

A new framework for integrated distributed data management, processing, and advanced analytics supporting Digital Twin systems: a Proof of Concept*

Nicola Masi¹, Chiara Maria Capizzi¹, Francesco Patanè¹, Giuseppe Veneziano¹, Davide Profeta¹, Sergio Comella¹, Giada Sajeva¹, Lorenzo Gori¹, Susanna Bonura^{1°}

¹ *Engineering Ingegneria Informatica S.p.A., Piazzale dell'Agricoltura 24, 00144 Roma.*

Abstract

The digital revolution has driven exponential growth in ICT technologies, enabling resource scalability and the emergence of digital monopolies. In this context, the European AVANT project (dAta and infrastructural serVices for the digital coNTinuum) aims to foster a decentralized, sustainable, and secure digital ecosystem by leveraging technologies such as edge computing and Digital Twins (DTs). Its goal is to develop a new AI-powered software infrastructure for the resilient and distributed management of complex systems, enhancing interoperability, technological sovereignty, and energy efficiency across Europe. AVANT is structured around a set of modular platforms, including intelligent resource orchestration, distributed data management, advanced analytics, tools for DT creation and cybersecurity, and domain-specific applications in industry, cultural heritage, smart cities, energy, and healthcare. The document also presents proof of concept for a smart building DT, using IoT sensors, predictive modeling, explainable AI techniques, and a federated infrastructure for dynamic machine learning training.

Keywords

Digital twin, data management, data analytics, data engineering, big data

1. Introduction

Over the last decades, we witnessed transformations enabled by the digital revolution. Thanks to ICT, resources were scalable and abundant, indeed reducing margins for many firms and creating several de facto monopolies.

We believe that edge computing and decentralization in a trusted digital ecosystem will play a crucial role, and that the EU society can gain from this new innovative trend. Companies like Google, Facebook, and Amazon created an abundance: they were able to scale globally, winning the race when the market was at its beginning. Their effective management of massive amounts of data made them monopolists, and they created innovative ecosystems, gaining control of sensitive customer data.

Our scope is to establish our dominion in decentralized computing through the AVANT project, especially in decentralization and edge computing, where data produced at the edge is organized, and made useful for Digital Twins (DTs), which will become actionable representations of real-world entities. This new generation of technologies, through capillary and highly efficient networks, will enable public administrations, firms, and people to benefit from secure, flexible, and dynamic networked value chains. The AVANT project will seek the following challenges:

1. The high failure rate (around 85%) of big data/AI projects is due to the lack of appropriate tools for improving models, evaluating issues, and improving forecast quality.[1][2]

* *ITADATA2025: The 4th Italian Conference on Big Data and Data Science, September 9-11, 2025, Turin, Italy.*

° These authors contributed equally.

Nicola.masi@eng.it (N. Masi); chiaramaria.capizzi@eng.it (C.M. Capizzi); francesco.patanè@eng.it (F. Patanè); giuseppe.veneziano@eng.it (G. Veneziano); davide.profeta@eng.it (D. Profeta); sergio.comella@eng.it (S. Comella); giada.sajeva@eng.it (G. Sajeva); lorenzo.gori@eng.it (L. Gori); susanna.bonura@eng.it (S. Bonura).



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2. Europe is missing major original platforms, technology (e.g., iOS, or Android), social (Facebook, TikTok) or commercial (Amazon).[3]

3. Effective digital transformation is affected by the non-interoperability of IoT platforms, where much data remains unused and best practices for service delivery, such as DevOps, are not adopted; and finally, the Total cost of ownership (TCO) is high.

4. Make the computing infrastructure federated, dynamic, simple, elastic, and resilient. [4]

5. Sustainable computing; while data center technologies are progressing, their energy consumption due to software (SW) not using resources adequately remains a problem. [5]

6. Exploit the edge data: the solutions should exploit data spaces based on resources distributed geographically and a new SW orchestration architecture, flexible enough to adapt to the context. Such architecture will make AI and big data analytics at the edge an abundant resource. The information flow should be developed in open cloud environments.

7. Protect privacy and assure cyber security and resilience, tailoring services to user needs and requirements. [6]

8. Deliver a data platform for the continuum, from containers to web assembly and new data processing technologies to extract information in decentralized data contexts.

