HANSER

Leseprobe

Jonas Freiknecht

Big Data in der Praxis

Lösungen mit Hadoop, HBase und Hive. Daten speichern, aufbereiten, visualisieren

ISBN (Buch): 978-3-446-43959-7

ISBN (E-Book): 978-3-446-44177-4

Weitere Informationen oder Bestellungen unter

http://www.hanser-fachbuch.de/978-3-446-43959-7

sowie im Buchhandel.

Inhalt

Vorwort

Die Verfügbarkeit von Daten hat sich in den vergangenen zehn Jahren drastisch verändert. Immer neue Datenquellen, die zunehmende Verbreitung mobiler, internetfähiger Geräte und natürlich alle Entwicklungen, die sich aus dem *Web 2.0* ergeben haben, tragen dazu bei, dass sich Unternehmen heute mit einer sehr viel größeren Datenmenge konfrontiert sehen, die es zu erfassen, zu speichern und auszuwerten gilt, als bisher. Dabei ist es nicht nur die Datenmenge selbst, die den Unternehmen Probleme bereitet, sondern darüber hinaus auch die Struktur und die Art der Daten sowie die Geschwindigkeit, mit der sie anfallen.

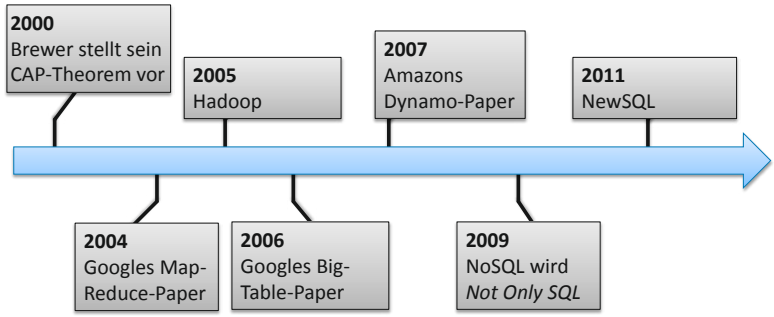
Die meisten Unternehmen sehen sich noch immer in einer klassisch relationalen Welt verhaftet, die sie sich selbst in den letzten Jahrzehnten aufgebaut haben. Über mehr als 40 Jahre hinweg haben relationale Datenbanken und SQL als standardisierte Abfragesprache sowie die dazugehörigen Architekturen die Datenhaltung, -verarbeitung und -auswertung beherrscht. Und auch heute stellt niemand infrage, dass der Großteil der Datenbanken, die wir in den Unternehmen vorfinden, diesem klassischen Paradigma entsprechen: Zeilenorientierte, weitestgehend normalisierte RDBMS-Systeme, die die Daten in Tabellen halten und über Beziehungen miteinander verknüpfen, um auf diese Weise ihre Datenbankschemata aufzubauen.

Viele der Problemstellungen, die sich in den vergangenen zehn Jahren entwickelt haben, passen jedoch nicht in ein solch relationales Modell – was die Anwender zum Teil aber nicht davon abhält zu versuchen, ihre Welt mit eher mäßigem Erfolg in das relationale Modell zu zwingen.

Seit einigen Jahren aber gibt es Alternativen zu den klassischen Architekturen. Als Stichwörter fallen dabei immer wieder Termini wie Hadoop, NoSQL oder auch Map-Reduce. Letztlich aber ist es der Begriff Big-Data, der diese Stichwörter zusammenführt. Ein Begriff, der leider in den vergangenen Monaten zu oft für alles herhalten musste, was mit Daten zu tun hatte. Die hier zu beobachtenden Entwicklungen sind dabei vielfach eng miteinander verwoben, auch wenn sie nicht explizit voneinander abhängen. Getrieben werden sie in allen Bereichen vor allem von großen Internetkonzernen wie Google, Facebook, Yahoo oder Amazon, die maßgeblich an den grundlegenden Konzepten mitgearbeitet haben.

Über den Begriff Big-Data und seine Definition ist in der Fachwelt viel und sehr kontrovers diskutiert worden. Unabhängig davon, ob man von den klassischen drei V (*Volume*, *Variaty* und *Velocity*) ausgeht oder noch weitere (wie etwa *Value* oder *Veracity*) einbezieht, lässt sich feststellen, dass es sich bei Big-Data um zumeist große, polystrukturierte Datenmengen

handelt, deren Verarbeitung mit konventionellen Mitteln, wie etwa den traditionellen relationalen Datenbanken, kaum oder gar nicht mehr möglich ist.



Meilensteine in der Big-Data-Entwicklung

Die *Hadoopisierung* in den Unternehmen schreitet darüber hinaus zunehmend voran. Dabei geht es letztlich um die Verarbeitung von Big-Data mit Hadoop oder ähnlichen Frameworks (z. B. *Disco* oder *BashReduce*) in einem verteilten System. Die hier zum Einsatz kommenden Konzepte umfassen etwa das sogenannte Map-Reduce-Verfahren, welches die Verarbeitung großer Datenmengen in einem Cluster ermöglicht, und für die Datenhaltung werden in der Regel NoSQL-Datenbanken eingesetzt. Der Grund dafür, dass im Geflecht von Big-Data und Hadoop NoSQL-Datenbanken eingesetzt werden, ist unter anderem in der Skalierbarkeit der Systeme zu sehen.

Das vorliegende Buch zeigt nun, wie man eine entsprechende Architektur tatsächlich aufbauen kann. Und dies nicht nur theoretisch, sondern sehr praxisnah, stets mit konkreten Beispielen und bei Bedarf auch mit passenden Code-Schnipseln. Die entsprechenden theoretischen Grundlagen und Konzepte werden vorgestellt und in einem nächsten Schritt gleich zur Anwendung gebracht. Angefangen bei einem entsprechenden Hadoop-Cluster mit den typischen in der Praxis häufig zu findenden Tools über die verwendete NoSQL-Datenbank bis hin zur Visualisierung der Ergebnisse. Dabei werden auch die unterschiedlichen Aspekte, die Big-Data ausmachen, behandelt. So wird nicht nur der Größenaspekt allein adressiert, sondern darüber hinaus auch, wie man etwa das Problem der Unstrukturiertheit der Daten angehen kann. Hier werden Methoden aus dem Data- und Text-Mining ebenso vorgestellt wie neue Aspekte, beispielsweise das Natural Language Processing.

