Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Gerrit Thede

Big Data - Datenquellen und Anwendungen Ausarbeitung Grundlagen Vertiefung und Anwendungen 1

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Gerrit Thede

Big Data - Datenquellen und Anwendungen Ausarbeitung Grundlagen Vertiefung und Anwendungen 1

Eingereicht am: 10.03.2014

1 Big Data

1.1 Einleitung

Das Datenaufkommen in der Informationstechnik ist in der letzten Zeit rasant gestiegen und produziert eine Menge und Vielfalt an Daten, die es erforderlich gemacht haben, mit neuen Werkzeugen und Algorithmen an das Problem der Informationsgewinnung heranzugehen. Dieses Phänomen ist mit dem Begriff „Big Data“ klassifiziert, dessen Eigenschaften Volume, Velocity und Variety im 3-V Modell [12] beschrieben werden. Es bezeichnet die enorme Größe der Datenmenge und ihr stetiges Wachstum (Volume), die Geschwindigkeit der Verarbeitung und Informationsgewinnung um zeitnahe Analysen zu finden (Velocity) und die vielen verschiedenen Datentypen und Datenstrukturen (Variety), die verarbeitet werden müssen.

Anwendungen im Einsatzgebiet von „Big Data“ haben das Ziel, mit Hilfe der Analyse und in Kombination mit anderen Datenquellen in einem Prozess der Informationsgewinnung neue, wertvollere Informationen zu erschaffen, sogenannte „Smart Data“.

Die „Big Data“-Strategien sollen es ermöglichen, kosteneffizienter, schneller und besser skalierend Daten zu verarbeiten, als es mit den bisherigen Herangehensweisen der Fall war.

Es gibt keine Grenze, ab welcher Datenmenge der Übergang zur „Big Data“ verläuft, aber wenn die Menge der Daten nicht mehr in den Arbeitsspeicher oder Festplattenspeicher passen und es einen schnellen Eingangstrom neuer Daten gibt, kann die Verarbeitung der Daten auf einer verteilten und skalierenden „Big Data“-Plattform davon profitieren. Der Einsatz eines verteilten

Systems mit vielen Rechnern in einem Cluster und der Parallelisierung der Datenverarbeitung kann „Big Data“ wieder verarbeitbar machen.

1.2 Datenquellen

Es besteht die Möglichkeit, dass Unternehmen noch nicht alle Datenquellen für sich erschlossen haben, aus dem Grund, dass sie wegen der anfallenden Menge an Daten nicht gespeichert oder nicht weiter verarbeitet werden können. Um den eigenen Datenbestand zu erweitern und mit anderen Informationen zu verschmelzen, kann eine Vielzahl von neuen Datenquellen erschlossen werden.

Als Quelle kann beispielsweise das Social Web (Twitter, Facebook, Google+) herangezogen werden, um zum eigenen Datenbestand relevante Informationen zuzuordnen. Aus diesem kontinuierlichen Strom an Daten, der eine zeitliche Komponente und eventuell auch Geodaten enthält, können in Kombination wertvolle Informationen gewonnen werden.

Der verbreitete Einsatz von Smartphones bietet ein neue Datenquelle, die Mobilitätsdaten und Nutzungsprofile direkt von den Nutzern liefern kann. Das Anbieten von Anwendungen für Smartphones bietet einem Unternehmen die Möglichkeit das Nutzungsprofil seiner Kunden auszuwerten. Dies kann intransparent für die Nutzer geschehen, aber es bietet auch die Möglichkeit, dass Nutzer freiwillig Daten preisgeben und in einer Art „Crowd Sourcing“ als Datenquelle zur Verfügung stehen. Anwendungen sind zum Beispiel die Auswertung von Bewegungsdaten im Straßenverkehr um das Verkehrsaufkommen zu bestimmen.

Es bietet sich auch die Auswertung von Sensormessdaten an, die bisher zwar schon erfasst wurden, aber wegen der Menge des Datenstroms nicht gespeichert und analysiert wurden. Ein Beispiel sind Stromzählerdaten, deren Werte zu Stichtagen erfasst werden und damit nur in geringer zeitlicher Auflösung gespeichert werden.
Eine wichtige Komponente bei der Erschließung der Datenquellen ist die

Verlässlichkeit und Qualität der Daten. Gerade das Socialweb bietet nicht unbedingt einen repräsentativen Ausschnitt aller Nutzergruppen.

