Big Data Management Challenges in SUPERSEDE

Sergi Nadal, Alberto Abelló, Oscar Romero, Jovan Varga
Universitat Politècnica de Catalunya, BarcelonaTech
Barcelona, Spain
{snadal,aabello,oromero,jvarga}@essi.upc.edu

1. INTRODUCTION

The H2020 SUPERSEDE (www.supersede.eu) project aims to support decision-making in the evolution and adaptation of software services and applications by exploiting end-user feedback and runtime data, with the overall goal of improving the end-users quality of experience (QoE). Such QoE is defined as the overall performance of a system from the point of view of users, which must consider both feedback and runtime data gathered. End-user's feedback is extracted from online forums, app stores, social networks and novel direct feedback channels, which connect software applications and service users to developers. Runtime data is primarily gathered by monitoring environmental sensors, infrastructures and usage logs. Hereafter, we discuss our solutions for the main data management challenges in SUPERSEDE.

2. CHALLENGES

2.1 Big Data Governance

One well-known problem of NOSQL repositories is the lack of semantics caused by their schemaless properties. This lack of schema prevents the system from knowing which data is stored and how they interrelate. Thus, data analysts are hindered with data management tasks, like understanding the specific structure and parsing it, before writing their analytical pipelines. In SUPERSEDE, this gets more challenging as it aims at performing integrated analysis over multiple, evolving and heterogeneous data sources. A challenge that current Big Data technologies fail to address.

Big Data ecosystems demand complex metadata governance processes spanning throughout all data management phases, from ingestion to analysis [2]. Semantic Web technologies have proven to be a valid asset for such purpose. The Resource Description Framework (RDF) allows to flexibly define concepts and their relationships in the form of a semantic graph. Furthermore, it can leverage on the Linked Data initiative to (a) reuse existing vocabularies, (b) make data self-descriptive, and (c) publish such data to faciliate on-the-fly data crossing [1]. In SUPERSEDE, an integration-oriented RDF graph is used to represent and integrate the data related to monitoring and user feedback, as well as crossing it with contextual data from the use cases. Also, the analytical processes to support decision making are represented on top of such concepts.

2.2 Big Data Architectures

The λ-architecture [3] is currently the most widespread reference architecture for scalable and fault-tolerant Big Data processing. While succeeding at managing humongous amounts of data (i.e., in the Batch layer), as well as near-real time data streams (i.e., in the Speed layer), it has two main drawbacks. First, it completely overlooks semantics, as discussed before, as it uses NOSQL technologies as its baseline components. Second, its vaguely defined, which hinders its instantiation.

(i) Refining the λ-architecture, by defining its components as well as their interconnections, would facilitate its instantiation and allow a simpler deployment of SUPERSEDE's Big Data ecosystem. (ii) To accommodate the requirements on governance, metadata should be considered as first-class citizen throughout the data management processes.

3. PARTICIPATION BENEFITS

Our objective is twofold. Firstly, we aim at presenting our approach to tackle the previously described challenges. Secondly, by leading a round table, we aim at discussing pros and cons of this and other solutions pursued by other reserachers in similar settings.

4. ACKNOWLEDGEMENTS

This work has been partly supported by the SUPERSEDE project, funded by the European Union's Information and Communication Technologies Programme (H2020) under grant agreement number 644018.

5. REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[2] E. Kandogan, M. Roth, P. M. Schwarz, J. Hui, I. Terrizzano, C. Christodoulakis, and R. J. Miller. LabBook: Metadata-driven Social Collaborative Data Analysis. In *IEEE Big Data*, 2015.

[3] N. Marz and J. Warren. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*. Manning, 1st edition, 2015.