Beitrag K: Petra Zimmer, Frank Reussner

Analyse-Lifecycle heterogener Informationen auf Basis von Hadoop und Visual Analytics

Petra Zimmer, TU Darmstadt, petra.zimmer@deutschebahn.com
Frank Reussner, Deutsche Bahn AG, frank.reussner@deutschebahn.com

Abstract

Gaining an insight on the company's mass of data was a common goal in the last few years. But information is growing exponentially and companies yearn for a data management system that is able to work with heterogenic data from different sources. A possible answer is the Hadoop Data Platform. With its diverse components, it makes several ways of data management as a foundation for the analysis. The possibilities of Hadoop range from parallelized SQL-Queries to machine learning algorithms. Combined with a visual analytics tool you can gain a deep insight in your own data. This article illustrates an analysis lifecycle of the visual analytics tool Qlik with a connected Hadoop system using the example of European air quality data.

Zusammenfassung

Einblicke in die großen Datenmengen des Unternehmens zu erlangen, wird in den letzten Jahren immer häufiger fokussiert. Die Datenmengen wachsen exponentiell und sind zudem meist in ihrer Struktur heterogen.

Eine Datenverwaltung, die mit heterogenen Informationen aus unterschiedlichen Quellen arbeiten kann, ist daher wünschenswert. Ein mögliches Datenmanagementsystem ist die Hadoop-Plattform. Die vielfältigen Komponenten von Hadoop ermöglichen verschiedene Arten des Datenmanagements als Grundlage für die Analyse. Von parallelisierten SQL-Abfragen bis Machine Learning Algorithmen spannen sich die Möglich-

keiten von Hadoop. In Kombination mit Visual Analytics Tools bietet sich ein tiefer Einblick in die eigenen Daten. Anhand des Visual Analytics Tools Qlik mit angeschlossenen Hadoop zeigt dieser Artikel einen Life-Cycle einer Analyse am Beispiel von europäischen Luftqualitätsmessungen.

Bei der Deutschen Bahn AG ist die Digitalisierung und damit das Thema des Umgangs mit diesen Datenbergen ein Vorstandsanliegen. Die Abteilung DB Analytics, die in der Konzernleitung bzw. Konzernentwicklung angesiedelt ist, unterstützt in diesem Zusammenhang die verschiedensten Geschäftsfelder der Bahn. Kernaufgaben von DB Analytics sind: Analyse, Prognose, Simulation und Optimierung. Auch der Bereich Visual Analytics wird als Leistung in den Konzern angeboten. Diese Arbeit wurde in Zusammenarbeit mit DB Analytics realisiert.

1 Einleitung

Immer mehr Unternehmen versuchen einen Mehrwert aus ihren Daten zu generieren und die daraus resultierenden Informationen in ihre Entscheidungsprozesse einfließen zu lassen. Laut eine repräsentativen Umfrage von Bitkom Research basieren 80% der gefragten Unternehmen ihre relevanten Entscheidungen zunehmend auf Erkenntnissen aus der Analyse von Daten. Gut zwei Drittel der Unternehmen benennen Datenanalysen als einen zunehmend entscheidenden Baustein für ihre Wertschöpfungskette, vgl. [Bitkom, 2016].

Alleiniges Ansammeln von Daten genügt nicht, um Fragestellungen gezielt zu beantworten und somit Wissen aus den Daten zu extrahieren.

Dieses Paper zeigt eine Vorgehensweise mit verschiedenen Datenhaltungs- und Analyse-Tools, die dem Big Data Umfeld zugeordnet werden, um iterativ mehr Wissen durch die Analyse der Daten zu erhalten.

2 Analyse-Lifecycle

Ein Analyse-Lifecycle zur Beantwortung einer bestimmten Fragestellung enthält die vier Schritte Datengenerierung bzw. -sammlung, Datenmanagement, Datenanalyse und Interpretation (siehe Abbildung K-1). Meistens stellen sich durch die ersten Ergebnisse neue Fragen, die durch weitere Iterationen des Zyklus zunehmend beantwortet werden können. Expertenwissen über die Daten und insbesondere

Geschäftswissen sind allerdings unverzichtbar, um die richtigen Schlüsse aus den Ergebnissen zu ziehen [Zikopoulos, 2011].

