Big Data Pipelines on the Computing Continuum: Ecosystem and Use Cases Overview

Dumitru Roman*, Nikolay Nikolov*, Ahmet Soylu†, Brian Elvesæter*, Hui Song*, Radu Prodan‡, Dragi Kimovski‡, Andrea Marrella§, Francesco Leotta§, Mihhail Matskin¶, Giannis Ledakis‖, Konstantinos Theodosiou‖, Anthony Simonet-Boulogne**, Fernando Perales††, Evgeny Kharlamov‡‡, Alexandre Ulisses[x], Arnor Solberg[xi], Raffaele Ceccarelli[xii]

*SINTEF AS, Norway
†OsloMet – Oslo Metropolitan University, Norway
‡University of Klagenfurt, Austria
§Sapienza University of Rome, Italy
¶KTH Royal Institute of Technology, Sweden
‖UBITECH, Greece
**iExec, France
††JOT, Spain
‡‡Bosch Center for Artificial Intelligence, Germany
[x]MOG, Portugal
[xi]Tellu, Norway
[xii]Ceramica Catalano, Italy
Contact email: *dumitru.roman@sintef.no*

***Abstract*—Organisations possess and continuously generate huge amounts of static and stream data, especially with the proliferation of Internet of Things technologies. Collected but unused data, i.e., Dark Data, mean loss in value creation potential. In this respect, the concept of Computing Continuum extends the traditional more centralised Cloud Computing paradigm with Fog and Edge Computing in order to ensure low latency pre-processing and filtering close to the data sources. However, there are still major challenges to be addressed, in particular related to management of various phases of Big Data processing on the Computing Continuum. In this paper, we set forth an ecosystem for Big Data pipelines in the Computing Continuum and introduce five relevant real-life example use cases in the context of the proposed ecosystem.**

***Index Terms*—Big Data, Computing Continuum, Dark Data, Data Pipelines, Cloud-Fog-Edge Computing.**

I. Introduction

Big data pipelines are composite processing flows for data characterised by the so-called "Vs" of Big Data, such as volume, velocity, and variety. Big Data pipelines are traditionally set up to process large amounts of real-time data in the Cloud, which offers elastic on-demand resource provisioning. As the Internet of Things (IoT) devices generate massive amounts of data that can overwhelm the relatively centralised Cloud data centres, low latency pre-processing and filtering of data close to the data sources is needed [1]. Thus, Cloud Computing cannot fully meet the requirements of Big Data processing applications and their data transfer overheads. As a consequence, the collected but unused data assets become Dark Data, which represents not only untapped value, but also could pose a risk for organisations. In this respect, the Computing Continuum creates a fluid ecosystem by extending the Cloud Computing with the emerging Fog and Edge Computing paradigms by pushing the services that are traditionally bounded within data centres towards remote network nodes [2]. Big Data pipelines on the Computing Continuum require management and usage of heterogeneous computing resources [3]. This opens up challenges spanning the entire lifecycle of Big Data pipelines, including effective discovery, modelling, simulation, and deployment of Big Data pipelines over heterogeneous resources from a diverse set of providers (i.e., using trustworthy resources).

In this paper, we propose an ecosystem for Big Data pipelines processing on the Computing Continuum. We then present and analyse five real-life example use cases from a set of different application domains that provide concrete insights on the challenges concerning Big Data processing on the Computing Continuum.

The rest of the paper is organized as follows. Section II introduces the envisioned ecosystem. Section III presents a high-level overview of the five use cases, followed by a preliminary analysis in Section IV. Finally, Section V summarises the paper and outlines future work.

II. Big Data Pipeline Ecosystem

An ecosystem for managing Big Data pipelines on the Computing Continuum needs to cover various phases of their lifecycle with corresponding methods and tools, and to involve the complete set of relevant stakeholders. This requires novel

This is the author accepted version of an article published in
2021 IEEE Symposium on Computers and Communications (ISCC)
https://doi.org/10.1109/ISCC53001.2021.9631410

methods to support the complete Big Data pipeline processing, enabling their discovery, definition, model-based analysis and optimisation, simulation, deployment, adaptive run-time provisioning, and monitoring on top of decentralised heterogeneous infrastructures on the Computing Continuum. The ecosystem aims to make the execution of Big Data pipelines traceable, trustable, manageable, analysable, and optimisable.

