 DOI 10.1109/SERVICES.2019.00120

Towards Specification of a Software Architecture for Cross-Sectoral Big Data Applications

I. Arapakis*, Y. Becerra†, O. Boehm‡, G. Bravos§, V. Chatzigiannakis§, C. Cugnasco†,G. Demetriou¶, I. Eleftheriou‖, J. E. Mascolo**, L. Fodor**, S. Ioannidis‡‡, D. Jakovetic††, L. Kallipolitis[x], E. Kavakli‖[xi], D. Kopanaki‡‡, N. Kourtellis*, M. M. Marcos[xii],R. M. de Pozuelo[xii], N. Milosevic**, G. Morandi¶, E. P. Montanera[xiii], G.H. Ristow[xiv], R. Sakellariou‖, R. Sirvent†, S. Skrbic**, I. Spais[x], G. Vasiliadis‡‡, M. Vinov‡

*Telefonica Research, Spain†Barcelona Supercomputing Center, Spain‡IBM, Israel§Information Technology for Market Leadership, Greece ¶Ecole des Ponts ParisTech, France‖University of Manchester, UK**Centro Ricerche FIAT, Italy††University of Novi Sad, Serbia‡‡ Foundation for Research and Technology Hellas, Greece[x]Aegis IT Research LTD, UK[xi]University of the Aegean, Greece,[xii]CaixaBank, Spain [xiii]ATOS, Spain[xiv]Software AG, Germany, corresponding author: dusan.jakovetic@dmi.uns.ac.rs

***Abstract*—The proliferation of Big Data applications puts pressure on improving and optimizing the handling of diverse datasets across different domains. Among several challenges, major difficulties arise in data-sensitive domains like banking, telecommunications, etc., where strict regulations make very difficult to upload and experiment with real data on external cloud resources. In addition, most Big Data research and development efforts aim to address the needs of IT experts, while Big Data analytics tools remain unavailable to non-expert users to a large extent. In this paper, we report on the work-in-progress carried out in the context of the H2020 project I-BiDaaS (Industrial-Driven Big Data as a Self-service Solution) which aims to address the above challenges. The project will design and develop a novel architecture stack that can be easily configured and adjusted to address cross-sectoral needs, helping to resolve data privacy barriers in sensitive domains, and at the same time being usable by non-experts. This paper discusses and motivates the need for Big Data as a self-service, reviews the relevant literature, and identifies gaps with respect to the challenges described above. We then present the I-BiDaaS paradigm for Big Data as a self-service, position it in the context of existing references, and report on initial work towards the conceptual specification of the I-BiDaaS software architecture.**

***Keywords*-big data value chain; big data as a self-service solution; software architecture**

I. Introduction

The emerging concept of Big-Data-as-a-Self-Service [1] refers to empowering non-expert Big Data users to easily utilise and interact with Big Data technologies. Big Data as a self service has a great potential to broaden the scope of applicability of Big Data analytics. However, there are several barriers in adopting Big Data as a self-service. First, current efforts are mainly focused on the needs of IT experts, administrators and application developers, and there is a need to develop more platforms that are easy to use by non-experts. Second, sensitivity of data and strict regulations in several application domains such as banking, make impossible to upload real data to cloud-based Big Data platforms for experimentation and testing. To overcome these challenges, as detailed below and in [13], we propose a paradigm for Big Data analytics as a self-service wherein a user: 1) can fabricate realistic synthetic data; 2) can experiment and test with synthetic data in external cloud resources; and 3) once satisfied with the solution, can export a satisfactory configuration to internal, local premises.

II. Literature review

There has been significant scientific and technological effort in the area of Big Data analytics over the past decade, with new emerging concepts like Big Data as a service and Analytics as a service [9], as well as applications in various domains such as business intelligence [8]. Specifically, significant effort has focused on defining software architectures for Big Data analytics. Big Data software architectures have to deal with several challenges. First, they need to handle huge amounts of data that cannot be processed with standard infrastructures [3]. Software solutions like Apache Spark and Hadoop come with obvious advantages in designing Big Data systems [4]. A second challenge lies in accommodating real time, streaming data [4]. As data has to be stored in a continuous manner, database engines and file systems have to be redesigned to be able to handle such scenarios. Components like asynchronous data processing and quick, lightweight web services are immediately necessary for adequate system performance. Third, Big Data architectures need to handle data that comes from heterogeneous, autonomous sources with distributed and anonymous platforms [5]. To accommodate this challenge, service composition is often a requirement. In addition, standards for storing and accessing data are being defined [5]. To address

the above challenges, several Big Data software architectures have been described, e.g., [6], [2], and [7].