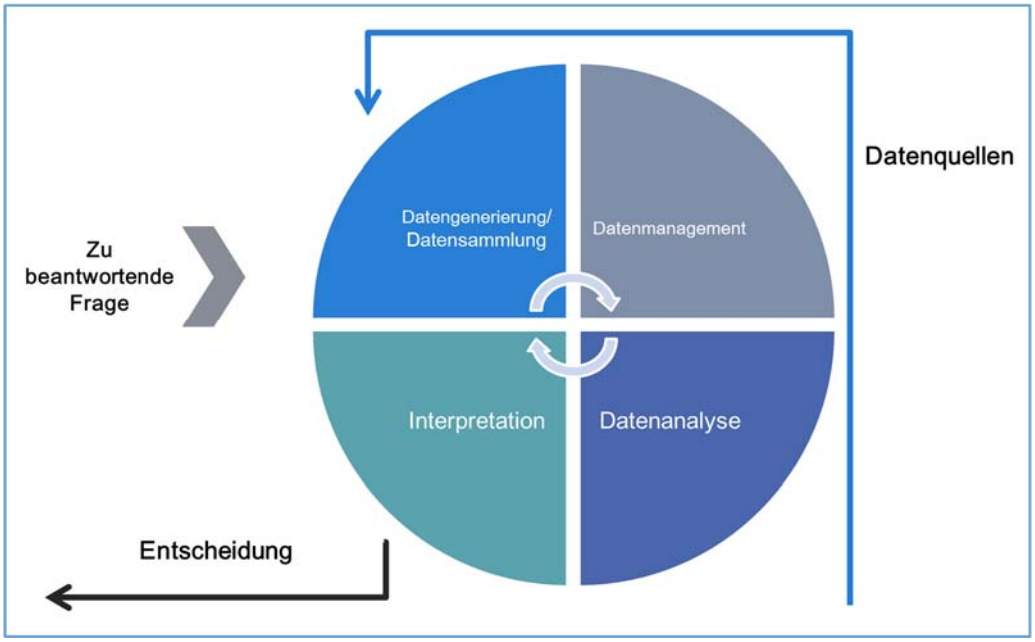

Abbildung K-1: Übersicht über das iterative Vorgehen des Analyse-Lifecycles, vgl. [Feras, 2016]

Als Beispiel wird eine Ansammlung von frei verfügbaren Umweltdaten von eurostat verwendet, die Bevölkerungsdaten, Luftmesswerten, sowie Daten zur Flächennutzung und Flächenbedeckung von EU-Ländern umfassen [Eurostat, 2016].

Ausgehend von der Fragestellung, wie die Luftwerte mit den Einwohnerzahlen und der Bewirtschaftung der Landfläche zusammenhängen, wurde dieser Analyse-Lifecyle konstruiert.

Im Folgenden werden zunächst mögliche Datenmanagement- und Analyse-Tools beschrieben und anschließend wird die Verwendung derer exemplarisch mit den Daten gezeigt.

2.1 Hadoop-System – Hortonworks

Das verwendete Hadoop-System ist ein Cluster mit der *Hortonworks* open-source Dataplattform [Hortonworks, 2016]. Die *Hortonworks* Dataplattform besteht aus verschiedenen Komponenten, die sowohl im Hadoop-System als auch nach außen viele Schnittstellen für verschiedene Anbindungen bietet (siehe Abbildung K-2).

Die einzelnen Komponenten unterscheiden sich in der Geschwindigkeit der Datenverarbeitung, ob die Daten parallel bearbeitet werden und ob inhomogene Daten auch verarbeitet werden können. Insbesondere zeichnet sich Hadoop durch das zugrundeliegende *Hadoop File System (HDFS)* aus, das die Daten auf den einzelnen Knoten des Clusters so organisiert, dass die Berechnungen daten-parallel ausgeführt werden können.

