



Big Data

Extended Abstract

Vortrag am 19. April 2017

Inhaltsverzeichnis

1 Motivation	1
2 Begriff „Big Data“	1
3 Wertschöpfungskette	3
4 Technologien	3
5 Anwendungen	4
6 Kritik & Schattenseiten	5
Quellenverzeichnis	I

Jonathan Gruber, Matrikelnr: 68341

Oberseminar Datenbanksysteme – Aktuelle Trends

Hochschule für Technik, Wirtschaft und Kultur Leipzig

1 Motivation

Mit der fortschreitenden Digitalisierung nahezu aller Lebensbereiche geht ein rasantes Wachstum elektronischer Datenbestände einher. Während anfangs ausschließlich analoge Medien wie beispielsweise Magnetbänder oder Schallplatten den gesamten Wissensstand der Menschheit abbildeten, nahm der relative Anteil digitaler Daten in den letzten drei Jahrzehnten stark zu. Infolgedessen liegen heute nahezu ausschließlich digitale Medien vor. Auch die Gesamtkapazität der Daten ist explosionsartig angewachsen: 1986 lagen weltweit 2,6 Exabyte¹ analoge und 0,2 Exabyte digitale Daten vor. Bereits 19 Jahre später hatte sich die Gesamtdatenmenge etwa ver Hundertfacht und die digitalen Daten vereinnahmten mit 280 Exabyte einen deutlich größeren Anteil, als die analogen mit 19 Exabyte [9]. 2011 wurden alle zwei Tage 1,8 Zettabyte² an neuen Daten generiert. Gründe hierfür sind u. A. der Siegeszug des World Wide Webs und weiterhin die enorme Zunahme von Sensordaten des *Internets der Dinge*.

Diese anhaltende Entwicklung und das Aufkommen derartig gigantischer Datenmengen stellt aktuelle Datenverarbeitungssysteme vor große Herausforderungen. Es konnte vielfach praktisch gezeigt werden, dass großes wissenschaftliches und ökonomisches Potential in solchen riesigen Datensätzen steckt. Deshalb entstanden innerhalb der letzten Jahre eine Vielzahl neuer Technologien und Forschungsprojekte, um die Problemstellungen von *Big Data* zu lösen und Wert aus der Datenflut zu schöpfen.

2 Begriff „Big Data“

Der Begriff „*Big Data*“ ist nur vage definiert und wandelt sich kontinuierlich. Als medial sehr häufig verwendetes *Buzzword* umfasst er verschiedene Technologien wie *Cloud Computing*, *Parallel Computing*, *Data Mining*, *Business Intelligence* uvm. Grundsätzlich werden Datensätze, die so gigantisch groß sind, dass sie mit herkömmlichen Datenverarbeitungssystemen nicht bewältigt werden können als *Big Data* bezeichnet. Es existieren mehrere Beschreibungsmodelle, die unterschiedliche Aspekte des Begriffs hervorheben.

¹Exabyte = 10^{18} Bytes.

²Zettabyte = 10^{21} Bytes. Dies entspricht der **gesamten** Datenmenge der menschlichen Zivilisation bis 2003.

V-Modell

Eines dieser Modelle wurde bereits 2001 von *Dough Laney*, einem Analysten des US-amerikanischen Marktforschungsunternehmens Gartner, in einem Forschungsbericht [11] definiert. Das traditionelle sog. V-Modell beschreibt Big Data wie folgt:

- **Volume:** Der riesige Umfang der generierten und gespeicherten Daten im Peta-, Exa- und Zettabytebereich³.
- **Velocity:** Die enorme Datenrate und ausreichende Geschwindigkeit bei der Datenverarbeitung und -analyse.
- **Variety:** Die Vielzahl unterschiedlicher Datenformate, -typen und -quellen.

Einige Quellen erweitern das Modell um ein viertes V **Veracity**. Diese beschreibt die Glaubwürdigkeit und Akkuratess der Daten. 2011 fügte die *International Data Corporation* (IDC) dem Modell ein weiteres V **Value** hinzu, um die Sinnhaftigkeit von Big Data zu unterstreichen [7]: Die großen Datenmengen sind in ihrer *Gesamtheit* nach deren Analyse wertvoll und ermöglichen neue Erkenntnisse.

F-Modell

Während das V-Modell Big Data aus technischer Sicht beschreibt, betrachtet das F-Modell die **Anwendersicht**:

- **Fast:** Das Big-Data-System soll Ergebnisse möglichst schnell bereit stellen. Mögliche Flaschenhälse sind die Heterogenität der Daten, die verfügbare Ressourcen und die Problemkomplexität.
- **Flexible:** Es muss mit geringem Aufwand möglich sein das System an veränderte Bedingungen anzupassen (weitere Datenquellen, neue Algorithmen oder statistische Modelle).
- **Focused:** Es ist möglich nur relevante Datenquellen auszuwählen.

