NEARDATA

Funded by the European Union

HORIZON EUROPE FRAMEWORK PROGRAMME

NEARDATA

(grant agreement No 101092644)

Extreme Near-Data Processing Platform

D6.3 Final dissemination, exploitation, and standardization report

Due date of deliverable: 30-11-2025
Actual submission date: 30-11-2025

Start date of project: 01-01-2023 Duration: 36 months

Summary of the document

Document Type	Report
Dissemination level	Sensitive
State	v1.0
Number of pages	63
WP/Task related to this document	WP6 / T6.1, T6.2 ,T6.3
WP/Task responsible	SCO
Leader	Andre Miguel (SCO)
Technical Manager	Andre Miguel (SCO)
Quality Manager	Raúl Gracia (DELL), Pedro García (URV), Albert Cañadilla (URV), Aaron Call (BSC)
Author(s)	Max Kirchner (NCT), Sebastian Bodenstedt (NCT), Raúl Gracia (DELL), Alan Cueva (DELL), Hossam Elghamry (DELL), Sean Ahearne (DELL), Ger Hallissey (DELL), Pedro García (URV), Albert Cañadilla (URV), Vanesa Ruana (URV), Usama Benabdelkrim (URV), Paolo Ribeca (UKHSA), Maciej Malawski (SANO)
Partner(s) Contributing	SCO, URV, DELL, BSC, TUD, KIO, NCT, SANO, UKHSA
Document ID	NEARDATA_D6.3_Sen.pdf
Abstract	Thorough description of all dissemination activities and specifically our impact on global standards. The standardization report will contain all main contributions of the project's partners to relevant standards both in terms of input documents submitted for consideration and software components integrated to the reference software implementations.
Keywords	communication, dissemination, exploitation, engagement, near data processing, OMICs, genomics, transcriptomics, metabolomics, surgery.

History of changes

Version	Date	Author	Summary of changes
0.1	10-09-2025	Andre Miguel (SCO)	First draft.
0.2	14-10-2025	Andre Miguel (SCO), Raúl Gracia (DELL), Max Kirchner (NCT), Aaron Call (BSC), André Martin (TUD), Paolo Ribeca (UKHSA), Maciej Malawski (SANO), Pedro García (URV), Albert Cañadilla (URV), Vanesa Ruana (URV), Usama Benabdelkrim (URV)	Contributions.
0.3	17-11-2025	Raúl Gracia (DELL), Pedro Garcia (URV), Albert Cañadilla (URV), Aaron Call (BSC)	Internal review of the deliverable
1.0	30-11-2025	Andre Miguel (SCO)	Final version.

Table of Contents

List of Abbreviations and Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
BSC	Barcelona Supercomputing Center
CAS	Configuration and Attestation Service
CC	Creative Commons
CI/CD	Continuous Integration / Continuous Delivery
CIDR	Classless Inter-Domain Routing
CLI	Command Line Interface
CNCF	Cloud Native Computing Foundation
CNN	Convolutional Neural Netowork
COG	Cloud-Optimized GeoTIFF
COPC	Cloud-Optimized Point Cloud
CORE	Computing Research and Education Association of Australia
CPU	Central Processing Unit
CSV	Comma-separated values
DB	DataBase
DL	Deep Learning
DOI	Digital Object Identifier
EBS	Elastic Block Store
ECS	Elastic Container Service
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
FaaS	Function as a Service
FL	Federated Learning
FPGA	Field-Programmable Gate Array
GPU	Graphics Processing Unit
HDFS	Hadoop Distributed File System
HPC	High-Performance Computing
I/O	Input/Output
IDE	Integrated Developing Environment
INSDC	International Nucleotide Sequence Database Collaboration

K8S	Kubernetes
KPI	Key Performance Indicator
LiDAR	Light Detection And Ranging
LLD	Last-Level Defence
LTS	Long-Term Storage
ML	Machine Learning
MPI	Message-Passing Interface
MSI	Mass Spectometry Imaging
NCBI	National Center for Biotechnology Information
NCT	National Center for Tumor Diseases
NIH	National Institutes of Heatlth
NVMe	Non-Volatile Memory Express
PoC	Proof of Concept
RAG	Retrieval-Augmented Generation
SaaS	Software as a Service
SCO	Scontain GmbH
SDK	Software Development Kit
SGX	Intel® Software Guard Extensions (Intel® SGX)
SSD	Solid State Drive
TEE	Trusted Execution Environments
TUD	Technische Universität Dresden
UC	Use Case
UCSC	University of California Santa Cruz
UKSHA	UK Health Security Agency
URL	Uniform Resource Locator
URV	Universitat Rovira i Virgili
VM	Virtual Machine

1 Executive summary — read this introduction to understand the current TOC

This deliverable aims to present the communication, dissemination, and exploitation strategies and activities carried out during the second and final half of the project. We present the many artifacts produced to materialize the communication of the project to the various stakeholders, and we bring the complete list of the many activities carried out to disseminate NEARDATA throughout Europe and to congresses worldwide and highlight the most relevant of them.

An assessment of the results from the consortium partner's efforts to publicize the projects' realizations has been made, with special mention to the scientific work executed, and the impact on the *omics fields they represent and what they aim for the future. But not only this: use case leaders bring an important point of view of how their respective use cases can influence and benefit the society and the European economy.

Finally, we close this report with the consortium's partners exploitation plans. Everyone has a valuable perspective on how to explore not only the project's outcome, but also to employ the experience gained from working together with many people to put forth such a complex platform.

2 Dissemination and Communication Activities

This section presents a conclusion report and analysis of communication and dissemination plans. It provides a description of the activities occuring from month 17 onward to the end of the project in month 35. The complete list of events and publications from this period can be found at the appendix [6].

2.1 Communication Strategy

Throughout the project, the communication strategy implemented by the consortium has five main initiatives to facilitate engagement with different audiences: 1) project brand and identity; 2) project website; 3) social networks; 4) promotional material; and 5) newsletters. The consortium has made substantial efforts to keep communication materials and tools regularly updated, covering both technical information and content accessible to the general public.

2.1.1 Project Brand and Identity

Establishing a clear brand and cohesive project identity was a fundamental first step in enhancing the project's visibility and recognition. To develop a distinctive and consistent visual presence, the WP6 dissemination team designed a unified graphic identity and associated materials.

Brand assets, including guidelines, logos, and templates have been shared with all consortium partners through the internal repository. All communication materials are aligned with the official color scheme and visual style. A detailed overview of the graphic identity was provided in deliverable *D6.1 Communication Plan*. Project partners have systematically applied this brand identity across their communication and dissemination activities.

2.1.2 Project Website

The official project website is available at `https://neardata.eu/`. Managed by the Universitat Rovira i Virgili (URV), it is regularly updated to accurately reflect the consortium's progress and achievements. As the primary public-facing communication channel, the website provides comprehensive information on the project, including details about partners, objectives, use cases, branding elements, and project outputs such as publications, deliverables, and demonstrations. It also features recent news and events related to the project and its partners.

During the second half of the project, the website was enhanced in accordance with the reviewers' recommendations from the first review, regarding impact and use cases sections, among others.

2.1.3 Social Networks

The project maintains an active presence on the social media platform, X (formerly Twitter). This platform is key tool for increasing awareness of project activities and driving traffic to the website. It also plays an essential role in extending outreach to broader audiences and fostering engagement with stakeholders from academia, industry, and general domains.

This platform is strategically used to target specific audiences and convey tailored messages. X is utilized to reach a wider audience, promote partner participation in events, and share concise updates on project milestones. The brevity of X posts makes it particularly effective for short, non-technical announcements.

This channel makes use of hashtags such as #compute, #continuum, #extremedata, #cloud, #edge, and #HorizonEurope. The project actively follows and engages with related initiatives within its cluster (HORIZON-CL4-2022-DATA-01-05), as well as organizations such as Xartec Salut and the Big Data Value Association (BDVA), among others.

TWITTER.

Reaching a global audience of around 368 million users, this platform enables the project to communicate effectively with diverse groups, from the general public to policymakers and industrial actors. The NEARDATA account has been active since the start of the project and has demonstrated

gradual engagement growth. Figure 1 illustrates the official presence of NEARDATA on the X platform, which, by Month 35, had accumulated 45 followers, was following 50 accounts, and had published 113 posts.

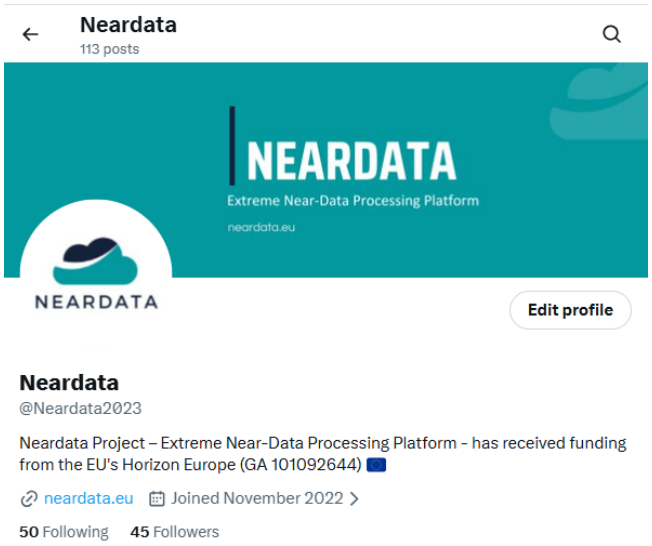


Figure 1: `https://x.com/Neardata2023`

The relatively modest number of followers may be attributed to several factors not anticipated in deliverable *D6.1 Communication Plan*. Since the project's inception, the platform has experienced notable changes in ownership, business strategy, and operational structure, leading to periods of instability and shifting priorities. More recently, the rebranding from Twitter to X may have prompted some organizations and users to reduce their engagement or discontinue their activity on the platform altogether.

To assess the impact of its communication efforts, the project has utilized X analytics. However, the analytical capabilities of the platform are somewhat limited, allowing only three months of historical data for comparison. This restriction constrains the project's ability to perform long-term trend analyses and evaluate the sustained effectiveness of its communication strategy.

Figure 2: Examples of X posts.

2.1.4 Promotional Material

Throughout the project, we have developed a set of communication and dissemination materials to support partners in creating and sharing content efficiently, consistently, and in full alignment with the project's branding and acknowledgment requirements. These resources help ensure that partners can effectively raise awareness of the project and its objectives in a coherent and engaging way. Some of the most important dissemination materials are available on the project website[1].

Video.

Promotional videos are a compelling and succinct method for showcasing the project. The first NEARDATA video[2] is a 1.44-minute clip that will serve as an introduction to the project's technology and its application in the designated use cases. Figure 3 depicts video snapshot from the first NEARDATA video.

Figure 3: First NEARDATA promotional video.

We are currently working on additional project videos, which we expect to be ready by December 2025. These will include a short video for each use case, explaining its context and the results achieved, as well as an updated project video focusing more specifically on the overall outcomes. These materials will help provide a clearer understanding of the project and its achievements. Once the videos are finalized, we will share them on the project website and social media platforms, and actively promote them to ensure they reach a broad audience.

Flyer.

A project brochure has been created to provide comprehensive information about NEARDATA, its objectives, and the implemented use cases. It was designed to be clear and accessible, ensuring that both specialized and general audiences can easily understand the project's scope and goals. The official NEARDATA brochure is available on our website[3]. The brochures have been distributed at several events, including the Cloud-Edge Continuum Workshop, the European Big Data Value Forum, Data Week, and the Mobile World Congress, among others.

Poster.

In addition to the initial poster, an updated and comprehensive poster was developed for use by all partners. This poster highlights the project's objectives, partner organizations, and use cases. Partners are encouraged to integrate it into their own communication and dissemination activities. Several copies have been printed and used at various events and other occasions. Figure 4 shows the new NEARDATA poster[4].

[1] https://neardata.eu/materials
[2] https://youtu.be/GmQetMrwCQw?si=Oe-1w5ElPLT3HbfP
[3] https://neardata.eu/assets/dissemination/NEARDATA_Brochure_DEF.pdf
[4] https://neardata.eu/assets/dissemination/NEARDATA_poster.png

Figure 4: NEARDATA poster

Other material.

The consortium continues to use the official presentation templates, quarterly management report (QMR) templates, logos, and the style guide (including font specifications) available on the shared cloud drive. All this material was showed in the previous deliverables *D6.1 Communication plan* and *D6.2 Communication and standardization report*.

2.1.5 Newsletters

Throughout 2025, the final year of the NEARDATA project, when most of the project results were achieved, we created and distributed a monthly project newsletter. Each issue was published on the project website and on social media, and also sent by email to all partners and subscribers. The newsletters covered various topics related to the project, such as high-level technical insights, new publications, project meetings, and participation in relevant events. The main goal of these newsletters is to enhance communication and outreach, ensuring that project updates and achievements reached a wider and more engaged audience. Figure 5 shows an example of one of the NEARDATA newsletters[5].

[5] `https://neardata.eu/assets/dissemination/newsletters/NEARDATA_Newsletter_October_2025.pdf`

Figure 5: NEARDATA Project Newsletter - October 2025

2.2 Dissemination Strategy

The project's dissemination strategy extends beyond communicating objectives, focusing on actively promoting outcomes to external stakeholders. The consortium has engaged diverse audiences, research and innovation organizations, industry players, and service providers across Europe, through multiple communication channels.

From Months 17–35, partners participated in *76* events and produced 25 publications, establishing the project's visibility and research direction. Activities intensify in the last part of the project as results mature, targeting communities that can benefit most. In the final phase, the WP6 team will focus on technology adoption and sustainability through industry engagement, and strategic partnerships within and beyond the NEARDATA community. The NEARDATA's visibility has been strengthened by partners everywhere they gave talks, exposed their work, and met with stakeholders.