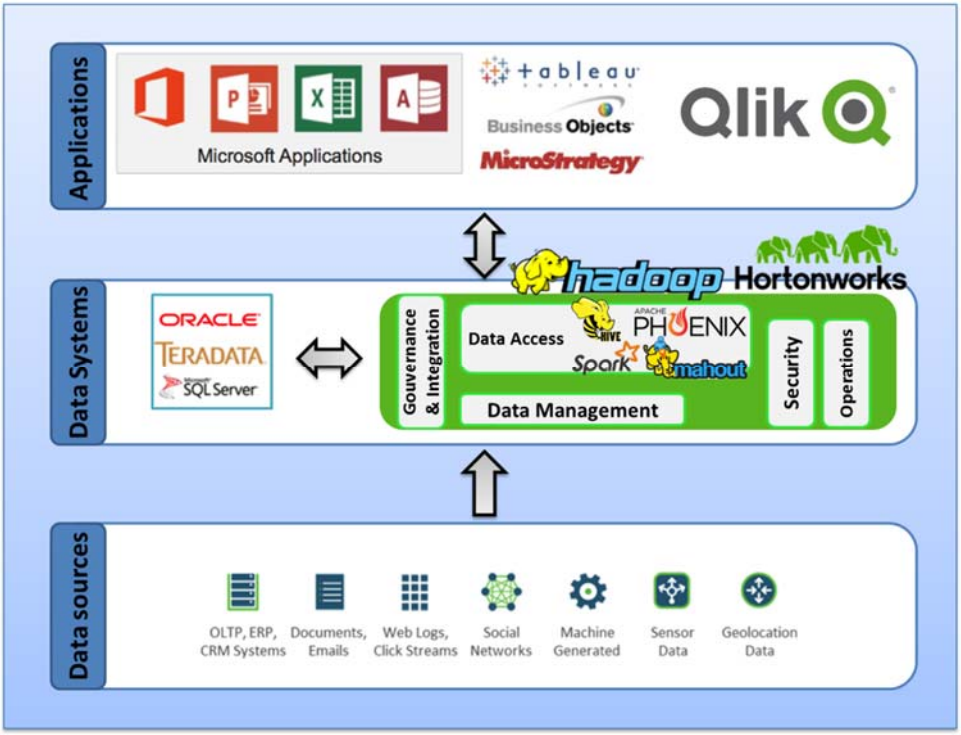

Abbildung K-2: Übersicht über die Hortonworks Dataplattform und einige der möglichen Schnittstellen [Hortonworks, 2016]

2.2 Visual Analytics – QlikView bzw. Qlik Sense

Das Visual Analytics Tool *QlikView* bietet dem Benutzer die Möglichkeit die Daten interaktiv und visuell zu erkunden. Es gehört zu den sogenannten Self-Service Business Intellegence Tools, eignet sich aber ebenfalls für eine geführte Datenanalyse. Im Folgenden gilt das Gezeigte neben *QlikView* auch für *Qlik Sense*, welches den Self-Service-Gedanken gegenüber *QlikView* weitertreibt.

2.3 Datenmanagement mit Hadoop

Die *Hive* Komponente des Hadoop-Systems bietet ein Data-Warehouse, auf das mit verschiedenen Komponenten des Hadoop Clusters zugegriffen werden kann. Zudem

gibt es eine SQL-ähnliche Abfragesprache für ad-hoc Analysen, die im Hadoop-System als ein Map-Reduce-Job ausgeführt werden, sodass die Anfrage über die auf dem Cluster verteilten Daten eine Antwort liefert, vgl. [Thusoo, 2010].

Hbase ist die NoSQL-Datenbank der Hadoop-Plattform, die sich auch mit *Hive* verbinden lässt. Mit *Hbase* können sowohl NoSQL-Anfragen bearbeitet werden als auch über *Phoenix* sehr schnelle SQL-Abfragen auf der NoSQL-Datenbank ausgeführt werden.

3 Analyse

Apache Spark ist eine in-memory Datenverarbeitung mit vielfältigen Bibliotheken. Die Bibliotheken lassen sich mit den Programmiersprachen Scala, Python oder Java mit Map-Reduce-Funktionen, Schnittstellen zu *Hive* oder *Hbase* mit SQL oder Machine-Learning-Bibliotheken nutzen, vgl. [Owen, 2012].

