Evolving Big Data Analytics Towards Data Science

Carlos Ordonez, University of Houston, USA
Il-Yeol Song, Drexel University, USA

1 Introduction

Data Warehousing is becoming an old topic, but it represents the root big data analytics and data science. Data warehouses evolved and went through a disruptive transformation to become Big Data repositories (so-called Data Lakes), where data exploded in Volume, Velocity and Variety. This evolution brought faster approaches, tools and algorithms to load data, querying beyond SQL considering semantics, mixing text with tables, stream processing and computing machine learning models. A latter revolution brought Veracity in the presence of contradictory information and even Value, given so many options and the investment to exploit big data. Nevertheless, having so much information in a central repository enabled more sophisticated exploratory analysis, beyond multivariate statistics and queries. Over time people realized that managing so much diverse data required not only database technology, but also a more principled approach laying its foundation on one hand in mathematics (probability, machine learning, numerical optimization) and on the other hand, more abstract, highly analytic, programming (combining multiple languages, pre-processing data, integrating diverse data sources), giving birth to Data Science. This new trend is not another fad: data science is now considered a competing discipline to computer science and applied mathematics.

The Data Warehousing and Knowledge Discovery (DaWaK) conference was the "child" of the marriage between data warehouse and knowledge discovery. DaWak was launched in 1999 aimed at bringing together researchers, analysts and developers to discuss research issues and experience in developing and deploying data warehousing and knowledge discovery systems, applications, and tools. From 1999 till 2014, the DaWaK conference series received and accepted papers related to the topics covered by these two technologies. Thus in 2015, DaWak its full name was replaced by Big Data Analytics and became Big Data Analytics and Knowledge Discovery, but keeping the well-established Dawak acronym. Later, starting in 2015 the scope was expanded to accept big data papers, a trend which is now morphing again given the explosion of data, but also the evolution of software and hardware. Towards 2020 Dawak made another major leap forward to become a data science conference, heeding the big volume aspect, but also looking back at it data warehousing and data mining roots.

This special issue contains selected papers from the 21st International Conference on Big Data Analytics (DaWaK 2019). These papers reflect an expanded scope truly focusing on big data analytics and large-scale data science, instead of the ultra popular trend today: machine learning on benchmark (but generally small) data sets. DaWaK 2019 attracted 61 submissions from which 22 papers were accepted. After their presentation at DaWaK in Linz, Austria, August 26-29, 2019, and further discussion among the PC Chairs, we invited 4 out of those 22 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. Following our initiative from Dawak 2018, we made no distinction between full and short papers in order to select novel, but still good, promising, papers. This initiative allowed all papers to be on level ground. Our goal was basically to avoid republishing full papers with minor technical extensions, which does not help advancing the field. Our paper selection was based mainly on the presentation at the conference (clear contribution), the reviews (especially at least one strong accept), and the authors response to reviewers during the conference (an extra slide). 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, only two papers made the final cut. These papers provide a glimpse of important research issues in Big Data Analytics and Data Science today: a new architecture to store big data and process modeling.

Here we provide a summary of the two selected papers:

- The first article [1], titled "Design and Implementation of ETL Processes using BPMN and Relational Algebra", formalizes and extends ETL workflows using elegant BPMN diagrams, using the relational database model as a theoretical foundation.

- The second article [3], titled "Mo.Re.Farming: A Hybrid Architecture for Tactical and Strategic Precision Agri-

culture" presents an interesting big data architecture to capture spatial data sets and decision support for farmers in Italy.

2 Conclusions

Big data has brought a new research angle, including not requiring a database model [2], innovative storage beyond rows (e.g. columns, arrays [4]), and scale-out parallel processing [5]. 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 evolved leaving data warehousing and data mining research as history, paving the way for data science. The papers included in our special issue show this trend.

We hope DKE 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 in the evolution of Big Data Analytics towards Data Science. 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 Peter Chen, for supporting this special issue at DKE and his collaboration with the Dawak conference series.

References

[1] Judith Awiti, Alejandro A. Vaisman, and Esteban Zimanyi. Design and implementation of etl processes using bpmn and relational algebra. *Data & Knowledge Engineering*, 2020.

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

[3] Enrico Gallinucci, Matteo Golfarelli, and Stefano Rizzi. Mo.re.farming: A hybrid architecture for tactical and strategic precision agriculture. *Data & Knowledge Engineering*, 2020.

[4] 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.

[5] 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.

Carlos Ordonez got a Ph.D. degree in Computer Science from the Georgia Institute of Technology, USA, in 2000. From 2001 to 2006 Dr Ordonez worked on extending the Teradata DBMS with machine learning and cube techniques. Then in 2006 Dr. Ordonez joined the University of Houston, where he condcuts research on parallel data processing systems. Dr Ordonez was a visiting researcher at MIT from 2014 to 2016 working on array and columnar database systems. The Dr Ordonez worked research scientist at ATT Labs from 2014 to 2015, focusing on the R language. His research is centered on large-scale data science, parallel database systems and big data. Dr Ordonez research has produced over 120 papers, over 3500 citations and has been funded by NSF grants.

Il-Yeol Song is professor in the College of Computing and Informatics of Drexel University. He served as Deputy Director of NSF-sponsored research center on Visual and Decision Informatics (CVDI) between 2012-2014. His research topics include conceptual modeling, data warehousing, big data management and analytics, and smart aging. He is an ACM Distinguished Scientist and an ER Fellow. He is the recipient of 2015 Peter P. Chen Award in Conceptual Modeling. Dr. Song published over 200 peer-reviewed papers in data management areas. He is a co-Editor-in-Chief of Journal of Computing Science and Engineering (JCSE) and Consulting Editor for Data and Knowledge Engineering. He won the Best Paper Award in the IEEE CIBCB 2004. He won four teaching awards from Drexel, including the most prestigious Lindback Distinguished Teaching Award. Dr. Song served as the Steering Committee chair of the ER conference between 2010-2012. He delivered a keynote speech on big data at the First Asia-Pacific iSchool Conference in 2014, ACM SAC 2015 conference, ER2015 Conference, EDB 2016, and A-LIEP 2016 Conference.