Distrib Parallel Databases
DOI 10.1007/s10619-017-7204-2

# Special issue on in-database analytics

**Dan Olteanu**[1] · **Florin Rusu**[2]

© Springer Science+Business Media, LLC 2017

Recent years have witnessed a sustained effort towards big data analytics, such as regression, classification, and recommendation, on massive datasets of varied types and formats. Although most of the methods to perform these tasks have been studied before in machine learning and data mining, scalability and generality have risen as primary challenges not addressed before. As a result, many libraries, frameworks, and platforms have been recently developed to provide scalable support for distributed and parallel statistical analytics. While mostly scalable, several of the newly proposed approaches fall short on generality. In-database solutions seek to reuse and extend the query language and processing infrastructure of a database system with generic constructs and operators required to perform these considerably more complicated tasks. They are generic, but not necessarily scalable.

This special issue focuses on conceptual, systems, and algorithmic aspects of large-scale data analytics, both inside and outside the database. It features a collection of four articles, selected out of 12 submissions, that provide a snapshot of the state of the art and a useful point of reference for research in large-scale database analytics. The topics of the articles cover in-database linear algebra algorithms for massive graphs, privacy-preserving classification with missing data, parallel skyline computation on MapReduce, and indexing in a federated cloud. We present a short summary of the accepted articles in the following:

B Florin Rusu
frusu@ucmerced.edu

Dan Olteanu
dan.olteanu@cs.ox.ac.uk

1 University of Oxford, Oxford, UK

2 University of California Merced, Merced, CA, USA

Published online: 11 September 2017

- *Scalable parallel graph algorithms with matrix-vector multiplication evaluated with queries* by Wellington Cabrera (University of Houston) and Carlos Ordonez (University of Houston) presents a unified algorithmic pattern to express matrix-vector multiplications with SQL queries. They apply this methodology to four graph problems—reachability from a source vertex, single source shortest path, weakly connected components, and PageRank— and propose efficient optimizations for parallel evaluation in a shared-nothing database.
- *Privacy-preserving of SVM over vertically partitioned with imputing missing data* by Mohammed Z. Omer (University of Electronic Science and Technology of China), Hui Gao (University of Electronic Science and Technology of China), and Nadir Mustafa (University of Electronic Science and Technology of China) introduces a privacy-preserving SVM model over vertically-partitioned data based on the Gram matrix. This model supports multiple imputations of missing or incomplete data through the Multivariate Imputation by Chained Equations (MICE) technique and the Paillier Cryptosystem.
- *A parallel computation of skyline using multiple regression analysis-based filtering on MapReduce* by Miyoung Jang (Chonbuk National University of Korea), Youngho Song (Chonbuk National University of Korea), and Jae-Woo Chang (Chonbuk National University of Korea) proposes a parallel skyline query processing algorithm for MapReduce using multiple regression analysis-based filtering. The filtering threshold line is accurately computed from a sample of the data. An angle-based partitioning scheme is applied for local skyline computation, while the dominance relationship among grid-based partitions is used for the global skyline evaluation.
- *An efficient distributed search solution for federated cloud* by Yongqing Zhu (A*STAR Singapore), Quanqing Xu (A*STAR Singapore), Haixiang Shi (A*STAR Singapore), and Juniarto Samsudin (A*STAR Singapore) presents a Distributed Search Index (DS-index) for distributed search capability in federated cloud. The target is multi-attribute range queries with secondary indexes. DS-index deploys a three-layer architecture with a multi-attribute index overlay, a tree-based P2P network layer, and a federated cloud layer. Dynamic mapping between these layers and automatic node selection are realized with a Markov Chain-based cost model.

We appreciate all the authors who submitted papers to this special issue for their contributions. We also thank the reviewers for their generous help and valuable comments. And last, we are grateful to Prof. Divyakant Agrawal, Prof. Amit P. Sheth, and Prof. Mohamed Mokbel, the Editors-in-Chief of DAPD, for their strong support for this special issue.