Industrial-Driven Big Data as a Self-Service Paradigm for Industry 4.0

Evangelos Georgiou, Nikolaos Michos, Alexandros Papanikolaou, Dimitrios Vasilopoulos

Evangelos Georgiou and Nikolaos Michos - Department of Informatics, Aristotle University of Thessaloniki, Greece;

Abstract The convergence of Internet of Things (IoT), Cloud, and Big Data, creates new challenges and opportunities for data analytics. Human- and machine-created data is being aggregated continuously, transforming our economy and society. To face these challenges, companies call upon expert analysts and consultants to assist them.

In this paper, we present I-BiDaaS, a European Union Horizon 2020 research and innovation project that proposes a self-service solution for Big Data analytics. The solution will be transformative for companies that aim to extract knowledge from big data. It will empower their employees with the right knowledge, and give the true decision-makers the insights they need to make the right decisions. It will shift the power balance within an organization, increase efficiency, reduce costs, improve employee empowerment, and increase profitability. I-BiDaaS aims to empower users to easily utilize and interact with Big Data technologies, by designing, building, and demonstrating, a unified solution that significantly increases the speed of data analysis while coping with the rate of data asset growth, and facilitates cross-domain data-flow towards a thriving data-driven EU economy.

Keywords Big Data, Batch Processing, Stream Processing

1 Introduction

Organizations leverage data pools to drive value, while it is variety, not volume or velocity, which drives big-data investments. The convergence of IoT, cloud, and big data, creates new opportunities for self-service analytics [3] towards big data analytics. Human and machine created data is being aggregated, transforming our economy and society. The aforementioned trends

lead us to one of the main challenges of the data economy [5], Big-Data-asa-Self-Service. A self-service solution will be transformative for organizations, it will empower their employees with the right knowledge, and give the true decision-makers the insights they need to make the right decisions. It will shift the power balance within an organisation, increase efficiency, reduce costs, improve employee empowerment, and increase profitability. The domains that can exploit such self-service solutions are numerous, including among others banking, manufacturing, and telecommunications.

In this paper we present the project I-BiDaaS [2], that aims to address the above challenges and deficiencies. I-BiDaaS targets to empower users to easily utilize and interact with big data technologies, by designing, building, and demonstrating, a unified framework that: significantly increases the speed of data analysis while coping with the rate of data asset growth, and facilitates cross-domain data-flow towards a thriving data-driven EU economy. I-BiDaaS will be tangibly validated by three real-world, industry-lead experiments with significant challenges and requirements: banking, manufacturing, and telecommunications.

I-BiDaaS is a European Union (EU)-funded H2020 3 year project that has started on January 1, 2018, and will last for 3 years. The project consortium comprises the following institutions: Foundation for Research

and Technology Hellas, Greece (FORTH) – project coordinator; Barcelona Supercomputing Center (BSC), Spain; IBM Israel (IBM); Centro Ricierche FIAT SCPA Italy (CRF); Software AG (SAG), Germany; Caixabank, SA, Spain (CAIXA);
The University of Manchester, UK (UNIMAN); Ecole Nationale des Ponts et
Chausees, France (ENPC); ATOS, Spain (ATOS); AEGIS IT Research LTD, UK (AEGIS); Information Technology for Market Leadership, Greece (ITML); University of Novi Sad, Faulty of Sciences, Serbia (UNSPMF); and Telefonica Investigacion y Desarrollo SA, Spain (TID). The main objectives of I-BiDaaS are:

- Develop, validate, demonstrate, and support, a complete and solid big data solution that can be easily configured and adopted by practitioners.
- Break inter- and intra-sectorial data-silos, create a data market and offer new business opportunities, and support data sharing, exchange, and interoperability.
- Construct a safe environment for methodological big data experimentation, for the development of new products, services, and tools
- Develop data processing tools and techniques applicable in real-world settings, and demonstrate significant increase of speed of data throughput and access.
- Develop technologies that will increase the efficiency and competitiveness of all EU companies and organisations that need to manage vast and complex amounts of data.

Current project stage. So far, the project work has focused on: 1) investigating the industrial challenges of the data economy in the fields of Finance, Manufacturing, and Telecommunications; 2) carrying out a through review

of state of the art in the scientific and technological domains relevant to the project; and 3) defining initial data management policy for the data that will be consumed and generated within the project.

2 Approach and Methodology

Based on the challenges and requirements of the three critical domains (i.e., banking, manufacturing, and telecommunications), we aim to develop I-BiDaaS, a solution to enable Big Data as a self-service. It will offer an integrated, fullstack solution for processing and extracting actionable knowledge from big data, that includes: (i) configuration of the underlying infrastructure resources (commodity/public clusters or private clouds), (ii) efficient and automatic usage of computational and storage resources (resource provisioning, data transfers, etc.), (iii) data capture and integration from a variety of different sources and formats (unstructured, noisy, incomplete, etc.), (iv) batch and real-time data processing analytics for fast-growing data, and (v) simple, intuitive, and effective visualization and interaction capabilities for the end-users.

