Datenbank Spektrum (2015) 15:171–174
DOI 10.1007/s13222-015-0202-5

EDITORIAL

Editorial

Theo Härder

Online publiziert: 7. Oktober 2015

Schwerpunktthema: Ausgewählte Beiträge von den Workshops der BTW 2015

Die Konferenz „Datenbanksysteme für Business, Technologie und Web (BTW)“ wird vom Fachbereich Datenbanken und Informationssysteme (DBIS) der Gesellschaft für Informatik (GI) in einem zweijährigen Turnus seit 1985 organisiert und stellt das zentrale Forum der Datenbank-Communities in Deutschland, in Österreich und in der Schweiz dar. Mit der 16. Auflage feierte die BTW, die in der Woche vom 2. bis 6. März 2015 an der Universität Hamburg stattfand, ihren 30. Geburtstag. Seit 2001 werden in der BTW-Woche neben dem Studierendenprogramm stets auch Workshops zu aktuellen oder sich abzeichnenden Themen der Datenbankforschung durchgeführt. Bei der BTW 2015 wurden solche zu folgenden Themen organisiert: Data Streams and Event Processing (DSEP), Joint Workshop on Data Management for Science (DMS), Databases in Biometrics, Forensics and Security Applications (DBforBFS).

Zusammen mit den Workshop-Organisatoren haben wir vier Workshop-Beiträge und einen Beitrag aus dem Studierendenprogramm ausgewählt, um aktuelle Forschungstrends im DBIS-Bereich aufzuzeigen. Für dieses Heft des Datenbank-Spektrums haben die eingeladenen Autoren ihre Beiträge im Vergleich zur Workshop-Version substanziell erweitert und verbessert. Diese „Best Papers“ wurden erneut streng begutachtet, bevor sie nach zwei Revisionsrunden zur Publikation angenommen wurden.

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

Das Volumen der verfügbaren Daten veränderte die Art und Weise, wie im geo-wissenschaftlichen Bereich Forschung betrieben wird, und bestimmt somit auch die Anforderungen an Systeme, die raumbezogene Daten verarbeiten. Zur Unterstützung von datengetriebener Forschung und von explorativen Workflows sind vor allem die Funktionen zur Visualisierung, Analyse und Transformation (VAT) wichtig, wozu im ersten Beitrag mit dem Titel *VAT: A System for Visualizing, Analyzing and Transforming Spatial Data in Science* Christian Authmann, Christian Beilschmidt, Johannes Drönner, Michael Mattig und Bernhard Seeger (Universität Marburg) ein entsprechendes System vorschlagen. Dazu identifizieren die Autoren zunächst 10 Anforderungen, die den gesamten Bereich von räumlichen Datentypen über effiziente Rechenverfahren bis hin zu den Visualisierungstechniken überspannen. Basierend auf diesen Anforderungen bewerten sie State-of-the-Art-Systeme aus verschiedenen Bereichen wie beispielsweise Geo-Informationssysteme, Workflow-Systeme und wissenschaftliche Datenbanken. Mit Hilfe dieser Evaluationsergebnisse können die Autoren dann zeigen, dass das VAT-System deren Begrenzungen durch einen ganzheitlichen Ansatz für Raster- und Vektordaten, bedarfsgesteuerter Verarbeitung und effizienter Nutzung von heterogenen HW-Architekturen überwindet. Durch empirische Experimente können sie bereits in ihrem frühen Projektstadium das Potenzial heterogener Systemarchitekturen durch den kombinierten Einsatz von CPUs und GPUs für die Verarbeitung raumbezogener Daten aufzeigen.

Im folgenden Beitrag mit dem Titel *Genome sequence analysis with MonetDB – A case study on Ebola virus diversity* beschreiben Robin Cijvat (1), Stefan Manegold (2), Martin Kersten (1, 2), Gunnar Klau (2), Alexander Schönhuth (2), Tobias Marschall (3) und Ying Zhang (1, 2) ((1) MonetDB

Solutions, Amsterdam; (2) Centrum Wiskunde & Informatica, Amsterdam; (3) Saarland University & Max Planck Institute for Informatics, Saarbrücken) schlagen einen DBMS-basierten Ansatz vor, mit dem die Genomsequenz-Analyse beschleunigt und substanziell vereinfacht werden soll, um die Arbeit der Biologen bei diesen *Big Data*-Anwendungen zu unterstützen. Dazu wurde MonetDB um einen BAM-Modul (Block Availability Map, eine Datenstruktur zur Organisation der Datenblöcke auf einem Massenspeicher) erweitert, der eine einfache, flexible und schnelle Verwaltung und Analyse von Sequenzalignment-Daten ermöglicht. Mit Hilfe einer Fallstudie basierend auf Ebola-Virus-Genomen beschreiben die Autoren die wesentlichen Eigenschaften von MonetDB/BAM.

