Big Data Pipelines on the Computing Continuum: Tapping the Dark Data

Dumitru Roman
SINTEF AS, Norway

Radu Prodan
University of Klagenfurt, Austria

Nikolay Nikolov
SINTEF AS, Norway

Ahmet Soylu
Oslo Metropolitan University, Norway

Mihhail Matskin
Royal Institute of Technology, Sweden

Andrea Marrella
Sapienza University of Rome, Italy

Dragi Kimovski
University of Klagenfurt, Austria

Brian Elvesæter
SINTEF AS, Norway

Anthony Simonet-Boulogne
iExec, France

Giannis Ledakis
UBITECH, Greece

Hui Song
SINTEF AS, Norway

Francesco Leotta
Sapienza University of Rome, Italy

Evgeny Kharlamov
Bosch Center for Artificial Intelligence, Germany

***Abstract*—Big Data pipelines are essential for leveraging Dark Data, i.e., data collected but not used and turned into value. However, tapping their potential requires going beyond existing approaches and frameworks for Big Data processing. The Computing Continuum enables new opportunities for managing Big Data pipelines concerning efficient management of heterogeneous and untrustworthy resources. This article discusses the Big Data pipelines lifecycle on the Computing Continuum, its associated challenges and outlines a future research agenda in this area.**

This is the authors accepted manuscript of an article published in
Computer, volume 55, issue 11, 2022
https://doi.org/10.1109/MC.2022.3154148

CLOUD COMPUTING has been a major disruptive technology in the last decade, providing resources-as-a-service for diverse Internet applications and offering elastic capacity and customisable connectivity over a large-scale network. However, Big Data processing applications' resilience, sustainability, and human-centric collaborative requirements demand an interoperable end-to-end ecosystem that pushes the data centre infrastructure services towards remote nodes closer to the data sources. In this context, it is crucial to employ serverless computing and blockchains technologies to develop a new generation of scalable and secure Big Data-aware Cloud infrastructures [1]. The **Computing Continuum** extends the Cloud services with emerging Edge and Fog computing paradigms, reducing overheads for transferring distributed data into remote data centres [2]. In the Big Data domain, eminent challenges in supporting the processing of Big Data remain, including effective discovery, modelling and simulation of Big Data pipelines and their trustworthy deployment over heterogeneous resources from different providers [3].

Big Data pipelines are composite processing and communication streams with non-trivial properties and characteristics. Examples of the so-called Big Data "Vs" include volume, velocity, variety, veracity, validity, value, variability, venue, vocabulary, vagueness, etc. Big Data pipelines require management and usage of heterogeneous computing resources on the Computing Continuum; however, providing a general-purpose solution for transparent Big Data pipelines characterisation and control across the Computing Continuum is still an open research problem [4].

This article introduces research and design challenges related to the lifecycle of the Big Data pipelines on the Computing Continuum, while not aiming to give a comprehensive review. Its ultimate purpose is to outline a research agenda for future work in this area.

1. TAPPING THE DARK DATA

The Internet of Things (IoT) allows seamless cyber-physical integration of computing services in disparate application domains. The IoT devices generate massive amounts of data that can overwhelm the centralised Cloud data centres and require low latency pre-processing and filtering close to the data sources [5]. Without proper handling, the collected assets often become Dark Data, comprising the mass of text, tables, images, and other unstructured and untapped data stored for compliance purposes. If not exploited, Dark Data brings organisations more risks than added value. Additional primary sources of Dark Data are organisations running their core services on data generated and (pre-) processed outside the boundaries of their (Cloud) data centres.[1]

To take advantage of the untapped Dark Data, the Computing Continuum offers organisations a dynamic infrastructure for adaptive resource management and data processing strategies, tailored according to the application needs. However, several barriers related to infrastructure management, Big Data processing and Dark Data analysis hinder the effective use of the Computing Continuum.

Dark Data and the Computing Continuum are tightly related concepts. Dark Data produced across the Computing Continuum in such a distributed setting is technically impractical to collect and process in a centralised manner. Organisational and legal obstacles that hinder the data transfer, storage and analysis in a centralised location aggravate the problem and require novel paradigms for managing data pipelines.

Cloud, Fog, Edge infrastructures enable the generation of massive amounts of **unused and devalued Dark Data**.

[1] https://www.gartner.com/smarterwithgartner/what-edge-computing-means-for-infrastructure-and-operations-leaders

Big Data pipelines are *essential for leveraging Dark Data*, but valorising and tapping their potential requires going beyond the current state-of-the-art in Big Data processing.

