FernUniversität in Hagen

Fakultät für Mathematik und Informatik

Sommersemester 2015

Seminar
Big Data Management

Prof. Dr. Ralf Hartmut Güting

Fabio Valdés

Einführung

Angesichts der rasant anwachsenden weltweiten Informationsmenge, verursacht u.a. durch Standort- und Verbindungsdaten von Mobiltelefonen, Logdaten von IT-Systemen, Social-Media-Daten, Videoaufzeichnungen oder Kreditkartentransaktionen – im Jahr 2012 wurden laut IBM weltweit täglich 2,5 Trillionen Bytes produziert; auf Facebook werden wöchentlich 1,75 Mrd. Bilder hochgeladen; die weltweite Datenmenge soll bis 2020 auf über 100 Trilliarden Bytes angewachsen sein –, müssen Datenbanksysteme große Datenmengen schnell für komplexe Anfragen von immer mehr Nutzern bereitstellen. SQL-Datenbanken sind dafür nur eingeschränkt geeignet.

Big Data Management ist der Oberbegriff für neue Methoden und Technologien zur Erfassung, Speicherung und Analyse von Daten verschiedener Strukturen in nichtrelationalen verteilten Datenbanken. Zum Einstieg in das Thema seien die Werke [LLC+11] und [Cat10] empfohlen. Zudem sollten sich alle Seminarteilnehmer einen umfassenden Überblick über den aktullen Stand der Forschung zu MapReduce [DN14] verschaffen.

Zugang zu Quellen

Die im Literaturverzeichnis erwähnten Paper finden Sie auf der Internetseite des Seminars im Abschnitt Literatur.

Der Zugriff auf die Seite ist aus urheberrechtlichen Gründen auf Seminarteilnehmer beschränkt.

Themenauswahl

Bitte senden Sie bis zum 15.03.2015 eine Liste mit Ihren Prioritäten für die einzelnen Themen an `fabio.valdes@fernuni-hagen.de`. Die Liste soll 16 Einträge enthalten. Vergeben Sie Priorität 1 für das Thema, welches Sie am liebsten bearbeiten möchten. Markieren Sie das Thema, welches Sie auf keinen Fall bearbeiten möchten, mit Priorität 16. Dazwischenliegende Zahlen werden entsprechend gewertet. Alle Zahlen von 1 bis 16 sollen genau einmal vergeben werden.

Bei der Verteilung der Themen werden Ihre Prioritäten so weit wie möglich berücksichtigt.

Themen

Die Seminarthemen gliedern sich in fünf Vorträge über MapReduce, zwei Vorträge über Key-Value-Stores, zwei Vorträge zur Anfrageoptimierung, zwei Vorträge über aktuelle Datenbanksysteme von Google sowie fünf weitere Vorträge. Jeder Absatz beschreibt dabei genau ein Vortragsthema.

1 MapReduce – Skalierbare Datenauswertung

1.1 MapReduce

Das von Google entwickelte MapReduce Framework [DG04, DG08, DG10] dient dazu, die Verarbeitung sehr großer Datenmengen effizient auf zahlreiche Rechner zu verteilen. Dieser Vortrag beleuchtet die technischen Hintergründe des MapReduce-Verfahrens. MapReduce ist äußerst populär geworden, vor allem auch durch die frei verfügbare Implementierung Hadoop.

1.2 Google File System

Beim Google File System [GGL03] handelt es sich um ein verteiltes Dateisystem für datenintensive Anwendungen, welches die technische Grundlage für MapReduce liefert (analog zum Hadoop File System, welches grundlegend für Hadoop ist) und vor allem für Googles Websuche optimiert ist und verwendet wird. Die Aspekte Fehlertoleranz und Korrektheit sollten in diesem Vortrag eine wichtige Rolle spielen.

1.3 HadoopDB

HadoopDB [ABPA$^+$09] ist ein hybrides System, welches die Vorteile traditioneller paralleler Datenbanksysteme mit denen des MapReduce-Konzepts vereinigt. Es bietet eine komplette Open-Source-Lösung für die verteilte Speicherung und parallele Auswertung großer Datenmengen durch MapReduce-Techniken in Rechnerclustern. Der Vortrag soll zudem zwei reale Anwendungen von HadoopDB [ABH$^+$10] sowie Möglichkeiten zur Optimierung [BC14] thematisieren.

