Distributed and Parallel Databases
https://doi.org/10.1007/s10619-019-07258-x

Special issue on scientific and statistical data management

Kesheng Wu[1] · **Florin Rusu**[2]

The recent success of "Big Data" is partly the result of decades of research work on large scientific and statistical data. Indeed, scientific datasets, such as those from large physics experiments, and statistical datasets, such as those from national censuses, are the original forms of "Big Data". The tools and techniques developed for these forms of "Big Data" have led to the discovery of the Higgs Boson (aka, the god particle) and the emergence of behavioral economics. As the form and variety of scientific and statistical data evolve, there is a thriving series of work on organizing, analyzing, publishing, and preserving the datasets and their derived artifacts. In particular, advances in computer architecture and High Performance Computing (HPC) techniques are opening new opportunities to accelerate the analysis of even larger datasets.

This special issue intends to provide a snapshot of the active research topics in the field of statistical and scientific data. We solicited contributions from a wide range of sources, including at the 29th International Conference on Scientific and Statistical Database Management (SSDBM 2017). With the help of dozens of reviewers, we present seven of them for publication. Here is a brief summary of the accepted articles.

- *On Effective and Efficient Graph Edge Labeling* by Oshini Goonetilleke (RMIT), Danai Koutra (Univ of Michigan), Kewen Liao (Swinburne University of Technology), and Timos Sellis (Swinburne University of Technology) presents a set of techniques to label graph edges instead of nodes. Since graph algorithms generally follow the edges during their operations, ordering edges can lead to better locality in data access and reduce the overall execution time. One of the proposed labeling algorithms was specifically designed for streaming graph partitioning.

✉ Kesheng Wu
kwu@lbl.gov

Florin Rusu
frusu@ucmerced.edu

[1] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[2] University of California, Merced, CA, USA

Published online: 28 February 2019

- *Detecting Global Hyperparaboloid Correlated Clusters: A Hough-Transform Based Multicore Algorithm* by Daniyal Kazempour (LMU Munich), Markus Mauder (LMU Munich), Peer Kröger (LMU Munich) and Thomas Seidl (LMU Munich) introduces a method to detect global non-linear correlated clusters, focusing on quadratic relations, while existing techniques can only extract linear correlations. Furthermore, the authors present an effective way of parallelizing the algorithm to significantly improve the runtime.
- *MDCUT2: A Multi-Density Clustering Algorithm with Automatic Detection of Density Variation in Data with Noise* by Soumaya Louhichi (MIRACL), Mariem Gzara (MIRACL), and Hanêne Ben Abdallah (King Abudulaziz University) proposes a new approach of multiple density clustering to overcome the shortcomings of common density-based clustering approaches. The new technique is highly effective in identifying clusters of varying densities.
- *High-Throughput Publish/Subscribe on Top of LSM-based Storage* by Mohiuddin Abdul Qader (UC Riverside) and Vagelis Hristidis (UC Riverside) presents a high-throughput publish/subscribe system that also supports efficient self-joining subscriptions. Tests on real datasets demonstrate that the new system is able to serve subscriptions much more efficiently than state-of-art implementations.
- *Incrementally Updating Unary Inclusion Dependencies in Dynamic Data* by Nuhad Shaabani (Hasso-Plattner-Institut) and Christoph Meinel (Hasso-Plattner-Institut) presents an incremental approach for updating inclusion dependencies. This is particularly important because none of the existing techniques for detecting inclusion dependencies in data sets can work effectively with dynamic data. Tests show that the incremental approach is able to reduce the runtime by five orders of magnitude compared to static approaches.
- *PLI+ Efficient Clustering of Cloud Databases* by Dai-Hai Ton That (DePaul University), James Wagner (DePaul University), Alexander Rasin (DePaul University), and Tanu Malik (DePaul University) introduces the Physical Location Index Plus (PLI+) for large databases hosted on commercial cloud systems. It maps a range of physical co-locations with a range of attribute values to create approximately sorted buckets. Tests show that PLI+ is able to answer queries effectively while keeping the index sizes modest.
- *DeStager: Feature Guided In-Situ Data Management in Distributed Deep Memory Hierarchies* by Xuechen Zhang (Washington State Univ, Vancouver), Fang Zhang (IBM Watson), and Bai Nguyen (Washington State Univ, Vancouver) explores the deep memory hierarchy available on large high-performance computing systems to improve in situ data analyses. It captures features of the data dynamically to assist with adaptive index creation and data placement. These feature-guided optimizations allow the proposed DeStager to substantially improve the in situ processing pipelines.

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