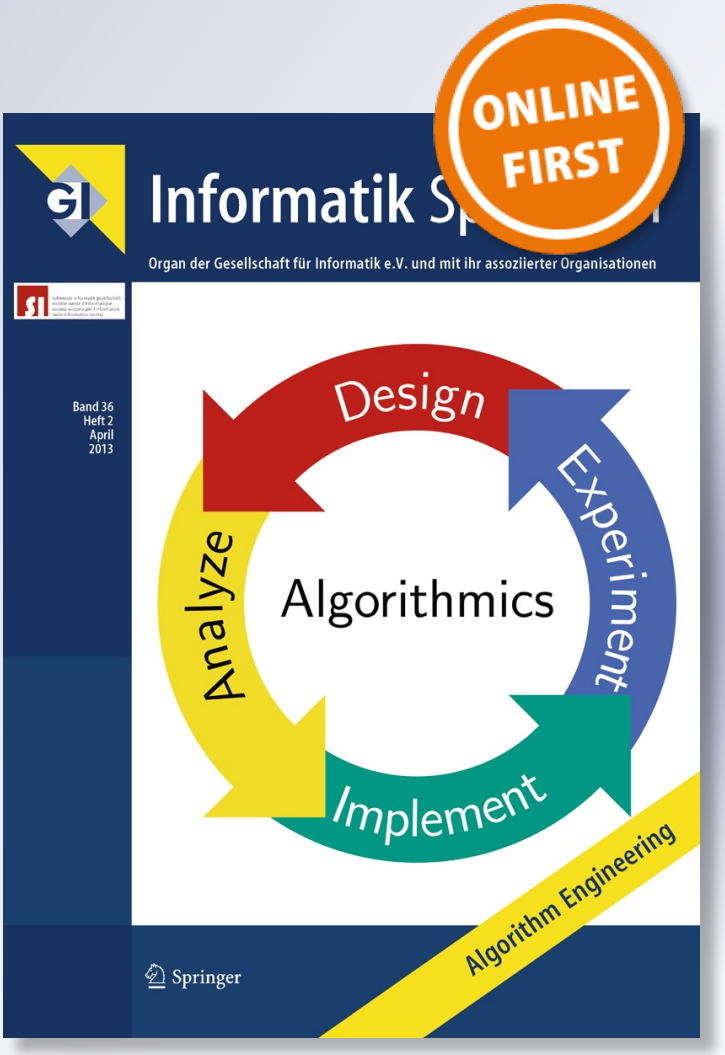
Big Data

Dominik Klein, Phuoc Tran-Gia & Matthias Hartmann

Informatik-Spektrum
Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen

ISSN 0170-6012

Informatik Spektrum
DOI 10.1007/s00287-013-0702-3



Springer

Big Data

Dominik Klein · Phuoc Tran-Gia
Matthias Hartmann

Einleitung

In den letzten Jahrzehnten hat sich das Internet von einem Forschungsnetz zu einem weltumspannenden Kommunikationsnetz entwickelt. Während dieser Zeitspanne sind eine Vielzahl von Diensten und Anwendungen entstanden, die insbesondere auch auf mobilen Endgeräten nicht mehr aus dem alltäglichen Leben wegzudenken sind. Laut einer aktuellen Studie der International Telecommunication Union (ITU) [1] gibt es derzeit ca. sechs Milliarden Mobilfunkteilnehmer, was einem Prozentsatz von 85,7 % der Weltbevölkerung entspricht. Der Prozentsatz der aktiven mobilen Teilnehmer mit Zugang zu Breitband-Netzen beläuft sich immerhin noch auf 15,7 % Prozent der Weltbevölkerung oder 1,1 Mrd. Menschen. Inklusive Haushalten mit kabelgebundenem Internetanschluss nutzen 32,5 % Prozent der Weltbevölkerung oder 2,3 Mrd. Menschen das Internet.

