Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Big Data

- Notwendige Technologien und Analysetechniken für Unternehmen-

Sarah Bock

Ausarbeitung im Rahmen der Ringvorlesung
„InnovationCity 2030“

Fakultät Technik und Informatik
Department Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Sarah Bock

Big Data

- Notwendige Technologien und Analysetechniken für Unternehmen-

Ausarbeitung im Rahmen der Ringvorlesung „InnovationCity 2030“ SS/2015

im Studiengang Next Media (M.A.)
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Abgegeben am 31.08.2015

Sarah Bock

Thema der Ausarbeitung
Big Data – Notwendige Technologien und Analysetechniken für Unternehmen

Stichworte
Big Data, Data Mining, Metadaten, Cloud Computing, Business Analytics, Datenintegration, Datenanalyse

Kurzzusammenfassung

Der Umgang mit einer enorm großen Menge an Daten stellt Unternehmen vor große Herausforderungen. Unternehmen müssen eine geeignete IT-Infrastruktur schaffen um die Analyse und darauf ausgerichtete Wertschaffung zu ermöglichen. Die Ausarbeitung erläutert die gängigen Technologien, die auf Big Data spezialisiert sind und zeigen zudem mögliche Analysetechniken für den Umgang mit Big Data auf.

Sarah Bock

Inhaltsverzeichnis

1 Einleitung

Die letzten Jahre und bereits Jahrzehnte haben gezeigt, dass Unternehmen, die durch die digitale Disruption zu enormen Erfolg in unmittelbar kurzer Zeit gekommen sind, digital anders aufgestellt sind als herkömmliche traditionelle Konzerne. Sie verfügen oft über wenig Kapital und produzieren meist keine physischen Produkte, nutzen aber die Möglichkeiten, welche die Analyse von Daten bieten und haben enormen Erfolg. Big Data, eine enorme Menge an Daten, die gesammelt, generiert, gespeichert und analysiert werden, sollen die Welt vorhersagbarer machen. Jeder von uns produziert täglich eine unsagbare Menge von Daten. GPS Daten verraten wo wir uns täglich bewegen, welche Seiten wir im Internet aufrufen, was wir bei Google suchen und selbst wo wir was kaufen. Durch die gezielte Nutzung der Analyseergebnisse von Big Data können Unternehmen ganz neue Erkenntnisse gewinnen die zu erheblichen Wettbewerbsvorteilen führen können.

Um diese Vorteile jedoch optimal nutzen zu können, müssen Unternehmen Ihre Unternehmensstrategie auf die Anforderungen der Big Data Architekturen ausrichten. Die in den letzten Jahren immer mehr angewachsene Menge von Daten stellt Unternehmen vor große Herausforderungen.

Im Mittelpunkt der vorliegenden Arbeit steht die Voraussetzung für Unternehmen eine skalierbare und flexible IT-Infrastruktur zu schaffen. Die Arbeit soll somit zu einem differenzierten Verständnis der technologischen und methodischen Software-Voraussetzungen beitragen und einen Überblick über die gängigsten Technologien und Analysemethoden geben. Zu diesem Zweck wird zunächst der Begriff Big Data definiert und auf dessen Bedeutung für Unternehmen eingegangen. Es folgt eine Beschreibung und Bewertung der gängigsten Big Data Technologien, sowie der Analyseverfahren. Abschließend erfolgt eine kurze Zusammenfassung sowie ein Ausblick auf die zukünftige Entwicklung von Big Data und dessen Bedeutung für Unternehmen.

2 Definition und Abgrenzung

In der vorherrschenden Literatur herrscht keine einheitliche Definition des Begriffs „Big Data“, nicht zuletzt weil es sich um eine relativ junge Begrifflichkeit handelt. Big Data wird als eine Kombination aus den Bereichen der BI (Business Intelligence) und Data Warehouse bezeichnet.

In einem der Leitfäden des Bitkom(2014) des Arbeitskreises Big Data wird der Begriff Big Data als „[...]Einsatz großer Datenmengen aus vielfältigen Quellen mit einer hohen Verarbeitungsgeschwindigkeit zur Erzeugung wirtschaftlichen Nutzens bezeichnet.“ Big Data Methoden greifen somit dort wo herkömmliche Technologien der Erfassung, der Speicherung und der Analyse von Daten aufgrund der hohen Datenmengen nicht mehr ausreichen (Bagnoli, Marten & Wagner, 2012).