2.2.1 Event Participation

Events serve as a key channel for dissemination, providing essential opportunities to present project achievements and engage diverse audiences. In particular, leading peer-reviewed conferences offer the consortium a platform to showcase the latest project developments and foster early-stage, technology-oriented discussions. A detailed list of dissemination events (period M1-M16) was included in Deliverable *D6.2 Communication and standardization report*, and the WP6 team has continued to identify new opportunities to connect with target audiences.

During Months 17–35, the consortium participated in *76* events addressing academic, research, industrial, and general public audiences. Highlights of the main contributions are summarized below, while full details are available in Appendix 6 Dissemination and Meeting Activities.

INTERNATIONAL MIDDLEWARE CONFERENCE – MIDDLEWARE'24.

Organized by ACM (Association for Computing Machinery) and the IFIP (International Federation for Information Processing), the congress Middleware[6] is an annual academic and scientific forum for discussing innovations and advancements in middleware systems and distributed computing, covering the design, implementation, deployment, and evaluation of distributed systems and platforms.

[6] `https://middleware-conf.github.io/2025/`

In 2024, MIDDLEWARE was held in Hong Kong, China[7], from December 2nd to 6th. Partners brought NEARDATA once more to Asia and presented two papers produced in the consortium: "StreamSense: Policy-driven Semantic Video Search in Streaming Systems" and "Serverful Functions: Leveraging Servers in Complex Serverless Workflows". Figure 6 shows the presentations.

Figure 6: Pictures of MIDDLEWARE 2024 – Horizon projects and research papers presented by URV, NCT, and Dell.

Our partners EMBL, URV, together IBM Research joined important Asian and European organizations, like China Mobile Suzhou Software, Chinese University of Hong Kong, Reykjavik University, Universidade de Lisboa, and many others to present advancements on distributed systems technologies, like cloud systems, distributed machine learning, and data and stream processing.

CLOUD-EDGE CONTINUUM WORKSHOP 2024 & 2025 – CEC'24 & CEC'25.

Organized by our partner, Dell Technologies, CEC is held co-located with IEEE International Conference on Network Protocols (IEEE ICNP), which is a well-established scientific conference in the field of computer networks, IoT and data centers, as well as network architectures, security and AI/ML in networks. CEC is part of the *European Cloud, Edge and IoT Continuum Initiative*[8], which promotes collaboration between researchers, developers, and users to advance the Cloud, Edge, and IoT (CEI) continuum. CEC brings together researchers and practitioners from academia and industry to discuss, partner, and develop innovations on interconnecting and abstracting the computing power and storage capacity of the Cloud with the data processing capabilities of CEI artifacts.

In 2023, it was held in Reykjavik, Iceland[9], at October10th. Counting with the presence of researchers, staff, project managers from various other universities, research laboratories, and companies, such as University of Oulu, University College Cork, Aristotle University of Thessaloniki, NVIDIA, L3S Research, INESC-ID Lisbon, TU Berlin, and many others, the event not only was successful in its purpose but prompted a continuation for the following years.

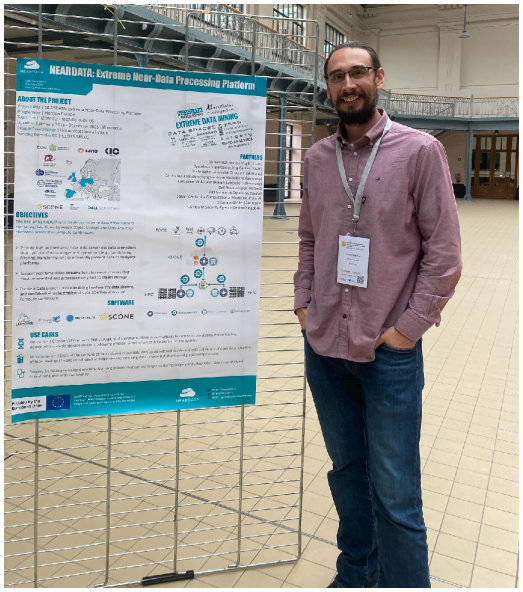
The following realization of CEC was hosted in Charleroi, Belgium[10], at October 28th, 2024. Heavy-weight players, like IBM Research, Fraunhofer Institute for Microelectronic Circuits and Systems, TU Eindhoven, and Huawei Technologies, joined BSC, University College Cork, NVIDIA, TU Dresden and many others to present their works [7a].

The poster session [7b][7c] is a great place to expand your networking, start conversations with prospect partners, share experiences, and maybe put forward some cooperation.

The 3rd CEC Workshop took place in Seoul, South Korea[11], at September 22nd, 2025. Joining our partners BSC, Dell, KIO Netowrks, and TU Dresden, organizations like Yonsei University, University of Oulu, Cellnex Telecom, Nearby Computing, Universitat Politecnica de Catalunya, Intel, and others, shared their experiences on distributed systems and brought NEARDATA as far as Asia [7d].

[7]MIDDLEWARE 2024 – `https://middleware-conf.github.io/2024/`
[8]`https://eucloudedgeiot.eu/event/cec-workshop-2025/`
[9]CEC Workshop 2023 – `https://cec23.github.io/index.html`
[10]CEC Workshop 2024 – https://cec24.github.io/index.html
[11]CEC Workshop 2025 – `https://cec25.github.io/index.html`

(a) Cloud-Edge Continuum Workshop 2024 Prof. Christof Fetzer – keynote on "*Using Confidential Computing to Protect Applications in a Cloud-Edge Continuum*".



(b) NEARDATA's Dr. Raúl Gracia.

(c) Attendants discussing a poster, with NEARDATA's at the back.

(d) Cloud-Edge Continuum Workshop 2025 took place in Seoul, South Korea.

Figure 7: Cloud-Edge Continuum Workshop 2024 and 2025.

INTERNATIONAL WORKSHOP ON SERVERLESS COMPUTING EXPERIENCE – WOSC.

The WOSC is the result of a partnership between URV and IBM Research, that, since 2023, has gathered important players on serverless technologies to show their work and share experiences. SANO and BSC joined URV to promote the event, along with other prominent attendees such as Telefonica Research, Imperial College London, ETH Zürich, University of Würzburg, and many others. Current challenges on serverless computing, proposed solutions and the future in the field are addressed in the talks[8].

All the talks are available online in `https://www.youtube.com/@serverlessworkshops/playlists`

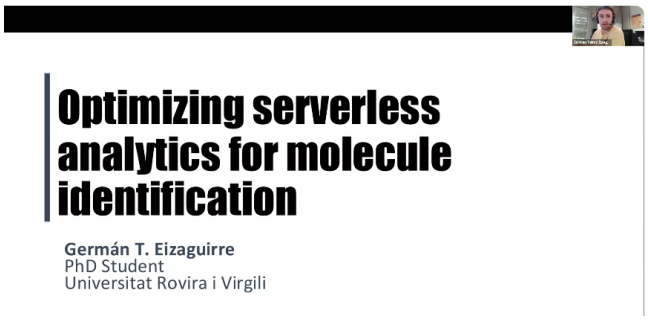


Figure 8: WoSCX4 Session 2 – papers presentation.

EUROPEAN BIG DATA VALUE FORUM – EBDVF 2024 and 2025.

Organized by the Big Data Value Association (BDVA)[12], the EBDVF is the central event for the European Data and AI research and innovation community, where professionals from industry, research, and policy meet to advance data-driven innovation.

NEARDATA has always been very well represented by our coordinator. Together with the other projects from the DataNexus cluster, we have participated in two editions to present the cluster and the work we are carrying out.

In the 2024 edition (held in October 2024 in Budapest, Hungary), the DataNexus team delivered a presentation entitled "*Innovative Approaches to Extreme Data Challenges*"[9][13]. In addition, the DataNexus cluster hosted a booth to further enhance the cluster's visibility.

Figure 9: Pictures of DataNexus session and booth at EBDVF 2024 (October 2024).

And in the 2025 edition (held in November 2025 in Copenhagen), the cluster presented the projects in a talk entitled "*Extreme Data Architectures Empowering AI Innovation Across the Compute Continuum*"[10][14]

[12] `https://bdva.eu/`

[13] `https://european-big-data-value-forum.eu/2024-edition/programme/?edition_session_id=13076`

[14] `https://european-big-data-value-forum.eu/2025-edition/programme/`

Figure 10: Pictures of DataNexus session at EBDVF 2025 (November 2025).

2.2.2 Other Activities

In addition to the key activities outlined in previous sections, consortium partners also participated in various events and meetings and contributed news articles, blog posts, and white papers featuring the NEARDATA project. A complete list of these dissemination activities is provided in Appendix 6 Dissemination and Meeting Activities.

Moreover, BSC granted priority access to Supercomputing resources of up to 10 million CPU hours to run the experiments related to the Epistasis use-case under grant IDs BCV-2024-2-0004 and BCV-2025-2-0008. We have been able to run the full dataset, which has contributed to enabling a new analysis method of T2D beyond the state-of-the-art GWAS.

2.2.3 Publications

Peer-reviewed publications serve as a key channel for disseminating the consortium's research findings and ensuring that NEARDATA's technical results are openly and widely shared. This section highlights major conferences where NEARDATA partners participated. Between Months 17 and 35, 25 publications were released, with selected works summarized below. A complete list of these publications is provided in Appendix 6.2 Publications. In the previous deliverable *D6.2 Communication and standardization report* 17 publications were reported for Months 1–16. In total, 42 publications have been produced to date.

USENIX Annual Technical Conference (USENIX ATC '25)

The USENIX Annual Technical Conference (USENIX ATC '25) is a premier, peer-reviewed conference organized by the USENIX Association that brings together researchers, practitioners, and industry experts to present and discuss cutting-edge advances in systems software and engineering. It covers a wide range of topics, including operating systems, cloud infrastructure, storage, networking, security, and performance, emphasizing practical, innovative, and technically rigorous work.[15]. The following the paper was presented:

- **Burst Computing: Quick, Sudden, Massively Parallel Processing on Serverless Resources**[16]
  Abstract: We present burst computing, a novel serverless solution tailored for burst-parallel jobs. Unlike Function-as-a-Service (FaaS), burst computing establishes job-level isolation using a novel group invocation primitive to launch large groups of workers with guaranteed simultaneity. Resource allocation is optimized by packing workers into fewer containers, which accelerates their initialization and enables locality. Locality significantly reduces remote communication compared to FaaS and, combined with simultaneity, it allows workers to communicate synchronously with message passing and group collectives. Consequently, applications unfeasible in FaaS are now possible. We implement burst computing atop OpenWhisk and

[15] https://www.usenix.org/conference/atc25
[16] `https://zenodo.org/records/17662627`

provide a communication middleware that seamlessly leverages locality with zero-copy messaging. Evaluation shows reduced job invocation and communication latency for a 2× speed-up in TeraSort and a 98.5% reduction in remote communication in PageRank (13× speed-up) compared to standard FaaS.

Middleware Industrial Track '24: Proceedings of the 25th International Middleware Conference Industrial Track.

An annual conference and a major forum for the discussion of innovations and recent scientific advances of middleware systems with a focus on the design, implementation, deployment, and evaluation of distributed systems, platforms and architectures for computing, storage, and communication[17]. The following the papers were presented:

- **Serverful Functions: Leveraging Servers in Complex Serverless Workflows (industry track)**[18]
  Abstract: The scalability of cloud functions makes them a convenient backend for elastic data analytics pipelines where parallelism changes drastically from one stage to the next. However, cloud functions require intermediate storage systems for communication, which limits the efficiency of stateful operations. Furthermore, cloud functions are expensive, which reduces the cost-effectiveness of pure serverless architectures. We propose a hybrid architecture for data analytics that uses cloud functions for embarrassingly parallel stages and virtual cloud instances for stateful operations under a unified serverless programming framework. Extending Lithops, a serverless programming library, we implement a parallel programming interface that proactively provisions serverless and serverful cloud resources with minimal user intervention. We validate the feasibility of a hybrid architecture, by comparing it to fully serverless and serverful versions of a production-level metabolomics pipeline. We show that mixing cloud functions with virtual instances increases the cost-effectiveness of the execution by up to 188.23% over the serverless implementation, while achieving a speedup of 3.64 compared to the serverful one.

- **StreamSense: Policy-driven Semantic Video Search in Streaming Systems**[19]
  Abstract: Streaming systems are an increasingly appealing substrate for managing video data via the stream abstraction. However, if we consider a large stream collection, it can be hard for data scientists to discover and locate relevant videos, let alone specific video fragments. In this paper, we propose StreamSense: a policy-driven, semantic video search solution for streaming systems. StreamSense allows users to deploy AI models that generate embeddings from video frames via policies. Our system uses such embeddings for building a two-level index in a vector DB that efficiently handles inter/intra video queries. StreamSense abstracts users from vector DB interactions so they can perform semantic search using images as input and visualize the results. We built our prototype on top of a tiered streaming storage system (Pravega) and validated it on a health-related use case. We show that StreamSense allows data scientists to search for video fragments in real surgery datasets in < 30ms. StreamSense also reduces data ingestion related to AI training data loading in +80% compared to simple bulk loading video streams.

- **"Back to the Byte": Towards Byte-oriented Semantics for Streaming Storage**[20]
  Abstract: Event streaming systems, such as Apache Kafka, are increasingly used for data-intensive applications due to the proliferation of streaming sources of data and the requirement of responding with low latency. These systems center their internal design and external APIs around the concept of events. In this paper, we challenge the event-centric design as the primary design option for event streaming systems. Instead, we describe a streaming storage design based on a byte-oriented streaming primitive that supports key properties such as

[17] https://middleware-conf.github.io/2024/
[18] `https://zenodo.org/records/15019182`
[19] `https://zenodo.org/records/17348282`
[20] `https://zenodo.org/records/17348225`

atomicity, conditional writes, and durability. Such a primitive allows us to implement not only a traditional event API, but also other APIs supporting, e.g., byte streams, key-values, and state synchronization. A byte stream API is relevant for storage workloads requiring large data transfers or low-latency byte buffer IO. We implement this concept in Pravega: a tiered storage system for data streams. Via experiments on AWS, we show the practicality of exposing byte APIs for storage workloads compared to events.

