INTEND: Intent-Based Data Operation in the Computing Continuum

Donatella Firmani[1], Francesco Leotta[1], Jerin George Mathew[1], Jacopo Rossi[1,*], Lorenzo Balzotti[1], Hui Song[2], Dumitru Roman[2], Rustem Dautov[2], Erik Johannes Husom[2], Sagar Sen[2], Vilija Balionyte-Merle[2], Andrea Morichetta[3], Schahram Dustdar[3], Thijs Metsch[4], Valerio Frascolla[4], Ahmed Khalid[5], Giada Landi[6], Juan Brenes[6], Ioan Toma[7], Róbert Szabó[8], Christian Schaefer[8], Cosmin Udroiu[9], Alexandre Ulisses[10], Verena Pietsch[11], Sigmund Akselsen[12], Arne Munch-Ellingsen[12], Irena Pavlova[13], Hong-Gee Kim[14], Changsoo Kim[15], Bob Allen[15], Sunwoo Kim[16] and Eberechukwu Paulson[16]

[1]*Sapienza University*
[2]*SINTEF Digital*
[3]*TU Wien*
[4]*Intel Deutschland GMBH*
[5]*Dell Technologies*
[6]*Nextworks*
[7]*Onlim GmbH*
[8]*Ericsson*
[9]*CS Group Romania*
[10]*MOG Technologies*
[11]*Fill GmbH*
[12]*Telenor ASA*
[13]*GATE Institute Sofia University*
[14]*Seoul National University*
[15]*AiM Future*
[16]*Hanyang University*

Abstract

The European Commission (EC) Digital Decade strategy to gain by 2030 autonomy in the digital economy requires more and more data to be processed in the Cloud-Edge-IoT computing continuum, instead of only in the central cloud. This requires advanced automation and intelligence of the continuum. At the same time, recent breakthroughs in Artificial Intelligence (AI) research have shown unprecedented results in handling creative tasks. Such human-like intelligence will eventually disrupt how people use the cloud and continuum. The European Union (EU) -funded project INTEND aims at bringing such human-like intelligence into the cognitive continuum, to achieve the novel concept of intent-based data operation. The project will deliver 11 novel software tools, which integrate into an INTEND toolbox. The outputs pave the way of migrating EU's data industry from cloud to the continuum, and implement EC's strategy of human-centric AI in the domain of data processing and computing continuum.

RPE@CAiSE'24: Research Projects Exhibition at the International Conference on Advanced Information Systems Engineering, June 03–07, 2024, Limassol, Cyprus
jacopo.rossi@uniroma1.it (J. Rossi)

1. Introduction

Processing large amounts of data creates immeasurable value for Artificial Intelligence (AI) and Machine Learning (ML)-based applications [1] but also generates huge cost, with much of the cost concentrated to the few public cloud providers [2]. The Digital decade policy program of the European Commission (EC), aiming at reaching an EU digital autonomy by 2030, pushes to exploit resources at the edge of the telecommunication network for data processing, to reduce the cost and the dependency to central cloud providers [3]. More and more EU organizations should have their data pipelines running in the computing continuum instead of the central cloud, but this requires advanced automation and intelligence of the continuum [4, 5], since doing so at the edge is much more complicated than in the cloud.

The EC has recently funded 9 projects under the call HORIZON-CL4-2023-DATA-01-04 for applying AI to achieve cognitive computing continuum, with promising outcomes towards high-level automation [6]. However, at the same time, many people worry that AI lacks sufficient intelligence to do creative work, e.g., in the context of the continuum, to use the heterogeneous and unconventional devices in unpredicted ways, to handle resources at different places from different providers in a strategic way, and to understand what the human stakeholders really need. Despite a decade of effort on improving automation, central cloud vendors still offer human service representatives to their large customers.

In the meantime, recent breakthroughs in AI research have shown unprecedented human-like intelligence in the direction of generative AI [7], neural-symbolic AI [8], and deep reinforcement learning [9]. Such human-like intelligence has the potential to eventually disrupt how people use the cloud-edge computing continuum. By exploiting these latest AI breakthroughs, it is possible to bring the next-level human-like intelligence into the *cognitive* computing continuum, allowing the latter to adapt, think and talk like humans: continually learn how to adapting data pipelines to heterogeneous and unconventional resources in an effective way, make strategic decisions at different places in the continuum like the human brain thinks in a multi-objective way, and chat with human stakeholders in natural language to understand their intents and explain what was done.