9. Use at the best the power of new communication networks that will democratize low latency services for the mass market.

Section 2 will present the AVANT project, its framework, and the platforms used to address the challenges related to Digital Twins. Using the building use case as a reference, Section 3 will describe the features of its Digital Twin and illustrate how the proof-of-concept (PoC) demonstrates the entire process from the collection of real-world data to the generation of predictive models. This PoC also introduces some of the very first approaches to interoperability, with a focus on the management of integrable and extensible connectors (for example, by adopting IDSA-certified connectors for interaction with data spaces). Computational workloads are dynamically and elastically delegated to federated clusters.

This opens scenarios in which resource selection and deployment configurations can be managed by integrated systems capable of considering the characteristics of the environment and the processes involved, thereby enabling optimizations such as energy consumption reduction.

2. The AVANT Project

AVANT (dAta and infrastructural serVices for the digitAl coNTinuum)¹ is the project started in 2024 by ENGINEERING within the *Important Project of Common European Interest on Next Generation Cloud Infrastructure and Services (IPCEI-CIS)*², which saw a total of 19 implemented projects with an overall budget of 3.5 billion euro.

Differently to current state-of-the-art solutions, the AVANT approach faces the digital world through a new perspective, which takes into account the latest challenges, as requested also by the European Commission. It foresees the following intertwined issues:

- Seamless users experience different resource/service providers in the continuum, with flexible, automated resource discovery, and elastic, cognitive resource, and data management. It involves the creation of a Distributed Data Ecosystem for distributed and orchestrated analytics, offering user-oriented, low/no-code tools to move analytics where data is produced, at the edge, where IoT deployments are not creating value due to interoperability barriers; and the development and deployment of innovative systems and services based on DTs.

¹ <https://www.eng.it/it/insights/stories/research-projects/ipcei-cis-avant>

² https://competition-policy.ec.europa.eu/state-aid/ipcei/approved-ipceis/cloud_en

- Increasing the energy efficiency through the use of DTs.
- Increasing the cybersecurity in DTs, together with data protection data processing, consent and privacy management and digital identification.
- Supporting the interoperability at resource and data levels, through the creation of a uniform interoperability model across the continuum, and within specific domains as energy and healthcare.
- Making the continuum appear as a seamless technological platform.

In this context, our project creates cutting-edge solutions for the management of complex systems through the use of Digital Twins, actively involving stakeholders and open-source communities, aiming to strengthen Europe's technological leadership. The goal is to facilitate the transition towards distributed and cognitive Clouds, capable of adapting to the ever-changing needs of users. AVANT also aims to improve coordination in the fragmented European market, reducing the time to implement advanced resources for data analysis and the development of DTs. The main objectives are:

1. Digital Twin as a Service: making DTs scalable and accessible in various sectors.
2. Cognitive Cloud: creating flexible ICT infrastructures, improving sustainability and efficiency

AVANT fits into this context as a key element in the creation of an advanced digital ecosystem, capable of ensuring interoperability, security and technological sovereignty at European level, committing to realize cutting-edge solutions for the future of cloud computing in Europe.

In particular, it aims at delivering a novel AI-powered software infrastructure supporting and boosting the development, deployment, and operations of trustworthy and cyber-resilient, decentralized, and data-intensive systems in the digital continuum as an evolution of the cloud-edge continuum. On top of such an infrastructure, AVANT will deliver six domain-specific specialized frameworks and platforms aiming at designing, developing, deploying, and operationalizing cloud-edge applications and services based on and exploiting the potential of DTs at several different levels of complexity and granularity, involving (and digitally representing and twinning) multiple and heterogeneous types of entities, both physical (e.g. objects, devices, machines, systems, facilities, people) and logical (e.g. processes, organizations). The following picture shows the logical architecture of the whole AVANT outcome, with a more detailed representation of the infrastructural building blocks.

