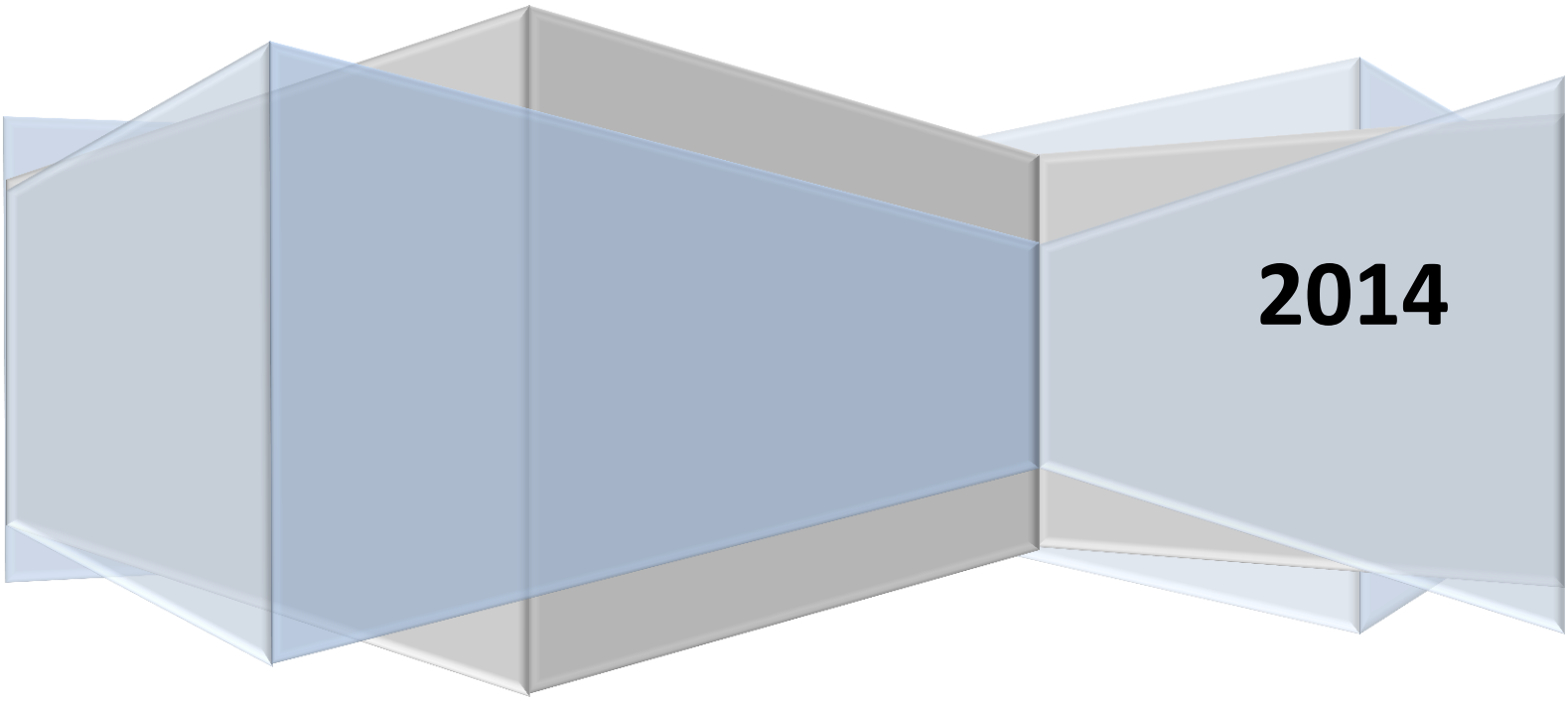
Hochschule München

Big Data – Hype oder Chance

Autor: Fuchs Dominik
Dozent: Michael Theis



2014

Inhaltsverzeichnis

Abbildung 1: Big Data[1]

1. Aufgabenstellung

Ziel dieser Arbeit ist es dem Leser das Thema Big Data näher zu bringen. Es soll ein Verständnis für das Ausmaß des Begriffes vermittelt werden. Dazu werden dem Leser einige Unternehmen als Beispiel aufgeführt. Im Weiteren werden dem Leser Mittel aufgezeigt mit dem sich Big Data bewältigen lässt.

Die Frage, ob sich durch Big Data Chancen eröffnen oder ob es sich lediglich um einen Hype handelt, soll untersucht werden. Es werden die Schmerzen aufgezeigt die jedem Unternehmen im Umgang mit Big Data über kurz oder lang erwarten.

[1] http://blog.joelrubinson.net/wp-content/uploads/2014/04/big-data-cloud.jpg (aufgerufen am 17.4.2014)

Da das Thema sehr umfangreich ist, werden einzelne Bereiche nur oberflächlich angeschnitten. Um ein genaueres Bild zu dem durchaus spannenden Thema zu bekommen, empfehle ich Fachliteratur. Ein paar gute Quellen befinden sich im beigefügten Literaturverzeichnis

2. Der Begriff Big Data

2.1. Begriffsklärung

Zu Beginn einer wissenschaftlichen Arbeit ist es üblich den zentralen Begriff des Themas zu erläutern, beziehungsweise zu definieren. Im Fall des Begriffes Big Data fällt das ganze etwas langwieriger aus, da sich die Bedeutung des Begriffes in vielen Quellen unterscheidet.

Fangen wir mit der Herleitung des Wortes an. Big steht im Deutschen für groß und Data für Daten. Also ist es sehr naheliegend, dass es sich um große Datenmengen handelt. In der Regel Datenmengen die zu groß sind um sie händisch oder mit klassischen Methoden zu bewältigen. Zusätzlich wird Big Data aber auch mit der umfangreichen Analyse und Überwachung von Nutzdaten in Verbindung gebracht. Vor allem seit dem Aufkommen der NSA-Affäre. Letztendlich unterliegt der Begriff einem kontinuierlichen Wandel.[2]

Laut Pavlo Baron beschäftigt sich Big Data mit dem Sammeln von Daten. Aus diesen sollen dann wertvolle bzw. nützliche Informationen gewonnen werden. Menge sowie Form der Daten sind hierbei beliebig. Zusätzlich gehört auch die Findung und Schaffung von neuen Datenquellen zum Begriff Big Data. Durch den Umgang mit diesen Daten kann der Unternehmenserfolg verbessert werden. Es geht bei Big Data um die Zusammenarbeit des menschlichen Gehirns und den Fähigkeiten die ein Computer zur Verfügung stellt. Hierbei stellen Computer die Informations- und Technologiegrundlage für Empfehlungs- bzw. Entscheidungsunterstützungssysteme. Einfach gesagt ein Computer filtert aus einer enormen Menge aus Daten(z.B. mit Hilfe eines Algorithmus) eine Entscheidungsgrundlage heraus. Hier kommt nun das menschliche Gehirn zum Einsatz. Aufgrund seiner kognitiven Fähigkeiten sowie seinem „Bauchgefühl“ trifft es die endgültige Entscheidung. Baron kritisiert in seinem Buch, dass der Begriff stark gehyped wurde und er empfiehlt nicht auf den Begriff an sich zu achten sondern auf die Inhalte und Herausforderungen die sich dahinter verbergen.[3]

2.2. Relevanz des Begriffs

Der Hintergrund des Begriffs begründet sich im starken Anstieg des weltweiten Datenvolumens. Verantwortlich dafür ist eine Menge von Quellen. Um nur ein paar zu nennen: Sensordaten, Maschinendaten, Log Daten und viele mehr. 2011 knackte das weltweite Datenvolumen die

[2] http://de.wikipedia.org/wiki/Big_Data (aufgerufen am 17.4.2014) (Wikipedia)
[3] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.1 (Baron, 2013)

Zettabyte Barriere, was für eine 1 mit 21 Nullen steht. Zusätzlich zu der schieren Menge an Daten spielt aber auch noch die mangelnde Struktur eine Rolle.[4]

Nicht umsonst spricht man bei Daten mittlerweile schon vom „4. Produktionsfaktor“. Daten werden heutzutage als wirtschaftlich sehr wertvoll erachtet. Der Wert von Erkenntnissen, der durch Auswertung der Daten gewonnen werden kann gilt als potenziell gewaltig. Nur als kurzes Beispiel um den wirtschaftlichen Nutzen greifbar zu machen, greife ich hier ein Beispiel aus Stockholm auf. Durch Integration von Wetter- und Verkehrsdaten(Unfall und Staumeldungen) konnte das Verkehrsaufkommen und die Emissionen um 20%, die Fahrzeiten gar um 50 % reduziert werden.[5] Im nächsten Kapitel gehe ich genauer auf Big Data Szenarien ein.

2.3. Die „3“ V’s

Um die Probleme mit Big Data ein bisschen einordnen zu können, ist es hilfreich die 3 V’s zu betrachten. Sie stellen die 3 wichtigsten Dimensionen dar, mit denen Unternehmen im Bezug auf Big Data konfrontiert werden. Die drei in der Überschrift des Kapitels ist bewusst in Anführungszeichen gesetzt, aber dazu später mehr.

