Distributed and Parallel Databases (2019) 37:323–327
https://doi.org/10.1007/s10619-019-07262-1

A special issue in extending data warehouses to big data analytics

Ladjel Bellatreche[1] · **Sharma Chakravarthy**[2]

Published online: 26 February 2019

In a very short time (from 1992 to date), the data warehouse technology has gone through all the phases of a technological product's life: research, introduction on the market, growth, maturity and paving way for new extensions. Maturity means there is a clearly identified design life cycle plus a race and competition between companies for their use, to increase their decision-making power. Another specificity of this technology is its adaptability over time, its capacity to address new types of data sources. Maturity and the need to move on was signaled by the appearance of Big Data. As always, some people have already announced the demise of the data warehouse, in the same way as it was predicted for relational DBMS after the appearance of NoSQL. It is therefore essential to find other challenges that will contribute to the revival and subsumption of data warehouses into newer requirements and challenges.

Big Data is one of the challenges of companies owning data warehousing technology, since they are obliged to align their business solution to Big Data requirements. This alignment comes from facing the V's brought by Big Data (Volume, Variety, Velocity, and Veracity). This situation pushes these companies to enhance their data warehouse environment with Big Data technology, including disparate data types, distributed programming, cloud computing, parallel processing and so on. As a consequence, the data warehousing community has to deal with data lakes (schema-free repositories), data warehouse design (data curation, data flow management and optimization), Big Data Management (structured, unstructured, and varied data types), modeling, query languages (SQL and beyond), analysis, parallel systems technology (Spark, HDFS), etc.

✉ Ladjel Bellatreche
ladjel.bellatreche@ensma.fr; bellatreche@ensma.fr

Sharma Chakravarthy
sharma@cse.uta.edu

[1] LIAS/ISAE-ENSMA – Poitiers University, Poitiers, France

[2] Information Technology Laboratory (IT Lab), The University of Texas at Arlington, Arlington, TX, USA

This special issue discusses efforts deployed by research community allowing data warehousing technology to benefit from Big Data dimensions and spectacular technological advances in terms of emerging hardware and parallel and distributed computations to satisfy performance requirements of decision-makers.

Originally, this special issue has been associated to the 19th International Conference on Big Data Analytics and Knowledge Discovery (DaWak), that was held in Lyon, France from 28 to 31 August 2017. Over the past years, DaWaK has become one of the most important international scientific events to bring together researchers, developers and practitioners to discuss the latest research issues and experiences in developing and deploying new generation of data warehouses and knowledge discovery systems, applications, and solutions. This year DaWak received 97 papers and the program committee finally selected 24 full papers and 11 short papers, making an acceptance rate of 25%. The accepted papers cover a number of broad research areas on both theoretical and practical aspects of new generations of data warehouses and knowledge discovery.

As a motivation to attract good papers, we have chosen highly ranked papers from DaWaK for our special issue for Distributed and Parallel Databases Journal. Out of the 24 full papers accepted in DaWak2017, three of them related to the topics of our special issue, were invited to extend their paper by at least 40% new content. Also, an open call for papers has been organized and attracted seven papers covering the different topics of DaWak 2017. In total, our special issue got 12 papers. After a second round of reviews, five papers made the final cut. Thus, the relative acceptance rate for the papers included in this special issue is extremely competitive. We congratulate the authors who submitted articles to DaWak 2017 and our special issue.

There was great response to the call for papers for this special issue. We received 12 papers from 10 countries (Algeria, Australia, Brazil, China, France, Greece, India, Saudi Arabia, Tunisia, USA.) This was a great response to the call for papers, but due to the limited space only five papers could be accepted for this special issue. These papers are authored by an outstanding roster of experts in their respective fields, and tackle various issues from different angles, requirements and interests. Their topics include: data curation, data flow management and optimization in parallel environments, query processing and performance, cost models, data intensive computation, parallel processing, machine learning, data summarization matrix, parallel array databases. These topics cover several application domains: social data analytics, data lakes, data integration, non-functional requirement measurement metrics, etc. It is useful to note that most of these papers were results of funded projects by national agencies such as Data to Decision CRC and Cooperative Research Centres Program of Australia, and National Institutes of Health and the Air Force Office of Scientific Research and NSF Career Awards, USA.

The five selected papers are summarized below:

The paper titled "Scalable machine Learning Computing a Data Summarization Matrix with a Parallel Array DBMS", authored by Carlos Ordonez, Yiqun Zhang and S. Lennart Johnson, tackles the problem of parallelizing the computation of a generalized summarization matrix, a fundamental intermediate computation for many Machine Learning algorithms. This work is a pioneer in reducing data set summarization to a simple matrix multiplication. The authors propose a system integrating a parallel array DBMS (SciDB) and the R language. This system leverages the capabilities of SciDB to accelerate and scale the computation of several fundamental machine learning models in a parallel cluster of computers, by exploiting multiple cores (CPUs or GPUs) to speed up the computation of the summarization matrix. Theoretical results showing linear speedup and low time/space complexity are given. The system was programmed in C++ and it is compared with R and Spark, showing significant performance improvements on both fronts.