³10¹⁵ - 10²¹ Bytes.

3 Wertschöpfungskette

Um aus Big Data Wert neue wertvolle Erkenntnisse zu schöpfen werden typischerweise folgende Phasen durchlaufen:

1. Datengenerierung
2. Datenerfassung
3. Datenspeicherung
4. Datenanalyse

4 Technologien

Zur Bewältigung der großen technischen Herausforderungen bei der Verarbeitung derartig riesiger Datenbestände haben sich im Big-Data-Umfeld viele neue Disziplinen und Technologien entwickelt:

- **Cloud Computing** stellt die riesigen, benötigten Speicher- und Rechenkapazitäten zur Verfügung und realisiert damit Skalierbarkeit (*Infrastruktur*). Big Data und Cloud Computing sind eng miteinander verknüpft und treiben deren jeweilige Entwicklung voran.
- **Verteilte Dateisysteme** ermöglichen die Speicherung enormer Datenmengen und erhöhen die Ausfallsicherheit durch Redundanz der Dateien. Vertreter sind beispielsweise das *Google File System* (GFS) oder das *Hadoop Distributed File System* (HDFS).
- **NoSQL-Datenbanksysteme**⁴ zur Speicherung der Datenmengen als bessere geeignete Alternativen zu klassischen relationalen Datenbanksystemen (RDBMS). Schemafreiheit, gute horizontale Skalierbarkeit und die bewusste Verletzung mancher ACID⁵-Prinzipien ermöglichen die effiziente Verarbeitung riesiger Datenmengen. NoSQL-Datenbanken können wie folgt klassifiziert werden:
 - Key-Value-Datenbanken wie *Berkeley DB*.
 - Dokumentenorientierte Datenbanken wie *MongoDB* oder *Apache Couch DB*.
 - Spaltenorientierte Datenbanken wie *Google Big Table* oder *Apache Cassandra*.

⁴*Not only SQL.*

⁵Atomicity, Consistency, Isolation, Durability.

- **MapReduce**-Programmierung ermöglicht die Parallelverarbeitung von Daten im Petabytebereich⁶ auf verteilten Systemen.
- **Apache Hadoop** ist ein beliebtes Open-Source Framework, das eine vollständige Big-Data-Plattform realisiert. Es stellt u. a. ein verteiltes Dateisystem (HDFS) und eine Implementierung von *Map Reduce* bereit.

5 Anwendungen

Big Data ist in der praktischen Anwendung für verschiedenste Unternehmen, Organisation und akademischen Einrichtungen hochinteressant:

- **Business und E-Commerce:** Vorhersage von Kundenverhalten, Marktforschung, Preisvergleiche, Personalwesen, Analyse des Kaufverhaltens und Finanztransaktionen.
- **Social Media:** Analyse der Interessen, Beziehungen, Verhaltensmuster, Demografie etc. der Mitglieder des sozialen Netzwerks.
- **Gesundheitswesen:** Analyse und Vorhersage von Pandemien (z. B. *Google Flu*), Auswertung medizinischer Daten (Beispiel Vitaldaten von Frühgeborenen zur frühzeitigen Erkennung von Infektionen).
- **Crowdsourcing:** Kollektive Problemlösung durch Verteilen von Teilaufgaben an freiwillige Helfer und Rekombination dieser Teillösungen (z. B. *BOINC-Projekt*⁷: Software-Plattform für verteiltes Rechnen).
- **Naturwissenschaften:** Auswertung hochkomplexer Experimente mit gigantischen Datenaufkommen wie beispielsweise des ATLAS-Detektors im *Large Hadron Collider* am CERN⁸.
- **Internet der Dinge:** Beispiel *Smart Cities* (Verkehrssteuerung, Überwachung von Umweltfaktoren, Energiemanagement, öffentliche Sicherheit usw.).

⁶Petabyte = 10^{15} Bytes.

⁷Berkeley Open Infrastructure for Network Computing.

⁸Conseil Européen pour la Recherche Nucléaire (Europäische Organisation für Kernforschung).

6 Kritik & Schattenseiten

Während mit Big-Data-Anwendungen viele interessante Fragestellungen beantwortet werden können, birgt die Technologie auch viele gesellschaftliche und ethische Probleme:

1. Gefahr des „gläsernen Bürgers“, der Datenschutz ist oftmals sehr fragwürdig. Im Extremfall droht ein orwellscher Überwachungsstaat.
2. Zu wenig Regulierungen und Normierung. Die Politik reagiert und reguliert langsamer, als die tatsächliche Entwicklungen vonstatten geht.
3. Mangelnde Transparenz für den Anwender im World Wide Web. Nahezu jede Interaktion hinterlässt heute digitale Fingerabdrücke und wird potentiell verwertet.

