Analytics over Large-Scale Multidimensional Data: The Big Data Revolution!

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

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

Karen C. Davis
University of Cincinnati
Cincinnati, OH, USA
karen.davis@uc.edu

ABSTRACT
In this paper, we provide an overview of state-of-the-art research issues and achievements in the field of *analytics over big data*, and we extend the discussion to *analytics over big multidimensional data* as well, by highlighting open problems and actual research trends. Our analytical contribution is finally completed by several novel research directions arising in this field, which plays a leading role in next-generation Data Warehousing and OLAP research.

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
Analytics over Big Data, Analytics over Big Multidimensional Data, Data Warehousing, OLAP

1. INTRODUCTION
"*Big Data*" refers to enormous amounts of *unstructured data* produced by high-performance applications falling in a wide and heterogeneous family of application scenarios: from scientific computing applications to social networks, from e-government applications to medical information systems, and so forth. Data stored in the underlying layer of all these application scenarios have some specific characteristics in common, among which we recall: (*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.

Cloud computing is a successful computational paradigm for managing and processing big data repositories, mainly because of its innovative metaphors known under the terms "*Database as a Service*" (DaaS) [10] and "*Infrastructure as a Service*" (IaaS). DaaS defines a set of tools that provide final users with seamless mechanisms for creating, storing, accessing and managing their proper databases on remote (data) servers. Due to the naïve features of big data, DaaS is the most appropriate computational data framework to implement big data repositories [2]. *MapReduce* [8] is a relevant realization of the DaaS initiative. IaaS is a provision model according to which organizations outsource infrastructures (i.e., hardware, software, network) used to support ICT operations. The IaaS provider is responsible for housing, running and maintaining these services, by ensuring important capabilities like *elasticity*, *pay-per-use*, *transfer of risk* and *low time to market*. Due to specific application requirements of applications running over big data repositories, IaaS is the most appropriate computational service framework to implement big data applications.

Among the collection of open problems and research challenges deriving from the latest *big data revolution*, analytics over big data play a relevant role in the context of Data Warehousing and OLAP research. Let us focus on this research challenge in a greater detail. Analytics can be intended as complex procedures running over large-scale, enormous-in-size data repositories (like big data repositories) whose main goal is that of extracting useful knowledge kept in such repositories. Two main problems arise, in this respect. The first one is represented by the issue of conveying big data stored in heterogeneous and different-in-nature data sources (e.g., legacy systems, Web, scientific data repositories, sensor and stream databases, social networks) into a structured, hence well-interpretable, format. The second one is represented by the issue of managing, processing and transforming so-extracted structured data repositories in order to derive *Business Intelligence* (BI) components like diagrams, plots, dashboards, and so forth, for decision making purposes. Actually, both these aspects are of emerging interest for a wide spectrum of research communities, and more properly for the Data Warehousing and OLAP research community. As a consequence, this has generated a rich literature. At the industrial research side, *Hadoop* [3] and *Hive* [13] are two fortunate implementations of the ETL layer and the BI layer of big data applications, respectively.

Although analytics over large-scale data repositories have been deeply investigated recently, the problem of extending actual models and algorithms proposed in this respect to the specific *big multidimensional data* context plays a leading role, as multidimensional data naturally marry with analytics [6].

Inspired by this main motivation, in this paper we provide an overview of state-of-the-art research issues and achievements in the field of analytics over big data, and we extend the discussion to analytics over big multidimensional data as well, by highlighting open problems and actual research trends, and drawing novel research directions in this field.

DOLAP'11, October 28, 2011, Glasgow, Scotland, UK.
ACM 978-1-4503-0963-9/11/10.

2. ANALYTICS OVER BIG DATA: STATE-OF-THE-ART

Analytics over big data repositories has recently received a great deal of attention from the research communities. One of the most significant application scenarios where big data arise is, without doubt, scientific computing. Here, scientists and researchers produce huge amounts of data per-day via experiments (e.g., think of disciplines like high-energy physics, astronomy, biology, bio-medicine, and so forth) but extracting useful knowledge for decision making purposes from these massive, large-scale data repositories is almost impossible for actual DBMS-inspired analysis tools.

In response to this "computational emergency", the Hadoop system [3] has been proposed. Hadoop runs MapReduce [8] tasks over big data, and also it makes available the *Hadoop Distributed File System* (HDFS) [3] for supporting file-oriented, distributed data management operations efficiently. It has been highlighted that Hadoop is a kind of *MAD* system [7] meaning that (*i*) it is capable of attracting all data sources (*M* standing for *Magnetism*), (*ii*) it is capable of adapting its engines to evolutions that may occur in big data sources (*A* standing for *Agility*), (*iii*) it is capable of supporting depth analytics over big data sources much more beyond the possibilities of traditional SQL-based analysis tools (*D* standing for *Depth*). In a sense, Hadoop can be reasonably considered as the evolution of next-generation Data Warehousing systems, with particular regards to the ETL phase of such systems. MapReduce is the core of Hadoop. MapReduce is a programming model with the associated computational framework that is inspired to the primitives *Map* and *Reduce* of functional languages. Basically, *Map* partitions computational tasks into smaller computational tasks (this involves in the partition of the target data domain as well) and assigns to then appropriate ⟨*Key*, *Value*⟩ pairs. These smaller computational tasks are executed very efficiently, even by exploiting parallelism. The final result of the overall computational task (i.e., the output, processed big data) is obtained via a *Reduce* operation that combines all the values sharing the same *Key* value.

