Fakultät IMN

Medieninformatik (Master)

Oberseminar Datenbanksysteme – Aktuelle Trends

Sommersemester 2014

Prof. Dr.-Ing. Thomas Kudraß

Big Data

Martin Wilmer (13MIM)

11. Juli 2014

Inhaltsverzeichnis

1. Motivation

Der Begriff *Big Data* ist derzeit in aller Munde. Im Zeitalter der umfassenden und alle Bereiche des Lebens erreichenden Digitalisierung werden mehr und mehr Daten systematisch gesammelt, gespeichert, analysiert und Nutzern zugänglich gemacht. Im World Wide Web werden pro Minute zwei Millionen *Google* Suchanfragen gestellt, 30 Stunden Videomaterial bei *YouTube* hochgeladen und 650.000 *Facebook*-Posts abgesetzt. Laut der Nutzeranzahl wäre Facebook das drittgrößte Land der Welt nach China und Indien.

Forscher schätzen das derzeitige Gesamtdatenvolumen der Welt auf drei bis fünf Zettabyte. Ein Zettabyte ist entspricht 10^{21} Byte oder einer Milliarde Terabyte. Prognosen zufolge soll diese Datenmenge bis zum Jahr 2020 auf 40 Zettabyte wachsen. Allein die *Europäische Organisation für Kernforschung CERN* erzeugt pro Jahr 30 Petabyte neue Forschungsdaten.

Diese unvorstellbaren Datenmassen zu erfassen, zu speichern und effizient weiterzuverarbeiten stellt Wissenschaft und Forschung derzeit vor zahlreiche neue Herausforderungen.

2. Grundlagen

Big Data wird häufig als Schlagwort oder Überbegriff für verschiedene informatische Disziplinen wie beispielsweise Data Mining, Information Extraction, Data Analytics, Complex Event Processing und Data Warehousing verwendet. Der Begriff ist stark überladen. Demnach fällt eine eindeutige Definition schwer.

Madden vom *Massachusetts Institute of Technology* formulierte in einer Veröffentlichung aus dem Jahr 2012 folgende etwas salopp klingende Definition: (vgl. [Madden])

> Big Data is data that's too big, too fast, or too hard for existing tools to process.

Die drei genannten Adjektive *big, fast* und *hard* finden sich auch in der Big Data-Definition von *Doug Laney* aus dem Jahr 2011 wieder. Er beschrieb die wichtigsten Charakteristika von Big Data mit den sogeannten drei Vs, welche häufig in der Fachliteratur Verwendung finden. Sie sind wie folgt beschrieben:

Volume meint die Datenmengen im Bereich von Peta- (10^{15} Byte), Exa- (10^{18} Byte) und Zettabyte (10^{21} Byte).

Velocity meint die notwendige Geschwindigkeit der Datenerfassung und -verarbeitung.

Variety meint die Vielzahl unterschiedlicher Datenformate, -typen und -quellen.

Big bezieht sich demnach auf das riesige Datenvolumen (engl. *Volume*), *fast* auf die Geschwindigkeit der Datenerfassung und -verarbeitung (engl. *Velocity*) und *hard* auf die starke Varianz von Datenformaten, -typen und Quellen (engl. *Varierty*), für welche aktuell verfügbare Tools nicht ausreichend leistungsfähig bzw. flexibel sind.

Einige wissenschaftliche Veröffentlichungen ergänzen dieses Modell um ein viertes V: *Veracity*. Damit soll die Schwierigkeit hinzugenommen werden, die Qualität, Genauigkeit und Glaubwürdigkeit von Daten automatisiert zu bewerten. Kritiker bemängeln jedoch, dass es sich dabei um keine quantitativ messbare Größe handelt.

Quellen für Big Data sind unter anderem Prozess- und Produktionsdaten, Web-Daten, Wissenschafts- und Geschäftsdaten, Sensordaten, Daten mobiler Geräte und RFID-Chips, sowie Daten aus sozialen Netzwerken. Die erfassten Daten sind häufig stark heterogen und nur semi- oder gar nicht strukturiert.

