Toward Model-based Big Data-as-a-Service: The TOREADOR Approach

Ernesto Damiani[1,2], Claudio Ardagna[1,3], Paolo Ceravolo[1,3], and
Nello Scarabottolo[1,3]

[1] Consorzio Interuniversitario Nazionale per l'Informatica, Italy
[2] EBTIC/Khalifa University of Science and Technology, UAE
[3] Università degli Studi di Milano, Italy

Abstract. The full potential of Big Data Analytics (BDA) can be unleashed only by overcoming hurdles like the high architectural complexity and lack of transparency of Big Data toolkits, as well as the high cost and lack of legal clearance of data collection, access and processing procedures. We first discuss the notion of *Big Data Analytics-as-a-Service* (BDAaaS) to help potential users of BDA in overcoming such hurdles. We then present TOREADOR, a first approach to BDAaaS.

1 Introduction

Big Data technology has recently become a major market estimated to reach \$203 billion in 2020, growing at a CAGR of 11.7% [11]. According to [9], every human in the world is producing over 6 megabytes for minute, a total of 1.7 million billion bytes of data. Also, the Compliance, Governance and Oversight Council claimed that the information volume doubles every 18-24 months for most organizations [5]. Many organizations have discovered that, in order to remain competitive, they have to deal with business cases where the volume of data reaches terabytes and even petabytes, and whose requirements include low latency and handling a variety of datatypes [2]. Still, Big Data applications are complex systems whose design and deployment poses challenges at multiple levels, ranging from data representation and storage issues to choice and adaptation of analytics, parallelisation and deployment strategies as well as display and interpretation of results. ICT companies propose to their customers to tackle Big Data application development using a mix of technologies going from NoSQL ("notonlySQL") databases like Cassandra or HBase, data preparation utilities like Paxata, and distributed, parallel computing systems like Apache Hadoop, Stark or Flink. However, the high architectural complexity and lack of transparency of Big Data toolkits leads many customers to use them as blackboxes, with little or no insight on how analytics are actually executed. Another major factor hindering Big Data Analytics (BDA) adoption is the "regulatory barrier:" concerns about violating data access, sharing and custody regulations when using BDA, and the high cost of obtaining legal clearance for their specific scenario are discouraging for many organizations. Finally, the limited size of the BDA talent pool makes Big Data scientists, architects, and developers costly and in high demand internationally. Even outsourcing BDA to a service provider and/or engaging consultants does not eliminate the need of costly in-house skills. Having to tackle these challenges from scratch creates an entry barrier for small organisations, particularly SMEs, that cannot access the data science and data technology competence pools. This paper presents TOREADOR, a model-based approach for fast specification and roll-out of Big Data applications, fostering reuse via a Software Product Line approach. The TOREADOR methodology and toolkit support collecting user requirements and preferences in a declarative format, converting them into a procedural model of the Big Data computation, and compiling the latter into low-level specification directly deployable on a number of execution platforms and technologies. In this paper, we first give an overview of Big Data concepts (Section 2); we then define and compare BDA and BDAaaS providing some relevant application scenarios (Section 3); we finally discuss the TOREADOR approach and design to BDAaaS (Section 4).

2 Big Data: Overview

In the rest of the paper, we shall use the term *Big Data* to refer to data sets and flows whose size and update frequency cannot be handled by traditional database systems [15]. Big Data have been defined using the so called 5V model: Volume, Variety, Velocity, Value, Veracity [13]. The term *Big Data Analytics* refers to the implementation of analytics over architectures that automatically adapt to the volume, variety and velocity of data, making it possible to extract valuable results within strict deadlines. Following [4], we now describe some major hurdles that Big Data Analytics is facing today.

The technology opacity hurdle. While Big Data analytics can in principle support existing or new value propositions in a number of business domains, choosing and deploying the "right" analytics on the "right" computational infrastructure is still more an art than an engineering practice [8, 12]. Today, only large organizations with deep pockets can afford going trial-and-error for weeks on failure-prone, resource intensive Big Data Analytics projects. If SMEs and other limited budget actors, like start-ups and no-profit organization, have to join the Big Data ecosystem, provisioning a Big Data analysis process must become fast, transparent, affordable, repeatable and robust.

