

Editorial

Peter M. Fischer · Daniela Nicklas · Bernhard Seeger

Online publiziert: 4. Februar 2012
© Springer-Verlag 2012

Viele Anwendungen basieren bereits heute auf aktiven Datenquellen. Sensoren werden kleiner und günstiger, die Ausführungsumgebungen werden dynamischer – und die Anwendungen sollen „smarter“ werden, d.h., sich ein aktuelles Bild der Welt machen und darauf geeignet und rechtzeitig reagieren. Damit entstehen aber auch neue Anforderungen an die Datenverarbeitung. Den Komfort und die Effizienz, den uns Datenbankmanagementsysteme (DBMS) für die Verwaltung und Abfrage von gespeicherten Daten bieten, wünschen wir uns auch für Datenströme und Ereignisse.

Der Titel des Themenhefts „Data Stream and Event Processing“ zeigt, dass hier Beiträge aus zwei Blickwinkeln zusammen kommen. Während in der Datenstromverarbeitung kontinuierliche Anfragen verarbeitet werden, werden in der Ereignisverarbeitung aus eintreffenden Basisereignissen komplexe Ereignisse erkannt. Beide Forschungsrichtungen stehen mittlerweile in engem Dialog miteinander.

Seit 2009 trifft sich die deutschsprachige wissenschaftliche Gemeinschaft zu diesem Thema halbjährlich. In der informellen Workshopreihe „Stream Data Management and

Applications“, die ausgehend von Oldenburg mittlerweile in Erlangen, Marburg und Zürich stattfand, findet über Vorträge und Diskussionsrunden ein reger Austausch über aktuelle Forschungsarbeiten und zukünftige Entwicklungen statt. Ausgehend von diesen Aktivitäten organisierten die Herausgeber auf der BTW 2011 einen Workshop zum Thema „Data Streams and Event Processing“, dessen thematische Ausrichtung auch Kern der Ausschreibung für dieses Themenheft war.

In diesem Heft finden sich nun fünf Beiträge, die verschiedene Herausforderungen in der Verarbeitung von Datenströmen und Ereignissen behandeln:

Frank Lauterwald, Michael Daum, Niko Pollner und Klaus Meyer-Wegener berichten in *Umgang mit semantischer Heterogenität bei der Integration stromverarbeitender Systeme* von einem Ansatz, in dem verschiedene bestehende Datenstromsysteme zu einer gemeinsamen Anfrage-schnittstelle föderiert werden. Dabei kann der Anwendungsentwickler wahlweise präzise definieren, wie eine Anfrage verarbeitet werden soll, oder dem System bestimmte Teilaspekte freistellen, um so Optimierungspotentiale zu nutzen. Die Kombination und Föderation verschiedener Datenstrommanagementsysteme (DSMS) ist von Interesse, um unterschiedliche Quellen und Dienste zu integrieren, z.B. bei Sensornetzwerken oder Realtime Business Intelligence. Eine besondere Herausforderung entsteht dabei durch die unterschiedlichen Semantiken einzelner DSMS im Bezug auf ihre Operationen und Verarbeitungslogik, sodass selbst bei „gleichen“ Anfragen unterschiedliche Ergebnisse erzeugt werden.

Gerade in sicherheitsrelevanten und sensitiven Anwendungen ist es wichtig, dass die Datenströme selbst vor unerlaubtem Zugriff geschützt werden. Mögliche Beispiele sind die Beschränkung der Verbreitung beim Börsenhandel oder die Einhaltung der Privatsphäre bei Ortsbezoge-

P.M. Fischer (✉)
Institut für Informatik, Universität Freiburg, Freiburg,
Deutschland
e-mail: peter.fischer@informatik.uni-freiburg.de

D. Nicklas
Department für Informatik, Universität Oldenburg, Oldenburg,
Deutschland
e-mail: dnicklas@acm.org

B. Seeger
Fachbereich Mathematik und Informatik, Universität Marburg,
Marburg, Deutschland
e-mail: seeger@informatik.uni-marburg.de

nen Diensten. Nazario Cipriani, Oliver Dörler und Bernhard Mitschang stellen in ihrem Artikel *Sicherer Zugriff und sichere Verarbeitung von Kontextdatenströmen in einer verteilten Umgebung* ein flexibles Rahmenwerk zur sicheren Verarbeitung von Kontextdaten vor, das es Anbietern von Daten in Datenstromverarbeitungssystemen ermöglicht, den Zugriff und die Verarbeitung schützenswerter Daten zu kontrollieren. Das vorgeschlagene Konzept erlaubt feingranularen Zugriff und unterstützt die Einbindung benutzerdefinierter Operatoren.

Die Verarbeitung von Datenströmen kann selbst als Dienst angeboten werden. Wie bei klassischen Services sollten nun Garantien für die Leistung abgegeben werden, für deren Verletzung der Dienstanbieter finanzielle Folgen zu tragen hat. Thomas Vogelgesang, Dennis Geesen, Marco Grawunder, Daniela Nicklas und H.-Jürgen Appelrath untersuchen in ihrer Arbeit *Service Level Agreement basiertes Scheduling von Datenströmen*, welche Kriterien dafür bei DSMS in Frage kommen und wie diese mit Hilfe von Ressourcen-Scheduling erreicht werden können. Die Latenz zwischen ankommenden und ausgehenden Daten hat sich dabei als einzige relevante und effektiv steuerbare Größe herausgestellt.