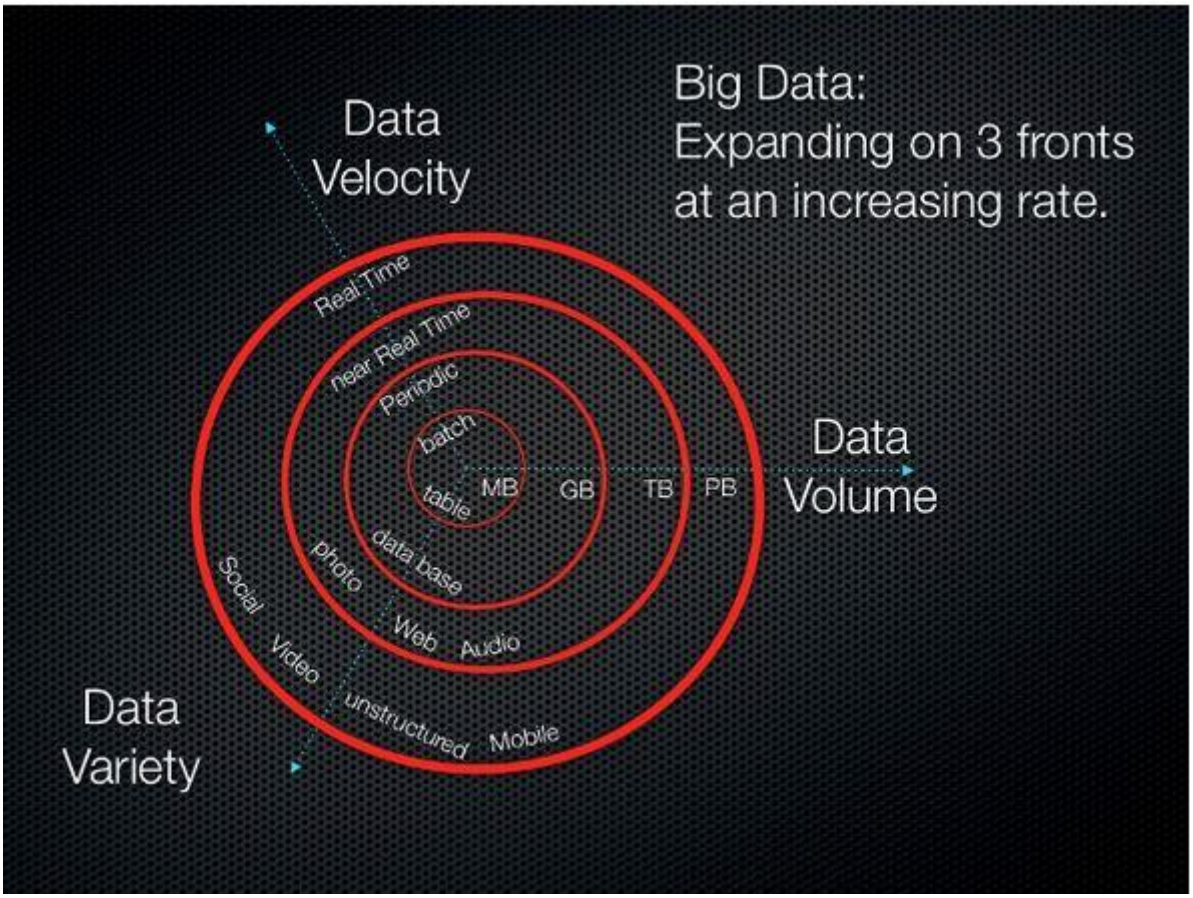


Abbildung 2 : Die 3 V's[6]

[4] http://www.softselect.de/wissenspool/big-data (aufgerufen am 17.4.2014) (Gottwald)
[5] Vgl. Aktueller Begriff Big Data – Sabine Horvath S.1 (Horvath, 2013)
[6] http://whatis.techtarget.com/definition/3Vs (aufgerufen am 18.4.2014)

Velocity (Geschwindigkeit)

Fangen wir mit der Geschwindigkeit an. Diese ist relativ konstant beziehungsweise nach oben hin unbeschränkt. Im Klartext: Es gibt keine Grenze wie schnell man seine Daten bekommen, auswerten und verwenden will. Umso schneller man in der Lage ist seine Entscheidungen zu treffen umso höher ist der Erfolg am Markt.

Es gibt ganze Branchenzweige in denen es auf Milli- teilweise sogar Nanosekunden ankommt. Hier ist es essenziell die Strecken zwischen Maschinen zu verkürzen, Netze zu optimieren und das Maximum aus den Maschinen herauszuholen.[7]

Früher fielen Daten in bestimmten Abständen an, heute ist man aufgrund von Vernetzung und elektronischer Kommunikation dem Datenfluss ununterbrochen ausgesetzt. Das führt dazu dass die Daten heute teilweise in „Echtzeit“ aufgenommen und analysiert werden müssen.[8]

In der vorrangehenden Abbildung sieht man eine immer schneller werdende Abstufung nach außen.

Volume (Volumen)

Als 2. Kriterium gehe ich auf das Volumen ein. Heutzutage muss ein Unternehmen darauf vorbereitet sein mit unlimitierten Datenmengen hantieren zu können. Je mehr Daten man besitzt umso qualifizierter sind die Entscheidungen aus den resultierenden Informationen. Die Analyse lebt von großen Datenmengen weil sie nur so immer genauer trainiert werden kann.[9]

Man geht davon aus, dass sich die Datenmenge alle 2 Jahre verdoppelt. Es wird geschätzt, dass bereits 2013 weltweilt 2 Trilliarden Bytes gespeichert wurden. Auf iPads gespeichert und gestapelt ergäbe das eine 21.000 km lange Mauer.[10]

Eine wahre Flut von Daten findet sich beispielsweise schon alleine im Gesundheitswesen. 20 Terrabyte Daten pro Kunde sind hier keine Seltenheit. Auch die sozialen Medien tragen ihren Teil bei. Spitzenreiter ist hierbei Facebook mit 900 Millionen Mitgliedern.[11]

In der Abbildung sieht man, dass es mittlerweile gängig ist über Petabytes zu sprechen. Wichtig hierbei ist nicht zu vergessen, dass der Speicherplatz für die Dateien immer billiger wird. Er ist im Vergleich zu anderen Ausgaben von größeren Unternehmen marginal.

[7] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.23
[8] Vgl. Aktueller Begriff Big Data – Sabine Horvath S.2
[9] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.24
[10] Vgl. Aktueller Begriff Big Data – Sabine Horvath S.2
[11] Vgl. Big Data – Dr. Martin Wolfgang S.4

Variety (Variabilität)

Last but not least, die Variabilität in den Daten. Anders bezeichnet auch als Chaos. Menschen neigen dazu chaotisch zu sein, also muss man zwangsweise mit entsprechenden Datensätzen leben. Variety lässt sich also nicht als Variable sondern eher als eine Konstante bezeichnen. Sie ist eine Tatsache in der modernen IT-Welt.[12]

Man geht davon, dass heute etwa 90% der Daten unstrukturiert sind.[13] Gerade diese mangelnde Struktur macht es schwer die Daten händisch oder klassisch auszuwerten. Die Einführung von Freitextfeldern oder das Analysieren von Dateien die menschliche Sprache beinhalten sind Beispiele für gängige Probleme.

Verbraucher

Jetzt kommen wir zu dem Grund für die in Anführungszeichen gesetzte drei in der Überschrift. Mittlerweile hat sich in manchen Kreisen ein viertes „V“ herausgebildet. Und zwar in Form der Verbraucher. Die Anzahl dieser ist auch sehr stark am steigen.[14] Der Feldzug des Internets schreitet immer weiter voran. Stand Juni 2012 gibt es weltweit ca. 2.408.520.000 Internet-Nutzer. Bemerkenswert ist, dass im Schnitt jeden Tag 218.000 neue Nutzer hinzukommen. Das macht einen Zuwachs von 2,5 Nutzern/Sekunde.[15]

3. Big Data Szenarien

Da ich nun dargestellt habe, dass die Bedeutung des Begriffes nicht ganz einfach und unklar strukturiert ist, erachte ich es als notwendig in Fallbeispielen ein paar Big Data Szenarien zu präsentieren. Ich habe mich hier wieder auf die Recherche von Pavlo Baron gestützt, der in seinem Buch zeigt, wie die erfolgreichsten der Firmen mit dem Thema Big Data umgegangen sind, ohne den Begriff zu hypen. Hier fällt auf, dass Big Data an sich nichts Neues ist. Einige Unternehmen beschäftigen sich schon lange mit großen Datenmengen, sowie der zugehörigen Analyse. Lediglich der Begriff ist in den letzten paar Jahren vermehrt aufgekommen.

---

[12] Vgl. Aktueller Begriff Big Data – Pavlo Baron S.25
[13] Vgl. Big Data – Dr. Martin Wolfgang S.4
[14] Vgl. Big Data – Dr. Martin Wolfgang S.5
[15] http://www.live-counter.com/internetnutzer-weltweit/ (aufgerufen am 29.04.2014)

3.1. Der Fall Google

Baron erwähnt in seinem Buch immer wieder den Begriff Big Data Schmerzen. Was er damit meint sind die im Umgang mit Big Data auf die Unternehmen zukommenden Probleme. Bei Google ist es nicht schwer sich vorzustellen was etwaige Schmerzen sein könnten. Man muss ja nur mal bei Google ins Suchfeld einen x-beliebigen Begriff eingeben. Beispielsweise eine Suche nach dem Begriff „München“ ergibt sage und schreibe 84.400.000 Ergebnisse. In 0,33 Sekunden wohlgemerkt. Aber das ist ja mittlerweile lange nicht mehr alles was Google im Angebot hat. Google ist zur Riesenplattform geworden auf der Nutzer alles Mögliche an Daten ablegen können. Ich persönlich nutze beispielsweise Google Drive für das Studium, es ermöglicht es unserem Team in Softwareengineering, Dateien wie Berichte oder Listings auszutauschen. Man sieht also, dass Google in kurzer Zeit in der Lage sein muss Unmengen von Daten zu verarbeiten. Kein Nutzer möchte lange warten, wenn er eine Suchanfrage startet. Es liegt also nahe, die hier entstehenden Big Data Schmerzen als groß anzusehen.