Several studies, like [9,12], have provided recommendations for further improving the computational capabilities of Hadoop, whereas [1] proposes *HadoopDB*, a novel hybrid architecture that combines MapReduce and traditional DBMS technologies for supporting advanced analytics over large-scale data repositories. Furthermore, *Starfish* [11] is a recent self-tuning system for supporting big data analytics that is still based on Hadoop but it incorporates special features trying to achieve higher performance by means of *adaptive metaphors*.

By looking at BI aspects of analytics over big data, the state-of-the-art research result is represented by Hive [13], a BI system/tool for querying and managing structured data built on top of the Hadoop's HDFS. Hive which allows us to obtain the final analytics components (in the form of diagrams, plots, dashboards, and so forth) from the big data processed, materialized and stored via Hadoop. Also, Hive introduces a SQL-like query language, called *HiveQL* [13], which runs MapReduce jobs immersed into SQL statements.

3. OPEN PROBLEMS AND ACTUAL RESEARCH TRENDS OF BIG DATA ANALYTICS

There are a number of open problems and actual research trends related to big data analytics. In the following, we provide an overview on some of the most significant of them.

(*a*) Data Source Heterogeneity and Incongruence. Very often, data sources storing data of interest for the target analytics processes (e.g., legacy systems, Web, scientific data repositories, sensor and stream databases, social networks, and so forth) are *strongly heterogeneous and incongruent*. This aspect not only conveys in typical integration problems, mainly coming from active literature on *data and schema integration issues*, but it also has deep consequences on the kind of analytics to be designed.

(*b*) Filtering-Out Uncorrelated Data. Due to the enormous size of big data repositories, dealing with large amount of data that are *uncorrelated* to the kind of analytics to be designed occurs very frequently. As a consequence, filtering-out uncorrelated data plays a critical role in the context of analytics over big data, as this heavily affects the *quality* of final analytics to be designed.

(*c*) Strongly Unstructured Nature of Data Sources. In order to design meaningful analytics, it is mandatory that input big data sources are transformed in a suitable, structured format, and finally stored in the HDFS. This poses several issues that recall classical ETL processes of Data Warehousing systems, but with the additional challenges that data alimenting big data repositories are *strongly* unstructured (e.g., social network data, biological experiment result data, and so forth) in contrast with less problematic unstructured data that are popular in traditional BI tools (e.g., XML data, RDF data, and so forth). Again, here transformations from unstructured to structured format should be performed on the basis of the analytics to be designed, according to a sort of *goal-oriented methodology*.

(*d*) High Scalability. High scalability of big data analytics is one of the primer feature to be ensured for a MAD-inspired big data analytics system. To this end, exploiting the cloud computing computational framework seems to be the most promising way to this end [2]. The usage of the IaaS-inspired cloud computing computational framework is meant with the aim of achieving some important characteristics of highly-scalable big data analytics systems, among which we recall: (*i*) *"true" scalability*, i.e. the effective scalability that a powerful computational infrastructure like clouds is capable of ensuring; (*ii*) *elasticity*, i.e. the property of rapidly adapting to massive updates and fast evolutions of big data repositories; (*iii*) *fault-tolerance*, i.e. the capability of being robust to faults that can affect the underlying distributed data/computational architecture; (*iv*) *self-manageability*, i.e. the property of *automatically* adapting the framework configuration (e.g., actual load balancing) to rapid changes of the surrounding data/computational environment; (*v*) *execution on commodity machines*, i.e. the capability of scale-out on thousands and thousands of commodity machines when data/computational peaks occur.

(*e*) Combining the Benefits of RDBMS and NoSQL Database Systems. One of the more relevant features to be achieved by big data analytics systems is represented by *flexibility*, which refers to the property of covering a *large collection* of analytics scenarios over the *same* big data partition. In order to obtain this critical feature, it is necessary to combine the benefits of traditional RDBMS database systems and those of next-generation *NoSQL database systems*, which propose representing and managing data via *horizontal data partitions* by renouncing to fixed table schemas and, consequentially, resource-expensive join operations [4].

