DataCloud: Enabling the Big Data Pipelines on the Computing Continuum

Dumitru Roman[1], Nikolay Nikolov[1], Brian Elvester[1], Ahmet Soylu[2], Radu Prodan[3], Dragi Kimovski[3], Andrea Marrella[4], Francesco Leotta[4], Dario Benvenuti[4], Mihhail Matskin[5], Giannis Ledakis[6], Anthony Simonet-Boulogne[7], Fernando Perales[8], Evgeny Kharlamov[9], Alexandre Ulisses[10], Arnor Solberg[11], and Raffaele Ceccarelli[12]

[1] SINTEF AS, Norway
[2] Oslo Metropolitan University, Norway
[3] University of Klagenfurt, Austria
[4] Sapienza University of Rome, Italy
[5] KTH Royal Institute of Technology, Sweden
[6] UBITECH, Greece
[7] iExec, France
[8] JOT, Spain
[9] Bosch Center for Artificial Intelligence, Germany
[10] MOG, Portugal
[11] Tellu, Norway
[12] Ceramica Catalano, Italy

1 Summary of the project

With the recent developments of Internet of Things (IoT) and cloud-based technologies, massive amounts of data are generated by heterogeneous sources and stored through dedicated cloud solutions. Often organizations generate much more data than they are able to interpret, and current Cloud Computing technologies cannot fully meet the requirements of the Big Data processing applications and their data transfer overheads [1]. Many data are stored for compliance purposes only but not used and turned into value, thus becoming *Dark Data*, which are not only an untapped value, but also pose a risk for organizations [3].

To guarantee a better exploitation of Dark Data, the **DataCloud project**[13] aims to realize novel methods and tools for effective and efficient management of the Big Data Pipeline lifecycle encompassing the Computing Continuum.

Big Data pipelines are composite pipelines for processing data with non-trivial properties, commonly referred to as the Vs of Big Data (e.g., volume, velocity, value, etc.) [4]. Tapping their potential is a key aspect to leverage Dark Data, although it requires to go beyond the current approaches and frameworks for Big Data processing. In this respect, the concept of *Computing Continuum*

[13] DataCloud is a Research and Innovation project funded by the European Commission under the Horizon 2020 program (Grant number 101016835). The project runs for three years, between 2021-2023.

extends the traditional centralised Cloud Computing with Edge[14] and Fog[15] computing in order to ensure low latency pre-processing and filtering close to the data sources. This will prevent to overwhelm the centralised cloud data centres enabling new opportunities for supporting Big Data pipelines.

2 Objectives and Expected Results

The main objective of the project is to develop a software ecosystem for managing Big Data pipelines on the Computing Continuum. The ecosystem consists of new languages, methods and infrastructures for supporting Big Data pipelines on heterogeneous and untrusted resources. Six lifecycle phases are covered:

1. *Pipeline discovery* concerns discovering Big Data pipelines from various data sources.
2. *Pipeline definition* deals with specifying pipelines featuring an abstraction level suitable for pure data processing.
3. *Pipeline simulation* aims to evaluate the performance of individual steps to test and optimise deployments.
4. *Resource provisioning* is concerned about securely provisioning a set of (trusted and untrusted) resources.
5. *Pipeline deployment* is concerned with deployment of pipelines across the provisioned resources.
6. *Pipeline adaptation* deals with optimised run-time provisioning of computational resources.

The ecosystem separates the design-time from the run-time deployment of pipelines and complements modern serverless approaches [2]. A set of research challenges related to each pipeline phase will be tackled within the project, such as the advancement of process mining techniques to learn the structure of pipelines, the definition of proper DSLs for pipelines, novel approaches for pipeline containerisation and blockchain-based resource marketplaces, etc.

The expected impact of DataCloud is to lower the technological entry barriers for the incorporation of Big Data pipelines in organizations' workflows and make them accessible to a wider set of stakeholders regardless of the hardware infrastructure. To achieve this, DataCloud will validate its results through a strong selection of complementary business cases offered by four SMEs and a large company targeting higher mobile business revenues in smart marketing campaigns, reduced live streaming production costs of sport events, trustworthy eHealth patient data management, and reduced time to production and better analytics in Industry 4.0 manufacturing.

Acknowledgments. This work has been supported by the Horizon 2020 project DataCloud (Grant number 101016835).

[14] Edge Computing is a paradigm that brings computation and data storage closer to the location where it is needed to improve response times and save bandwidth.

[15] Fog Computing uses edge devices to carry out a substantial amount of computation, storage, and communication locally and routed over the Internet backbone.

References

1. Barika, M., Garg, S., Zomaya, A.Y., Wang, L., Moorsel, A.V., Ranjan, R.: Orchestrating big data analysis workflows in the cloud: research challenges, survey, and future directions. ACM Computing Surveys (CSUR) **52**(5), 1–41 (2019)
2. Castro, P., Ishakian, V., Muthusamy, V., Slominski, A.: The Rise of Serverless Computing. Communications of the ACM **62**(12) (2019)
3. Gimpel, G.: Bringing dark data into the light: Illuminating existing IoT data lost within your organization. Business Horizons **63**(4), 519–530 (2020)
4. Plale, B., Kouper, I.: The centrality of data: data lifecycle and data pipelines. In: Data Analytics for Intelligent Transportation Systems, pp. 91–111. Elsevier (2017)