Das Apache Projekt *Mahout* bietet skalierbare Machine-Learning-Algorithmen, die nicht nur mit Map-Reduce im Hintergrund laufen, sondern auch Standalone arbeiten können. Mahout ist auf Algorithmen zur Klassifikation und zum Clustern spezialisiert, mittlerweile werden aber auch andere mathematische Funktionen implementiert, vgl. [Meng, 2016].

QlikView ist wie bereits erwähnt ein Business Intelligence Tool, das ebenfalls in-memory arbeitet und mit dessen Hilfe sich ein Mehrwert aus Daten generieren lässt. Als Datenquellen können verschiedene textbasierte Dateiformate oder aber auch Datenbanken, darunter auch das *Hadoop Data Warehouse*, genutzt werden. Nach dem Importieren und Modellieren der Daten können diese ohne weitere Duplikationen in verschiedenen Diagrammarten visualisiert werden. Das Besondere sind die interaktiven Selektionen, die zum einen schnell die Auswahl auf alle Diagramme übertragen und zum anderen nicht hierarchisch strukturiert sind, sodass die Daten leicht in beliebigen Kategorien verglichen werden können.

Visuelle Analysen bieten den Vorteil, dass man sich schnell einen Überblick über große Datenmengen verschaffen kann, Trends grafisch erkennen kann und mit dem notwendigen Fachwissen auch diese Trends gezielt untersuchen kann.

RHadoop bietet die Möglichkeit *R* Befehle auf dem Hadoop-Cluster datenparallel auszuführen bzw. teilweise sind auch einige Algorithmen parallelisiert implementiert. Somit kann man verschiedene statistische Modelle auch mit größeren Datenmengen antrainieren oder eine Dimensionsreduktion auf sehr großen Datensets ausführen. Genauso wie bei der Standalone *R* Variante werden die meisten Operationen in-memory ausgeführt.

4 Datenmodellierung und Transformation

Zunächst wurden verschiedene Datentabellen von [Eurostat, 2016] heruntergeladen und anschließend mit *Qlik* so modelliert, dass zu einer Luftmessstation die gemessenen Werte, die Flächennutzung der Region, die landwirtschaftliche Produktion, die Bevölkerungsdichte und Bebauung der Landschaft jeweils für die Jahre 2010 bis 2012 verfügbar sind.

Diese vielfältigen Daten sind auf unterschiedlichen Unterteilungsebenen der Länder vorhanden und werden so modelliert, dass am Ende eine große Tabelle mit vergleichbarer Skalierung für die einzelnen Messwerte entsteht, d.h. die Werte liegen für dieselbe Unterteilungsebene vor.

Die Ergebnistabelle wird dann im *Hive* Datawarehouse abgespeichert und anschließend von dort als Quelle für das Machine Learning mit *Hadoop Spark* genutzt. Zur Klassifizierung wird der Decision-Tree-Algorithmus aus der mllib von *Spark* verwendet. Anschließend werden die Ergebnisse mit *QlikView* visualisiert.

Aufgrund der Berechnung des Luftqualitätsindexes aus dem Ozon- (O_3), Stickstoffdioxid- (NO_2), Schwefeldioxid- (SO_2), Kohlenstoffmonoxid- (CO) und dem Feinstaubwert (PM_{10}) haben wir uns dafür entschieden, im Folgenden nur diese Luftwerte zu betrachten. Des Weiteren haben wir uns dafür entschieden, mit Hilfe von *Apache Spark* und dem dort implementierten Decision-Tree-Algorithmus für die fünf verschiedenen Luftwerte jeweils ein Modell zur Klassifizierung zu trainieren. Dafür wurden die kontinuierlichen Luftwerte, die als täglicher Jahresdurchschnitt vorhanden sind, von den kontinuierlichen Werten auf diskrete Werte nach den Wertegrenzen für den Luftqualitätsindex abgebildet [LUBW, 2016], sodass die Werte jeweils in sechs Kategorien, Schulnoten, eingeteilt sind. Nach der Modellierung mit *Qlik* wurden die Daten auf das *HDFS* kopiert und anschließend mit *Apache Spark* weiter verarbeitet.