The data diversity hurdle. According to the current Big Data hype, the world is awash in readily accessible Big Data having common time, location, and identity references. Reality is very different. A few Over-The-Top (OTT) operators, like Google, have proprietary, semantically rich, and homogeneous data sources; they can conceivably expand their scope, adding uniform location and identity metadata. For others, data diversity is much higher [1, 16, 21], as data are independently collected and supplied by multiple actors: utility companies own sensor, management and billing information, telecommunication operators offer location and identity data, while public administrations supply open data on their territory and urban environments. With respect to relatively uniform OTT-style data, these multi-owner data sources are highly diverse: they differ in volume (involving small giga-scale and large peta-scale data sizes), granularity and veracity.

The compliance hurdle. Vertical domains where BDA can make a real difference (healthcare, transportation and energy) are highly regulated [7, 14]. Regulatory peculiarities cannot be addressed on a project-by-project basis. Rather, certified compliance of each BDA (e.g., in the form of a Privacy Impact Assessment) should be made available from the outset to all actors that use BDA in their business model. Also, BDA comes with legal issues that may trigger unwanted litigation. How to account intellectual property and how to shape the economical exploitation of BDA in multi-party environments [23]? How to provide evidence that data processing is compliant to ethics, beyond norms and directives [18]? Those are among the questions that still require mature and reliable solutions.

We believe these hurdles to have played a major role in hindering BDA acceptance [20, 22]. For example, IDC [10] reports that 60% of organizations are hampered by too little business intelligence and only 10% of employees are satisfied with the Big Data technology resources available [10]. Big Data Analytics-as-a-service can play a role in bringing Big Data to the mass, representing the entry point also for companies lacking Big Data skills and competences.

3 Big Data Analytics-as-a-service

The BDAaaS paradigm [4] represents the next evolution step of Big Data to accomplish the hurdles discussed in Section 2. It consists of a set of automatic tools and methodologies that allow customers lacking Big Data expertise to manage BDA and deploy a full Big Data pipeline addressing their goals. BDAaaS can be seen as a function that takes as input users' Big Data goals and preferences, and returns as output a ready-to-be-executed Big Data pipeline.

Users with different skills and expertise can benefit by using a BDAaaS paradigm. Users lacking expertise proper of data scientists (e.g., modeling, analysis, problem solving) can use a BDAaaS solution for preparing the real analytics, reason on data to find out hidden patterns and information, and solve business problems. Users lacking data engineering expertise (e.g., build a robust data pipeline, install a Big Data toolkit) can use a BDAaaS to automatically identify and deploy the proper set of technologies that accomplish their requirements. Users lacking both expertise can still use BDAaaS solutions for a proper initiation in the Big Data realm.

Users' requirements are in the form of platform-independent declarative goals, which are then transformed in low-level platform-dependent configurations of the Big Data pipeline. Requirements can be defined in five different conceptual areas as follows:

- **Data preparation** specifies all activities aimed to prepare data for analytics. For instance, it defines how to guarantee data owner privacy.
- **Data representation** specifies how data are represented and expresses representation choices for each analysis process. For instance, it defines the data model and data structure.
- **Data analytics** specifies the analytics to be computed. For instance, it defines the expected outcome and the type of analytics.
- **Data processing** specifies how data are routed and parallelized. For instance, it defines the processing type and the parameters driving a map-reduce processing.
- **Data visualization and reporting** specifies an abstract representation of how the results of analytics are organized for display and reporting. For instance, it defines visualization type and visual density.

BDAaaS paradigm applies to Big Data scenarios involving enterprises that, for different reason, cannot rely on the adequate level of Big Data competences and/or on skilled data scientists and engineers. In the following, we discuss the issues and challenges introduced by the BDAaaS paradigm.