III. Positioning and contributions

Despite significant work on defining Big Data software architectures, there are still several gaps that need to be bridged. As pointed in [2], Big Data solutions are usually developed bottom-up, and it is often technologies and not user requirements that drive application development. Reference architectures [10], [11] attempt to provide generic views and solutions, i.e., generalize and harmonize requirements, concerns, etc. However, they are not concerned with technical details of a specific end-to-end architecture. And although, cross-sectoral Big Data requirements have been well understood [12], their mapping onto concrete architectures can be further developed.

Furthermore, existing solutions are not primarily concerned with the barrier of exposing proprietary real data to external cloud resources. These platforms can either be used at local premises or work with non-sensitive data. However, in many cases, local resources are occupied with current operational processes, and the resources for experimentation and testing need to be external. This is the scenario which the proposed architecture aims to resolve. In particular, the architecture aspires to make Big Data technologies more agile in scenarios where it is very difficult to upload real data to cloud resources. Finally, an open challenge is making Big Data technologies accessible to non-expert users and this is another requirement targeted by the proposed solution. To address the above challenges, we introduce a paradigm for Big Data analytics as a self-service based on data fabrication and sequential programming models [13]. This paradigm involves a generic workflow/pipeline of Big Data analytics wherein a user: 1) fabricates realistic synthetic data; 2) experiments and tests with synthetic data in external cloud resources; and 3) once satisfied with the solution, the user exports the satisfactory configuration to internal, local premises. At the same time, the described process is envisioned to be easy to realize, even in-house in companies by non experts, thus materializing the "Big Data as a self-service" concept. The proposed system is aimed to support Big Data as a self-service, as well as to provide a safe experimentation environment for the methodical development of new Big Data products, services, and tools. Specifically, we have carried out a systematic requirements analysis [13] which allowed us to specify the functional requirements (FRs) and non-functional concerns (NFC) which the solution should satisfy; we refer to [13] for details on how the proposed paradigm, FRs, and NFCs, as well as the FRs and NFCs can be mapped into a concrete software architecture.

Acknowledgment. This work is supported by the I-BiDaaS project, funded by the European Commission under Grant Agreement No. 780787. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

References

[1] C. A. Ardagna, P. Ceravolo, and E. Damiani, "Big data analytics as-a-service: Issues and challenges," in 2016 IEEE Int. Conf. on Big Data (Big Data).

[2] C. A. Ardagna, V. Bellandi, M. Bezzi, P. Ceravolo, E. Damiani, and C. Hebert, "Model-based big data analytics-as-aservice: Take big data to the next level," IEEE Trans. Services Computing, pp. 11, 2018.

[3] P. Basanta-Val, "An efficient industrial big-data engine," IEEE Tran. Ind. Inform. 14.4 (2018): 1361-1369.

[4] D. Xu, D. Wu, X. Xu, L. Zhu, and Len Bass, "Making real time data analytics available as a service," in Quality of Software Architectures (QoSA), 11th Int. ACM SIGSOFT Conf., pp. 73-82. IEEE, 2015.

[5] T.H.A.S. Siriweera, I. Paik, B. T.G.S. Kumara, K.R.C. Koswatta, "Intelligent big data analysis architecture based on automatic service composition," in Big Data (BigData Congress), 2015 IEEE Int. Congress on, pp. 276-280. IEEE, 2015.

[6] A. A. Munshi, Y. Abdel-Rady I. Mohamed, "Data Lake Lambda Architecture for Smart Grids Big Data Analytics," IEEE Access 6 (2018): 40463-40471.

[7] E. R. Sparks, S. Venkataraman, T. Kaftan, M. J. Franklin, B. Recht, "KeystoneML: Optimizing Pipelines for Large-Scale Advanced Analytics," 2017 IEEE 33rd Int. Conf. Data Engineering, Apr. 2017, pp. 535546.

[8] C. Pickering, M. Gupta, "Self Service Business Intelligence (SSBI) for Employee Communications and Collaboration (ECC)," in 2015 Int. Conf. on Collaboration Technologies and Systems (CTS), Atlanta, GA, 2015, pp. 302-304.

[9] Z. Zheng, J. Zhu, and M. R. Lyu, "Service-generated Big Data and Big Data-as-a-Service: An Overview," 2013 IEEE Int. Congress on Big Data, pp. 403-410, 2013.

[10] NIST Big Data Interoperability Framework: Volume 6, Reference Architecture. National Institute of Standards and Technology, Special Publication 1500-6r1, 2018.

[11] European Big Data Value Strategic Research and Innovation Agenda, version 4.0, Oct 2017.

[12] T. Becker, E. Curry, A. Jentzsch, W. Palmetshofer, "Cross-sectorial Requirements Analysis for Big Data Research," New Horizons for a Data-Driven Economy, 2016, pp 263-276.

[13] Positioning of I-BiDaaS, the I-BiDaaS project consortium, 2018, http://ibidaas.eu/deliverables