Diese enorme Zahl an Teilnehmern in Kombination mit der stetig wachsenden Anzahl an unterschiedlichen Diensten verursacht ein unvorstellbares Datenaufkommen sowohl an Nutzdaten, als auch an Profildaten und statistischen Daten. Insgesamt wird das in 2012 erzeugte Datenvolumen in einer Studie der International Data Corporation (IDC) [2] auf 2,7 Zettabytes (entspricht 2,7 Mrd. Terabyte) geschätzt, was ein Wachstum von 48 % gegenüber 2011 darstellt. In den erzeugten Daten schlummert ein enormes finanzielles Potenzial, das allerdings oft noch im Verborgenen bleibt. Diese Daten werden auch als das neue Öl des Digitalen Zeitalters bezeichnet: in unverarbeiteter Form sind sie relativ nutzlos, aber wenn es gelingt, durch aufwendige Verfahren und Analysen Struktur in die Daten zu bekommen, dann können sie zur Beantwortung von neuen Fragestellungen genutzt werden und ihr finanzielles Potenzial entfalten. Genau in diesem Umfeld kommt der Begriff „Big Data“ ins Spiel. Dies ist ein abstrakter Oberbegriff für jegliche Art und Anzahl von Daten, die mit traditionellen Datenanalyseverfahren nicht mehr handhabbar sind und deshalb neuer Techniken und Technologien bedürfen.

Big Data ist neben Cloud Computing und Crowdsourcing eine der wichtigsten neuen Technologie-Treiber und wird daher im Aktuellen Schlagwort näher beleuchtet. Zu Beginn gehen wir auf die Definition von Big Data ein und erläutern die Unterschiede zu traditionellen Verfahren. Im Anschluss daran stellen wir zugrundeliegende Technologien vor und geben einen kurzen Überblick über wissenschaftliche Herausforderungen in diesem Bereich.

Begriffsdefinition

Der Ursprung und die erstmalige Verwendung des Begriffes Big Data im aktuellen Kontext sind nicht ganz eindeutig und es werden unterschiedliche Quellen genannt, die den Begriff in der aktuellen

DOI 10.1007/s00287-013-0702-3

Dominik Klein · Phuoc Tran-Gia · Matthias Hartmann
Universität Würzburg, Institut für Informatik,
Lehrstuhl für Kommunikationsnetze,
Am Hubland, 97074 Würzburg
E-Mail: dominik.klein@informatik.uni-wuerzburg.de

*Vorschläge an Prof. Dr. Frank Puppe
<puppe@informatik.uni-wuerzburg.de> oder
Prof. Dr. Dieter Steinbauer <dieter.steinbauer@schufa.de>

Alle „Aktuellen Schlagwörter“ seit 1988 finden Sie unter:
www.ai-wuerzburg.de/as

Published online: 10 April 2013

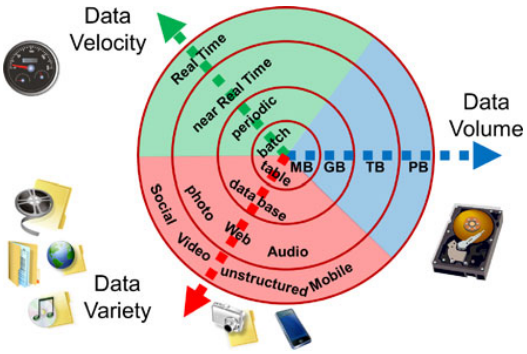


Abb. 1 3-V-Model für Big Data

Verwendung geprägt haben könnten [3]. Relativ unumstritten jedoch ist die Definition der Eigenschaften von Big Data durch Gartner im Jahr 2011 [4]. Das darin verwendete 3-V-Modell geht auf einen Forschungsbericht das Analysten Doug Laney zurück, der die Herausforderungen des Datenwachstums als dreidimensional bezeichnet hat [5]. Die drei Dimensionen beziehen sich auf ein ansteigendes Volumen (engl. *volume*) der Daten, auf eine ansteigende Geschwindigkeit (engl. *velocity*), mit der Daten erzeugt und verarbeitet werden und auf eine steigende Vielfalt (engl. *variety*) der erzeugten Daten (siehe Abb. 1). Im Folgenden gehen wir näher auf die Bedeutung der drei Dimensionen und der daraus resultierenden Herausforderungen ein.

