Data Warehousing and OLAP over Big Data: Current Challenges and Future Research Directions

Alfredo Cuzzocrea
ICAR-CNR and University of Calabria
Rende, Cosenza, Italy
cuzzocrea@si.deis.unical.it

Ladjel Bellatreche
LIAS/ISAE-ENSMA
Futuroscope, France
bellatreche@ensma.fr

Il-Yeol Song
Drexel University
Philadelphia, PA, USA
songiy@drexel.edu

ABSTRACT
In this paper, we highlight open problems and actual research trends in the field of *Data Warehousing and OLAP over Big Data*, an emerging term in Data Warehousing and OLAP research. We also derive several novel research directions arising in this field, and put emphasis on possible contributions to be achieved by future research efforts.

Categories and Subject Descriptors
H.2 [Database Management]: H.2.7 Database Administration – Data Warehouse and Repository

General Terms
Algorithms, Design, Management, Performance, Theory

Keywords
Big Data, Big Multidimensional Data, Data Warehousing, OLAP

1. INTRODUCTION
A strong interest towards the term "*Big Data*" is arising in the literature actually (e.g., [12,17,18]). This term identifies specific kinds of data sets, manly of *unstructured data*, which populate the data layer of scientific computing applications (e.g., [21]). Data stored in the underlying layer of all these application scenarios have some specific characteristics in common, among which we recall [1]: (*i*) *large-scale data*, which refers to the size and the distribution of data repositories; (*ii*) *scalability issues*, which refers to the capabilities of applications running on large-scale, enormous data repositories (i.e., big data, for short) to scale over growing-in-size inputs rapidly; (*iii*) supporting *advanced Extraction-Transformation-Loading* (ETL) *processes* from low-level, raw data to somewhat *structured information*; (*iv*) designing and developing *easy and interpretable analytics* over big data repositories in order to derive intelligence and extract useful knowledge from them.

DOLAP'13, October 28, 2013, San Francisco, CA, USA.

Due to the intrinsic nature of Big Data application scenarios (e.g., [25]), it is natural to adopt *Data Warehousing and OLAP methodologies* [2] with the goal of collecting, extracting, transforming, loading, warehousing and OLAPing such kinds of datasets, by adding significant add-ons supporting *analytics over Big Data* (e.g., [11,1,21]), an emerging topic in Database and Data Warehousing research.

Data Warehousing and OLAP are classical scientific fields which have been addressed since several decades by the Database and Data Warehousing research community. Symmetrically, the fundamental problem of *computing OLAP data cubes* has been contextualized in a wide family of types, ranging from classical *relational data sets* (e.g., [3]) to *graph data sets* (e.g., [4]), and from *XML data* (e.g., [5]) to novel *social network data* (e.g., [6]), and so forth.

With the advent of the Big Data research context, it is natural to think of the problem of *computing OLAP data cubes over Big Data* as one of the top-interesting challenges in the research community, with also powerful technological achievements to be reached within the scope of real-life *large-scale data-intensive applications and systems*.

Unfortunately, despite the clear convergence, state-of-the-art solutions are not capable to deal with computing OLAP data cubes over Big Data, mainly due to two intrinsic factors of Big Data repositories: (*i*) *size*, which becomes really explosive in such data sets; (*ii*) *complexity* (of *multidimensional data models*), which can be very high in such data sets (e.g., cardinality mappings, irregular hierarchies, dimensional attributes etc.).

As a consequence, there emerge the forceful needs of designing novel models, techniques, algorithms and computational platforms for supporting the problem of computing OLAP data cubes over Big Data, which, indeed, literally represents an effective call to arms for next-generation Data Warehousing and OLAP research.

Inspired by this main motivation, in this paper we highlight open problems and actual research trends in the field of Data Warehousing and OLAP over Big Data (Section 2), (2) derive several novel research directions arising in this field, and (3) put emphasis on possible contributions to be achieved by future research efforts.

2. OPEN RESEARCH PROBLEMS OF DATA WAREHOUSING AND OLAP OVER BIG DATA

Several research problems arise when computing OLAP data cubes over Big Data. Among these, we identify the following ones:

