Special issue on best papers of VLDB 2016

Surajit Chaudhuri¹ · Jayant R. Haritsa²

Published online: 24 September 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

The 2016 VLDB Conference was held in New Delhi, India, during September 5–9, 2016. We received a record number of 719 research submissions, of which around 16% were accepted for presentation by the Program Committee. From these high-quality manuscripts, a Best Papers Committee consisting of Jignesh Patel, Jeffrey D. Ullman, and Gerhard Weikum, selected six outstanding papers and also picked the best paper award for the conference. We invited the authors of the six selected papers to submit an extended version for consideration in this Special Issue, and they all accepted our call. The reviewers for the manuscripts were a mix of those who had reviewed the conference versions, as well as additional experts who reviewed only the journal submissions. After two rounds of reviewing, all six papers were accepted for publication in this issue, covering a diverse spectrum of topics ranging from core database engines to knowledge management.

The paper “Adding Data Provenance Support to Apache Spark” by Interlandi, Ekmekji, Shah, Gulzar, Tetali, Kim, Millstein, and Condie facilitates the debugging of programs written in dataflow systems such as Spark, where intermediate results are routed through a DAG-structured network of data-parallel operators. They introduce a library, called Titian that enables building, for each type of Spark RDD, a matching lineage RDD that enables the programmer seamlessly and interactively navigate both backward and forward in the Spark program dataflow. Titian also features a specialized join algorithm for efficiently executing time-travel queries on the lineage, thereby greatly facilitating debugging. A detailed experimental evaluation demonstrates that Titian’s provenance capture incurs only modest overheads on the overall program execution time.

The paper “Efficient Generation of Query Plans Containing Group-By, Join, and Groupjoin” by Eich, Fender, and Moerkotte revisits an old idea (referred to as “eager aggregation”) of pushing aggregations below joins in query execution plans. Eager aggregation was originally proposed in the 1990s and has since been refined and deployed in database system implementations. The paper makes an in-depth study on the impact of eager aggregation on the search space for optimizers. Specifically, the effect of keys, equivalence classes, functional dependencies and not-null attributes are studied for different join operators. This analysis leads to the key contribution of a refinement of the search space pruning strategies for early aggregation.

The paper “Query Optimization Through the Looking Glass, and What We Found Running the Join Order Benchmark” by Leis, Radke, Gubichev, Mirchev, Boncz, Kemper, and Neumann meticulously examines the roles of cardinality estimation, cost estimation and search strategy on the quality of the plans produced by relational query optimizers. Although primarily focused on main-memory systems, the paper also studies disk-based systems. We expect that even those knowledgeable about query processing engines will gain fresh insights from their careful analysis of optimizer errors. An equally important contribution of the paper is the proposal and rationale for benchmarks to evaluate the quality of query optimizers.

The paper “Many-Query Join: Efficient Shared Execution of Relational Joins on Modern Hardware” by Makreshanski, Giannikis, Alonso, and Kossmann focuses on concurrent query execution, a common feature in contemporary data processing environments. Fast query execution in such scenarios necessitates execution efforts across concurrent queries. The paper leverages the considerable literature in this area by intelligently combining a variety of prior techniques to improve the performance of concurrent query execution. These techniques include mechanisms for the sharing of join results, and the deployment of efficient data structures well suited to multi-threaded query executions. Collectively, this bricolage approach results in extremely impressive performance accelerations on main-memory platforms.

✉ Jayant R. Haritsa
haritsa@iisc.ac.in

Surajit Chaudhuri
surajitc@microsoft.com

¹ Microsoft Research, Redmond, USA

² IISc Bangalore, Bangalore, India

The paper “Package Queries: Efficient and Scalable Computation of High-order Constraints” by Brucato, Abouzied, and Meliou is in the broad area of augmenting database query languages with the power of specifying constraints. They develop a new query model wherein each query returns a set of tuple sets, called packages, with the tuples in each set matching the specified constraints. They show that when an optimization criterion is added to choose among qualifying packages, the query evaluation can be efficiently accomplished by a novel combination of the relational query execution engine and a generic ILP solver. The authors also explain how such “package queries” may be specified declaratively. We are happy to share the news that the abridged version of this paper has been selected for publication in a forthcoming “Research Highlights” section of CACM.

The paper “Compressed Linear Algebra for Large-Scale Machine Learning” by Elgohary, Boehm, Haas, Reiss, and Reinwald addresses the problem of compressing matrices for linear algebra operations in the context of machine learning tasks. Unlike prior work which focused solely on sparse matrices, the paper also tackles dense matrices. The

solution covers all relevant aspects for practical viability: sampling-based fast compression by adapting techniques from column-store databases, careful planning for applying different compression techniques to different matrix partitions, cache-conscious operations over compressed data, and full integration into the Apache SystemML software. A comprehensive set of experiments demonstrate substantial savings in the memory footprint, while retaining speed comparable to computations on uncompressed data. This work is a textbook example of judiciously adapting and cleverly combining a variety of techniques, and provides a solution for a problem of wide applicability and contemporary relevance.

In closing, we thank all the authors for their diligent efforts in substantively extending their conference manuscripts, and for submitting them in a timely manner. We also thank our team of reviewers for their insightful comments and suggestions, which enhanced the final quality of the papers. Finally, we reiterate our profound gratitude to the Best Papers Committee of VLDB 2016 for tackling the formidable task of selecting these six papers from a large and competitive group of conference publications.

PVLDB Volume 9 Editors-in-Chief
VLDB 2016 Program Committee Chairs