Volume: Das prominenteste und wohl auch größte soziale Netz Facebook verzeichnet weltweit über eine Mrd. Nutzer, von denen monatlich über 600 Mio. über ein mobiles Endgerät auf das soziale Netz zugreifen. Pro Minute generieren die aktiven Nutzer in Facebook über 650.000 verschiedene Inhalte oder verteilen ca. 35.000 „Likes“ an Hersteller oder Organisationen [6]. Weitere Beispiele für das erzeugte Volumen sind die mehr als 200 Mio. Emails, die pro Minute verschickt werden oder die 175 Mio. Kurznachrichten bzw. Tweets, die über Twitter von den über 465 Mio. Accounts pro Tag gepostet werden. Diese enorme Ansammlung an Daten stellt für traditionelle Datenbanksysteme eine Herausforderung dar. Es gibt zwar bereits Datenbanksysteme im Petabyte Bereich, diese werden jedoch schnell teuer und daher besteht hier die Herausforderung, abzuwägen, welchen Wert Daten haben und ob diese die Kosten für große Datenbanksysteme aufwiegen.

Velocity: Der Aspekt Geschwindigkeit kann in zweierlei Hinsicht betrachtet werden. Erstens bezieht sich dies auf die enorme Rate, mit der Daten aktuell in den verschiedensten Anwendungsfeldern erzeugt werden. Zweitens muss diese rasch wachsende Datenmenge auch zeitnah weiterverarbeitet werden, um möglichst schnell darauf reagieren zu können. Je nach Anwendung kann dies bis in den Minuten oder gar Sekundenbereich gehen. Pro Minute werden zum Beispiel über Google mehr als 2 Mio. Suchanfragen abgesetzt, über Amazon mehr als 80.000 Dollar umgesetzt oder in YouTube 30 Stunden Videomaterial hochgeladen und 1,3 Mio. Videos konsumiert.

Variety: Die Vielzahl an Daten ist der wichtigste Aspekt in der Definition von Big Data. Die stark unterschiedlichen und oft nicht strukturierten Daten stellen gerade für traditionelle Datenbanksysteme ein Problem dar und können nicht effizient verarbeitet werden. In traditionellen relationalen Datenbanksystemen werden Datensätze mit Hilfe von Relationen abgespeichert. Dies kann man sich vereinfacht als Tabelle vorstellen, in der jede Zeile einem Datensatz entspricht. Die abzuspeichernden Daten müssen dazu eine Struktur besitzen. Ein Beispiel für strukturierte Daten könnten Kundenstammdaten sein (siehe linke Spalte in Abb. 2). Halbstrukturierte Daten besitzen zwar auch bis zu einem gewissen Grad eine Struktur, jedoch besitzen sie auch einen unstrukturierten Teil. Ein Beispiel hierfür könnten E-Mail-Nachrichten sein. Der Kopf der Nachricht besitzt eine klare Struktur wie Absender, Adressat oder Betreff. Der Rumpf der Nachricht jedoch kann jeglichen Inhalt oder auch beliebige Anhänge enthalten und ist daher ohne Struktur. Im Rahmen von Big Data werden jetzt alle vorhandenen Daten, ob strukturiert oder nicht, zusammengefasst und gemeinsam analysiert. Das hierdurch erzeugte, in sich unstrukturierte Datenkonglomerat kann in drei Kategorien klassifiziert werden. Die erste Kategorie enthält Daten, die aus der Kommunikation zwischen Personen entstehen. Beispiele hierfür wären Daten aus sozialen Netzen oder auch Web Protokolldateien. Die zweite Kategorie enthält Daten aus der Kommunikation zwischen Personen und Diensten oder Maschinen. Beispiele hier wären Daten aus E-Commerce Anwendungen oder auch Daten aus der Nutzung bestimmter Geräte wie zum Beispiel Geldautomaten. In der dritten Kategorie schließlich finden sich Daten zwischen Diensten oder Maschinen wie zum Beispiel Sensordaten, GPS Positionsinformationen oder Überwachungsbilder.