(*f*) Query Optimization Issues in HiveQL. Several open issues arise with respect to *query optimization aspects* of HiveQL. Among

these, noticeable ones are the following: (*i*) moving towards more expressive, complex aggregations, e.g. OLAP-like rather than SQL-like, hence enforcing the *User Defined Function* (UDF) and the *User Defined Aggregate Function* (UDAF) [5] paradigms; (*ii*) covering advanced SQL statements such as *nested queries* and *order-by predicates*; (*iii*) incorporating *data compression paradigms* in order to achieve higher performance; (*iv*) devising novel *cost-based optimizations*, e.g. based on table or column statistics; (*v*) integration with third-part BI tools.

4. TURNING INTO ANALYTICS ON BIG MULTIDIMENSIONAL DATA

It has been already studied that multidimensional data naturally marry with analytics [6]. Indeed, analytics significantly extend typical OLAP operators (e.g., roll-up, drill-down, and so forth) hence it is natural to think of multidimensional data as an *add-on value* for analytics models and methodologies. Multidimensional data finally allow us to enhance the *expressive power* and the *capabilities* of analytics, and actual research experiences in the context of big data analytics are mature enough to launch a novel paradigm for Data Warehousing and OLAP research: *analytics over big multidimensional data*. Basically, this innovative paradigm aims at integrating the classical, well-known benefits of multidimensional data models (such as multidimensional abstractions, hierarchy-based dimensional tables, multi-resolution fact tables, multi-way aggregations, OLAP tools, and so forth) with analytics, in order to achieve more powerful analytics capable of enhancing actual models [13,11] by means of typical amenities deriving from such multidimensional data models. In order to realize this sort of *second-generation big data revolution*, it is necessary to face-off a number of open research problems, some of which can be summarized by the following questions.

(1) ***How To Build Multidimensional Structures On Top Of The HDFS?*** This problem refers to the issue of building multidimensional data structures on top of the structured HDFS repositories of Haadop, as a first step towards directly integrating multidimensional data models with analytics over big data. A promising direction to this end consists in exploiting *array-based in-memory representation methods*.

(2) ***How To Directly Integrate Multidimensional Data Sources Into The Hadoop Lifecycle?*** Hadoop populates the underlying structured big data repositories from heterogeneous and different-in-nature data sources, such as legacy systems, Web, scientific data sets, sensor and stream databases, social networks, and so forth. Despite this, no research efforts have been devoted to the yet-relevant issue of *directly integrating multidimensional data sources* into the Hadoop lifecycle, which is an exciting research challenge for next-generation Data Warehousing and OLAP research.

(3) ***How To Model and Design Multidimensional Extensions of HiveQL?*** In order to achieve an effective integration of multidimensional data models with analytics over big data, the query language HiveQL must be enriched with multidimensional extensions as well. These extensions should take into consideration language syntax aspects as well as query optimization and evaluation aspects, perhaps by inheriting lessons learned in the context of actual *MDX-like languages* for multidimensional data.

(4) ***How to Design Complex Analytics over Hadoop-Integrated Multidimensional Data?*** Multidimensional data provide add-on value to big data analytics. In this respect, design complex analytics over Hadoop-integrated multidimensional data plays a critical role. Actual analytics, although quite well-developed, still do not go beyond classical BI components, like diagrams, plots, dashboards, and so forth, but complex BI processes of very large organizations demand for *more advanced BI-oriented decision support tools*, perhaps by integrating principles and results of different-in-nature disciplines like statistics.

(5) ***How To Deal with Visualization Issues Arising From Big Multidimensional Data Analytics?*** Visualization issues represent a leading problem in Data Warehousing and OLAP research. These issues get worse when re-visited in the context of big multidimensional data analytics, as here visualization must kept a stronger *decision-support value*. More complex techniques, such as *multidimensional space exploration approaches*, must be investigated to this end.

5. CONCLUSIONS

Starting from state-of-the-art research issues and achievements in analytics over big data, in this paper we have provided critical discussion over open research issues and achievements arising in this scientific field, and we have extended the discussion to the emerging context of analytics over big multidimensional data. Open problems and actual research trends have been highlighted, and novel research directions have been proposed.

6. REFERENCES

[1] 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.

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

[3] *Apache Hadoop*. http://wiki.apache.org/hadoop

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

[5] Chen, Q., Hsu, M., and Liu, R. Extend UDF Technology for Integrated Analytics. *Proc. of DaWaK*, 2009.

[6] Chen, Z., and Ordonez, C. Efficient OLAP with UDFs. *Proc. of DOLAP*, 2008.

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

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

[9] Dittrich, J., Quiané-Ruiz, J.-A., Jindal, A., Kargin, Y., Setty, V., and Schad, J. Hadoop++: Making a Yellow Elephant Run Like a Cheetah. *PVLDB 3(1)*, 2010.

[10] Hacigumus, H., Iyer, B., and Mehrotra, S. Providing Database as a Service. *Proc. of ICDE*, 2002.

[11] 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.

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

[13] 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.