International Conference on Service-Oriented Computing (ICSOC)

The International Conference on Service-Oriented Computing (ICSOC) is a leading, peer-reviewed research conference focused on the science and technology of service-oriented systems, cloud services, and related paradigms such as microservices, service orchestration, BPM, and service-based applications. It brings together researchers, practitioners, and industry experts to present advances in models, methods, architectures, and tools for designing, deploying, and managing service-oriented solutions [21]. The following the paper will be presented in December:

- **Quantifying Serverless Elasticity: The gumeter Benchmark Suite**[22]
  Abstract: Serverless computing has emerged as a powerful paradigm for distributed workflows, offering fine-grained, low-latency resource provisioning to precisely meet job demands. While workflows often utilize hundreds of concurrent CPUs, existing serverless benchmarking suites frequently concentrate on small-scale parallelism and microservice workloads. Furthermore, these benchmarks typically consider only Function-as-a-Service (FaaS) backends, overlooking their natural cloud successors, Container-as-a-Service (CaaS). These limitations have created a significant gap in evaluating a crucial feature of serverless platforms: their ability to accurately handle sudden changes in resource allocation, also known as elasticity. To address this, we introduce gumeter, a new benchmarking suite specifically designed to evaluate the elasticity of serverless platforms (both FaaS and CaaS) for highly parallel distributed workflows. gumeter facilitates a thorough assessment of an underlying platform using a "fire-and-forget" execution model and minimal user intervention. It leverages a set of comprehensive pipelines that sample various typical scaling behaviors, providing an in-depth analysis of elasticity, execution time, cost, and efficiency. We apply gumeter to evaluate the elasticity of three popular serverless platforms: AWS Lambda, Google Cloud Run, and IBM Code Engine. Our results reveal significant differences in elasticity across these platforms, showing that FaaS offerings still out-perform CaaS in elasticity by a factor of up to 6.5x. However, despite a lower performance, CaaS can be up to 64.5% less expensive than FaaS in specific scenarios, thereby unveiling an interesting optimization space for cloud workflows.

Publication Guideline.

Under Horizon Europe, Open Science has become a mandatory requirement. This entails that both publications and research data must comply with Open Access requirements. As publishing in open access may raise certain questions or uncertainties, a guideline document was developed to support project partners in following the correct procedures. This guide has been presented and distributed among the consortium partners to summarize and clarify the most relevant concepts, thereby assisting them in meeting all Open Access requirements effectively. Figure 11 shows two slides of this guideline.

2.2.4 Community Building

Community building focuses on deepening stakeholders' understanding of project outcomes and encouraging wider engagement within the NEARDATA ecosystem. The WP6 team, together with the consortium, has actively worked to identify and expand the project's community, fostering synergies and collaborations with broader European networks in AI, data, industry, academia, and initiatives

[21] https://icsoc2025.hit.edu.cn/main.htm
[22] https://zenodo.org/records/17701734

Figure 11: Screenshots of the Open Access guideline.

such as the Horizon Results Booster.

Horizon Results Booster.

The European Commission's Horizon Results Booster (HRB)[23] initiative aims to accelerate the translation of research into market-ready innovation and maximize the impact of publicly funded research across the EU. Its purpose is to help projects go beyond their Dissemination and Exploitation (D&E) requirements, guiding research toward tangible societal benefits and reinforcing the value of Research and Innovation (R&I) in addressing societal challenges.

To support these objectives, HRB provides free consultancy services to both completed and ongoing research projects funded under FP7, Horizon 2020, or Horizon Europe. The NEARDATA project had leveraged Module A of this service, together with the Graph Massivizer and EXTRACT projects, to develop a portfolio dissemination and exploitation strategy aimed at identifying and consolidating research and innovation outputs. Then, thanks to HRB, the group expanded with four new projects to form the DataNexus cluster, explained in more detail below.

DataNexus Cluster.

NEARDATA joined other projects in the DataNexus Cluster [12]. Seven EU projects collaborate around shared mission to extract meaningful insights from extreme data. We seek to answer the Horizon Europe call for 'Extreme data mining, aggregation and analytics technologies and solutions' [24].

Figure 12: DataNexus logo.

DataNexus projects develop solutions for managing extreme data-characterised volume, speed, and complexity – to securely extract meaningful insights from raw data. These insights help support advanced decision-making, leveraging big data, artificial intelligence (AI), Internet of Things (IoTs) and advanced computing paradigms.

The cluster is composed by seven EU-funded projects: *Graph-Massivizer*[25], *EXTRACT*[26], *NEAR-*

[23] `https://www.horizonresultsbooster.eu/`

[24] Extreme data mining, aggregation and analytics technologies and solutions (RIA) - `https://cordis.europa.eu/programme/id/HORIZON_HORIZON-CL4-2022-DATA-01-05`

[25] Graph-Massivizer - `https://graph-massivizer.eu/`

[26] EXTRACT - `https://extract-project.eu/`

DATA[27], *EXA4MIND*[28], *EMERALDS*[29], *SYCLOPS*[30], and *EFRA*[31].

The projects came together to form the cluster thanks to the Horizon Results Booster initiative. Since then, we have collaborated on multiple activities, including a flyer[32], a video[33], several press releases[34] [35] [36], and presentations at On-Demand Solutions with AI, Data, and Robotics event[13] and various editions of the EBDVF and the Data Week[37] [14]. They are also preparing a joint paper, among other initiatives. All these efforts aim to exchange experiences and join forces to achieve better outcomes and greater visibility.

Figure 13: DataNexus at Future-Ready event, held on February 18, 2025 in Brussels.

The DataNexus Cluster projects have led to significant improvements in data processing, analysis, and visualization. These advancements have made data handling more accurate, faster, and more usable. The benefits are expected to be felt across a wide range of fields, including crisis management, healthcare, and environmental protection, thanks to a human-centered, user-friendly design.

2.2.5 Science for Society

Project partners have a responsibility to communicate the project's activities and results to the general public. The consortium has made dedicated efforts to raise awareness and promote NEARDATA within society. A few examples include:

URV shared information and a poster about the NEARDATA project on the European Corner[38] as part of the **European Researchers' Night 2025**, an annual event held in over 300 cities across 30 European countries. The initiative aims to bring research and innovation closer to citizens through interactive activities such as workshops, talks, demonstrations, and games, engaging audiences of all ages.

URV's **T-Systems Cloud Computing Chair**[39] focuses on analyzing the impact of cloud technologies on business and society through research, training, and outreach. During the second half of the

[27] NEARDATA - https://neardata.eu/
[28] EXA4MIND - https://exa4mind.eu/
[29] EMERALDS - https://emeralds-horizon.eu/
[30] SYCLOPS - https://www.syclops.org/
[31] EFRA - https://efraproject.eu/
[32] https://neardata.eu/assets/img/datanexus/DATANEXUS_Flyer.pdf
[33] https://www.youtube.com/watch?v=n3xACMCbUw4
[34] https://neardata.eu/assets/img/datanexus/Data_Sister_Projects_press%20release_FINAL.pdf
[35] https://neardata.eu/assets/img/datanexus/DataNexus_EBDVF24_press_release.pdf
[36] https://neardata.eu/assets/img/datanexus/DataNexus_Future-Ready.pdf
[37] https://data-week.eu/2025-edition/programme/
[38] https://lanitdelarecerca.cat/neardata-extreme-near-data-processing-platform-2
[39] https://www.urv.cat/en/society-business/chairs/cloud-computing/

Figure 14: DataNexus at DataWeek'25, held on May 28 in Athens.

project, and as part of its dissemination activities, the Chair delivered 21 presentations on cloud computing topics, introducing the NEARDATA project to approximately 525 secondary school students in the Tarragona region. Figure 15 depicts one of these sessions, held on February 10, 2025, focusing on Cloud Computing, Big Data, and Artificial Intelligence. In addition, seven training courses were organized for university students, providing knowledge about cloud computing to around 150 participants. The Chair also maintains a blog[40] where posts related to cloud topics are regularly published. This space serves to share updates, analyses, and insights aligned with its areas of work.

Figure 15: One of the 21 presentations of the T-Systems Cloud Computing Chair.

The NEARDATA project has been actively involved in organizing two hackathons aimed at engaging university students and fostering their skills in Cloud Computing through hands-on challenges.

Led by the T-Systems Cloud Computing Chair, the first **URV–T-Systems Cloud Computing Hackathon 2025**[41] was held in February 2025 [16]. The objective of the activity, which brought together around forty engineering students, was to design a cloud-based architecture for managing socially impactful data. The hackathon was open to senior undergraduate and master's students from various engineering disciplines, such as Computer Engineering and Telecommunications Systems and Services Engineering. The challenge involved developing cloud software of public interest

[40] https://cloudlab.urv.cat/catedracloud/blog/
[41] https://cloudlab.urv.cat/catedracloud/hackato/

using regional datasets and Amazon Web Services (AWS) capabilities. Teams were tasked with migrating data to the cloud, analyzing it, and creating a socially relevant application with an intuitive visual interface.

Figure 16: Participants in URV–T-Systems Cloud Computing Hackathon 2025.

In June 2025, the **Cloud Meets Innovation Hackathon**[42] took place, co-organized by Universität Innsbruck (UIBK) and Universitat Rovira i Virgili (URV)[17]. Co-hosted on June 6–7 across both campuses, each site addressed its own challenge while participating in hybrid sessions that allowed teams to showcase demonstrations and share results in real time. The event encouraged participants to test their skills, leverage live cloud services and data streams, and engage in hands-on innovation inspired by leading European research projects, with opportunities to compete for exclusive awards.

Figure 17: Participants in Cloud Meets Innovation Hackathon.

2.3 Audience Assessment and Publications Impact

Partners have attended many congresses, meetings, workshops that made them have contact with heterogeneous audiences, albeit our major public has been the scientific community. NEARDATA has been the motivation for presentations, was mentioned in others, and was discussed with other peers. Additionally, publications produced by the consortium sum up tens of citations – something

[42] `https://cmi-hackathon.github.io/`

that makes NEARDATA contribute indirectly to other scientific projects.

We now present an assessment of the dissemination audiences, according to the Key Performance Indicator (KPI) to maximize impact - dissemination, exploitation and communication (*Table 2.2a: Communication* from the document of work). And closing this sub section it is presented an assessment of publications sponsored by the consortium.

2.3.1 Dissemination Performance

The project has accomplished all dissemination KPIs and reached the different scientific, industrial, and societal audiences.

Regarding scientific audiences, the project has obtained a record number of publications (42) in scientific conferences and high impact journals. This means that thousands of scientists in the fields of computer science and bioinformatics are aware of the achievements of NEARDATA project.

Regarding industrial impact, partners in the consortium have presented results in industrial conferences with massive audiences like the Mobile World Congress in Barcelona, in global industrial events with DELL, but also in well-know developer conferences in Python, Big Data, and AI like SciPy, EuroScipy or PyCONEs. We also participated in EU events organized by BDVA in AthenS, Budapest and Copenhagen and ADRA event in Brussels where numerous industrial partners were present.

We also outline the exemplar collaboration with other extreme data projects in the same call in the context of the DATA NEXUS CLUSTER that presented results together in many events. This also increased our exposure to other european key industrial partners. In those events, the consortium established contacts with a number of european cloud providers that were exposed to NEARDATA results including KIO Networks, Scaleway, OVH, CloudFerro, Nebius or AppDistrict among others.

Consortium members also exposed results to a number of hospitals in Europe including for example: Institute of Mother and Child, University Clinical Hospital Białystok in Poland; the German university hospitals in Heidelberg, Pediatric Surgery Department of the University Hospital Carl Gustav Carus in Dresden; and the St. Mary's hospital of the Imperial College London. We also disseminated results to hospitals and clinics like Klinikum Rechts an der Isaar TUM in Munich, University Hospital of Krakow, Asklepios-ASB Krankenhaus in Radeberg, Hospital Mainz Frankfurt, Diakonissenkrankenhaus in Dresden, Krankenhaus St. Joseph Stift Dresden, st. Elisabethen-Krankenhaus in Ravensburg, Basel Hospital, IHU Strasbourg innovative image-guided therapies, Hospital Clínic de Barcelona, Institut Guttmann, Hospital Sant Pau, Germans Tries i Pujol, and Sant Joan de Déu in Barcelona.

We also made efforts to bring information to government authorities, such as to Petra Köpping, the former Saxony's Minister of State for Social Affairs and to government representatives in Spain and Poland.

Category of audience	Completed
Scientific community	✓
European Cloud providers	✓
Hospitals	✓
General Public	✓

Table 1: Target audiences

Finally, the project also targeted the general society through different publications in mass media (newspapers, TVs) and in the websites and social networks of the different institutions in the consortium. We were present in European researchers night, in science demonstrations in the different partners (Tarragona, Dresden, Barcelona, Krakow), and even in talks in schools addressed to young students. We also organized different hackathons (Tarragona, Innsbruck) that demonstrated project software results to computer engineering students.

In summary, we have achieved all the dissemination KPIs expected in the beginning of the project in the different scientific, industrial and societal communities.

2.3.2 Impact of Publications on *omics Industries

In the second period 25 publications were written sponsored by the consortium (with acknowledgment for NEARDATA's code "**101092644**"), with multiple citations. For example, we have many papers cited in multiple journals published by IEEE, by Springer, and by ACM – like "*IEEE Transactions on Sustainable Computing*", "*International Journal of Computer Assisted Radiology and Surgery*", and "*International Conference on Management of Data*" respectively. This demonstrates how far the knowledge produced by NEARDATA can travel. Table [2] presents the publications ranked according to the number of citations (with 5 or more citations). For a complete list, refer to both *D6.2 Communication report and the standardization report*, for the first half of the project, and table [9], for the second and final period.