Nach der Lektüre bleibt für den Leser nur noch die Frage offen, ob er tatsächlich *Big-Data-Schmerzen* hat, wie Pavlo Baron (2012) es nennt, und die vorgestellten Konzepte einsetzen möchte. Wie diese konkret umzusetzen sind, zeigt Jonas Freiknechts Buch anschaulich und erfreulich praxisorientiert.

Professor Dr. Uwe Haneke, Hochschule Karlsruhe

2 Big-Data

Zu Beginn wurde bereits eine kurze Einführung in das Thema Big-Data gegeben. In diesem Kapitel möchte ich diese noch etwas vertiefen und verschiedene Definitionen vorstellen, die diesbezüglich in den letzten Jahren entstanden sind. Bitte lassen Sie mich zuvor noch einmal den Gedanken der Präsenz von Big-Data in den verschiedenen Industrien aufgreifen, die eben in der Einleitung bereits in den Raum gestellt wurde. Bild 2.1 zeigt eine Statistik, die eine Umfrage von Gartner (Kart, 2012) aus dem Jahre 2012 verbildlicht. In dieser wurde evaluiert, welche Industrien sich mit dem Thema Big-Data in welchem Ausmaß auseinandersetzen.

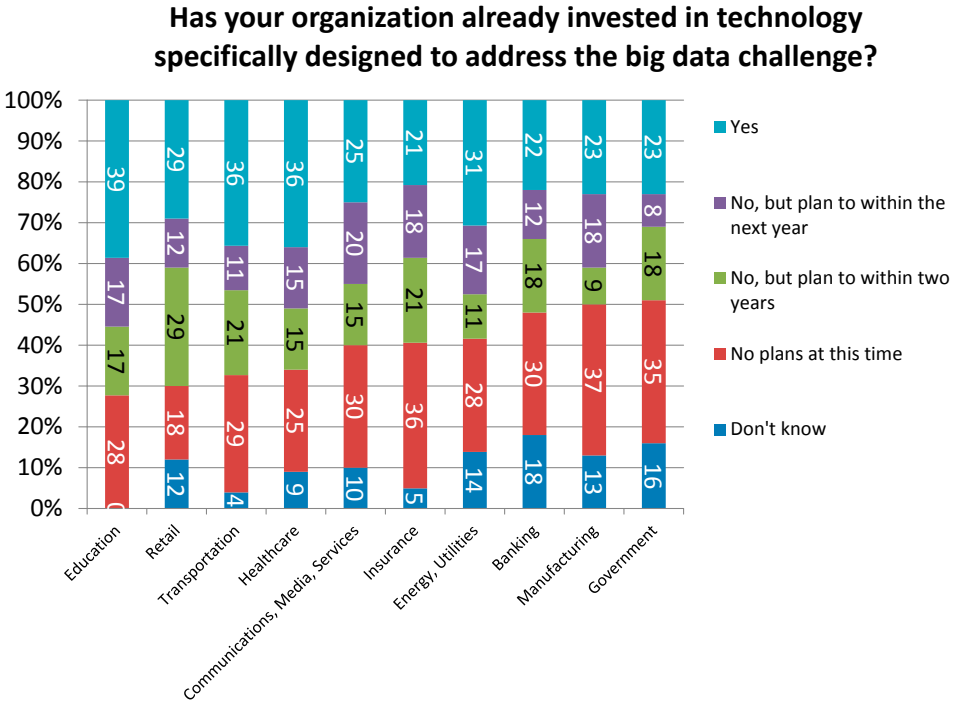


Bild 2.1 Big-Data regt das Interesse sämtlicher Industrien an.

Das wenig verblüffende Ergebnis ist, dass sich alle bereits damit beschäftigt haben, allen voran die Sparte *Education*. Zwar kann man davon ausgehen, dass es sich häufig noch um einfache Evaluationen handelt, die bestimmen sollen, ob der Einsatz passender Technologien im Unternehmen infrage kommt und notwendig ist bzw. sein wird, jedoch ist eine gewisse Grundeinstellung zu erkennen, die aussagt, dass Big-Data als wichtige Neuerung wahr- und ernst genommen wird. Erfolgte noch keine intensivere Auseinandersetzung, so ist sie jedoch für die Hälfte aller Unternehmen und Einrichtungen in den nächsten zwei Jahren geplant. Natürlich bleibt bei vielen eine Restunsicherheit, die dadurch hervorgerufen wird, dass man nicht weiß, ob Big-Data nun ein kurzlebiger Hype bleibt oder nicht. Dem sei jedoch entgegenzusetzen, dass uns die Diskussion rund um große Datenmengen nun schon einige Jahre beschäftigt und immer noch zu der Veröffentlichung neuer Artikel, Bücher und neuer Software führt. Wir können also bald aufhören, von einem bloßen Trend zu sprechen, und beginnen, Big-Data als integralen Bestandteil der heutigen IT zu sehen.

Bevor es in den folgenden Abschnitten an die Begriffsdefinitionen von Big-Data geht, soll zuvor ein kurzer Einblick in dessen historische Entstehung gegeben und skizziert werden, wo die Big-Data-Bewegung ihren Ursprung hat.