1.4 Hive

Basierend auf Hadoop wurde von Facebook das System Hive [TSJ$^+$09, TSJ$^+$10] entwickelt. Hive bietet u.a. zusätzlich eine deklarative SQL-ähnliche Abfragesprache namens HiveQL, deren Befehle in MapReduce-Kommandos übersetzt und vom zugrundeliegenden Hadoop-System ausgeführt werden.

1.5 Pig

Das von Yahoo! entwickelte System Pig [ORS$^+$08, GNC$^+$09] basiert ebenfalls auf Hadoop und ist frei verfügbar. In der höheren Sprache Pig Latin lassen sich Programme entwickeln, die MapReduce-Befehle in Hadoop ausführen. Es sollen zudem Ansätze zur Optimierung [GDN13] beleuchtet werden.

2 Key-Value-Stores und Verwandte – Skalierbare dynamische Datenbanken

2.1 Dynamo

Dynamo [DHJ+07] ist ein von Amazon eingeführtes Datenbanksystem auf Basis von Key-Value-Stores, das für die Zuverlässigkeit der Amazon-Internetplattform verantwortlich ist. Um höchstmögliche Verfügbarkeit zu gewährleisten, die für Amazon maßgeblich ist, verzichtet Dynamo in bestimmten Fehlerszenarien auf die Aufrechterhaltung der Datenkonsistenz.

2.2 Bigtable

Das verteilte Datenbanksystem Bigtable [CDG+08] wurde von Google entwickelt, um riesige Datenmengen auf tausende Standardrechner zu verteilen. Eine Bigtable ist eine verteilte multidimensionale Tabelle, in der jeder Wert durch einen Zeilenschlüssel, einen Spaltenschlüssel sowie einen Zeitstempel eindeutig identifiziert wird. Diese Technologie wird von Google selbst u.a. für die Suchmaschine, Google Earth und Google Finance verwendet.

3 Anfrageoptimierung für MapReduce

3.1 AQUA & YSmart

Die Systeme Hive und Pig können zwar SQL-ähnliche Anfragen in MapReduce-Befehle übersetzen, legen dabei jedoch keinen großen Wert auf Optimierung. Dieser Vortrag soll die beiden Anfrageoptimierer AQUA [WLMO11] und YSmart [LLH+11] thematisieren, die neben der Übersetzung auch eine effiziente Ausführung garantieren.

3.2 Jaql & Tenzing

Jaql [BEG+11] ist eine von IBM entwickelte erweiterbare deklarative Skriptsprache zur Analyse großer halbstrukturierter Datenmengen, die MapReduce verwendet. Ebenfalls im Jahr 2011 hat Google das Anfragesystem Tenzing [CLL+11] präsentiert, das SQL- (und weitere) Anfragen auf Basis von MapReduce ermöglicht. Beide Systeme beinhalten Lösungen zur Anfrageoptimierung.

4 Spanner und F1 – Googles verteilte Datenbanksysteme

4.1 Spanner

Nach über fünf Jahren Entwicklungszeit wurde im Jahr 2012 das Datenbanksystem Spanner [CDE+12, CDE+13] von Google vorgestellt. Es kombiniert Stärken traditioneller Datenbanksysteme wie Transaktionen und eine SQL-ähnliche Anfragesprache mit Aspekten wie Skalierbarkeit, Fehlertoleranz und weltweiter Verteilung und kann somit als Weiterentwicklung gegenüber Bigtable angesehen werden.

4.2 F1

Basierend auf Spanner wurde – ebenfalls von Google – das verteilte Datenbanksystem F1 [SVS+13] entwickelt. Es bietet ähnliche Vorteile wie Spanner und kommt bei Googles Werbesystem AdWords zum Einsatz.

5 Weitere Themen

5.1 Optimierung von Hadoop durch Indizierung

Gegenüber traditionellen Datenbanksystemen hat Hadoop oft Leistungsnachteile. Daher wurde das System Hadoop++ [DQJ+10] eingeführt, das auf Hadoop aufsetzt und neue Indizierungs- und Joinverfahren verwendet. Zwei Jahre später ist die Weiterentwicklung HAIL [DQR+12] erschienen. Im Vergleich zu Hadoop++ werden in HAIL lange Upload- und Indizierungszeiten vermieden.

