Hochschule für Technik, Wirtschaft und Kultur Leipzig
Fakultät Informatik, Mathematik und Naturwissenschaften

Big Data - Überblicksvortrag

im Oberseminar „Datenbanksysteme - Aktuelle Trends“

Extended Abstract

Von: Martin Kolbe

Studiengang: Informatik Master | 17-INM

Datum: 25. Juni 2018

Inhaltsverzeichnis

1 Was ist Big Data?

Für den Begriff *Big Data* gibt es auf den ersten Blick keine einheitliche Definition. Im deutschen Wikipedia-Eintrag findet man jedoch einen ersten Versuch dazu, den Begriff in einem Satz zu definieren:

„Big Data [...] bezeichnet Datenmengen, welche bspw. zu groß, zu komplex, zu schnelllebig oder zu schwach strukturiert sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten."

Jedoch wird deutlich, dass hier vielmehr einzelne Merkmale aufgelistet werden, welche die Datenmenge von Big Data beschreiben. Dies fußt auf einer Definition der Eigenschaften von Big Data, welche das Unternehmen Gartner 2011 aufstellte, und die bislang als relativ unumstritten gilt. Grundlage hierfür war ein Forschungsbericht von Doug Laney. Er entwickelte ein erstes *3-V-Modell*, welches fortan kontinuierlich weiterentwickelt wurde.

2 Die 5 Vs und 3 Fs von Big Data

Das Modell der drei Vs ist die gängigste Betrachtungsweise der Eigenschaften von Big Data. Teilweise wird dieses Modell sogar in vier bis fünf Vs aufgebrochen. Zu den grundlegenden drei Vs zählen:

- **Volume:** Die schiere Menge an erzeugten Daten pro Minute im Netz stellt eine Herausforderung für klassische Datenbanksysteme dar. Hier muss abgewogen werden, welchen Wert bestimmte Daten haben und ob es sich lohnt, diese zu speichern.
- **Velocity:** Daten werden zum einen mit einer enormen Geschwindigkeit in verschiedenen Anwendungsfeldern erzeugt, zum anderen ist eine zeitnahe (teilweise sekundenschnelle) Weiterverarbeitung der Daten erforderlich.
- **Variety:** Die große Vielfalt an oft nicht strukturierten Daten stellt für klassische Datenbankmanagementsysteme eine Herausforderung dar und ist gleichzeitig der wichtigste Aspekt von Big Data. Dabei entstehen die Daten sowohl zwischen Personen als auch zwischen Diensten/Maschinen sowie zwischen Personen und Diensten untereinander.

Zusätzlich werden noch zwei weitere Vs für die Betrachtung herangezogen, welche jedoch nicht im ursprünglichen Bericht von Laney vorkommen. Diese ergeben sich vielmehr durch die praktische Betrachtung von Big Data. Sie lauten:

- **Veracity/Validity:** Die Zuverlässigkeit bzw. der Wahrheitsgehalt von Daten aus unterschiedlichen Quellen können sich stark unterscheiden, weshalb die Daten vorab gefiltert werden sollten. Dies steht jedoch im Zwiespalt zur schnellen Verfügbarkeit von Datenanalysen.

- **Value:** Hiermit ist der unternehmerische Mehrwert der Daten, welcher die Analyse der großen Datenmengen rechtfertigt, gemeint. Es müssen die Entscheidungsfindung und die Ergebnisse im Unternehmen verbessert werden, um einen tatsächlichen Mehrwert zu erzielen.

Neben dem V-Modell gibt es auch ein Modell mit drei Fs, welches die Anforderungen an Big Data aus Kundensicht betrachtet. Dieses Modell hat sich jedoch nicht gegen das V-Modell durchsetzen können. Die drei Fs lauten:

- **Fast:** Das Big-Data-System sollte Ergebnisse trotz der Heterogenität der Daten möglichst schnell dem Kunden bereitstellen.
- **Flexible:** Das System sollte mit geringem Aufwand an sich verändernde Bedingungen anpassbar sein. Neue Datenquellen oder Algorithmen sollten beispielsweise problemlos implementiert werden können.
- **Focused:** Nur die jeweils für den Anwendungsfall relevanten Datenquellen sollten betrachtet werden.