Um den Begriff Big Data genauer von dem Begriff der Analytics und den bisher üblichen Datenanalysen abzugrenzen werden in der Literatur vier wesentliche Merkmale genannt, die Big Data charakterisieren. Darunter fallen die Datenmenge (Volume), die Datenvielfalt (Variety), die Geschwindigkeit (Velocity) sowie die Analyse dieser Daten und die Wertschaffung aus diesen (Value) (Oracle, 2015). Big Data handhabt die Verarbeitung von Datensätzen die bis in den Yottabyte[1]-Bereich führt. Diese fast unvorstellbaren Größen von Datensätzen und –mengen scheinen zudem in den kommenden Jahren und Jahrzehnten weiterhin zu steigen. Klausnitzer (2013, s. 84) beschreibt den Datenzuwachs in den letzten Jahren als enorm und überproportional. Er sagt, dass 90 Prozent aller heute auf der Welt existierenden Daten erst in den letzten zwei Jahren generiert wurden. Die zur Verwendung stehende Menge an Daten wird somit zu einem so erheblichen Volumen anwachsen, dass immer größere Anforderungen an dessen Analyse gestellt werden muss. Gleichzeitig nimmt nicht nur die Menge an Daten zu, sondern ebenfalls die Vielfalt der Datenstrukturen. Es muss daher eine Vielzahl unstrukturierter Daten berücksichtigt werden, die aus einer Vielzahl unterschiedlichster Quellen resultieren. Des Weiteren ist eine hohe Geschwindigkeitsrate sowohl bei der Speicherung und Generierung der Daten, als auch deren Verwertung und

[1] Yottabyte (YB) umfasst umgerechnet 10^{24} Byte; eine entsprechende Vergleichstabelle findet sich im

Verarbeitung wichtig. Angestrebt wird hierbei die Echtzeitverarbeitung. Als letztes entscheidendes Merkmal von Big Data wird der Wert der Daten genannt. Schließlich sind die großen Mengen an unstrukturierten Daten, die möglichst in Echtzeit gesammelt und generiert werden lediglich nützlich, wenn aus Ihnen entscheidende Schlüsse gezogen werden können.
Aus einer großen Menge Daten Informationen zu filtern birgt großes Potenzial für Unternehmen. Fragen zu beantworten, die vor der Generierung der Daten nicht gestellt werden können und aus den vorliegenden Daten die entsprechenden nützlichen Schlüsse zu ziehen bedarf jedoch der richtigen Analysemethoden und technischen Voraussetzungen.

3 Big Data für Unternehmen

Nachdem der Begriff Big Data erläutert und abgegrenzt wurde, soll im Folgenden auf die wirtschaftliche Entwicklung und Bedeutung von diesen enormen Datenmengen für Unternehmen eingegangen werden. Es wird kurz erläutert wieso der Umgang mit Big Data für Unternehmen notwendig ist und diese überlebenswichtige Entscheidungen lediglich auf Basis dieser Daten treffen können bzw. sollten. Es soll dabei kurz auf die möglichen Potenziale eingegangen werden, in welchen Bereichen eines Unternehmens Big Data eine entscheidende Rolle spielen kann und um welche Daten es sich dabei handelt.

„Alles was digitalisiert werden kann, wird digitalisiert!“ (Bloching, Luck & Ramge, 2015. S. 38)

Der enorme Technologiefortschritt führt zu immer mehr Daten, die schnell erzeugt und gespeichert werden. Der digitale Wandel wird in vielen Bereichen der Wirtschaft deutlich, was es für Unternehmen notwendig macht sich dieser Entwicklung anzupassen. Tun Unternehmen dies nicht, ist die Wahrscheinlichkeit von Markt gedrängt zu werden lediglich eine Frage der Zeit. Der enorme Zuwachs von Online-Handel, App-Nutzung sowie andere mobile Anwendungen und Nutzung

des World Wide Web lässt eine enorme Datenmenge entstehen, aus der Generierung, Speicherung und Verwertung sich Unternehmen nicht nur die Verbesserung Ihrer vorhandene Produkte erhoffen, sondern vor dem Hintergrund der digitalen Disruption, die Schaffung neuer Geschäftsfelder (BITKOM, 2014). Nicht umsonst werden Daten bereits als das Öl des 21 Jahrhunderts bezeichnet.
Pein und Schoeneberg (2014) geben einen praktikablen Einblick in die mögliche Nutzung und Vorteilsgenerierung aus Daten. „Aus Daten lassen sich Informationen, aus Informationen lässt sich Wissen generieren“ (Pein und Schoeneberg, 2014. S. 310).

Bisherige Analysen bedarf es zunächst einer entsprechenden Hypothese. Man muss somit wissen welche Frage man stellen möchte um eine entsprechende Antwort zu erhalten. Die Datenmenge im Big Data Bereich und die notwendigen Algorithmen und deren Auswertungen ermöglichen jedoch eine Mustererkennung und somit eine Beantwortung von Fragen die bisher nicht gestellt wurden. Es können sich für ein Unternehmen folglich Chancen ermöglichen, die ohne Big Data nicht erkannt worden wären. Dabei stehen die Unternehmen jedoch vor einer Menge Herausforderungen um mit der enormen Datenflut und deren Komplexität umzugehen. Dabei müssen gewisse Voraussetzungen bezüglich der notwendigen Technologien und Methoden zur Analyse dieser Daten erfüllt werden, auf die im Folgenden etwa genauer eingegangen wird.