Eine qualitativ hochwertige Datenquelle kann in Zukunft mit dem Einzug von „Open Data“ und „Open Government“ [9] entstehen, wenn öffentliche Verwaltungen dazu verpflichtet werden, öffentliche Daten den Bürgern zur Verfügung zu stellen. Die Aufbereitung und Verarbeitung dieser Daten stellt eine große Herausforderung dar, kann aber zu einer wertvollen Ressource werden. „Data is the new raw material of the 21st century.“ (Nigel Shadbolt, Tim Berners-Lee)

1.3 Big Data Werkzeuge

Die Software Komponenten, die eingesetzt werden, um „Big Data“ zu analysieren bestehen häufig aus „Open Source“-Projekten, die auch von großen Firmen wie Facebook, Twitter oder Yahoo eingesetzt und unterstützt werden.

Als Grundlage einer „Big Data“-Plattform dient häufig Apache Hadoop [3], ein Java-Software-Framework, das die verteilte Verarbeitung von Daten auf hochskalierenden PC-Clustern ermöglicht.

Bestandteile von Hadoop sind:

- **Hadoop File System (HDFS)**: Fehlertolerantes verteiltes Dateisystem, dass darauf optimiert ist, einen hohen Datendurchsatz beim Lesen von großen Dateien und Datenkonsistenz beim Schreiben zu ermöglichen. Hohe Verfügbarkeit und Schutz gegen Datenverlust wird durch Replikation auf mehreren Rechner-Knoten gewährleistet.

- **Hadoop MapReduce**: System zur parallelen Verarbeitung von großen Datensätzen. Ein MapReduce Auftrag teilt die Datensätze in unabhängige Teile auf, die im „Map“-Verarbeitungsschritt auf verschiedenen Knoten parallel verarbeitet werden. Danach werden im „Reduce“ Ar-

beitsschritt die Ergebnisse gesammelt und zu einem Ergebnis zusammengefasst. Durch die Verteilung über HDFS kann eine hohe Verarbeitungsgeschwindigkeit erreicht werden, da nach dem Prinzip der Datenlokalität die Algorithmen zur Datenanalyse auf den Knoten ausgeführt werden, die die Datensätze vorhalten.

- **Apache Hive**: Data Warehouse System für Hadoop, das eine SQL ähnliche Sprache bietet, um ad-hoc Anfragen an ein Hadoop System zu stellen.
- **Apache Pig**: High-Level Sprache zur Analyse von Datensätzen, mit der verteilte und parallelisierte MapReduce Aufträge erstellt werden.
- **Apache Mahout**: Library mit Algorithmen für Data Mining und maschinelles Lernen, die gut skalierbar sind.

Das Open Source Projekt „Apache Spark“ [4] ist eine Alternative zu Hadoop, das eine modernere Entwicklung darstellt. Es baut ebenfalls auf der Grundlage HDFS auf, arbeitet im Gegensatz zu Hadoop aber mit einer „In-Memory“-Datenbanktechnik, die eine höhere Verarbeitungsgeschwindigkeit verspricht. Die Vorteile von „Spark“ liegen in der Verarbeitung und der Analyse von Stream-Daten. Spark ist in der Programmiersprache Scala entwickelt und bietet den Vorteil, dass Aufträge für die verteilte Analyse auch in Scala programmiert werden können. Einen weiteren Vorteil bietet das Caching der Daten, denn um Algorithmen der Datenanalyse zu verfeinern, stehen die Daten bei mehreren Durchläufen bereits im Speicher zur Verfügung.

1.4 Data Scientists

Für die Bewältigung von „Big-Data“ sind nicht nur neue Werkzeuge nötig, sondern auch neue Spezialisten, die diese einsetzen können. Es hat sich die Berufsbezeichnung des „Data Scientist“ für Spezialisten entwickelt, die ihre Fähigkeiten in Mathematik, Statistik, maschinellem Lernen und „Natural Language Processing“ benötigen und verbinden. Um aus „Big Data“-Datensätzen neue Erkenntnisse zu ziehen ist ein spielerischer Umgang mit den Daten und Algorithmen notwendig.
Mit „Kaggle“ [8] ist eine Data Science Community entstanden, die Unternehmen einen Marktplatz bietet, Problemstellungen und Datensätze zu veröffentlichen, für die die Community im Wettbewerb Lösungen entwickeln kann. Die besten Lösungen werden mit Preisgeldern belohnt. Ein Vorteil für die Teilnehmer sind die bereitgestellten Trainingsdaten und die Möglichkeit, eigene Lösungen mit anderen Teilnehmern vergleichen zu können. Ein Beispiel ist die Problemstellung von General Electric zur Flugrouten Optimierung basierend auf Wetter und Verkehrsdaten.