Für alle fünf Luftwerte wurden aus allen Datensätzen mit nichtleeren Luftwerten zufällig 70% der Daten als Trainingsdaten gezogen, die zum Trainieren des jeweligen Modells genutzt wurden. Nach dem Trainieren des Decision-Tree-Modells wurden die distinkten Testdaten dazu genutzt, die Fehlerrate der Modelle jeweils zu bestimmen. Anschließend wurden die Ergebnisse mit einem *Hortonworks ODBC Treiber* von *Qlik* abgerufen, um so den Zusammenhang der Kenngrößen mit den vorhergesagten Werten zu evaluieren und validieren.

5 Ergebnisauswertung mit Visual Analytics

Ein wichtiges Ergebnis des Lifecycles ist, dass die verschiedenen erlernten Modelle alle eine Abhängigkeit zu der Bevölkerungsdichte zeigen. Dies entspricht der intuitiven Vermutung, da größere Städte mehr Kraftverkehr haben und somit der Feinstaubwert auch höher sein sollte.

Die erlernten Modelle für die einzelnen Messwerte haben Probleme mit ungleichen Häufigkeitsverteilungen der ursprünglichen Werte für den Messwert, da bei der zufälligen Ziehung die Häufigkeitsverteilung der Werte das Ergebnis der Ziehung beeinflusst. So zeigt zum Beispiel die Konfusionsmatrix für den Luftwert O_3 in Abbildung K-3, dass der Wert 2 insgesamt viel häufiger auftritt, als die Werte 1 oder 3. Somit ist es nicht verwunderlich, dass tendenziell eher der Wert 2 zugeordnet wird. In einer weiteren Iteration sollte man somit die Trainingsdaten passend gewichten, um die unterschiedlichen Häufigkeiten auszugleichen.

Auf alle Daten mit nicht-leerem O_3-Wert ergibt sich der Fehler des Modells zu 16,9%. Aber auch wenn die Modelle nicht alle Werte abbilden, so kann dennoch Wissen über die Messwerte erlangt werden.

Wert	Zuordnung	1	2	3	Summe
	1	1	144	0	145
	2	0	3957	146	4103
	3	0	569	264	833
Summe		1	4670	410	5081

Abbildung K-3: Konfusionsmatrix für den Luftwert O_3, Färbung nach prozentualem Anteil der Werte-Klasse

Bei dem Feinstaubwert PM10 kann trotz der Konfusionsmatrix in Abbildung K-4, die zeigt, dass die Werte 1 und 5 vom Modell nicht erkannt wurden, ein Trend vorhergesagt werden, ob der Messwert eher zu den besseren Werten gehört (Werte 1 und 2) oder zu den schlechteren (Werte 4 und 5).

Es ergibt sich der resultierenden Gesamtfehler von 31,6%. Die Werte, von denen der Entscheidungsbaum seine Zuordnung abhängig macht, sind die Populationsdichte, der Getreidelandanteil und die damit erzielte Ernte. Wenn man betrachtet, wie die wahren Werte mit den angegeben Parametern zusammenhängen, lässt sich feststellen, dass bei 73% der Datenreihen mit PM10-Wert von 3 und 56% der Datenreihen mit PM10-Wert von 4 die Populationsdichte größer als 0.14 Personen/km² und das Getreideland kleiner als 33.1 ha ist.

Wert	Zuordnung	2	3	4	Gesamt
	1	36	35	0	**71**
	2	398	1067	3	**1468**
	3	211	3828	131	**4170**
	4	7	460	251	**718**
	5	0	70	48	**118**
Gesamt		**652**	**5460**	**433**	**6545**

Abbildung K-4: Konfusionsmatrix für den Luftwert PM10, Färbung nach prozentualem Anteil der Werte-Klasse

Für Stickstoffdioxid lassen sich die Werte 2, 3 und 4 durch eine Waldfläche kleiner als 45 ha klassifizieren, wobei damit 91% der Daten abgedeckt sind.

Andererseits liefert der gebildete Entscheidungsbaum für PM10 auch eine Zuordnung zu den Ländern Tschechien, Romänien, Ungarn, Italien und Polen für den erlernten Wert 4 für alle wahren Werte größer gleich 4, die somit dadurch klassifiziert werden. Damit lassen sich also durch das Machine Learning neue Ähnlichkeiten zwischen Ländern finden.