Google hat einiges richtig gemacht auf seinem Weg. So hat das Unternehmen früh bemerkt, dass es darauf ankommt die Nutzer für sich arbeiten zu lassen. Google nutzt das Feedback der Nutzer um eine möglichst hohe Qualitätsgarantie für die Treffer zu erzielen. Dies geschieht aber nicht durch umständliche, ellenlange Web-Formulare. Ein paar Klicks lösen eine Kaskade an maschinell analytischen Prozessen aus, um dies zu erreichen steht die Analyse bei Google ganz oben. [16]

3.2. Der Fall Amazon

Als 2. Fallbeispiel nehme ich Amazon. Warum? Weil Amazon eines der erfolgreichsten, dynamischsten Unternehmen unserer Zeit ist. Sie haben dazu beigetragen die IT-Welt zu revolutionieren. (Cloud, Dynamo) Amazon hat diese Dinge zwar nicht inhaltlich erfunden, aber sie haben es geschafft die Wissenschaft erfolgreich und gewinnbringend zu nutzen, zu branden und massentauglich zu machen. Amazon ist in der Lage ihre Plattformen mehrfach am Tag zu aktualisieren. Somit ist es möglich die Kunden mit den neuen und verbesserten Funktionen sofort zu erreichen.

Was sind die Big Data Schmerzen bei Amazon? Ich denke da muss man auch nicht lange überlegen. Allein die Menge der Artikel in ihrem Shop ist gewaltig. Zusätzlich ist Amazon omnipräsent, das heißt es gibt eine enorme geographische Verteilung. Das zieht Regionen mit verschiedenen Nutzungs-Peaks nach sich. Weiterhin leistet Amazon enorme Arbeit was Analyse, Navigationsoptimierung und Empfehlungssysteme betrifft. So sieht man bei einer Shoppingtour permanent was Nutzer die den im Moment betrachteten Artikel gekauft haben zusätzlich erworben haben. Weiterhin bewerten Nutzer die einzelnen Produkte und schreiben Rezessionen. Man sieht also Datenmengen soweit das Auge reicht.[17]

[16] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.11 ff
[17] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.8 ff

4. „W“ Fragen für den Umgang mit Big Data

Um ein erfolgreiches Vorgehen im Umfang mit Big Data zu gewährleisten muss man sich selbst die richtigen Fragen stellen. Wenn man in einem Unternehmen mit Unmengen an Daten konfrontiert ist, aber keinerlei strategisches Vorgehen hat, ist man mehr oder weniger verloren. In diesem Kapitel wird ein Vorgehen skizziert, das dabei helfen kann eine Struktur in seine Big Data Problematik zu bringen.

Eine Übersicht über die Zusammenhänge der Big Data Architektur sieht man in Abbildung 3. Hier wird ersichtlich, dass beim Thema Big Data ein komplexes Zusammenspiel aus vielen Einflüssen und Faktoren zusammenläuft. Um die Antworten auf alle Fragen zu bekommen, muss man sich also mit vielen Disziplinen und Themen auseinandersetzen.

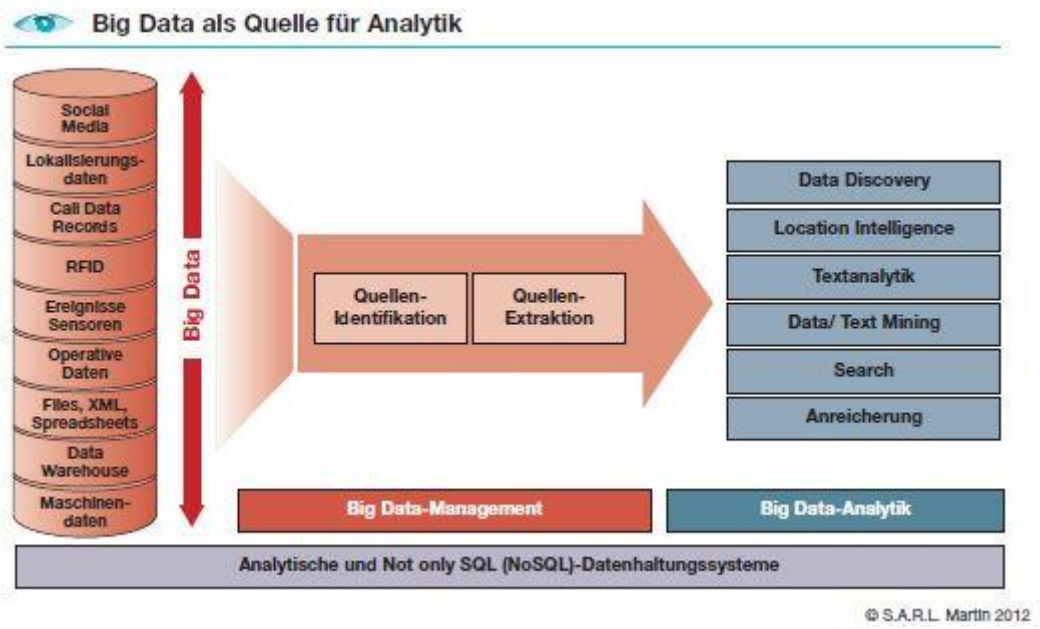


Abbildung 3 : Big Data Architektur[18]

4.1. Das „Was?“

Am Anfang steht das „Was?“. Bei Big Data ist das die schwierigste Frage, die es zu klären gilt. Wenn man die Antwort auf diese erste zentrale Frage kennt, fällt es leichter die folgenden zu beantworten. Man muss sich zuerst klarmachen, dass Big Data vorrangig ein zentrales Tool im Werkzeugkasten der IT ist.

[18] Vgl. Big Data –Dr. Wolfgang Martin S.14

Einen kleinen Ausblick habe ich schon mit den Fallbeispielen von Google und Amazon gegeben. Auch weitere große Firmen wie Facebook, Twitter, Xing und viele weitere sind Vorreiter in diesem Themengebiet.

Daten sind nutzlos wenn sie nicht in Informationen umgewandelt werden. Sie liegen nur rum und kosten Geld. Wichtig ist es also Daten zuerst schnell zu Informationen zu verarbeiten und anschließend diese wiederrum schnell zu nutzen.

Als Beispiel nehmen wir hierfür Kundendaten. Wir gehen davon aus, dass diese in einem Unternehmen schon in großer Menge vorhanden sind. Nun ist die Frage was wir mit diesen Daten anfangen. Man könnte damit beginnen, diese zu zählen. Das gibt erste Einblicke in Bezug auf Kundenanzahl. Interessanter wird es aber erst, wenn wir die beiliegenden Informationen verwerten. Alter, Umsatz, Einkaufszeit, Bestellhistorie, angesehene Artikel. Man könnte hier immer weiter fortfahren. Aus diesen Daten lassen sich bestimmt schon mehrere interessante Informationen herausfiltern. Vertriebsmitarbeiter können hier tolle Graphen erstellen. Diese zeigen beispielsweise prognostizierte Umsätze für kommende Quartale oder auch welche Artikel von Kunden, in welchem Alter gekauft werden. Wenn man noch einen Schritt weiter geht und den Aufwand erweitert, wäre es doch sicher interessant persönliche Vorlieben ihrer Kunden zu kennen. Ich denke da an die Lieblingsfarbe, das Lieblingsgericht oder was man sonst so für sein Unternehmen gebrauchen kann. Wie gesagt hier ist der Aufwand größer, man könnte etwa soziale Medien wie Facebook etc. anzapfen. Aber dazu später mehr. [19]

Resultierend aus diesen Informationen wäre es nun möglich die Werbung und die Auswahl der Produkte genauer auf die Kunden zuzuschneiden. Wenn ein Kunde einen Vorschlag bekommt, der seinen Geschmack trifft, ist dieser begeistert. Die Chance auf einen Verkauf steigt.

Eine weitere sehr interessante Möglichkeit eröffnet die Gruppendynamik. Die meisten Menschen bilden Gruppen mit gemeinsamen Interessen. Diese Gruppen bilden schnell eine Dynamik, die sich für ein Unternehmen nutzen lässt. Ermöglichen sie es den Nutzern auf ihren Portalen Communities zu bilden. Beispielsweise über ein Forum in dem sich Gleichgesinnte austauschen können. Wenn sie Zugriff auf persönliche Informationen besitzen, lassen sich diese ausnutzen. Wenn sie bemerken, dass 2 ihrer Kunden Fans des FC Bayern Münchens sind, empfehlen sie den beiden den Umgang. Am einfachsten wäre es wenn sie es ermöglichen einen Facebook oder Twitter Account mit ihrem Portal zu vernetzen. Diese Gruppendynamik sorgt dafür dass sich die Leute wohlfühlen. Das Portal wird zu einem Teil ihres Lebens.[20]

Zusammenfassend lassen sich mit Big Data also Wettbewerbsvorteile, Einsparungen und Umsatzsteigerungen erzielen und innovative neue Geschäftsfelder erschließen.

Hier nochmal ein paar ausgewählte Beispiele um die Vorstellung der Frage „Was?“ zu verdeutlichen:

- Optimierung und Personalisierung von Werbemaßnahmen und Steigerung von Cross- und Up-Selling aufgrund von besserem Kunden- und Marktwissen
- Besseres Risiko-Management in Zahlungs- und Handels-Strömen durch Entdeckung von Abweichungen und Unregelmäßigkeiten

[19] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.21
[20] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.27

- Aufbau flexibler und intelligenter Abrechnungssysteme in der Versorgung (Strom, Wasser, Gas) und Telekommunikation
- Erkennen von Interdependenzen und automatisierte Hypothesenbildung in Wissenschaft und Forschung[21]

4.2. Das „Wie?“

In diesem Kapitel sollen dem Leser nahe gebracht werden„ Wie“ sich das „Was“ realisieren lässt. Aber auch welche limitierenden Faktoren es gibt und welche wissenschaftlichen und technischen Voraussetzungen erfüllt sein müssen.

Im Kapitel „Was?“ ging es darum die Big Data Schmerzen, also die Probleme, aufzuzeigen. Im „Wie?“ werden allgemeine Mittel erklärt mit denen man sich diesen Problemen widmet.