Freytag kategorisiert die Verarbeitungsprozesse, die mit Big Data durchgeführt werden, wie folgt: (vgl. [Freytag])

Nachverfolgen und Auswerten Prozesse und Zustände erfassen und bewerten, um eventuelle Korrekturen vorzunehmen (möglichst in Echtzeit).

Suchen und Identifizieren Gezieltes Herausfiltern von Objekten aus einer großen Vielzahl von Objekten anhand weniger bekannter Merkmale zur weiteren Verarbeitung.

Analysieren Analyse großer Datenmengen mit Techniken aus Bereichen der künstlichen Intelligenz und Statistik (z.B. Regressionsverfahren, Verfahren zur Cluster-Bestimmung, Herleitung von Assoziationsregeln); Ziel ist es, aus Daten Informationen bzw. Wissen herzuleiten.

Vorhersagen und Planen Aus den anderen Verarbeitungsprozessen gewonnenen Erkenntnisse für die Zukunft nutzen, um beispielsweise Geschäfts- oder Fertigungsprozessen zu optimieren.

3. Anwendungsgebiete

Der Anwendungsbereich von Big Data ist nahezu unbegrenzt. Vor allem in datenintensiven und datengetriebenen Bereichen verspricht die Verwendung von Big Data-Technologie große, zum Teil wirtschaftliche Vorteile. Zu den wichtigsten Anwendungsgebieten zählen:

- naturwissenschaftliche Forschung
- Wasser- und Energiemanagement
- Medizin und Gesundheitsmanagement
- Produktlebenszyklusmanagement (*Industrie 4.0*)

- Smart Ecosystems, Smart Cities, Smart Home
- personalisierte Online-Werbung
- Trendanalysen in Social Media
- Cyber Security

3.1. Astronomie

Als ein Pionier der Big Data-Bewegung gilt der Amerikaner *Jim Gray*, der es sich mit seinem Forschungsteam im *Sloan Digital Sky Survey Project* zur Aufgabe gemacht hat, den Himmel vollständig digital zu erfassen. Die im Rahmen des Projekts entstandene Analyse- und Kollaborationsplattform für Astronomen erfasst Himmelsobjekte anhand von Bildaufnahmen und Messungen. Bisher konnten mehr als 930.000 Galaxien und 120.000 Quasare dokumentiert werden. Pro Nacht entstehen dabei etwa 250 Gigabyte neue Daten. (vgl. [Freytag])

3.2. Medizin

Auch in der Medizin ergeben sich aus der Big Data-Technologie vielversprechende Anwendungen. So werden derzeit Möglichkeiten untersucht, Ursachen für Krebserkrankungen aus genetischen Veränderungen zu schließen. Allein die digitale Erfassung des menschlichen Genoms mit seinen 3,2 Milliarden Basenpaaren und deren Kombinationen erzeugt diagnostische Rohdaten im Bereich von 300 bis 500 Gigabyte. Sollen diese gegen die 80 Millionen bekannten Mutationen auf 25.000 Genen geprüft werden, entstehen weitere riesige Datenmengen. Die konventionelle Datenverarbeitung benötigt für die Auswertung der Daten eines Patienten Zeit im Bereich von mehreren Wochen bzw. Monaten. Mit Big Data-Technologie ist es erst mals möglich, diese Berechnungen in Echtzeit durchzuführen. Am *Hasso Plattner Institut* wurden dafür spezielle Ansätze entwickelt. Dank In-Memory-Technologie und Multicore-Architekturen mit 1.000 Kernen sind flexible Echtzeitanalysen möglich. (vgl. [Meinel])

4. Big Data-Plattformen

Um den Herausforderungen von Big Data gewachsen zu sein, ist es notwendig neue Konzepte und spezialisierte Softwarelösungen, sogenannte Big Data-Plattformen, zu entwickeln.