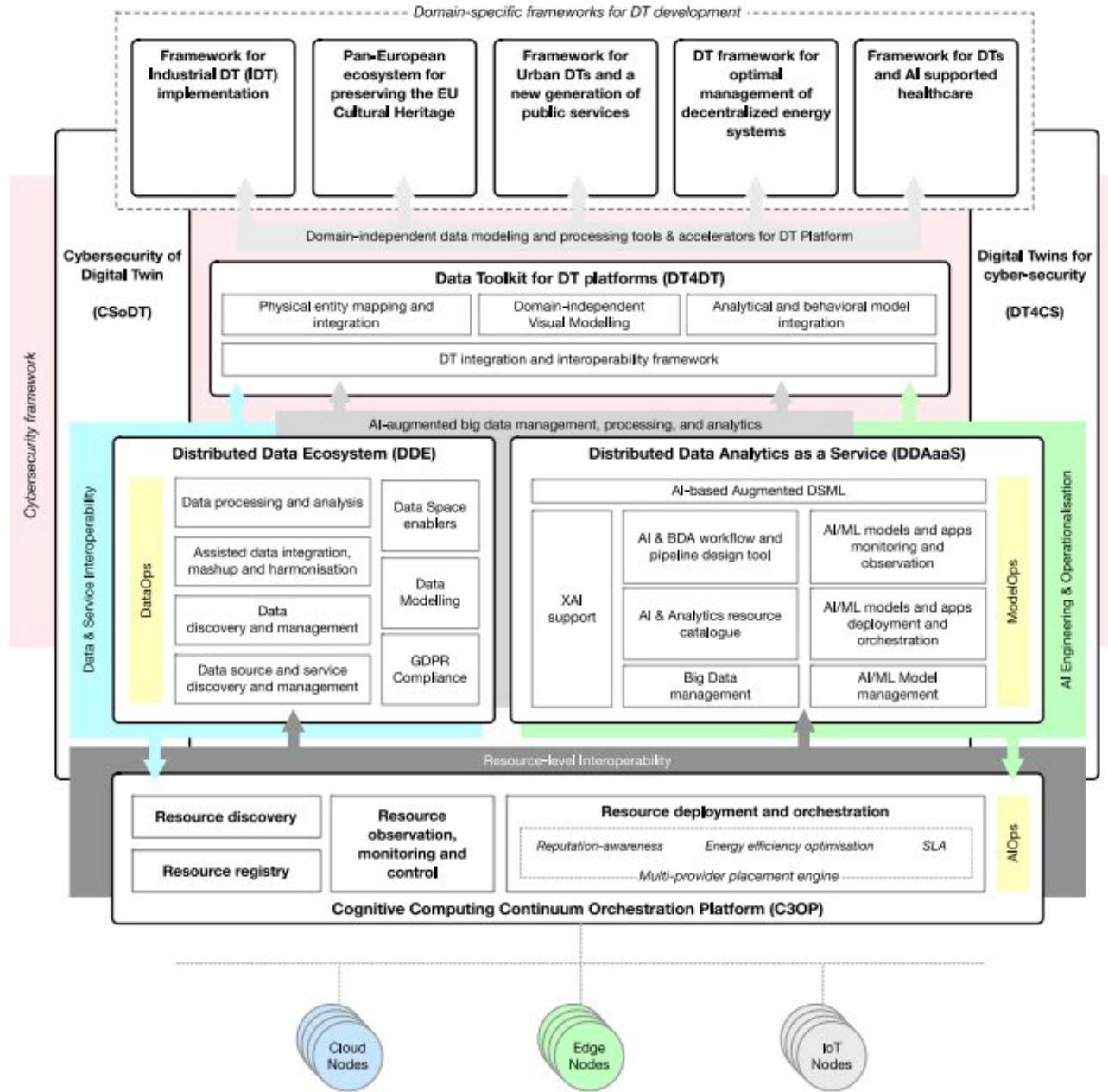


Figure 1: AVANT project architecture.

Specifically, the following main integrable and interoperable, yet independent infrastructural, domain-independent frameworks and platforms will be delivered:

- Cognitive Computing Continuum Orchestration Platform (C3OP): it supports the intelligent and optimal usage of computational cloud-edge resources in the digital continuum, by providing AIOps (Artificial Intelligence for IT Operations) capabilities for (1) effective discovery in heterogeneous, multi-domain, dynamic, mobile environments; (2) monitoring, observing, and controlling those distributed resources; (3) orchestrating, deploying, and executing them according to energy-efficiency, Service Level Agreement (SLA), and reputation considerations and optimization. Such capabilities aim at optimizing the usage and the potential of the elastic continuum infrastructure, towards trusted and energy efficient deployments.[7]

