EDITORIAL

Special issue on DOLAP 2017: Design, Optimization, Languages and Analytical Processing of Big Data

This special issue of Information Systems contains the best papers from the Nineteenth International Workshop On Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2017), held in conjunction with the 20th International Conference on Extending Database Technology (EDBT 2017), in Venice, Italy, during March 21-24, 2017.

This 19th edition marked a turning point in the history of DOLAP, with a change of name reflecting the evolution of Business Intelligence, from the decision support systems that are "data presenting" to more dynamic systems that allow the semi-automation of the decision making process. In an era where data analysts spend 60% of their time on data wrangling [5], complain about the lack of tools that scale well with their dataset size for handling data heterogeneity and ad-hoc grouping [6], cite SQL as the third commonly used tool in data science projects and relational data as the most commonly type of data used at work [5], and where supporting Interactive Data Exploration receives increasing attentions [7, 8], it is obvious that DOLAP still has significant roles to play in data analysis on Big Data.

After presentation at DOLAP in Venice, 6 of the 7 full papers have been invited for submission to this Information Systems special issue. After three rigorous review rounds, only 4 papers have been accepted for final publication in this special issue. These papers embody the new acronym of DOLAP, in that they address issues related to Design [1,2], Optimization [3], Language [2], Analytical Processing [4] and various facets of Big Data [1,3,4].

The first paper [1] titled *An Integration-Oriented Ontology to Govern Evolution in Big Data Ecosystems* by Sergi Nadal, Oscar Romero, Alberto Abelló, Panos Vassiliadis and Stijn Vansummeren tackles the problem of data integration in the presence of schema changes. They use the ontology-based data access (OBDA) approach to enable end-users to pose ontology-mediated queries to an integrated set of heterogeneous and disparate data sources. Their approach couples OBDA with local-as-view (LAV) schema mappings that characterize elements of the source schemata in terms of a query over the ontology. Practically, the approach introduces a structured RDF-phrased ontology to model and integrate evolving data. The data sources are accommodated via a traditional mediator/wrapper architecture, where wrappers with relational schema handle most of the query complexity while the ontology determine how to join sources and what attributes are projected. The authors provide an algorithm for translating queries against the global schema into queries against wrappers. Queries are restricted to a subset of SPARQL queries, to ensure that this translation can be computed without computationally expensive reasoning. A theoretical and practical study of the algorithm's complexity and limitations is presented.

The second paper [2] titled *The Percentage Cube* by Yiqun Zhang, Carlos Ordonez, Javier García-García, Ladjel Bellatreche and Humberto Carrillo revisits data cube definition and implementation to show the relationship between measures aggregated at different grouping levels. The authors define a Percentage Cube that, contrarily to the classical data cube containing aggregations along possible aggregation dimensions, adds percentage information as measures relating other measures with aggregated totals along different possible combinations. They show efficient methods to build the percentage cube, which consists of the pre-calculation of a set of cuboids that include the percentage measures. They characterize the difficulty of this problem as doubly exponential in the number of dimensions. They introduce minimal SQL syntax extensions to compute percentage queries and to materialize the percentage cube. They investigate two different alternatives for the computation of the percentage cube, namely OLAP functions and GROUP-BY using standard aggregate functions. They optimize the computation by extending previous pruning techniques coming from the evaluation of iceberg queries. They use the classical TPC-H Benchmark to evaluate the approaches, comparing the alternatives and concluding that the GROUP-BY method is the most appropriate one.

The third paper [3] titled *MapReduce Performance Models for Hadoop 2.x* by Daria Glushkova, Petar Jovanovic and Alberto Abello is a database processing paper that proposes a cost model for the processing of map-reduce queries in the architecture introduced by Hadoop 2.x. As this second version of Hadoop decouples the programming model from the resource management infrastructure, resources are dynamically allocated, making the cost models relying on a static resource allocation as in Hadoop 1.x no longer applicable. The authors then identify the factors that potentially affect the cost of job execution in Hadoop 2.x, and define a performance model accordingly, by extending the performance model for MapReduce workloads proposed for Hadoop 1.x. The new performance model is oriented to capture the scheduling tree of the tasks, i.e., the precedence of different tasks of MapReduce jobs as well as the synchronization delays due to shared resources, giving less attention to the determination of the actual cost of each task. The model of performance is implemented and its theoretical complexity is given. The model is empirically tested on two classical types of jobs, Wordcount and Sort, showing an accurate estimation of these job costs.

Finally, the fourth paper [4] titled *EXODuS: Exploratory OLAP over Document Stores* by Mohamed L. Chouder, Stefano Rizzi, Rachid CHALAL introduces an approach enabling the analytical processing of schemaless document stores. This approach, named EXODuS, follows the schema-on-read method to leave data unchanged in their structure until they are accessed. The approach operates in three phases. The first phase extracts a draft multidimensional schema from the documents, classifying attributes into levels and measures based on their type and relating measures to the subset of levels that determine their granularity by checking multidimensional constraints. The second phase dynamically checks the well-formedness of an initial user query by extracting approximate functional dependencies from the data. The schema obtained during phase one guides the extraction thus avoiding the exponential complexity of checking all dependencies. This schema also allows to recommend levels to the user. Finally, to enable the use of OLAP primitives, the third phase incrementally builds portion of hierarchies based on the levels appearing in the initial user query by extracting approximate functional dependencies. The approach is implemented by translating user queries into the native query language of the MongoDB document store to retrieve JSON data. The performance of EXODuS is evaluated over three real-world datasets, revealing that the approach is in line with the requirements of a real-time user interaction.

The Guest Editors would like to express their gratitude to the Editors-In-Chief of Information Systems, Prof. Felix Nauman, Prof. Dennis Shasha and Prof. Gottfried Vossen, for accepting their proposal of this special issue and for assisting them whenever required. The Guest Editors would also like to warmly thank the reviewers and the authors who together contributed to the quality of the papers published in this special issue.

REFERENCES

[1] Sergi Nadal, Oscar Romero, Alberto Abelló, Panos Vassiliadis, Stijn Vansummeren. An Integration-Oriented Ontology to Govern Evolution in Big Data Ecosystems. Information Systems.

[2] Yiqun Zhang, Carlos Ordonez, Javier García-García, Ladjel Bellatreche, Humberto Carrillo. The Percentage Cube. Information Systems.

[3] Daria Glushkova, Petar Jovanovic, Alberto Abello. MapReduce Performance Models for Hadoop 2.x. Information Systems.

[4] Mohamed L. Chouder, Stefano Rizzi, Rachid Chalal. EXODuS:

Exploratory OLAP over Document Stores. Information Systems.

[5] https://www.kaggle.com/surveys/2017

[6] S.  Kandel,  A.  Paepcke,  J.  M.  Hellerstein,  and  J. Heer. Enterprise  data analysis and visualization:  An interview study. IEEE Trans. Vis. Comput.Graph., 18(12):2917–2926, 2012.

[7] S. Idreos, O. Papaemmanouil, and S. Chaudhuri.  Overview of data exploration techniques. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, pages 277–281, 2015.

[8] T. Milo and A. Somech. Next-step suggestions for modern interactive data analysis platforms. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 576–585, 2018.

GUEST EDITORS

Patrick Marcel

University of Tours, France

E-mail address: patrick.marcel@univ-tours.fr

Il-Yeol Song

Drexel University, USA

E-mail address: songiy@drexel.edu