4.1. Anforderungen

Da die drei bzw. vier zuvor beschriebenen Vs eher technisch orientiert sind, werden die Anforderungen an Big Data-Plattformen mit drei Fs aus der Nutzerperspektive beschrieben:

Fast Die Ausführungsumgebung sollte in der Lage sein, das Ergebnis so schnell wie möglich zu erzeugen. Dies ist jedoch stark abhängig von der Komplexität der Aufgabe, dem Umfang und der Heterogenität der Daten, sowie den verfügbaren Ressourcen.

Flexible Bereits existierende Verarbeitungsaktivitäten sollten sich mit geringem Aufwand an veränderte Bedingungen anpassen lassen. Dies kann z.B. das Einbeziehen neuer Datenquellen oder neuer Werkzeuge und Algorithmen sein.

Focused Es sollte mit geringem Aufwand möglich sein, relevante Datenquellen und Verarbeitungsschritte auszuwählen, welche für die Erfüllung einer Aufgabe sinnvoll sind.

4.2. Komponenten und Funktionen

Freytag beschreibt folgende Komponenten und Funktionen, welche eine anwenderfreundliche Big Data-Plattform implementieren sollte: (vgl. [Freytag])

Datenvisualisierung Ergebnisse sollen übersichtlich und verständlich dargestellt werden. Hier wird unter anderem auf Ansätze aus dem Data Warehousing-Bereich zurückgegriffen.

Datenintegration Die strukturellen Unterschiede verschiedener Daten und deren Strukturbeschreibung sollen im Datenzugriff berücksichtigt werden.

Entitätsintegration Informationen aus verschiedenen Quellen über dasselbe Objekt aus der realen Welt sollen zusammengeführt und verschmolzen werden können.

Datenqualität Es soll eine automatisierte Bewertung der Qualität von Daten stattfinden. Fehlerhafte Daten sollen erkannt und ausgeschlossen werden.

Datenherkunft bzw. -abstammung Es sollen Metadaten über die Herkunft der Daten, über die Datenquelle und den Erzeugungsprozess gespeichert werden. Des Weiteren soll eine Versionierung der Daten für die spätere Weiterverarbeitung erfolgen.

Prozess- und Workflow-Management Komplexe Datenanalysen und -transformationen setzen sich aus atomaren Aktionen zusammen, welche als Einheit in Form einer Prozesskette (Workflow) behandelt werden sollen.

Metadaten-Management Im Laufe eines Lebenszyklus anfallende Metadaten über die Datenqualität, -herkunft und -transformation durch Prozesse sollen gespeichert werden. Diese enthalten Hinweise zur Nutzbarkeit und Brauchbarkeit der Daten.

4.3. Technische Entwicklung

Um die riesigen Datenmengen von Big Data zu speichern, ist ein enormer Speicherplatz notwendig. Die Entwicklung von preisgünstigen und schnellen Massenspeichertechnologien ermöglicht es, dieser Anforderung zu entsprechen. Grundlage dafür sind Terabyte *Hard Disk Drives* und die *Solid State Drive*-Technologie. Datenbanken und Dateisysteme werden zunehmend auf Cluster verteilt. Dafür sind horizontal skalierbare Hard- und Softwaresysteme im Einsatz. Ein weiteren Trend ist die Dezentralisierung der Datenhaltung in Cloud-Speichern. Dienste wie *Amazon S3* und *Google Cloud Plaform* stellen gegen Bezahlung Speicherplatz in der Cloud bereit. Um die Anschaffung und Wartung von Hardware, sowie die Partitionierung der Daten muss sich der Anwender dabei nicht kümmern.

Der Umgang mit Big Data erfordert extrem hohe Verarbeitungsgeschwindigkeiten. Um Zugriffszeiten zu minimieren werden Daten zunehmend im schnellen Hauptspeicher gehalten (In-Memory-Technologie). Dabei werden bereits Hauptspeicher mit Volumen im Terabyte-Bereich eingesetzt. Zudem ermöglichen Multicore-Technologien die massive Parallelisierung der Datenverarbeitung. Grundlage für die Parallelisierung sind neue Programmiermodelle wie *MapReduce* von *Google*. Ebenso wie Speicherbedarf, kann auch Rechenleistung in die Cloud ausgelagert werden. Durch Virtualisierung kann die Big Data-Verarbeitung alternativ auch ortsunabhängig stattfinden.