We consider the following phases for the lifecycle of Big Data pipelines on the Computing Continuum, supported by corresponding techniques and tools:

1) *Pipeline discovery* concerns discovery of Big Data pipelines based on available metadata and traces;
2) *Pipeline definition* concerns specification of 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 pipeline deployments;
4) *Resource provisioning* concerns secure provisioning of (trusted and untrusted) resources;
5) *Pipeline deployment* concerns deployment of pipelines across the provisioned resources;
6) *Pipeline adaptation* concerns optimised run-time provisioning of computational resources.

These phases depicted in Figure 1 involve relevant stakeholders in a feedback loop, as follows:

1) *Data providers* provide static and stream data for the pipelines, including event data that could be used for pipeline discovery;
2) *Data scientists* provide the necessary implementation details for the pipelines' steps and test them through simulation;
3) *Resource providers* provide a pool of hardware and software resources as part of the Computing Continuum;
4) *DataOps operators* automatically deploy designed and tested pipelines over the provisioned resources;
5) *Data consumers* exploit the insights and knowledge generated by the execution of Big Data pipelines.

The ecosystem separates the design-time from the run-time deployment of pipelines and complements modern serverless approaches [4]. At design-time, pipelines are learned/discovered from the data sources, designed, customised, and simulated based on the provisioned resources, and deployed as-a-service. At run-time, new data from the data providers are served as input to the deployed pipelines, which execute and deliver data as output in the form of smart data and insights that represent actionable knowledge for data consumers.

III. Use Cases

We exemplify the proposed ecosystem with an overview of a set of real-life use cases, carefully selected to cover the stringent requirements imposed by Big Data pipelines lifecycle phases from different companies aiming at utilising Big Data pipelines on the Computing Continuum for their businesses.

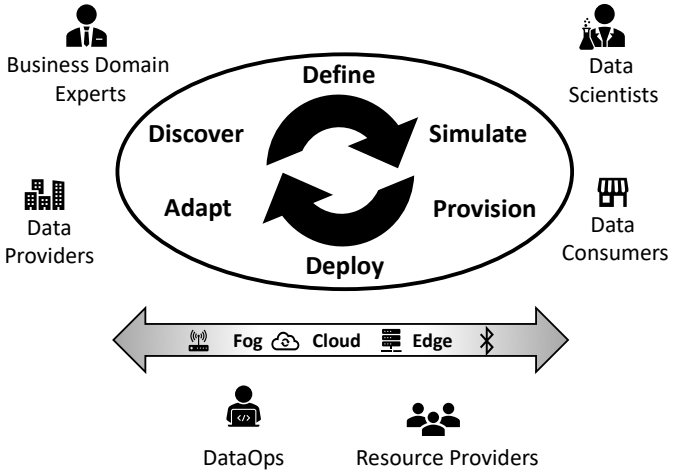

Fig. 1. Ecosystem overview for Big Data pipeline lifecycle on the Computing Continuum.

A. Smart Mobile Marketing Campaigns

Digital marketing represents a critical industry connecting companies offering product, services, and content to their targeted audience. This connection takes place by defining and launching digital marketing campaigns, whose core raw data are the keywords (i.e., words, short sentences) defining the company's offers used by Search Ads Platforms (e.g., Google, Bing) to correlate with the user interest/queries.

A marketing company implements massive data-driven management of marketing campaigns, which enables the generation of a big historical data set containing the main performance indicators. However, due to the high amount and diversity of data, and the enormous amount of potential configuration and AI libraries, the company is not capable of implementing a trusted, robust, and reliable data pipeline to generate the required insights guiding the marketing investment strategy. In addition to the Cloud storage, the company has an on-premise cluster composed of several servers. The company has experienced many problems related to node configuration or memory capacity distribution when implementing novel data pipelines for new insights generation.

In the proposed ecosystem, pipelines are defined and simulated by domain experts and data scientists to decide the "best" configuration in terms of extract-transform-load and analytical tools (e.g., algorithms, models, training data sets). Once the model is selected, the "best" technological architecture can be deployed to generate the expected information.

B. Automatic Live Sports Content Annotation

The transmission of live sport events is one type of audio-visual content that reaches a higher number of final spectators. Spectators are ever more demanding in the way they expect to get involved in the event. Therefore, it is crucial to provide the public with new resources and experiences, such as enhancing viewer engagement by automatically annotating large quantities of media data from various sources.