The **Computing Continuum** enables new opportunities for supporting Big Data pipelines but *requires efficient management of heterogeneous and untrustworthy resources*.

2. AN ECOSYSTEM FOR MANAGING BIG DATA PIPELINES ON THE COMPUTING CONTINUUM

Tapping the full potential of Dark Data requires critical research related to efficient usage of the Computing Continuum and Big Data pipeline processing and analysis. The envisioned ecosystem for managing the Big Data pipeline lifecycle on Computing Continuum comprises a feedback sequence of six phases, depicted in Figure 1: *Discover* → *Define* → *Simulate* → *Provision* → *Deploy* → *Adapt*. These phases involve a set of relevant stakeholders discussed in this section.

1) ***Discover***: The Big Data pipeline definition process starts by analysing a *provider*'s Dark Data that consists of various sources (static data and event streams). The goal is to discover the structure and properties of the Big Data pipelines and provide input to their definition.

2) ***Define***: *Business domain experts* use the domain-specific processing requirements extracted from the Dark Data to structure, define, configure, and design Big Data pipelines. For this purpose, they use a Domain Specific Language (DSL) designed explicitly for the pipelines before deploying them. *Data scientists* with artificial intelligence (AI) and machine learning (ML) expertise inject the implementation details of the Big Data pipelines, such as data-specific analytical models and processing codes.

3) ***Simulate***: Pipeline simulation (used by business domain experts and data scientists) tests the Big Data pipeline definition before deployment on the Computing Continuum. Simulation is essential to estimate the resource needs for the deployment and execution of pipelines.

4) ***Provision***: A decentralised blockchain-based marketplace provides a pool of Computing Continuum resources (hardware and software) belonging to untrustworthy third-party *resource providers* for pipeline deployment (e.g., Cloud virtual machines, integrated access devices, sensors) and engaging *DataOps operators*.

5) ***Deploy***: After designing and testing the Big Data pipeline, the DataOps operators automatically deploy it across the provisioned Computing Continuum resources.

6) ***Adapt***: An intelligent and data-aware adaptive scheduling mechanism addresses the dynamic Big Data pipeline runtime (i.e., failures, velocity fluctuations, infrastructure drifts) under the DataOps supervision.

A two-stage approach hides the technical complexity between the data consumers and data providers, making the process of managing Big Data pipelines more transparent, efficient and effective:

1) At *design time*, pipelines are discovered (and learned) from the data sources, designed (and customised), simulated based on the provisioned marketplace Continuum resources, and deployed as-a-Service.

2) At *runtime*, pipelines are monitored and adapted as new data from the data providers is served as input to them. They execute and deliver valuable data outputs and insights that represent actionable knowledge for *data consumers*.

In contrast to the concurrent Big Data ecosystems and ETL (extract, transform, and load) platforms, such as xPlenty[2] and Airflow[3], the envisioned ecosystem encompasses the entire lifecycle of managing Big data pipelines across the Computing Continuum [3]. This lifecycle allows clear separation of the design and runtime deployment of pipelines and enables modern serverless execution [6]. The ecosystem further employs decentralised control to support the integration of untrustworthy third-party resources from various control domains, which frees the stakeholders from the monopoly of the Cloud providers [7].