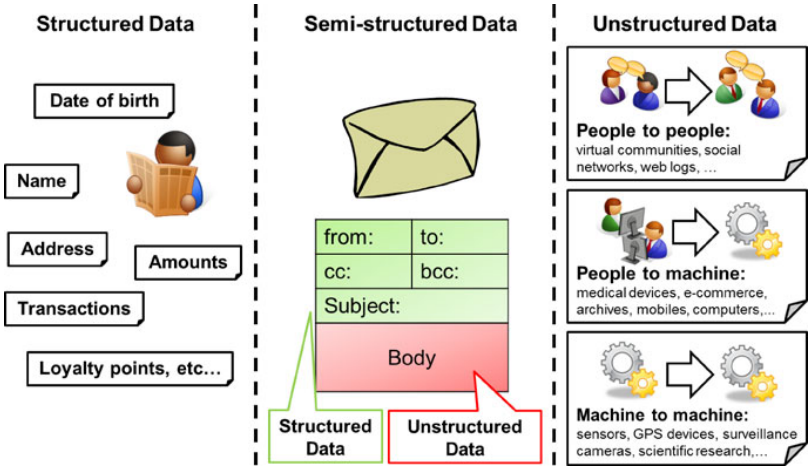


Abb. 2 Datenvielfalt

Ein viertes Attribut das ebenfalls häufiger zur Beschreibung von Big Data Verwendung findet, ist die Zuverlässigkeit (engl. *veracity*) der Daten, welches durch IBM geprägt wurde [7]. Die Herausforderung hierbei liegt darin, dass die Daten häufig aus unterschiedlichen Quellen kommen und daher eventuell zweifelhaft oder ungenau sind. Aufgrund der ebenfalls hohen Anforderungen an die schnelle Verfügbarkeit der Analysen können die Daten oft auch nicht rechtzeitig bereinigt werden. Somit haftet den gesammelten Daten häufig eine gewisse Unsicherheit oder Ungenauigkeit an, die es ebenfalls zu berücksichtigen gilt.

Unterschied zu Business Intelligence

Business Intelligence (BI) ist ein Begriff, der häufiger zusammen mit Big Data auftaucht, und daher wollen wir im Folgenden die Unterschiede kurz beleuchten. Der Begriff Business Intelligence wurde bereits 2009 in einem Aktuellen Schlagwort behandelt und wird als betriebliche Entscheidungsunterstützung durch einen integrierten, aufs Unternehmen bezogenen IT-basierten Gesamtansatz definiert [8]. Dabei werden die Daten aus unterschiedlichen Abteilungen extrahiert, transformiert und in einem zentralem Datenlager (engl. Data Warehouse (DW)) abgelegt. Selektionen aus den gesammelten Daten bezüglich eines bestimmten Kriteriums werden Data-Marts genannt und stellen einen nicht persistenten Zwischenspeicher zum Data Warehouse dar. Über definierte Schnittstellen können Business Intelligence Anwendungen, wie Leistungsanalysen oder Berichterstattungen, darauf zugreifen und die Daten weiterverarbeiten.

Das hauptsächliche Unterscheidungsmerkmal zwischen Business Intelligence und Big Data ist die Ausrichtung auf die gesammelten und verarbeiteten Daten. Business Intelligence Lösungen setzen strukturierte, konsistente und beständige Daten voraus, wohingegen Big Data Lösungen speziell auf unstrukturierte und möglicherweise nicht konsistente Daten hin optimiert sind. Dementsprechend sind die eingesetzten Technologien auch ausgerichtet. Business Intelligence Lösungen setzen auf traditionelle Datenbanksysteme wie relationale Datenbanken, wohingegen Big Data Lösungen auf neuen Konzepten wie zum Beispiel Not Only SQL (NoSQL) Datenbanken oder dem Hadoop Framework basieren, die effizienter mit unstrukturierten und großen Datenmengen umgehen können.