4 Technologien

Bisher wurden Daten in Datenbanken gespeichert. In diesen konnten Sie zudem geändert, gelöscht und jederzeit angerufen werden. Diese Systeme zur elektronischen Datenverwaltung entsprechen jedoch den großen Anforderungen der enormen Datenmengen nicht mehr. Die Verarbeitungskapazitäten, die notwendig sind, um mit Datenmengen umzugehen und zu arbeiten, welche bis in den Yottabyte-Bereich reichen, stehen jedoch eher wenigen Unternehmen zur Verfügung. Chris Anderson beschreibt das nun angetretene Zeitalter als Petabyte-

Zeitalter, in dem es üblich ist mit Datenmengen in dieser Größenordnung umgehen zu müssen. „Kilobytes speicherte man auf Disketten, Megabytes auf Festplatten, Terabytes auf Disk-Arrays. Petabytes speichert man in der Cloud“ (Anderson, 2008). Das Cloud-Computing macht den Umgang und die Verarbeitung dieser Datenmengen erst möglich.

4.1 Cloud Computing

„Cloud-Computing [...] stellt eine Ansammlung von Diensten, Anwendungen und Ressourcen dar, die dem Nutzer flexibel und skalierbar über das Internet angeboten werden, ohne eine langfristige Kapitalbindung und IT-spezifisches Know-How voraussetzen“ (Pannicke ,Repschläger, Zarnekow, 2010).

Der notwendige Speicherplatz, die notwendige Rechenleistung für die Datenverarbeitung sowie die zur Verarbeitung dieser Daten notwendigen Software-Programme werden in die „Cloud“ ausgelagert. Mit dieser Lösung kann jedes digitale Gerät auf fast unbeschränkte Rechen- und Speicherleistung zugreifen (Klausnitzer, 2013). Es werden generell drei wesentlichen Funktionen bzw. Serviceebenen der Cloud genannt, die in Anlehnung an Münzl, Pauly und Reti (2015) kurz erläutert werden.

- Infrastructure as a Service (IaaS)

Diese Funktion bietet dem Nutzer Zugriff auf skalierbare Rechen-, Speicher- und Netzkapazitäten. Notwendig dabei ist selbstverständlich ein hoher Automatisierungs- und Standardisierungsgrad. Die physische IT-Infrastruktur liegt dabei außerhalb der Verantwortung des Nutzers und wird dabei lediglich als einen Service bzw. wie eine Dienstleistung in Anspruch genommen.

- Platform as a Service (PaaS)

Zur Verfügung stehen auf dieser Ebene optimierte Middleware, wie Datenbank-Services, Services für die Integration, Zugriffskontrolle, Sicherheit, Synchronisation und Datenhaltung. Es entstehen so Cloud-basierte Plattformen für den gesamten Prozess der Erstellung und Bereitstellung webbasierter Anwendungen (Klausnitzer, 2013).

- Software as a Service (SaaS)

Auf dieser Ebene werden dem Nutzer Anwendungsservices zur Verfügung gestellt. Die Software läuft dann auf der technischen Infrastructure eines externen Anbieters und kann stetig und mobil abgerufen und genutzt werden.

Es wird des Weiteren zwischen Private Cloud Computing, sowie Public Cloud Computing unterschieden. Bei dem Private Cloud Computing stehen die jeweiligen Services lediglich dem einen Nutzer (Unternehmen) zur Verfügung, und wird nicht selten von diesem selbstständig betrieben, wobei bei dem Public Cloud Computing die Ressourcen einer Vielzahl an Nutzern zur Verfügung stehen und Eigentum des Dienstleisters sind. Eine Kombination dieser beiden Formen ist die Hybrid Cloud. Es handelt sich dabei um eine Kombination und organisatorischen Verknüpfung von Clouds mit einer traditionellen IT-Umgebung (Münzl, Pauly und Reti, 2015).

Frameworks wie MapReduce und Hadoop machen es des Weiteren möglich große Datenmengen zu Clustern und diese separat (be)rechnen zu lassen um diese nach dem Rechenvorgang wieder zusammenzufügen. Es wird dadurch die parallele Analyse großer semistrukturierter Daten möglich.

4.2 Apache Hadoop

„Hadoop ist ein Framework der Apache-Foundation für das verteilte Ausführen von Berechnungslogik auf sehr große Datenmengen“ (Neumann, 2015). Bei Hadoop handelt es hauptsächlich um zwei zusammenhängende Kernkomponenten, welche die Arbeitsweise des sogenannten Ökosystems[2] ermöglichen. Dazu gehören das verteilte Dateisystem zur Speicherung und Verwaltung der Daten HDFS (Hadoop Distributed File System), sowie das Hadoop MapReduce, das hauptsächlich zur verteilten und parallelen Verarbeitung der Daten dient (Kiese, 2015). HDFS ist ein

[2] Das Apache Hadoop Ökosystem besteht aus weiteren wichtigen Komponenten, auf die aufgrund der Kürze der Ausarbeitung nicht weiter eingegangen wird. Eine bildliche Übersicht über die wichtigsten Komponenten des Apache Hadoop Ökosystems findet man in Anhang 2.