3 Datenquellen

Seit 2014 betrachtet haben fast alle großen Unternehmen ihre Big-Data-Projekte gestartet. Dazu zählen u.a. IBM, Oracle, EMC, Microsoft, Facebook, Google und Amazon. Allein in einer Minute werden im Internet 16 Millionen Textnachrichten verschickt, 40.000 Stunden Musik auf Spotify gehört und 452.000 Tweets auf Twitter verfasst (s. Abb. 1)

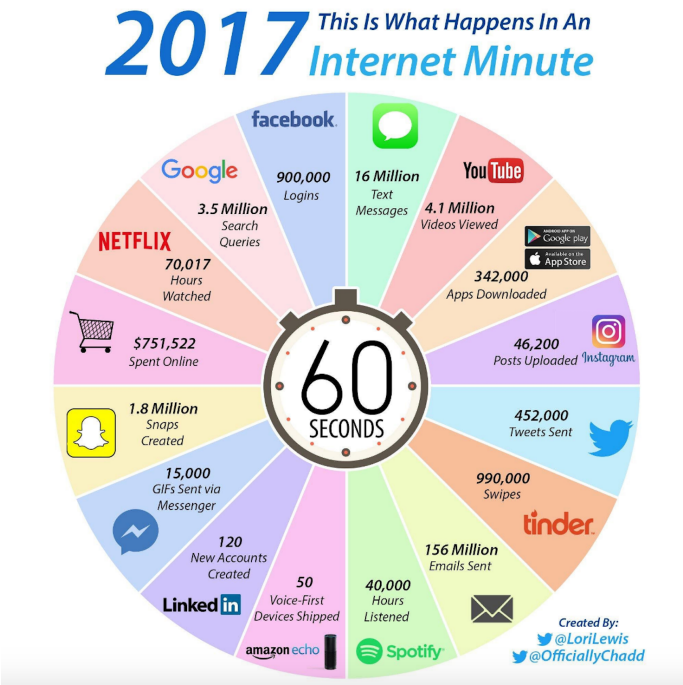

Abbildung 1: Was 2017 in einer Internet-Minute passiert. Quelle: Visual Capitalist

Jedoch zählen zu den Hauptquellen von Big Data ebenfalls Transaktionsdaten in Unternehmen, z.B. im Bereich der Warenwirtschaftssysteme. Hier entstehen auch logistische Daten. Das *Internet of Things* erweitert diese außerdem um sensorisch erfasste Daten.
Durch die im Internet stark geförderte Interaktion der Menschen untereinander werden zum einen eine große Menge an Kommunikationsdaten, zum anderen aber auch immer mehr Positionsdaten erfasst, welche u.a. der Personalisierung von Werbung dienen.
Als letzte große Quelle für Big Data seien ebenfalls Forschungseinrichtungen genannt, welche eine Vielzahl an Messdaten generieren.

4 Herausforderungen bei der Verarbeitung von Big Data

Die in Abschnitt 3 genannten Datenquellen umfassen verschiedene Anwendungsbereiche von Big Data. Diese haben jeweils verschiedene Probleme bei der Verarbeitung ihrer Datensätze zu bewältigen. Grundlegende Herausforderungen von Big Data lassen sich jedoch wie folgt zusammenfassen:

- Die **Datenrepräsentation** der stark heterogenen Daten sollte eine effiziente Analyse dieser ermöglichen.
- **Redundanzverringerung** und **Datenkompression** sind wichtig, um die Kosten des Big-Data-Systems (sowohl im technischen als auch im wirtschaftlichen Sinne) zu senken.
- Der **Lebenszyklus von Daten** muss verwaltet werden, um häufig benötigte Daten stets zum Abruf bereithalten zu können.
- Die **analytischen Mechanismen** müssen an die heterogenen Datensätze angepasst werden.
- Die **Vertraulichkeit der Daten** muss auch gewährleistet bleiben, selbst wenn bspw. Fremdfirmen mit ihrer Analyse beauftragt werden.
- **Energiemanagement** spielt als wirtschaftlicher Faktor eine große Rolle.
- Die **Erweiterbarkeit** und **Skalierbarkeit** in Hinblick auf die künftige Menge und Komplexität der Datensätze muss z.B. bei der Entwicklung von Analyseverfahren beachtet werden.
- Eine **Zusammenarbeit** von Experten verschiedener Fachbereiche ist notwendig, um gemeinsam die Verarbeitung von Big Data zu realisieren.

5 Anwendungsbereiche von Big Data

Als Hauptquellen für große Datenmengen werden häufig interne Unternehmensdaten und Daten aus dem Internet of Things genannt. Diese stellen einen Großteil der Anwendungsbereiche von Big Data dar. Jedoch entstehen auch in der Medizin große Mengen an Daten, wie das Beispiel des *Human Genome Project* zeigt.

Das Projekt hat sich im Bereich der Genforschung auf das weltweite Sammeln von Daten aus Genanalysen spezialisiert. Eine menschliche Gensequenz generiert ca. 100 bis 600 Gigabyte an Rohdaten. Die *China National Genebank* allein speichert 1,3 Millionen Genproben, wovon 1,15 Millionen menschlicher und 150.000 tierischer Herkunft sind. Das Sequenzieren von Genen wird in Zukunft voraussichtlich schneller durchführbar sein, wodurch auch das Datenaufkommen hierfür stark anwachsen wird. Hinzu kommen außerdem die Daten jeder einzelnen medizinischen Einrichtung. Das *University of Pittsburgh Medical Center* allein besaß schon 2012 knapp 2 Terabyte an medizinischen Daten. Das voraussichtliche durchschnittliche Datenvolumen eines Krankenhauses wurde demnach für 2015 auf 167 bis 665 Terabyte geschätzt.
Dass Unternehmen wie Google und Microsoft, welche bereits Erfahrungen im Umgang mit Big Data besitzen, in diesen Markt einsteigen wollen, verdeutlicht außerdem das Wachstumspotenzial dieses Anwendungsfeldes in Hinblick auf die Verarbeitung von Big Data.

6 Beziehung zum Internet of Things (IoT)

Internet of Things bezeichnet ein Netzwerk verschiedener physischer Geräte, Fahrzeuge, Heimapplikationen u.a. Gegenstände, die über Elektronik, Software, Sensoren, Aktoren (Antriebselemente) und Verbindungshardware verfügen, womit sie sich untereinander austauschen können.
Hier werden viele Eingabedaten erfasst, welche zwar teilweise durch die eingebauten Rechensysteme der Geräte vorab gefiltert werden können, aber dennoch im Rahmen von Big-Data-Systemen verarbeitet werden müssen. Schließlich handelt es sich hierbei um heterogene, unstrukturierte bzw. semistrukturierte Daten, welche gerade im Bereich der Sensorik auch oft redundant oder mit „unnützen“ Daten (sog. *Noise*) angereichert sein können (beispielsweise Videoaufnahmen von Überwachungskameras). Diese Daten sind nur dann brauchbar, wenn sie zeitnah analysiert werden.
Auch wenn es heute noch keinen dominanten Anteil einnimmt, gehen Schätzungen u.a. des Unternehmens HP davon aus, dass bis 2030 das Internet of Things mit ca. 1 Billion Sensoren einen Großteil von Big Data ausmachen wird.