Ein generelles Problem in der IT-Welt ist, dass das Motto „Never touch a running System“ sich in den Köpfen von IT-Entscheidern eingebrannt hat. So wird anstatt neues zu wagen jede Veränderung mit Angst beäugt und oft abgestoßen. So auch beim Thema Big Data. Aber gerade hier gilt, dass nur die Unternehmen gewinnen können die etwas wagen.[22]

Mathematik – ein Muss

Big Data dreht sich viel um angewandte Wissenschaften. So spielt für die Frage „Wie?“ die Mathematik eine zentrale Rolle. Erfolg beruht hierbei bei dem Kennen bzw. Beherrschen von Algorithmen und dem Trainieren von diesen. Man muss sich zwangsweise mit Wahrscheinlichkeiten, Statistik und Prognosen auseinandersetzen. Bereiche die hier besondere Bedeutung haben sind das Machine Learning[23] und Natural Language Processing.[24]

Machine Learning gibt es Supervised und Unsupervised. Supervised bedeutet überwachtes Lernen. Der Algorithmus lernt eine Funktion aus gegebenen Paaren von Ein- und Ausgaben. Dabei stellt während des Lernens ein „Lehrer“ den korrekten Funktionswert zu einer Eingabe bereit. Ein Teilgebiet des *überwachten Lernens* ist die automatische Klassifizierung. Anwendungsbeispiel: Handschrifterkennung. (Wikipedia)

Unsupervised steht für unüberwachtes Lernen. Hier ist es möglich fast ganz ohne Training von Algorithmen auszukommen. Hier werden Cluster Verfahren zur Hilfe genommen um charakteristische Eigenschaften herauszufiltern.

Beim Natural Language Processing geht es konkreter zu. Hier gibt es Möglichkeiten der Maschine Textverständnis beizubringen. Methoden sind hier beispielsweise die Sentiment Analysis (Text in positiv oder negativ einteilen um zu klassifizieren), die Document Classification (um welchen Art von Text handelt es sich) und einige weitere. Bei der Sentiment Analysis kann man die Struktur der Sätze

[21] Vgl. Big Data – Dr. Wolfgang Martin S.4
[22] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.37
[23] http://de.wikipedia.org/wiki/Maschinelles_Lernen (aufgerufen am 17.4.2014)
[24] http://en.wikipedia.org/wiki/Natural_language_processing (aufgerufen am 18.4.2014)

und Paragraphen ignorieren und sich darauf beschränken mithilfe von vorklassifizierten Daten die Anzahl der positiven und negativen Wörter zu ermitteln. Schwierigkeit beim Natural Language Processing Learning ist aber, dass es sehr schwer ist mit sarkastischen oder ironischen Texten umzugehen.[25]

Polyglott[26] – mehrschichtig orientieren

Die nächste Frage, um dem „Wie?“ gerecht zu werden, zielt auf die passende Programmiersprache ab. In der komplexen Big Data Welt wird es nicht möglich sein alles auf einer Plattform mit einer Programmiersprache abzuwickeln. Wie oben erwähnt spielt die Mathematik eine große Rolle, hierbei sind 2 der effizientesten Systeme Phyton und R. Auch funktionale Programmiersprachen wie Scala, Clojure und F# sind aufgrund ihrer mathematischen Prägung relevant. All diese Sprachen reizen das menschliche Gehirn mehr als Java, was für wissenschaftliche Entwicklungen ein Vorteil ist. Vor allem im Bereich Big Data, wo die Mathematik eine große Rolle spielt.[27]

Gleiches Spiel hat man beim Thema Datenbanken, auch hier ist es selten sinnvoll sich auf eine einzige Datenbank zu fixieren. Aufgrund der oben genannten Varietät ist es zwangsweise notwendig verschiedene Speicherarten zu lernen.

Man darf sich also nicht scheuen neue Programmiersprachen und Datenbanksysteme kennenzulernen.

Logs – effiziente Analyse

Um Big Data zu betreiben ist es unumgänglich jede Aktion die ein Nutzer auf einem System durchführt zu protokollieren. Am besten nicht die einzelne Tätigkeit, sondern den ganzen Fluss der Aktionen. Dadurch fallen zwangsweise viele Daten an. Aus diesen lassen sich interessante Informationen gewinnen, diese werden in oft sehr simplen, leicht lesbaren Logs, abgespeichert. Die Analyse dieser Informationen kann interessante Fakten offenlegen.

Beispiele :

- Zeit die der Nutzer auf der Seite verbracht hat -> kurze Zugriffszeit = Kunde hat schnell das Interesse verloren (mangelnde Übersicht)
- Produkte, die der Nutzer angesehen hat -> Rausfiltern der Interessen des Kunden (Produktempfehlungen)
- Nutzer musste sich einloggen, um Produkte zu sehen -> Reine Einsicht ohne Login ermöglichen
- Zugriffszeiten auswerten -> Vermeidung von möglichen Peaks
- Klick-Zähler -> Interessante Produkte ermitteln und hervorheben[28]

[25] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.42
[26] http://de.wikipedia.org/wiki/Polyglott (aufgerufen am 19.4.2014)
[27] Vgl. Big Data für IT-Entscheider –Pavlo Baron S.48-49
[28] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.50ff

Nutzung sozialer Netzwerke

In Kapitel 2 habe ich schon die Nutzung von sozialen Medien angeschnitten. Die Daten die Unternehmen aus Facebook, Twitter, Xing, usw. herausziehen könnten, sind Gold wert. Ich schreibe hier bewusst könnten, weil die Beschaffung dieser Daten mit einigen Hürden verknüpft ist.

Fangen wir mit Facebook an. Welche Daten könnten Unternehmen hier interessieren? Hier einige Beispiele :

- Die „Likes“ = hiermit lässt sich leicht verfolgen für was sich eine Person interessiert. Das geht vom Lieblingsfussballverein, über Johnny Deep, bis hin zum Urlaub in Spanien. Wenn man mit diesen Informationen in ein Verkaufsszenario kommt, hat man einen großen Wettbewerbsvorteil. Man kann der Person Sachen anbieten von denen man bereits weiß, dass sie sie begeistern.
- Statusmeldungen = hier lässt der User persönliche Beiträge in Facebook einfließen. In denen bindet er noch zusätzlich Orte und Personen ein, mit denen er diese erlebt.
- Fotos = auch durch Fotos kann man einiges über seine potenziellen Kunden herausfinden. Alleine schon durch die Kleidung die er gerne trägt. Oder aber auch Hobbys lassen sich ableiten.

Nun zu den Hürden beim Beschaffen dieser Daten. Wenn ein Facebook-Nutzer weiß worauf er sich einlässt, kann er den Riegel vorschieben. Durch ein paar simple Konto-Einstellungen entscheidet der Nutzer wer Zugriff auf seine Informationen hat. Wenn er den kompletten Zugriff auf sich sperrt, lassen sich nur noch ein paar statistische Daten ermitteln. Tatsache ist aber das viele Nutzer ein reges Mitteilungsbedürfnis haben und aus diesem Grund ihre Accounts öffentlich einsehbar lassen. Vor allem Personen die im öffentlichen Interesse stehen.

Als Nächstes sehen wir uns Xing an. Ich nenne es hier einfach mal das Facebook für Geschäftsleute. Ich nehme Xing, weil wir bei uns im Unternehmen (Adesso AG) hiermit Ansprechpartner von interessanten Firmen für unser CRM[29]-System recherchieren. Auch hier gibt es wieder ein paar Restriktionen auf den Zugriff der Daten. Als erstes braucht man einen Premium-Account, ohne diesen sind die Möglichkeiten auf Xing extrem limitiert. Beispielsweise kann man sogar nur begrenzt Leute anklicken, wenn man nur über einen Standard-Account verfügt. Die monatlichen Kosten für einen Premium Account sind allerdings sehr überschaubar. Die zweite Hürde ist ähnlich wie bei Facebook, man muss mit den Nutzern befreundet sein um alle Informationen einsehen zu können. Vorteil ist bei Xing aber das man den „Header“ sehen kann. Aus dem Header lassen sich folgende Informationen ableiten:

- Der vollständige Name
- Der akademische Titel
- Das aktuell als Arbeitsplatz eingetragene Unternehmen
- Den Ort

[29] http://de.wikipedia.org/wiki/Customer-Relationship-Management (aufgerufen am 21.4.2012)

In Abbildung 4 ist der Header meines inaktiven Xing Profils zu sehen. Diese Daten sind für jeden sichtbar der mich anklickt. Selbst bei diesem kaum gepflegtem Account könnte man das Stichwort CMS herausfiltern. CMS steht für Content-Management-System, und Firmen die diese Systeme anbieten könnten hier einhaken. Natürlich ist es den Nutzern möglich hier alles anzugeben. Aber das ist im Internet fast überall so. Die eventuell mangelnde Richtigkeit muss man bei der Nutzung sozialer Netzwerke wohl zwangsweise in Kauf nehmen.

Wenn der User mehr Informationen preisgibt, lässt sich auch seine Historie in der Arbeitswelt begutachten. Zusätzlich gibt es noch die Felder „ich suche“ und „ich biete“, aus denen sich Fachgebiete und persönliche Interessen ableiten lassen. So ist es möglich mit den richtigen Begriffen in der Suchmaske alle Führungskräfte der BMW AG in München zu suchen, die in ihrem Profil an irgendeiner Stelle den Begriff JAVA angegeben haben.

