Supporting OLAP-Based Big Data Analytics over Data-Intensive Business Processes: Issues, Models, Proposals, and a Real-Life Framework

Alfredo Cuzzocrea
University of Trieste and ICAR-CNR
Trieste, 34127
alfredo.cuzzocrea@dia.units.it

Abstract

This paper focuses the attention on the problem of supporting *big data analytics* over so-called *data-intensive business processes*, i.e. business processes connected to big data sources. This applicative setting is now more and more of great interest in the community, also due to emerging computational paradigms like *Cloud Computing*. The paper explores issues, models and proposals in the field, and finally provides the architecture of a real-life framework that supports big data analytics over data-intensive business processes via fortunate OLAP metaphors.

1 Introduction

Nowadays, the problem of supporting *big data analytics* (e.g., [CSD11, Cuz13, CS14, Rus11, RR14]) over so-*called data-intensive business processes* (e.g., [ALRM17, SMM17, GK18]) plays a relevant role. This because, on one hand, business processes still keep the most of the data, information and knowledge of very-large enterprises and organizations, and, on the other hand, perfectly marry with the emerging characteristics of *big data* (e.g., [CSU13, CBS13, LJYC15, ZE11, MCB+11]).

An important solution for supporting big data analytics concerns with applying fortunate *multidimensional metaphors and abstractions*, mainly falling in the well-known OLAP context, thus originating an evolving trend that can be safely recognized within the term "*OLAP-based big data analytics*" (e.g., [Cuz17, CMF+16]).

Inspired by this research context, in this paper we focus the attention on the problem of supporting OLAP-based big data analytics over data-intensive business processes, and we describe a real-life framework inspired developed in the context of a real-life project, called REMS.PA, which has produced the corresponding framework, mainly designed on top of open-source technologies, and that, particularly, focuses on business processes of the Public Administration.

The remaining part of this paper is organized as follows. In Section 2, we report on main research issues of supporting OLAP-based big data analytics over data-intensive business processes. In Section 3, we describe the proposed framework. Finally, in Section 4, we provide conclusions and future work for our research.

2 OLAP-Based Big Data Analytics over Data-Intensive Business Processes: Emerging Research Issues

OLAP-based big data analytics over data-intensive business processes opens the door to several emerging research issues, among which some noticeable ones are the following:

- computing multidimensional OLAP aggregations over data-intensive business processes;
- supporting OLAP querying, operators and operations over so-computed OLAP *cubes*;
- effective and efficient in-memory representation of business process cubes;
- supporting flexible big data prediction methodologies over so-computed OLAP cubes.

How to aggregate a collection of data-intensive business processes? This is a relevant question that has attracted the attention of several studies. Basically, classical OLAP aggregation algorithms cannot be applied as they are, but suitable adaptations must be devised. A possibility consists in considering the graph-like nature of business processes in this respect. Doing this, the *scalability* property, which is relevant for big data management and processing (e.g., [WXGM18, SYGZ18, YLHC14, CMX13]), must be taken into account.

After computing aggregations, the support for OLAP querying, operators and operations must be ensured. Among queries, *range queries* are very significant in this context. In addition, supporting *roll-up* and *drill-down* operators is, for instance, a first-class problem in this respect. At the same, *slice* and *dice* operations are significant in order to provide a comprehensive support to ad-hoc big data analytics procedures.

Effectively and efficiently supporting in-memory representation of business process cubes conveys on several challenges to be faced-off. Indeed, so-computed OLAP cubes can achieve very large sizes when stored in suitable Cloud storage systems. Therefore, specialized approaches must be devised in order to tame such enormous sizes. Partition-based approaches seem a promise trend to this end.

Finally, another critical problem is represented by the issue of supporting flexible big data prediction methodologies over target OLAP cubes, as the final goal is that of discovering useful knowledge from data-intensive business processes (e.g., [BCC$^+$14, WQL$^+$18, She18]). Again, multidimensional paradigms, such as *multidimensional clustering* (e.g., [Mur85]), can be successfully applied to this end.

3 An Innovative Framework for Supporting OLAP-Based Big Data Analytics over Data-Intensive Business Processes

The proposed framework aims at supporting OLAP-based big data analytics over data-intensive business processes. It combines two main assets: analysis and prediction of business processes, with focus on the case of business processes in the Public Administration, and intends to reach the definition of the framework for the automated management and optimization of business processes in the Public Administration. From a strictly technological point of view, the fundamental components of the framework are the following:

- tools to support multidimensional analysis of business process schemes using the OLAP paradigm;
- visual analytics tools for business processes based on multidimensional abstractions;
- tools to support the prediction of executions of business processes based on a data-driven approach.