2.1 Historische Entstehung

Einen konkreten Ursprung für den Begriff Big-Data gibt es nicht. Zwar ist bekannt (Press, 2012), dass der Begriff selbst von Michael Cox und David Ellsworth verhältnismäßig früh (im Jahr 1997) öffentlich in einem Paper genannt wurde (Cox et al., 1997), jedoch wird auf *Wikipedia* der Ursprung des Begriffs kontrovers diskutiert und andere Quellen werden genannt (Mashey), die behaupten, dass John Mashey den Terminus *Big* bereits 1994 im Zusammenhang mit Datenmengen verwendete. Dies geschah jedoch vorwiegend als stilistisches Mittel, um die Komplexität einiger Teilbereiche der IT hervorzuheben. Darunter waren auch *Big Bandwidth*, *Big Physics*, *Big Latency* und eben auch *Big Data*, sodass der Begriff zwar gleich dem heutigen verwendet wurde, die Bedeutung jedoch nicht oder nur in Teilen übereinstimmte.

Im November 2009 entstand der erste Wikipedia-Artikel zu Big-Data und wurde vom Benutzer John Blackburn prompt wieder gelöscht. Die Begründung war folgende:

> *"Delete as per nom – it is simply a combination of big and data, dictionary words which have no place here. I'm not even sure it's a neologism, and even if it was it doesn't need an article."* (John Blackburne)

Nach einigen Diskussionen, die besagten, dass *Big Band* oder *Big Bang* ebenso gelöscht werden müssten, wenn Big-Data doch auch nur ein zusammengesetzter Begriff aus einem Adjektiv und einem Substantiv sei, wurde der Artikel letztendlich zugelassen und sogar später (2011) in den *Gartner Hype Cycle* aufgenommen (McBurney, 2012).

Bild 2.2 Frühe Nennung des Begriffs Big-Data auf der IEEE Supercomputing Conference 1996 in Pittsburgh; Foto: Michael Woodacre

Besonders interessant an dieser Betrachtung ist, dass Doug Cutting bereits 2006 unter der Schirmherrschaft von Yahoo an Hadoop arbeitete (Cutting). Die Definition des Begriffs Big-Data hat also von der ersten Open-Source-Implementierung von Hadoop in 2006 bis hin zum Wikipedia-Artikel fünf Jahre benötigt, um zu reifen und um Big-Data als eigenes Teilgebiet der Informatik anzuerkennen.

In den letzten zwei Jahren wurde der Begriff von den bekannten IT-Häusern wie IBM, SAP oder Oracle aufgenommen und wird heutzutage in Zusammenhang mit Speicher, Business-Intelligence (BI), Data-Warehousing (von Data-Warehouse, DWH) und Data Analytics verwendet. *Visible Technologies* diagnostizierte einen Anstieg der Zahl der Begriffsnennungen in Social Media Channels von 2009 bis 2012 um 1211 % (Press, 2012). Es ist also kaum mehr möglich, vor dem Begriff die Augen zu verschließen. Wie aber wird nun Big-Data eigentlich definiert?

2.2 Big-Data – ein passender Begriff?

In diesem Abschnitt sollen zunächst verschiedene Quellen herangezogen werden, die Big-Data aus verschiedenen Perspektiven betrachten. Ohne in die Tiefe zu gehen, lässt sich Big-Data so definieren: Big-Data sind Datenmengen, die zu groß für traditionelle Datenbanksysteme sind, eine hohe Halbwertszeit besitzen und in ihrer Form nicht den Richtlinien herkömmlicher Datenbanksysteme entsprechen (Dumbill, 2012). Ziel ist es nun, diese Daten dennoch zu speichern und zu verarbeiten, sodass aus ihnen zeitnah wertvolle, neue Informationen gewonnen werden können. Diese neu gewonnenen Informationen können etwa passende Produktempfehlungen in E-Commerce-Lösungen sein, empfohlene Kontakte

in sozialen Netzwerken oder Artikelvorschläge auf Nachrichtenseiten. Sind diese Daten nun wirklich *big* gemäß der oben gegebenen Definition?

Auf den ersten Blick wirkt es tatsächlich nicht so, denn Freundschaften können als klassische n:m-Relation gespeichert werden, Artikelvorschläge werden anhand der Tags des gelesenen Artikels erstellt und Produktempfehlungen entstehen auf Basis der Produktkategorie der vorher betrachteten Artikel. Zieht man aber nun vom Benutzer generierte Inhalte (*User-Generated Content*) hinzu, trifft man auf Inhalte wie:

- Rezensionen und Bewertungsschreiben für Güter und Dienstleistungen
- Foren-Posts und Kommentare
- Pinnwandeinträge
- Blogeinträge
- (Wissenschaftliche) Artikel
- Tweets

Diese enthalten oft subjektive Meinungen von Benutzern und Konsumenten und sind dementsprechend wertvoll. Allerdings müssen sie vorher analysiert werden, um sie für Maschinen lesbar zu machen und somit dem Datenhalter einen Mehrwert zu liefern. Diese Datenbeschaffenheit geht mit einem Datenumfang einher, der die traditionelle Datenverarbeitung vor eine scheinbar unlösbare Aufgabe stellt. So sammeln Twitter beispielsweise 8 Terabyte an Daten am Tag, Facebook 500 Terabyte und Google verarbeitet pro Tag etwa 20 Petabyte an User-Generated Content.

2.2.1 Die drei V

Die Frage nach der am meisten verbreiteten Definition von Big-Data lässt sich wohl am ehesten durch die drei V beantworten, die der Anbieter von Marktanalysen Gartner 2001 einführte (Lancy, 2001). Zwar wurde hier noch kein Zusammenhang zu Big-Data hergestellt, jedoch erkannte man bereits die zukünftigen Herausforderungen der Datenverarbeitung (hier im E-Commerce-Sektor), die den Big-Data-Begriff später prägen sollten. Es werden nun die einzelnen V genauer beschrieben.