- *size*: fact tables can easily become huge when computed over Big Data sets – this adds severe computational issues as the size can become a real bottleneck from practical applications (e.g., [22]);
- *complexity*: building OLAP data cubes over Big Data also implies complexity problems which do not arise in traditional OLAP settings (e.g., in relational environments) – for instance, the number of dimensions can really become *explosive*, due to the strongly unstructured nature of Big Data sets, as well as there could be *multiple* (and *heterogeneous*) *measures* for such data cubes;
- *design*: *design methodologies* for OLAP data cubes have been of relevant interest for Database and Data Warehousing research. In the specific case of designing methodologies of OLAP over Big Data, the *performance aspect* must be taken into greater consideration, due to obvious spin-offs given by such design task – in this case, designers must move the attention on the following critical questions: (*i*) what is the *overall building time* of the data cube to be designed (computing aggregations over Big Data may become prohibitive)?; (*ii*) how the data cube should be updated? – which *maintenance plan* should be selected?; (*iii*) which *building strategy* should be adopted (e.g., *divide & conquer* (e.g., [7])?;
- *computing methodologies*: due to the enormous size, computing OLAP data cubes over Big Data will turn (again!) into a challenging research problem, similarly to what happened for early OLAP data cube computing experiences (e.g., [31]) – in this case, the most promising technology to follow seems to be the emerging *Cloud Computing* paradigm, perhaps inspired by classical *parallel computing* methodologies (e.g., [8,16]);
- *in-memory representation*: how an OLAP data cube over Big Data should be mapped in memory? This is a critical challenge to be considered, due to the fact that the very high number of dimensions in such cubes easily convey to explosive cell cardinalities – as a consequence, solutions based on tertiary memory should be deeply investigated;
- *innovative hardware support*: it is natural to figure-out that *innovative hardware solutions*, such as *GPU-based data processing* (e.g., [9]), will play an important role with respect to the issue of computing OLAP data cubes over Big Data;
- *query languages and optimization*: classical *MDX approaches* do not incorporate *optimization solutions* prone to deal with Big Data needs; future investigations must focus the attention on optimization issues given by processing Big Data in a multidimensional fashion;
- *end-user performance*: OLAP data cubes computed over Big Data tend to be huge, hence end-user performance easily becomes poor on such cubes, especially during the *aggregation* and *query phases* – therefore, it follows that end-user performance must be included as a critical factor within the *design process* of OLAP data cubes over Big Data;
- *quality*: quality aspects will more and more become a critical factor in next generation Data Warehousing and OLAP methodologies over Big Data – in fact, due to the strongly unstructured nature of Big Data sources, aggregations computed on such data sources can easily turn out to be "poor"; hence, it is easy to understand how much important controlling the quality of final data cubes will become;
- *usability*: OLAP data cubes over Big Data must, prominently, be processed and managed to extract and build useful analytics – this aspect opens the door to a wide family of research problems, such as devising methodologies to "measure" how much usable an OLAP data cube built on Big Data repositories is;
- *visualization*: as Big Data expose explosive size, visualization issues of OLAP data cubes (e.g., [27,28]) over Big Data play a first-class role in this research field – as a consequence, a novel class of *visualization metaphors*, *methodologies* and *solutions* must be devised, in order to cope with emerging challenges posed by visualizing massive OLAP data cubes over Big Data; real-time visualization of extracted core data, visualization of mashuped data, and effective visualization over mobile devices should also be considered;
- *interactive exploration*: coupled with visualization issues, *interactive exploration issues* are severe milestones to traverse in the context of OLAP data cubes over Big Data research – in fact, enormous-in-size data cubes are difficult to explore (e.g., under the execution of a fixed analytical process) while extracting useful knowledge, with important implications such as *conceptual navigation*, *concept drift*, *interaction metaphors* (e.g., [11]), and so forth;
- *analytics*: analytics over Big Data (cubes [1]) represent a topic of emerging interest for the Database and Data Warehousing research community – in this case, there exist several problems to be investigated, running from how to design an analytical process over OLAP data cubes computed on top of Big Data to how to optimize the execution of so-obtained analytical processes, and from the seamless integration of OLAP (Big) data cubes with other kinds of unstructured information (within the scope of analytics);
- *integration with classical data-intensive platforms*: an important issue is represented by how to *integrate* models, techniques, algorithms and computational platforms devised for OLAP over Big Data with classical data-intensive platforms, in the view of a seamless vision of comprehensive large-scale data-intensive systems;
- *development tools*: last but not least, *suitable tools* for supporting the design and the development of OLAP data cubes over Big Data, according to a methodology able of

incorporating all the aspects discussed above, represent a non-secondary challenge to be dealt with.