I-BiDaaS will offer Big Data as a Self-Service to enterprises by allowing seamless integration and injection of streaming and batch heterogeneous data, and facilitate the adoption of big data analytics to enterprises that possess big data, but may not have in-house expertise to extract the required actionable knowledge. To achieve this, it will allow the development of new applications or tasks via standard sequential programming, alleviating the burden of dealing with sophisticated analytics techniques (that requires data mining expertise), thus lowering enterprise costs. Also, the platform will be extendible to other application scenarios, as well as compatible with existing platforms such as OpenStack.

2.1 The three-layer architecture: A layer-by-layer description

We now present our layered system architecture, and provide a detailed workflow and user interface description. Conceptually, the architecture is divided into three principal layers: the infrastructure layer, the distributed large-scale layer, and the application layer.

2.1.1 Infrastructure layer

The infrastructure layer includes the actual underlying storage and processing infrastructure of the I-BiDaaS solution, nominally provided and managed by ATOS and FORTH. This includes (i) a private cloud infrastructure provided by ATOS, (ii) a commodity cluster provided by FORTH (which consists of highend GPUs, Intel Phi accelerators, and powerful multi-core CPUs that contain secure enclaves that are able to protect the code and the sensitive data). We note that the I-BiDaaS solution will be deployable to other infrastructure
4premises as well; for instance, in the end user scenario that involves CAIXA, the platform will be deployed within CAIXA proprietary private cloud.

2.1.2 Distributed large-scale layer

The distributed large-scale layer is responsible for the orchestration and management of the underlying physical computational and storage infrastructure. It allows the effective and efficient use of the infrastructures and enables the application layer to provide effective big data analytics. The distributed largescale layer is responsible for the following tasks: (i) task and data dependency capturing, (ii) data transfer optimization, (iii) task and data scheduling, (iv) resource provisioning and management, and (v) capturing, integrating, and preparing data from heterogeneous, distributed sources,

2.1.3 Application layer

The application layer sits on top of the distributed large-scale layer. It refers to the architecture aspects and components that are involved in the actual workflow of extracting actionable knowledge from the big data, starting from data preparation and analytics, to delivering results for supporting decision making. The data analytics include both batch and real time processing of streaming data. The data from heterogeneous sources are ingested in the solution. For early development scenarios when not sufficient real data is available, we will use the IBM's data fabrication platform [1].

In terms of interleaving batch and stream processing, the proposed solution goes beyond the traditional lambda architecture [4]. It uses a complex event analysis system, combined with a hardware-based implementation of streaming analytics that uses many different many-core accelerators (GPUs, Intel Phi, etc.). This design allows us to offload parts of the streaming analytics that can be parallelized and gives us the opportunity to partition the analytics queries, between the high-level stream processing engine and the low-level, hardware-optimized implementation. By carefully performing part of the queries at the lowest level (especially for the filtering), only the required data will be forwarded to the stream-processing engine for a more sophisticated analysis, while the remainder will be ignored at the earliest possible. The partition of the queries (between the complex event analysis system and the hardware-based streaming analytics) can be done either statically (i.e., during the implementation of a specific user query) or dynamically, at run-time (i.e., by monitoring the execution of a user-defined query, and deciding if the offloading to a manycore processor would lead to better performance.

2.1.4 User interface

In order to make our solution ease to use by end-users, we will build a multipurpose interface (AEGIS), that can be used by different categories of users. The interface will provide different levels of abstractions, tailored to different

categories of user expertise. First, we will offer a programming API for access to every level of our software stack. This will give the flexibility to experienced IT users to utilise every aspect of our solution, and fine-tune their applications. The API will give access to the high-level application components — such as the advanced machine-learning modules and the streaming analytics — as well as to the low-level infrastructure layer, such as the scheduling and the resource management of the underlying infrastructure. Second, we will provide a domain language for access to the application layer. The purpose of this language is to offer an easy way to program data analytics (either batch or stream processing) without caring about scalability issues and infrastructure placement.

3 Conclusions

In this paper, we presented I-BiDaaS, a European Union Horizon 2020 project that proposes a solution for Big Data as a self-service. Once achieved, the solution will be transformative for enterprises that seek to extract actionable knowledge from Big Data, as it will allow their employees to easily utilize and interact with Big Data technologies. This can lead to increased efficiency, reduced costs, and increased profitability within an enterprise.

References

1. Creating secure test data to test systems. https://www.ibm.com/blogs/research/2014/ 07/creating-secure-test-data-to-test-systems/.
2. I-BiDaaS: Industrial-Driven Big Data as a Self-Service Solution. https://www.ibidaas. eu.
3. Self-Service Analytics. https://www.gartner.com/it-glossary/ self-service-analytics/.
4. Marz, N., and Warren, J. *Big Data: Principles and Best Practices of Scalable Realtime Data Systems*, 1st ed. Manning Publications Co., Greenwich, CT, USA, 2015.
5. Passlick, J., Lebek, B., and Breitner, M. H. A Self-Service Supporting Business Intelligence and Big Data Analytics Architecture. In *Wirtschaftsinformatik* (2017).