Nicht nur durch die Untersuchung der Abhängigkeit der Messwerte zu dem Decision-Tree-Vektor, sondern auch die Untersuchung der Extremwerte kann neue Erkenntnisse liefern.

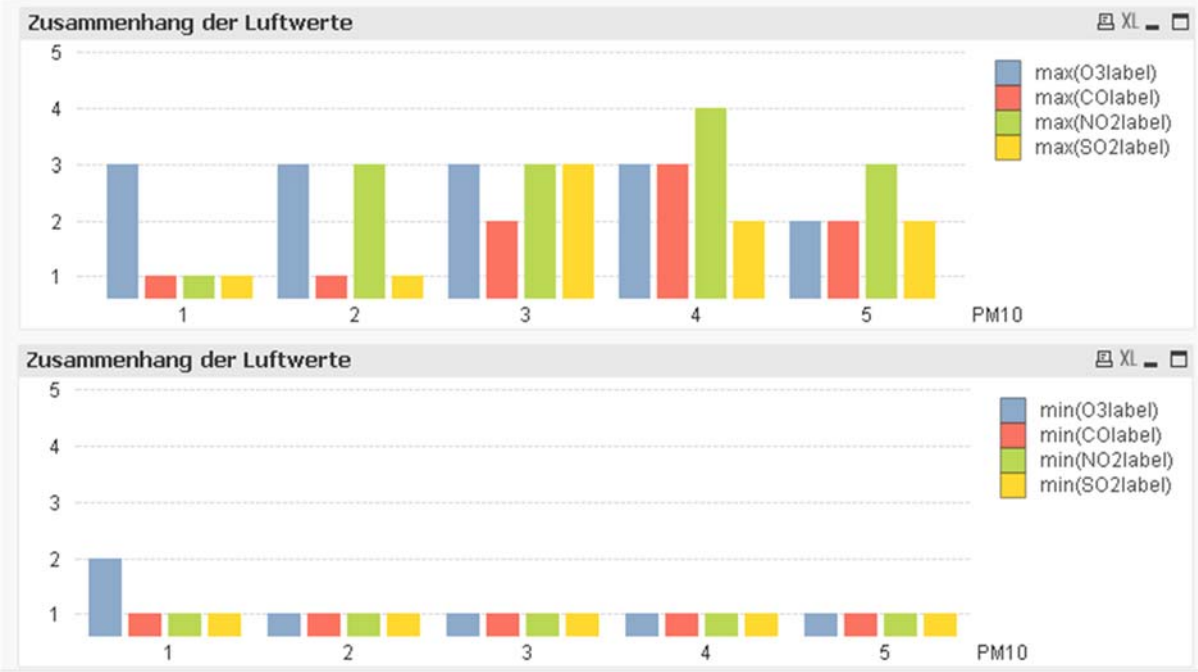

Abbildung K-5: Zusammenhang von PM10 zu den Extremwerten der anderen Luftwerte

Aus dem Zusammenhang zwischen dem PM10 Wert und dem O_3-Wert aus Abbildung K-5 wird ersichtlich, dass der Wert 1 für PM10 nur auftritt, wenn der O_3-Wert mindestens 2 beträgt.

In Folge dessen lässt sich schlussfolgern, dass mit einem sehr naiven Ansatz schon erste Zusammenhänge als Mehrwert gefunden werden können. Um die Ergebnisse des Machine-Learning-Ansatzes akkurater zu machen, wäre es denkbar, eine Dimensionsreduktion durchzuführen, sodass es keine Abhängigkeiten mehr zwischen beschreibenden Attributen mehr geben kann. Des Weiteren kann eine Verbesserung der Zuordnung dadurch erreicht werden, dass die Verteilung der Attribute im Raum untersucht wird, sodass ein Trainingsset aus annährend gleichverteilten, raumabdeckenden Datenreihen besteht.