3. FUTURE RESEARCH DIRECTIONS OF DATA WAREHOUSING AND OLAP OVER BIG DATA

From the analysis of open research problems of Data Warehousing and OLAP over Big Data, several future research directions to be considered turn out. Among these, we highlight the following ones:

- *innovative methodologies for designing OLAP data cubes over Big Data*: there is a strong need for methodologies capable of dealing with requirements posed by designing and modeling OLAP data cubes over Big Data;
- *innovative solutions for computing aggregations*: computing aggregations, which is an annoying problem in classical Data Warehousing and OLAP research, gets worse when considered in the context of OLAP data cubes over Big Data – from this, it follows an emerging need for innovative solutions capable of dealing with challenging requirements of OLAP (Big) data cubes, such as *curse of dimensionality*, *"irregular" data sets* (e.g., by numerousness of dimensional members), *multi-way aggregations*, and so forth;
- *novel computational paradigms for effectively and efficiently computing OLAP data cubes over Big Data*: computing OLAP data cubes over Big Data is very resource-consuming, hence computational paradigms for innovative aspects (e.g., *context-aware resource scheduling* (e.g., [23]) are necessary to this end;
- *powerful high-performance architectures for implementing OLAP data cubes over Big Data*: the exploitation of hardware solutions for supporting the implementation of OLAP data cubes over Big Data (e.g., GPU [9]) is a promising direction for next generation scientific computing applications;
- *complex OLAP data cubes over Big Data*: due to the intrinsic complexity of Big Data sets, it follows the need for defining and exploiting *complex* OLAP data cubes over Big Data, tailored to support advanced data-intensive large-scale scientific applications (e.g., [26]);
- *"customizable" MDX predicates*: OLAP data cubes over Big Data must also be *flexible*, due to the requirements posed by modern scientific applications – in this respect, the classical MDX language for querying multidimensional data should be extended as to incorporate *"customizable" predicates* catching the necessary flexibility in OLAP (Big) data cube processing;
- *semantically-rich OLAP (Big) data cubes*: exploiting *semantics-based techniques* for modeling classical OLAP data cubes has been a successful experience in the context of next-generation complex information systems – similarly, we believe that these techniques can provide significant achievements in the context of OLAP (Big) data cubes as well, due to the fact semantics-based methods (e.g., *Ontologies* [19,20]) can improve the access, browsing and delivery experiences over such data cubes;
- *process-oriented definition languages for analytics*: analytics define complex functions over very-large amounts of data, even with the exploitation of modern *NoSQL platforms* [14] – this activity may turn to be very problematic when OLAP (Big) data cubes are considered so that it follows the need for devising rigorous process-oriented definition languages for analytics over OLAP (Big) data cubes with the goal of achieving *standardization* and *interoperability*;
- *security and privacy issues*: security and privacy aspects, which are relevant for classical OLAP data cubes as well (e.g., [13,15]), play a very relevant role for the case of OLAP data cubes over Big Data, due to the fact these structures are "open" by definition – as a consequence, future efforts must focus on the emerging requirement of enforcing and emphasizing the security and the privacy of OLAP (Big) data cubes in open information systems and over the Cloud;
- *applications*: finally, due to the same nature of OLAP (Big) data cubes, applications over these structures play a first-class role – mostly, future research efforts should be focused on testing the effectiveness and the reliability of OLAP (Big) data cubes in a wide range of application scenarios ranging from bio-medical applications to social networks, from Cloud-enabled systems to sensor- and stream-based frameworks (e.g., [29,30]), and so forth.

4. CONCLUSIONS

In this paper we have provided critical discussion over open research issues and future research directions in the context of Data Warehousing and OLAP over Big Data, by highlighting actual limitations and possible solutions. Several research directions for future investigation have been drawn as well.

5. REFERENCES

[1] Cuzzocrea, A., Song, I.-Y., and Davis, K.C. Analytics over Large-Scale Multidimensional Data: The Big Data Revolution! *Proc. of ACM DOLAP*, 2011.

[2] Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichart, D., Venkatrao, M., Pellow, F., and Pirahesh, H. Data Cube: A Relational Aggregation Operator Generalizing Group-by, Cross-Tab, and Sub Totals. *Data Mining and Knowledge Discovery 1(1)*, 1997.

[3] Harinarayan, V., Rajaraman, A., and Ullman, J.D. Implementing Data Cubes Efficiently. *Proc. of SIGMOD Conference*, 1996.

[4] Chen, C., Yan, X., Zhu, F., Han, J., and Yu, P.S. Graph OLAP: A Multi-Dimensional Framework for Graph Data Analysis. *Knowledge and Information Systems 21(1),* 2009.

[5] Jensen, M.R., Møller, T.H., and Pedersen, T.B. Specifying OLAP Cubes on XML Data. *Proc. of SSDBM*, 2001.

[6] Zhao, P., Li, X., Xin, D., and Han, J. Graph Cube: On Warehousing And OLAP Multidimensional Networks. *Proc. of ACM SIGMOD*, 2011.

