Datenbank Spektrum (2013) 13:1–3
DOI 10.1007/s13222-013-0116-z

EDITORIAL

Editorial

Theo Härder

Online publiziert: 1. Februar 2013

1 Schwerpunktthema: MapReduce Programming Model

MapReduce ist ein Programmiermodell für die parallele Verarbeitung großer Datenmengen auf einer Vielzahl von Rechnern, das im Jahr 2004 durch den Beitrag *MapReduce: Simplified Data Processing on Large Clusters* von den Google-Mitarbeitern Jeffrey Dean und Sanjay Ghemawat auf dem 6th Symposium on Operating System Design and Implementation (OSDI 2004) vorgestellt wurde. Seither hat dieser Aufsatz eine Lawine von Forschungsansätzen und Systementwicklungen zur Analyse und Verarbeitung von *Big Data* ausgelöst.

Das MapReduce-Programmiermodell ermöglicht die skalierbare Analyse und Transformation großer, verteilter und heterogener Datenmengen. Um die Entwicklung spezifischer MapReduce-Anwendungen zu vereinfachen und zu beschleunigen, stellt eine MapReduce-Implementierung ein Framework zur Verfügung, das sich um die Aspekte der Datenverteilung und um das Scheduling paralleler Rechenaufgaben kümmert. Im Wesentlichen muss der Benutzer dieses Framework nur vervollständigen durch Spezifikation einer *Map*-Funktion, die aus einer Liste von Schlüssel/Wert-Paaren als Zwischenergebnis eine neue Liste von Schlüssel/Wert-Paaren erzeugt, und einer *Reduce*-Funktion, die alle Sätze mit demselben Schüssel in einer Zwischenergebnisliste gruppiert, alle Werte solcher Gruppen mischt und durch Berechnungen „reduziert“.

Mit diesem Ansatz der funktionalen Programmierung können Programme automatisch hohe Parallelitätsgrade nutzen und dadurch in perfekter Weise skalieren. Da sich das MapReduce-Programmiermodell auch für ein breites Spektrum verschiedenartiger Berechnungsprobleme eignet, war es als Folge dieser Eigenschaften in den letzten Jahren bei der Verarbeitung von Big Data in vielen verschiedenen Anwendungsgebieten enorm erfolgreich.

In diesem Themenheft beschreiben vier Beiträge interessante Fragestellungen im Kontext von MapReduce und der Analyse großer Datenmengen. Sie zeigen auch die vielfältigen Einsatzmöglichkeiten des Programmiermodells und seine Variationsbreite bei Anwendungen in verschiedenartigen Bereichen.

Im ersten Beitrag *Compilation of Query Languages into MapReduce* von Caetano Sauer und Theo Härder wird ein Brückenschlag von SQL zu Datenbankprogrammiersprachen für *Big Data* versucht. Es wird insbesondere die Frage aufgeworfen, warum SQL in vielen Belangen der Analyse von *Big Data* zu restriktiv ist. Beschränkungen von SQL, die in diesem Kontext besonders auffallen, führten zur Entwicklung von besser geeigneten Datenprogrammiersprachen, von denen PigLatin, HiveQL, Jaql und XQuery näher untersucht werden. Wichtige Spracheigenschaften bilden eine Art Wunschliste für die Verarbeitung von Big Data, nach der dann qualitativ bewertet wird, wie gut diese Sprachen die mangelnde Flexibilität und die Einschränkungen von SQL überwinden. Basierend auf dieser Wunschliste von Spracheigenschaften wird in abstrakter Weise die Kompilation in das MapReduce-Programmiermodell beschrieben, die deutlich werden lässt, das der Übersetzungsprozess für alle vier Sprachen im Wesentlichen gleich ist. Einfache Generalisierungen des ursprünglichen MapReduce-Programmiermodells erlauben die Wiederbenutzung der bewährten Techniken zur Anfrageverarbeitung, die dann die Generierung von optimierten Anfrageausführungsplänen für MapReduce-Analysen erleichtern.

T. Härder (✉)
AG Datenbanken und Informationssysteme, TU Kaiserslautern,
67663 Kaiserslautern, Deutschland
e-mail: haerder@cs.uni-kl.de