[2] https://xplenty.com
[3] https://airflow.apache.org

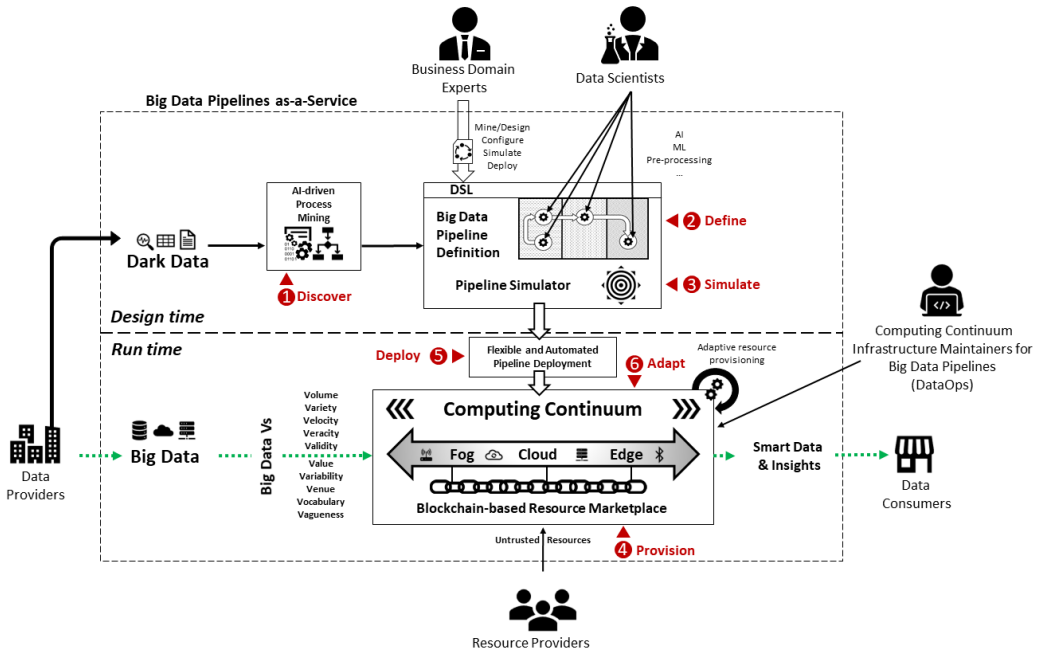

Figure 1. Envisioned ecosystem for managing Big Data pipelines on the Computing Continuum.

However, this benefit raises new challenges for securing the ecosystem from third-party misuse and lacks a common data governance framework.

Example scenario: In modern manufacturing industries (e.g., sanitary ceramic manufacturing), business domain experts typically know the abstract workflows executed by their organisations but are unaware of the generated Big Data pipelines underneath. These industries employ advanced technologies such as smart robotic arms, sensors, etc., ensuring high production precision in manufacturing parts (e.g., ceramic-based sanitary items). However, they do not use the Dark Data collected within the production line (e.g., environment parameters like temperature, humidity, etc.) for flexible real-time operation and trouble-free manufacturing.

To date, experts often perform the quality control of a manufacturing part manually at the end of the production process. If a component is subject to a defect that changes its shape in a non-recoverable way, it is impossible to predict it before completing the production process. On the other hand, employing, for example, Big Data pipeline discovery techniques allow data scientists to identify the relevant data pipelines underlying the production process of the ceramic industry. Thus, monitoring their progress over time to pinpoint the most influential data in case of defects. Pushing this data in a pipeline ecosystem (benefiting from a pool of resources on the

Computing Continuum) enables deep data analysis and allows data scientists to infer, for example, significant changes in the values that recurrently bring to a defect. This pipeline allows manufacturing experts predict potential (recurring) defects before the production process by only monitoring the behaviour of the involved data pipelines. Consequently, this empowers manufacturing industries to reduce the time to production of a new component targeting a zero-error manufacturing process.

3. BARRIERS

It is essential to lower the barriers discussed here to enable organisations incorporate Big Data pipelines in their business processes and make them accessible to a broader set of stakeholders regardless of the hardware infrastructure. Although Grid computing addressed these barriers at varying maturity levels, the solutions targeted primarily scientific applications without considering Dark Data for industrial application pipelines. Moreover, Grid infrastructures used to connect homogeneous bare-metal supercomputers lack the heterogeneity and mobility present in today's virtualised continuum platforms.

3.1. Discovery barrier

Big Data processing often relies on the assumption of an already known data pipeline anatomy at the outset before running any Big Data processing feature [8]. Therefore, the core challenge in discovering data pipelines is mapping and analysing their structure (and significant variations) by turning torrents of event data generated by complex Computing Continuum into valuable insights related to workflow performance and compliance. The discovery activity is crucial to identify bottlenecks, inefficiencies and risks hidden behind the complexity of data pipelines, which prevent or delay the proper pipeline enactment in Cloud, Fog and Edge environments.

3.2. Definition barrier

The diversity and complexity of data modelling and processing pipelines and the heterogeneity of the Computing Continuum require a multidisciplinary effort using expert domain and technical knowledge of the computational environment. However, the collaboration between domain and technical experts requires repeated, time-consuming, and error-prone communication cycles and introduces a learning gap with significant overhead. Therefore, it is critical to bridge the gap between domain and technical experts by using simple and easy to use high-level definition languages and supporting tools [9].