[7] Yuan, Y., Lin, X., Liu, Q., Wang, W., Yu, J.X., and Zhang, Q. Efficient Computation of the Skyline Cube. *Proc. of VLDB*, 2005.

[8] Dehne, F.K.H.A., Eavis,T., and Rau-Chaplin, A. The cgmCUBE Project: Optimizing Parallel Data Cube Generation for ROLAP. *Distributed and Parallel Databases 19(1)*, 2006.

[9] Sitaridi, E.A., and Ross, K.A. Ameliorating Memory Contention of OLAP Operators on GPU Processors. *Proc. of ACM DaMoN*, 2012.

[10] Sarawagi, S., Agrawal, R., and Megiddo, N. Discovery-Driven Exploration of OLAP Data Cubes. *Proc. of EDBT*, 1998.

[11] Abouzeid, A., Bajda-Pawlikowski, K., Abadi, D.J.. Rasin, A., and Silberschatz, A. HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads. *PVLDB 2(1)*, 2009.

[12] Agrawal, D., Das, D., and El Abbadi, A. Big Data and Cloud Computing: Current State and Future Opportunities. *Proc. of EDBT*, 2011.

[13] Cuzzocrea, A., and Bertino, E. Privacy Preserving OLAP over Distributed XML Data: A Theoretically-Sound Secure-Multiparty-Computation Approach. *Journal of Computer and System Sciences 77(6)*, 2011.

[14] Cattell, R. Scalable SQL and NoSQL Data Stores. *SIGMOD Record 39(4)*, 2010.

[15] Cuzzocrea, A., and Saccà, D. Balancing Accuracy and Privacy of OLAP Aggregations on Data Cubes. *Proc. of DOLAP*, 2010

[16] Bellatreche, L., Cuzzocrea, A., and Benkrid, S. Effectively and Efficiently Designing and Querying Parallel Relational Data Warehouses on Heterogeneous Database Clusters: The F&A Approach. *Journal of Database Management 23(4)*, 2012.

[17] Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J.M., and Welton, C. MAD Skills: New Analysis Practices for Big Data. *PVLDB 2(2)*, 2009.

[18] Dean, J., and Ghemawat, S. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM 51(1)*, 2008.

[19] Khouri, S., Bellatreche, L., and Berkani, N. MODETL: A Complete MODeling and ETL Method for Designing Data Warehouses from Semantic Databases. *Proc. of COMAD*, 2012.

[20] Khouri, S., and Bellatreche, L. DWOBS: Data Warehouse Design from Ontology-Based Sources. *Proc. of DASFAA*, 2011

[21] Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., and Babu, S. Starfish: A Self-Tuning System for Big Data Analytics. *Proc. of CIDIR*, 2011.

[22] Jiang, D., Ooi, B.C., Shi, L., and Wu, S. The Performance of MapReduce: An In-depth Study. *PVLDB 3(1)*, 2010.

[23] Thusoo, A. Sarma, J.S., Jain, N., Shao, Z., Chakka, P. Zhang, N., Antony, S., Liu, H., and Murthy, R. Hive – A Petabyte Scale Data Warehouse Using Hadoop. *Proc. of ICDE*, 2010.

[24] Bizer, C., Boncz, P.A., Brodie, M.L., and Erling, O. The Meaningful Use of Big Data: Four Perspectives - Four Challenges. *SIGMOD Record 40(4)*, 2011

[25] Chen, Y., Alspaugh, S., and Katz, R.H. Interactive Analytical Processing in Big Data Systems: A Cross-Industry Study of MapReduce Workloads. *PVLDB 5(12)*, 2012

[26] Cuzzocrea, A., Saccà, D., and Serafino, P. Semantics-Aware Advanced OLAP Visualization of Multidimensional Data Cubes. *International Journal of Data Warehousing and Mining 3(4)*, 2007

[27] Cuzzocrea, A., Saccà, D., and Serafino, P. A Hierarchy-Driven Compression Technique for Advanced OLAP Visualization of Multidimensional Data Cubes. *Proc. of DaWaK*, 2006.

[28] Cuzzocrea, A. Retrieving Accurate Estimates to OLAP Queries over Uncertain and Imprecise Multidimensional Data Streams. *Proc. of SSDBM*, 2011.

[29] Cuzzocrea, A., and Chakravarthy, S. Event-based Lossy Compression for Effective and Efficient OLAP over Data Streams. *Data and Knowledge Engineering 69(7)*, 2010

[30] Cuzzocrea, A. Providing Probabilistically-Bounded Approximate Answers to Non-Holistic Aggregate Range Queries in OLAP. *Proc. of ACM DOLAP*, 2005.