Volume

Hinter *Volume* (deutsch: Volumen) verbirgt sich der Begriff, den man auf den ersten Blick am wahrscheinlichsten erwartet. Wo früher 500 Megabyte große Festplatten eine Sensation waren, sind es heute die Gigabyte großen USB-Sticks in Fingernagelgröße. Dieses Beispiel mag zwar etwas einfach erscheinen, spiegelt aber dennoch die rasante Entwicklung der Speicherhardware wider. Mit den Möglichkeiten steigen nämlich auch die Anforderungen. Wo vor 15 Jahren 20 Gigabyte für einen PC völlig ausreichend waren, ist heute ein bis zwei Terrabyte Speicher der Stand der Dinge. Dieses Wachstum fasst eine Studie von IDC (Gantz et al., 2011) passend in Worte:

> *"Like our physical universe, the digital universe is something to behold — 1.8 trillion gigabytes in 500 quadrillion 'files' — and more than doubling every two years."*

Projiziert man diese Behauptung nun auf die Anforderungen an Suchmaschinen, Reporting, Online-Marketing und irgendwann auch Suchfunktionen auf dem lokalen Computer, so wird schnell klar, dass die Art, die Daten zu speichern und zu verarbeiten, überdacht werden muss. Wann man nun Volumina als groß bezeichnet, ist immer stark von der verfügbaren Hardware abhängig. Öffnet man eine 100 Megabyte große CSV-Datei (*Comma-Separated Value*) in Microsofts Excel, so benötigt der Ladevorgang bereits einige Minuten auf einem durchschnittlich ausgestatteten Desktop-PC. Auf einem mobilen Gerät oder sehr simpler Hardware (*RaspberryPI*, *FritzBox*, *Aruduino* …) sind solche Vorgänge überhaupt nicht praktikabel. Klar ist, dass hier die Leistung der CPU (Central Processing Unit) und des RAM (Random Access Memory) den Flaschenhals darstellt. Der Speicher an sich reicht in den meisten Fällen für 100 Megabyte ohne Weiteres aus. Es ist also ersichtlich, dass bei großen Datenmengen die Verarbeitung das *Big* ausmacht (Baron, 2013), nicht der Speicherbedarf (siehe Abschnitt 2.2.3).

Velocity

Das zweite V beschreibt die Geschwindigkeit der Daten. Damit wird einmal deren Aktualität angesprochen sowie die Geschwindigkeit der Verarbeitung. Ein passendes Beispiel sind etwa Twitter-Meldungen oder Blog-Posts, die zu Zeiten von Wahlkämpfen verfasst werden. Dabei stellt das Internet ein starkes, aber schwer zu kontrollierendes Werkzeug dar. Bekommt ein Politiker beispielsweise eine schlechte Presse, so muss das bereitstehende Marketing-Team entsprechend zügig reagieren und diese durch korrigierende Meldungen relativieren. Die Betonung liegt auf *zügig*, denn ein einziges Gerücht zur falschen Zeit kann zu ungewollten Auswirkungen führen. So sorgte ein am 23. April 2013 geschriebener Tweet über einen Bombenanschlag auf das Weiße Haus für einen Börsen-Crash an der Wall Street. Bleibt man dem Beispiel der Börse treu, finden sich schnell noch weitere Szenarien, in denen die Geschwindigkeit der Erfassung von Daten eine große Rolle spielt. So sind zum Beispiel Ankündigungen über Firmenübernahmen Gold wert, wenn diese direkt nach der Bekanntgabe verifiziert werden können. Für beide Szenarien gilt: Je schneller die Auswertung stattfindet, desto höher ist der Wert der Information.

Variety

Hier kommt nun die bereits häufig angesprochene Abwesenheit von festen Strukturen und Normalisierungen zur Sprache. Das sicherlich beste Beispiel, um die Datenvielfalt, mit der wir es zu tun haben, zu beschreiben, ist das Internet, das – außer vielleicht den Wiki-Seiten – keine feste Struktur vorweisen kann, aber doch einige Ähnlichkeiten aufweist, die eine maschinelle Verarbeitung ermöglichen. So können etwa bei der Analyse des HTML-Codes die Titel-Tags, z. B. *<h1>*, durchsucht werden, um eine thematische Einordnung des Inhalts vorzunehmen. Jedoch sind HTML-Seiten nicht die einzigen Daten, die verarbeitet werden. IBM befragte Mitte 2012 (Schroeck et al., 2012) einige Unternehmen, die an einer Big-Data-Initiative teilnahmen, welche Quellen sie für ihre Analysen verwenden.

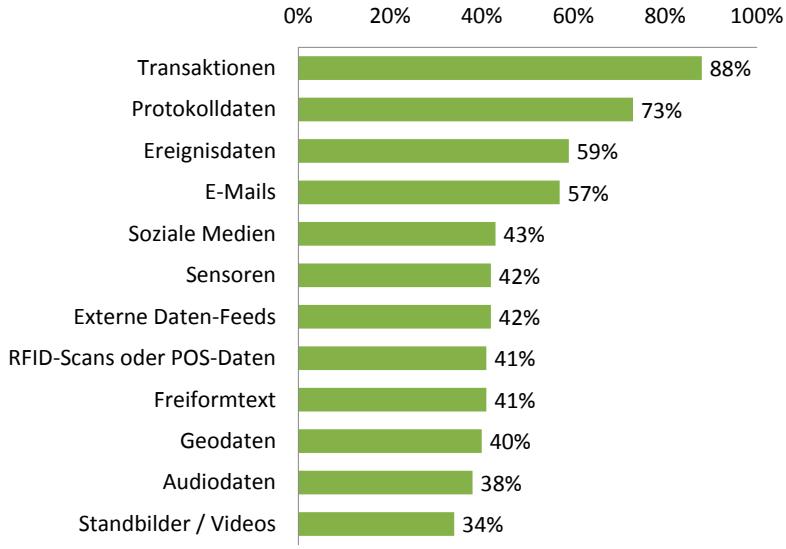


Bild 2.3 Quellen für Big-Data-Analysen auf Basis einer Umfrage von IBM im Jahr 2012

