Towards Bringing Human-Like Intelligence in the Computing Continuum: the INTEND Project

Lorenzo Balzotti[1], Donatella Firmani[1], Francesco Leotta[2], Jerin George Mathew[2] and Jacopo Rossi[2]

[1]*Department of Statistical Science, Sapienza University of Rome, p.le Aldo Moro 5, 00185 Rome, Italy.*

[2]*Department of Computer, Control, and Management Engineering "Antonio Ruberti", Sapienza University of Rome, Via Ariosto 25, 00185, Rome, Italy.*

Abstract

The growing complexity of the Cloud-Edge-IoT continuum requires advanced automation and intelligence to efficiently manage distributed resources and optimize data processing. The Horizon Europe project INTEND aims to introduce human-like intelligence into this continuum, enabling intent-based data operations through AI-driven decision making and natural language interactions. By integrating 11 novel software tools into the INTEND toolbox, the project facilitates the transition from a cloud-focused approach to a more distributed data management model within the computing continuum. This aligns with the European Commission's Digital Decade strategy to achieve digital economy autonomy by 2030, which requires an increasing amount of data to be processed within the Cloud-Edge-IoT computing continuum rather than relying solely on the central cloud.

1. Introduction

The European Commission's (EC) Digital Decade policy program aims to achieve EU digital autonomy by 2030, promoting the use of telecommunication network edge resources for data processing. This strategy aims to lower costs and decrease the dependence on centralized cloud providers [1]. However, processing data at the edge presents significantly greater challenges than in traditional cloud environments, requiring advanced automation and intelligence within the computing continuum [2, 3]. Despite these challenges, an increasing number of EU organizations are expected to transition their data pipelines from central cloud infrastructures to a more distributed continuum. Large-scale data processing generates immense value for Artificial Intelligence (AI) and Machine Learning (ML)-driven applications [4], yet much of the associated cost is currently concentrated among a few public cloud providers [5].

To advance the development of AI in the cognitive computing continuum, the European Commission (EC) has recently funded nine projects under the HORIZON-CL4-2023-DATA-01-04, aiming to leverage AI for achieving higher levels of automation [6]. While these initiatives show promising progress, concerns persist regarding AI's ability to perform creative tasks and take human-like decisions. In the context of the computing continuum, AI still struggles with utilizing heterogeneous and unconventional devices in unforeseen ways, strategically

SEBD 2025: 33rd Symposium on Advanced Database Systems, June 16-19, 2025, Ischia, Italy

lorenzo.balzotti@uniroma1.it (L. Balzotti); donatella.firmani@uniroma1.it (D. Firmani); leotta@diag.uniroma1.it (F. Leotta); mathew@diag.uniroma1.it (J. G. Mathew); j.rossi@diag.uniroma1.it (J. Rossi)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-03-13

managing distributed resources across different providers, and fully grasping the needs of human stakeholders. Even after a decade of advancements in automation, central cloud vendors still offer human service representatives to their large customers.

In the meantime, recent advancements in AI research have demonstrated unprecedented human-like intelligence in areas such as generative AI [7], neural-symbolic AI [8], and deep reinforcement learning [9]. These breakthroughs have the potential to fundamentally reshape how the cloud-edge computing continuum is utilized. By integrating these cutting-edge AI developments, the *cognitive* computing continuum could evolve to exhibit next-level human-like intelligence—enabling it to adapt, reason, and communicate similarly to humans. This would allow it to continuously refine data pipelines by effectively leveraging heterogeneous and unconventional resources, make strategic decisions across different layers of the continuum in a manner similar to multi-objective human reasoning, and engage with human stakeholders in natural language to both comprehend their needs and provide explanations of its actions. Achieving a seamless interplay between AI and human stakeholders is crucial for the adoption of cognitive computing systems.

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. Section 3 describes the results obtained so far and the expected results in the upcoming years. Finally Section 4 reports conclusive remarks.

This paper is primarily based on [10, 11], summarizing their key insights and discussing their applications.