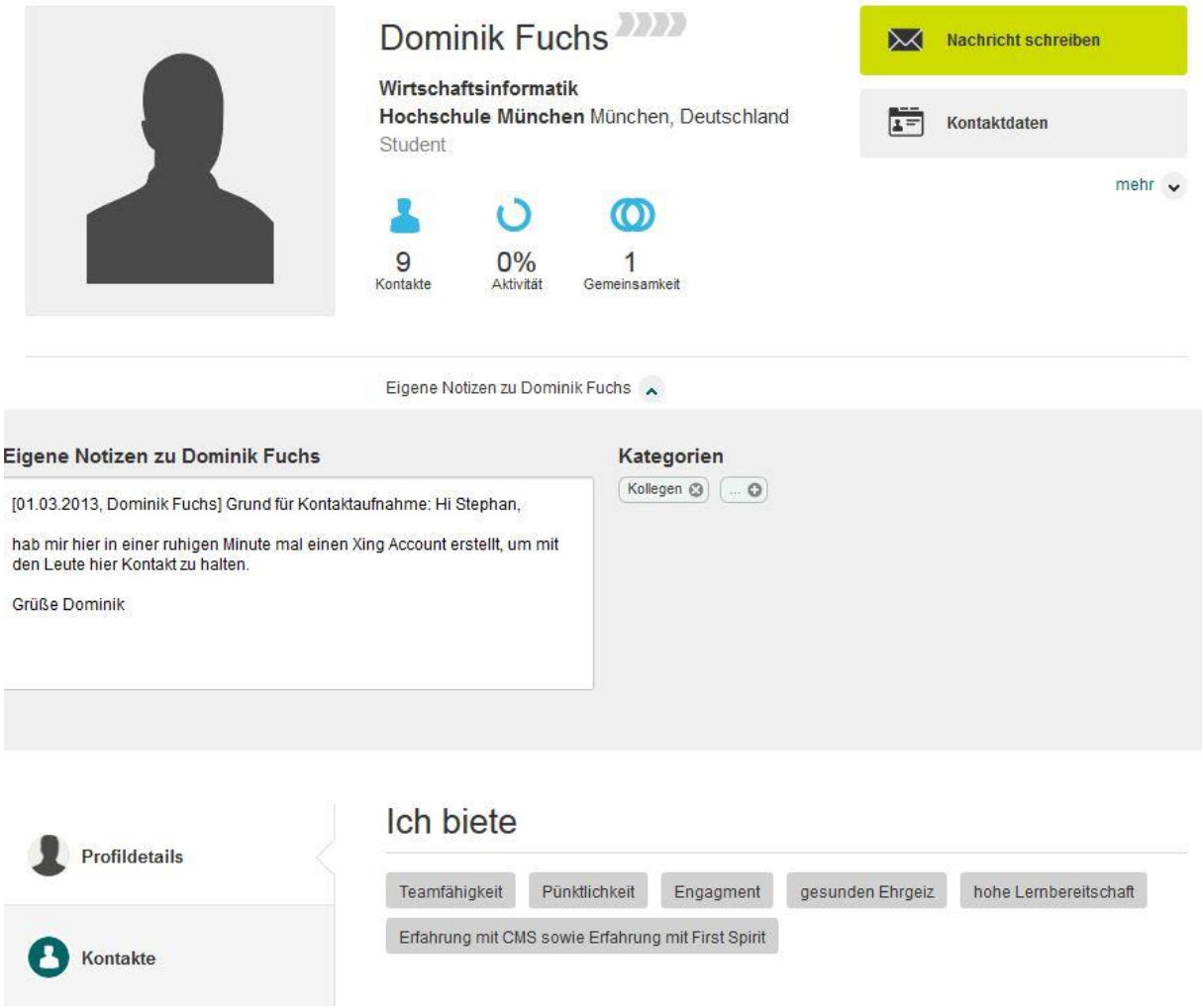


Abbildung 4 : Xing Header

Persönlich sehe ich die Nutzung dieser Daten als sehr kritisch an. Man überschreitet hier sehr schnell eine moralische Grenze. Viele Leute, mich eingeschlossen, sehen es als kritisch wenn fremde Personen alle möglichen persönlichen Daten über sie sammeln. Da das vor allem unwissentlich geschieht wird einem sozusagen auch das Mitspracherecht genommen. Das erzeugt Misstrauen und auch teilweise ein Unwohlbefinden. Meiner Meinung nach sollte eine Grenze eingehalten, die die Privatsphäre der Personen berücksichtigt. Ich kann aber auch die Verlockung der Unternehmen verstehen, die sich hinter den ganzen persönlichen Daten verbirgt. Man sollte sich hier aber die Frage stellen ob die Firmenethik ein so extremes Maß an Kundentransparenz gutheißt. Vor allem Ehrlichkeit sollte ein wesentlicher Schritt im Umgang mit den Daten sein. Wenn ein Kunde genau darüber informiert wird, was mit seinen Daten geschieht, kann er eine Wahl treffen. Wenn er der Seite/dem System vertraut ist er eventuell bereit, zusätzliche Daten bereitzustellen. Nicht ist schädlicher als schlechte Publicity.

4.3. Das „Womit?“

Beim „Womit?“ handelt es sich um die leichteste aber auch um die konkreteste Frage. Hier zeige ich einige Tools mit denen hier gearbeitet werden kann. Im speziellen Tools, die auf der Programmiersprache Java aufsetzen.

Verteilte Systeme

Verteilte Systeme gelten als der einzige Weg um flexibel und weit skalieren zu können. Diese nehmen eine zentrale Position in der Handhabung der Big Data Problematik ein.

Ein **Verteiltes System** ist nach der Definition von Andrew Tanenbaum ein Zusammenschluss unabhängiger Computer, die sich für den Benutzer als ein einziges System präsentieren. Peter Löhr definiert es etwas grundlegender als „eine Menge interagierender Prozesse (oder Prozessoren), die über keinen gemeinsamen Speicher verfügen und daher über Nachrichten miteinander kommunizieren“. (Wikipedia)

Für die Big Data Thematik sind vor allem verteilte Datenspeicher, verteilter Cache und verteilte Worker Cluster wichtig. Zwischen verteilten Systemen findet idealerweise ein Nachrichtenaustausch statt. Am besten erflogt dieser asynchron. Vorteil hierbei ist, dass das Empfängersystem zum Zeitpunkt des Funktionsaufrufes nicht verfügbar sein[30]

[30] Vgl. Big Data für IT-Entscheider –Pavlo Baron S.118

Datenhaltung

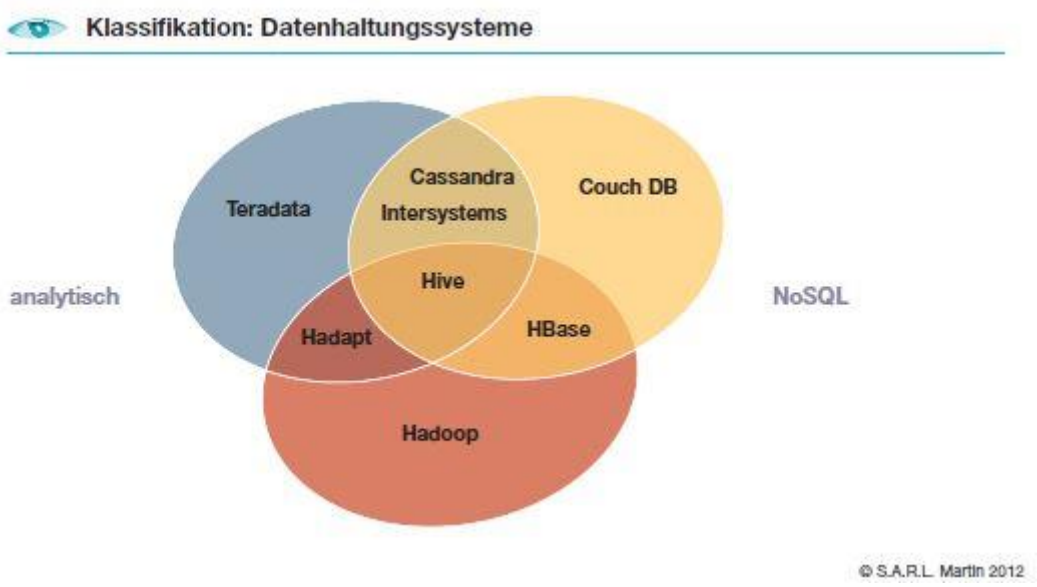


Abbildung 5 : Datenhaltungssysteme[31]

Man geht heute davon aus, dass die zu analysierende Datenmenge in Form von Volumen und Quellenanzahl schneller steigt als die Leistung von traditionellen Datenbanken zunimmt. Das hat erhebliche Performance Probleme zur Folge. Aus diesem Grund haben sich neue Technologien zur Datenhaltung entwickelt, wie in Abbildung 5 sichtbar.

Interessant ist hier vor allem **Hadoop**. Dies ist ein Framework für skalierbare, verteilt arbeitende Software, mit der Zielsetzung intensive Rechenprozesse mit großen Datenmengen auf Clustern von Rechnern durchzuführen. Aber auch NoSQL-Datenbanken gewinnen an Bedeutung. [32]

NoSQL (englisch für *Not only SQL*) bezeichnet Datenbanken, die einen nicht-relationalen Ansatz verfolgen und damit mit der langen Geschichte von relationalen Datenbanken brechen. Diese Datenspeicher benötigen keine festgelegten Tabellenschemata und versuchen, Joins zu vermeiden, sie skalieren dabei horizontal. Im akademischen Umfeld werden sie häufig als „strukturierte Datenspeicher“ (engl. *structured storage*) bezeichnet. (Wikipedia)

Als dritte Möglichkeit gibt es noch **analytische Datenbanken**. Diese weisen folgende Konzepte auf:

- Spalten-Orientierung
- Daten-Komprimierung
- Zugriffsmethoden und Algorithmen (Map Reduce)
- Parallelisierung
- In Memory Verarbeitung[33]

[31] Vgl. Big Data – Dr. Wolfgang Martin S.15
[32] Vgl. Big Data – Dr. Wolfgang Martin S.15
[33] Vgl. Performance Managment und Analytik –Dr. Wolfgang Martin S.8

In diesem Kapitel gebe ich nur eine grobe Übersicht über verschiedene Datenhaltungssysteme, die im Big Data Umfeld eine Rolle spielen. Später in der Arbeit gehe ich noch genauer auf ein Paar von ihnen ein.

NoSQL

Man spricht im Thema NoSQL von sogenannten Data Stores, diese ersetzen die sonst benutzen relationalen Datenbanksysteme. Die Data Stores bauen auf mehreren Grundprinzipien auf. Nach denen gliedere ich die verschiedenen Systeme im Umfang dieser Arbeit. Es gibt zwar diverse Überlappungen zwischen den einzelnen Vertretern der Arten, aber um einen Überblick zu geben ist die Aufteilung definitiv sinnvoll.