NoSQL Datenbanken

Not Only SQL (NoSQL) Datenbanksysteme sind für Aspekte konzipiert, bei denen relationale Datenbanksysteme an ihre Grenzen stoßen. Verwendet werden sie somit bei verteilten Systemen mit großen Datenmengen. Besonders bei Systemen, in denen vorhandene Werte nicht oft geändert, aber stets neue hinzugefügt werden, sind NoSQL Datenbanken von Vorteil. Beispiele für solche Anwendungsgebiete sind Twitter Posts oder Internet Server Protokolldateien. NoSQL Datenbanksysteme können in dokumentenorientierte Datenbanken, Graphen-Datenbanken und Key-Value-Datenbanken eingeteilt werden [9]. Dokumentenorientierte Datenbanken sind auf

das Speichern von halbstrukturierten Daten ausgelegt und erlauben ein Durchsuchen der Dokumentinhalte. Beispiele sind MongoDB und Apache CouchDB. Graphen-Datenbanken sind für das Speichern von Beziehungen zwischen verschiedenen Entitäten optimiert. Beispiele hierfür sind Neo4j und ArangoDB. Key-Value-Datenbanken schließlich speichern unter einem Schlüssel beliebige Werte und lassen sich noch einmal in zwei Untergruppen einteilen. In-Memory Varianten behalten die Daten komplett im Arbeitsspeicher und erzielen daher eine hohe Leistung. On-Disk Varianten hingegen speichern die Information auf der Festplatte und sind daher als Datenspeicher geeignet. Allgemeine Beispiele für Key-Value-Datenbanken sind Apache Cassandra oder BigTable.

Apache Hadoop Framework

Apache Hadoop ist ein auf Java basierendes freies Framework für die effiziente und hochskalierbare verteilte Berechnung von Aufgaben. Es basiert auf dem MapReduce Algorithmus und auf einem verteilten Dateisystem von Google und ermöglicht es, intensive Berechnungen auf Computerclustern durchzuführen. Der MapReduce Algorithmus nutzt den Teile-und-Herrsche-Lösungsansatz (engl. Divide and Conquer) und ist dementsprechend aus zwei Phasen aufgebaut. In der Map-Phase, welche dem Divide entspricht, werden die Eingabedateien in mehrere Fragmente unterteilt und jeweils einem Map-Task zugewiesen, der dann die gewünschte Berechnung auf seinem Fragment durchführt. Die Zwischenergebnisse aus der Map-Phase werden dann vom Hadoop Framework an die Reduce-Tasks in der Reduce-Phase verteilt. Hier wird dann aus den einzelnen Zwischenergebnissen das Gesamtergebnis berechnet, was dem Conquer entspricht.

Wissenschaftliche Herausforderungen und Ausblick in die Zukunft

Die wissenschaftlichen Herausforderungen im Bereich Big Data sind sehr vielfältig und stammen aus den unterschiedlichsten Forschungsgebieten. Die Herausforderungen im Bereich Datenmanagement beziehen sich auf eine effiziente Speicherung, Verteilung und Bereitstellung der großen Datenmengen. Hier geht es zum Beispiel um neue Speicherarchitekturen im Cloud Bereich oder auch geeignete Netzwerktopologien für Datenzentren, welche Big Data Analysen umsetzen. Im Bereich Datenanalyse sind geeignete statistische oder mathematische Algorithmen zur Modellierung und Darstellung der unterschiedlichen Daten sowie angepasste Mechanismen zur Wissensentdeckung auf großen und dynamischen Datenvolumina relevant. Ein weiteres Problem bei der Analyse ist die mögliche Unsicherheit in den Daten, die aus unterschiedlichen und eventuell unsicheren Quellen stammen. Im Bereich Netzwerktechnik sind vor allem Fragestellungen aus den Bereichen Datenübertragung oder Machine-to-Machine Kommunikation relevant. Hier geht es darum, wie große Datenmengen effizient übertragen werden können oder wie die Daten von einer Vielzahl von Sensoren an eine zentrale Datenhaltung gelangen.