4 The TOREADOR methodology

Let us consider a Big Data Analytics application from the point of view of the final user. Most of the times, some or all the following activities are needed (not necessarily in this order):

- *Define a business value proposition.* What income increase or cost reduction will the BDA results enable?
- *Identify the data.* Which inputs are needed to feed the BDA ?
- *Define the ingestion data flows (and the data lake where data will be ingested).* Where are (and who supplies) the data ingestion hooks (e.g., URLs or callbacks)? Are the ingestion flows stream or batch-like? Are the data flows to be processed immediately or will they feed a “lake” from which the BDA will periodically take its inputs?
- *Select and apply data preparation filters.* Will the flows be filtered (e.g., for anonymity/obfuscation or density(before ingestions?
- *Select and apply data protection measures.* Will after-ingestion access control be applied on the data lake, so each BDA will be able to access only the information it is entitled to?
- *Select analytics.* Given the available data and a description, classification or prediction goal, which algorithm or model should be employed to achieve the best results?
- *Define the analytics processing.* Will the BDA code be executed in a parallel way? Which HPC parallelization paradigm will be employed?
- *Define visualization, reporting and interaction.* Should the results be presented in a visual fashion, will they be stored in the data lake or will they feed other BDAs?

The questions above can be easily mapped into the five areas discussed in Section 3, which are then implemented by the TOREADOR methodology in five steps as follows:

1. A *declarative model* is used to describe the desired functional and non-functional goals of the BDA. The model is generated by a form where the questions listed above are answered by choosing among a closed list of answers, expressed in TOREADOR controlled vocabulary. In a nutshell, the declarative model is expressed as a list of lists, each list corresponding to one the five areas of the BDA. The atoms within each list are $\{$*property*, *value*$\}$ pairs where *property* expresses functional or non-functional indicators, while *value* can be either 0 or 1, for Boolean properties, or a value in an ordinal scale.
2. The TOREADOR declarative model is checked for consistency to eliminate conflicting requirements [3].
3. The declarative model is used to instantiate a platform-independent *procedural model* that describes the BDA computation. This model is a linear composition represented in OWL-S and consists of 5 pipeline stages,[4] one for each area of the BDA pipeline (i.e., preparation, representation, analytics, processing, reporting/visualization). For each stage, the TOREADOR toolkit generates an *intra-area procedural model*, specifying the composition of internal services within the stage. More in detail, the toolkit feeds each list of $\{$*property*, *value*$\}$ pairs into a SPARQL query pattern [19]. The query is then applied to a pre-defined OWL-S service ontology that lists the available services for each area (e.g., k-anonymity obfuscation service for area data preparation). The result is a list of services compatible with the preferences expressed in the declarative model.
4. The user is then called in to specify how the abstract services should be composed. These compositions are not necessarily pipelines; for instance, the analytics stage may involve disjunction, parallel execution of services and even loops [17]. To simplify this operation, the user can choose among a number of pre-defined composition patterns. Once the abstract service interface and their composition have been specified, the TOREADOR toolkit adds to the model the service interface's *grounding*, that is, the corresponding URLs in a target execution environment (e.g., Apache Flink, Spark) selected by the user.
5. Each intra-area procedural model is then compiled by the TOREADOR toolkit into a platform-dependent *deployment model*. The deployment model can be either executed via SaaS or PaaS service interface on the target deployment environment, according to the user preferences. In the former case, the composition workflow is translated in an XML incarnation (e.g., Apache Oozie), whose execution is supported by the target environment; in the latter case, an executable image (a Python bundle or a Docker image [6]) is generated ready for execution on a service provider's infrastructure.

[4] Again, the order of the stages depends on the specific BDA and is decided by the user.

5 Conclusions

In this paper, we defined the Big Data Analytics-as-a-Service (BDAaaS) paradigm as the next evolution step of Big Data domain. We then briefly outlined the TOREADOR approach to BDAaaS as a suitable driver bringing BDA to those organizations and SMEs lacking sufficient in-house competences. The TOREADOR toolkit and the project open deliverables describing the toolkit in detail are available on the site www.toreador-project.eu.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under the TOREADOR project, grant agreement No 688797. It was also partly supported by the program "piano sostegno alla ricerca 2016' funded by Università degli Studi di Milano.

