Datenbank Spektrum (2012) 12:157–159
DOI 10.1007/s13222-012-0104-8

EDITORIAL

Editorial

Michael Gertz · Wolfgang Müller

Online publiziert: 5. Oktober 2012

1 Schwerpunktthema: Scientific Data Management

Spricht man heutzutage von Herausforderungen an die Verwaltung und Analyse großer Datenmengen, so bezieht man sich dabei meist auf Anwendungen im Bereich des eCommerce sowie neuerdings insbesondere auf Analysen sozialer Netzwerke. Dieser Fokus ist sicherlich gut begründet, da hier typischerweise große internationale Firmen wie Facebook, Twitter, eBay oder Amazon mit Geschäftsmodellen im Vordergrund stehen, bei denen der Verkauf von Produkten sowie die Pflege und Analyse von Kunden- und Verkaufsdaten die Geschäftsgrundlage bilden. Ähnliches gilt für die Telekommunikationsindustrie, bei der in großen Data Warehouses Informationen zu Anwendern und deren Nutzung von Services verwaltet und zur Verbesserung jener Services analysiert werden. Schwerpunkte traditioneller Forschung und Entwicklung im Bereich Datenbanken lassen sich generell in den oben genannten Bereichen der Verwaltung und Analyse von Geschäftsdaten ansiedeln.

Der Menge an geschäftsorientierten Daten steht aber eine noch größere Menge an wissenschaftlichen Daten gegenüber, der bisher weniger Aufmerksamkeit im Rahmen der Mainstream-Datenbankforschung und -entwicklung gewidmet wurde. Gerade in den letzten Jahren haben die Naturwissenschaften immense Fortschritte in der Instrumentierung von Experimenten, Simulationen und Beobachtungen erfahren. Dies betrifft nahezu alle Bereiche in den Naturwissenschaften. Hierzu gehören u.a. die Physik, insbesondere die Astrophysik, Kosmologie und Teilchenphysik, die Biologie mit Schwerpunkten in der Genetik und Molekularbiologie sowie die in den Geowissenschaften angesiedelten Umweltwissenschaften und die Klimaforschung. In der Biologie nimmt zum Beispiel die Fähigkeit, mit Experimenten Daten zu generieren, schneller zu als die Rechenleistung zur Verarbeitung der Daten, was insbesondere bei der Gensequenzierung ein großes Bottleneck darstellt.

Aus Datensicht handelt es sich bei wissenschaftlichen Daten nicht um einfache transaktionale Daten, sondern um Daten, die typischerweise sehr heterogen und komplex sind und multiple Skalen beschreiben. Traditionelle Datenbanktechniken sind hierdurch teilweise nicht einfach auf diese Anwendungsbereiche zu übertragen. Die Datenintegration ist ein wichtiges Thema, das beispielsweise durch Scientific Workflows angegangen wird. Gleichzeitig ergeben sich neue Optimierungsmöglichkeiten durch die gegenüber universellen Datenbanken veränderte Domäne, denn neue Zugriffsmuster und andere vorherrschende konzeptuelle Datenstrukturen ermöglichen neue Optimierungen.

Die effiziente Verwaltung, Speicherung, Suche und Analyse wissenschaftlicher Daten stellt eine immense Herausforderung an diese und verwandte Bereiche der Naturwissenschaften dar. Wie kann man effektiv neues Wissen aus den Daten ableiten? Wo stoßen aktuelle Systeme an ihre Grenzen? Wo bieten sich neue und interessante Themen für die Datenbankforschung?

In unserem Themenheft werden in vier Artikeln interessante Themenstellungen angegangen, die einen Eindruck von der Vielfältigkeit der obigen Herausforderungen geben.

Der erste Beitrag *Data Management Challenges in Next Generation Sequencing* von Sebastian Wandelt, Astrid Rheinländer, Marc Bux, Lisa Thalheim, Berit Haldemann und Ulf Leser widmet sich dem Next Generation Sequen-

M. Gertz (✉)
Heidelberg University, Heidelberg, Deutschland
e-mail: gertz@informatik.uni-heidelberg.de

W. Müller
HITS gGmbH, Heidelberg, Deutschland
e-mail: wolfgang.mueller@h-its.org

cing, dem „Big Data“-Thema in der Biologie überhaupt. Häufig ist zu hören, dass wir uns auf einen Zustand zubewegen, in dem die Datenverarbeitung von Gensequenzen (technisch gesehen sind dies lange Strings) teurer wird als ihre eigentliche Gewinnung. Der Artikel gibt einen Überblick über die Herausforderungen sowie die wichtigsten Lösungsansätze.

Der zweite Beitrag *Handling Big Data in Astronomy and Astrophysics: Rich Structured Queries on Replicated Cloud Data with XtreemFS* von Harry Enke, Adrian Partl, Alexander Reinefeld und Florian Schintke befasst sich mit der Anfragebearbeitung in verteilt vorliegenden Datensammlungen. Das Anwendungsgebiet sind hier virtuelle Observatorien und deren astronomische Daten, die entweder experimentell oder durch Simulation gewonnen worden sind. Hier werden durch die Kombination eines verteilten Dateisystems und eines Anfrageoptimierers und -verteilers erhebliche Vorteile erzielt.

In dem dritten Artikel dieses Heftes befassen sich Dimitar Misev, Peter Baumann und Jürgen Seib unter dem Titel *Towards Large-Scale Meteorological Data Services: A Case Study* mit Daten, die im Rahmen von Wettersimulationen verwendet werden. Diese umfassen u.a. verschiedenste Formen von Echtzeit-Sensordaten, Satellitendaten sowie historische Daten, die in vieldimensionalen Arrays integriert werden. Gegenstand ist hier rasdaman, eine Datenbank, die zwar auf relationalen Datenbanken aufsetzt, aber auf dieser Basis für die Verwaltung von und Anfrage an meteorologischen Daten effiziente und praktisch relevante Anfragemöglichkeiten bereitstellt.

