Datenanalyse über Verkehrsdaten und Unfallstatistiken in United Kingdom mithilfe von Logstash, Elasticsearch und Kibana

Ziel:

In dieser Arbeit sollte ein Zusammenhang zwischen Verkehrsunfällen und Daten einer Verkehrszählung in United Kingdom erarbeitet werden.

Idee:

Durch immer mehr Assistenzsystemen in Autos, könnte es sein, dass die Zahl der Verkehrsunfälle im Vergleich zu Verkehrszählungen in den letzten Jahren zurück gehen.

Daten:

Als Basis des Projekts wurden zwei Datensätze von kaggle.com in Form von CSV-Files verwendet.

1. https://www.kaggle.com/silicon99/dft-accident-data
   Dieser Datensatz enthält alle erfassten Verkehrsunfälle in United Kingdom. Zu jedem Datensatz sind der exakte Zeitpunkt des Unfalls, die Längen- und Breitengrade des Unfallorts und die Anzahl der beteiligten Fahrzeuge enthalten. Es sind noch weitere Daten enthalten, die in diesem Projekt nicht verwendet wurden. Das sin zum Beispiel Infos darüber, ob der Unfall tödlich endete oder ob die Polizei dazu gerufen wurde.

2. https://www.kaggle.com/sohier/uk-traffic-counts
   Dieser Datensatz enthält alle durchgeführten Verkehrszählungen in United Kingdom. Für dieses Projekt wurden wieder die Koordinaten und der Zeitpunkt der Zählung verwendet. Außerdem wurde die Anzahl der gezählten Autos verwendet.

Tools:

Verwendete Tools zur Datenauswertung sind:

- Logstash in der Version 6.2.2
- Elasticsearch in der Version 5.6.8
- Kibana in der Version 5.6.8

Weitere Tools wurden zur Hilfe verwendet:

- Docker for Windows in der Version 18.03.0-ce-win59
- GitHub Desktop in der Version 1.1.1

Vorgehensweise:

1. Sowohl für Elasticsearch als auch für Kibana wurden zwei Docker Container installiert, die miteinander verknüpft wurden, um eine Kommunikation zwischen den Tools zu ermöglichen. Logstash hingegen wurde lokal auf dem Rechner verwendet.
2. Die heruntergeladenen Daten in Form von CSV-Files sollten mithilfe von Logstash in Elasticsearch geladen werden. Dazu wurden zwei Logstash-Config-Files geschrieben, die genau beschreiben, wie die CSVs aufgebaut sind. Es wurde angegeben durch welches Zeichen die Spalten getrennt sind, welche Spalten vorhanden sind und welchen Dateityp die Spaten haben. Zusätzlich mussten Besonderheiten angegeben werden, z. B. in welcher Form das Datum vorliegt und welche Integer-Werte die Längen- und Breitengrade sind, um einen Geo_Point Datentyp anzulegen.
3. Nachdem alle Daten in Elasticsearch vorhanden waren, konnten in Kibana die Indexe dazu angelegt werden, sowie ein gemeinsamer Index, der die Daten aus beiden CSVs beinhaltet. Danach konnten die Daten verwendet werden, um grafische Auswertungen durchzuführen.
4. Grafisch wurden zum einen zwei Heatmaps erstellt, die zeigen, wo in United Kingdom mehr und wo weniger Unfälle passiert sind, bzw. Wo mehr oder weniger Verkehr ist. Außerdem wurden Grafen angelegt, die einen Verlauf von Verkehr und Unfällen über die Jahre hinweg darstellt. Für alle Auswertungen zu den Unfällen wurde die Anzahl der beteiligten Fahrzeuge verwendet, analog wurden für alle Auswertungen zum Verkehrsaufkommen die Anzahl der gezählten Autos addiert, um Unfälle und Verkehr auf einer gemeinsamen Basis vergleichen zu können.

Probleme:

1. Das Konvertieren von Längen- und Breitenkoordinaten aus Float-Werten in den Geo_Point Datentyp, der von Kibana benötigt wird, um einen Heatmap zu erstellen.
   Lösung: Es wurde eine zusätzliche Konstruktion erstellt, in die die Werte von Länge und Breite eingefügt wurden und danach wurde diese Konstruktion in den Geo_Point Typ konvertiert.
2. Das Konvertieren von Strings in den Date Typ.
   Lösung: Man kann angeben, wie das vorhanden Datum im String aussieht durch die Verwendung durch Pattern wie zum Beispiel dd-MM-yyyy.
3. Datum war in zwei Verschiedenen Schreibweisen angegeben, sollte aber für eine gemeinsame Auswertung in den gleichen Typ konvertiert werden.
   Lösung: In beiden CSVs wurde jeweils das Datumsfeld geklont und ein neuer, in beiden Dateien gleicher Name vergeben.

Fazit:

Es ist ein Rückgang des Verkehrsaufkommens und der Unfälle zu erkennen. Allerdings fällt die Anzahl der Unfälle im Verhältnis zum Verkehrsaufkommen nicht signifikant stärker. Somit ist entgegen der Annahme, kein zusätzlicher Rückgang der Unfälle durch Assistenzsysteme wie zum Beispiel Rückfahrkamera oder Spurhalteassistent zu erkennen. Allerding hat sich bei der Auswertung eine andere Auffälligkeit gezeigt. In den Jahren 2009 und 2010 ist das Verkehrsaufkommen außergewöhnlich hoch.

Reflexion:

- Gelungenes Projekt
- Vieles gelernt – Arbeiten mit Docker und dem Elastic Stack
- Zeitaufwand wegen ungeplanter Hürden unterschätzt
- Freiraum bei Wahl der Tools und hohe Selbstorganisation
- Gute Zusammenarbeit im Team