1. Key/Value Stores: Diese Art von Stores baut auf einer Hash-Tabelle, also einer Streuwerttabelle, auf. Das Hashverfahren ist ein Algorithmus zum Suchen bestimmter Datenobjekte in großen Datenmengen. Über einen Schlüssel wird ein Objekt eindeutig definiert.
2. Document Store: Hier wird ein Stück weiter gegangen als bei den Key bzw. Value Stores. Die Daten werden in einem relationslosen Modell als Dokumente abgespeichert. Dort liegen diese redundant vor. Diese Stores dienen also vorwiegend der Speicherung von Dokumenten. Gegenüber relationalen Datenbanken hat man eine schwächere Struktur. Man muss sich nicht um die Relationen bei verschiedenen Tabellen kümmern, wenn man in einem Feld eine Liste speichern möchte. Auch das nachträgliche Eintragen von neuen Feldern, die bisher nicht existiert haben, ist unproblematisch. Beim Auslesen der Daten wird üblicherweise mit einem Map/Reduce-Algorithmus gearbeitet. Dieser wird im nachfolgenden Kapitel im Detail beschrieben.[34]
3. In-Memory Stores: Diese Stores eignen sich vor allem für schnelles Laden und Speichern der Daten direkt aus dem Hauptspeicher der Platte, abgesichert durch einen Backup-Mechanismus. Zugunsten einer sehr hohen Performance können hier auch mal Daten verloren gehen. Wenn man also einen Anwendungsfall vorliegen hat, der diesen Verlust erlaubt, sind In-Memory Stores eine gute Alternative.[35]

Abschließend weise ich noch darauf hin, dass der NoSQL Markt extrem segmentiert ist. Es gibt noch sehr viele weiter Store-Arten. Und es kommen immer noch weitere hinzu.

Mir ist klar, dass die Beschreibungen hier nur an der Oberfläche kratzen. Im Sinne dieser Arbeit soll dem Leser aber auch nur ein Überblick über die Tools vermittelt werden. Es soll eine Vorstellung entwickelt werden, welche Möglichkeiten es gibt, um ein Gefühl für die Thematik zu entwickeln.

[34] http://eliteinformatiker.de/2011/05/18/nosql-document-store-couchdb-mongodb/ (aufgerufen am 25.04.2014)
[35] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.156

MapReduce

Dieser Begriff wurde schon ein paar Mal angeschnitten. Im Folgenden werde ich näher auf MapReduce eingehen.

Hier handelt es sich nicht um ein Tool, vielmehr um einen Ansatz. Es beinhaltet ein Programmiermodell für nebenläufige Berechnungen von Daten enormer Größe(Petabytes). Auf jedes Element der ursprünglichen Originalliste wird eine Funktion angewendet. Eine sogenannte Reduce-Funktion, die als Ziel hat die Liste auf weniger Werte zu reduzieren. MapReduce setzt auf Parallelität, um zu funktionieren werden mehrere Prozessoren bzw. Maschinen benötigt. Dem Reduce vorrausgehend findet erstmal das Mapping statt. Hier werden die eingehenden Daten mit Basisberechnungen durchforstet und lokal abgespeichert. Das sind die zwei Phasen die den Begriff geprägt haben, es gibt aber durchaus noch weitere Phasen. Zum einen der Split, dieser teilt die Daten auf die verschiedenen Maschinen im Cluster auf, oder der Combine, der direkt nach der Map-Phase greift um schon direkt dort die Daten zu reduzieren.[36]

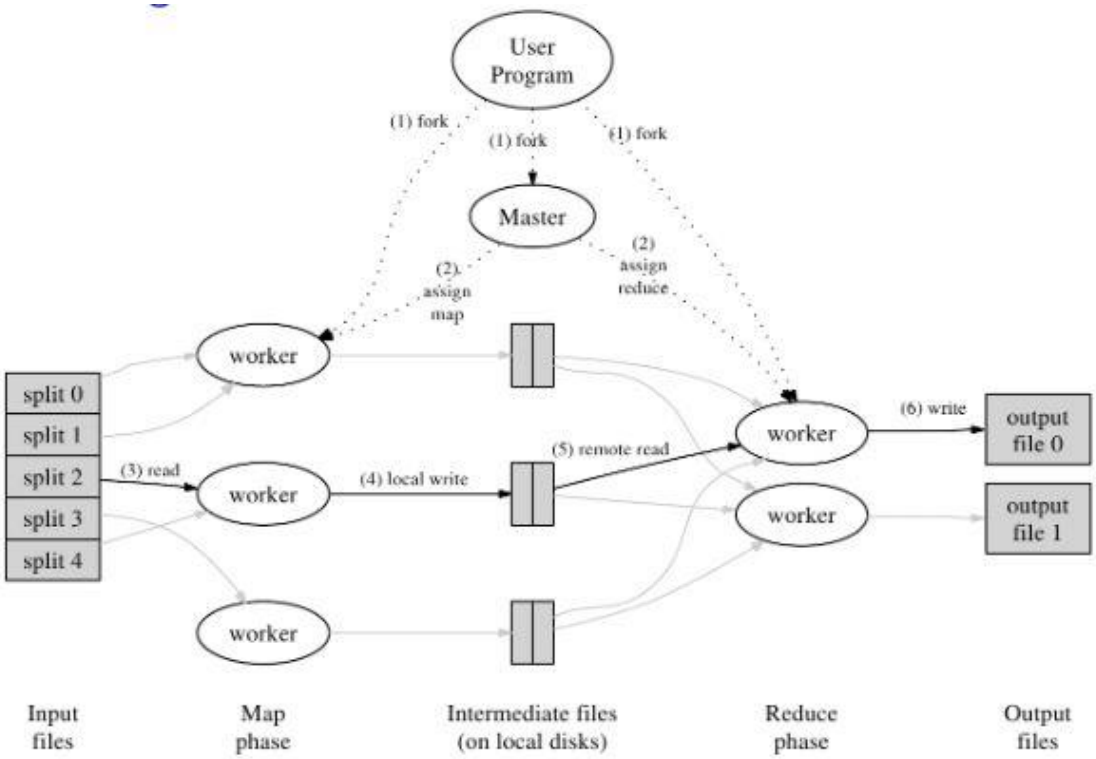


Abbildung 6 : MapReduce Konzept[37]

Abbildung 6 zeigt den von mir grob beschriebenen Vorgang nochmal auf. Das sollte helfen um die einzelnen Schritte nachzuvollziehen. Man sieht das in den einzelnen Phasen Worker-Threads die Arbeit erledigen. Der Master, ein Scheduler, verteilt Jobs an eben diese Worker. Hierbei wird die Lokalität der Daten beachtet.[38]

Ein sehr simpler Anwendungsfall ist das Zählen von Wörtern oder das Finden von Anagrammen.

---

[36] Vgl. Big Data für IT-Entscheider – Pavlo Baron S.163
[37] Vgl. Map Reduce Eine kleine Einführung – Simon Elsbrock
[38] Vgl. Map Reduce Eine kleine Einführung – Simon Elsbrock

In Abbildung 7 sieht man den Vergleich von traditionellen relationalen Datenbanksystemen(RDBMS) mit Map Reduce. Bei MapReduce werden die Daten im Stream bearbeitet, wobei sie linear skalieren. Bei RDBMS hingegen werden gezielte Suchanfragen gestartet, die deutlich langsamer arbeiten. MapReduce eignet sich für semistrukturierte bis unstrukturierte Daten. RDBMS benötigen aber strukturierte Daten.[39]

Viele wichtige Tools im Big Data Umfeld bauen auf dem MapReduce Prinzip auf. Eines der populärsten davon ist Apace Hadoop. Diesem Tool widme ich das nächste Kapitel.

	Traditional RDBMS	MapReduce
Data Size	Gigabytes	Petabytes
Access	Interactive and Batch	Batch
Updates	Read and write many times	Write once read many times
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear

Abbildung 7 : Vergleich von RDBMS mit MapReduce[40]

Eine ganz aktuelle Beobachtung zeigt aber das MapReduce langsam aber sicher veraltet. Hauptsächlich weil es eine hohe Latenz aufweist und deswegen nicht mehr performant ist. Alternativen, ich nenne hier Spark, haben einige Vorteile. Hat MapReduce eine Latenz von etwa 15 Sekunden pro Job, beeindruckt Spark mit einer Latenz 0.01 Sekunden. Zusätzlich hat Spark im Gegensatz zu MapReduce einen In-Memory Support. Das heißt beim wiederholten Ausführen der Jobs erreicht Spark eine noch bessere Performanz. Auch ein Streaming Verfahren wird unterstützt. Spark punktet somit vor allem im Echtzeit Bereich, aber auch im Batch Bereich setzt es sich immer mehr durch.[41]

Hadoop

Hadoop ist also eine Implementierung des MapReduce-Konzepts. Es handelt sich hierbei um ein Open Source Projekt der Apache Software Foundation, was bedeutet es ist auch offen für freiwillige Teilnehmer.[42]

[39] Vgl. MapReduce-Konzept – Thomas Findling & Thomas König S.22
[40] Vgl. MapReduce-Konzept – Thomas Findling & Thomas König S.23
[41] Vgl. Low-Latency Anwendungen mit Hadoop – Dr. Henrik Behrens
[42] http://hadoop.apache.org/ aufgerufen am 26.04.2014