Dateisystem, dass hauptsächlich zur skalierbaren und zuverlässigen Speicherung von sehr großen Datenmengen dient. Daten werden auf unterschiedlichen Servern bzw. Knoten gespeichert. Dieses verwaltet eingehende Datenanfragen und speichert des Weiteren hilfreiche Metadaten.

Bei MapReduce handelt es sich um Programmiermodell zur Verarbeitung dieser großen Datenmengen. Jede Anfrage besteht dabei aus zwei Vorgängen, dem Map und dem Reduce. Zunächst werden dabei alle möglichen Ergebnisdaten gesammelt und in Zwischenspeichern angelegt. Im Anschluss daran wird der Reduce Vorgang angestoßen, bei dem die Zwischenspeicher parallel ausgelesen werden und anhand der angefragten Kriterien die entsprechenden Ergebnisdaten ausgeben. Die bereits erwähnte Weiterentwicklung von MapReduce Yarn, teilt den beschriebenen Vorgang in mehrere seperate Prozesse und optimiert diese Vorgänge somit weiterin (Kiese, 2015). Apache Hadoop bietet bereits eine gute Möglichkeit mit den Herausforderungen der enorm großen Datenmengen umzugehen und hat sich aus diesem Grund bereits als Kern der modernen Datenarchitektur etabliert (BITKOM, 2014).

4.3 NoSQL Datenbanken

NoSQL[3] Datenbanken sind speziell für sehr große Datenmengen designed. Verfolgt wird die Möglichkeit nicht-relationale Konstrukte abzubilden. Herkömmliche Datenbanksysteme sollen dahingehend erweitert werden und nicht vollständig ersetzt werden (Manhart, 2013). NoSQL Datenbanken werden jedoch Vorteile zugesprochen, die explizit bei der Arbeit mit sehr großen Datensätzen von großem Vorteil sein können. Einer der wichtigsten Vorteile ist die mögliche horizontale Erweiterung der Datenbanken. Bei bisher üblichen relationalen Datenbanken war lediglich die vertikale Skalierung möglich. Durch diese Möglichkeit muss ein vorhandener Server nicht mit weiterem Speicher aufgerüstet werden, sondern es können weitere Server in das Datenbanksystem integriert werden. Die Daten werden dann auf die Systeme verteilt, was nicht nur eine kostengünstigere

[3] NoSQL steht für „Not Only SQL“, (Manhart, 2013)

Methode der Speichererweiterung ist, sondern ebenfalls eine flexiblere Methode der Skalierung (Dietl, 2011).

4.4 Enterprise Data Warehouse

Als Data Warehouse wird eine Datenbank bezeichnet, welche die Speicherung von Daten aus sehr heterogenen Quellen ermöglicht. Diese werden in dieser Datenbank zu einem einheitlichen Format zusammengefasst, was wiederrum den Zugriff und das Abrufen der Daten erleichtert. „Ein Data Warehouse ist ein „Datenlager“, das nach einem bestimmten Konzept strukturiert ist, um flexible und schnelle Auswertungen zu ermöglichen“ (Riggert, 2015). Data Warehouse bildet zunächst eine geeignete Basis zur Aggregation von heterogenen Daten und betrieblichen Kennzahlen und ermöglicht damit Analysen und bildet häufig die Grundlage des Data Mining. Üblich ist das Betreiben von Data Warehouse auf relationalen Datenbanken. Wie bereits beschrieben, ist genau dies oft eine Schwierigkeit bei der Arbeit mit sehr großen Datenmengen. Zudem treten weitere Schwierigkeiten bei der Verarbeitung von unstrukturierten Daten, sowie bei zunehmenden Antwortzeiten bei sehr großen Datenmengen auf.

Data Warehouse beschreibt jedoch weitestgehend eine Datenarchitektur und kann somit mit weiteren bereits beschriebenen Technologien kombiniert werden um dem Anspruch von sehr großen Datenmengen zu entsprechen. Das Einführen von In-Memory Datenbanken in ein Data Warehouse System hat die Einsatzmöglichkeiten dessen, explizit in Bezug auf die Abfrageperformance, deutlich verbessert (Welker, 2015). Ralph Kimball betont zudem die Flexibilität, Performance und Kostenersparnis eines zukünftigen Hadoop Data Warehouses und sieht in dieser Kombination großes Potenzial (Kimball, 2014).

4.5 In Memory Systeme

In Memory Datenbanken haben sich ebenfalls aufgrund der notwendigen Verarbeitung enorm großer Datenbanken etabliert. Besonders relevant sind diese Systeme bei der Analyse von großen Datenmengen, da diese auf die höhere

Geschwindigkeit, beim Speichern auf und Abrufen von Daten aus dem Arbeitsspeicher zurückgreifen. Bei In Memory Datenbanken wird das gesamte Dateivolumen inklusive die notwendigen Datenbankanwendungen in den Hauptspeicher geladen. Dadurch kann dann die Analyse schneller erfolgen, da auf das lange Laden der Daten von der Festplatte verzichtet werden kann (Manhart, 2013). Bei in Memory Datenbanken kann es sich zudem sowohl um SQL und NoSQL Datenbaken handeln.