3.3. Simulation barrier

Simulating complex and dynamic Big Data pipelines running on top of many heterogeneous services and infrastructure resources is essential for increasing their performance and accuracy. Existing simulation approaches in business process and workflow modelling rely on prior performance knowledge, such as throughput, time and resource utilisation, and are rarely accurate. Since Big Data pipeline execution and throughput are non-deterministic (i.e., vary significantly on the heterogeneous computing and network resources), it is critical to provide pipeline simulation frameworks that help predict their overall performance across the Computing Continuum analytically. The simulation tool needs to implement a step-level performance analysis to automatically predict the necessary resources without the need for expensive deployments and to measure the performance of individual steps for scaling the pipeline steps [10].

3.4. Provisioning barrier

Provisioning resources in the Cloud requires selecting and trusting the provider and its resources, difficult to achieve in the Computing Continuum with a myriad of small providers and limited resources [11]. This challenge requires new models for open hardware and software resource marketplaces in the Computing Continuum, together with novel trust mechanisms. Recent advancements in blockchain technologies enable heavy computations on untrustworthy servers whilst delegating the trust to smart contracts. This principle of off-chain computing builds on repli-

cation, reputation and decentralised consensus algorithms, hardware cryptography and monetary incentives to extend some properties of smart contracts (trust, traceability and transparency) to regular Cloud resources. Off-chain computing, however, is still limited to simple applications, a pay-per-task model, significant deadlines, and offers no service-level agreement (SLA). To support modern Big Data pipelines and compete with the leading Cloud computing platforms, it is critical that provisioning mechanisms for Fog and Edge computing cater towards supporting the same applications and providing similar SLA.

3.5. Deployment barrier

The traditional pipelines management methods use the Cloud, Fog, and Edge resources in isolation and only explore static metrics for their characterisation, such as data processing speed, load-balancing, and data transmission overheads. Therefore, the fragmented approaches lead to inefficient pipeline deployment and overly complex adaptation processes. With the advent of microservices architectures and containerisation, Big Data pipelines deployments must consider the proper installation sequence, facilitate networking connectivity, and provide optimised tasks configuration with guaranteed security and scalability [12]. For this reason, the deployment of complex Big Data pipelines needs orchestrators capable of exploiting the provisioned Cloud, Fog and Edge computing infrastructure, automating the software deployment on a large scale of heterogeneous and distributed resources and adequately handling the various contexts, status and lifecycle of these resources.

3.6. Adaptation barrier

Big Data pipelines execution across the Computing Continuum commonly depends on provisioning and composition of resources and services across multiple providers. Several interoperability issues prevent efficient resource provisioning across different network domains, such as incompatible transport protocols, data formats, and missing open platform communications support. Within a single control domain, resources and service providers usually offer proprietary interfaces for infrastructure definition and resource provisioning, limiting interoperability and locking users to a single provider. Regrettably, existing Cloud interoperability solutions are mostly limited to heterogeneous networked service interfaces, are not compatible with the Cloud standards, and ignore software component portability across deployment interfaces, critical for Big Data pipelines. Numerous other ad-hoc solutions rely on manual, tedious and error-prone development of mediators to resolve heterogeneity problems [13].

4. GOING BEYOND STATE OF THE ART IN RESEARCH

Overcoming these barriers requires going beyond the current state of the art and answering several core research questions. Table 1summarises the key aspects.

4.1. Discovery

Operations managers can use advanced analytics to take a deep dive into historical process data, identify relationships among process steps, and optimise the factors that prove to have the most significant effect on yield. Nonetheless, existing solutions cannot align executed processes and their associated data pipelines to meet requirements related to compliance and efficiency. To this aim, organisations need a more granular approach to discovering and diagnosing data-intensive processes. Process mining aims to provide a solution.

State of the art While many organisations collect vast troves of event data, they typically use event data only for tracking purposes, not as a basis for improving operations. There are many hurdles to overcome for extracting event data suitable for process mining. These include merging event data distributed over various sources, events that do not point to process cases, event data containing unusual behaviour, and events at different granularity. While many successful process discovery solutions in several application domains exist, they are suitable for processes with no incorporated data perspective [14]. Conformance checking compares a model and an event log for the same process to understand the presence and nature of deviations. Most conformance checking techniques use ad-hoc implementations of traditional searching algorithms in specific domains.