Wie im vorigen Kapitel erwähnt eignet es sich zur Bearbeitung sehr großer Datenmengen, also im Petabyte-Bereich. Recheneinheiten werden zu Cluster-Einheiten zusammengefasst und auf den Recheneinheiten werden die Jobs parallel abgearbeitet. Hadoop hat eine hohe Fehlertoleranz. Pro Job geschehen durchschnittlich 1-2 Fehler. Einschränkung ist das ein Download nur unter Unix/Linux verfügbar ist. Die Programmierung ist mit Java, C++, Phyton und vielen weiteren Sprachen möglich. Hadoop an sich benötigt aber Java 1.6 . Für das Cluster wird ein gemeinsames Dateisystem benutzt, das Hadoop Distributed File System(HDFS). In dieses müssen die Eingabedateien vor der Verarbeitung zuerst reinkopiert werden.[43]

Bei Hadoop gibt es einen Job-Tracker. Dieser koordiniert Jobs und teilt sie in Tasks auf. Für die Tasks wiederrum gibt es einen Task-Tracker. Hier werden die Map- sowie die Reduce Funktion durchgeführt. Es wird sichergestellt das pro Prozessorkern nur ein Task läuft.[44]

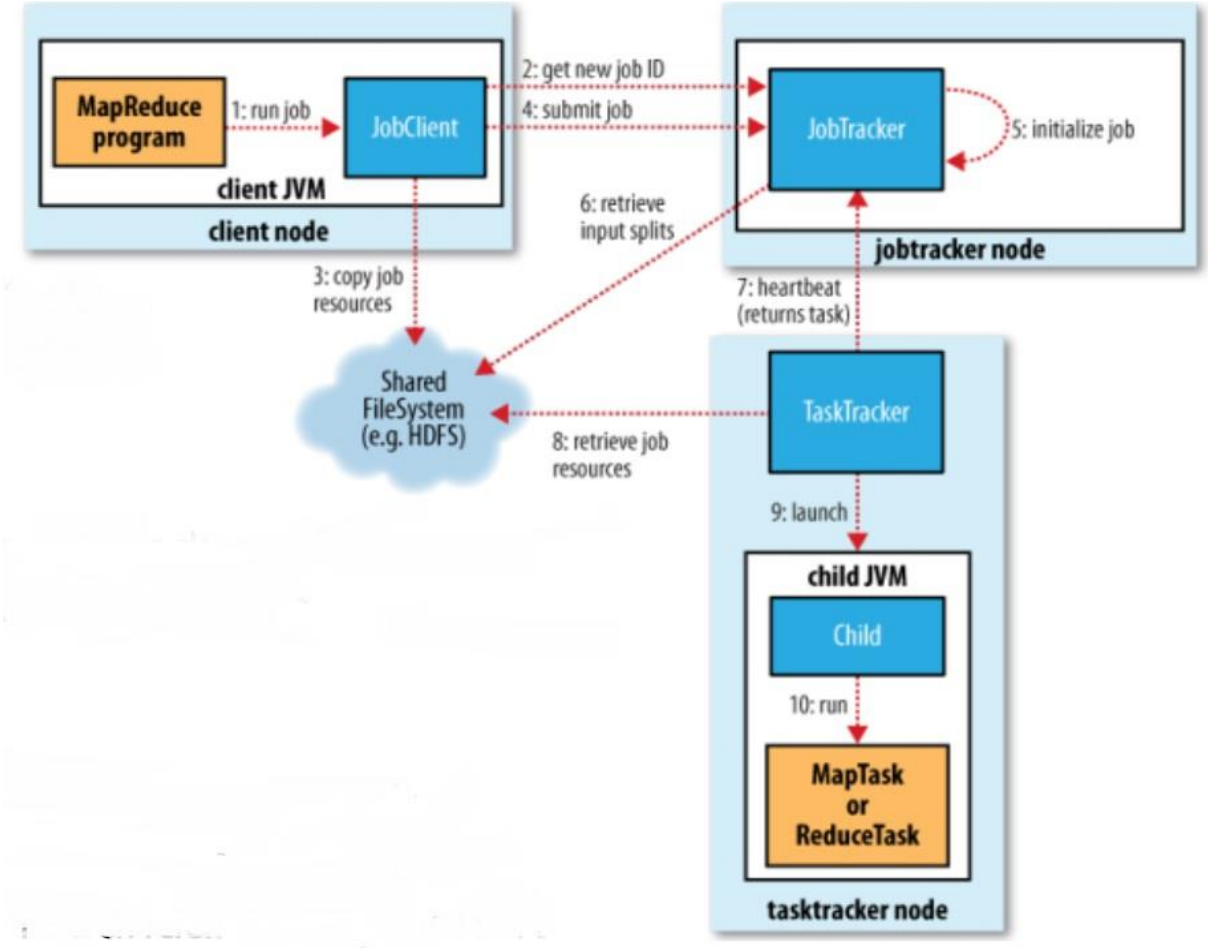


Abbildung 8 : Hadoop Funktionsweise[45]

Abbildung 8 zeigt die Funktionsweise von Hadoop. Ich breche im Folgenden nur kurz die einzelnen Schritte herunter:

1. Run job: Eine neue Job-Client Instanz wird erzeugt
2. Get new job ID: Eine neue Job-ID wird erzeugt

[43] Vgl. MapReduce-Konzept – Thomas Findling & Thomas König S.25ff
[44] Vgl. MapReduce-Konzept – Thomas Findling & Thomas König S.25ff
[45] Vgl. MapReduce-Konzept – Thomas Findling & Thomas König S.27

3. Copy job resources: alle Job-Daten wird in das gemeinsame HDFS kopiert
4. Submit job: Job wird an den Tracker weitergeleitet, dabei werden In- und Output Spezifikationen überprüft
5. Initialize job: Job wird initialisiert und ggf. in mehre Tasks unterteilt
6. Retrieve input splits: Abholen der Input-Splits(logische Referenzen auf einen Block)
7. Heartbeat: anhand des Job-Scheduling wird den Task-Trackern ein spezieller Task zugeordnet
8. Retrieve job resources: Daten und Bibliotheken werden für aktuellen Task abgeholt
9. Launch: für jeden Task wird eine neue JVM zur Ausführung gestartet
10. Run: Hier wird der Task nun letztendlich ausgeführt, je nach Erfolg wird ein „finished“ oder ein „failed“ zurückgeben. Nach Abschluss aller Tasks wird „finished“ zurückgegeben.[46]

5. Fazit & Ausblick in die Zukunft

5.1. Entwicklung des Big Data Marktes

Wie jeder Markt unterliegt auch der Big Data Markt stetigen Veränderungen. Trends kommen und gehen, Begriffe tauchen auf und verschwinden. Wie bei Map Reduce beschrieben kommt es vor, dass Technologien aus der Mode kommen und durch neue ersetzt werden. Es ist wichtig diese ganzen Entwicklungen zu verflogen und immer ein Auge auf die Zukunft zu haben.

Annäherung des SQL Marktes

Im Big Data Umfeld herrscht eine Menge Chaos. Chaos in den Technologien, sowie Chaos im Hype. Nahezu jeden Tag werden neue Datenhaltungssysteme angepriesen, das neue Wunderheilmittel für die Big Data Probleme zu sein. Genauso tauchen immer wieder neue „ultimative“ Wege auf, mit denen sich Daten mit mehr als Echtzeit analysieren lassen. Trotzdem lassen sich langsam immer mehr Muster erkennen. Beispielsweise nähern sich die Welten von NoSQL, NewSQL und RDBMS an.[47]

Bei NewSQL handelt es sich um ein Zusammenspiel von klassischer SQL mit den neuen Big Data Technologien. Diese Technologie gewinnt immer mehr an Bedeutung, da es eben auf einer wohlbekannten interaktiven Schnittstelle aufbaut, aber zusätzlich für große Datenmengen tauglich ist.

[46] MapReduce-Konzept – Thomas Findling & Thomas König S.27-36
[47] Big Data für IT-Entscheider – Pavlo Baron S.194

Umsatzprognosen

Interessant ist auch wie der Markt für Big Data prognostiziert wird. In Abbildung 9 sieht man den erwarteten Umsatz der sich hinter Big Data verbirgt. Hierzu zählen sowohl Hardware, Software als auch Services. Es wird von einem kontinuierlichen Wachstum ausgegangen. 2017 soll sich der Umsatz verzehnfacht haben, von 2012 $5,4 Milliarden auf $53,4 Milliarden. Das liegt daran das Unternehmen quer über alle Branchen von den hier in der Arbeit bereits erwähnten Vorteilen wie Kosteneinsparungen und Umsatzsteigerungen profitieren wollen.

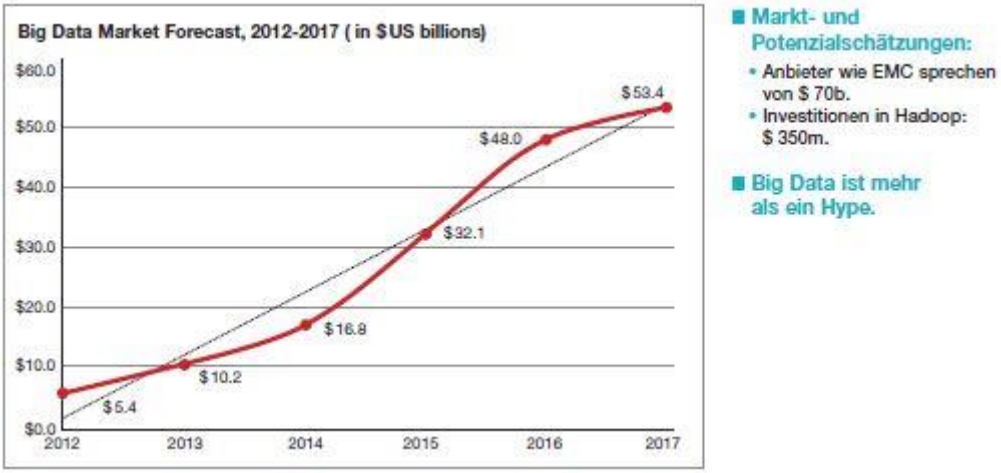


Abbildung 9 : Vorhersage der Big Data Gewinnspanne[48]

Im Gegensatz zu dieser gut aussehenden Entwicklung steht Abbildung 10. Hier sieht man die Top Unternehmen, die mehr als $100 Millionen Umsatz mit Big Data machen. Mit der Ausnahme von TerraData machen die Big Data Umsätze nur einen marginalen Anteil vom Gesamtumsatz(0-1%).