- Distributed Data Ecosystem (DDE): it is a data management framework and platform supporting and providing data-level and service-level interoperability and capabilities for the DataOps³ practice. This includes data sources and services discovery, modelling, integration, mashup, harmonization, transformation, and processing, while managing data quality and GDPR compliance as well as enabling and supporting data spaces and sharing.
- Distributed Data Analytics as a Service (DDAaaS): it is a data processing platform providing engineering and operationalization of AI-based data analytics capabilities by leveraging on and supporting Machine Learning Operations (MLOps)⁴ practice. This includes DSML (Data Science and Machine Learning) engineering tools and execution modules for big data management, catalogue of AI and analytics resources and building blocks, workflow and pipeline design, management of ML models, orchestration, deployment, monitoring and observation of ML models and AI and analytics applications, while supporting explainability of AI models.[8]
- Data Toolkit for DT platforms (DT4DT): leveraging on DDE and DDAaaS, it provides the underlying data-oriented (management, transformation & homogenization, and processing) capabilities and domain-independent technological tools common to and exploitable by all domain-specific DT platforms, in terms of integration and mapping of physical and logical entities to digital models, general components and tools to model and implement generic and common physical entities' dynamics and behaviors, and a basic general-purpose 3D visual modelling tool that will be the basis for more sophisticated and specific DT visual modelling tools.
- Cybersecurity of Digital Twin (CSoDT): an integrated set of AI-based cybersecurity measures for the DTs service delivery platforms that would allow trusted, reliable, and resilient operations of the DTs, supporting cyber risk assessment and management, enabling dynamic identification of the critical vulnerabilities and threats, enhancing cyber-attack detection capabilities, while supporting effective response, across the digital continuum and different organizations, providing near real-time situational awareness on the cyber-resilience of DTs.
- Digital Twins for cyber-security (DT4CS): a set of novel cybersecurity measures and tools that use and exploit the DT as an innovative means to increase the cybersecurity of the cyber-physical system distributed across the continuum. Our approach will make the DT itself a unique digital collaborative playground to protect the cyberphysical twin against cyber-attacks, supporting and improving cyber risk assessment and management, threat and attack detection, modelling and testing response and recovery, training, and awareness.

³ DataOps is a set of practices, processes and technologies that combines an integrated and process-oriented perspective on data with automation and methods from agile software engineering to improve quality, speed, and collaboration and promote a culture of continuous improvement. (Ereth, Julian (2018). "DataOps-Towards a Definition" (PDF). Proceedings of LWDA 2018: 109.)

⁴ MLOps is a set of practices that combines Machine Learning, DevOps, and Data Engineering to deploy and maintain machine learning models in production reliably and efficiently.

3. The AVANT framework: Proof-of-Concept on smart-buildings Digital Twins systems

The use case focuses on the development of a DT for a building. Within the PoC, two rooms were simulated, each equipped with IoT sensors (temperature, humidity, people counting, etc.) and actuators (air conditioning, sprinkler system, smart blinds).

The goal is to monitor indoor environmental conditions through a dashboard, using real-time data streams from the physical world. The system is designed to predict the conditions one hour ahead, providing interpretable explanations for the forecast and to suggest optimal actions on the actuators, in order to maintain a comfortable and controlled environment. Actuators can also be operated directly from the dashboard.

Within the PoC, predictive models were trained, also leveraging services based on eXplainable Artificial Intelligence (XAI).

Moreover, collaboration with ENEA, TIM, and Arsys enabled the federation of new clusters, which were then used during the model training phase.

Before going into the details of the implemented functionalities, the adopted standards, and the used metamodels, it is useful to provide an overview of the models that define a Digital Twin in the AVANT project:

- Event Models: describe how to collect data from the real world.
- Data Models: represent the physical structure of the Digital Twin.
- Computation Models: define the behavior of the Digital Twin.
- Visualization Models: allow the visualization of the obtained results.
- Command Models: enable actions to be taken in the real world.