Outline. This paper presents the INTEND research project[1] towards cognitive computing continuum with advanced intelligence to achieve the novel *intent-based data operation*. Section 2 describes the participants, the main objectives and the relevance of INTEND for the CAiSE community. Section 3 describes the results obtained so far and the expected results in the upcoming years. Finally Section 4 reports conclusive remarks.

2. Summary of the project

INTEND "Intent-based data operation in the computing continuum" is a Horizon Europe collaborative research project funded by the EU. It started in January 2024 and is expected to last 36 months.

Partners. SINTEF AS (Norway) is the project leader. The project involves both academic and industry partners. Academic partners are Sapienza Università di Roma (Italy), Technische

[1] https://intendproject.eu

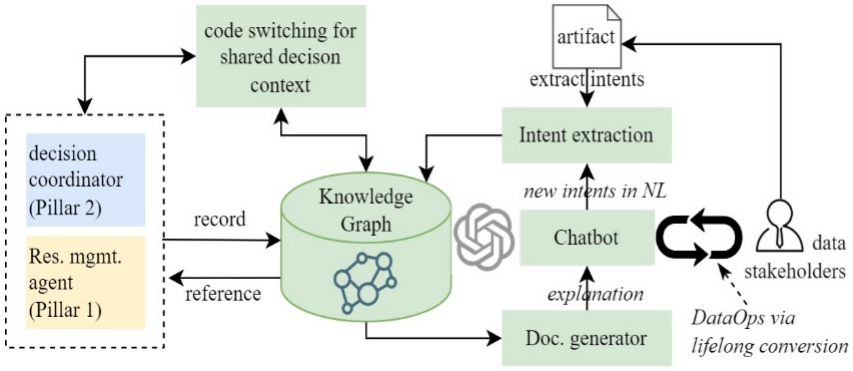


Figure 1: Components of the intent-based human-AI interaction

Universitaet Wien (Austria), GATE Institute Sofia University (Bulgary), Seoul National University (Korea) and Hanyang University (Korea). The industry partners provide coverage of the complete supply chain of compute continuum, from chips (Intel Deutschland GMBH, AiM Future), servers (Dell Technologies), cloud infrastructure (Ericsson), telecom (Telenor ASA), software (CS-Group), conversational AI (Onlim GmbH) to consultance (NEXTWORKS).

Objectives and expected outputs. INTEND's research will lead to 11 novel software tools for the cognitive continuum, with a focus on intelligent operation of data pipelines, organized in three main research pillars, each corresponding to a specific objective. Pillar 3 is shown in detail in Figure 1. The key idea behind Pillar 3 tools is to use knowledge graph to provide a machine-readable representation of intents, and a novel code-switching approach to connect the common knowledge graph with multiple decision makers. Between the knowledge graph and stakeholders, the natural language interface extracts stakeholders' intents from direct dialogs or existing artifacts, and explains to the stakeholders how and why the AI models made certain adaptations based on the existing intents. Tools in Pillar 2 will handle the distributed and dynamic nature of computing continuum by decentralized and federated decision coordination, to compare and combine the adaptation decisions made by different AI models, into globally optimal adaptation. Finally, tools in Pillar 1 research will handle the hardware diversity by continual learning, i.e., to learn autonomously what is the best way to use the resources in the continuum, based on observing how the data pipelines perform on the current and similar resources. We now discuss pillars/objectives and related software tools in detail.

- **Objective/Pillar 1.** Achieve intelligent management of computing, storage, network, and neural processing resources based on continual learning, to better exploit diverse and unconventional hardware on edge and cloud. Expected output consists of 4 tools: **inStore**, intelligent data placement and storage management in cloud and edge, the Kubernetes[2] based **inOrch** [10] tool for hardware-aware intelligent orchestration of data processing services, the **inNet** for AI-powered intent-based networking, and the **inNeural** tool for intelligent adaptation of concurrent multimodal workloads on AI accelerators.
- **Objective/Pillar 2.** Enable strategic and federated decision making covering multiple

[2]https://kubernetes.io/

AIs from different perspectives and places across the continuum, to achieve end-to-end data security and sustainability. Expected output consists of 3 tools: **inCoord**, a decentralized and federated decision coordinator for global adaptation, the **inSustain** tool to comprehensively measure the sustainability of data pipelines, and **inSec** to assess end-to-end data security on multi-provider security and identity management mechanisms.