Table 1. State of the art and research challenges

Area	State of the art	Research challenges
Discovery	Process discovery assumes the availability of a well-formed event log. Process discovery with no incorporated data perspective. Process mining cannot accomplish the analytics task in the presence of Big Data.	Techniques to extract well-formed event logs from heterogeneous sources. Discovery techniques to learn the data pipelines' underlying processes. Scalable AI-based Big Data pipeline-driven analytics.
Definition	Big Data analysis solutions lack user-friendly pipeline definition and reuse with pluggable components.	Graphical compositional DSLs for defining Big Data pipelines, provisioning meta-pipelines libraries, and developing extensible and pluggable container templates for pipeline steps supporting different programming environments and computational resources.
Simulation	There are no frameworks for testing and simulating Big Data pipelines before deploying on the Computing Continuum.	Approaches for simulating pipelines and finding the most cost-effective distribution of resources.
provisioning	Current blockchain-based decentralised Cloud computing approaches support stateless pipeline steps with no dependencies and no hardware guarantees for performing individual steps.	Decentralised oracles capable of assessing the hardware properties for executing the pipelines. Improvement of off-chain computing protocols on the blockchain for supporting interdependent and stateful pipeline steps.
Deployment	Current orchestrators are suitable to deploy, scale, and manage applications on single data centres rather than many distributed data centres in the Cloud.	Orchestration mechanisms with auto-scaling features to deploy and execute pipelines on the Computing Continuum and adapt them to data drifts or exogenous events.
Adaptation	Platforms for pipeline deployment lack support for automated deployment with ad-hoc Edge resources and adaptive resource provisioning with infrastructure drift-awareness.	User and provider-centred approaches for improved and adaptive pipeline deployment, utilising resources across various control domains.

However, when process models and event logs are considerable, the existing approaches do not scale efficiently due to their ad-hoc nature and often cannot accomplish the analytical tasks [15].

Research questions for pipeline discovery

- How can we extract Big Data pipeline event logs considering distributed data over various sources, incomplete data (i.e., with events that do not explicitly point to any pipeline instance), data outliers (i.e., unusual behaviour also referred to as noise), and different event granularity levels?
- How can we learn the structure of Big Data pipelines from event logs containing traces with many events, based on process mining and AI techniques?
- How can we visualise pipelines based on a flow-based notation, providing detailed diagnostics about their execution (bottlenecks, critical waiting times)?

4.2. Definition

Ensuring a smooth usage of Cloud resources for Big Data analysis requires developing languages and graphical interfaces that hide technical details of Cloud technologies. A practical solution to solving such problems is DSLs that does not target universality but a specialised set of problems efficiently. On the other side, Big data pipelines involve processing various inputs and outputs between steps and data preparation, which are highly domain-dependent. Their setup and maintenance are complex, time-consuming, and require multiple technologies. The heterogeneity of necessary computing resources provided by the Cloud and Edge for deploying such pipelines leads to ad-hoc processing models usable only on a specific technological stack.

State of the art Many of today's Big Data pipelines orchestration solutions, such as Pachyderm, Apache Airflow, Snakemake, Apache NiFi, NextFlow, Tekton Pipelines and ARGO, lack a simple DSL or graphical interface to define data pipelines for broad application experts [16], who require deep technical knowledge of the Cloud software. The existing solutions also lack knowledge reusability and debugging support directed towards domain experts. In most cases, exposing data pipelines "as-a-service" is impossible out-of-the-box. For example, Apache Oozie and Luigi

support Hadoop-based pipeline management but do not provide a flexible solution implementable in different technologies. Other partial solutions, such as Amazon GreenGrass, Azure IoT Edge and Apache Edgent, focus on supporting data processing on the Edge devices but do not span across Cloud and Edge. Applications such as Apache NiFi and Node-RED provide pipeline design features but do not natively support data tasks on the Cloud and Edge devices. Some efforts address these challenges by using container technologies for data pipelines. In the early days of containers, scientific workflows provided preliminary solutions [17] to build and deploy pipelines using Docker. Other platforms, such as Tekton Pipelines, Argo Workflows, and Kubeflow, have runtime container support for individual pipeline component steps but cannot define dynamic pipeline executions.

Research questions for pipeline definition

- How can we capture distributed data pipeline syntax and semantics using DSLs on the Computing Continuum infrastructure along with a format for data serialisation and deserialisation?
- How can we configure, deploy, and simulate data pipelines and corresponding DSL compilers capturing its definitions using graphical notations?
- How can we use an extensible library of predefined container templates, handling various data formats and types and allowing for injecting business logic and automated processing of input and output types?