References

1. Abadi, D., Agrawal, R., Ailamaki, A., Balazinska, M., Bernstein, P.A., Carey, M.J., Chaudhuri, S., Dean, J., Doan, A., Franklin, M.J., Gehrke, J., Haas, L.M., Halevy, A.Y., Hellerstein, J.M., Ioannidis, Y.E., Jagadish, H.V., Kossmann, D., Madden, S., Mehrotra, S., Milo, T., Naughton, J.F., Ramakrishnan, R., Markl, V., Olston, C., Ooi, B.C., Ré, C., Suciu, D., Stonebraker, M., Walter, T., Widom, J.: The beckman report on database research. ACM SIGMOD Record 43(3), 61–70 (December 2014)
2. Ardagna, C., Damiani, E.: Network and storage latency attacks to online trading protocols in the cloud. In: Proc. of the International Conference on Cloud Computing, Trusted Computing and Secure Virtual Infrastructures. Amantea, Italy (October 2014)
3. Ardagna, C.A., Bellandi, V., Bezzi, M., Ceravolo, P., Damiani, E.: Model-driven methodology for big data analytics-as-a-service. In: Proc. of the 6th IEEE International Congress on Big Data (BigData Congress 2017). Honolulu, HI, USA (June 2017)
4. Ardagna, C.A., Ceravolo, P., Damiani, E.: Big data analytics as-a-service: Issues and challenges. In: Proc. of the IEEE International Conference on Big Data (Big Data 2016). Washington, DC, USA (December 2016)
5. Austin, D.: eDiscovery Trends: CGOCs Information Lifecycle Governance Leader Reference Guide (May 2012), http://www.ediscoverydaily.com
6. Boettiger, C.: An introduction to docker for reproducible research. ACM SIGOPS Operating Systems Review 49(1), 71–79 (2015)
7. Eckhoff, D., Sommer, C.: Driving for big data? privacy concerns in vehicular networking. IEEE Security Privacy 12(1), 77–79 (January 2014)
8. Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., Suri, V.R., Tsou, A., Weingart, S., Sugimoto, C.R.: Big data, bigger dilemmas: A critical review. Journal of the Association for Information Science and Technology 66(8), 1523–1545 (2015)
9. European Commission: Helping SMEs Fish the Big Data Ocean (July 2014), http://ec.europa.eu/digital-agenda/en/news/helping-smes-fish-big-data-ocean

10. IDC: Six patterns of big data and analytics adoption (3 2016), http://www.oracle.com/us/technologies/big-data/six-patterns-big-data-infographic-2956541.pdf
11. IDC: Worldwide Semiannual Big Data and Analytics Spending Guide (October 2016), http://www.idc.com/getdoc.jsp?containerId=prUS41826116
12. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J.M., Ramakrishnan, R., Shahabi, C.: Big data and its technical challenges. Communication of the ACM 57(7), 86–94 (July 2014)
13. Lomotey, R.K., Deters, R.: Analytics-as-a-service framework for terms association mining in unstructured data. International Journal of Business Process Integration and Management (IJBPIM) 7(1), 49–61 (2014)
14. Lu, R., Zhu, H., Liu, X., Liu, J.K., Shao, J.: Toward efficient and privacy-preserving computing in big data era. IEEE Network 28(4), 46–50 (July 2014)
15. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity (2011), http://tinyurl.com/z9wjhuw
16. Markl, V.: Breaking the chains: On declarative data analysis and data independence in the big data era. Proc. of VLDB Endowment 7(13), 1730–1733 (August 2014)
17. Martin, D., Paolucci, M., McIlraith, S., Burstein, M., McDermott, D., McGuinness, D., Parsia, B., Payne, T., Sabou, M., Solanki, M., et al.: Bringing semantics to web services: The owl-s approach. In: Proc. of the International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004). San Diego, CA, USA (July 2004)
18. Martin, K.E.: Ethical issues in the big data industry. MIS Quarterly Executive 14, 2 (2015)
19. Prud, E., Seaborne, A., et al.: Sparql query language for rdf (2006)
20. Rahman, N.: Factors affecting big data technology adoption. http://pdxscholar.library.pdx.edu/cgi/viewcontent.cgi?article=1099 (2016)
21. Russom, P.: Big Data Analytics. TDWI best practices report, TDWI Research (2014), http://www.iso.org/iso/home/news_index/news_archive/news.htm?refid=Ref1821
22. Salleh, K.A., Janczewski, L.: Adoption of big data solutions: A study on its security determinants using sec-toe framework. In: Proc. of the International Conference on Information Resources Management (CONF-IRM 2016). Cape Town, South Africa (May 2016)
23. Wu, D., Greer, M.J., Rosen, D.W., Schaefer, D.: Cloud manufacturing: Strategic vision and state-of-the-art. Journal of Manufacturing Systems 32(4), 564–579 (2013)