- **Objective/Pillar 3.** Achieve human-centric data operation, via a novel natural language interface for data stakeholders to efficiently express their data operation intents, and understand and trust the AI-made decisions. Expected output consists of 4 tools: the **inGraph** tool, to manage the intent knowledge graph for cognitive data operation, **inSwitch** to perform code-switching between intent knowledge graph and decision-making contexts, **inGen** to extract intents and generate decision-making explanations, and **inChat**, a chat-based interface for continual interaction.

Based on the prototype toolbox, we will create an open software and hardware platform with open APIs and marketplace to support the integration of new types of devices, new AI models and new types of data operation intents. We will validate INTEND platform in five vertical use cases.

- **Video streaming.** The global market of video stream is estimated at USD 375.1 billion in 2021, with CAGR of 18.45% (Precedence), and emits almost 1% of global GHG emissions (UpToUs). MOG Technology expects to apply the INTEND techniques to their video streaming products and therefore lower 30% of the overall emission by spending less resources for content ingestion, transportation, and storage in the cloud.
- **Machine data.** The market size for predictive maintenance is at USD 4.2 billion in 2021, with a CAGR of 30.6% till 2026 (Markets&Markets). Manufacturing is the most representative application in this market. Fill GmbH will lower the cost and time spent on operating predictive maintenance pipelines customized for their custom factories.
- **5G data infrastructure.** The market of edge data centres is estimated at USD 7.2 billion in 2021 and 21.4 CAGR until 2026 (Markets&Markets). Telenor will enter this market based on their position in telecommunication infrastructures.
- **Urban dataspace.** Data space is a key emerging concept, quickly gaining economical and political attention. GATE Institute is building the first data space in Bulgaria, which is a flagship data space for smart cities. INTEND will be used to support the 16 data owners already singed in the Urban Data Space, with many more interested to join.
- **Robotic AI.** The global market of robotics systems will be USD 225 billion by 2025, with a significant part of it on development and operation of data pipelines on the robots. Hanyang University's experiment is expected to reduce 30% effort on operating AI pipelines on robots, allowing researchers and companies to focus on technical innovation.

All use cases demand advanced data operation from different perspectives, involving stakeholders like data engineers, AI researchers and non-IT consultants.

Relevance for CAiSE. The INTEND project directly aligns with the following topics in the CAiSE call for paper.

- "Cloud- and edge-based Information Systems engineering". The INTEND platform will provide a ready-to-use solution for realizing the novel concept of intent-based data

operation in the continuum: data stakeholders chat with the toolbox about how they intend their data pipelines to perform in the continuum.
- "Context-aware, autonomous and adaptive Information Systems". Understanding the intents, the platform will keep adapting the data pipelines in the continuum.
- "Privacy, security, trust, and safety management". The platform will explain to the stakeholders what it is planning to do or what cannot be achieved, in order that the stakeholders can trust the AI and collaborate with AI for safer data operation.
- "Sustainability and social responsibility management". Finally, INTEND will significantly improve the cost-efficiency and sustainability data processing, and lower the effort and skill barrier for stakeholders in handling data in the continuum.

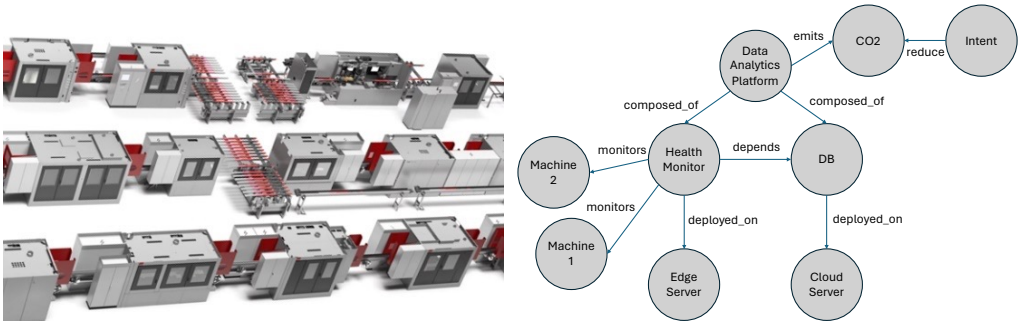
3. Current Project Status

The project is still at its inception, that is at the requirement definition phase, focusing on designing starting features of tools, identifying techniques, AI models, datasets and interaction between tools. The initial demo will be released in June 2024 with state-of-the-practice data operation, showing demands of intent-based data operation on sample scenarios and running "on the paper".