A media company wants to use video and audio frames to identify patterns and events to summarise the entire show later on. Initially, cameras generate media data (at 3Gbit/s) spread around a sports venue, covering different viewpoints of the sports event and of the players. The live stream of video and audio data flows in an unstructured way to a Cloud-based platform for analysis to find relevant events or players during the match. Dedicated trusted connections between the cameras and the Cloud could lower the data vulnerability. If available, computational resources near the event facility could lower the latency, reduce costs, and improve the general experience. Audio-visual content needs to pass through a combination of complex AI processing pipelines, such as audio recognition, pattern extraction, and classification.

In the proposed ecosystem, data scientists and video experts discover, define, and simulate multiple events pipelines to obtain the "best" algorithms for detecting events, and the "best" processes to enrich the videos with the obtained metadata. Pipelines can then be deployed to generate the expected information and annotate the videos. This process must be supported through intelligent resource management.

C. Digital Health System

Next generation IoT-enabled medical devices and remote supervision sensors will enable new advanced telecare services. Continuous progress in the remote healthcare domain has led to the invention of new wireless devices and wearables, which can obtain a comprehensive view of the peoples' overall health based on their vitals. They collectively trace diseases of individuals and the overall community to understand the evolution of a viral infection, for example.

A digital health system helps care providers to remotely supervise and follow their patients while staying at home during treatment and care and supports elderly patients to accomplish self-care living at home. The system needs to gather data from medical devices and cameras and perform local processing, such as filtering, encryption, anonymisation and local storage for security and privacy reasons. The system also needs to orchestrate a set of Software-as-a-Service applications for patients and health personnel and to integrate and exchange data to third party services such as national Electronic Health Record systems. Overall, the system has to manage and orchestrate data pipelines in a trustworthy way across the Edge, Fog, and Cloud resources.

In the proposed ecosystem, data processing pipelines on Edge devices are learned and specified based on the various IoT patient monitoring sensors. Such pipelines are then tested through simulation and deployed on personal health gateways and eHealth platforms, which make smart use of resources through adaptation.

D. Predicting Deformations in Ceramics Manufacturing

Smart manufacturing combines information, technology and human knowledge, bringing a rapid revolution in the development and application of manufacturing intelligence to every aspect of business. A typical manufacturing plant uses information technology, sensors, actuators, and computerised controls to manage each specific stage or operation of a manufacturing process.

In the ceramic industry, the deformation of ceramic materials during the drying and baking steps of the ceramics is a well-known issue, as it changes the shape of the ceramic element and makes the final product different from the desired one at design time. To this end, skilled technicians perform the mould design of ceramic products using a trial-and-error monitoring of all the steps of the production process. This makes the production of a new sanitary item a cyclic process that continuously monitors the mould model to contrast the deformation and reach the exact design dimensions for the final shape. Consequently, the time and the related costs required for a new component to enter in production are high and strongly limit the productive capacity of ceramic factories.

In the proposed ecosystem, three-dimensional scanning, coupled with the information flow from the data sources already active in the factory, such as sensors, computer-aided design models, and technicians' inputs, are used to discover, define, and simulate data pipelines underlying the production processes. Leveraging the support of the skilled technicians and data scientists, the models "best" reflecting the products can be deployed with adaptation support.

E. Analytics of Manufacturing Assets

Fully computerised and automated Industry 4.0 production processes require a stack of analytical applications for constant monitoring, diagnostics, and optimisation of production assets, including assembly lines and manufacturing robots. This requires equipping production assets with sensors reporting characteristics, such as temperature, state, and operation errors.

Production of electronic panels involves the placement and soldering of electronic components on print circuit boards of several machines, including surface mount devices (SMD) that place components on boards, electric ovens that solder components to the boards, and solder joint inspection (SJI) devices that detect flaws in joints and misplacement of the components. All these machines are equipped with sensors that produce data of various complexity and heterogeneity. For example, SMDs usually use multiple (vacuum) nozzles to pick (possibly multiple) components for placement, which triggers a sequence of picking and placing actions and generates the corresponding temporal data. The analytics of manufacturing assets solutions offered by a producer of sensors and related chips include failure root-cause analyses, process quality monitoring, production quality monitoring, and anomaly detection.

In the proposed ecosystem, data collected from different sources, such as SMD and SJI, are used to discover, define, and simulate data pipelines. Data scientists specify appropriate AI methods for the simulation and deployment of the pipelines on a pool of resources, which are actively adapted over the provisioned trusted resources.

