Guest Editorial-DaWaK 2018 Special Issue-Trends in Big Data Analytics

Carlos Ordonez, University of Houston, USA
Ladjel Bellatreche, ISAE-ENSMA, France

1. Introduction

From the 90s 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 including functional and non-functional requirements, data sources, conceptual design, Extract, Transform, Load, logical design, deployment phase, physical design, exploitation phase, and tuning phase [5]. Knowledge discovery is a process that requires a lot of data, and that data needs to be in a reliable state before it can be subjected to the data mining process. The accumulation of enterprise data within a data warehouse that has been properly validated, cleaned, and integrated provides the best source of data that can be subjected to knowledge discovery. Data Warehousing and Knowledge Discovery (DaWaK) conference is the fruit of the marriage between data warehouse and knowledge discovery. DaWak has been launched in 1999 aimed at bringing together researchers and practitioners to discuss research issues and experience in developing and deploying data warehousing and knowledge discovery systems, applications, and solutions. From 1999 till 2014, DaWaK conferences received and accepted papers related to the topics covered by these two technologies.

With the arrival of Big Data, it was therefore essential to find other challenges that will contribute to the revival and subsumption of data warehouses and knowledge discovery 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 system technology (Spark, HDFS), etc. As a consequence, Big Data has

become an opportunity for data warehouse and knowledge discovery community to go further in proposing solutions for Big Data Analytics. In 2015, DaWak the first part of its name has been replaced by Big Data Analytics and became Big Data Analytics and Knowledge Discovery, and keeping the same acronym.

This special issue contains selected papers from the 20th International Conference on Big Data Analytics (DaWaK 2018). These papers reflect an improved topics scope truly focusing on big data analytics, instead of the ultra popular trend today: machine learning on benchmark (mostly small) data sets. DaWaK 2018 attracted 76 submissions from which 30 papers were accepted. After presentation at DaWaK in Regensburg, Germany, September 3-6, 2018, and further discussion among the PC Chairs, we invited 6 out of 30 papers to this special issue with a strict requirement to extend their paper with at least 40% new content and to carefully consider conference reviewers feedback. For this special issue, in an exploratory manner, we made no distinction between full and short papers in order to select novel, but still good, papers. Our goal was basically to avoid republishing full papers with minor technical extensions as it frequently happens. Our paper selection was based mainly on the presentation at the conference (clear contribution), the final published version (significantly better than the submission), authors response to reviewers (heeding advice) and to a lesser extent on reviews (but still carefully considered). That is, we gave authors an opportunity to improve their work considering reviewers suggestions in order to write a novel, high quality, *journal article*. After the second round of reviews, four papers made the final cut. They provide a glimpse of important research issues in Big Data Analytics today: cloud computing, graphs, privacy-preserving algorithms, and sequences.

The four selected papers are summarized below:

- The first paper [1], titled "A Cryptographic Ensemble for Secure Third Party Data Analysis: Collaborative Data Clustering Without Data Owner Participation" combines security and data mining in the context of Data Mining as a Service.

- The second article [2], whose title is "Query Processing on Large Graphs: Approaches to Scalability and Response Time Trade Offs", studies parallel processing of queries in a distributed system. This paper proposes to partition a graph into subgraphs for parallel processing instead of redistributing edges by vertex.

- The third article [4], titled "Discovering Rare Correlated Periodic Patterns in Multiple Sequences", revisits the sequence mining problem from an outlier perspective. This work takes a step beyond previous work on discovering frequent itemsets.

- Finally, the fourth paper [8], titled "CloudDBGuard: A Framework for encrypted data storage in NoSQL Wide Column Stores", studies security aspects in a cloud Hadoop data analytic system, which is the most common platform today to store big data.

2. Conclusions

Big data has brought a new research angle, including not having a database model [3], innovative storage beyond rows (e.g. columns, arrays [6]), and scale-out parallel processing [7]. Many assumptions based on a centralized data warehouse or rigid database have been weakened and even disappeared. It is fair to say big data analytics has left data warehousing and data mining research behind. The papers included in our special issue show this trend.

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 Professor Il-Yeol Song, Consulting Editor of Data and Knowledge Engineering Journal (DKE), for accepting our proposal of a special issue and for assisting us whenever required.

References

[1] N. Almutairi, F. Coenen, and K. Dures. A cryptographic ensemble for secure third party data analysis: Collaborative data clustering without data owner participation. *Data & Knowledge Engineering*, 2019.

[2] S. Das, A. Santra, J. Bodra, and S. Chakravarthy. Query processing on large graphs: Approaches to scalability and response time trade offs. *Data & Knowledge Engineering*, 2019.

[3] Xin Luna Dong and Divesh Srivastava. Big data integration. In *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, pages 1245–1248. IEEE, 2013.

[4] P. Fournier-Viger, P. Yang, Z. Li, J.C. Lin, and R.U. Kiran. Discovering rare correlated periodic patterns in multiple sequences. *Data & Knowledge Engineering*, 2019.

[5] S. Khouri, K. Semassel, and L. Bellatreche. Managing data warehouse traceability: A life-cycle driven approach. In *27th International Conference on Advanced Information Systems Engineering (CAiSE)*, pages 199–213, 2015.

[6] C. Ordonez, W. Cabrera, and A. Gurram. Comparing columnar, row and array DBMSs to process recursive queries on graphs. *Information Systems*, 63:66–79, 2017.

[7] C. Ordonez, Y. Zhang, and W. Cabrera. The Gamma matrix to summarize dense and sparse data sets for big data analytics. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 28(7):1906–1918, 2016.

[8] L. Wiese, T. Waage, and M. Brenne. Clouddbguard: A framework for encrypted data storage in nosql wide column stores. *Data & Knowledge Engineering*, 2019.