Table 2: Number of citations publications from M1 - M35 (obtained from Google Scholar – at 25/11/2025)

Publication title	Citations
MLLess: Achieving Cost Efficiency in Serverless Machine Learning Training	33
Serverless End Game: Disaggregation enabling Transparency	27
Trustworthy confidential virtual machines for the masses	20
Pravega: A Tiered Storage System for Data Streams	10
METASPACE-ML: Metabolite annotation for imaging mass spectrometry using machine learning	8
Glider: Serverless Ephemeral Stateful Near-Data Computation	8
Novel Approaches Toward Scalable Composable Workflows in Hyper-Heterogeneous Computing Environments	8
One model to use them all: Training a segmentation model with complementary datasets	7
Triad - Trusted Timestamps in Untrusted Environments	6
Challenges and Opportunities for RISC-V Architectures towards Genomics-based Workloads	5
Practical Storage-Compute Elasticity for Stream Data Processing	5
A Comprehensive Study on the Impact of Vulnerable Dependencies on Open-Source Software	5
Dataplug: Unlocking extreme data analytics with on-the-fly dynamic partitioning of unstructured data	5

Ranking journals according to the CORE Conference Ranking.

A good measure to evaluate how well received have been the partners' articles, is to look at how the journals where they got published rank in the CORE Conference Ranking. CORE – *Computing Research and Education Association of Australia* – is a well-known system used to assess the quality of academic conference in computer science and related fields[3; thorough explanation]. Table [4] presents all the venues with NEARDATA's publications since the start of the project.

Table 3: CORE grades system explained

Grade	Description
A*	**Flagship**; highly visible and well known both within their own community. 7.65% of 784 ranked venues
A	**Excellent**; many of the characteristics of an A*; less widely known and less visible outside their community. 14.92% of 784 ranked venues
B	**Good to Very Good**; program committees may include a lower ratio of established researchers. 28.06% of 784 ranked venues
C	**Sound and Satisfactory**; genuine academic venues with a selection process involving full papers referred by academics prior to acceptance. 46.17% of 784 ranked venues
Unranked	**Not yet evaluated**; possibly new, interdisciplinary or local workshops

Figure [18] shows the total number of articles NEARDATA had published throughout its lifetime according to the grade the journals have in the CORE ranking scale. With *24* published articles in well established venues, the intellectual production has been remarkable, especially when it is considered the number of citations [2] for such a fresh set of papers. We have *43%* published in unranked venues, which means the venues are not yet classified by CORE, but they also have strict selection of papers. The ISI JCR ranking for journals brings the information that we also had published in the top *Q1* and *Q2* quartile venues [5], again supporting the high quality of intellectual production.

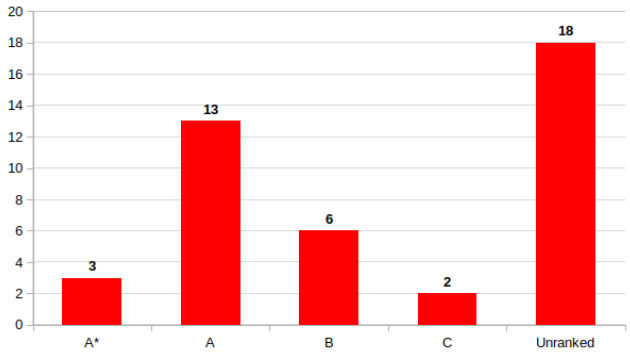


Figure 18: Number of articles published per venues according to CORE grade system.

Table 4: CORE Conference Ranking from M1 - M35

Title	Acronym	Core Rank
IEEE International Conference on Computer Communications	INFOCOM	A*
IEEE Conference on Computer Vision and Pattern Recognition	CVPR	A*

Continued on next page

Table 4 (continued)

Title	Acronym	Core Rank
ACM SIGMOD-SIGACT-SIGART Conference on Principles of Database Systems	PODS	A*
International Symposium on Software Reliability Engineering	ISSRE	A
International Conference on Service Oriented Computing	ICSOC	A
ACM/IFIP/USENIX International Middleware Conference	Middleware	A
International Conference for High Performance Computing Networking Storage and Analysis (was Supercomputing Conference)	SC	A
Usenix Annual Technical Conference	USENIX	A
IEEE International Conference on Cloud Computing	CLOUD	B
IEEE International Symposium on Cluster Cloud and Grid Computing	CCGRID	B
International Conference on Network Protocols	ICNP	B
IEEE International Conference on Cluster Computing	CLUSTER	B
International Conference on Network Protocols	ICNP	B
IEEE International Conference on Cloud Computing Technology and Science	CloudCom	C
ISC High Performance (was International Supercomputing Conference)	ISC	C
International Workshop on Serverless Computing	WoSC	Unranked
European Dependable Computing Conference	EDCC	Unranked
IEEE International Conference on eScience	eScience	Unranked
Cloud-Edge Continuum (co-located with IEEE ICNP)	CEC	Unranked
International Conference on Information Processing in Computer-Assisted Interventions	IPCAI24	Unranked
Workshop on Serverless Systems, Applications and Methodologies	SESAME	Unranked

Table 5: ISI JCR Journal Ranking

Title	ISI JCR Rank
Journal of Parallel and Distributed Computing	Q1
Nature Scientific Reports	Q1
Nature Communications	Q1
Electronics 2025	Q2

3 Use Cases Impact

NEARDATA is the story of a set of use cases on the *omics macro segments, namely *genomics*, *surgomics*, and *metabolomics*. In each of them, while managing to solve questions and advance technology in their respective fields, the use case leaders kept an eye on the impact the work developed here had on six aspects: *scientific*, *economic*, *societal*, *health data spaces*, *industry standards*, and *future standards*. The impact that the project had on global standards and the software adoption as part of the reference stack is addressed below by the leaders.

3.1 Use Case: Genomics Epistasis

3.1.1 Societal Impact

The implementation of predictive models for complex disease prevention represents a transformative step toward improving public health outcomes and fostering healthier, longer lives across Europe. Complex disorders, such as cardiovascular disease, diabetes, cancer, and neurodegenerative conditions, are highly prevalent and rank among the top ten causes of death worldwide, underscoring their profound societal burden. By identifying individuals with a heightened genetic predisposition to these conditions, predictive models enable the timely deployment of targeted prevention strategies. These strategies often center on lifestyle interventions, including dietary adjustments, physical activity, and behavioral support, which are both cost-effective and scalable. Crucially, such proactive measures can delay disease onset or, in certain cases, prevent it entirely. This shift from reactive treatment to anticipatory care not only reduces the burden on healthcare systems but also empowers citizens to take control of their health, reinforcing the EU's commitment to equitable, person-centered care.

3.1.2 Scientific Impact

Despite notable progress in the genomic characterization of complex disorders over the past two decades, significant gaps remain in our understanding of their multifactorial nature. These diseases often arise from intricate interactions between genetic variants and environmental exposures, necessitating analytical frameworks that can capture this complexity. The integration of artificial intelligence and machine learning into large-scale genomic analyses offers a powerful avenue for uncovering hidden patterns and associations. By applying massively parallel models to diverse datasets, researchers can refine polygenic risk scores, identify novel biomarkers, and stratify populations based on individualized risk profiles. This scientific advancement lays the groundwork for precision medicine approaches, enabling the development of more effective, personalized prevention and treatment strategies that align with the EU's vision for innovation-driven healthcare.

3.1.3 Economic Impact

Complex disorders, such as cancer, metabolic syndrome, and mental health conditions, affect millions of individuals across Europe and are among the leading drivers of healthcare expenditure. Their chronic nature and associated comorbidities often require long-term management, resulting in substantial direct and indirect costs for national health systems. Early detection and prevention models offer a compelling economic solution by shifting the focus from costly late-stage interventions to proactive risk mitigation. Identifying high-risk individuals before disease onset allows for the implementation of tailored prevention protocols, which can significantly reduce hospitalization rates, pharmaceutical dependency, and productivity losses. Over time, this approach contributes to a more sustainable healthcare ecosystem, freeing up resources for innovation and improving the overall resilience of public health infrastructure.

3.1.4 Health Data Spaces

The development of health data spaces is essential to accelerating scientific progress and delivering societal impact. While full-scale data sharing remains a challenge, recent advances have laid important groundwork. Due to the sensitive nature of genomic and clinical data, access to individual-level information requires robust ethical and legal frameworks, as well as global standards for secure and

interoperable exchange. Initiatives such as ELIXIR and GA4GH have contributed significantly by proposing governance models and developing tools like the Beacon protocol, which enables federated genomic data discovery. Foundational infrastructures such as EGA, dbGaP, and UK Biobank continue to play a critical role by providing controlled access to high-value datasets. In this project, we have leveraged these resources and will contribute to international health data spaces by publishing summary statistics and research outputs on open-access platforms such as Zenodo. This supports FAIR data principles and facilitates ethical reuse across borders and disciplines, including applications in public health modeling.

3.1.5 Industry Standards

The core of the NEARDATA project is to build connectors to aid the aforementioned challenges. Thus, the published connectors help any party member perform wider analysis with their own data and aid in the results. Therefore, while we can't contribute to data sharing, we do contribute to generating connectors for the extreme processing of chromosome variations.

The connectors of NEARDATA are publicly available along with their documentation, and they intend to provide a mechanism to build custom pipelines based on those. Therefore, reinventing the wheel is no longer necessary, and each use case can be re-implemented via a collection of standard connectors. The behavior of such connectors, along with their documentation, is well documented in deliverables 3.2 and 5.2, while the reference architecture can be found in deliverable 2.3.

3.1.6 Future Standards

Future standards include improvements for the already designed connectors as well as developing new ones, allowing the execution of the use-case outside supercomputers easily. While the already developed connectors may allow that, work is required to modify and tune them for proper behavior in cloud providers.

3.2 Use Case: Computer-Assisted Surgery

Pioneering work in computer-assisted video surgery, led by the National Center for Tumor Diseases (NCT), is making significant strides across societal, scientific, economic, and healthcare landscapes. These innovations are not only enhancing surgical precision and patient outcome but also shaping the future of medical technology and data sharing, with a clear trajectory towards influencing global standards.

3.2.1 Societal Impact: Fostering Future Generations and Public Engagement

A strong commitment to societal outreach is evident through a variety of initiatives aimed at engaging the public and inspiring the next generation of scientists. The co-organization of the CeTI "Girls for Robots" workshop for three consecutive years (2023-2025)[43] is actively encouraging young women to pursue careers in STEM fields. Public engagement is further fostered through the "Lange Nacht der Wissenschaft Dresden" (2023-2025)[44], where the doors of research are opened to the public, showcasing cutting-edge medical techniques.

Efforts to cultivate interest in STEM studies are also a priority, with direct communication and tailored programs for school students. University students are provided with unique insights into medical research through lectures, lab projects, and collaborations with institutions like the Hans-Riegel Stiftung[45], the SECAI-CeTI summer school[46], and students from the Dualen Hochschule Baden-Württemberg (DHBW). These exchanges focus on human-robot interaction and computer science in a medical context. A notable recognition of this societal contribution is the selection of Prof. Stefanie Speidel as one of the faces of the SPIN2030 Saxony science initiative[47].

[43] https://x.com/TSO_Lab/status/1864312385772691881

[44] https://www.linkedin.com/posts/translational-surgical-oncology_at-this-years-dresdner-lange-nacht-der-wissenschaften-activity-7342872060353761280-cq8O?...rcm=ACoAADJyVz4B7_V9rAYi8wdKLS6goQZd3_0DEs8

[45] https://www.linkedin.com/posts/translational-surgical-oncology_surgicalailab-translationalsurgicaloncology-activity-7370490654377820160-ono3?...rcm=ACoAADJyVz4B7_V9rAYi8wdKLS6goQZd3_0DEs8

[46] https://ceti.one/international-summer-school-on-ai-applications-for-medicine/

[47] https://spin2030.com/faces-of-spin/prof-dr-ing-stefanie-speidel/

3.2.2 Scientific Impact: A Foundation for Future Research and Patient Benefit

The scientific contributions of this work are underscored by the creation and public release of the Appendix300 dataset, the related EndoVis Challenge FedSurg24, FL-EndoViT, "One model to use them all"-paper, and the demos using GStreamer, Pravega, Scone, and Federated Learning.

The Appendix300 dataset [1] addresses the critical barrier of limited data in surgical AI by providing a comprehensive, multi-center collection of video footage from numerous laparoscopic appendectomies in both pediatric and adult patients. This diverse dataset also includes control recordings of non-inflamed appendices and is enriched with extensive clinical metadata and crucial annotations of the intraoperative grade of appendicitis. By offering this high-quality, representative data, Appendix300 enables new, clinically relevant validation tasks for computer vision. It significantly enhances the breadth and translational relevance of AI-based surgical video analysis, pushing the field beyond single-institution and single-procedure benchmarks.

Federated Learning for Surgical Vision in Appendicitis Classification: Results of the FedSurg EndoVis 2024 Challenge [2] looks into how well current federated learning (FL) methods can classify appendicitis inflammation from surgical videos without sharing private patient data. The challenge benchmarked various strategies on a new multi-center dataset, evaluating both generalization to unseen hospitals and performance after local fine-tuning. Results showed that while generalization to new centers was limited, all approaches improved significantly after local adaptation, with spatiotemporal models proving most effective. This work establishes the first benchmark for FL in surgical video, highlighting key challenges like class imbalance and the trade-off between personalization and global robustness, thereby setting a reference point for future development of privacy-preserving clinical AI.

The paper "One model to use them all: Training a segmentation model with complementary datasets" [3] shows how to overcome the critical need for massive, fully-annotated datasets in surgical AI. Our new method intelligently combines multiple, partially labeled datasets to train a single, comprehensive AI segmentation model. We successfully trained one model to identify numerous different anatomical structures from separate incomplete datasets, which not only significantly improved overall accuracy but also substantially reduced confusion between critical organs like the stomach and colon. This breakthrough demonstrates a new path forward, enabling researchers to build more powerful surgical AI systems faster by leveraging the complementary data that already exists.