Transaktionsdaten beziehen sich dabei beispielsweise auf den klassischen Börsenhandel, in dem jeder Ver- und Einkauf gespeichert wird. Protokolldaten sind unter anderem Serverlogs, die entsprechend dem Log-Level der jeweiligen Architektur sehr groß ausfallen können und häufig dazu genutzt werden, um Klickpfade durch komplexere Anwendungen zu ermitteln und Benutzer möglichst lange im System zu halten bzw. zu einer bestimmten Aktion (Kauf, Registrierung, Empfehlung) zu bewegen. Ereignisdaten werden etwa in der Automobilindustrie protokolliert, in der Fahrzeugteile produziert, an die Logistik übergeben und verfrachtet werden. Das Feld dieser Daten ist jedoch weder auf die Automobilindustrie noch auf die Produktherstellung im Allgemeinen beschränkt, sondern beschäftigt sich in der IT mit allen Systemen, die bestimmte Ereignisse aufzeichnen und zur Auswertung bereitstellen. So kann etwa die Ausschussquote einer Produktreihe überprüft und ggf. fehlerhafte Teile im gesamten Produktionszyklus, über mehrere Hersteller hinweg, ausfindig gemacht werden. Platz vier auf der Liste belegen E-Mails, die von Mail-Service-Anbietern gescannt, auf Muster von Malware oder Spam durchsucht und für gezielte Produktvorschläge für den jeweiligen Empfänger hin untersucht werden. Ob und wie dieses Vorgehen mit den Datenschutzbestimmungen des jeweiligen Landes und der Moral der Betreiber vereinbar ist, sei einmal dahingestellt. Dass soziale Medien, externe Daten-Feeds und Freitextformen noch keine so starke Beachtung finden, mag daran liegen, dass Restriktionen für die Sichtbarkeit von Daten, etwa in Facebook, die Akquise erschweren oder aber dass die Szenarien für eine Nutzung der Daten noch nicht gefunden sind, um einen produktiven Mehrwert daraus zu ziehen.

Da nun einige Quellen für Big-Data vorgestellt wurden, lässt sich auch gleich auf die Vielgestalt der Formatierung eingehen. Neben klassischen, unformatierten Texten kommen häufig JSON (*JavaScript Object Notation*), XML (*Extensible Markup Language*), HTML- oder sogar Byte-Code vor. Gerade wenn man an den Aspekt der Visualisierung denkt, ist es wichtig, Relationen zwischen einzelnen Datensätzen herzustellen, um diese in Abhängigkeit voneinander zu präsentieren. Was in relationalen Datenbanken über simple Queries erreicht

werden kann, bedarf bei Plain-Text-Analysen eines erheblichen Aufwands (siehe Kapitel 8). Viel früher trifft man jedoch bei der Analyse auf die Herausforderung, die gewünschte Information aus jedem einzelnen der vielen Formate herauszufiltern. Des Weiteren gilt zu bedenken, dass sich Formate im Laufe der Zeit auch ändern können. Gerade bei der Auswertung fremder, externer Datenquellen erfolgt meist keine Benachrichtigung über eine Anpassung der Datenstruktur seitens des Datenhalters. Hier ist es wichtig, Auswertungen entsprechend zu überwachen, um Abweichungen frühzeitig festzustellen.

Ist es denn nun gerechtfertigt, von unstrukturierten Daten zu reden? Schließlich weisen ja viele Datensätze eine Struktur auf, nur eben keine feste, einheitliche. Als besserer Begriff wäre hier vielleicht polystrukturiert anzuführen, wie es der Analyst Mike Ferguson in seinem Blog beschreibt (Ferguson). Im späteren Verlauf des Buches, wenn wir zu den Eingabeformaten für die diversen Diagramme kommen, wird sich zeigen, dass diese Vielgestalt einen großen Teil der Arbeit eines Datenanalysten ausmacht, denn die Interpretation der Eingangsdaten einer Analyse variiert nur allzu häufig. Zu diesem Umstand gesellen sich ebenso Fehler in Daten, die schon vor dem Big-Data-Hype bekannt waren. Nathan Yau (Yau, 2010) benennt sechs davon wie folgt:

- Fehlende Werte
- Falsche Beschriftung
- Inkonsistenz
- Tippfehler
- Werte ohne Kontext
- Verteilte Datensätze (über mehrere Quellen hinweg)

Die Komplexität einer Verarbeitung von Daten, die in mehreren, ggf. unbekannten, Strukturen vor- und diesen sechs Umständen unterliegen, erfordert also einen erheblichen Mehraufwand gegenüber der Aufbereitung von normalisierten Daten, z. B. aus einer relationalen Datenbank.

2.2.2 Das vierte V – Veracity

IBM führt ein viertes V ein, das die Richtigkeit und die Echtheit von Daten beschreibt (Zikopoulos et al., 2013). Zwar steigt die Menge an zur Verfügung stehenden Daten nachweislich an, jedoch werden diese häufig durch generierte Inhalte verfälscht, die da sein können:

- Werbung und Spam, die eine einseitige Sicht auf Personen, Produkte oder Vorkommnisse wiedergeben.
- Per Automatismus übersetzte Texte, die häufig grammatikalische, sprachliche und inhaltliche Fehler aufweisen.
- Veraltete oder falsch kategorisierte Suchergebnisse oder Forenindizes.
- Gezielte Falschaussagen oder Fehlinformationen.

Das beste Beispiel sind die klassischen Falschmeldungen, die sich im Internet manchmal in wenigen Minuten verbreiten. So streute etwa ein junger Brite am 26. Februar 2012 das Gerücht, dass der Schauspieler Rowan Atkinson gestorben sei (Gardner, 2012). In nur drei

Stunden wurde das Gerücht so schnell verteilt, dass sogar auf Wikipedia der Todestag des Schauspielers eingetragen wurde. Einmal mehr zeigt sich hier die Notwendigkeit, die gesammelten Daten vor der Nutzung zu verifizieren und auszusortieren.