Aber nicht nur die Auswertung der Daten ist wichtig, auch die Visualisierung der Ergebnisse ist ein wichtiger Bestandteil der Analyse. Zusammenhänge komplexer Daten sind in Zahlenreihen und Tabellen nicht so leicht erfassbar, wie in Grafiken. Ergebnisse müssen daher attraktiv und verständlich aufbereitet werden, z.B. in Geovisualisierungen und Diagrammen. So lassen sich visuell Muster und neue Zusammenhänge erkennen.

1.5 Data Mining

Die Konzepte, die für das Auswerten der Datensätze nötig sind, stammen aus der Künstlichen Intelligenz und dem Bereich des maschinellen Lernen. Mit überwachtem maschinellen Lern-Algorithmen können Modelle gebildet werden, die Datensätze analysieren können, ohne alle Datensätze in den Speicher laden zu müssen.

- Regression wird eingesetzt um Werte einer Variable vorhersagen zu können, die über einen Zusammenhang basierend auf historischen Daten berechnet werden können.
- Klassifizierung: Eine vorgebene Anzahl von Klassen wird anhand eines Merkmals den Datensätzen zugeordnet. (z. B. Nearest Neighbour Classification)
- Clustering: Für eine vorgebene Anzahl von Clustern ermittelt der Algorithmus Gruppen, die eine möglichst hohe Ähnlichkeit der Merkmal Vektoren zueinander haben (z.B. K-Means Algorithmus).

1.6 Analyseprozess

Der Analyseprozess besteht in der Indentifizierung des Problems, dem Entwurf und der Sammlung der Datensätze, der Datenanalyse und der Aufbereitung und Visualisierung der Ergebnisse.

Die Datensätze müssen formatiert und in die Big Data Plattform, z.B. Hadoop geladen werden.
Die Analyse kann mit verschiedenen maschinellen-Lern-Algorithmen oder modellbasiert durchgeführt werden.
Die Ergebnisse müssen validiert werden und die Algorithmen und Modelle Schritt für Schritt verfeinert werden.

1.7 Forschung und Konferenzen

Die aktuellen Forschungsthemen in Big Data werden in den Gruppen ACM SIGKDD Knowledge Discovery and Data Mining [1] und ACM SIGMOD Management of Data [2] und auf Konferenzen wie BigMine [5] und ECML-PKDD [6] behandelt.

Auf der KDD 2013 Konferenz war eines der aktuellen Paper eine Studie über die Skalierung der Infrastruktur bei Twitter („Scaling Big Data Mining Infrastructure: The Twitter Experience“ [13]. Es gibt Einblicke in die Infrastruktur und Erfahrungen im Einsatz der Analyse-Tools mit Twitter-Daten. Es zeigt, dass die Werkzeuge noch nicht so weit sind, dass Mining-Algorithmen einfach eingesetzt werden können und viel Aufwand bei der Datenvorbereitung und ihrer Modelle nötig sind.

Der Einsatz von Data Mining und maschinellem Lernen bei dem Video-Ondemand Dienst Netflix wird in „Mining Large Streams of User Data for Personalized Recommendations“ [10] beschrieben. Hier werden Techniken dargestellt, mit denen Empfehlungen und Personalisierung des Webdienstes realisiert werden. Insbesondere wird diskutiert, ob für die Verbesserung der Empfehlungen mehr Daten oder bessere Modelle nötig sind.

„Predictive Analytics“ ist eine Technik, die durch die Verbindung eines Modells und dem Einfluss von aktuellen Datenströmen die Vorhersage von Fragestellungen zur Entscheidungsfindung ermöglicht. Dazu wird das Modell mit maschinellem Lernen und Data Mining auf historischen Datensätzen entwickelt.