4.3. Simulation

Designing and implementing Big Data pipelines requires cost and performance optimisation of business data services over heterogeneous resources. A viable approach is to ensure an a-priori evaluation of the deployment strategy. The application of simulation techniques, based on previously identified resource and infrastructure patterns, allows the assessment of the distributed data services, ensures the selection of cost-optimising deployment strategies and identifies potential bottlenecks (e.g., unexpected rate of data inputs/outputs, inefficient pipeline steps).

State of the art While related problems such as simulating Cloud deployments, Fog and Edge computing systems, Grid computing and others offer solutions with varying levels of maturity, there is still a gap in state of the art in data pipeline simulation. Several scientific workflow simulators on computational Grids and Clouds, such as GroudSim, GloudSim, and DynamicCloudSim, scale up to hundreds of thousands of heterogeneous machines [18], but do not provide execution models for Edge and Fog resources. Furthermore, their performance models focus on virtual machine is application performance, varying parameters related to scalable hardware resources and estimation of instructions per second. Another approach in the state of the art is to simulate workflow execution by predicting the total execution time of each task based on machine learning [19] and or evolutionary programming [20]. None of the workflow prediction approaches applies to Big Data pipelines since they do not address the dynamic nature (differing workloads and Continuum resources during execution) or continuous nature of the pipelines.

Research questions for pipeline simulation

- How can we simulate data pipelines to evaluate and predict the individual step performance?
- How can we design a performance model based on step measurements (e.g., performed in a sandbox) for estimating the aspects of throughput and data transfer of Big Data pipelines defined using a DSL?
- How can we design an execution model to support the identification of bottlenecks through the parameters provided by a Big Data pipeline performance model?

4.4. Provisioning

Fog and Edge computing aim to decongest the network by reducing the data transfers towards data centres and enabling applications with low latency requirements. This decentralisation promises to improve the service offering and remove the vendor lock-in. Off-the-shelf solutions such as OpenStack and Kubernetes are not fit for geo-distributed environments. Their controllers cannot maintain a complete state of the resources

where network latency is much higher than in regular data centres. Moreover, these controllers represent a single point of failure and a performance bottleneck that defeats the purpose of Fog and Edge computing.

State of the art Recent projects attempt to combine volunteer computing (a.k.a. Desktop Grids) with blockchain technologies to manage distributed Cloud infrastructures [21]. Desktop Grids such as Xtremweb and BOINC typically implement a pull-based execution model that supports independent deterministic tasks. This model works well with Internet latency (e.g., BOINC pooled an average of 36.6 PetaFLOPS in March 2020) because no knowledge of the state of resources is required to schedule tasks. However, the control of Desktop Grids remains centralised and prohibits the development of free economic models for rewarding the workers. iExec implemented a decentralised Cloud computing infrastructure based on Xtremweb featuring a marketplace for determining the price of a given execution and a replication and majority voting (needed to protect requesters from malicious workers) on the Ethereum blockchain. This off-chain computing technique, pursued by some projects such as Golem and Gridcoin, shares a pay-per-task approach and implements payment in smart contracts, which removes the need for trusted parties between the requesters and the workers. One challenge of blockchain-based Cloud computing infrastructures is their dynamic nature, which causes extreme performance variability swings, sometimes even after applying lean techniques to analytic pipelines [22].

Research questions for resource provisioning

- How can we design a decentralised marketplace for software appliances (i.e., virtual machines, containers) with pre-installed Big Data frameworks for publishing and monetising proprietary, original cryptographically signed software?
- How can we design a decentralised blockchain-based marketplace for Cloud/Fog/Edge hardware devices (e.g., servers, sensors) with hardware specifications, performance and reputations metrics certified by decentralised oracles accessible through a REST interface?
- How can we create a decentralised execution infrastructure of trusted servers advertised on a marketplace and accepting authenticated client workloads?

4.5. Deployment

The advent of Cloud and Big Data systems and microservices brought forth the needs of environment provisioning and auto-scaling. Hence, the management of applications' lifecycle orchestration became an integrated part of resource managers, i.e., orchestrators. Tools such as Mesos, Yarn, OpenShift, Docker Swarm and Kubernetes enable the deployment of containers and the applications' lifecycle management. Kubernetes is currently the dominant tool for managing containerised applications.