7 Big-Data-Technologien

Neben den Internet of Things gibt es noch vier weitere Technologien, welche eine große Präsenz im Bereich von Big Data aufweisen. Diese sind:

- **Cloud Computing:** Eine Lösung zur Speicherung und Verarbeitung von Big Data, welche aus der Virtualisierung verschiedener Technologien entstanden ist. Hierbei werden die Daten in eine andere IT-Architektur überführt.
- **Data Center:** Eine Plattform zur zentralen Datenspeicherung, welche jedoch auch die Akquisition, Verwaltung und Organisation von Daten übernimmt. Ein Data Center bildet momentan

noch den Grundbaustein eines erfolgreichen Big-Data-Netzwerks. Es muss dabei stabil laufen, ausreichend Kapazität und schnelle Übertragungswege bieten. Ein Data Center muss außerdem immer auch mit der Größe des Big-Data-Netzwerkes und seiner Anwendungen wachsen können.

- **Apache Hadoop:** Ein in der Industrie weit verbreitetes Tool, u.a. zur Spamfilterung, Netzwerksuche, Clickstream-Analyse und für Social Recommendations. Im Juni 2012 lief Hadoop bei Yahoo bereits auf 42.000 Servern in vier Data Centers. Cluster mit bis zu 10.000 Knoten sind mit Hadoop 2.0 möglich.
- **CloudView:** Eine Plattform, welche vermischte Architekturen, lokale Knoten und ausgelagerte Hadoop-basierte Cluster nutzt, um Daten zu analysieren.

8 Komponenten und Aufgaben einer Big-Data-Plattform

Der Aufbau einer Big-Data-Plattform lässt sich in vier Bereiche unterteilen, welche schrittweise durchlaufen werden:

- Datenerstellung (data generation)
- Datenakquisition (data acquisition)
- Datenspeicherung (data storage)
- Datenanalyse (data analysis)

Die **Datenerstellung** ist der erste Schritt für Big Data. Es werden über mehrere Quellen verteilte Daten erzeugt. Da hierbei hauptsächlich das Internet als Datenquelle dient, finden sich viele Daten aus dem persönlichen Leben der Menschen, z.B. Sucheinträge, Chatverläufe oder Microblog-Nachrichten. Diese besitzen einen hohen Wert, aber eine niedrige Informationsdichte. Aber auch Transaktionsdaten aus Unternehmen sowie logistische, sensorische und Forschungsdaten werden hier erstellt.
Im Anschluss folgt die **Datenakquisition** bei der ein entsprechender Übertragungsweg der Daten zum Datenspeicher gefunden werden muss. Bereits hier sollten redundante Daten aussortiert werden, um die Übertragungsgeschwindigkeit effizient zu nutzen. Die Datenakquisition unterteilt sich in die drei Schritte der *Datensammlung (data collection)*, *Datenübertragung (data transmission)* und *Datenvoraufbereitung (data pre-processing)*.
Nach der Übertragung der Daten erfolgt die **Datenspeicherung** (bspw. in einem Data Center). Hierfür gibt es drei Arten von Speicherverfahren, die sich in ihrer Funktionsweise unterscheiden:

- **Speichersysteme für Massive Data (massive storage systems):** Es wird sich Speichermanagementsystemen wie DAS, NAS oder SAN bedient, welche die Daten innerhalb eines Netzwerks vorhalten.

- **Verteilte Speichersysteme (distributed storage systems):** Die Speicherung der Daten erfolgt auf Servern verschiedener Netzwerke, was das Speichersystem dem CAP-Theorem unterwirft.
- **Big-Data-Speicher-Mechanismen (big data storage mechanisms):** Dies umfasst Dateisysteme, Datenbanken und Programmiermodelle, welche im Big-Data-Bereich eingesetzt werden.

Schlussendlich erfolgt die **Datenanalyse**, bei der verschiedene analytische Verfahren über Big-Data-Datensätze aber auch statistische Verfahren aus der traditionellen Datenanalyse angewandt werden. Beispiele für analytische Methoden sind: *Hashing*, *Bloom Filter*, *Indexierung* oder *Parallel Computing*. Als statistische Methoden werden hingegen bspw. *Cluster Analalysis*, *Factor Analysis*, *Correlation Analysis* oder *Regression Analysis* durchgeführt. Die Analyse der Daten kann hierbei in real-time oder offline erfolgen.