Techniques. Modern data processing in the continuum is based on the virtualization of resources, to enable the seamless movement of data and workloads. FogFlow [11] uses Docker containers as the building block of data flow from IoT to cloud. DataCloud [12] builds tools for the discovery, development, and optimization of big data pipelines on top of container technology. Recent approaches introduce Function as a Service (FaaS) to further shield the resource complexity underneath data pipelines [13], and handle many scattered edge resources together as a fleet [14]. We are also planning to use containers as the backbone. With containers to simplify deployment, the orchestration of data pipelines will focus on the placement of data and workloads [15] using AI technologies.

AI models. Various ML approaches have been used in continuum management [16], but mostly for single and isolated purposes, such as predicting future loads and optimizing service locations. Despite some attempts toward unified approaches based on specific AI techniques in the cloud era [17], what is still missing is the capability to create joint and coordinated learning and reasoning, according to a previous roadmap [18]. Moreover, existing ML models incomprehensibly learn their own knowledge from data, leaving human stakeholders unable to interact, understand, or trust what the AI is doing. Instead of applying ML for a one-shot optimization of data pipelines, our plan is to investigate continual reinforcement learning that keeps adapting the data pipelines and improving their adaptation effects at the same time. We will also introduce stakeholders' intents as the objective of ML-based adaptation.

Demo scenario. The sample scenario is inspired by the *Machine data* use case, lead by Fill GmbH, an internationally leading special machinery and plant engineering company. Fill's CYBERNETICS ANALYZE is the data analytics platform that Fill provides to their customers, i.e., factories manufacturing different goods, as cylinder heads, battery trays, or the ski production line in Figure 2a. The platform is used for processing, storing, and sharing various machine data to monitor the health and efficiency of the machine and processes for production and



(a) (b)

Figure 2: FILL ski production line (partial) and example knowledge graph.

maintenance and to continuously increase the machine's efficiency as well as the quality of parts manufactured. In such scenario, our demo will improve how data pipelines are operated in their data platform based on intents expressed by salespeople (who have more business than IT background) for data pipelines, e.g. what new data analytics components are needed to meet the customers' requirements, e.g. what is the required data quality and latency and what is the expected energy consumption. Afterwards, the demo will generate Docker commands to automatically adjust the pipeline orchestrations and the resource usage according to these intents. Figure 2b shows a simplified Knowledge Graph data-structure in this scenario, representing user intents and resources as entities (i.e., nodes of the graph) and *triples* (i.e., edges) such as <*Data Analytics Platform, composed of, Health Monitor*>. The Knowledge Graph can record linked knowledge of stakeholder's intents, such as quality of service, affordability, security and privacy concerns, and sustainability ethics, etc., as well as the semantic knowledge that the stakeholders know about the pipeline.

4. Conclusions

We presented the INTEND research project towards cognitive computing continuum with advanced human-like intelligence, to achieve the novel concept of intent-based data operation. We described the objectives of the project, the expected results and discussed the main concepts and envisaged technologies to achieve these objectives. The project is still at its inception and the first demo, planned for June 2024, will provide an illustrative prototype based on Generative AI to explain the key properties of our approach and show the potential directions. INTEND is a EU-funded research and innovation project with 16 partners, including universities, research institutes and companies from 10 European countries and South Korea. The integrated INTEND platform will be applied and demonstrated in 5 use cases from the domains of video streaming, digital manufacturing, telecommunication, smart cities and robotics systems.

Acknowledgments

This work is partly funded by the HORIZON Research and Innovation Action 101135576 INTEND "Intent-based data operation in the computing continuum". Jerin George Mathew is financed by the Italian National PhD Program in AI. Jacopo Rossi is supported by Thales Alenia Space and Regione Lazio, through the fellowships 35757-22066DP000000041-A0627S0031 *Advanced Software Based on Cloud Computing and Machine Learning for Space Systems*.