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 Universitaet Wien (Austria), GATE Institute Sofia University (Bulgary) and Seoul National University (Korea). The academic partners complement each other in key research directions on computing continuum, data science, human interaction, and AI. The Distributed System Group in Technische Universitaet Wien is a leading group in the domain of Edge AI. The Sapienza and the SINTEF teams have strong expertise in the domain of knowledge representation and Chatbots, while the Seoul National University team is dedicated research group on neuro-symbolic AI for collaborated decision making.

Industry partners are Intel Deutschland GMBH (Germany), DELL (Ireland), Ericsson (Hungary and Sweden), Telenor ASA (Norway), CS-Group (Romania) and FILL GMBH (Austria). Small enterprise partners include NEXTWORKS (Italy), Onlim GMBH (Austria), MOG Technologies (Portugal) and ad AiM Future, Inc (Korea). The industry partners provide coverage of the

[1] https://intendproject.eu

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. The overview of the project is depicted in Figure 1.

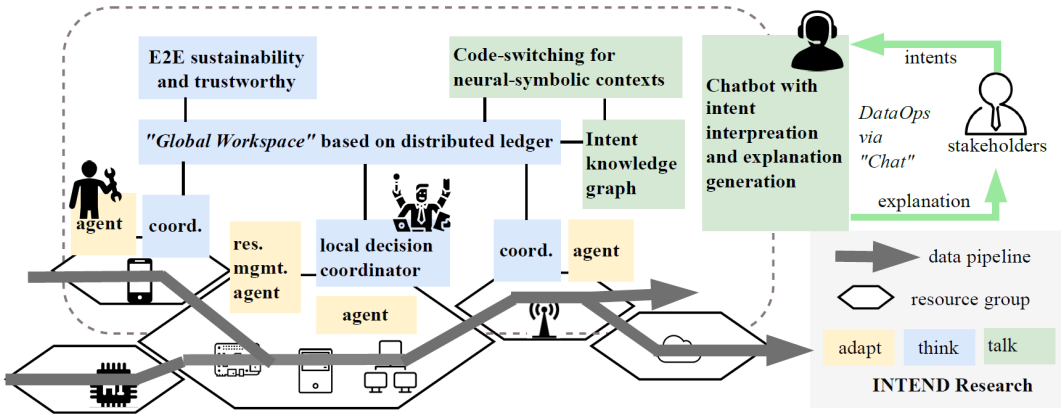

Figure 1: Intend pipeline.

Pillar 1, Pillar 2 and Pillar 3 are highlighted in yellow, blue and green, respectively.

Research Pillar 1 The cognitive continuum relies on continual ML for managing and dynamically adapting resources in data processing pipelines. We propose developing continual learning agents for four resource types: computation, storage, network, and neural processing. These ML agents operate within resource groups, logical subsets of the continuum (e.g., Kubernetes clusters), where they monitor, reallocate, and reconfigure resources while orchestrating data pipelines. Each resource group hosts one or more data pipelines, with competing data pipes requiring dynamic resource allocation. The agents adapt based on various runtime contexts, including data orchestration, resource availability, data status, performance, energy consumption, stakeholder requirements, and coordination with other groups. When conditions change, agents adjust resource allocations to find an optimal balance using knowledge from historical data (supervised learning) and real-time behavior (continual reinforcement learning). Research focuses on defining adaptation actions, selecting effective ML algorithms, and establishing continual learning pipelines, varying across resource types.

Research Pillar 2 As a large, distributed, and dynamically federated system, the continuum relies on multiple resource management agents for decision-making. These agents operate across resource groups with distinct objectives (e.g., computation orchestration, data placement) and must coordinate for strategic global adaptation. Inspired by cognitive architecture theories, we propose a federated decision coordinator, structured as a distributed system with

local and regional coordinators. Local coordinators oversee resource groups, resolve conflicts among adaptation decisions, and report to a virtual "global workspace," where decisions are evaluated, revised, and re-broadcasted for further optimization. Regional coordinators, covering multiple resource groups, provide a broader perspective and long-term strategic planning. Using techniques from long-short-term memory and federated learning, we will design efficient decision-sharing mechanisms and explore swarm learning with lightweight distributed ledger technology for a reliable, traceable global workspace. This cognitive-inspired approach enables decentralized, competitive decision-making, ensuring optimal resource management without requiring a complete view of the continuum.