2.2.3 Der Verarbeitungsaufwand ist big

Ein weiterer interessanter Ansatz, den Aufwand der Verarbeitung großer Datenmenge als *big* zu sehen, liefert der Autor Pavlo Baron:

> *„Ich hatte z. B. einen Fall, bei dem es um lächerliche Datenmengen ging, die problemlos auf einen USB-Stick gepasst hätten. Man erwartete allerdings simultane Zugriffszahlen im zweifachen Millionenbereich pro Sekunde. […] das ist definitiv Big […].“* (Baron, 2013)

Es ist also aus dieser Perspektive nicht die bloße Größe der Daten, sondern die Komplexität der Aufbereitung und der Informationsgewinnung. Ein gutes Beispiel sind dafür etwa Video-Streams in Kaufhäusern, die das Kaufverhalten von Kunden auswerten sollen. Auch wenn eine einstündige Aufnahme lediglich ein paar Hundert Megabyte groß ist, ist die Schwierigkeit der Implementierung und des Trainierens von situationserkennenden Algorithmen sehr hoch und steht beispielsweise im Gegensatz zu einem einfachen Algorithmus, der lediglich alle *<h1>-Tags* aus einigen Millionen HTML-Seiten auslesen muss. Sind die notwendigen Daten extrahiert, müssen ggf. noch Beziehungen zu anderen Datensätzen hergestellt werden, etwa über einen übereinstimmenden Datumswert, Quellenübereinstimmungen oder, im Optimalfall, über vorliegende IDs. Relationale Daten hingegen verfügen über Schlüsselattribute, die eine Zuordnung von Datensätzen erheblich vereinfachen.

Roger Magoulas von O'Reilly Media gibt eine weitere sehr schöne Definition für Big-Data.

> *„Big-Data ist, wenn die Daten selbst Teil des Problems werden.“*

Dieses Zitat passt besonders gut, da Magoulas gar nicht erst versucht, eine Größendefinition zu geben, sondern einfach sagt, dass man, wenn die Datenmenge für *aktuelle* Verarbeitungsmethoden zu umfangreich wird, von Big-Data spricht. Wenn ein Kunde zu Ihnen kommt und Sie fragt, ob seine Daten nun *big* sind oder nicht, dann antworten Sie doch einfach mit einem der beiden hier gegebenen Zitate.

2.2.4 Sicht der Industrien auf Big-Data

IBM befragte in der bereits in Abschnitt 2.2.1 erwähnten Studie (Schroeck et al., 2012) mehrere Unternehmen nach deren Definition von Big-Data. Eine Auswertung nach Schlagworten bestätigte die in den vorigen Abschnitten gegebene Sicht auf den neuen Trend weitestgehend. Auffällig ist, dass der Größenbegriff nicht immer im Sinne von Datengröße verwendet wird. Stattdessen wird etwa in dieser Studie das Schlagwort *Größere Bandbreite an Informationen* verwendet, das sowohl auf große Datenmengen als auch auf mehr oder vielfältigere Informationsquellen hindeutet. 16 Prozent der befragten Unternehmen stellen neue Datenarten und Analysemethoden in den Vordergrund, was wieder für das *big* im Sinne des Verarbeitungsaufwands hindeutet (siehe Abschnitt 2.2.3).

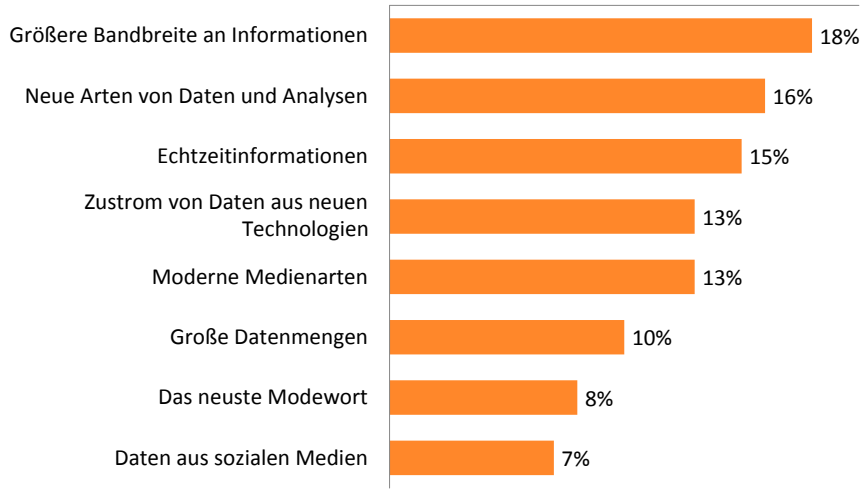


Bild 2.4 Definition von Big-Data (Schroeck et al., 2012)

Überraschend ist, dass Big-Data von einer Mehrheit im Jahre 2012 mehr als Modewort gesehen wurde, als dass es Daten bezeichnet, die aus sozialen Medien stammen. Zwar weist der Begriff *Big-Data* den typischen Buzz-Word-Charakter auf, jedoch ist hier das Ranking entscheidend, dass Daten aus Social-Network-Daten dem Modewortcharakter hintanstellt. Auch diese Studie bestätigt durch die verschiedenen Antworten, dass eine einheitliche Sichtweise auf Definition, Verwendung und Ausprägung des Begriffs noch nicht vorherrscht. Nichtsdestotrotz zeigt sich, dass bestimmte Themengebiete bei einem Eingrenzungsversuch immer wieder auftauchen.