Es ließe sich nun auf weitere technische Innovationen und Verbesserungen bezüglich der Verarbeitung von Big Data eingehen. Aufgrund der Kürze der Ausarbeitung wird jedoch lediglich auf die meist diskutiertesten Themen eingegangen. Big Data basiert nicht auf einer technischen Lösung, sondern ist auf das Zusammenwirken einer Vielzahl von Technologien angewiesen. „Insgesamt erlauben diese Fortschritte, aus immer mehr Daten einen immer höheren betriebswirtschaftlichen Nutzen zu ziehen.“ Es kommen dabei unterschiedlichste Technologien, die auf das jeweilige Anwendungsszenario spezialisiert ist, zum Einsatz (BITKOM, 2014).

5 Analyseverfahren

Nicht nur die Daten selbst und die entsprechend angewandten Technologien sind entscheidend ob aus Big Data auch Smart Data gewonnen werden können, sondern ebenfalls die richtige Wahl des jeweiligen Analysemodells. Klausnitzer (2013) beschreibt drei wesentliche Analysemodelle, die heute angewendet werden.

Descriptive Analytics - ist die beschreibende Analyse von Daten die ein Unternehmen in der Vergangenheit bezüglich des Auftretens am Markt gewonnen hat. Sie beschreibt hauptsächlich die Aufgaben der BI (Business Intelligence) in Unternehmen, die auf die, in den wettbewerbs- und Marktdaten enthaltenen Informationen spezialisiert sind und diese für Handlungsempfehlungen bezüglich der zukünftigen Performance des Unternehmens nutzen. Im Mittelpunkt der

Auswertung der Daten liegt folglich die Optimierung der Unternehmensperformance mit Grundlage der vergangenen Performance am Mart.

Predictive Analytics – beschreibt die vorhersagende Analyse von Daten. Diese Form der Analyse beinhaltet eine Vielzahl von unterschiedlichsten statistischen Modellen, das Maschinenlernen, das „Data Mining“ um aus einer großen Anzahl von Daten Wahrscheinlichkeiten zukünftiger Ereignisse und Entwicklungen zu berechnen. Zumeist angewendet wird diese Form der Analyse im CRM um nicht nur höhere Gewinne zu erlangen sondern insbesondere um die Kundenbeziehung zu verstärken und dem Kunden gezielte jene Produkte anzubieten die er kaufen wird. Die Analysen zielen direkt auf mögliche Verhaltensmuster der Kunden ab um Angebote und den Service des Unternehmens stetig kundenindividueller zu gestalten.

Prescriptive Analytics – beschreibt die empfehlende Analyse. Genutzte Analyseverfahren sollen bei diesem Analysemodell vor Allem gezielte Handlungsempfehlungen für das Unternehmen erarbeiten. Es wird dabei gezielt versucht nicht nur vorherzuschauen, wie eine Entwicklung, sondern vor allem warum diese Entwicklung fortschreitet.

Bei allen Analysemodellen ist die kontinuierliche Sammlung der Daten sehr bedeutend, da diese auch parallel zur Analyse selbststattfindet und gleichzeitig alle neu generierten Daten mit in die Analyse mit einbezieht. Je mehr Daten gesammelt werden, desto genauer wird somit das Ergebnis der Analyse.

5.1 Data Mining

Der Begriff „Data Mining“ ist laut BITKOM (2014) ein Oberbegriff für eine Vielzahl von verschiedenen Methoden, Verfahren und Techniken, die dazu genutzt werden aus einer Menge an Daten verwertbares Wissen zu fördern und zu verwerten. Der ursprünglich aus dem Bereich der Statistik stammende Begriff, der die selektive Methodenanwendung zur Bestätigung vorformulierter Hypothesen verwendet wurde, wird heute mit dem Begriff der Datenmustererkennung gleichgesetzt

(Bensberg & Grob, 1999). Bensberg & Grob (1999) beschreiben aufgrund vielfältiger Abgrenzungsschwierigkeiten gängiger Definitionen des Data Mining Begriffs, diesen als „integrierten Prozess [...], der durch die Anwendung von Methoden auf einen Datenbestand Muster identifiziert." Des Weiteren beschreiben Sie den Data Mining Prozess in fünf Phasen. Diese sind die Extraktion der relevanten Daten aus den jeweiligen Datenquellen, anschließend die Selektion der Datensätze und Attribute (vertikale und horizontale Selektion), worauf die Phase der Vorverarbeitung folgt. In dieser Phase wird die Datenqualität der selektierten Datensätze untersucht um Fehler zu vermeiden. Die vierte Phase ist daraufhin die Transformation der Daten. In dieser werden die relevanten Daten in ein Datenbankschema transferiert, dass von dem vorliegenden Data Mining System verarbeitet werden kann. In der fünften Phase erfolgt die Methodenauswahl und Anwendung zur Identifikation von Mustern und Relationen in dem untersuchten Datenbestand. Angewandte Methoden sollen daraufhin Muster und Relationen erkennen und erarbeiten, die es ermöglichen Aussagen über die untersuchten Daten und Objekte zu treffen.