Der zweite Beitrag *Efficient OR Hadoop: Why not both?* von Jens Dittrich, Stefan Richter und Stefan Schuh widmet sich der Anfrageoptimierung im Kontext von Big Data und beschreibt verschiedene Ansätze, die in der Forschungsgruppe Informationssysteme der Universität des Saarlandes mit dem Ziel verfolgt wurden, Hadoop effizienter zu machen. Das Projekt Hadoop++ konzentrierte sich, ohne den Code von Hadoop Distributed File System (HDFS) und Hadoop MapReduce zu ändern, auf die Flexibilisierung der einzelnen Schritte (Pipeline) der Anfrageverarbeitung in Hadoop. Dabei konnte eine Reduktion der Laufzeiten von bis zu einem Faktor 20 erzielt werden. Im Projekt Trojan Layouts wurden verschiedene Daten-Layouts (tupelweise, spaltenweise, Partition Attributes Across (PAX)) im Kontext der MapReduce-Verarbeitung untersucht mit dem Ziel, ein geeigneteres Daten-Layout von Hadoop für die analytische Anfrageverarbeitung zu finden. Mit dieser Optimierungsmaßnahme konnte gezeigt werden, dass sich damit die Laufzeiten der Anfrageverarbeitung im Vergleich zu Tupel- und PAX-Layouts um bis zu einem Faktor 5 verbessern lassen. Im Projekt HAIL (Hadoop Aggressive Indexing Library) wurde die Nutzung von verschiedenen Indexstrukturen mit Clusterbildung evaluiert. Auch dabei konnten enorme Leistungsgewinne gegenüber der Anfrageverarbeitung in Hadoop und Hadoop++ nachgewiesen werden.

Im dritten Beitrag dieses Heftes beschreiben Lars Kolb und Erhard Rahm unter dem Titel *Parallel Entity Resolution with Dedoop* ein Tool zur Identifikation von Entities, auch als Entity Resolution (ER) bezeichnet, in Cloud-Infrastrukturen auf der Basis von Hadoop. Besonders bei heterogenen Datensammlungen ist das Erkennen von Duplikaten bei Objekten (z. B. Autoren, Kunden oder Produkten), die mit ähnlichen Strukturen und Werten repräsentiert sind, zur Absicherung der Verarbeitungsqualität von großer Wichtigkeit. In herkömmlichen Verfahren müssen Entities paarweise mithilfe verschiedenartiger Ähnlichkeitsmaße in aufwändiger Weise ausgewertet werden, um möglichst genaue Vergleichsentscheidungen zu erzielen. Zur Verbesserung der Effizienz wird normalerweise der Suchraum durch Einsatz sogenannter Blocking-Techniken verkleinert. Das von den Autoren entwickelte System Dedoop besitzt eine umfangreiche Bibliothek von Blocking- und Matching-Verfahren und setzt trainingsbasierte Methoden des Machine Learning ein, um für eine gegebene Anwendung geeignete ER-Strategien zu konfigurieren. Nach Auswahl solcher Verfahren werden diese automatisch in MapReduce-Programme übersetzt, die dann parallel auf verschiedenen Hadoop-Clustern ausgeführt werden können. Eine Verbesserung der Leistung erzielt Dedoop durch den Einsatz von Multi-Pass Blocking und effektiven Methoden zur Lastbalancierung. Die Vielseitigkeit und Leistungsfähigkeit von Dedoop wird durch die vergleichende Auswertung verschiedener ER-Strategien auf realen Datensammlungen nachgewiesen.

Schließlich beschäftigt sich der Aufsatz *Inkrementelle Neuberechnungen in MapReduce* von Johannes Schildgen, Thomas Jörg und Stefan Deßloch mit einer für die Datenverwaltung wichtigen MapReduce-Anwendung. Es wird zunächst gezeigt, dass sich der Ansatz von MapReduce als Lösungskonzept für ein breites Spektrum von Berechnungsproblemen eignet – z. B. die Erstellung von Worthistogrammen für einen Text, die Ableitung eines invertierten Link-Graphen aus einer Sammlung von Web-Seiten oder die Berechnung von Freundesfreund-Beziehungen in Sozialen Netzwerken. Solche MapReduce-Berechnungen erfolgen typischerweise auf großen Datensammlungen, die normalerweise in verteilten Dateisystemen vorliegen. Ändern sich diese Datensammlungen, werden die Berechnungsergebnisse mit der Häufigkeit der Aktualisierungen ungenauer und müssen von Zeit zu Zeit neu erstellt werden. Eine vollständige Neuberechnung ist dabei in der Regel keine effiziente Lösung. Deshalb schlägt der Beitrag einen Ansatz zur inkrementellen Neuberechnung in MapReduce vor, der auf den Ideen und Konzepten zur inkrementellen Wartung materialisierter Sichten in relationalen Datenbanksystemen basiert. Dazu wird das auf Map-Reduce basierende Marimba-Framework vorgestellt, das der einfachen Entwicklung von MapReduce-Programmen dient, die nach Änderungen im Datenbestand nur inkrementelle Neuberechnungen vornehmen und dadurch eine vollständige Wiederholung des MapReduce-Ablaufs vermeiden. Die Entwicklung solcher inkrementellen MapReduce-Programme wird für mehrere Anwendungen gezeigt; für zwei verschiedene Strategien wird ihr Leistungsverhalten abhängig vom Änderungsgrad des Datenbestandes bestimmt und mit dem der vollständigen Neuberechnung verglichen.