The FL-EndoViT: Pretraining Vision Transformers via Federated Learning on Endoscopic Image Collections [4] paper shows how to build powerful surgical AI foundation models without violating data privacy regulations. It introduces a federated learning (FL) approach, FL-EndoViT, which allows multiple institutions to train a model collaboratively without sharing sensitive patient data. By using a special adaptive optimizer, this method successfully overcomes the challenges of data differences between hospitals, a problem that caused standard FL to fail. The results prove that this privacy-preserving model performs comparably to a centralized model trained on all data, establishing a viable framework for creating robust, scalable surgical AI.

The developed demos in the NEARDATA project showcase a powerful semantic search engine, StreamSense, for surgical video archives. This system combines various existing frameworks like Pravega, GStreamer, Milvus, Python, PyTorch, and different AI Models. This system uses AI to create a "digital fingerprint" for video frames, building a highly efficient and scalable index to find relevant clips. Surgeons and data scientists can now instantly query vast video libraries for specific events or anatomies, which dramatically accelerates AI training by eliminating the need to download entire datasets. This breakthrough unlocks the potential of surgical data, paving the way for advancements in AI-assisted surgery and training to improve patient outcomes.

In parallel, the project validated high-security AI training by combining Federated Learning with Trusted Execution Environments (TEEs). Using the SCONE platform, adversarial experiments showed that even privileged attackers with full memory access could not read sensitive data from the training process. This approach, enhanced with features like network shielding, dynamic attestation,

and confidential orchestration, provides a robust, end-to-end secure environment for collaborative AI model development, ensuring that data, code, and models remain confidential and unmodified throughout the training workflow.

3.2.3 Economic impact: Driving Innovation through Industry Collaboration

Our innovations are driven by strong collaborations with leading industry partners. We've forged valuable ties with the ZEISS Group, particularly their Health Solution & Central Marketing division, bridging the gap between research and industry application. A close cooperation with DELL and SCONTAIN during the NEARDATA project was critical, facilitating significant resource sharing that accelerated our research and development. This success has directly led to discussions for a sustained partnership, with a future DKFZ-DELL research project now being proposed to tackle key challenges in cancer research.

Our research outcomes are already driving innovation in other major projects, such as the Surgical AI Hub Germany[48] and new initiatives within the broader Surgical Data Science community. Within the Hub, AI is being leveraged to scale the assessment, improvement, and assurance of surgical quality. This impact is amplified by strong collaborations with industry partners, including Karl Storz SE & Co. KG, Chimaera GmbH, mbits imaging GmbH, and formigas GmbH. We anticipate our Federated Learning results will significantly accelerate model development, particularly for training powerful foundation models across multiple institutions within the European Union.

3.2.4 Pioneering Health Data Spaces and Open Science

In a move to foster open science and contribute to the growing landscape of health data spaces, the Appendix300 dataset will be made publicly available. This initiative provides a valuable, high-quality dataset for researchers worldwide. Complementing this, NCT maintains a repository of open-source code, further enabling the scientific community to build upon their work and accelerate the development of new technologies.

3.2.5 Influencing Industry and Future Standards

Interactions with industry leaders are crucial for translating research into real-world applications and influencing industry standards. The ongoing networking and exchange with ZEISS are instrumental in shaping the next generation of surgical technologies. The collaboration with DELL and SCONTAIN on the NEARDATA project is a testament to the value of shared resources and expertise in pushing the boundaries of what's possible. The potential for a follow-up DKFZ-DELL research project highlights the long-term commitment to this collaborative approach.

Looking towards the future, NEARDATA's research outcomes from the collaboration between DKFZ and DELL explore promising technologies that are not yet standard in hospital environments but hold immense potential. The use of Pravega for robust, scalable, real-time data streaming and GStreamer for flexible video processing pipelines are being investigated as foundational elements for the operating room of the future. While these technologies are not yet ready for widespread clinical deployment, they offer capabilities that current solutions do not support.

Furthermore, Federated Learning, particularly using the Flower (flwr)[5] framework, is being explored as a groundbreaking approach for training surgical foundation models. This method allows for the collaborative training of models on data from multiple institutions without the need to share sensitive patient data, addressing critical privacy concerns. While further research into areas like regularization is needed, federated learning presents a viable and powerful tool for developing robust and generalizable AI models for surgery.

Regarding the adoption of these innovations by global standards bodies, while there is no record of formal adoption into reference stacks at this stage, the commitment to open-source data and code, coupled with close collaborations with major industry players, represents a critical pathway toward influencing and helping to define the future standards for computer-assisted surgery and surgical data science.

[48] `www.saihg.de`

3.3 Use Case: Metabolomics

The Metabolomics Use Case demonstrates how a federated, scalable and privacy-preserving ecosystem can be built on top of heterogeneous metabolomics repositories, combining the capabilities of DataCockpit, DATOMA and PyRun. By addressing fragmentation, scalability constraints, and confidentiality requirements, this use case delivers a concrete implementation of NEARDATA's vision for unified, FAIR-aligned and computation-ready biomedical Data Spaces. The impact of this work spans scientific, societal, economic, and standardization domains, and has been positively received by communities working on international data spaces, cloud-native analytics, and future omics data standards.

3.3.1 Societal Impact

The Metabolomics Use Case provides societal impact by accelerating access to high-quality metabolomics data and enabling faster biomedical research workflows. Through the federation of key repositories—METASPACE, MetaboLights, Metabolomics Workbench, and AWS Open Registry—the use case reduces fragmentation and democratizes access to open molecular data resources. Clinicians, researchers and public health actors benefit from faster processing pipelines and improved reproducibility, supporting advancements in diagnostic methodologies and population-scale studies. The integration of privacy-preserving computation further strengthens public trust by ensuring that sensitive biomedical information can be processed securely across jurisdictions.

3.3.2 Scientific Impact

Scientifically, the use case advances the state of the art by providing an interoperable and reproducible framework for large-scale mass-spectrometry imaging (MSI) analysis. DataCockpit enables standardized dataset discovery and automated partitioning; DATOMA provides repository integration and application availability; and PyRun orchestrates scalable workflows across cloud infrastructures. These components collectively enable researchers to perform reproducible analyses with elastic resource scaling and transparent data provenance. The incorporation of Trusted Execution Environments (TEEs) demonstrates NEARDATA's contribution to confidential computing for biomedical pipelines, opening a new research direction for privacy-preserving multi-omics processing.

3.3.3 Economic Impact

Economically, the use case demonstrates the viability of serverless and cloud-native metabolomics pipelines that minimize operational expenses by scaling resources only when required. By offering reusable, modular pipelines through PyRun and app-store-like access to data through DATOMA, the use case reduces entry barriers for SMEs, laboratories, and biotech companies. This modular ecosystem fosters innovation, decreases infrastructure costs and supports the emergence of commercial and open scientific services built upon unified metabolomics data. The cost-efficient design aligns with European strategic goals for sustainable scientific computing and digital competitiveness.

3.3.4 Health Data Spaces

The NEARDATA Metabolomics Use Case contributes directly to the development of international Health Data Spaces, aligning with WP2 objectives and KPI-5. By harmonizing access to heterogeneous metabolomics repositories through federated connectors, the use case provides a concrete example of how FAIR-compliant biomedical data can be exposed within a distributed Data Space. The design echoes the architectural principles promoted by the International Data Spaces Association (IDSA), especially regarding interoperability, secure data exchange and governance. Additionally, the integration of confidential computing supports future cross-border biomedical workflows by enabling privacy-preserving analytics. These contributions position the metabolomics federation approach as a reusable blueprint for upcoming European Health Data Space (EHDS) initiatives and other international health data collaborations.

3.3.5 Industry Standards

The use case advances industry standards by introducing reusable patterns for API-driven data access, cloud-optimized data partitioning, and portable containerized workflows. DataCockpit's ingestion and chunking mechanisms provide a repeatable methodology for preparing large omics datasets for distributed computation, while DATOMA demonstrates a standardizable abstraction layer for integrating multiple metabolomics data providers. PyRun's modular pipeline packaging aligns with industry expectations for reproducible, shareable and cloud-portable workflows. External interest from research groups, cloud-service providers and data-management communities demonstrates the broader relevance of these outcomes. Moreover, several components have been recognized by international data-spaces communities as potential contributions to reference stacks in cloud-native scientific processing.

3.3.6 Future Standards

The architecture of the NEARDATA metabolomics ecosystem demonstrates readiness for emerging international standards in multi-omics data representation, including those inspired by ISO/IEC 23092 (MPEG-G). Although originally developed for genomic information, MPEG-G principles—such as data streaming, selective access, privacy enforcement, metadata association, compressed file concatenation, and incremental metadata updates—are directly applicable to large metabolomics datasets. The federated Data Space supports partition-aware access, distributed ingestion, structured metadata alignment and selective privacy controls, making it well positioned to adopt or influence future standardization efforts in biomedical data compression and streaming. The integration of TEEs anticipates evolving requirements for selective encryption and secure computation, ensuring long-term adaptability as global standard bodies incorporate privacy-preserving computation into their reference architectures.

3.4 Use Case: Transcriptomics Atlas Use Case

Transcriptomics Atlas groups the uniformly processed data from a representative set of human tissues. This resource can be of use in a wide range of scientific applications, including pharmacogenomics and biomarker discovery, where transcriptomic analyses are often performed in a comparative framework, examining different health states, diseases, or stimuli relative to baseline conditions.

3.4.1 Societal Impact

Transcriptomic Atlas can be used as a reference database for future clinical practice when gene expression profiles will be used in disease diagnosis and treatments. The use of the Transcriptomics Atlas should result in reduced (monetary and operational) costs of transcriptomics and wet lab experiments. Moreover, faster results in the processing of RNA-sequences with the STAR aligner can improve medical services if STAR is used for diagnosis by medical specialists. This, in turn, has impact on the more accurate and timely diagnosis and treatments, which is key for development of healthcare systems in modern society.

3.4.2 Scientific Impact

In this project multiple publications have been published and presented in international conferences. The publications are presented below:

- Bader J. et al. (incl. Kica P., Lichołai S., Malawski M.), "Novel Approaches Toward Scalable Composable Workflows in Hyper-Heterogeneous Computing Environments," SC-W '23: Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis
  `https://doi.org/10.1145/3624062.3626283`
- P. Kica, S. Lichołai, M. Orzechowski, M. Malawski, "Optimizing Star Aligner for High Throughput Computing in the Cloud," 2024 IEEE CLUSTER Workshops
  `https://doi.org/10.1109/CLUSTERWorkshops61563.2024.00039`

- Kica, P., Orzechowski, M., & Malawski, M. (2025). Serverless Approach to Running Resource-Intensive STAR Aligner. CCGrid2025
  https://doi.org/10.1109/CCGridW65158.2025.00039
- P. Kica, S. Lichołai, M. Orzechowski, M. Malawski, "Accelerating Cloud-Based Transcriptomics: Performance Analysis and Optimization of the STAR Aligner Workflow," ICCS2025
  https://doi.org/10.1007/978-3-031-97635-3_31

Within the project, multiple optimizations were introduced to speed up processing and reduce costs. The optimizations implemented for the Transcriptomics Atlas pipeline can be applied to other bioinformatics pipelines and compute environments. This experience will thus have broader impact on other scientific domains which require scientific computing.

3.4.3 Economic Impact

The combined impact of optimizations within the Transcriptomics Atlas reduced the total compute time and cost by two orders of magnitude, which is very significant. For example, the early stopping feature helps to prevent computing of input files that would result in a low mapping rate and, therefore, increases the overall throughput by about 20% and reduces costs by a similar amount.

The list of optimizations is presented below:

- 20% reduction with early stopping feature
- 50% cost reduction with spot instances
- 25% cheaper and faster with the newest processors on EC2 (r7a.2xlarge) compared to the older generation
- 11× cheaper and faster with the newer STAR index (release 111, toplevel type from Ensembl)
- Efficient STAR index distribution solution for worker nodes

Combined optimizations allow for processing RNA-sequences cheaply and quickly, which results in in larger Transcriptomics Atlas and reduced cloud spending.

3.4.4 Health Data Spaces

The data used in the Transcriptomics Atlas are publicly available RNA-sequences obtained from National Center for Biotechnology Information (NCBI) and European Nucleotide Archive (ENA). Those combined have collected tens of petabytes of sequencing data from which we selected SRA/-FASTQ files related to human tissues. By processing the input dataset with the Transcriptomics Atlas pipeline, we created a new dataset — the Transcriptomics Atlas — that consists of hundreds of high-mapping-rate files corresponding to the given tissue. The Atlas itself will be useful for researchers in the Transcriptomics domain, e.g., to reduce wet lab experiments, and is planned to be published on the Zenodo platform.

3.4.5 Industry Standards

The solution created for this project was influenced by tools widely used in the industry, such as Terraform for infrastructure management and Docker for portability and improved build workflow. Moreover, we adhered to standards for the cost-efficient use of cloud services (*AWS Well-Architected Framework* - Performance efficiency and Cost Optimization pillars). Also, cloud-related optimizations, such as spot instances, are a recommended way to reduce costs for batch processing scenarios.

3.4.6 Future Standards

One of the crucial optimizations developed within the Transcriptomics Atlas is the early stopping feature, which keeps track of the mapping rate of the currently processed FASTQ file. This is possible due to the STAR aligner's feature of providing such a metric; however, alternative (pseudo)aligners such as Salmon do not yet support this. This finding shows that not every aligner has this ability; therefore, it should be covered in order to be compatible with the early stopping feature.

Another development which can contribute to the future standards was implementation of SRA extension for DataPlug, which allows for on-the-fly partitioning of SRA files for distributed computation. The solution enables possibility of generalizing to other compressed data formats as well. This will enable new data types to be used with the advanced data partitioning technology designed for modern cloud infrastructure.