9 Gesellschaftliche Aspekte von Big Data

Die aktuelle Anklage an Facebook, dass Cambridge Analytica ihre Nutzerdaten zweckentfremdet verwendet habe, um den US-Wahlkampf zu manipulieren, zeigt die Aktualität des Themas Big Data. Es wirft vor allem die Frage des politischen Einflusses auf, den Big-Data-Auswertungen und -Analysen haben.
Plattformen wie Facebook und Google bestimmen immer mehr, was wir sehen, wie wir die Welt wahrnehmen und vor allem, was wir nicht sehen. Durch Methoden wie *Microtargeting* werden Nutzer zu homogenen Gruppen zusammengefasst, um ihnen gezielt Werbeanzeigen und Angebote zu schalten. Man spricht in dem Fall von einer „Echokammer“ oder „Filterblase“, da die Nutzer nur noch bestimmte Informationen erreichen, die für sie von Analyse-Algorithmen vorausgewählt wurden.
Hinzu kommen *Habit-Forming-Technologies*, deren Ziel es ist, den Nutzer möglichst lange auf einer bestimmten Plattform zu halten und ihm Anreize zum Interagieren auf dieser zu bieten. Diese sind oft mit einem hervorgerufenem Glücksgefühl verbunden.
Dies führt unweigerlich auch zur Frage der sozialen und staatlichen Verantwortung. Das grundlegende Problem besteht schließlich im Missverhältnis von technischer Entwicklung und gesellschaftlicher Verantwortung. Möglicherweise sehen sich Unternehmen zukünftig in der Beweislast, zu erklären, welchem konkreten Zweck ihre Datenanalysen dienen, um einen Missbrauch wie im Falle des US-Wahlkampfes zu vermeiden.

Literaturquellen

Bücher

[1] Hrushikesha Mohanty, Prachet Bhuyan und Deepak Chenthati, Hrsg. *Big Data - A Primer*. Bd. 11. Studies in Big Data. Springer India, 2015. ISBN: 978-81-322-2493-8. DOI: 10.1007/s11036-013-0489-0.

Artikel

[2] Mi Chen, Shiwen Mao und Yunhao Liu. "Big Data: A Survey". In: *Mobile Networks and Applications* 19.2 (Apr. 2014), S. 171–209. DOI: 10.1007/s11036-013-0489-0.

[3] Dominik Klein, Tran-Gia Phuoc und Hartmann Matthias. "Big Data". In: *Informatik-Spektrum* 36.3 (Juni 2013), S. 319–323. DOI: 10.1007/s00287-013-0702-3.

[4] Stefan Wrobel u. a. "Big Data,Big Opportunities - Anwendungssituation und Forschungsbedarf des Themas Big Data in Deutschland". In: *Informatik-Spektrum* 38.5 (Okt. 2015), S. 370–378. DOI: 10.1007/s00287-014-0806-4.

Weblinks

[5] Thomas Beschorner und Martin Kolmar. *Die Gefahr durch Facebook wurde zu lange ignoriert*. 2018. URL: https://www.zeit.de/wirtschaft/2018-03/plattformkapitalismus-internetplattformen-regulierung-facebook-cambridge-analytica.

[6] DPA. *Big Data und Microtargeting: Was passiert mit unseren Daten?* Deutsche Presse-Agentur, 2018. URL: https://www.zeit.de/news/2018-03/21/big-data-und-microtargeting-was-passiert-mit-unseren-daten-180321-99-572947.

[7] Christoph Salzig. *Was ist Big Data? – Eine Definition mit fünf V*. 2016. URL: https://blog.unbelievable-machine.com/was-ist-big-data-definition-f%C3%BCnf-v.

[8] Roberto Simanowski. *Sie manipulieren, wir kollaborieren*. 2018. URL: https://www.zeit.de/kultur/2018-03/soziale-netzwerke-facebook-nutzerdaten-schutz-bedeutung-gesellschaft.