Stand 2012 geht man davon, dass der Big Data Markt noch ganz am Anfang steht. Man rechnet auch in diesem Markt mit einer großen Übernahmewelle ausgehend von den großen, führenden Unternehmen.

[48] http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues (aufgerufen am 23.04.2014)

Big Data-Umsätze großer IT-Anbieter

Total 2012 Big Data Revenue by Vendor

Vendor	Big Data Revenue (in $US millions)	Total Revenue (in $US millions)	Big Data Revenue as Percentage of Total Revenue
IBM	$1,100	$106,000	1%
Intel	$765	$54,000	1%
HP	$550	$126,000	0%
Oracle	$450	$36,000	1%
Teradata	$220	$2,200	10%
Fujitsu	$185	$50,700	1%
CSC	$160	$16,200	1%
Accenture	$155	$21,900	0%
Dell	$150	$61,000	0%
Seagate	$140	$11,600	1%
EMC	$140	$19,000	1%
Capgemini	$111	$12,100	1%
Hitachi	$110	$100,000	0%

Abbildung 10 : Anteil von Big Data am Gesamtumsatz[49]

Kritik am Big Data Ansatz[50]

- Größere Datenmengen bedeuten nicht zwangsweise auch eine größere Qualität der Daten. Im Big Data Umfeld sucht man beispielsweise in den Daten nicht nach den Ausreißern um sie zu bereinigen, sondern um sie zu analysieren.
- Nicht alle Datenquellen sind gleich beziehungsweise vergleichbar. Statistische Hilfsmittel wie das Erheben einer repräsentativen Stichprobe werden oft vernachlässigt
- Ein großer Punkt ist auch das Überschreiten von ethnischen Grenzen. Die Frage ist inwieweit die gewollte Transparenz der Kunden mit der Unternehmensethik vereinbar ist. Auf die ethnische Grenze gehe ich im Kapitel „Mein persönliches Fazit“ nochmal ein

Hype

Um den Begriff des Hypes besser zu verstehen habe ich mit Abbildung 11 einen traditionellen, technischen Hype Kreislauf eingebunden. Es ist eine graphische Darstellung die die Reife von Technologien oder Anwendungen zeigt.

[49] Vgl. Big Data – Dr. Wolfgang Martin S.12
[50] Vgl. Big Data –Dr. Wolfgang Martin S.37

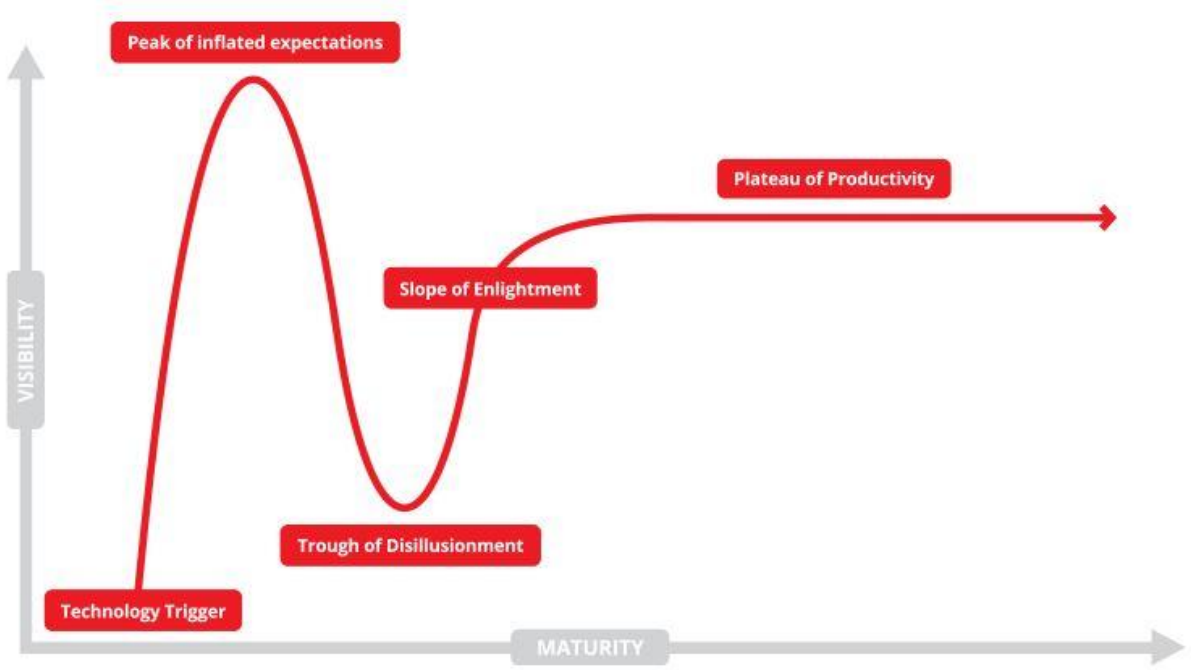


Abbildung 11 : Hype Kreislauf[51]

Es zeigt verschiedene Phasen, ausgehend von einem Auslöser bis hin zu einer Produktivität. Die Technologie muss die Phase des Hypes überwinden. Das gelingt ihr wenn sie von der generellen Masse genutzt wird. Die Phase der Ernüchterung setzt ein, wenn die gehypten Vorstellungen nicht eintreten oder zumindest nicht sofort. Laut Experten sollte Big Data den Abgrund/die Phase der Ernüchterung 2013 überwunden haben und nun Richtung Akzeptanz der Masse steuern.[52]

Persönliches Fazit

Handelt es sich bei Big Data um einen Hype? Ich würde sagen ja und nein. Ja in Bezug auf den Begriff, der nahezu wahllos auf alles angewandt wird was nur im Entferntesten mit Big Data zu tun hat. Nein in Bezug auf die sich dahinter verbergenden Technologien und Lösungsansätze. Hier erwarten Unternehmen die sich darauf einlassen eine Menge Chancen Kunden zu gewinnen oder an sich zu binden. Wenn Unternehmen beginnen auf die Inhalte hinter der Phrase Big Data zu achten, können die hier in der Arbeit erwähnten Vorteile realisiert werden. Unternehmen müssen sich den sogenannten Big Data Schmerzen also stellen. Durch Berücksichtigung der 3 Fragen „Was?“, “Wie?“ und „Womit?“ ist es möglich sich dem Thema angemessen zu nähern. Allem voran steht die Vorstellung was sich mit den Daten anstellen lässt und welche Daten relevant sein können. Wenn diese Frage ausführlich und gut beantwortet wird hat man in meinen Augen den größten Schritt geschafft.

Wie in der Arbeit beschrieben, die Empfehlung für das ultimative Tool oder die beste Programmiersprache ist nicht sinnvoll. Man muss das richtige für sich finden oder eine Mischung aus

[51] http://www.datameer.com/blog/big-data-analytics-perspectives/big-data-crossing-the-chasm-in-2013.html (aufgerufen am 29.04.2014)
[52] http://www.datameer.com/blog/big-data-analytics-perspectives/big-data-crossing-the-chasm-in-2013.html (aufgerufen am 29.04.2014)

verschiedenen. In meinen Augen ist Hadoop ein sehr gutes Tool. Es ist weit verbreitet, wird ständig erweitert und aktualisiert.

Wie im Abschnitt „Nutzung sozialer Medien“ beschrieben, stehe ich dem unbeschränkten Anzapfen von Plattformen wie Facebook oder Twitter kritisch gegenüber. Hauptsächlich wenn es um das Beschaffen und Nutzen persönlicher Daten geht. Das ist in Amerika aufgrund milderer Gesetze ein größeres Problem. Aber auch hier in Deutschland wird versucht jede Lücke auszunutzen. Seit dem NSA-Skandal ist die große Mehrheit der Bevölkerung aber nun alarmiert. Ich kann mir gut vorstellen, das es zukünftig immer schwerer wird, die Daten von Leuten zu überwachen.

In der Literatur und im Internet wird das Thema Big Data sehr kontrovers diskutiert. Teilweise wird Big Data nur als Technologie aufgefasst, teilweise nur als BI Lösung. Ich denke, dass sich das was hinter Big Data steckt in den nächsten Jahren immer weiter etablieren wird. Mich persönlich hat das Thema im Laufe der Recherchen immer mehr in seinen Bann gezogen.

Literaturverzeichnis

Anonymus, S. (18. 5 2011). *eliteinformatiker* . Abgerufen am 25. 4 2014 von http://eliteinformatiker.de/2011/05/18/nosql-document-store-couchdb-mongodb/

Baron, P. (2013). *Big Data für IT-Entscheider* . Hanser.

Behrens, D. H. (2014). *Low-Latency Anwendungen mit Hadoop.* München: SHS Viveon AG.

Elsbrock, S. (2011). *Map Reduce Eine kleine Einführung.* Mannheim: RaumZeitLabor.

Gottwald, M. (kein Datum). *SoftSelect*. Abgerufen am 17. 4 2014 von http://www.softselect.de/wissenspool/big-data

Horvath, S. (2013). *Aktueller Begiff Big Data.* Deutscher Bundestag.

Kelly, J. (15. 2 2012). *Wikibon*. Abgerufen am 23. 4 2014 von http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues

Live-Counter. (kein Datum). Abgerufen am 29.04.2014 von http://www.live-counter.com/internetnutzer-weltweit/

Martin, D. W. (2011). *Performance Managment und Analytik.* iBond.

Martin, D. W. (Juli 2012). *Big Data.* Strategic Bulletin.

Thomas König, T. F. (kein Datum). *MapReduce-Konzept.* Leipzig.

Wikipedia. (kein Datum). Abgerufen am 17. 4 2014 von http://de.wikipedia.org/wiki/Big_Data

Abbildungsverzeichnis