Quellenverzeichnis

- [1] Eric A. Brewer. „Towards Robust Distributed Systems“. In: *Proceedings of the Nineteenth Annual ACM Symposium on Principles of Distributed Computing*. PODC '00. Portland, Oregon, USA: ACM, 2000, S. 7–. ISBN: 1-58113-183-6. DOI: 10.1145/343477.343502. URL: <http://doi.acm.org/10.1145/343477.343502>.
- [2] Min Chen, Shiwen Mao und Yunhao Liu. „Big Data: A Survey“. In: *Mob. Netw. Appl.* 19.2 (Apr. 2014), S. 171–209. ISSN: 1383-469X. DOI: 10.1007/s11036-013-0489-0. URL: <http://dx.doi.org/10.1007/s11036-013-0489-0>.
- [3] IBM Deutschland. „IBM 5 in 5“: Innovationen, die unser Leben verändern werden. Dez. 2013. URL: <http://www-03.ibm.com/press/de/de/pressrelease/42779.wss> (besucht am 17. 04. 2017).
- [4] Andreas Dewes und Stephanie Rohde. *Es wird immer schwieriger, sich zu schützen*. Deutschlandfunk. Jan. 2017. URL: http://www.deutschlandfunk.de/datensicherheit-es-wird-immer-schwieriger-sich-zu-schuetzen.694.de.html?dram:article_id=377536 (besucht am 16. 04. 2017).
- [5] Klaus-Peter Eckert und Radu Popescu-Zeletin. „Smart Data als Motor für Smart Cities“. In: *Informatik-Spektrum* 37.2 (2014), S. 120–126. ISSN: 1432-122X. DOI: 10.1007/s00287-014-0769-5. URL: <http://dx.doi.org/10.1007/s00287-014-0769-5>.
- [6] Johann-Christoph Freytag. „Grundlagen und Visionen großer Forschungsfragen im Bereich Big Data“. In: *Informatik-Spektrum* 37.2 (2014), S. 97–104. ISSN: 1432-122X. DOI: 10.1007/s00287-014-0769-5. URL: <http://dx.doi.org/10.1007/s00287-014-0769-5>.
- [7] John Gantz und David Reinsel. *Extracting Value from Chaos*. 2011. URL: <http://www.emcgrandprix.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf>.
- [8] Tony Hey. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft External Research. 2016. URL: <http://fiz1.fh-potsdam.de/volltext/fhpotsdam/10445.pdf> (besucht am 15. 04. 2017).
- [9] Martin Hilbert und Priscila López. „The World’s Technological Capacity to Store, Communicate, and Compute Information“. In: *Science* 332.6025 (2011), S. 60–65. ISSN: 0036-8075. DOI: 10.1126/science.1200970. eprint: <http://science.sciencemag.org/content/332/6025/60.full.pdf>. URL: <http://science.sciencemag.org/content/332/6025/60>.
- [10] Telecommunication Standardization Sector of ITU. *Overview of the Internet of things*. Techn. Ber. International Telecommunication Union, Juni 2012.

- [11] Douglas Laney. *3D Data Management: Controlling Data Volume, Velocity, and Variety*. Techn. Ber. META Group, Feb. 2001. URL: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>.
- [12] Sam Madden. „From Databases to Big Data“. In: *IEEE Internet Computing* 16.3 (Mai 2012), S. 4–6. ISSN: 1089-7801. DOI: 10.1109/MIC.2012.50. URL: <http://dx.doi.org/10.1109/MIC.2012.50>.
- [13] Viktor Mayer-Schönberger und Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. UK: John Murray Publishers, 2013.
- [14] World Health Organization. *How Much Information?* University of California at Berkeley. 2000. URL: <http://www2.sims.berkeley.edu/research/projects/how-much-info/how-much-info.pdf> (besucht am 16. 04. 2017).
- [15] World Health Organization. *Influenza (Seasonal). Fact sheet*. 2016. URL: <http://www.who.int/mediacentre/factsheets/fs211/en/> (besucht am 13. 04. 2017).
- [16] World Health Organization. *Preterm Birth. Fact sheet*. 2016. URL: <http://www.who.int/mediacentre/factsheets/fs363/en/> (besucht am 14. 04. 2017).
- [17] Dominik Ryžko u. a. *Machine Intelligence and Big Data in Industry*. 19. Springer International Publishing, 2016. DOI: 10.1007/978-3-319-30315-4.
- [18] Chris Strohm. *Predicting Terrorism From Big Data Challenges U.S. Intelligence*. Okt. 2016. URL: <https://www.bloomberg.com/news/articles/2016-10-13/predicting-terrorism-from-big-data-challenges-u-s-intelligence> (besucht am 16. 04. 2017).