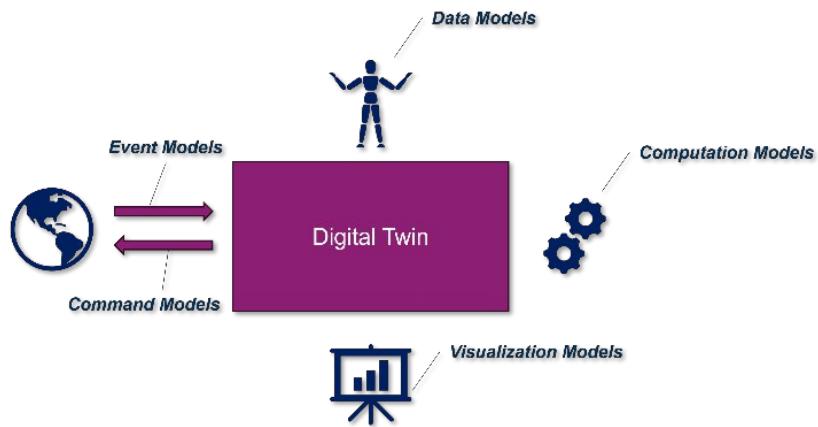


Figure 2: AVANT DT Models.

The PoC implements a process which involves the use of several tools: real world, digital twin accelerators, distributed data service and analytics, and multi-provider interop layer. The PoC is fully described in the picture below:

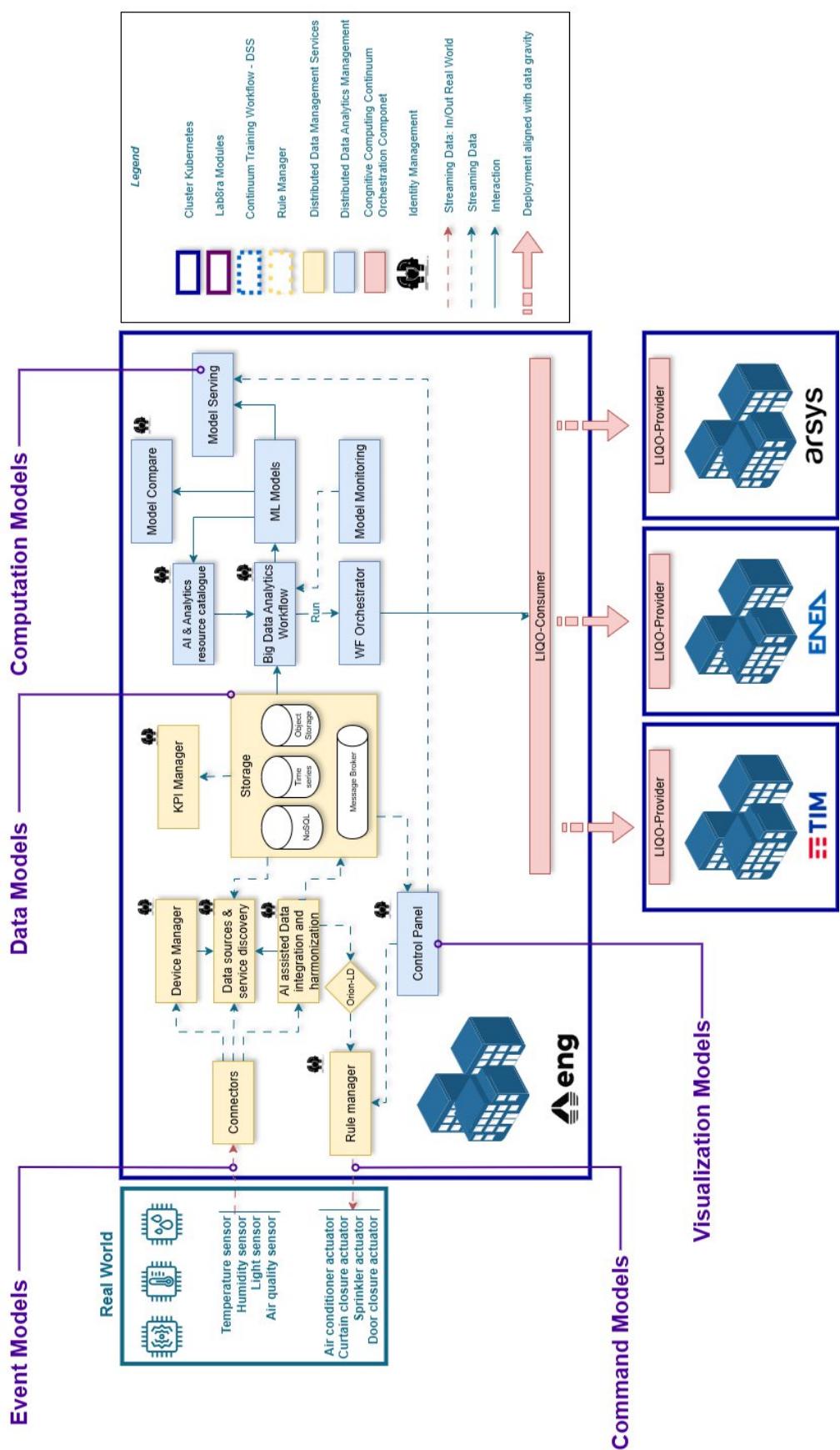


Figure 3: AVANT PoC architecture.

Real-time data collection begins with environmental sensors measuring temperature, humidity, light levels, and air quality. These sensors transmit data through dedicated connectors to the platform, maintaining continuous data flow and enabling real-time system control.

The collected data is processed through the Data Mashup Editor, which converts the building structure along with its sensors and actuators into DTDL (Digital Twins Definition Language) format. This standardization step transforms the physical building representation into a unified digital model that can be consistently processed across the system.

Once the DT is standardized in DTDL format, the data flows to multiple components in parallel. The Rule Manager monitors data conditions and creates control rules for real-world actions (for example, if temperature exceeds 30°C, the system automatically closes blinds, or if humidity levels drop below threshold, sprinklers may be activated). Simultaneously, the KPI (key performance indicators) Manager processes the same data to monitor key performance indicators and system efficiency metrics.

In parallel with these processing activities, data persisted across multiple storage targets to ensure availability for other systems within the architecture. This distributed storage approach creates a comprehensive historical dataset for machine learning model training, system optimization, and cross-system integration.

Once the data is stored in a distributed manner, the DDAaaS system can take over the tasks required for model training. These tasks include the definition of algorithms, the prototyping of data processing workflows through no-code tools, the execution of training algorithms using federated clusters (in this case, ENEA's cluster), the use of XAI services, the comparison of metrics from trained models, and, more generally, the management of datasets and machine learning models involved. Specifically, dedicated services were created within workflows designed for training ML models aimed at predicting environmental conditions. As the data is gradually harmonized by the DDE, it becomes available for use in these training processes.

The workflow aims for classification of PMV (Predicted Mean Vote), an index representing thermal comfort in indoor environments. It starts with a problem-specific labeling service that computes the PMV class based on environmental and physiological variables. The target label is then shifted forward by one hour to support time-ahead prediction. A stratified subsampling step is applied to balance the class distribution, followed by training of both a predictive model and its corresponding XAI component using anchor-based explanations.

Following the same architectural pattern, another model is the training of a classifier for IAQ (Indoor Air Quality), an index reflecting perceived air quality based on environmental conditions. It includes a dedicated labeling service tailored to IAQ classification, followed by the same target shifting logic, subsampling for class balance, and training of both a predictive model and its corresponding XAI component using anchor-based explanations. The use of modular BDA services ensures reusability and consistency across applications. [9]

After their implementation, the results are sent to a backend service that feeds the control *panel* user interface. It is possible to select, directly from the graphical interface, the namespace in which to activate the services dedicated to the training of Machine Learning models, choosing among those exposed by federated providers through *Liqo*⁵. This capability is particularly relevant for the Continuum, as it allows for the dynamic allocation of computational components to the most suitable infrastructure. TIM, ENEA and ARSYS clusters could be selected as most appropriate on the basis of their hardware resources availability.

⁵ Liqo is an open-source project that enables Kubernetes clusters to seamlessly connect and collaborate, creating a larger, virtualized Kubernetes network. It allows clusters to share resources and services across different environments (on-premise, cloud, edge) while maintaining individual cluster control. <https://docs.liqo.io/en/v1.0.0/#>

As data is processed and stored, they are trained by ML models, as previously described in DDAaaS, in order to give predictions and anchor-based explanations. Among them there is the *Big Data Analysis Application* (BDA), which was developed using the workflow editor, and enables the composition of modular BDA services for predictive model training. The workflow aims for classification of PMV (Predicted Mean Vote), an index representing thermal comfort in indoor environments. It starts with a problem-specific labeling service that computes the PMV class based on environmental and physiological variables. The target label is then shifted forward by one hour to support time-ahead prediction. A stratified subsampling step is applied to balance the class distribution, followed by training of both a predictive model and its corresponding XAI component using anchor-based explanations.