Event Processing kann aber auch dazu genutzt werden, um den Herausforderungen von Service- und Ressourcen-Management, insbesondere im Umfeld von Cloud Computing, begegnen zu können. Bastian Hoßbach, Bernd Freisleben und Bernhard Seeger legen in ihrem Beitrag *Reaktives Cloud Monitoring mit Complex Event Processing* dar, dass die Techniken aus dem Bereich Complex Event Processing (CEP) gute Möglichkeiten bieten, einige der Hauptrisiken von Cloud Computing besser zu kontrollieren: Unzuverlässigkeit, Unsicherheit, und mangelnde Überwachung der Dienste.

Das Mining von Datenströmen führt gegenüber klassischen Miningverfahren zu neuen Herausforderungen. Insbesondere ist dabei auf die Verfügbarkeit von Ressourcen wie CPU und Speicher zu achten, da Ankunftsraten und Verteilung der Daten stark variieren können. Philipp Krannen, Ira Assent und Thomas Seidl verwenden in ihrer Arbeit *An Index-inspired Algorithm for Anytime Classification on Evolving Data Streams* einen sogenannten „Anytime-Algorithmus“, welcher die für ein ankommendes Datentupel zur Verfügung stehende Zeit flexibel ausnutzt, indem er stets ein Ergebnis produziert, dieses jedoch mit mehr Ressourcen graduell verbessert.

Unser Dank gilt allen Autoren, die Beiträge zu diesem Themenheft eingereicht haben, auch wenn wir diese nicht berücksichtigen konnten. Besonders aber möchten wir den Gutachtern danken, ohne deren ausführliche und konstruktive Gutachten das Themenheft in dieser Qualität nicht möglich gewesen wäre: Ludger Fliege, Stefan Floering, Wolfgang Lehner, Rainer Manthey, Bernhard Mitschang, Gero

Mühl, Kai Sachs, Kai-Uwe Sattler, Thorsten Schöler, Nesime Tatbul und Michael Wurst.

Diese Schwerpunktbeiträge werden ergänzt durch einen Fachbeitrag *Doppelte Virtualisierung am Beispiel einer Datenbank-Cluster-Konfiguration für den Aufbau von Testumgebungen* von Markus Bräuning und Rüdiger Steffan. Weiterhin finden Sie unter der Rubrik „Kurz erklärt“ einen Beitrag von Michael Seibold und Alfons Kemper zu *Database as a Service* und einen Beitrag *Information Retrieval an der Universität Hildesheim: Optimierung, Evaluierung und Informationsverhalten* unter der Rubrik „Datenbankgruppen vorgestellt“. Die Rubrik „Dissertationen“ ist erfreulicherweise recht umfangreich; sie enthält in diesem Heft sieben Kurzfassungen von Dissertationen. Um in dieser Rubrik ein möglichst vollständiges Bild von erfolgreich abgeschlossenen Promotionsarbeiten in unseren Fachgruppen bieten zu können, erbitten die Hauptherausgeber die Zusendung der entsprechenden Kurzfassungen direkt nach Abschluss der Promotionsverfahren. Schließlich erscheinen in der Rubrik „Community“ aktuelle Berichte und Nachrichten aus der jüngsten Zeit.

Künftige Schwerpunktthemen

1 Information Extraction

Database technology as tool to structure, retrieve, and analyze information has proven to be of great value. However, today's information repositories continuously grow and diversify and range from rather unstructured document collections on the Web, over partially structured information in business warehouses, to structured, yet huge experimental data collections, e.g., in the life sciences. Extracting information from such heterogeneous sources and making it available for analysis is a hard problem needing complex algorithms and mining technologies. This special issue aims to provide an overview over the spectrum of techniques addressing different types of information sources.

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

- Web information extraction
- Data mining and visualization
- Information extraction from data streams
- Business intelligence and information mining

Guest editor:

Wolf-Tilo Balke, TU Braunschweig,
balke@ifis.cs.tu-bs.de

2 Scientific Data Management

The past decade has witnessed a dramatic increase in scientific data being generated in the physical, earth, and life sciences. This development is primarily a result of major advancements in sensor technology, surveying techniques, computer-based simulations, and instrumentation of experiments. In a special issue of the „Datenbank-Spektrum“, we want to publish original work on different aspects related to the management and analysis of scientific data. The objective of this special issue is to exchange ideas between academia and industry and to discuss recent developments, challenges, and future directions in scientific data management.

Topics of interest include (but are not limited to)

- Modeling and representation of data, metadata, ontologies, and processes for scientific application domains
- Integration and exchange of scientific data
- Design, implementation, and optimization of scientific workflows
- Architectures and components for scientific computing and eScience, including Web portals, repositories, and digital libraries
- Annotation and provenance of scientific data
- Mining and analysis of large-scale scientific datasets
- Case studies and applications related to scientific data management in all domains, with a particular focus on biology, physics, chemistry, medicine, and geography

Paper format: 8–10 pages, double column

Notice of intent for a contribution: March 15th, 2012

Guest editors:

Michael Gertz, Heidelberg University,
gertz@informatik.uni-heidelberg.de
Wolfgang Müller, HITS gGmbH,
wolfgang.mueller@h-its.org

Deadline for submissions: June 1st, 2012

3 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 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