3.5 Use Case: Pathogen Genomics

As demonstrated by recent global pandemics and epidemics (such as COVID-19, Mpox, and the current worldwide surveillance of avian flu), routine genome sequencing of pathogens coupled with downstream analysis has become an indispensable technical tool to track and control emerging pathogens. A number of applications based on genomics have become standard tools in the field of public health and even household names sometimes, good examples being variant tracking for viruses and the surveillance of anti-microbial resistance (AMR) in bacteria. Genomic-based profiling of pathogens is becoming commonplace in the official regulatory frameworks for a number of medical and industrial applications, for instance when enforcing or investigating food security. As a result, UKHSA generates an impressive volume of sequencing data that needs to be stored, analyzed and made ready for use in downstream applications — for instance, the *Salmonella* national reference laboratory, which is just one among many hosted at UKHSA, sequences roughly 10,000 samples every year, equivalent to 5-10 TB of raw sequencing data. Finally, UKHSA is a complex federated system putting together a number of actors — doctors, hospitals, the NHS, laboratory operators, bioinformaticians, epidemiologists and decision makers.

3.5.1 Societal Impact

The social impact of the UKHSA use case is perhaps the easiest one to argue, as a better provision of technological solutions in the field of pathogen genomics directly translates into better public health. This is due to a number of reasons, such as the possibility to implement better analysis workflows for the same price, more resilience due to the use of more robust and scalable infrastructure, and a better overall information flow resulting in improved coordination, management and decision making when public or personal health are at stake.

This is well illustrated by a number of results obtained by the project. The KPop algorithm for genome comparison and classification will provide better databases and more informative results to the reference labs operating at UKHSA; use of data connectors can offer faster access to data closer to storage, resulting in more throughput and reduced costs; and stream-based, event-driven workflows will enable all the actors operating in the UKHSA ecosystem to interact seamlessly and follow progress of data analysis in real time.

3.5.2 Scientific Impact

UKHSA's participation in NEARDATA has resulted in several direct technological innovations and a number of scientific publications. In particular:

- "Scaling a Variant Calling Genomics Pipeline with FaaS" [6] describes a serverless genomics pipeline that can be used as a template for calling variant on large-scale datasets. Note that a number of workflows for pathogen genomics, including most techniques used to establish anti-microbial resistance (AMR), do rely on variant calling as the first step.

- "`KPop`: accurate and scalable comparative analysis of microbial genomes by sequence embeddings" [7] presents an innovative way of comparing microbial genomes by embedding them

into a virtual space of moderate dimensionality. This paves the way to establish a robust algorithm and vector databases that can efficiently be queried to quantify the magnitude of outbreaks and the speed of pathogen evolution. Recognizers and classifiers can also be easily set up by applying standard machine- and deep-learning techniques to the vector embeddings produced by `KPop`.

- Other papers describing in more detail the development of adaptors and data connectors to manipulate sequencing data in the cloud and at the edge (for instance [8]). It should be noted that they solve common problems that are likely to be encountered by most bioinformatic workflows, which makes them interesting as building blocks even beyond NEARDATA.

3.5.3 Economic Impact

The KPIs demonstrated during the implementation of the UKHSA use case (such as `KPop` and its re-implementation in terms of facilities offered by the NEARDATA platform) have the potential to lead to a significant reduction in the computational time and resources required to run complex genomic workflows.

If consistently adopted, this will have a very measurable impact not only by improving scalability, by providing shorter response times and by democratizing access to pathogen genomic workflows; it will also represent a much greener approach to what are usually compute- and memory-hungry algorithms. Ultimately, this will translate into significant savings of taxpayer's money, which is one of the many fronts on which UKHSA strives to improve its everyday performance.

3.5.4 Health Data Spaces

Most of the sequencing data produced by UKHSA can be found on the NCBI Short Read Archive — at the moment there are roughly 50,000 non-SARS-CoV-2 samples deposited there. This represents a significant resource and positions the UK as perhaps the second largest contributor of pathogen genomics data worldwide after the US. While guidelines encourage data produced with public money to be uploaded to public repositories, in this case doing so is made easier by the fact that strictly speaking pathogen genomics data is not subject to the most stringent regulations that apply to human data. So, after human reads have been filtered out in order to avoid possible re-identification, most of the samples can be uploaded to public repositories. Having the sequences on the SRA is especially convenient as the database is mirrored in the cloud and can be accessed directly from AWS. For this project, we have reanalyzed some of these sequences.

3.5.5 Industry Standards

Better data connectors for the manipulation of sequencing data, such as Pravega's byte-oriented streams [9] to ingest FASTQ files and run cutting-edge methods such as `KPop` on them, can enable streaming analytics while maintaining efficient, tiered storage for batch processing. They allow FASTQ data produced by sequencers to be accessed both via Pravega and directly from object storage.

There is a growing interest from several players in the field (notably NVIDIA and AWS) to provide a set of standardized tools to be used in the creation and deployment of workflows on popular platforms. Genomic data is especially in the focus due to its relevance to a number of fields that are currently experiencing a boom in terms of investment and number of users. The data connectors developed for genomics by NEARDATA represent good example of helpful and technologically meaningful solutions; hopefully they will find their way into these novel toolkits currently being developed by the leading companies in the field.

3.5.6 Future Standards

Since 2016, the UKHSA PI has been personally involved in the ongoing work of ISO/IEC 23092 [10] (also known as MPEG-G) on the standardization of Genomic Information Representation and Compression. In particular, he has served at different times as editor of several of its parts (notably part 3,

for which he received an ISO/IEC Excellence Award in April 2022), and has been the main proponent of the upcoming part 6 on genomic annotations.

This represents an interesting direct communication channel, and an opportunity for NEARDATA technological results to be fed back into MPEG-G work. We plan on proposing MPEG-G to evaluate the integration of facilities to perform data sketching and event-based streaming into future parts of the standard. Interestingly, MPEG-G already provides an architecture and design (in particular considering its streaming and transport capabilities) upon which the proposed additions might be built relatively easily.

4 Exploitation

Both commercial and non-commercial partners have already exposed their initial exploitation ideas in the previous deliverable "*D6.2 Communication and standardization report*". Commercial partners will now explain how they can convert the knowledge produced in NEARDATA into products and how this gives them a competitive advantage. Non-commercial partners will detail their plans and enrich their initial ideas on the exploitation of its own technologies; they will state the challenges and opportunities ahead; open room for new partnerships and developments, new research topics; advance systems and protocols for the benefit of the public interest; strengthen the European protagonism in the *omics industries and on the ever stronger AI adoption.

4.1 Dell's Exploitation Plans

Dell Technologies has played a strategic role in the NEARDATA project, contributing to the development and validation of advanced data streaming, serverless analytics, and AI-driven connectors tailored for biomedical and clinical workloads. The project's focus on real-time data processing, semantic search, and scalable AI infrastructure aligns directly with Dell's mission to deliver intelligent, edge-to-core platforms for data-centric industries. NEARDATA's outcomes offer Dell a unique opportunity to extend its infrastructure portfolio with validated, real-world solutions that address pressing challenges in healthcare, genomics, and surgical data science.

The NEARDATA project has produced several key innovations that are directly exploitable by Dell, including:

- *StreamSense*: A semantic video search engine built on Pravega and vector DBs, enabling content-aware retrieval of surgical video data. It supports inter- and intra-video queries with sub-30ms latency and up to 99.8% data transfer reduction for AI training (see D5.2).
- *FaaStream*: A serverless framework for unified batch and stream analytics, validated in real-time video ingestion and AI inference scenarios. It supports sub-12ms latency and up to 22.5× throughput improvements over traditional cloud-native setups (see D3.2).
- *Serverless Vector DBs*: A novel architecture for scalable AI workloads using stateless cloud functions, validated against Milvus and Vexless. It achieves 3.5× to 5.8× indexing performance improvements and 56% to 63% cost reduction (see D3.2).

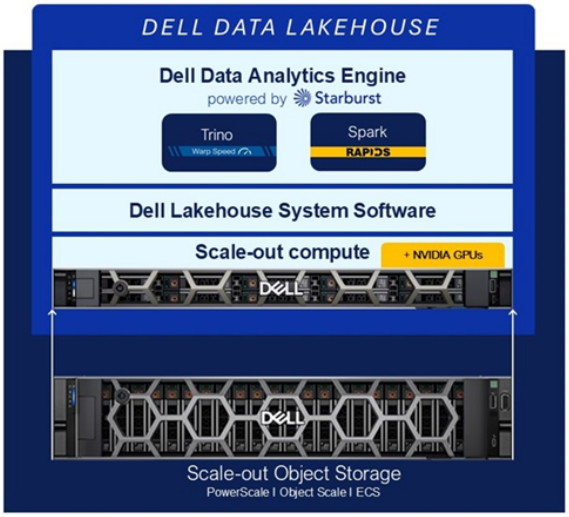
These outcomes are not only technically innovative but also validated in real-world use cases, including genomics, metabolomics, and AI-assisted surgery. They provide Dell with a strong foundation for productization and market differentiation.

4.1.1 Business Plan

Dell's exploitation strategy focuses on integrating NEARDATA's validated technologies into its infrastructure stack, particularly through the following products (see Fig. 19):

- *Dell NativeEdge*: NativeEdge is Dell's edge operations platform designed to deploy, manage, and secure workloads across distributed environments. It supports containerized applications, real-time analytics, and AI inference at the edge, making it ideal for latency-sensitive use cases

(a) Dell NativeEdge.



(b) Dell Data Lakehouse.

(c) Dell ObjectScale.

Figure 19: Dell product portfolio aligned with NEARDATA research outcomes.

such as industrial automation, retail, and healthcare. NativeEdge provides centralized orchestration, zero-touch provisioning, and integration with Kubernetes and other cloud-native technologies.

- *Dell Data Lakehouse*: Dell's Lakehouse architecture combines the scalability of data lakes with the performance and structure of data warehouses. It enables unified analytics across structured and unstructured data, supporting advanced AI/ML workloads, business intelligence, and real-time decision-making. The Lakehouse integrates with Dell ECS, ObjectScale, and external compute engines, offering flexibility and performance for enterprise-scale data operations.

- *Dell ObjectScale*: ObjectScale is Dell's software-defined object storage platform built for modern, cloud-native workloads. It supports S3-compatible APIs, multi-tenancy, and geo-distributed deployments. ObjectScale is optimized for scalability, durability, and performance, making it suitable for AI training datasets, video archives, and tiered storage solutions.

The convergence of NEARDATA outcomes with these products enables Dell to deliver AI-native, serverless-ready, and healthcare-compliant platforms. These are the avenues of exploitation we are currently exploring with Dell engineering business units:

- *StreamSense on NativeEdge*: StreamSense enables real-time semantic indexing and querying of surgical video streams. When deployed on NativeEdge, it allows hospitals and surgical centers to perform AI-assisted video search and retrieval directly at the edge, reducing latency and avoiding unnecessary data transfers to the cloud. This integration supports intelligent surgical assistance, training, and quality control in clinical environments.

- *Serverless Vector DB on ObjectScale*: The serverless vector DB architecture developed in NEARDATA can be deployed on top of ObjectScale, leveraging its scalable, S3-compatible object storage for storing and querying vector embeddings. This enables elastic, cost-efficient AI workloads for genomics, RAG pipelines, and biomedical search, while maintaining data locality and compliance with healthcare data regulations.
- *FaaStream for Dell Data Lakehouse*: FaaStream provides a unified serverless engine for batch and streaming analytics, ideal for processing large-scale datasets. Integrated into Dell's Data Lakehouse, FaaStream enables real-time ETL, AI inference, and data transformation pipelines across structured and unstructured data.

These integrations are being explored through a strategic partnership with NCT (National Center for Tumor Diseases), whose use case presents unique opportunities for Dell to co-develop validated designs and productize them in the medium term. The goal is to deliver Dell-certified platforms for AI-assisted surgery, genomics, and federated learning, aligned with International Health Data Spaces and EHDS initiatives.

Importantly, during the second half of the project we have increased the internal dissemination of NEARDATA outcomes across multiple Dell engineering and research teams. We have produced a set of internal studies and presentations which can be inspected upon review, as they are internal material that cannot be shared publicly.

4.1.2 Competitors

The NEARDATA outcomes provide Dell with a clear advantage over several categories of competitors:

- *Medical Storage Vendors*: Traditional vendors (*e.g.*, Illumina) offer proprietary systems with limited extensibility. Dell's use of open source platforms (*e.g.*, Kubernetes, streaming) and standard APIs (*e.g.*, S3-like API in ObjectScale) enables open, scalable, and AI-ready storage architectures, reducing vendor lock-in and enabling integration with standard hardware and software.
- *Cloud Platforms*: Hyperscalers offer scalable AI services but require onboarding sensitive data outside the owner's infrastructure domains. Dell platforms, such as NativeEdge, support hybrid deployments enabling low-latency inference and data sovereignty, which is critical for clinical environments.
- *Vector DB Providers (e.g., Milvus, Weaviate)*: Dell's serverless vector DB architecture can offer better elasticity, operational simplicity, and cost-efficiency for bursty workloads, with seamless integration into Dell's ObjectScale and Lakehouse platforms.

By aligning NEARDATA innovations with its infrastructure products, Dell is positioned to lead in delivering next-generation platforms for biomedical and clinical applications, combining performance, scalability, and compliance with real-world validation.

4.2 KIO's Exploitation Plans

KIO Networks aims to leverage the outcomes of the NEARDATA project to strengthen its cloud and data services portfolio, focusing on **sovereign, high-performance, and data-centric infrastructures**. NEARDATA's framework aligns with KIO's strategic objective of delivering **secure, compliant, and efficient cloud platforms** for European enterprises and public institutions.