2.3 Eingliederung in BI und Data-Mining

Um die Begriffe BI und Data-Mining in Relation zu Big-Data setzen zu können, gilt es, diese im Vorfeld zu definieren. Kemper, Mehanna & Unger bezeichnen BI recht prägnant als Filter, der Daten in strukturierte Information umwandle (Kemper et al., 2010). Gartner hingegen konstatiert etwas ausführlicher, dass BI ein Überbegriff für Anwendungen, Infrastruktur, Werkzeuge und Best Practices sei, die den Zugriff auf und die Analyse von Informationen ermöglichen, um Entscheidungsfindung und Performance zu erhöhen (Gartner). Hält man sich nun strikt an die Definitionen, besteht der Unterschied zwischen BI und Big-Data darin, dass BI sich auf bereits vorliegende Informationen bezieht, die dazu noch strukturiert sind und sich auf einen eindeutigen Kontext beziehen. Das Ziel von BI und der Big-Data-Bewegung ist jedoch dasselbe, nämlich aus vorhandenen Daten neue Erkenntnisse zu gewinnen, die der Entscheidungsfindung bei vorher definierten Fragestellungen dienen. BI ist jedoch mittlerweile mehr als diese einfache Begriffsdefinition. Es hat sich in den letzten Jahren zu einem festen Prozess samt einem Set aus technischen Werkzeugen entwickelt, um das Berichtswesen in Unternehmen zu automatisieren. Dazu gehören die Datenaufbe-

reitung, die Datenspeicherung in DWHs sowie deren Darstellung aus verschiedenen Perspektiven.

Welche Techniken, Methoden und Arbeitsschritte werden aber nun angewandt, um Informationen aus vorliegenden Daten zu extrahieren? Die Antwort darauf gibt der sogenannte KDD-Prozess (*Knowledge Discovery in Databases*).

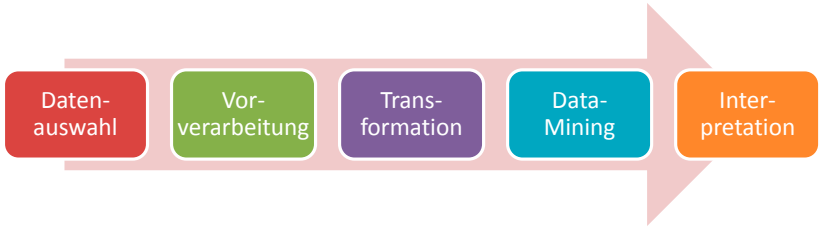


Bild 2.5 Der KDD-Prozess nach (Kononenko et al., 2007)

Der (iterative und interaktive) KDD-Prozess hat das Ziel, gültige, neue, nützliche und verständliche Muster in vorhandenen Daten zu erkennen (Fayyad et al., 1996). Wirft man nun einen Blick auf den vierten Schritt des in Bild 2.5 illustrierten Ablaufs, so ist zu erkennen, dass Data-Mining einen Teil des KDD-Prozesses darstellt. Dieser nimmt gesäuberte, korrigierte und transformierte Daten entgegen, extrahiert aus diesen Muster und Zusammenhänge und gibt diese zur Interpretation frei. Quellen müssen, anders als der Begriff KDD vermuten lässt, nicht zwingend Datenbanken sein, sondern können auch als simple Datensätze gesehen werden, z. B. als Flat-Files, CSV, XML oder Dateisystemstrukturen. Wichtig ist, dass diese bereits im Vorfeld aufbereitet wurden. Zu dieser Aufbereitung (*Preprocessing*) gehören:

- Formatanpassungen (z. B. Datums- und Zahlenformate)
- Korrigieren von Ausreißern (Messfehler, Verarbeitungsfehler, bewusste Falschangabe)
- Auffüllen dünn besetzter Daten

HINWEIS: Ohne nun zu viel verraten zu wollen, möchte ich hier kurz darauf hinweisen, den KDD-Prozess im Hinterkopf zu behalten, wenn wir in Kapitel 8 aus allen behandelten Themen unsere kleine Reporting-Anwendung konstruieren. Sie werden bei deren Entwicklung und Bedienung einige Parallelen feststellen. Ebenso lohnt sich ein Vergleich des KDD-Prozesses mit neuartigen verteilten Verarbeitungsframeworks wie Apache Spark (siehe Abschnitt 3.19), in denen sich einzelne Prozessschritte eins zu eins wiederfinden lassen.

Stellt man nun die drei Begriffserklärungen BI, Data-Mining und Big-Data einander gegenüber, so erkennt man schnell einige Gemeinsamkeiten sowie Unterschiede.

Einerseits werden die Daten bei der Big-Data-Verarbeitung vorher nicht aufbereitet und in einem eigens dafür eingerichteten System, wie beim Data-Mining dem DWH, abgelegt. Das macht den Prozess performanter und schlanker, erfordert allerdings wesentlich flexiblere, aufwendigere Algorithmen zur Mustererkennung. Ein weiterer Nachteil der Big-Data-Verarbeitung gegenüber dem Data-Mining ist, dass Daten bei Letzterem strukturiert in einem

DWH abgelegt sind und somit auch aus anderen Perspektiven betrachtet werden können (Stichwort *Online Analytical Processing Cube*). Soll also in einer Grafik der Gewinn eines Unternehmens samt aller Töchterfirmen angezeigt werden anstatt wie bisher ohne, so müssen im Reporting-Werkzeug lediglich vom Benutzer die bereits ermittelten Gewinne der Töchterfirmen dem Gesamtgewinn hinzuaddiert werden. Im Falle von einer Umsetzung auf Basis von Hadoop, müsste die Auswertung programmatisch angepasst werden, sie ist also weniger flexibel, was Änderungen in Abfragen angeht[1]. Die Verarbeitung der Daten zeigt deutliche Übereinstimmungen bei Big-Data und Data-Mining. Zwar nutzt Big-Data den Map-Reduce-Algorithmus zum Extrahieren der Daten, jedoch werden für Klassifizierung, Clusteranalyse etc. dieselben Algorithmen (*K-Means*, *CLARA*) verwendet. Die Apache Foundation hat den Vorteil der verteilten Verarbeitung für Data-Mining bereits aufgegriffen und implementiert eine entsprechende Lösung mit Namen Apache Mahout (siehe Abschnitt 4.8).