Es existiert eine Vielzahl von Analysemethoden innerhalb des Data Mining. Im Folgenden sollen lediglich kurz auf die am weitesten verbreiteten Methoden eingegangen werden.

5.1.1 Clusteranalyse

Sinn einer Clusteranalyse ist es eine große Anzahl von heterogenen und unstrukturierten Daten in homogene Gruppe, sogenannte Cluster zu sortieren. Dabei werden die Daten anhand von Variablen und Typologien nach Ähnlichkeiten sortiert. Es muss daher zunächst bestimmt werden nach welchen Merkmalen dies vorgenommen werden soll. Auf dieser Grundlage muss während des Analysevorgangs zunächst die Ausprägung des Merkmals jeder Datei überprüft und bewertet werden. Es muss zudem ein „Fusionierungsalgorithmus" ausgewählt werden, der die selektierten Daten zu Clustern zusammenfügt und gleichzeitig die Anzahl der Cluster angibt und auswählt. Es folgt eine Interpretation der Analyse und

Überprüfung der Güte dieser (Schäfer, 2009). Clusteranalysen können eine sehr hilfreiche Informationsgrundlage für Kundensegmente oder kundenspezifische Marktbearbeitungen sein (Bensberg & Grob, 1999).

5.1.2 Klassifikationsanalyse

Bei der Klassifikationsanalyse werden die Daten nach einem zuvor bekannten Merkmal in Klassen aufgeteilt und zusammengefasst. Es soll daraufhin ein Modell entwickelt werden, dass die Klassenzugehörigkeit neu gespeicherte Daten vorhersehen kann. Als Beispiel der Anwendungsmöglichkeit ist die Vorhersage der Kreditwürdigkeit von Bankkunden zu nennen. Es soll dann durch bereits gesammelte Daten vorhergesehen werden, mit welcher Wahrscheinlichkeit ein Kunde kreditwürdig ist oder nicht.
Zu möglichen Methoden der Klassifikationsanalyse gehören unter anderem Entscheidungsbaumverfahren, Neuronale Netze oder Navias Bayes, auf die aufgrund der begrenzten Ausführungsmöglichkeiten hier nicht genauer eingegangen wird.

5.1.3 Assoziationsanalyse

Die Assoziationsanalyse dient der Suche nach Abhängigkeiten zwischen den Daten. Identifizierte Muster können daraufhin in Wenn-Dann-Regeln übersetzt werden und genaue Handlungsmuster und Folge-Zusammenhänge verdeutlichen. Oft in der Praxis angewendet wird diese Analysemethode in Form von Warenkorbanalysen. Es können Produkte identifiziert werden, die mit einer bestimmten Wahrscheinlichkeit mit anderen Produkten zusammen gekauft werden. Die Ergebnisse dieser Analyse können in direkte Verkaufsstrategien einfließen und dementsprechend Handlungsempfehlungen beeinflussen.

5.1.4 Maschinelles Lernen

Dem maschinellen Lernen kommt gerade in Bezug auf Big Data eine hohe Bedeutung zu da dies das selbstständige Erwerben von Wissen durch Computerprogramme beschreibt. Durch die Automatisierung steht die gesuchte

Frage nach den Mustern welche innerhalb der Datensätze zu finden sind im Vordergrund der Analyse. Diese Eigenschaft ermöglicht es somit gegebenenfalls Muster zu identifizieren die, nicht wie in üblichen Analysemethoden gezielt gesucht werden. „Machine Learning“ beschäftigt sich mit Verfahren, um günstige Lösungsansätze für Probleme, die manuell nicht oder nur unter hohem Kostenaufwand lösbar sind, automatisch zu erlernen und in der Anwendung weiterzuentwickeln (BITKOM, 2014).

Der Umgang mit Big Data lässt, allein aufgrund der Vielfalt der zu Nutzenden Daten, sowie der Vielfalt der zu beantworteten Fragen, eine Menge an Analyseverfahren zu. Es muss abhängig von der verfolgten Strategie die geeigneten Methoden individuell projektbezogen ausgewählt werden. Gerade diese Tatsache macht es für viele Unternehmen schwierig die entsprechenden notwendigen IT-Architekturen bereitzustellen.

6 Ausblick

Es wird deutlich, dass Big Data überwiegend eine Schnittstellenkompetenz aus unterschiedlichen Bereichen wie die Informationstechnologie, Mathematik, Künstlicher Intelligenz, Design und Wirtschaftswissenschaften ist. Big Data hat stets eine ganzheitliche Konzeption zum Ziel, die es ermöglichen soll aus einer enormen Menge von Daten Schlüsse zu ziehen um nachhaltige Werte zu generieren (BITKOM, 2015).