Query Optimization in Heterogenous Event Processing Federations wurde als bester Beitrag des Studierendenprogramms ausgewählt. Marcus Pinnecke (Otto-von-Guericke-Universität Magdeburg) und Bastian Hoßbach (Universität Marburg) betrachten in diesem dritten Beitrag wichtige Probleme bei der kontinuierlichen Verarbeitung von Ereignisströmen, bei der die Konzepte, Modelle und Algorithmen von zwei Forschungsgebieten zusammengeführt werden müssen. Da bisher die Forschungsarbeiten zu Datenströmen und Ereignisverarbeitung isoliert und ohne vorhandene Standards durchgeführt wurden, sind heutige Stromverarbeitungssysteme (SPS) in einem sehr hohen Maße heterogen. Um die daraus entstehenden Probleme zu überwinden, wurde in den letzten Jahren eine Middleware zur Ereignisverarbeitung, die Java Event Processing Connectivity (JEPC), vorgestellt, wodurch jedoch nur eine „oberflächliche" Homogenisierung des Zugriffs auf SPS-Systeme erreicht wurde. Die sich durch verschiedenartige Algorithmen und Systemimplementierungen ergebenden Leistungsschwächen blieben bestehen, so dass jetzt zusätzliche verbessernde Maßnahmen erforderlich sind. Die Autoren entwickelten deshalb einen neuartigen Anfrageoptimierer, der die technischen Heterogenitäten in einem Verbund verschiedener vereinheitlichter SPS ausnutzt. Um das gesamte Leistungsverhalten zu verbessern, betrachten sie vor allem die Partitionierung von Anfrageplänen, die Kandidaten-Selektion und die Reduzierung der Kommunikation im SPS-Verbund. Als Ergebnis wird eine Heuristik zur anfänglichen Abbildung von Teilplänen auf eine Menge von heterogenen SPS vorgeschlagen. Mit einer experimentellen Evaluierung konnte dann gezeigt werden, dass heterogene Verbunde im Allgemeinen homogenen Verbunden überlegen sind und dass die vorgeschlagene Heuristik sich gut für praktische Anwendungen eignet.

Datenstromverarbeitungssysteme ermöglichen Anfragen auf kontinuierlichen Daten, ohne diese vorher zu speichern. Andererseits können Anfragen auf Datenströmen Daten von verteilten Datenquellen betreffen, wie sie beispielsweise von verschiedenen Sensoren einer Anwendung zur Umweltüberwachung kommen, was verteilte Anfrageverarbeitung erfordert. Diese Art der Verabeitung auf möglicherweise heterogenen Plattformen erschwert die Anfrageoptimierung. In ihrem Beitrag *Placement-Safe Operator-Graph Changes in Distributed Heterogeneous Data Stream Systems* untersuchen Niko Pollner (1), Christian Steudtner (2) und Klaus Meyer-Wegener (1) ((1) Friedrich-Alexander-Universität Erlangen-Nürnberg; (2) Deutsche Anwaltshotline AG, Nürnberg) Möglichkeiten der Anfrageoptimierung durch Veränderungen des Operatorgraphs und ihre Wechselwirkungen mit der Operatorzuordnung in heterogenen verteilten Systemen. Da eine Vorabverteilung der Änderungen im Operatorgraph bestimmte Operatorzuordnungen verhindern kann, ist ein unerwartetes Ansteigen des Ressoucenverbrauch bei der Anfrageverarbeitung möglich. Basierend auf dem Operatorzuordnungsproblem, das als Aufgabenzuordnungsproblem (task assignment problem (TAP) modelliert wird, beweisen die Autoren, dass im Allgemeinen die Entscheidung, ob eine beliebige Operatorgraph-Änderung die bestmögliche TAP-Lösung negativ beeinflusst, NP-schwer ist. Sie können jedoch für verschiedene spezifische Operatorgraph-Änderungen Bedingungen angeben, die eine Erhaltung der bestmöglichen TAP-Lösung garantieren.