5.2 Direktes Arbeiten auf Dateien

Da die zu analysierenden Datenmengen immer weiter anwachsen, versucht man zunehmend auf den Import der Daten in eine Datenbank zu verzichten und direkt auf die zugrundeliegenden Dateien zuzugreifen. Entsprechende Ansätze sind NoDB [ABB+12], Invisible Loading [AAS13] und SDS/Q [BWB+14].

5.3 Effiziente Algorithmen in MapReduce

Die Autoren von [KSV10] vergleichen MapReduce mit dem System PRAM (Parallel Random Access Machine; Maschinenmodell zur Analyse paralleler Algorithmen) und zeigen, dass eine große Klasse von PRAM-Algorithmen effizient mit MapReduce ausgeführt werden kann. In [TLX13] werden vier Kriterien für eine ideale Ausführung von MapReduce-Jobs festgelegt. Anhand dieser Kriterien werden minimale Algorithmen für typische Aufgaben wie Sortieren, Gruppieren oder Semi-Join vorgestellt.

5.4 MapReduce & Geodaten

Große Mengen geographischer bzw. räumlicher Daten zu analysieren, stellt wegen der Mehrdimensionalität und der teilweise hohen Komplexität von Anfragen eine Herausforderung dar. Im Jahr 2013 wurde Hadoop-GIS [AWV+13], ein skalierbares System für Anfragen auf solchen Daten unter Hadoop, veröffentlicht. Die Autoren von [WPAH14] haben ein Verfahren entwickelt, das räumliche Daten im Hadoop File System indiziert und bestimmte Analysen (Bereichsanfrage, Nächste-Nachbarn-Suche) auf ihnen durchführt. Schließlich wurde mit SpatialHadoop [EM13] eine umfassende Erweiterung von Hadoop präsentiert, die für die Analyse räumlicher Daten vorgesehen ist.

5.5 Stratosphere

Stratosphere [ABE+14] ist ein System zur Analyse großer Datenmengen an der Schwelle zwischen MapReduce und relationalen Datenbanken. Es bietet u.a. eine deklarative Anfragesprache, benutzerdefinierte Funktionen sowie eine automatische Parallelisierung und Optimierung.

Verwendete Technologien & theoretische Grundlagen

Der Vollständigkeit halber sei auf folgende Werke hingewiesen, die als Grundlage für einige der vorgestellten Themen dienen.

Vector Clocks

Das Konzept zeitlicher (Teil-)Ordnung von Ereignissen in verteilten Systemen wird in [Lam78] vorgestellt. Außerdem wird ein verteilter Algorithmus präsentiert, der ein System logischer Uhren synchronisiert.

Consistent Hashing

Mit Hilfe von konsistenten Hashfunktionen wird in [KLL+97] das Problem überlasteter Netzwerkknoten behandelt.

Multiversion Concurrency Control

Die Autoren von [BG83] entwickeln eine theoretische Grundlage, um die Korrektheit von Synchronisationsalgorithmen für konkurrierende Datenbankzugriffe zu analysieren.

Literatur

[AAS13] Azza Abouzied, Daniel J. Abadi, and Avi Silberschatz. Invisible loading: access-driven data transfer from raw files into database systems. In *Joint 2013 EDBT/ICDT Conferences, EDBT '13 Proceedings, Genoa, Italy, March 18-22, 2013*, pages 1–10, 2013.

[ABB+12] Ioannis Alagiannis, Renata Borovica, Miguel Branco, Stratos Idreos, and Anastasia Ailamaki. Nodb: efficient query execution on raw data files. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 241–252, 2012.

[ABE+14] Alexander Alexandrov, Rico Bergmann, Stephan Ewen, Johann-Christoph Freytag, Fabian Hueske, Arvid Heise, Odej Kao, Marcus Leich, Ulf Leser, Volker Markl, Felix Naumann, Mathias Peters, Astrid Rheinländer, Matthias J. Sax, Sebastian Schelter, Mareike Höger, Kostas Tzoumas, and Daniel Warneke. The stratosphere platform for big data analytics. *VLDB J.*, 23(6):939–964, 2014.