Research Pillar 3 AI models from Research Pillars 1 and 2 will use stakeholder intents as a reference for making and coordinating adaptation decisions. Breakthroughs in Large Language Models (LLMs), such as ChatGPT, demonstrate AI's ability to understand human intents via natural language. Bringing this capability to the cognitive continuum enables direct interaction with stakeholders. The challenge lies in bridging general-purpose LLMs with domain-specific AI decision-makers. We will address this by integrating knowledge graphs for machine-readable intent representation and employing a code-switching approach to connect multiple decision-makers. A natural language interface will extract stakeholder intents from dialogues and artifacts while explaining AI-driven adaptations. Data engineers, as key stakeholders, embed intents in artifacts like source code, documentation, and configurations. We will explore LLMs to extract these intents and translate them into actionable decisions, ensuring AI models can explain their reasoning in alignment with Explainable AI principles. The details of Pillar 3 are shown in Figure 2.

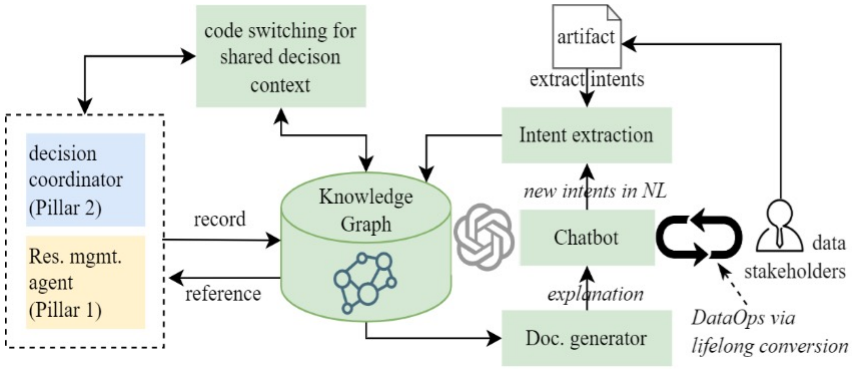

Figure 2: Components of the intent-based human-AI interaction (Pillar 3).

We now discuss pillars related software tools in detail.

- **Tools Pillar 1.** Expected output consists of 4 tools: **inStore**, intelligent data placement and storage management in cloud and edge, the Kubernetes[2] based **inOrch** [12] tool for hardware-aware intelligent orchestration of data processing services, the **inNet** for

[2]https://kubernetes.io/

AI-powered intent-based networking, and the **inNeural** tool for intelligent adaptation of concurrent multimodal workloads on AI accelerators.

- **Tools Pillar 2.** 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.
- **Tools Pillar 3.** 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 four vertical use cases.

- **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.
- **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 aims to integrate the INTEND techniques into its video streaming products, expecting to reduce overall emissions by 30% through optimized resource utilization in content ingestion, transportation, and cloud storage.
- **Urban dataspace.** The concept of data spaces is rapidly emerging, attracting significant economic and political interest. GATE Institute is developing Bulgaria's first data space, serving as a flagship initiative for smart cities. INTEND will support the 16 data owners already enrolled in the Urban Data Space, with many more expressing interest in joining.
- **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.

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

3. Current Project Status

The project is currently in the requirement definition phase, working on designing the initial features of tools, identifying techniques, AI models, datasets and interaction between tools. The initial demos will focus on 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 relies on the virtualization of resources to facilitate the smooth movement of data and workloads. Docker containers are the fundamental building blocks for data flow from IoT to the cloud, as demonstrated by FogFlow [13]. DataCloud [14] builds tools for discovering, developing, and optimizing big data pipelines using container technology. Recent advancements have introduced Function as a Service (FaaS) to simplify the complexity of resources beneath data pipelines [15], enabling the management of distributed edge resources as a unified fleet [16]. Our approach also leverages containers to streamline deployment, focusing on the orchestration of data pipelines with AI technologies to manage the placement of data and workloads [17].