Die Heterogenität der Datenformate und -typen von Big Data mündet im Bereich der Datenbanken in Weiterentwicklungen bestehender Systeme, als auch in zahlreichen Neuentwicklungen. *NoSQL*-Datenbankmanagementsysteme und -konzepte gewinnen dabei zunehmend an Bedeutung. Durch den Verzicht auf strikte *ACID*-Konformität und den Einsatz von Versionierungskomponenten (*Multi Version Concurrency Control*) ermöglichen sie einen hohen Datendurchsatz, sowie flexible Speicherstrukturen.

4.4. Produktübersicht

In den letzten Jahren sind zahlreiche neue Technologien für den Einsatz im Big Data-Bereich entstanden. Aber auch große Datenbank-Hersteller wie Oracle, IBM und Microsoft haben Erweiterungen für ihre bestehenden Datenbanksysteme entwickelt, um den Anforderungen von Big Data gerecht werden zu können. Die folgende Auflistung fasst einige bedeutende Produkte zusammen:

MapReduce-basierte Lösungen Hadoop (Apache), Stratosphere (TU Berlin), InfoSphereBigInsights (IBM)

NoSQL-DBMS HBase (Apache), HANA (SAP), Vertica (HP), Greenplum (EMC), Teradata, Cassandra (Apache)

Weiterentwicklungen von relationalen DBMS Oracle Database Server und Exadata (Oracle), DB2 (IBM), SQL Server (Microsoft)

Data Mining, Information Extraction SystemT (IBM), SAS Analytics, R, Matlab (Mathworks)

Indizierung, Volltextsuche Lucene (Apache), Solr (Apache)

A. Literaturverzeichnis

Alle Online-Quellen waren am 12. Juni 2014 unter der angegebenen Adresse verfügbar.

Apache Apache Software Foundation: *What Is Apache Hadoop?* 2014, URL: `http://hadoop.apache.org/`

Cardenas Cardenas, Alvaro A.; Manadhata, Pratyusa K. und Rajan Sreeranga P.:*Big Data Analytics for Security.* In: IEEE Security & Privacy, November/Dezember 2013, S. 74-76

Dimiduk Dimiduk, Nick und Khurana, Amandeep: *HBase in Action.* 2013, Manning Publications Co.

Fischer Fischer, Stephan: *Big Data: Herausforderungen und Potenziale für deutsche Softwareunternehmen.* In: Informatik Spektrum, Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen, Band 37, Heft 2, April 2014, S. 112-119

Freytag Freytag, Johann-Christoph: *Grundlagen und Visionen großer Forschungsfragen im Bereich Big Data.* In: Informatik Spektrum, Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen, Band 37, Heft 2, April 2014, S. 97-104

Liggesmeyer Liggesmeyer, Peter; Dörr, Jörg und Heidrich, Jens: *Big Data in Smart Ecosystems.* In: Informatik Spektrum, Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen, Band 37, Heft 2, April 2014, S. 105-111

Madden Madden, Sam: *From Databases to Big Data.* In: IEEE Internet Computing, Mai/Juni 2012, S. 4-6

Meinel Meinel, Christoph: *Big Data in Forschung und Lehre am HPI.* In: Informatik Spektrum, Organ der Gesellschaft für Informatik e.V. und mit ihr assoziierter Organisationen, Band 37, Heft 2, April 2014, S. 92-96

Redmond Redmond, Eric und Wilson, Jim R.: *Sieben Wochen, sieben Datenbanken. Moderne Datenbanken und die NoSQL-Bewegung.* Übersetzt von Peter Klicman. 2012, O'Reilly Verlag GmbH & Co. KG

Ruhmann Ruhmann, Ingo: *NSA, IT-Sicherheit und die Folgen. Eine Schadensanalyse.* In: Datenschutz und Datensicherheit - DuD, Asugabe 38 (1/2014), S. 40-46