[ABH+10] Azza Abouzied, Kamil Bajda-Pawlikowski, Jiewen Huang, Daniel J. Abadi, and Avi Silberschatz. Hadoopdb in action: building real world applications. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*, pages 1111–1114, 2010.

[ABPA+09] Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel J. Abadi, Alexander Rasin, and Avi Silberschatz. Hadoopdb: An architectural hybrid of mapreduce and dbms technologies for analytical workloads. *PVLDB*, 2(1):922–933, 2009.

[AWV+13] Ablimit Aji, Fusheng Wang, Hoang Vo, Rubao Lee, Qiaoling Liu, Xiaodong Zhang, and Joel H. Saltz. Hadoop-gis: A high performance spatial data warehousing system over mapreduce. *PVLDB*, 6(11):1009–1020, 2013.

[BC14] Cherif A. A. Bissiriou and Habiba Chaoui. Big data analysis and query optimization improve hadoopdb performance. In *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014*, pages 1–4, 2014.

[BEG+11] Kevin S. Beyer, Vuk Ercegovac, Rainer Gemulla, Andrey Balmin, Mohamed Y. Eltabakh, Carl-Christian Kanne, Fatma Özcan, and Eugene J. Shekita. Jaql: A scripting language for large scale semistructured data analysis. *PVLDB*, 4(12):1272–1283, 2011.

[BG83] Philip A. Bernstein and Nathan Goodman. Multiversion concurrency control - theory and algorithms. *ACM Trans. Database Syst.*, 8(4):465–483, 1983.

[BWB+14] Spyros Blanas, Kesheng Wu, Surendra Byna, Bin Dong, and Arie Shoshani. Parallel data analysis directly on scientific file formats. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, pages 385–396, 2014.

[Cat10] Rick Cattell. Scalable sql and nosql data stores. *SIGMOD Record*, 39(4):12–27, 2010.

[CDE+12] James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson C. Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. Spanner: Google's globally distributed database. Website, 2012. Presentation Slides.

[CDE+13] James C. Corbett, Jeffrey Dean, Michael Epstein, Andrew Fikes, Christopher Frost, J. J. Furman, Sanjay Ghemawat, Andrey Gubarev, Christopher Heiser, Peter Hochschild, Wilson C. Hsieh, Sebastian Kanthak, Eugene Kogan, Hongyi Li, Alexander Lloyd, Sergey Melnik, David Mwaura, David Nagle, Sean Quinlan, Rajesh Rao, Lindsay Rolig, Yasushi Saito, Michal Szymaniak, Christopher Taylor, Ruth Wang, and Dale Woodford. Spanner: Google's globally distributed database. *ACM Trans. Comput. Syst.*, 31(3):8, 2013.

[CDG+08] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E. Gruber. Bigtable: A distributed storage system for structured data. *ACM Trans. Comput. Syst.*, 26(2):4:1–4:26, June 2008.

[CLL+11] Biswapesh Chattopadhyay, Liang Lin, Weiran Liu, Sagar Mittal, Prathyusha Aragonda, Vera Lychagina, Younghee Kwon, and Michael Wong. Tenzing A SQL implementation on the mapreduce framework. *PVLDB*, 4(12):1318–1327, 2011.

[DG04] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. In *Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association.

[DG08] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.

[DG10] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: a flexible data processing tool. *Commun. ACM*, 53(1):72–77, 2010.

[DHJ+07] Giuseppe DeCandia, Deniz Hastorun, Madan Jampani, Gunavardhan Kakulapati, Avinash Lakshman, Alex Pilchin, Swaminathan Sivasubramanian, Peter Vosshall, and Werner Vogels. Dynamo: amazon's highly available key-value store. *SIGOPS Oper. Syst. Rev.*, 41(6):205–220, October 2007.

[DN14] Christos Doulkeridis and Kjetil Nørvåg. A survey of large-scale analytical query processing in mapreduce. *VLDB J.*, 23(3):355–380, 2014.

[DQJ+10] Jens Dittrich, Jorge-Arnulfo Quiané-Ruiz, Alekh Jindal, Yagiz Kargin, Vinay Setty, and Jörg Schad. Hadoop++: Making a yellow elephant run like a cheetah (without it even noticing). *PVLDB*, 3(1):518–529, 2010.