The framework has been realized by using and integrating open-source software technologies for the support of business process management with the aim of speeding up and simplifying the management of the operational workflows of the Public Administration, via defining and building the management processes in a rigorous and reliable way, and finally monitor the real status of their execution. More generally, the proposed framework aims at optimizing and automating the management of Public Administration processes through their analysis and prediction of their executions. Business process analysis and prediction are therefore the two central themes of the business process management framework, which aims, by recognizing in these two phases, critical elements for the improvement of the management of these Public Administration processes as well as the provision of services to the citizen. Therefore, the resulting optimizations tend towards the general objective of achieving efficiency and flexibility of the Public Administration processes. To this end, the proposed framework includes two innovative components to support the analysis and prediction phases: (*i*) visual analytics on business processes, which focuses on the analysis of business processes (and their execution traces) using multidimensional abstractions for the support of OLAP analysis on business process schemes; (*ii*) execution prediction on business processes, which focuses on the prediction of business process executions, to support their optimization, through an innovative data-driven approach. In short, this approach aims to predict execution of Public Administration business processes by resorting to the analysis of the variations that business-processes previous performances have produced on the data (focusing the attention, therefore, on the nature of the data distributions that characterize these variations). A software tool has been implemented, as to allow the Public Administration to optimize the management of internal processes, evaluate their effectiveness, and adopt the necessary corrections in order to make the service offered to the community efficient and transparent.

Indeed, the level of citizen satisfaction is a yardstick for the Public Administration with respect to public management. In this sense, the framework aims to ensure significant changes, including:

- improvement of administrative transparency (e.g., telematics desk for the citizen, and so forth);
- certainty of compliance with procedures and regulations and the traceability of activities;
- control and optimization of processes;
- reduction in the time required for administrative procedures;
- increase in "company productivity";
- global reduction of associated costs;
- automation of the planned activities;
- accountability and monitoring of the people involved.

The innovative features introduced by the proposed framework are the following.

Feature 1 – Innovative techniques and tools for OLAP analysis on business process schemes: Although OLAP is a methodology applied to many data models (such as graphs, sequences, text, etc.), in literature, as well as in industry, there are no proposals that offer an "explicit" OLAP support on business processes (for example: multidimensional browsing and exploration of aggregated business process schemes, coverage of the most common OLAP operators and operations - such as roll-up, drill-down, pivoting, etc., and so forth), in spite of the embryonic tools for multidimensional analysis made available by some tools (e.g., *ProM* [vDdMV+05]).

Feature 2 – Visual analytics tools and techniques on BP that exploit multidimensional abstractions: Even in this case, the visual analytics solution proposed by the framework directly exploit the power of multidimensional abstractions, for example thanks to multi-resolution analysis, which it is both powerful and very intuitive. It should be noted that, both in literature and in the field of industrial solutions, there are no approaches that propose this vision of visual analytics on business processes.

Feature 3 – Data-driven process mining: From a purely scientific and industrial point of view, the most valuable result that the framework introduces is represented by the innovative data-driven process mining methodology. This methodology is not only innovative in research (academic and industrial), but, despite its complexity, it effectively captures real-world application scenarios of business process management systems (which, in turn, are characterized by a certain intrinsic complexity) in a very powerful and flexible manner, thus imposing a sound methodology (based on multidimensional abstractions) as opposed to other approaches known in the state-of-the-art literature that solve the difficult problem of monitoring and optimizing business processes through solution-driven approaches (which introduce little flexibility and extensibility not only for application scenarios other than those for which they have been developed, but also for application scenarios characterized by execution settings that are not very different from the latter).

Summarizing, the main scientific and technical research issues addressed by the framework are the following:

- definition of methodologies, models and tools for supporting multidimensional analysis of business process schemes;
- effective and efficient representation of aggregated business process schemes in secondary storage;
- definition of paradigms for the support of OLAP functionalities and extensions on aggregated business process schemes;
- definition of methodologies, models and tools for supporting the multi-resolution OLAP analysis of business process schemes;
- optimization techniques for OLAP roll-up and drill-down operators on aggregated business process schemes;
- definition of appropriate multidimensional metaphors for the support of visual analytics for business process using OLAP methodologies and paradigms;
- efficient and scalable solutions for the support of visual analytics for business processes;
- definition of the predictive analysis method of data-driven process mining;
- cumulative similarity techniques between discrete data distributions;
- techniques for optimizing procedures for processing and analyzing discrete distributions on big business process data.

4 Logical Architecture of the Proposed Framework

Figure 1 shows the logical architecture of the proposed framework for supporting OLAP-based big data analytics over data-intensive business processes.