4.2.1 Business Plan

Expected Exploitable Outcomes

KIO will focus on the commercial exploitation of the following outcomes from NEARDATA:

- **Data Management and Processing Frameworks:** Integration of NEARDATA's data federation and lifecycle management technologies into KIO's cloud offering, enhancing multi-domain interoperability.

- **Energy-Efficient and Sustainable Infrastructure Models:** Adoption of green data management principles developed in NEARDATA to improve the efficiency of KIO's datacenters.
- **AI/ML-Enhanced Data Governance:** Incorporation of intelligent data governance models from NEARDATA into KIO's data services for compliance and value extraction.
- **Pilot Use Cases and Demonstrations:** Commercial adaptation of NEARDATA pilots into replicable service blueprints for sectors such as Smart Industry, Public Administration, and Research.

Exploitation Strategy

KIO's exploitation strategy involves:

- **Integration:** Embedding NEARDATA technologies into KIO's cloud management platform (IaaS/PaaS/SaaS).
- **Commercialization:** Packaging NEARDATA-inspired data services as premium offerings for clients requiring **data residency, AI-readiness, and interoperability**.
- **Partnership Expansion:** Using NEARDATA as a showcase to establish partnerships with research centers and EU institutions for future Horizon Europe initiatives.
- **Market Differentiation:** Promoting KIO as a **trusted European cloud alternative** with strong data sovereignty guarantees.

Business Impact

By integrating NEARDATA outcomes, KIO expects to:

- Expand its service portfolio beyond infrastructure to **data-centric value propositions**.
- Increase its **European market share** through compliance with data sovereignty frameworks (e.g., Gaia-X, EU Data Act).
- Generate **new revenue streams** via AI-assisted data services and energy-optimized infrastructure management.

4.2.2 Competitors and Market Position

Competitive Landscape

KIO operates in a highly competitive cloud and data services market dominated by hyperscalers (AWS, Microsoft Azure, Google Cloud) and specialized European providers (OVHcloud, Scaleway, CloudFerro). These actors are increasingly investing in **data sovereignty and sustainability**, which are also key goals of NEARDATA.

Competitive Advantages from NEARDATA

Participation in NEARDATA grants KIO several strategic advantages:

- **Access to cutting-edge European research** on federated data management, unavailable to non-participant competitors.
- **Technology differentiation** through open, interoperable, and sustainable approaches, contrasting with proprietary hyperscaler models.
- **Alignment with EU digital sovereignty goals**, improving eligibility for public tenders and projects requiring compliance with European data governance.
- **Enhanced brand reputation** as an innovative, research-active company engaged in the EU digital ecosystem.

Long-Term Positioning

KIO will position itself as a **data sovereignty enabler**, bridging the gap between large-scale hyperscalers and regional providers, by offering:

- Transparent, compliant, and efficient data services.
- Integration of NEARDATA's methodologies into managed infrastructure solutions.
- Strategic participation in future European data spaces initiatives.

4.3 Scontain's Exploitation Plans

Scontain has brought or produced a wide range of products to NEARDATA. There are five core products brought to the consortium: *sconification* is our flagship, conversion of a native application to confidential container; *file shield*, transparent encryption and/or authentication of files and volumes accessible by attested applications; *network shield*, transparent encryption of network communication; *Local Attestation Service* – LAS, used to mediate attestation on TEE [49]; and *Configuration and Attestation Service* – CAS, providing secrets after remote attestation for applications running inside TEEs/enclaves. There are the cloud applications used to deploy sconified services: *SCONE Operator*, an aide plugin for Kubernetes used to deploy CAS, LAS, and Scontain's SGX plugin, and *SCONECTL*, a special program to setup confidential mesh of services. And there are the sconified services: *Lithops*, *MinIO*, *Keycloak*, *MariaDB*, *Flower ML*, and *Pravega Client*.

4.3.1 Sconified Services

Sconified services are what users will mostly interact with on their daily business. These services are supported by Scontain's mature core systems and cloud applications. Many systems have been sconified in NEARDATA:

- **Lithops**, a FaaS – function as a service platform to deploy computation on the cloud;
- **MinIO**, a cloud object storage used by Lithops for synchronization of division of work and actual storage;
- **Metaspace**, a spatial metabolomics molecular annotation for imaging mass spectrometry datasets that employs Lithops as its engine;
- **Flower ML**, collaborative machine learning and federated AI framework supporting secure multi-party computation;
- **Pravega Client**, a Pravega Rust Benchmark for the Pravega storage service for data Streams.

Furthermore, two other sconified services have been brought to cover the largest extent of the computation continuum in NEARDATA:

- **Keycloak**, an identity and access manager used by other applications to outsource user management;
- **MariaDB**, a transactional database supporting data persistence for Keycloak.

4.3.2 Business Plan

The "full picture" of Scontain's activities in NEARDATA shows that the maturity of confidential computing products that Scontain has available can support new categories of cloud-native systems like streaming data transfer, function as a service, and machine learning that is closer to artificial intelligence aspects, and also cover gray areas in confidential computing, such as reaching the user layer of authorization to use systems and their resources. Suites of matching products can be exploited to

[49] TEE - Trusted Execution Environment, e.g. Intel SGX, Intel TDX, AMD SEV SNP, etc.

offer a variety of functions inside confidential computing and enforce privacy according to industry standards or government requirements.

Keycloak and CAS enable the implementation of **ZTA – Zero Trust Architecture** paradigm where the focus is on end-to-end resource and data protection that continuously evaluates identity, credentials, endpoints etc. to grant access to the least and essentially necessary actions needed to perform the mission. With confidential Keycloak, Scontain can offer a Single Sign-On alternative that can be integrated into other applications using OpenID Connect libraries or RESTful commands.

Metaspace and Lithops present an excellent opportunity to advance confidential computing support in *omics areas. Metaspace can be better integrated with the new Lithops Singularity backend, benefiting from the division of work using *threads* instead of *forks*, thus making pipeline executions in confidential computing much faster. Furthermore, the addition of Keycloak into this suite allows for the privacy of data manipulated to be reassured by revealing this new layer of traceability and accountability. Users with exclusive and personal Passkeys configured in Keycloak and linked to their smart phones or USB security devices can ascend to sensitive data processing while the system ensures *(C) confidentiality*, *(I) integrity*, *(A) authenticity*, and *(N) non-repudiation*.

MinIO is another great product that is now part of our portfolio of sconified services. Being *S3 API* compatible and Open Source developed and released, MinIO is a high performance object storage that Scontain can exploit in many other industries, including AI[50]

Artificial intelligence is a technology trend that will only grow in importance and breadth, thus strengthening participation in it as early as possible is the smart decision to make. With the sconification of Flower ML, Pravega Rust Client, and MinIO, Scontain is now deeper in the AI realm with the presence of confidential computing. Scontain already has a presence in AI, with the confidential **secureTF**, a distributed secure machine learning framework based on TensorFlow employed for classification or inference tasks. However, with the new suite of systems produced in NEARDATA, Scontain can now offer products for training, learning, analytics, evaluation, and storage. Additionally, SCONE will soon support GPU integration, making the confidential computing in AI much more appealing to major players.

Scontain is plotting its future towards Zero Trust Architecture and Artifitial Intelligence.

4.3.3 Competitors

Competitors in confidential computing focus on 1. *porting applications*, 2. *cluster deployment*, and 3. *enclaves synchronization*. Companies all around the world, like *Edgless*, *Fortanix*, *Opaque*, and *Anjuna* offer products covering these features, with each of them using their own approaches. Scontain excels in all three areas with:

1. *sconification lift-and-shift* approach, where a native application is converted and wrapped in a Docker image;
2. *SCONECTL* used to deploy confidential mesh of services using the sconified images; and
3. *CAS – Configuration and Attestation Service* that synchronizes containers of sconified images deployed in a Kubernetes cluster, by enforcing the state of the TCB (trusted computing base) and delivering secrets exclusively to attested applications.

As the artificial intelligence industry heats up very quickly, products like "*Privatemode AI*" from Edgless; "*Armet AI*" from Fortanix, and *Confidential AI Platform* form Opaque are strong contenders. However, the expertise obtained from NEARDATA with the suite of AI/ML + FaaS products along with the ZTA suite, Scontain can face this challenge with great confidence.

Moreover, SCONE was selected as a "*great EU-funded Innovation*"[51] in the Innovation Radar initiative of highlighting excellent innovations in 2024. This gives the company an edge in searching for investments and partnerships.

[50] AI Storage is Object Storage – `https://www.min.io/solutions/object-storage-for-ai`

[51] A confidential computing platform for porting non-confidential applications and protecting their data (SCONE) `https://innovation-radar.ec.europa.eu/innovation/58831`

4.4 URV's Exploitation Plans

PyRun's core mission is to eliminate the "Cloud Complexity Tax" for the millions of Python developers, data scientists, and researchers who need scalable cloud power but lack specialized infrastructure expertise. The process of setting up, configuring, and managing cloud environments is notoriously complex, costly, and time-consuming, diverting focus from core innovation.

PyRun solves this problem with a "Bring Your Own Cloud" (BYOC) platform, providing a unified, automated, and effortless environment for running complex workloads on a user's own AWS or IBM Cloud account.

Our core value proposition is an all-in-one, VS Code-like web interface that combines code editing, automated runtime management, and one-click execution. This paradigm shift reduces complex environment setup times from days or weeks to under five minutes. Unlike competitors that often focus on a single technology, PyRun offers first-class, integrated support for a wide range of distributed frameworks—including **Dask, Lithops (FaaS), Ray, and Cubed**—catering to a much broader set of use cases and a larger market.

Our commercialization strategy is a product-led growth (PLG) model with a tiered SaaS subscription (Community, Professional, Team, Enterprise). This model ensures predictable revenue while offering unparalleled value to our users. A foundational principle is transparency: users retain full control and cost visibility by paying for their own cloud compute, while PyRun charges for the value-added platform features that drive productivity, collaboration, and governance. Our primary focus is the vast, underserved community of over **3 million Python developers** actively working with data and AI.

4.4.1 Opportunities

PyRun is strategically positioned to capture a significant market share by capitalizing on several key, demonstrable competitive advantages that directly address the primary pain points of our target users. Our opportunities are not just theoretical; they are backed by verifiable performance and pricing data.

- **Disruptive Cost-Effectiveness and Price Leadership:** Our operational efficiency and innovative business model allow us to provide access to high-performance GPU computing at a fraction of competitor and even direct cloud-provider costs. This is our most significant market differentiator, democratizing access to expensive hardware for advanced AI/ML workloads that would otherwise be out of reach for many users.

Table 6: GPU Pricing Comparison ($/hour) - PyRun vs. The Market

GPU Configuration	PyRun Cloud	AWS	Coiled	Modal	Savings vs. AWS
L4 x4	**$1.80**	$4.60	$3.40	$3.20	**61%**
A100 x8	**$9.68**	$21.96	–	$19.99	**56%**
H100 x8	**$20.72**	$55.04	–	$31.59	**62%**

Our Pricing Advantage Explained: PyRun's price leadership stems from our core business model. We are not a cloud reseller that profits by marking up compute costs. Instead, we charge a predictable SaaS fee for our value-added platform that orchestrates resources on the user's own cloud. For specialized hardware like GPUs, we leverage operational efficiencies to secure capacity at rates far below the on-demand market, passing these substantial savings directly to our users. This creates an undeniable value proposition: supercomputing-scale power without the supercomputing-scale cost.

- **Quantifiable Performance Superiority:** PyRun's architecture is not just more affordable; it is verifiably faster. Direct benchmarks against our closest competitors (Coiled) prove that our platform translates into tangible time savings and increased productivity.

- *FaaS (Serverless) Workloads:* Up to **14.9x cheaper and 5.4x faster** for equivalent data processing tasks (e.g., generating Kerchunk references).
- *Dask Cluster Workloads:* Up to **2.5x faster cluster creation**, leading to a **1.6x faster total time-to-result**. This acceleration is critical for rapid iteration in data science and ML development.

PyRun's Unmatched Competitive Edge

Our holistic synthesis of a seamless user experience, broad framework versatility, and demonstrable performance creates a powerful moat. Below is a direct comparison of our advantages over each competitor segment.

Advantage over Direct SaaS Competitors (e.g., Coiled): While Coiled is a strong product, it is a Dask-first, single-framework solution. PyRun is decisively superior on three fronts:

1. **Broader Framework Support:** PyRun offers first-class, integrated support for **Dask, Lithops, Ray, Cubed and many more in the future**. This versatility addresses a much larger Total Addressable Market and makes PyRun the ideal platform for teams with diverse computational needs, who would otherwise need multiple tools.

2. **Superior Performance and Cost:** As our benchmarks prove, for both Dask and serverless (FaaS) workloads, PyRun is significantly faster and more cost-effective. Faster iteration is a critical driver of developer productivity.

3. **Integrated User Experience:** PyRun provides a unified, VS Code-like IDE where users code, configure, run, and monitor in one place. Coiled's workflow is fragmented, requiring users to integrate a separate IDE with their web dashboard and CLI, leading to context-switching and reduced productivity.

Advantage over Cloud Provider Platforms (e.g., AWS SageMaker, GCP Vertex AI): These platforms are powerful but represent the very "Cloud Complexity Tax" PyRun was built to eliminate.

1. **Simplicity and Speed:** These are complex ecosystems that require specialized expertise and often lock users into a proprietary way of working. PyRun is lightweight and Python-native; users bring their existing code and can be running at scale in minutes, not days or weeks. For many teams, provider platforms are "overkill."

2. **Freedom from Vendor Lock-In:** PyRun is a multi-cloud abstraction layer. We provide a consistent, user-friendly interface that runs on AWS and IBM Cloud today, with GCP and Azure on the roadmap. This gives our customers the freedom to choose the best cloud for their needs without having to re-learn a new platform, a critical advantage for modern enterprises.

Advantage over DIY / Open-Source Toolkits (e.g., Nebari): The "Do-It-Yourself" approach offers maximum flexibility but incurs the highest hidden costs in terms of time, complexity, and ongoing maintenance.