Later, a *control panel* starts the training in predictive models, specifically to estimate air and environmental quality one hour ahead, including explanations for each prediction.

The *panel* main goal is to suggest actuator actions that can help prevent unsuitable environmental conditions, and to control them to enact actions in the system (e.g., activating sprinklers in case of fire).

The image shows a Decision Support System (DSS) interface designed for monitoring and managing the environmental and safety conditions of a building in real time. The system focuses on two rooms equipped with IoT sensors and actuators.

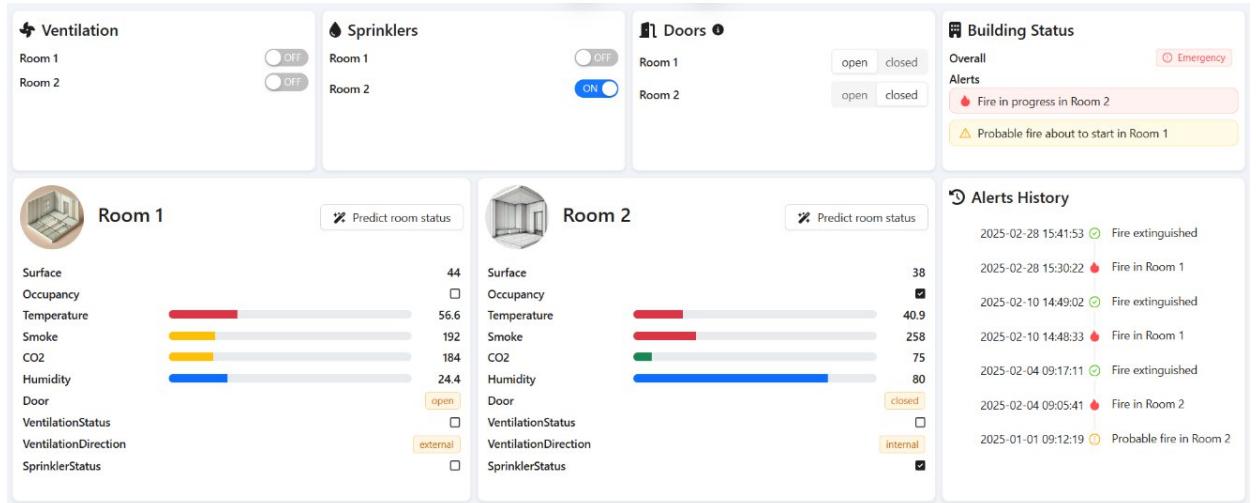


Figure 4: Decision Support System.

At the top of the dashboard, users can manually control ventilation, sprinklers, and doors for each room. For example, the sprinkler in Room 2 is currently active, and the doors are open in Room 1 and closed in Room 2. The "Building Status" section summarizes the overall situation. An emergency alert is currently active, with a fire in progress in Room 2 and a potential fire predicted in Room 1. Each room section displays information such as surface area, occupancy level, temperature, smoke, CO₂, and humidity. These values are presented both numerically and through colored bars for quick interpretation. Additional indicators show the state of the door, ventilation, and sprinkler systems. A button allows the user to trigger a prediction of the room's future condition, likely based on machine learning models. Overall, the DSS allows users to monitor current conditions, predict upcoming risks, receive explanations for predictions, and take immediate action through actuators. It supports both proactive and reactive decision-making for improved safety and energy efficiency.

This PoC presents the first version of the AVANT architecture: starting from the acceptance of real data, their harmonization and training, and obtaining several predictive models of future behaviors through the use of DT components, DDE and DDAaaS services, and interaction with external Liqo-providers.

Thanks to the use of MLOps, AI & Analytics, and orchestrator tools, the use case of the *building with two rooms* has been faced off, and produced concrete solutions to environmental challenges.

