Special Issue on DOLAP 2015: Evolving Data Warehousing and OLAP Cubes to Big Data Analytics

Carlos Ordonez, University of Houston
Carlos Garcia-Alvarado, University of Houston
Il-Yeol Song, Drexel University

1. Introduction

We welcome the reader to the special issue containing best papers from the ACM International Workshop on Data Warehousing and OLAP (DOLAP) 2015. We have revamped DOLAP to be a research venue for big data analytics, expanding its scope, but maintaining its high quality and applied focus.

2. Best papers from DOLAP 2015

DOLAP attracted 31 submissions from which 8 papers were accepted as full papers. After presentation at DOLAP in Melbourne, Australia and further discussion among the PC Chairs, we have chosen to include three high quality papers in this special issue [1, 6, 9] that are representative of the "Big Data Analytics" evolution.

- The first paper [1], titled "Category- and selection-enabled nearest neighbor joins", is a database processing paper. The paper proposes an efficient algorithm (roNNJ) to compute a nearest neighbor join with grouping attributes and selection predicates. The authors extend the state-of-the-art for which nearest neighbor joins that are not category-enabled must process each category independently. As a result state-of-the-art algorithms end up fetching, either from disk or memory, the blocks of the input relations multiple times. The newly proposed algorithm finds, for each outer tuple all the inner tuples that satisfy the equality on the category and have the smallest distance to the outer tuple. The new algorithm does only one scan of both inputs. The proposed algorithm is evaluated as a new physical operator with a similar set of states to those of the traditional Sort Merge Join operator (implemented in PostgreSQL). roNNJ computes the results in a single scan of the input relations as the result of a roNNJ query tree that is not dependent on the physical organization of the fact table. Finally, the authors prove that their approach has an upper bound time complexity of $O(n \log n)$ and their solution is applicable in both row and column stores.

- The second article [6], titled "SETL: A Programmable Semantic Extract-Transform-Load Framework for Semantic Data Warehouses" bridges database design and database processing. This paper presents research on data warehouse design and how organizations are dealing with diverse external data coming from multiple sources in a structured, semi-structured, or unstructured form. The authors extend the state-of-the-art by considering semantic issues during the Extract-Transform-Load (ETL) phase. The proposal presents a programmable Semantic ETL (SETL) framework that builds on top of the Semantic Web standards and produces a set of modules, classes, and methods for building an integrated data warehouse. The framework defines a target ontology based on the domain of interest, extracts data from multiple heterogeneous data sources, transforms the source data into RDF triples, loads the data into a triple store, and publishes the semantic (multidimensional) data warehouse as a knowledge base.
- Finally, the third paper [9], titled "Eco-physic: Eco-Physical design initiative for very large databases" is a database processing paper showing how to support "green computing." This paper revisits the physical design of a data warehouse considering energy consumption. This optimization problem is relevant for data processing at-scale (e.g data centers) and has become a resource and a budget concern for organizations. As such, the authors extend the state-of-the-art by proposing a methodology called Eco-DMW that integrates the energy dimension into the physical design when selecting materialized views. In a nutshell, the authors' proposal is based on capturing the power consumption cost model that is implemented in conjunction with PostgreSQL's cost model in order to predict query power consumption. Furthermore, the authors adapted a genetic algorithm to solve the multi-objective optimization problem that results from selecting a set of views constrained by an upper bound on power consumption. The outcome of this approach is a set of views that provide high performance within user-specified power consumption limits.

3. Conclusions

Big data has brought new perspectives and unforeseen problems, including the absence of a database model [3], novel storage techniques (e.g. columnar, array [7]), and scale-out parallel processing [8]. Many assumptions about data warehousing have been relaxed and even disappeared, depending on the specific problem, system, application or tool: big data analytics has subsumed data warehousing research. The papers presented in this special issue reflect this exciting trend.

References

[1] F. Cafagna and M.H. Böhlen. Category- and selection-enabled nearest neighbor joins. *Information Systems*, 2017.

[2] Z. Chen and C. Ordonez. Efficient OLAP with UDFs. In *Proc. ACM DOLAP Workshop*, pages 41–48, 2008.

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

[4] H. Garcia-Molina, J.D. Ullman, and J. Widom. *Database Systems: The Complete Book*. Prentice Hall, 1st edition, 2001.

[5] E. Malinowski and E. Zimányi. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. Springer, 1st edition, 2008.

[6] R.P. Deb Nath, K. Hose, T.B. Pedersen, and O. Romero. SETL: A programmable semantic extract-transform-load framework for semantic data warehouses. *Information Systems*, 2017.

[7] 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.

[8] 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.

[9] A. Roukh, L. Bellatreche, S. Bouarar, and A. Boukorca. Eco-physic: Eco-Physical design initiative for very large databases. *Information Systems*, 2017.