The paper titled "DataSynapse: A Social Data Curation Foundry", by Amin Beheshti, Boualem Benatallah, Alireza Tabebordbar, Hamid Reza Motahari-Nezhad, Moshe Chai Barukh and Reza Nouri. The contributions of the paper aim to advance the field of social data curation: the process of breathing meaning into raw data generated on social networks and transforming it to contextualized data and knowledge, for effective consumption in social analytics and insight discovery. Authors presented the novel notion of "Knowledge Lake", i.e., a contextualized Data Lake, to provide the foundation for big data analytics by automatically curating the raw social data and to prepare them for deriving insights. Authors developed a scalable curation algorithm to transform raw social data (e.g., a Tweet in Twitter) into contextualized data and knowledge through extracting, enriching, linking, annotating and summarizing social data. The proposed general purpose social data curation foundry, namely DataSynapse, offered as an extensible and scalable microservice-based architecture and can be used in various scenarios such as detecting and/or predicting cyber bullying, fake news and intelligence/terrorism activities.

The paper titled "Optimization of Data Flow Execution in a Parallel Environment", authored by Georgia Kougka and Anastasios Gounaris focuses on accurately reflecting the response time of a data flow provided by modern business intelligence tools, such as Pentaho Data Integration, that run jobs executed in parallel in a multi-core machine environment. Data flow management and optimization are one of the crucial tasks in data warehousing applications, since they are complex, expensive, and time-consuming tasks. To capture the response time of real data flows execution in parallel and distributed environments, the authors propose a mathematical cost model that can also drive task re-ordering overcoming the limitation of existing models that suffer from their simple assumptions and generates inaccurate results in real analytical applications. On top of that, algorithms are proposed to select the best order of task executions. Intensive experiments were conducted showing improvements of up to 59% compared to the state-of-art in data flow task ordering. An initial version of this paper has been published in DaWak 2017.

The paper titled "Abstract Cost Models for Distributed Data-Intensive Computations", authored by Rundong Li, Ningfang Mi, Mirek Riedewald, Yizhou Sun and Yi Yao deals with the problem of minimizing makespan for a distributed data-intensive computation. To do so, the authors proposed simplified functional structures that capture execution time for individual phases of user applications that run on top of distributed big-data batch processing frameworks such as Hadoop MapReduce and Spark. Those functional structures can be used to automatically tune application-specific parameters, such as block sizes in matrix multiplication, as well as more general parameters such as task number and parallelism. Compared with existing machine learning approaches that model the dependency between makespan and partitioning parameters as a blackbox, the proposed approach can greatly reduce the search space and thus optimization cost. And contrary to existing cost models used by commercial and academic DBMS that require estimating the actual number of system-level operations such as random and sequential input–output operations, the proposed approach focused on input size, output size, and computation cost, simplifying modeling for the user. Low-level system details were automatically accounted for through model coefficients associated with those three variables. To evaluate the soundness of the general model and homogeneous-task model, intensive experiments were conducted on eleven systems with different hardware and software configurations to evaluate join, sorting and matrix multiplication. An initial version of this paper has been published in DaWak 2017.

The paper titled "Horizontal Fragmentation for Fuzzy Querying Database", authored by Asmaa Drissi, Safia Nait-Bahloul, Karim Benouaret and Djamal Benslimane revisits the data fragmentation problem that has been widely studied in several generations of databases. With the explosion of Big Data applications, it has become a pre-condition to design any distributed and parallel solution. The authors deal with this problem in the context of fuzzy databases that are supposed to store very large datasets in order to meet the query performance requirements. Therefore, they propose a complete data fragmentation methodology supporting a such fragmentation. This methodology is inspired from the traditional ones applied in the relational and object oriented databases by borrowing some concepts such as a predicate usage matrix and predicate affinity matrix and adapting them to fuzzy tuples. Five-steps fragmentation algorithm is proposed that includes: data preparation, construction of the tuple-predicate matrix, construction of the predicate-similarity matrix, clustering predicates and building the final fragments. Query rewriting algorithm based on generated fragments and query execution strategies are given and experimentally evaluated. This proposal is deployed into PostgreSQL DBMS (https://github.com/postgresqlf/PostgreSQL_f).

We hope readers will find the content of this special issue interesting and that it will inspire them to look further into the challenges that are still ahead before designing extended data warehouse and analytics applications in the Big Data era. We would like to thank all the authors who submitted their papers to this special issue. In addition, we are grateful for the support of various reviewers who ensured high quality of this special issue. Last but not least, we would like to thank Professors D. Agrawal and M. Mokbel, Editors-In-Chief of Distributed and Parallel Databases Journal, for accepting our proposal of a special issue focused on Extending Data Warehouses to Big Data Analytics, and for assisting us whenever required. We would like to thank very much Aiswarya Satheesan for their endless help and support. The complete International Program Committee of this special issue is listed next.

International Program Committee

- Reza Akbarinia, INRIA, Montpellier, France
- Julien Aligon, University Toulouse 1 Capitole, France
- Djamal Benslimane, LIRIS, Lyon, France
- Sadok Ben Yahia, Tallinn university of Technology, Estonia

- Maria Luisa Damiani, Università degli Studi di Milano
- Soumyava Das, Teradata, USA.
- Haytham Elghazel, Polytech Lyon, France
- Noura Faci, LIRIS, Lyon, France
- Abdelkader Hameurlain, IRIT/Toulouse, France
- Stéphane Jean, LIAS/Poitiers University, Poitiers, France
- Mehdi Kaytoue, INSA, Lyon, France
- Sofian Maabout, Labri, Bordeaux, France
- Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
- Patrick Marcel, University of Tours, France
- Rim Moussa, University of Carthage, Tunisia
- Carlos Ordonez, Houston University, USA
- Sherif Sakr, University of New South Wales, Australia
- Panos Vassiliadis, University of Ioannina, Greece
- Jia Zou, Rice University, USA.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.