IV. Preliminary Use Case Analysis

A preliminary analysis of the use cases provides a good coverage of Big Data pipeline phases and stakeholder involve-

TABLE I
USE CASES AND THEIR DEGREE OF COVERAGE FOR PIPELINE LIFECYCLE.

Case	Pipeline Discovery	Pipeline Definition	Pipeline Simulation	Resource Provisioning	Pipeline Deployment	Pipeline Adaptation
Marketing	Low	Moderate	High	Moderate	High	Low
Media	Moderate	Moderate	Moderate	High	High	High
Healthcare	Low	High	High	Moderate	High	Moderate
Manufacturing	High	High	Moderate	Moderate	Moderate	High
Industry 4.0	High	High	Moderate	High	Moderate	Low

TABLE II
STAKEHOLDERS AND THEIR DEGREE OF INVOLVEMENT IN EACH USE CASE.

Case	Data providers	Business domain experts	Data scientists	Resource providers	DataOps operators	Data consumers
Marketing	High	Moderate	High	Moderate	Moderate	High
Media	High	Low	Moderate	High	High	High
Healthcare	High	Moderate	Moderate	Moderate	High	High
Manufacturing	High	High	Low	Moderate	Moderate	High
Industry 4.0	High	High	High	Moderate	Moderate	High

ment, as outlined in the proposed ecosystem. Each lifecycle phase is covered by at least two use cases, and each use case covers at least two steps in the lifecycle. Furthermore, each use case covers at least one type of other stakeholder with high importance, besides a high level of involvement of data provides and consumers.

Table I presents a summary of the preliminary analysis that highlights the importance of each Big Data pipeline phase to each selected use case. Table II presents an analysis of the degree of involvement of various stakeholders. The tables highlight a variety of cases with different emphasis on various lifecycle steps and relevance for specific stakeholders.

In collaboration with use case providers, we identified the following benefits of the proposed ecosystem for the use cases:

- Improved product quality and operation performance through simulation, adaptation, and monitoring;
- Lower production and operation costs through on-demand adaptive provisioning of resources;
- Easy monitoring and visualisation of run-time operation and production data pipeline performance by system engineers;
- Effective use of provisioned trusted Cloud-Fog-Edge resources and infrastructure for the orchestration of Big Data pipelines;
- Enabling direct collaboration between stakeholders, such as domain experts and data scientists;
- Improved trustworthiness of data management for delivering information to the correct stakeholders, taking into account security and privacy concerns;
- Improved data workflow management and deployment, such as management and exploitation of data pipelines that span across largely heterogeneous resources;
- Possibility to optimise the performance of AI algorithms through defining and simulating data pipelines.

V. SUMMARY AND OUTLOOK

In this article, we proposed an ecosystem for managing Big Data pipelines on the Computing Continuum including several phases, ranging from discovery to adaptation, and a set of stakeholders ranging from data providers to consumers. Our ecosystem separated the design-time and run-time aspects of pipeline execution. Finally, we provided a set of real-life use cases from different application domains to exemplify the possible use of the proposed ecosystem. We plan to realise the ecosystem by developing novel techniques and tools to support each phase in the pipeline lifecycle, and by further specifying and implementing the proposed use cases.

ACKNOWLEDGEMENT

This work was partly funded by the European Commission Horizon 2020 project *"DataCloud: Enabling The Big Data Pipeline Lifecycle on the Computing Continuum"* (Grant number 101016835).

REFERENCES

[1] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog Computing and Its Role in the Internet of Things," in *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing (MCC 2012)*. ACM, 2012, p. 13–16.

[2] D. Kimovski, R. Matha, J. Hammer, N. Mehran, H. Hellwagner, and R. Prodan, "Cloud, Fog or Edge: Where to Compute?" *IEEE Internet Computing*, vol. (in press), 2021.

[3] M. Barika, S. Garg, A. Y. Zomaya, L. Wang, A. V. Moorsel, and R. Ranjan, "Orchestrating Big Data Analysis Workflows in the Cloud: Research Challenges, Survey, and Future Directions," *ACM Computing Survey*, vol. 52, no. 5, 2019.

[4] P. Castro, V. Ishakian, V. Muthusamy, and A. Slominski, "The Rise of Serverless Computing," *Communications of the ACM*, vol. 62, no. 12, p. 44–54, Nov. 2019.