AI models. While various ML approaches have been applied to continuum management [18], they are typically focused on isolated tasks such as predicting future loads or optimizing service locations. Although there have been efforts toward unified AI-based approaches in the cloud era [19], a key missing element is the ability to create coordinated learning and reasoning, as outlined in a previous roadmap [20]. Additionally, existing ML models often generate knowledge in ways that are opaque to human stakeholders, making it difficult to interact with, understand, or trust their actions. Instead of using ML for one-time optimization of data pipelines, our approach aims to explore continual reinforcement learning, which continuously adapts data pipelines while improving their performance. We will also incorporate stakeholders' intents as the guiding objectives for ML-based adaptation.

Demo scenarios. *(Machine data)* The first scenario considered is derived from the *Machine data* use case, lead by Fill GmbH, a globally recognized leader in special machinery and plant engineering. Fill offers CYBERNETICS ANALYZE, a data analytics platform designed for its customers—factories engaged in manufacturing various products, such as battery trays, cylinder heads, and even ski production lines. The platform's primary function is to collect, store, and analyze machine data, enabling monitoring of operational efficiency and machine health. This, in turn, supports production, maintenance, and continuous improvements in both machine efficiency and the quality of manufactured components.

In this context, our demo enhances the management of data pipelines within Fill's data platform by leveraging intent-based operations. These intents, articulated by salespeople with a stronger business rather than IT background, outline specific needs for data pipelines. Examples include defining new data analytics components to meet customer requirements, determining expected energy consumption, and specifying necessary data quality and latency. Following this, the demo automatically generates Docker commands to adjust pipeline orchestrations and resource usage dynamically, aligning with the specified intents.

(Urban dataspace) The second demo scenario, is based on the *Urban dataspace* use case, lead by GATE Institute. In this scenario a data consumer collects air quality data from two providers, one through an API and the other from a database. This data is then processed through a preprocessing pipeline, which cleans and structures the information before storing it in a DVC bucket. Once the data is ready, two distinct machine learning models are trained: one predicts air pollution over time using historical data, while the other predicts spatial air pollution levels based on real-time sensor inputs. The trained models are made available in an App Store, where data consumers can download and deploy them within their own infrastructure. A consumer selects one of these models and executes it within their data connector, enabling real-time

predictions. When new data is fed into the system, it is automatically preprocessed, and the model generates predictions, which are then returned to the user.

The demo will showcase an end-to-end data pipeline for air quality prediction. In the current implementation, data engineers play a key role in managing and fine-tuning models. However, within the project, our objective is to replace manual intervention with AI-driven automation, enabling intelligent pipeline orchestration, automated model adaptation, and self-optimizing resource allocation taking into account the user's requirements and objectives. This shift will enhance scalability, reduce operational overhead, and improve efficiency in managing data pipelines across the urban dataspace.

4. Conclusions

We introduced the INTEND research project, aiming to advance the cognitive computing continuum with human-like intelligence, and to establish the novel concept of intent-based data operation. The project's objectives, expected outcomes, and the key concepts and technologies required to achieve them were outlined. Currently in its early stages, the first demo will feature a prototype based on Generative AI to illustrate the core features of our approach and explore potential directions. INTEND is an EU-funded research 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 across five use cases in video streaming, digital manufacturing, telecommunications, smart cities, and robotics.

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.*

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] 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.

[2] 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.

[3] 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.
[4] 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.
[5] 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.
[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] D. Firmani, F. Leotta, J. G. Mathew, J. Rossi, L. Balzotti, H. Song, D. Roman, R. Dautov, E. J. Husom, S. Sen, et al., Intend: Intent-based data operation in the computing continuum, in: CEUR Workshop Proceedings, volume 3692, CEUR-WS, 2024, pp. 43–50.
[11] R. Dautov, H. Song, D. Roman, E. J. Husom, S. Sen, V. Balionyte-Merle, D. Firmani, F. Leotta, J. G. Mathew, J. Rossi, et al., Intend: Human-like intelligence for intent-based data operations in the cognitive computing continuum, in: (RuleML+RR 2024), volume 3816, 2024.
[12] 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.
[13] 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.
[14] 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.
[15] 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.
[16] 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.
[17] 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.
[18] 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.

[19] 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.
[20] 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.