References

[1] C. Bernardos, et al., European vision for the 6G network ecosystem, White Paper of the 5G-IA Association (2021). doi:DOI:10.13140/RG.2.2.19993.95849.

[2] H. K. Hallingby, S. Fletcher, V. Frascolla, G. Anastasius, I. Mesogiti, F. Patzysz, 5G Ecosystems, White Paper of the 5G-IA Association (2021). doi:doi.org/10.5281/zenodo.5094340.

[3] E. Kartsakli et al., AI-Powered Edge Computing Evolution for Beyond 5G Communication Networks, in: 2023 Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit), 2023, pp. 478–483. doi:10.1109/EuCNC/6GSummit58263.2023.10188371.

[4] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, J. Cao, Edge computing with artificial intelligence: A machine learning perspective, ACM Computing Surveys 55 (2023) 1–35.

[5] Y. Wu, Cloud-edge orchestration for the internet of things: Architecture and ai-powered data processing, IEEE Internet of Things Journal 8 (2020) 12792–12805.

[6] V. Frascolla, et al., Intelligent Edge-Embedded Technologies for Digitising Industry, River Publishing, 2022.

[7] M. Xu, et Al., Unleashing the power of edge-cloud generative AI in mobile networks: A survey of AIGC services, IEEE Communications Surveys & Tutorials (2024) 1–1. doi:10.1109/COMST.2024.3353265.

[8] A. Sheth, K. Roy, M. Gaur, Neurosymbolic artificial intelligence (why, what, and how), IEEE Intelligent Systems 38 (2023) 56–62. doi:10.1109/MIS.2023.3268724.

[9] J. Huang, e. a. Yang, Reinforcement Learning based resource management for 6G-enabled mIoT with hypergraph interference model, IEEE Transactions on Communications (2024) 1–1. doi:10.1109/TCOMM.2024.3372892.

[10] T. Metsch, M. Viktorsson, A. Hoban, M. Vitali, R. Iyer, E. Elmroth, Intent-driven orchestration: Enforcing service level objectives for cloud native deployments, SN Computer Science 4 (2023). doi:10.1007/s42979-023-01698-0.

[11] J. Sendorek, T. Szydlo, M. Windak, R. Brzoza-Woch, Fogflow-computation organization for heterogeneous fog computing environments, in: Computational Science–ICCS 2019: 19th International Conference, Faro, Portugal, June 12–14, 2019, Proceedings, Part III 19, Springer, 2019, pp. 634–647.

[12] D. Roman, R. Prodan, N. Nikolov, A. Soylu, M. Matskin, A. Marrella, D. Kimovski, B. Elvesæter, A. Simonet-Boulogne, G. Ledakis, et al., Big data pipelines on the computing continuum: tapping the dark data, Computer 55 (2022) 74–84.

[13] B. Oliveira, N. Ferry, H. Song, R. Dautov, A. Barišić, A. R. Da Rocha, Function-as-a-service for the cloud-to-thing continuum: a systematic mapping study, in: 8th International Conference on Internet of Things, Big Data and Security-IoTBDS, 2023, pp. 82–93.
[14] H. Song, R. Dautov, N. Ferry, A. Solberg, F. Fleurey, Model-based fleet deployment in the iot–edge–cloud continuum, Software and Systems Modeling 21 (2022) 1931–1956.
[15] F. A. Salaht, F. Desprez, A. Lebre, An overview of service placement problem in fog and edge computing, ACM Computing Surveys (CSUR) 53 (2020) 1–35.
[16] Z. Zhong, M. Xu, M. A. Rodriguez, C. Xu, R. Buyya, Machine learning-based orchestration of containers: A taxonomy and future directions, ACM Computing Surveys (CSUR) 54 (2022) 1–35.
[17] C.-Z. Xu, J. Rao, X. Bu, Url: A unified reinforcement learning approach for autonomic cloud management, Journal of Parallel and Distributed Computing 72 (2012) 95–105.
[18] A. Morichetta, V. C. Pujol, S. Dustdar, A roadmap on learning and reasoning for distributed computing continuum ecosystems, in: 2021 IEEE International Conference on Edge Computing (EDGE), IEEE, 2021, pp. 25–31.