Die notwendigen Technologien, die dazu dienen die aufkommenden Datenmengen entsprechend auszuwerten und zu analysieren, bergen zunächst einige Schwierigkeiten für Unternehmen. Diese müssen die informationstechnologischen Architekturen schaffen um so die entsprechende Analyse dieser Daten zu ermöglichen. Die Ausführungen bezüglich der Big Data Technologien zeigen, dass es bisher keine einheitlichen Technologien und Architekturen gibt, sondern dass diese je nach Anspruch an die Analysemöglichkeiten und Zielsetzungen stark variieren. Dabei muss die Generierung und Sammlung dieser Daten ebenso berücksichtigt

werden, wie die Speicherung dieser Daten, die Analyse und die Präsentationsmöglichkeiten der durchgeführten Analyse. Die beschriebenen Technologien ermöglichen jedoch im Gegensatz zu bisherigen Analysetools und herkömmlichen Datenbanken den Umgang mit einer enormen Menge von Daten und machen Sie somit für Unternehmen unumgänglich.

Um auf einem immer stärker global ausgerichtetem Markt weiterhin effizient und effektiv zu agieren und Wettbewerbsvorteile zu generieren, müssen Unternehmen den Markt genauestens kennen. Dies macht den Umgang und die effiziente Verwertung der Daten notwendig. Daten werden zu einer existenziellen Ressource für Unternehmen.

Durch die zunehmenden Verfügbarkeit von Daten und die durch die Technologischen Möglichkeiten diese auch zu Nutzen, können Unternehmen weltweit neue Wege finden erhebliche Differenzierungsmöglichkeiten zu schaffen. Die Entscheidungsfindung sowie die Leistungsfähigkeit von Unternehmen können mit Hilfe von Big Data enorm verbessert werden. Ohne Big Data ist es somit kaum noch möglich sich gegen Unternehmen, die diese Möglichkeiten nutzen durchzusetzen. Big Data ist Information und Information ist Wissen.