As shown in Figure 1, the proposed framework introduces the following layers:

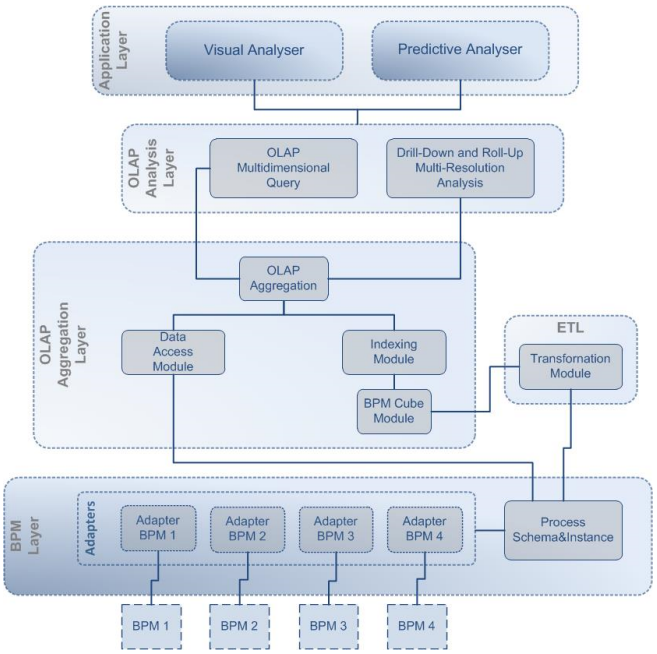

Figure 1: Logical architecture

- *BPM Layer*: is it the layer where the input business processes are located and exploited to populate the big data layer of the framework;
- *OLAP Aggregation Layer*: it is the layer where business processes are aggregated into cubes in order to supporting OLAP-based big data analytics;
- *OLAP Analysis Layer*: it is the layer where the OLAP querying, operators and operations over business processes are implemented;
- *Application Layer*: it is the layer where the consumer applications are located, being visual analytics and prediction analytics the main functionalities supported.

5 Conclusions and Future Work

This paper has focused the attention on the problem of supporting big data analytics over so-called data-intensive business processes, i.e. business processes connected to big data sources. We explored issues, models and proposals in the field, and finally the architecture of a real-life framework developed in the context of a real-life project has been provided.

Future work is mainly oriented to enrich the proposed framework via innovative big data properties, such as: *privacy preservation* (e.g., [CB11, CR09]), *open big data predicates* (e.g., [Kar17]), and *consistency checking* (e.g., [KWR+15]).

Acknowledgments

This research has been developed in the context of the *MISE Horizon 2020 – PON 2014/2020* project: "*REMS.PA (Resource in Engineering Management for Software process automation in Public Administration)*".

References

[ALRM17] Saima Gulzar Ahmad, Chee Sun Liew, M. Mustafa Rafique, and Ehsan Ullah Munir. Optimization of data-intensive workflows in stream-based data processing models. *The Journal of Supercomputing*, 73(9):3901–3923, 2017.

[BCC+14] Peter Braun, Juan J. Cameron, Alfredo Cuzzocrea, Fan Jiang, and Carson Kai-Sang Leung. Effectively and efficiently mining frequent patterns from dense graph streams on disk. In *18th International Conference in Knowledge Based and Intelligent Information and Engineering Systems, KES 2014, Gdynia, Poland, 15-17 September 2014*, pages 338–347, 2014.

[CB11] Alfredo Cuzzocrea and Elisa Bertino. Privacy preserving OLAP over distributed XML data: A theoretically-sound secure-multiparty-computation approach. *J. Comput. Syst. Sci.*, 77(6):965–987, 2011.

[CBS13] Alfredo Cuzzocrea, Ladjel Bellatreche, and Il-Yeol Song. Data warehousing and OLAP over big data: current challenges and future research directions. In *Proceedings of the sixteenth international workshop on Data warehousing and OLAP, DOLAP 2013, San Francisco, CA, USA, October 28, 2013*, pages 67–70, 2013.

[CMF+16] Alfredo Cuzzocrea, Carmen De Maio, Giuseppe Fenza, Vincenzo Loia, and Mimmo Parente. OLAP analysis of multidimensional tweet streams for supporting advanced analytics. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016*, pages 992–999, 2016.

[CMX13] Alfredo Cuzzocrea, Rim Moussa, and Guandong Xu. Olap*: Effectively and efficiently supporting parallel OLAP

over big data. In *Model and Data Engineering - Third International Conference, MEDI 2013, Amantea, Italy, September 25-27, 2013. Proceedings*, pages 38–49, 2013.

[CR09] Alfredo Cuzzocrea and Vincenzo Russo. Privacy preserving OLAP and OLAP security. In *Encyclopedia of Data Warehousing and Mining, Second Edition*, pages 1575–1581. 2009.

[CS14] Alfredo Cuzzocrea and Il-Yeol Song. Big graph analytics: The state of the art and future research agenda. In *Proceedings of the 17th International Workshop on Data Warehousing and OLAP, DOLAP 2014, Shanghai, China, November 3-7, 2014*, pages 99–101, 2014.