Das Forschungsprojekt FuturICT [7] unter Leitung von Dirk Helbing, ETH Zürich ist eine Plattform für Simulation, Visualisierung und Partizipation von sozialen Systemen, die auf einem „Predictive Analytics“ Ansatz aufbaut. Das Ziel ist die Entwicklung einer Spiegelwelt zur Beantwortung von „Was wäre, wenn ...?“ Fragen. Es soll aus den folgenden Komponenten bestehen:

- Planetary Nervous System: weltweit erfasste Sensordaten füttern die Modelle mit Daten

- Living Earth Simulator: Modelle für Gesundheit, Energie, Infrastruktur, Verkehr, Finanzen sollen erstellt werden, die auch auf die erfassten Sensordaten zurückgreifen.

Das Projekt soll eine globale Plattform bieten, die über ein offenes Framework Bürgern, Unternehmen und Organisationen die Teilnahme ermöglicht, Daten zu beziehen aber auch neue Daten bereitzustellen.

1.8 Privatsphäre

Ein wichtiges Thema im Zusammenhang mit Big Data ist der Umgang mit der Privatsphäre. Professor Alex Pentland, der Direktor des MIT Human Dynamics Laboratory und des MIT Media Lab ist ein führender Wissenschaftler, der sich im Gebiet Reality Mining speziell mit der Auswertung und Nutzung von persönlichen Daten befasst. Er fordert einen „New Deal on Data“ [14], der die Rechte für den Datenbesitz eines Nutzers formuliert. Firmen sollen Daten nur verwalten und entfernen wenn der Nutzer es will (immer Opt-In). Der Datenbesitzer muss die volle Kontrolle über die Daten haben und die Möglichkeit erhalten, darüber zu bestimmen, wie die Firma die Daten verarbeitet. Der Nutzer soll das Recht besitzen, Daten wiederzuerlangen, zu löschen oder woanders zu verwenden.

Das Ziel ist die Nutzung von Big Data im Einklang mit der Privatsphäre, um den positiven Nutzen der Allgemeinheit zu ermöglichen.

1.9 Herausforderungen

Die optimale Infrastruktur der Analyse-Architektur für die Auswertung von Echtzeit-Daten ist noch nicht gefunden und die Anforderungen und die Menge der Datenströme steigt. Viele Algorithmen zum Data-Mining sind nicht einfach parallelisierbar und eignen sich nicht auf verteilten Systemen, daher

ist die Entwicklung neuer Techniken notwendig.

Um Speicherplatz zu sparen könnten Daten komprimiert gespeichert werden. Ansätze liegen darin, Datensätze zu komprimieren, in Bereichen, die nicht interessant sind, aber hochauflösendere Datensätze zu behalten und zu untersuchen, die relevanter und repräsentativer sind. Der Ansatz von Core-Sets könnte viel Speicherplatz einsparen [11].

1.10 Ausblick

Ich möchte mich weiter mit den Herausforderungen der verteilten Infrastruktur befassen und Data-Mining-Algorithmen untersuchen, die sich parallelisieren lassen.

Literaturverzeichnis

[1] ACM SIGKDD Knowledge Discovery and Data Mining. http://www.kdd.org/.

[2] ACM SIGMOD Management of Data . http://www.sigmod.org/.

[3] Apache Hadoop project . http://hadoop.apache.org/.

[4] Apache Spark . Lightning-fast cluster computing . https://spark.incubator.apache.org/.

[5] Big Data Mining, International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications . http://bigdata-mining.org/.

[6] European Conference on Machine Learning and Principles and Practice of Knowledge Discovery . http://www.ecmlpkdd2013.org/.

[7] FuturICT . http://www.futurict.eu/.

[8] Kaggle Inc., Data Scientist Community . http://www.kaggle.com/.

[9] Open Data Institute . http://theodi.org/.

[10] Amatriain, Xavier: *Mining Large Streams of User Data for Personalized Recommendations*. SIGKDD Explor. Newsl., 14(2):37--48, April 2013.

[11] Feldman, Dan, Melanie Schmidt und Christian Sohler: *Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering*. In: *SODA*, Seiten 1434--1453, 2013.

[12] Laney, Doug: *3D Data Management: Controlling Data Volume, Velocity, and Va- riety, Application Delivery Strategies*. 2001.

[13] Lin, Jimmy und Dmitriy Ryaboy: *Scaling Big Data Mining Infrastructure: The Twitter Experience*. SIGKDD Explor. Newsl., 14(2):6--19, April 2013.

[14] Pentland, Alex (MIT): *Reality Mining of Mobile Communications: Toward a New Deal on Data*. The Global Information Technology Report 2008-2009 World Economic Forum, 2009.