State of the art Unification of Cloud and Edge processes requires evolving the software to deploy, scale, and manage applications in the Cloud, historically designed for a single data centre [23]. However, some are starting to integrate the Edge through proper abstraction mechanisms, resource disaggregation, and challenging network orchestration issues [24]. Ongoing efforts to adapt Kubernetes for the Edge include open-source attempts such as KubeEdge, startups such as Ori, Rafay Systems, and Volterra, and existing Cloud providers and initiatives like Google's Anthos, Microsoft's Azure Arc, and VMware's Tanzu.

Research questions for pipeline deployment

- How can we design an orchestrator capable of providing flexible, scalable and resilient deployment of Big Data pipelines over the Computing Continuum?
- How can we monitor and enforce data pipelines while providing automated elasticity and featuring real-time event detection and decision?
- How can we ensure secure and resilient orchestration of Big Data pipelines with online adaptation?

4.6. Adaptation

The resource heterogeneity in the Computing Continuum limits the possibility for provisioning and orchestrating resources belonging to different providers, hindering the execution of the complex data pipelines across resources residing in other control domains. The major Cloud providers (e.g., Amazon, Google, Microsoft) provide resources through an automated centralised infrastructure deployment. They do not support the utilisation and integration of third-party provisioning and orchestration algorithms that manage resources across multiple network domains.

State of the art Resource provisioning is essential for transparent pipelines execution in the Computing Continuum. Recently, several promising resource management and provisioning approaches in the Computing Continuum emerged, such as heterogeneous pooling of devices with unpredictable utilisation rates [2] and resource allocation based on Petri Nets [25]. These methods focus on resource provisioning by considering a limited set of non-functional parameters and optimising either a single constrained objective or a set of weighted objectives combined in a linear function. From the perspective of infrastructure description and automated deployment, several industrial solutions, such as CloudFormation, Terraform, and Heroku, address important issues, but support infrastructure description and automated deployment on ad-hoc Edge resources. Furthermore, they provide no adaptive pipeline resource provisioning support with infrastructure drift awareness.

Research questions for pipeline adaptation

- How can we realise smart data-aware provisioning of resources and services across the Computing Continuum with adaptation to infrastructure drifts?
- How can we design an engine for intelligent on-the-fly configuration of heterogeneous resources in the Computing Continuum with improved interoperability?
- How can we monitor dynamic resource provisioning with a time-critical assessment of SLA violations?

5. SUMMARY AND OUTLOOK

In this article, we argued for the need for an ecosystem to support the complete lifecycle of Big Data pipeline processing to release the potential of Dark Data. The ecosystem must enable their discovery, definition, model-based analysis and optimisation, simulation, deployment, adaptive runtime monitoring on a decentralised heterogeneous Computing Continuum infrastructure. We discussed the main concepts and state of the art for these phases, along with key considerations. Finally, we suggested several research problems to serve as a research plan for the future.

ACKNOWLEDGMENT

This work received partial funding from the European Commission's Horizon 2020 DataCloud project (Grant number 101016835) and the NFR BigDataMine project (Grant number 309691).

REFERENCES

1. R. Buyya et al., "A manifesto for future generation cloud computing: Research directions for the next decade," *ACM Computing Surveys*, vol. 51, no. 5, 2019.
2. A. Yousefpour et al., "All one needs to know about fog computing and related edge computing paradigms: A complete survey," *Journal of Systems Architecture*, vol. 98, pp. 289-330, 2019.
3. D. Balouek-Thomert et al., "Towards a computing continuum: Enabling edge-to-cloud integration for data-driven workflows". *The International Journal of High Performance Computing Applications*, vol. 33, no. 6, 2019.
4. M. Barika et al., "Orchestrating Big Data Analysis Workflows in the Cloud: Research Challenges, Survey, and Future Directions," *ACM Computing Surveys*, vol. 52, no. 5, 2019.
5. A. Kobusińska et al., "Emerging trends, issues and challenges in Internet of Things, Big Data and cloud computing," *Future Generation Computer Systems*, vol. 87, pp. 416-419, 2018.
6. P. Castro et al., "The rise of serverless computing," *Communications of ACM*, vol. 62, no. 12, pp. 44–54, 2019.
7. Y. Simmhan et al., "Cloud-based software platform for big data analytics in smart grids," *Computing in Science & Engineering*, vol. 15, no. 4, pp. 38-47, 2013.