Es stellt sich aber auch noch eine ganz andere Frage: Wenn alle vorhandenen Daten zusammengeführt und abfragbar sind und somit Antworten produziert werden könnten, fehlt oft noch die richtige Fragestellung. Welche Zusammenhänge in den Daten bestehen, welche Auswirkungen diese haben und was aus den aufbereiteten Daten gewonnen werden kann, übersteigt oft menschliches Fassungsvermögen und neue Verfahren müssen entwickelt werden, die dies greifbarer machen können.

Im Bereich Datenschutz ergeben sich ebenfalls interessante wissenschaftliche und gesellschaftliche Herausforderungen. Überall dort, wo personenbezogene Daten für Big Data Analysen verwendet werden, sollten diese natürlich mit geeigneten Verfahren anonymisiert werden. Hierbei sind die Verfahren davon abhängig, welche Arten von Daten gesammelt wurden. Bei GPS Positionsdaten zum Beispiel ist es nicht ausreichend, Name und Adresse zu löschen, da über den Tag-Nacht Zyklus hinweg sehr leicht herauszufinden ist, wo eine Person arbeitet oder wohnt. Gerade in Ländern wie Deutschland, wo der Datenschutz sehr ernst genommen wird, sind funktionierende Anonymisierungsverfahren eine Grundlage für den Erfolg von Big Data Projekten, können diese aber auch kompliziert und teuer machen. Allerdings gibt es gerade hier große Bedenken, da über das Internet Nutzerdaten auf Servern in der ganzen Welt gespeichert werden, die dann nicht mehr den strengen deutschen Datenschutzrichtlinien unterliegen.

In den nächsten Jahren werden durch die wachsende Anzahl und Diversität von Diensten immer größere Datenmengen anfallen. Zusätzlich steigern neue Mechanismen wie Gesichts- oder Gestener-

kennung noch die Information, die aus den Daten gezogen werden kann. Dieser große, unstrukturierte Datenhaufen wird in Zukunft durch Big Data Systeme strukturiert und für Analysen zugänglich gemacht werden. Hierfür sind einerseits geeignete Speicher- und Rechencluster notwendig und andererseits neue abstrakte und analytische Verfahren sowie statistische Methoden, um Zusammenhänge in den Daten zu verstehen und diese gewinnbringend nutzen zu können.

Literatur

1. ITU's ICT Data and Statistics Division (2012) Measuring the Information Society, 4th annual report on innovative and authoritative benchmarking tools. http://www.itu.int/ITU-D/ict/publications/idi/
2. Gens F (2012) International Data Corporation (IDC): Predictions 2012: Competing for 2020. http://www.idc.com/getdoc.jsp?containerId=prUS23177411#.UNGuw3fjH5m
3. McBurney V (2012) The Origin and Growth of Big Data Buzz, Toolbox for IT Blog. http://it.toolbox.com/blogs/infosphere/the-origin-and-growth-of-big-data-buzz-51509
4. Beyer M (o. J.) Gartner Says Solving „Big Data" Challenge Involcves More Than Just Managing Volumes of Data. http://www.gartner.com/it/page.jsp?id=1731916
5. Laney D (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety, Application Delivery Strategies published by META Group Inc.
6. James J (2012) How much data is created every minute? http://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/
7. Zikopoulos PC, deRoos D, Parasuraman K, Deutsch T, Corrigan D, Giles J, Melnyk RB (o. J.) Harness the Power of Big Data – The IBM Big Data Platform. http://www-01.ibm.com/software/data/bigdata/
8. Grünwald, Taubner D (2009) Business Intelligence, In: Informatik Spektrum: Aktuelles Schlagwort. http://link.springer.com/content/pdf/10.1007%2Fs00287-009-0374-1
9. heise.de: NoSQL im Überblick. http://www.heise.de/open/artikel/NoSQL-im-Ueberblick-1012483.html