Conclusions and next steps

The project development has led to the definition of use cases across various cross-domain areas, which have been instrumental in guiding the evolution of supporting components. Discussions held throughout the period have helped steer the work of the different macro-categories, while keeping security considerations constantly aligned. On the infrastructure side, access to hardware resources has been provided, already enabling some continuum scenarios. In parallel, an intelligent orchestration system has been developed to optimize deployment based on factors such as resource availability, process policies, and data management requirements. Within DataOps, existing components already provide initial support for harmonization and standardization, while developments are ongoing for Data Discovery and Data Management. Data exchanges have been enabled through Data Space connectors, along with tools for managing data models useful for defining Digital Twins and linking them with real-world data. Regarding DDAaaS, the first components have been delivered for the creation and execution of analytical workflows through a no-code designer and a service catalog that is already available and can be further expanded. Initial modules for Explainable AI and Data Quality tools have also been developed, and integration tests with the infrastructural components have been carried out to run specific workflows in a continuum environment.

Ongoing activities include the development of Data Governance tools and other strategic components not yet completed, which will be progressively integrated to enhance the platform's overall capabilities. The next steps will focus on continuing the development, extending functionalities based on the use cases, and introducing new modules and frameworks. Particular attention will be given to adopting standardized models for Digital Twin descriptions and real-world data integration, strengthening orchestration strategies for optimal resource allocation, and enriching the analytical capabilities with advanced Explainable AI, Data Quality, and Data Governance features—always in alignment with security requirements and the proposed infrastructure architecture.

In summary, the Digital Twin system described represents an advanced and integrated solution for smart building management. The AVANT project is currently in progress, in its second year of implementation, and this represents the first draft of the Proof of Concept. As the AVANT development framework is further defined, new versions will be delivered to enhance, consolidate, and extend the functionalities described. Future work will focus on enabling the management of DTDL across different standards and application contexts, introducing standard models for defining communication with the physical world, and adopting models for describing Digital Twin behavior. Additional objectives include integrating Data Governance tools to support the entire data lifecycle and embedding the intelligent orchestrator to optimize hardware resource allocation.

Acknowledgements

This work is framed within the Project "AVANT" – Project no. IPCEI-CL_0000005 - Application protocol no. 108421 of 14/05/2024 - CUP B89J24002920005 - Grant decree no. 1322 of August 8, 2024 - financed by the European Union – NextGenerationEU (IPCEI Funding).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

- [1] G. Reggio, E. Astesiano, Big-data/analytics projects failure: A literature review, 2020, pp. 246–255. sopra tutto
- [2] F. Mielli, N. Bulanda, Digital transformation: Why projects fail, potential best practices and successful initiatives, in: Conference Proceedings - IEEE-IAS/PCA Cement Industry Technical Conference, volume 2019-April, 2019.
- [3] Brühl, Volker, Big Tech, the Platform Economy and the European Digital Markets (May 2, 2023). Center for Financial Studies Working Paper No. 711, 2023.
- [4] Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: distributed ML for on-device intelligence.
- [5] Jessica Commins and Kristina Irion. Towards Planet Proof Computing: Law and Policy of Data Centre Sustainability in the European Union. TechReg. 2025. Vol. 2025:1-36.
- [6] Shokri, R. and Shmatikov, V. Privacy-Preserving Deep Learning. Conference on Computers and Communications Security (CCS), pp. 1310-1321, 2015. 1
- [7] Y. Wu, “Cloud-Edge Orchestration for the Internet-of-Things: Architecture and AI-Powered Data Processing,” IEEE IoT Journal, 2021. 2
- [8] Q. Yao, M. Wang, Y. Chen, W. Dai, H. Yi-Qi, L. Yu-Feng, T. Wei-Wei, Y. Qiang, Y. Yang, Taking human out of learning apps: A survey on automated ML, arXiv preprint arXiv:1810.13306 3
- [9] I. Stepin, et al., “A survey of contrastive and counterfactual explanation generation methods for explainable AI,” IEEE Access, 2021. 4
- [10] Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: distributed ML for on-device intelligence. arXiv preprint arXiv:1610.02527.
- [11] X. Ji, et al., “Visual exploration of neural document embedding in information retrieval: semantics and feature selection,” IEEE Trans. Vis. and Comp. Graph., 2019.
- [12] M. Babar, et al., “A secured data management scheme for smart societies in industrial internet of things environment,” IEEE Access, 2018.
- [13] Brethenoux, Erick. "Gartner Glossary". Gartner. Retrieved 16 December 2020.