8. B. Plalea and I. Kouper, "The centrality of data: data lifecycle and data pipelines," *Data Analytics for Intelligent Transportation Systems*, pp. 91-111, 2017.
9. M. Matskin et al., "A Survey of Big Data Pipeline Orchestration Tools from the Perspective of the DataCloud Project," *XXIII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2021)*, pp. 63-78, 2021.
10. C.-H. Hong and B. Varghese, "Resource Management in Fog/Edge Computing: A Survey on Architectures, Infrastructure, and Algorithms," *ACM Computing Surveys*, vol. 52, no. 5, 2019.
11. A. Shakarami et al., "Resource provisioning in edge/fog computing: A Comprehensive and Systematic Review," *Journal of Systems Architecture*, vol. 122, 2022.
12. S. Svorobej et al., "Orchestration from the Cloud to the Edge," *The Cloud-to-Thing Continuum*, 2020.
13. D. Weerasiri et al., "A Taxonomy and Survey of Cloud Resource Orchestration Techniques," *ACM Computing Surveys*, vol. 50, no. 2, 2018.
14. A. Augusto et al., "Automated Discovery of Process Models from Event Logs: Review and Benchmark," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 686–705, 2019.
15. M. de Leoni and A. Marrella, "Aligning Real Process Executions and Prescriptive Process Models through Automated Planning," *Expert Systems with Applications*. vol. 82, pp. 162–183, 2017.
16. R. Ranjan, et al., "Orchestrating BigData Analysis Workflows," *IEEE Cloud Computing*, vol. 4, no. 3, pp. 20–28, 2017.
17. P. Kacsuk, J. Kovács, and Z. Farkas, "The Flowbster Cloud-Oriented Workflow System to Process Large Scientific Data Sets," *Journal of Grid Computing*, vol. 16, pp. 55–83, 2018.
18. Bambrik, I., "A Survey on Cloud Computing Simulation and Modeling," *SN Computer Science*, vol. 1, 2020.
19. T. Pham, J. J. Durillo and T. Fahringer, "Predicting Workflow Task Execution Time in the Cloud Using A Two-Stage Machine Learning Approach," *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 256–268, 2020.
20. F. Nadeem and T. Fahringer, "Optimising execution time predictions of scientific workflow applications in the Grid through evolutionary programming," *Future Generation Computer Systems*, vol. 29, no. 4, pp. 926–935, 2013.
21. R. B. Uriarte and R. DeNicola, "Blockchain-Based Decentralised Cloud/Fog Solutions: Challenges, Opportunities, and Standards," *IEEE Communications Standards Magazine*, vol. 2, no. 3, pp. 22–28, 2018.
22. A. Uta and H. Obaseki, "A Performance Study of Big Data Workloads in Cloud Datacenters with Network Variability," *9th ACM/SPEC International Conference on Performance Engineering (ICPE '18)*, pp. 113–118, 2018.
23. M. Dias de Assuno, A. da Silva Veith, and R. Buyya, "Distributed data stream processing and edge computing," *Journal of Network and Computer Applications*. vol. 103, pp. 1–17, 2018.
24. X. Chen et al., "iDiSC: A New Approach to IoT-Data-Intensive Service Components Deployment in Edge-Cloud-Hybrid System," *IEEE Access*, vol. 7, pp. 59172–59184, 2019.
25. L. Ni et al., "Resource Allocation Strategy in Fog Computing Based on Priced Timed Petri Nets," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1216–1228, 2017.

Dumitru Roman is a senior research scientist at SINTEF AS, Norway.

Radu Prodan is a professor in distributed systems at the University of Klagenfurt, Austria.

Nikolay Nikolov is a research scientist at SINTEF AS, Norway.

Ahmet Soylu is a professor in computer science at Oslo Metropolitan University, Norway.

Mihhail Matskin is a professor in software engineering at KTH, Sweden.

Andrea Marrella is an assistant professor in computer science at Sapienza University of Rome, Italy.

Dragi Kimovski is an assistant professor at the University of Klagenfurt, Austria.

Brian Elvesæter is a research scientist at SINTEF AS, Norway.

Anthony Simonet-Boulogne is research scientist at iExec, France.

Giannis Ledakis is a software engineer at UBITECH, Greece.

Hui Song is a senior research scientist at SINTEF AS, Norway.

Francesco Leotta is an assistant professor in computer science at Sapienza University of Rome, Italy.

Evgeny Kharlamov is a research scientist at Bosch

Center for Artificial Intelligence, Germany.