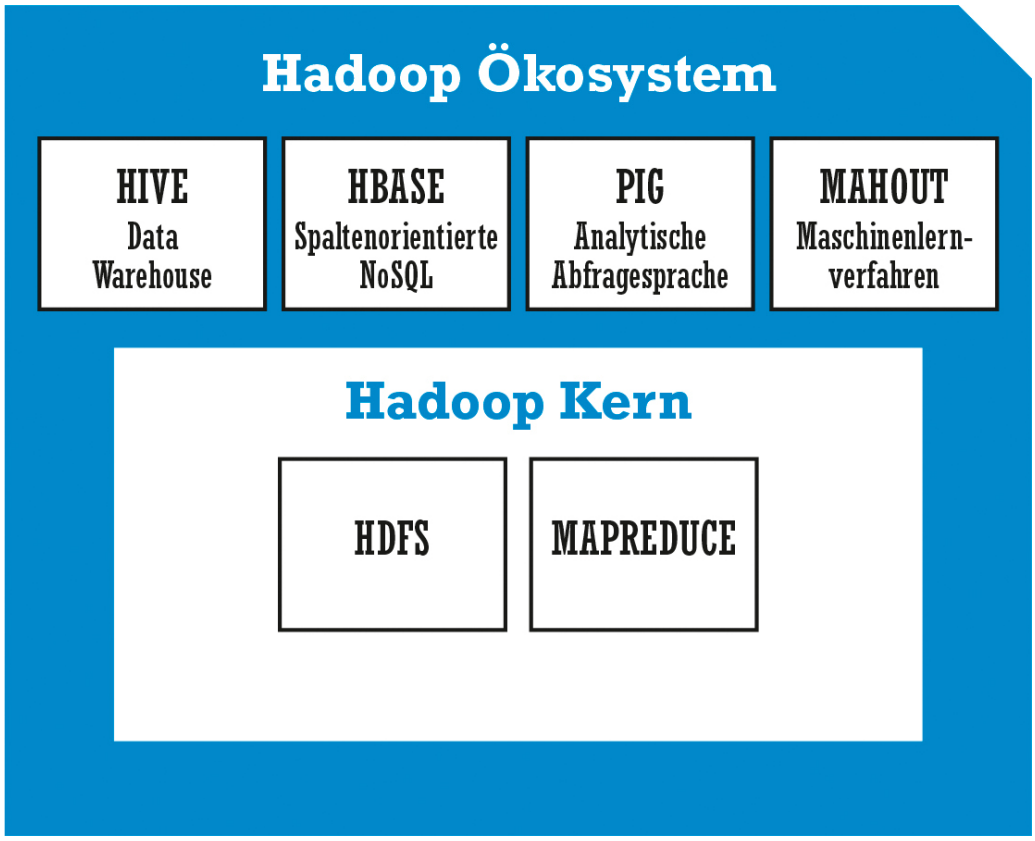
Anhang

Anhang 1

Dezimalpräfixe	
Name (Symbol)	**Bedeutung**[G 1]
Kilobyte (kB) [G 2]	10^3 Byte = 1 000 Byte
Megabyte (MB)	10^6 Byte = 1 000 000 Byte
Gigabyte (GB)	10^9 Byte = 1 000 000 000 Byte
Terabyte (TB)	10^{12} Byte = 1 000 000 000 000 Byte
Petabyte (PB)	10^{15} Byte = 1 000 000 000 000 000 Byte
Exabyte (EB)	10^{18} Byte = 1 000 000 000 000 000 000 Byte
Zettabyte (ZB)	10^{21} Byte = 1 000 000 000 000 000 000 000 Byte
Yottabyte (YB)	10^{24} Byte = 1 000 000 000 000 000 000 000 000 Byte

Quelle: https://de.wikipedia.org/wiki/Byte

Anhang 2



Apache Hadoop Ökosystem
Quelle: http://blog.isreport.de/wp-content/uploads/2013/11/white-duck-hadoop-grafiken-2.jpg

Literaturverzeichnis

Anderson, C. (2008), The End of Theory: The Datan Deluge Makes the Scientific Method Osolete. Wired Magazine 16.07, 2008

Bagnoli, V., Martel, E., Wagner, B., (2012), Big Data – Ausschöpfung von Businessdaten; in Trends in der IT; Mehler-Bicher, A., Steiger, L. (Hrsg.), Fachhochschule Mainz: Mainz.

BITKOM, (2014) – Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V, Big-Data-Technologien Wissen für Entscheider, Leitfaden

BITKOM, (2015), Kognitive Maschinen – Meilenstein in der Wissensarbeit. Leitfaden BITKOM, Bundesverband Informationswirtschaft, Telekommunikation und neue Median e.V.

Bloching B., Luck L. & Ramge T., (2015), Smart Data, Datenstrategien, die Kunden wirklich wollen und Unetrnehmen wirklich nützen, Redline Verlag: München.

Brensberg, F. & Grob, H. L. (1999), Arbeitsblatt Nr. 8, Das Data-Mining-Konzept, Institut für Wirtschaftsinformatik der westfälischen Wilhelms-Universität Münster

Dietl, M. (2011), NoSQL: Eine kurze Einführung in Theorie und Praxis, online im Internet unter < https://blog.seibert-media.net/blog/2011/08/31/nosql-datenbanken-theorie-und-praxis/> Stand 25.08.2015.

Kiese, T. (2015), Spaltenorientierte Datenbanken und HBase, in: : Eine praxisorientierte Bewertung von Architekturen und Techniken für Big Data, Berliner Schriften zu modernen Integrationsarchitekturen Band 13, Hrsg.: Prof. Dr.-Ing. habil. Andreas Schmietendorf, Hochschule für Wirtschaft und Recht Berlin; Ahaker Verlag: Aachen.

Kimball, R. (2014). Building a Hadoop Data Warehouse., Hadoop 101 For Enterprise Data Warehouse Professionals; Kimball Gruop, online im Internet <https://www.brighttalk.com/webcast/9059/109141 > Stand 30.08.2015.

Klausnitzer, R. (2013), Das Ende de Zufalls – Wie Big Data uns und unser Leben vorhersagbar macht, Salzburg: Econwin Verlag

Mahnhart, K. (2013), NoSQL und In Memory – die neuen Datenbanken für Big Data, online im Internet unter <http://blog.qsc.de/2013/10/nosql-und-in-memory-die-neuen-datenbanken-fur-big-data/> Stand 25.08.2015.

Münzl, G., Pauly, M., Reti, M. (2015), Cloud Computing als neue Herausforderung für Management und IT, Springer: Berlin Heidelberg

Neumann, R., (2015), Eine Einführung in Apache Hadoop, in: Eine praxisorientierte Bewertung von Architekturen und Techniken für Big Data, Berliner Schriften zu modernen Integrationsarchitekturen Band 13, Hrsg.: Prof. Dr.-Ing. habil. Andreas Schmietendorf, Hochschule für Wirtschaft und Recht Berlin; Ahaker Verlag: Aachen.

Oracle, (2015), An Enterprise Architect's Guide to Big Data – Reference Architecture Overview – Oracle Enterprise Architecture White Paper

Pannicke, D., Repschläger, J., Zarnekow, R. (2010), Cloud Computing: Definitionen, Geschäftsmodelle und Entwicklungspotenziale, HMD Praxis der Wirtschaftsinformatik (Journal), Volume 47, Issue 5, s. 6-15

Schäfer, T., (2009). Clusteranalyse, Vorlesung Methodenlehre 2, Sommersemester 2009, Technische Universität Chemnitz, Institut für Psychologie, Chemnitz.

Welker, P. (2015), Big Data oder Data Warehouse, online im Internet <http://www.computerwoche.de/a/big-data-oder-data-warehouse,3092517 > Stand 30.08.2015.

Versicherung über Selbstständigkeit

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, den _____ _____