Schließlich beschäftigt sich der Artikel *Scientific Workflows and Provenance: Introduction and Research Opportunities* von Víctor Cuevas Vicenttín, Saumen Dey, Sven Köhler, Sean Riddle und Bertram Ludäscher mit wissenschaftlichen Workflows. Wie eingangs erwähnt werden diese für die Verarbeitung wissenschaftlicher Daten immer wichtiger. Bei der Verarbeitung und Analyse wissenschaftlicher Daten stehen solche Workflows als komplexe Arbeitsabläufe im Hintergrund, die Daten auf verschiedenste Arten vorverarbeiten, integrieren und umstrukturieren, um sie beispielsweise einer Analyse zugänglich zu machen. Die Bedeutung von Workflows liegt darin, dass sie eine Perspektive bieten, auf einfachere Art und Weise Domänenwissen einzubringen. Im Bereich der wissenschaftlichen Datenbanken spielt das jeweilige Domänenwissen eine sehr große Rolle. Häufig arbeiten Domänenexperten und Entwickler zusammen, um konkrete Probleme zu lösen. Den Domänenexperten interessiert im Wesentlichen das Resultat, der Informatiker ist häufig an der Durchführung, der Flexibilität, Generalität und Wiederverwendbarkeit interessiert.

Während oben die Unterschiede zwischen wissenschaftlichen und „normalen“ Datenbanken betont wurden, sollte man aber auch darauf hinweisen, dass eine Vielzahl von Entwicklungen aus dem Datenbankbereich bei der Verarbeitung und Analyse wissenschaftlicher Daten erfolgreich eingesetzt werden. Hierzu gehören z.B. eine Vielfalt von Indexstrukturen für räumliche und zeitlich veränderliche Daten, Verfahren zur Analyse von Datenströmen, effiziente Data-Mining-Methoden zum Clustering hochdimensionaler Daten oder Techniken zur Analyse von Graphstrukturen (wie sie beispielsweise gerade in der Molekularbiologie von Interesse sind). Nichtsdestotrotz bietet der Bereich Scientific Data Management noch eine Vielzahl von interessanten Möglichkeiten, Methoden und Techniken aus der Datenbanktechnologie geeignet in den oben genannten und weiteren Anwendungsbereichen einzubringen und somit diesen Wissenschaften bei ihren Problemen mit der Datenflut („Data Deluge“) zu helfen. Ein guter Wegweiser hierzu war und ist immer noch der Artikel von Jim Gray et al. *Scientific data management in the coming decade* (SIGMOD Record 34(4): 34–41, 2005).

Diese Schwerpunktbeiträge werden ergänzt durch zwei Fachbeiträge *XPath and XQuery Full Text Standard and Its Support in RDBMSs* von Dušan Petković und *CityPlot: Colored ER Diagrams to Visualize Structure and Contents of Databases* von Martin Dugas und Gottfried Vossen. Die Rubrik „Dissertationen“ enthält in diesem Heft sechs Kurzfassungen von Dissertationen.

In der Rubrik „Community“ berichten Uwe Wloka und Gunter Gräfe von einer in der deutschen Datenbankgemeinde einmaligen und sehr lebendigen Wissenschaftseinrichtung, die seit fast 20 Jahren vom Fachinteresse der Dresdener Datenbankkollegen zeugt. In ihrem Beitrag *Der 175. Datenbankstammtisch an der HTW Dresden* beschreiben sie das Konzept und die Historie dieser Veranstaltungsreihe und untermauern deren Erfolg mit verschiedenen statistischen Daten. Weiterhin enthält diese Rubrik einen *Bericht über den 24. GI-Workshop „Grundlagen von Datenbanken“* von Ingo Schmitt und Hagen Höffner sowie aktuelle Informationen.

2 Künftige Schwerpunktthemen

2.1 MapReduce Programming Model

MapReduce (MR) is a programming model which facilitates parallel processing of large, distributed, and even heterogeneous data sets. To accelerate the development of specific MR applications, an MR implementation provides a framework dealing with data distribution and and scheduling of parallel tasks. The user only has to complement this framework by specifying a *map* function—processing key/value pairs to generate intermediate key/value pairs—and a *reduce* function which groups all records with the same intermediate key and merges all values of such groups.

Using this approach, programs written in such a functional style can automatically exploit large degrees of parallelism and thereby perfectly scale. As a consequence, the MR model had tremendous success in recent years covering many areas of *Big Data* processing. For this reason, the "Datenbank-Spektrum" wants to publish research contributions—especially of the German database community—providing an overview over ongoing work in this particular area.

Submissions covering topics from the following non-exclusive list are encouraged:

- Applications of the MR paradigm
- Optimization of the MR framework and its applications
- MR-conform compilation of DB languages
- Schema flexibility (key/value stores) and MapReduce
- Comparison of applications running under MapReduce/Hadoop and parallel DBMSs
- Cooperation of NoSQL and SQL when processing XXL data

Paper format: 8–10 pages, double column.
Notice of intent for a contribution: July 15th, 2012.

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

Deadline for submissions: October 1st, 2012.

2.2 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—especially when accessed over the Web
- Applications and usage scenarios
- Case studies and experience reports

Paper format: 8–10 pages, double column.
Notice of intent for a contribution: Nov. 15th, 2012.

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

Deadline for submissions: February 1st, 2013.

2.3 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.