[DQR+12] Jens Dittrich, Jorge-Arnulfo Quiané-Ruiz, Stefan Richter, Stefan Schuh, Alekh Jindal, and Jörg Schad. Only aggressive elephants are fast elephants. *PVLDB*, 5(11):1591–1602, 2012.

[EM13] Ahmed Eldawy and Mohamed F. Mokbel. A demonstration of spatialhadoop: An efficient mapreduce framework for spatial data. *PVLDB*, 6(12):1230–1233, 2013.

[GDN13] Alan Gates, Jianyong Dai, and Thejas Nair. Apache pig's optimizer. *IEEE Data Eng. Bull.*, 36(1):34–45, 2013.

[GGL03] Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. *SIGOPS Oper. Syst. Rev.*, 37(5):29–43, October 2003.

[GNC+09] Alan Gates, Olga Natkovich, Shubham Chopra, Pradeep Kamath, Shravan Narayanam, Christopher Olston, Benjamin Reed, Santhosh Srinivasan, and Utkarsh Srivastava. Building a highlevel dataflow system on top of mapreduce: The pig experience. *PVLDB*, 2(2):1414–1425, 2009.

[KLL+97] David R. Karger, Eric Lehman, Frank Thomson Leighton, Rina Panigrahy, Matthew S. Levine, and Daniel Lewin. Consistent hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web. In *STOC*, pages 654–663, 1997.

[KSV10] Howard J. Karloff, Siddharth Suri, and Sergei Vassilvitskii. A model of computation for mapreduce. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pages 938–948, 2010.

[Lam78] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7):558–565, 1978.

[LLC+11] Kyong-Ha Lee, Yoon-Joon Lee, Hyunsik Choi, Yon Dohn Chung, and Bongki Moon. Parallel data processing with mapreduce: a survey. *SIGMOD Record*, 40(4):11–20, 2011.

[LLH+11] Rubao Lee, Tian Luo, Yin Huai, Fusheng Wang, Yongqiang He, and Xiaodong Zhang. Ysmart: Yet another sql-to-mapreduce translator. In *2011 International Conference on Distributed Computing Systems, ICDCS 2011, Minneapolis, Minnesota, USA, June 20-24, 2011*, pages 25–36, 2011.

[ORS+08] Christopher Olston, Benjamin Reed, Utkarsh Srivastava, Ravi Kumar, and Andrew Tomkins. Pig latin: a not-so-foreign language for data processing. In *SIGMOD Conference*, pages 1099–1110, 2008.

[SVS+13] Jeff Shute, Radek Vingralek, Bart Samwel, Ben Handy, Chad Whipkey, Eric Rollins, Mircea Oancea, Kyle Littlefield, David Menestrina, Stephan Ellner, John Cieslewicz, Ian Rae, Traian Stancescu, and Himani Apte. F1: A distributed SQL database that scales. *PVLDB*, 6(11):1068–1079, 2013.

[TLX13] Yufei Tao, Wenqing Lin, and Xiaokui Xiao. Minimal mapreduce algorithms. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2013, New York, NY, USA, June 22-27, 2013*, pages 529–540, 2013.

[TSJ+09] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. Hive- a warehousing solution over a map-reduce framework. In *IN VLDB '09: PROCEEDINGS OF THE VLDB ENDOWMENT*, pages 1626–1629, 2009.

[TSJ+10] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Ning Zhang, Suresh Anthony, Hao Liu, and Raghotham Murthy. Hive - a petabyte scale data warehouse using hadoop. In *ICDE*, pages 996–1005, 2010.

[WLMO11] Sai Wu, Feng Li, Sharad Mehrotra, and Beng Chin Ooi. Query optimization for massively parallel data processing. In *ACM Symposium on Cloud Computing in conjunction with SOSP 2011, SOCC '11, Cascais, Portugal, October 26-28, 2011*, page 12, 2011.

[WPAH14] Randall T. Whitman, Michael B. Park, Sarah M. Ambrose, and Erik G. Hoel. Spatial indexing and analytics on hadoop. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas/Fort Worth, TX, USA, November 4-7, 2014*, pages 73–82, 2014.