6 Fazit

Durch *Qlik* konnten Zusammenhänge zwischen den Messwerten und verschiedenen Nutzungsdaten der Landflächen einfach dargestellt und als Muster erkannt werden. Es konnte ebenfalls ein Zusammenhang der einzelnen Messwerte gefunden werden.

Die erlernten Entscheidungsbäume mit *Apache Spark* bestätigten einen Zusammenhang zwischen den Messwerten und der Bevölkerungsdichte.

QlikView zeigte sich hier als gutes Tool, um die unbekannten Daten schnell zu erkunden. Es eignet sich gut, um die unterschiedlich skalierten Werten der diversen Datentabellen vergleichbar auf die Luftmessstationsebene zu modellieren. Für genauere Modelle des Machine Learnings sind sehr feingranulare Informationen für jede Station und jedes Jahr wünschenswert, sodass z.B. die Einteilung in Wasserfläche, Getreidefläche etc. auch auf Messstationsebene oder zumindest auf Regionsebene verfügbar ist, da die Einteilung der Klassifizierung bzw. Cluster vom Datendetailgrad abhängt.

Dieses Paper zeigt nur eine Iteration des klassischen Analyse-Lifecycles. Die Daten wurden gesammelt, transformiert und für eine Analyse aufbereitet. Ein prinzipieller Zusammenhang zwischen den unterschiedlichen Attributen und den Messwerten konnten mit einem ersten einfachen Ansatz gezeigt werden. Allerdings wirft es die Fragen auf, welche Attribute die Luftwerte optimal klassifizieren, ob es Ausreißer gibt, die man aussortieren muss und inwiefern eine abhängige Betrachtung der Luftwerte sinnvoll ist.

Es ist methodisch sinnvoll, den nächsten Lifecycle zunächst mit der Überprüfung der Thesen aus der Interpretation der vorangegangenen Analyseergebnisse zu beginnen, um diese zu validieren.

Mit einem tiefen Fachwissen und passenden Fragestellungen lassen sich jedoch schnell neue Erkenntnisse mit dem Analyse-Lifecycle generieren und validieren, sodass Entscheidungen im Unternehmen mit den Erkenntnissen aus der Datenanalyse begründet werden können.

7 Literaturverzeichnis

[Bitkom, 2016]

https://www.bitkom.org/Presse/Presseinformation/Jedes-dritte-Unternehmen-nutzt-Big-Data.html , abgerufen am 30.07.2016, eine repräsentative Umfrage von Bitkom Research im Auftrag von KPMG unter 704 Unternehmen und 102 Verwaltungen mit mehr als 100 Mitarbeitern ergeben

[Feras, 2016]

Feras A. Batarseh, Eyad Abdel Latif, Assessing the Quality of Service Using Big Data Analytics: With Application to Healthcare, Big Data Research, Volume 4, June 2016, Pages 13-24, ISSN 2214-5796

[Hortonworks, 2016]

http://hortonworks.com/products/data-center/hdp , abgerufen am 30.07.2016

[LUBW, 2016]

http://www4.lubw.baden-wuerttemberg.de/servlet/is/20152/ , abgerufen am 15.05.2016

[Zikopoulos, 2011]

Zikopoulos Paul, Eaton Chris, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data (1st ed.), 2011, McGraw-Hill Osborne Media.

[Thusoo, 2010]

Thusoo A. et al., Hive - a petabyte scale data warehouse using Hadoop, 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), Long Beach, CA, 2010, S. 996-1005

[Meng, 2016]

Meng, Xiangrui, et al. , Mllib: Machine learning in apache spark , JMLR 17.34 , 2016 S. 1-7.

[Owen, 2012]

Owen, Sean, et al. (2012); Mahout in action, Shelter Island, Manning 2012; ISBN-13: 978-1935182689

[Eurostat, 2016]

Eurostat Daten, abgerufen am 15.05.2016:

Bevölkerung: http://ec.europa.eu/eurostat/de/web/nuts/local-administrative-units

Lucas(Bodennutzung/-bedeckung): http://ec.europa.eu/eurostat/web/lucas/data/data-base

Messdaten (Airbase): http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-2

Einheiten und Erklärungen: http://www.eea.europa.eu/data-and-maps/daviz/sds/vocabularies-in-the-database/download.table