daten über mehrere Knoten gewährleisten zu können. Allerdings ist dies für den User faktisch nicht bemerkbar, da dieser immer mit den entsprechenden Pfaden arbeitet. Um nun den Kreis zu schliessen werden nicht nur Meta-Daten der Artikel gespeichert sondern auch der komplette Artikel im Rohformat um Datenverlust vorzubeugen und die Konsistenz der Daten später noch prüfen zu können. Die Rohdaten werden dann abgelegt unter dem Pfad: '/datakraken/articles/$bundesland$/$Zeitung$/$Datum$/$ArtikelId$_$TimeStamp$'. Dieser Pfad wird dann zu den Meta-Daten hinzugefügt. Nun zu den Meta-Daten, diese kommen werden vom entsprechenden Scraper erzeugt. Dies Daten werden dann im Elasticsearch Cluster abgelegt unter entsprechendem Index Namen. Dies hilft dabei eine Übersicht zu den Daten zu erhalte, einen Status zu bekommen in welchem Mass die Daten in das System kommen und ermöglicht eine rudimentäre Analyse. Die Daten an sich sind definiert durch die entsprechende Config

3.2 Corona RKI Daten

Die wichtigsten Basis Informationen über den Status der Corona-Pandemie in Deutschland sind höchst wahrscheinlich Inzidenz-Wert und Impfquote. Diese Informationen werden in diesem Projekt über das Robert-Koch-Institut bezogen. Über eine öffentliche API kann nicht nur Inzidenzwert pro Bundesland, sondern auch per Bezirk bezogen werden. Ausserdem sind alle Informationen zum derzeitigen Impfstand in Deutschland, sowie Informationen zu PCR- und Schnelltests in Deutschland verfügbar. Alle Informationen die hier von der API zur Verfügung gestellt werden, werden auch gespeichert. Hier folgt der Scraper dem Allgemeinen Konzept, die Original Daten werden im HDFS abgelegt und dann mit minimaler Veränderung im Elasticsearch indiziert.

3.3 Corona Maßnahmen

Das soeben erwähnte Konzept der Nachrichten wird genauso für die Massnahmen verwendet. Rohdaten werden im HDFS abgelegt, während die beschreibenden Daten im Elasticsearch indiziert werden. Diese Massnahmen werden mit folgendem Pattern abgelegt: '/datakraken/measures/$bundesland$/$Zeitung$/$Datum$/$ArtikelId$_$TimeStamp$'.

3.4 Corona Basis Daten

3.5 Wetterdaten

Auch die Wetterdaten werden im gleichen Stil abgelegt. Heisst die einzelnen Wetterlogs werden als Dokumente im Elasticsearch abgelegt und auch hierüber abgefragt. Allerdings werden die Wetterdaten auch noch entsprechend des allgmeinen Konzepts im HDFS abgelegt. Die Ablage Struktur im HDFS ist in folgendem Pattern: '/datakraken/weather/$Stadt$/$Datum$/$TimeStamp$'.

4 Problemstellung

5 Lösung

6 Evaluierung

7 Zusammenfassung

Abbildungsverzeichnis

Tabellenverzeichnis

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorgelegte Bachelorarbeit selbstständig verfasst und noch nicht anderweitig zu Prüfungszwecken vorgelegt habe. Alle benutzten Quellen und Hilfsmittel sind angegeben, wörtliche und sinngemäße Zitate wurden als solche gekennzeichnet.

N. N., am 21. August 2021

Zustimmung zur Plagiatsüberprüfung

Hiermit willige ich ein, dass zum Zwecke der Überprüfung auf Plagiate meine vorgelegte Arbeit in digitaler Form an PlagScan (www.plagscan.com) übermittelt und diese vorübergehend (max. 5 Jahre) in der von PlagScan geführten Datenbank gespeichert wird sowie persönliche Daten, die Teil dieser Arbeit sind, dort hinterlegt werden.

Die Einwilligung ist freiwillig. Ohne diese Einwilligung kann unter Entfernung aller persönlichen Angaben und Wahrung der urheberrechtlichen Vorgaben die Plagiatsüberprüfung nicht verhindert werden. Die Einwilligung zur Speicherung und Verwendung der persönlichen Daten kann jederzeit durch Erklärung gegenüber der Fakultät widerrufen werden.

N. N., am 21. August 2021