Der fünfte Schwerpunkt-Beitrag *Modulares Verteilungskonzept für Datenstrommanagementsysteme* von Timo Michelsen, Michael Brand und Hans-Jürgen Appelrath (Universität Oldenburg) untersucht die Verteilung kontinuierlicher Anfragen in verteilten Datenstrommanagementsystemen, für die es je nach Netzwerkarchitektur und Anwendungsfall unterschiedliche Strategien gibt. Eine einmal festgelegte Strategie kann jedoch bei einer Änderung der Netzwerkarchitektur oder des Anwendungsfalls zu Nachteilen führen. Deshalb entwickelt dieser Beitrag in drei Schritten einen Ansatz für eine flexible und erweiterbare Anfrageverteilung in verteilten Datenstrommanagementsystemen: 1) Partitionierung, 2) Modifikation und 3) Allokation. Die Partitionierung zerlegt eine kontinuierliche Anfrage in disjunkte Teilanfragen. Die optionale Modifikation erlaubt es, Mechanismen wie Fragmentierung oder Replikation zu verwenden, während die Allokation schließlich zur Ausführung der einzelnen Teilanfragen Knoten im Netzwerk zuweist. Da für jeden Schritt sich unabhängige Strategien verwenden lassen, ermöglicht ein solcher modularer Aufbau eine individuelle Anfrageverteilung. Zur Illustration stellen die Autoren für jeden der drei Teilschritte exemplarisch Strategien vor. Weiterhin zeigen drei Anwendungsbeispiele die Vorteile dieses modularen Ansatzes gegenüber einer festen Verteilungsstrategie.

Community-Beiträge in diesem Heft

Die Rubrik „Datenbankgruppen vorgestellt" enthält den Beitrag *Die Arbeitsgruppe Datenbanksysteme an der Philipps-Universität Marburg*, in dem Bernhard Seeger zunächst die Arbeitsschwerpunkte im Verlauf ihrer 20-jährige

Geschichte skizziert, bevor er die drei wichtigsten, aktuellen Forschungsarbeiten der Arbeitsgruppe beschreibt und einen Überblick über ihre Lehraufgaben und ihre vielfältigen Kooperationen gibt.

Die Rubrik „Kurz erklärt“ soll künftig im Datenbank-Spektrum wieder stärker belebt werden. Deshalb haben wir für dieses Heft zwei Beiträge zu aktuellen Stichworten akquiriert. Das Stichwort *Polyglot Persistence* wird von Felix Gessert und Norbert Ritter (Universität Hamburg) auf interessante Weise von verschiedenen Seiten beleuchtet. Weiterhin erklären Goetz Graefe (1), Caetano Sauer (2), Wey Guy (1) und Theo Härder (2) ((1) HP-Labs, Palo Alto; (2) TU Kaiserslautern) auf kompakte Weise *Instant recovery with write-ahead logging*. Die damit verbundenen Techniken sollen erreichen, dass Unterbrechungen der Transaktionsverarbeitung durch System- oder Gerätefehler nach außen nicht mehr „sichtbar“ werden.

Weiterhin werden in der Rubrik „Dissertationen“ in diesem Heft 4 Kurzfassungen von Dissertationen aus der deutschsprachigen DBIS-Community vorgestellt.

Die Rubrik „Community“ enthält schließlich unter *News* weitere aktuelle Informationen aus der DBIS-Gemeinde.

Künftige Schwerpunktthemen

1 Big Data & IR

The term *Big Data* refers to data and respective processing strategies, which, due to their sheer size, require a data center for the processing, and which become available through the ubiquitous computer and sensor technology in many facets of everyday life. Interesting scientific questions in this regard are the organization and management of Big Data, but also the identification of problems that now can be studied and better understood through the collection and analysis of Big Data. In the context of information retrieval as the purposeful search for relevant content, there are two main challenges: 1) retrieval in Big Data and 2) improved retrieval because of Big Data.

Retrieval in Big Data focuses on the organization, the management, and the quick access to Big Data, but also addresses the creative process of identifying interesting research questions that can only be understood and answered in Big Data. Besides the development of powerful frameworks for the maintenance and analysis of text, multimedia, sensor, and simulation data, an important research direction is the question of what kind of insights Big Data may give us today and in the future.

The second challenge in the context of Big Data & IR is the improvement of retrieval approaches through Big Data. Examples include the classic question of improved Web or eCommerce search via machine learning on user behavior data, the usage of user context for retrieval, or the exploitation of semantic data like Linked Open Data or knowledge graphs.

We are looking for contributions from researchers and practitioners in the above described context. The contributions may be submitted in German or in English and should observe a length of 8–10 pages in the Datenbank-Spektrum format (cf. the author guidelines at www.datenbank-spektrum.de).