Zu guter Letzt sollen die Resultate der drei Vorgänge betrachtet werden. Diese weisen auf den ersten Blick starke strukturelle Unterschiede auf. Der Benutzer eines BI-Reporting-Werkzeugs wünscht sich ein Diagramm, das einige Zahlenmengen so darstellt, dass er darin Informationen erkennt, die ihm vorher nicht zugänglich waren. Big-Data wird im Allgemeinen durch den Map-Reduce-Prozess auf Schlüssel-Wert-Paare reduziert und danach auf dem traditionellen Weg (Reporting-Werkzeuge, Geschäftsanwendungen …) weiterverarbeitet. Neue, flexiblere Frameworks wie YARN oder Spark sind jedoch stark im Kommen. DM ermittelt laut der Beschreibung des KDD-Prozesses Muster und Zusammenhänge, die dann als Rohdaten ebenso kontextabhängig aufbereitet werden müssen. So liefern diese aggregierte Einkäufe einer bestimmten Artikelgruppe über einen Zeitraum t sowie die Seitenaufrufe der Artikel *xyz* dieser Gruppe. Ein Algorithmus erkennt hier etwa einen Zusammenhang von Seitenaufrufen und Käufen über eine positive Kovarianz. Um dem Benutzer diese Information zugänglich zu machen, müssen diese in irgendeiner Form aufbereitet werden, z. B. als Tabelle, als Information in Textform oder als Diagramm. Auffällig ist, dass sich alle drei Ergebnisse in der fachlichen Betrachtung ähneln, denn jeder Einzelne hat das Ziel, neu gewonnene Informationen zu liefern, die der Entscheidungsfindung dienen. Der technische und fachliche Weg dorthin unterscheidet sich jedoch mitunter sehr.

Bild 2.6 (auf der nächsten Seite) zeigt noch einmal die in den Begriffsdefinitionen gefundenen Übereinstimmungen der drei Themenbereiche BI, Big-Data und Data-Mining. Diese Betrachtung war nun sehr theoretisch. In der Praxis ergänzen sich Big-Data, DWH und BI natürlich. So liefern Big-Data-Technologien die Daten für das Reporting und stellt Schnittstellen zur Verfügung, um die großen Datenmengen ähnlich einem DWH abzurufen. Die bisherige DWH-Infrastruktur wird mit großer Wahrscheinlichkeit noch viele Jahre parallel zu den neuartigen Methoden für die Verarbeitung von Big-Data existieren.

[1] Ausnahmen gibt es auch hier. So können SQL-ähnliche Abfragen über Apache Hive auch dynamisch zur Laufzeit angepasst werden. Dafür ist es allerdings notwendig, dass die Daten in einem festen Schema vorliegen, was jedoch nicht immer dem Normalfall entspricht.

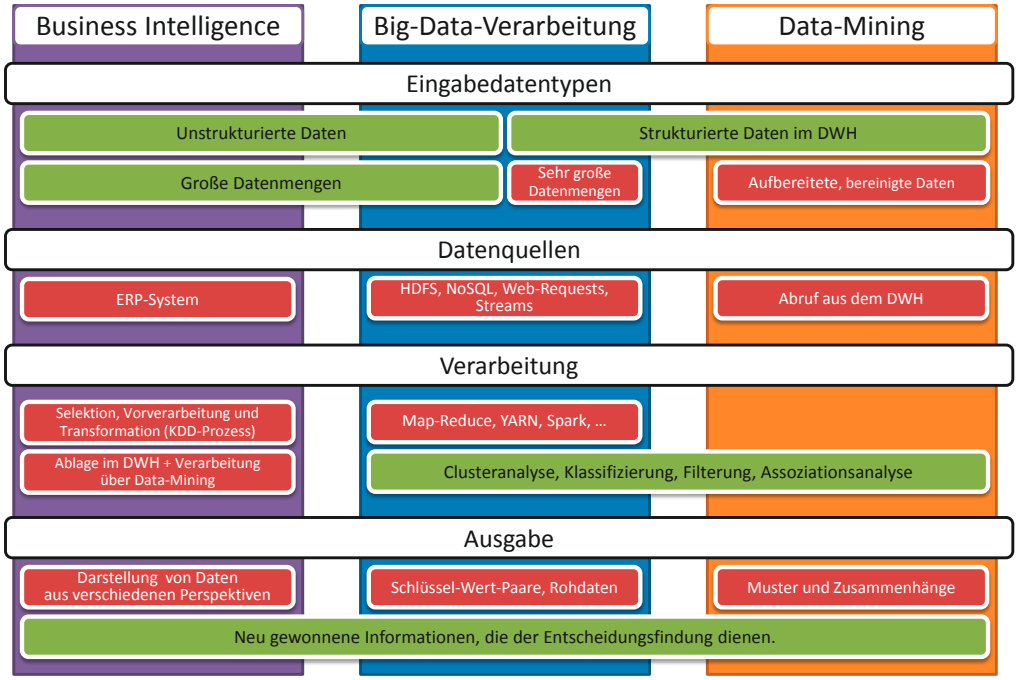


Bild 2.6 Definitionsvergleich von BI, Big-Data und Data-Mining

PRAXISTIPP: Der Trend bei der Bereitstellung großer Datenmengen und traditioneller Daten im DHW oder in relationalen Datenbanken geht dahin, dass alle Datenquellen über ein einheitliches Interface angeboten werden. So werden Abfragen nur noch an eine Datenbank-Engine gerichtet, die die Queries an die entsprechende Stelle weiterleitet. Die eigenständigen Systeme dahinter verlieren bei einem solchen sogenannten föderierten Informationssystem (*Federated Database System*) weder ihre Selbstständigkeit, noch müssen deren Daten zwischen den verschiedenen Systemen ausgetauscht werden.

Index

D

E

F

G

H

N

O

P

Q

R

S

T

U

V

W

X

Y

Z