[CSD11] Alfredo Cuzzocrea, Il-Yeol Song, and Karen C. Davis. Analytics over large-scale multidimensional data: the big data revolution! In *DOLAP 2011, ACM 14th International Workshop on Data Warehousing and OLAP, Glasgow, United Kingdom, October 28, 2011, Proceedings*, pages 101–104, 2011.

[CSU13] Alfredo Cuzzocrea, Domenico Saccà, and Jeffrey D. Ullman. Big data: a research agenda. In *17th International Database Engineering & Applications Symposium, IDEAS '13, Barcelona, Spain - October 09 - 11, 2013*, pages 198–203, 2013.

[Cuz13] Alfredo Cuzzocrea. Analytics over big data: Exploring the convergence of datawarehousing, OLAP and data-intensive cloud infrastructures. In *37th Annual IEEE Computer Software and Applications Conference, COMPSAC 2013, Kyoto, Japan, July 22-26, 2013*, pages 481–483, 2013.

[Cuz17] Alfredo Cuzzocrea. Scalable olap-based big data analytics over cloud infrastructures: Models, issues, algorithms. In *Proceedings of the 2017 International Conference on Cloud and Big Data Computing, ICCBDC 2017, London, United Kingdom, September 17 - 19, 2017*, pages 17–21, 2017.

[GK18] Janis Grabis and Janis Kampars. Application of microservices for digital transformation of data-intensive business processes. In *Proceedings of the 20th International Conference on Enterprise Information Systems, ICEIS 2018, Funchal, Madeira, Portugal, March 21-24, 2018, Volume 2.*, pages 736–742, 2018.

[Kar17] Holden Karau. Unifying the open big data world: The possibilities$_*$ of apache BEAM. In *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*, page 3981, 2017.

[KWR+15] Thanh Tran Thi Kim, Erhard Weiss, Christoph Ruhsam, Christoph Czepa, Huy Tran, and Uwe Zdun. Embracing process compliance and flexibility through behavioral consistency checking in ACM - A repair service management case. In *Business Process Management Workshops - BPM 2015, 13th International Workshops, Innsbruck, Austria, August 31 - September 3, 2015, Revised Papers*, pages 43–54, 2015.

[LJYC15] Kuan-Ching Li, Hai Jiang, Laurence T. Yang, and Alfredo Cuzzocrea, editors. *Big Data - Algorithms, Analytics, and Applications*. Chapman and Hall/CRC, 2015.

[MCB+11] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. Technical report, McKinsey Global Institute, 2011.

[Mur85] Fionn Murtagh. Multidimensional clustering algorithms. Physica-Verlag, 1985.

[RR14] Wullianallur Raghupathi and Viju Raghupathi. Big data analytics in healthcare: promise and potential. *Health Inf. Sci. Syst.*, 2(1):3, 2014.

[Rus11] Philip Russom. Big data analytics. Technical report, TDWI Research, Renton, WA, USA, 2011.

[She18] Bin Shen. Universal knowledge discovery from big data using combined dual-cycle. *Int. J. Machine Learning & Cybernetics*, 9(1):133–144, 2018.

[SMM17] Vladislav A. Shchapov, Aleksei G. Masich, and Grigorii F. Masich. The technology of processing intensive structured dataflow on a supercomputer. *Journal of Systems and Software*, 127:258–265, 2017.

[SYGZ18] Dawei Sun, Hongbin Yan, Shang Gao, and Zhangbing Zhou. Performance evaluation and analysis of multiple scenarios of big data stream computing on storm platform. *TIIS*, 12(7):2977–2997, 2018.

[vDdMV+05] Boudewijn F. van Dongen, Ana Karla A. de Medeiros, H. M. W. Verbeek, A. J. M. M. Weijters, and Wil M. P. van der Aalst. The prom framework: A new era in process mining tool support. In *Applications and Theory of Petri Nets 2005, 26th International Conference, ICATPN 2005, Miami, USA, June 20-25, 2005, Proceedings*, pages 444–454, 2005.

[WQL+18] Xinyang Wang, Deyu Qi, Weiwei Lin, Mincong Yu, Zhishuo Zheng, Naqin Zhou, and Pengguang Chen. A general framework for big data knowledge discovery and integration. *Concurrency and Computation: Practice and Experience*, 30(13), 2018.

[WXGM18] Yulei Wu, Yang Xiang, Jingguo Ge, and Peter Mueller. High-performance computing for big data processing. *Future Generation Comp. Syst.*, 88:693–695, 2018.

[YLHC14] Chao-Tung Yang, Jung-Chun Liu, Ching-Hsien Hsu, and Wei-Li Chou. On improvement of cloud virtual machine availability with virtualization fault tolerance mechanism. *The Journal of Supercomputing*, 69(3):1103–1122, 2014.

[ZE11] Paul Zikopoulos and Chris Eaton. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 1st edition, 2011.