Issue delivery: DASP-1-2016 (March 2016)
Guest editors:
Matthias Hagen, Universität Weimar
matthias.hagen@uni-weimar.de
Benno Stein, Universität Weimar
benno.stein@uni-weimar.de

2 Schutz der Privatsphäre in einer ubiquitären Welt

Mit immer mehr mobilen Geräten und Sensoren werden u. a. große Mengen an persönlichen Daten gesammelt, verarbeitet und transformiert. Solche Sammlungen personenbezogener Daten sind auf der einen Seite notwendig, um personenspezifische Angebote machen zu können, die dem Empfänger örtlich und zeitlich von Nutzen sind, oder um Trends zu erkennen und somit Planungen in unterschiedlichen Bereichen genauer und effizienter ausführen zu können. Auf der anderen Seite dienen sie häufig dazu, individuelle Personenprofile zu erstellen, die zum Vorteil oder Nachteil der beschriebenen Person genutzt werden können.

Aus den genannten Gründen wird es immer wichtiger, den Datenschutz in einer ubiquitären Welt im Kontext von Big Data nicht nur juristisch abzusichern (Bundesdatenschutzgesetz). Vielmehr wird es immer dringlicher, auch technische Möglichkeiten, Mechanismen und Ansätze zu entwerfen und zu realisieren, die es Personen ermöglichen, die Kontrolle über ihre Daten besser spezifizieren sowie ihre Nutzung und Weitergabe besser kontrollieren und nachvollziehen zu können. Trotz großer Fortschritte im Bereich des Schutzes der Privatsphäre durch unterschiedliche Techniken besteht weiterhin eine große Herausforderung darin, skalierbare Ansätze und Lösungen sowohl für die Nutzung personenbezogener Daten durch Dritte als auch deren Kontrolle durch den „Spender“ zu entwickeln und zu realisieren.

Somit ist es das Ziel des Themenheftes, neben einer Einführung in das Thema skalierbare Ansätze und Lösungen für das Sammeln, Verarbeitung und Analysieren personenbezogener Daten in unterschiedlichen Anwendungsdomänen zu beschreiben. Mögliche Themen sind in für dieses Themenheft sind (nicht ausschließlich):

- Schutz der Privatsphäre im Bereich Big Data generell beim Sammeln, Integrieren, Verarbeiten und Analysieren von Daten
- Technische Umsetzung juristischer (gesetzlicher) Vorgaben entsprechend des deutschen Rechts bzw. des EU-Rechts und Vorgaben in unterschiedlichen Bereichen wie beispielsweise dem Gesundheitsbereich (Arzt, Krankenhaus, Versicherer) oder dem Finanzbereich (Banken)
- Sprachen zur Beschreibung von Privacy-Präferenzen und deren Überprüfung (in skalierbarer Form)
- Datenaustausch unter Berücksichtigung von Privacy-Präferenzen und gesetzlichen Vorgaben
- Anfragebearbeitung in Datenbanksystemen unter Berücksichtigung von Privacy-Präferenzen
- Quantitative Bewertung von Ansätzen zum Schutz der Privatsphäre im Kontext der Nutzung geschützter Daten
- Anforderungsanalysen für den Schutz der Privatsphäre in Anwendungsdomänen – Schutz gegen Genauigkeit der Daten
- Anforderungen des Schutzes der Privatsphäre für räumliche und zeitbezogene Daten in verschiedenen Anwendungsbereichen
- Infrastrukturen zum Schutz der Privatsphäre
- Modelle zum Schutz der Privatsphäre bei Zugriff oder Datennutzung

Beitragsformat: 8–10 Seiten, zweispaltig

Ankündigung einer Beitragseinreichung bis zum 15. Dezember 2015

Gastherausgeber:
Johann-Christoph Freytag, HU Berlin
freytag@informatik.hu-berlin.de
Eric Buchmann, Karlsruher Institut für Technologie
eric.buchmann@kit.edu

Einreichung der Beiträge für DASP-2-2016 bis zum 1. Februar 2016

3 Data Management for Bio- and Geosciences

Like many other scientific disciplines, research in the bio- and geosciences follow more and more a data-driven approach. Big Data in the classical sense is only one of the issues, but probably more often the everyday problem of scientists is to cope with lots of „small“, heterogeneous data that needs to be integrated to answer complex questions.

This special issue addresses the arising challenges and solutions for data management in these areas. We are interested in both survey papers and papers describing original research dealing with the following or similar topics in the context of the bio- and geosciences:

- Data-intensive science
- Data management
- Data integration
- Spatio-temporal data processing
- Scientific workflows
- Semantic web technologies
- Visualization and visual analytics
- Data stream management
- Case studies and applications

Important dates:

- Notice of intent for a contribution: April 15th, 2016
- Deadline for submissions: June 1st, 2016
- Issue delivery: DASP-3-2016 (November 2016)

Paper format: 8–10 pages, double column

Guest editors:
Bernhard Seeger, Philipps-Universität Marburg
seeger@mathematik.uni-marburg.de
Birgitta König-Ries, Friedrich-Schiller-Universität Jena
Brigitta.Koenig-Ries@uni-jena.de