Diese Schwerpunktbeiträge werden ergänzt durch einen Fachbeitrag *Towards Integrated Data Analytics: Time Series Forecasting in DBMS* von Ulrike Fischer, Lars Dannecker, Laurynas Siksnys, Frank Rosenthal, Matthias Boehm und Wolfgang Lehner. Integrierte statistische Methoden gewinnen für Datenbankanwendungen immer mehr an Bedeutung, um mit den wachsenden Datenvolumina und der steigenden algorithmischen Komplexität bei der Datenanalyse fertig zu werden. Die Autoren plädieren für eine Tiefenintegration von ausgefeilten statistischen Methoden in Datenbankverwaltungssystemen. Speziell wird in diesem Beitrag die Integration der Zeitreihenvorhersage diskutiert, die in Entscheidungsfindungsprozessen in vielen Bereichen eine große Rolle spielt.

Weiterhin finden Sie unter der Rubrik „Datenbankgruppen vorgestellt“ einen Beitrag von Thomas Seidl zu *Datenmanagement und -exploration an der RWTH Aachen*. Die Rubrik „Dissertationen“ ist in diesem Heft mit acht Kurzfassungen von Dissertationen erfreulicherweise recht umfangreich.

In der Rubrik „Community“ geben Alfons Kemper, Tobias Mühlbauer, Thomas Neumann, Angelika Reiser und

Wolf Rödiger einen *Bericht vom Herbsttreffen der GI-Fachgruppe Datenbanksysteme* an der TU München. Das Treffen zum Thema „Scalabale Analytics“ hatte mit über 80 Teilnehmern eine erfreuliche Resonanz und stand unter dem Motto „Industry meets Academia“. Weiterhin enthält die Rubrik „Community“ einen Beitrag *News* mit aktuellen Informationen.

2 Künftige Schwerpunktthemen

2.1 RDF Data Management

Nowadays, more and more data is modeled and managed by means of the W3C Resource Description Framework (RDF) and queried by the W3C SPARQL Protocol and RDF Query Language (SPARQL). RDF is commonly known as a conceptual data model for structured information that was standardized to become a key enabler of the Semantic Web to express metadata on the web. It supports relationships between resources as first-class citizens, provides modeling flexibility towards any kind of schema, and is even usable without a schema at all. Furthermore, RDF allows to collect data starting with very little schema information and refining the schema later, as required. This flexibility led to a wide adoption in many other application domains including life sciences, multifaceted data integration, as well as community-based data collection, and large knowledge bases like DBpedia.

This special issue of the „Datenbank-Spektrum“ aims to provide an overview of recent developments, challenges, and future directions in the field of RDF technologies and applications.

Topics of interest include (but are not limited to)

- RDF data management
- RDF access over the Web
- Querying and query optimization over RDF data – especialy when accessed over the Web
- Applications and usage scenarios
- Case studies and experience reports

Guest editors:
Johann-Christoph Freytag, Humboldt-Universität zu Berlin, freytag@dbis.informatik.hu-berlin.de
Bernhard Mitschang, University of Stuttgart, Bernhard.Mitschang@ipvs.uni-stuttgart.de

2.2 Best Workshop Papers of BTW 2013

This special issue of the „Datenbank-Spektrum“ is dedicated to the Best Papers of the Workshops running at the BTW 2013 at the TU Magdeburg. The selected Workshop contributions should be extended to match the format of regular DASP papers.

Paper format: 8–10 pages, double column

Selection of the Best Papers by the Workshop chairs and the guest editor: April 15th, 2013

Guest editor:
Theo Härder, University of Kaiserslautern, haerder@cs.uni-kl.de

Deadline for submissions: June 1st, 2013

2.3 Information Retrieval

The amount of available information has increased dramatically in the last decades. At the same time, the way in which this information is presented has changed rapidly: Multimedia data such as audio, images, and video complements or even replaces textual information, user-generated content from blogs or social networks replaces static Web sites, and highly dynamic content such as tweets is published in real-time. Information Retrieval methods allow to quickly find relevant pieces of information for a possibly complex information need from this huge pile of data.

This special issue of the Datenbank-Spektrum aims to provide an overview of recent developments, challenges, and future directions in the field of Information Retrieval technologies and applications.

Topics of interest include (but are not limited to)

- Crawling, Indexing, Query Processing
- Information Extraction and Mining
- Interactive Information Retrieval
- Personalized and Context-Aware Retrieval
- Structured and Semantic Search
- Evaluation and Benchmarking
- Archiving and Time-Aware Retrieval Models
- Enterprise Search
- Realtime Search: Streams, Tweets, Social Networks
- Multimedia Retrieval

Paper format: 8–10 pages, double column

Notice of intent for a contribution: July 15th, 2013

Guest editors:
Ralf Schenkel, Max-Planck-Institut für Informatik, schenkel@mpi-inf.mpg.de
Christa Womser-Hacker, Universität Hildesheim, womser@uni-hildesheim.de

Deadline for submissions: October 1st, 2013