1. **Total Cost of Ownership (TCO):** The cost of manually setting up, securing, and maintaining a distributed computing environment is enormous. It requires weeks or months of a skilled (and expensive) engineer's time. PyRun eliminates this entirely. Our value proposition is crystal clear: *Nebari is for teams who want to build their own platform; PyRun is the platform you buy so you don't have to.*

2. **Focus and Opportunity Cost:** By abstracting away infrastructure, PyRun allows teams to focus on their core mission: innovation. The time saved translates directly into faster projects, quicker breakthroughs, and a higher ROI on data and AI initiatives.

4.4.2 Challenges

While the opportunities are significant, we have identified several challenges and have formulated clear strategies to address them.

- **Market Awareness and Brand Building:** As a new entrant, PyRun must build brand recognition against established names like Coiled and Databricks. Our go-to-market strategy is designed to mitigate this by demonstrating undeniable value. We will focus on:
  - *Aggressive Content Marketing:* Publishing high-quality tutorials, technical blog posts, and our compelling benchmark comparisons.
  - *Active Community Engagement:* Establishing PyRun as a thought leader in relevant communities (e.g., PyData, Dask/Ray Discords, Reddit) by offering genuine help and expertise.
  - *Product-Led Growth:* Leveraging our powerful free Community tier as our primary marketing engine, enabling users to experience a real-world win in under 15 minutes.
- **Competition from Incumbent Cloud Providers:** Hyperscalers (AWS, GCP, Azure) possess the resources to simplify their own complex offerings (e.g., SageMaker, Vertex AI). PyRun's strategic defense is our multi-cloud and framework-agnostic DNA. We offer users freedom from vendor lock-in—a critical and growing concern for modern enterprises—by providing a consistent application layer across different clouds.
- **Scaling User Acquisition and Support:** A successful PLG model requires a world-class, low-friction onboarding experience and an effective support system that can scale efficiently. Our priority is to invest in comprehensive documentation, video tutorials, and a responsive, community-driven support channel (Discord) to successfully convert free users to our paid tiers.
- **Team and Resource Constraints:** As a lean, expert-driven team, we must maintain a sharp focus on executing our strategic roadmap. The primary challenge is balancing the rapid development of new, high-demand features (e.g., first-class GCP/Azure support, integrated MLOps capabilities) with the sales and marketing efforts required to build market traction and establish PyRun as the default platform for Python-based cloud computing.

4.5 TUD's Exploitation Plans

Technische Universität Dresden (TUD) has been a key driver in NEARDATA's security research, leading the design, implementation, and validation of advanced modules for secure, policy-driven confidential data exchange. These outcomes provide TUD with a strong foundation for academic leadership, industry engagement, and impact on future standards.

TUD's main exploitable results include:

Academic Papers & Conferences TUD has contributed several key research results to NEARDATA, published in leading international venues:

- **CRISP:** Robust, hardware-backed storage security module ensuring confidentiality, integrity, and rollback protection for persistent data.
  Published at: IEEE Symposium on Security and Privacy (IEEE S&P 2024)
- **LLD (Last-Level Defense):** Enclave-based runtime providing application-level integrity and memory protection for sensitive workloads.
  Published at: ACM Conference on Computer and Communications Security (ACM CCS 2024)
- **TICAL:** Trusted compilation and attestation pipeline guaranteeing only verified, reproducible code is deployed and executed.
  Published at: IEEE Symposium on Security and Privacy (IEEE S&P 2024)

- **SinClave:** Singleton enforcement framework for critical TEE-based services, preventing state forking and ensuring service authenticity.
  Published at: arXiv preprint (2023)

- Comprehensive Study on the Impact of Vulnerable Dependencies on Open-Source Software.
  Published at: 2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)

Industry Collaboration Papers

- **Latency-Security Tradeoff Analytics:** TUD led empirical studies and tuning methods for balancing TEE-based security with real-time analytics performance, collaborating with industry partners to ensure practical relevance.

Workshops & Webinars

- **Telecom TV and Vodafone Events:**
  Presented on the importance and application of confidential computing in telecom, highlighting secure data processing for AI-driven network analytics during international industry broadcasts and expert roundtables.

- **Supermicro Collaboration:**
  Participated in hardware-focused workshops and technical discussions on building secure computing infrastructure, emphasizing confidential computing solutions for safe and efficient AI data processing.

- **Bank for International Settlements (BIS) Forums:**
  Contributed expertise to panels addressing data confidentiality, privacy, and regulatory compliance in financial systems, with a focus on confidential AI and secure computation for sensitive financial analytics.

- **Confidential Computing Summit:**
  Presented technical advances in confidential computing for AI applications and engaged in cross-sector networking sessions to foster industry adoption and discuss future trends in secure artificial intelligence.

Open-Source Contributions TUD has contributed code, tools, and documentation to open-source repositories, accelerating adoption and reproducibility of secure data processing solutions, including core NEARDATA modules and demonstrators for confidential data exchange.

4.5.1 Opportunities

The NEARDATA project offers TUD several promising exploitation avenues:

- **Academic Impact and Training**: Incorporation of NEARDATA results into graduate and postgraduate curriculum, including a new confidential computing course and hands-on labs with SCONE and TEEs.

- **Open Science Leadership**: Release of research artifacts, benchmarks, and demonstrators under open-source licenses, promoting reproducibility and community collaboration.

- **Industry Collaboration & Technology Transfer**: Joint pilots and demonstrators with industry partners (e.g., Scontain, Dell), fostering adoption in healthcare, finance, and regulated sectors.

- **Standardization & Policy Influence**: Participation in international standards efforts, shaping future attestation and secure data exchange frameworks.

- **Commercialization Pathways**: Providing migration guides and reference implementations to support enterprise adoption of TUD's security modules in real-world deployments.

4.5.2 Challenges

TUD also recognizes several challenges to effective exploitation:

- **Sustainability**: Long-term maintenance and active community support for open-source modules require dedicated resources and ongoing engagement.
- **Regulatory Alignment**: Ensuring continuous compliance with evolving privacy and security regulations, especially in cross-border data exchanges.
- **Educational Outreach**: Bridging the skills gap in confidential computing through effective curriculum integration and training programs.

By strategically leveraging NEARDATA's validated security modules and research outputs, TUD will strengthen its leadership in confidential computing, foster innovation through education and collaboration, and accelerate the adoption of trustworthy data brokering solutions in academia and industry.

4.6 BSC's Exploitation Plans

HPC users, especially people from the fields of life sciences, earth sciences, among others, typically do not have deep knowledge of how supercomputer infrastructures work. They also lack a deep understanding of how to optimize embarrassingly parallel workloads. Although they do have some knowledge, HPC infrastructures are commonly managed with tools like Slurm. Such tools request the user to provide information as the number of CPUs desired, wall clock time, etc. An example of a simple Slurm command HPC users need to provide could look like as following:

```
sbatch --cpus-per-task=4 --mem-per-cpu=1  --ntasks=128 my.bash -W 02:00:00
```

If the parameters are not properly set, the workload will crash, resulting in a waste of time and resources. To avoid this inconvenience, users typically request more resources (CPUs, memory, and wall time) than needed. Moreover, many times they make tests first using interactive sessions, which do not end whenever the workload ends, increasing even more the resource wastage.

In contrast, our proposition allows for managing infrastructure and running data analytics workloads leveraging HPC infrastructures with commands as simple as:

```
lithops runtime deploy hpc
python workload.py
```

Such an easy and simple interface reduces the learning curve for users, regardless of their background. Moreover, it can be further diminished via integrating it all into a Slurm script, such as the following example:

```
sbatch --parsable myworkload.slurm -A cns102 -q gp_resa
```

4.6.1 Opportunities

A growing trend in HPC user accessibility, aimed at simplifying programming, is driven by the increasing adoption of Python in supercomputing environments [11]. Understandably, most tools and libraries for scientific data processing and machine learning are available in Python, bringing them closer to non-expert users and allowing more people to explore them. Tools such as Open OnDemand [12] confirm this need by enabling a Jupyter Notebook interface to a supercomputer. Furthermore, this is motivated by recent initiatives like AI factories [13] in the EU and the National Artificial Intelligence Research Resource (NAIRR) [14] in the USA, designed to open supercomputers to the general public to drive innovation. The downside is that they do not provide an easy way for these non-experts to easily leverage supercomputing resources, resulting in ineffective utilization.

Frameworks such as Pegasus [15], COMPSs [16], Parsl [17], and cluster-based dataflow engines such as Dask [18], Spark [19], Flink [20], or Ray [21] aim to simplify these complexities for end-users. However, they still require users to manage resource allocation through the cluster manager (eg. SLURM). This process can be challenging, as users must specify parameters such as the number of cores, GPUs, and wall clock time.

For non-expert users, making these decisions is not straightforward, and incorrect estimations can lead to workloads either crashing due to insufficient resources or over-allocating resources. The latter scenario is particularly problematic, as it not only wastes valuable infrastructure resources but also prevents other users from accessing them, creating inefficiencies across the system.

BSC has developed the Lithops-HPC architecture [22], based upon the Lithops Serverless framework developed by [23]. The idea behind it is to combine the simplicity of the Serverless approach, which hides the complexities of infrastructure management while allowing for highly parallel workloads to run smoothly and efficiently on HPC resources. Typically, Serverless runs in the cloud, where resources and networking are not as high-performing as it is in HPC environments due to the costs. Cost is an important factor in the cloud, while in HPC, the priority is performance over cost.

Lithops-HPC brings the opportunity to combine both simplicity and performance. And achieves the following:

- **Simplified framework for HPC**: Lithops-HPC provides a data analytics framework with a familiar, simple, functional API to run massively parallel programs by coding simple functions, leveraging the Serverless paradigm.
- **Simplified management of HPC infrastructure**: applying the Serverless paradigm up to **2x** according to Yaqin's [24] and Cyclomatic's [25] metrics.
- **Improved data ingestion and processing capabilities**: we address key challenges in merging serverless and HPC, enabling non-experts to manage complex resources and execute large-scale workloads seamlessly. We demonstrated this via achieving a **24x** speed-up performance using our Genomics Epistasis use-case, which led to processing much more genomic data in order to unravel the variant interactions in the human genome that can lead to Type 2 Diabetes.
- **Auto-scaling mechanism**: we initiated a research prototype, Function as a Service Time Series (FaaSTs), aiming to allow for Lithops-HPC resources to scale according to current usage of the underlying HPC infrastructure. This brings an opportunity to improve time to service: users do not need to wait for resources to become available, rather start running immediately according to current offerings, and the framework scales transparently when more resources are available.

Those opportunities were not met by the aforementioned frameworks targeting the HPC simplified use experience. While they made it easier for developers to code parallel workloads in HPC, the infrastructure management was still a tedious job to be done by the end-user. Our framework simplifies this aspect as well, making a significant contribution to the current state-of-the-art data analytics frameworks.

4.6.2 Challenges

Despite the achievements, there are some challenges we still face in reaching a wider audience:

- **Performance overhead**: despite leveraging HPC resources effectively, the fact that the initial concept of the framework was targeting cloud environments implies some performance overheads. Currently, Lithops-HPC is far from achieving raw performance as MPI frameworks do. Thus, users needing to prioritize performance over simplicity may encounter Lithops-HPC not suitable for them. One of the challenges is consequently to reduce the overhead and get closer to the MPI-based frameworks' performance.
- **Multi-tenancy**: despite the Lithops-HPC design allowing for sharing resources across users, currently there is no isolation among them. Thus, although technically possible to share, there

may be security concerns. One of the steps to address in the future is to find a means to provide such isolation so multiple users can share the same Lithops-HPC backend transparently and trouble-free.

- **Programming language**: currently Lithops-HPC is entirely based on Python. However, the underlying idea and schema could be implemented on languages more fit for the HPC, such as C or C++, or even Fortran, which is a common language across earth scientists. Thus, a future step towards ensuring a wide audience is to support C and Fortran as well.

- **Interfunction communication**: due Lithops-HPC is based on the serverless paradigm, it means communication between functions is not possible. Although there are many approaches already being developed towards this end, we need to adapt them to Lithops-HPC. Not having this capability impedes to leverage our framework in some use-cases, while in others means re-reading data twice, which is inefficient.

4.7 NCT's Exploitation Plans

The exploitation strategy of the National Center for Tumor Diseases (NCT) builds on the substantial scientific, technical, and clinical advancements achieved within the NEARDATA project. The developments in computer-assisted surgery is ranging from multi-center datasets and Federated Learning (FL) workflows to secure AI training environments or semantic video search. This range is providing a strong foundation for long-term impact across healthcare, industry, and scientific research.

NCT's exploitation approach focuses on translating these research outputs into sustainable, real-world value by strengthening clinical adoption, enabling industry collaboration, and contributing to emerging European standards for surgical data science. By combining open data resources, privacy-preserving AI technologies, and digital operating room innovations, NCT aims to accelerate the development of clinically robust, interoperable, and secure AI systems for surgical applications.

The following sections outline the key opportunities for exploiting these results as well as the major challenges that must be addressed to ensure their successful translation into practice.

4.7.1 Opportunities

Building on the technical and scientific outputs described above, NCT identifies the following main opportunities for exploitation:

- **Establishing Foundational Infrastructure for EU Surgical AI**
  The project's developments, particularly the Appendix300 dataset, FL-EndoViT, and StreamSense, lay the groundwork for an emerging pan-European infrastructure for surgical AI. By providing a multi-center, richly annotated dataset and validated FL pipelines, partners can position themselves as reference providers for interoperable, privacy-preserving surgical data architectures. This is in a great way aligned with EU initiatives such as the European Health Data Space (EHDS) and future AI standardization frameworks.
  The ability to train robust models without sharing patient data directly addresses longstanding privacy and security barriers and creates a unique exploitation pathway: **offering scalable, compliant model-training solutions to hospitals, industry partners and research networks throughout Europe**. The consortium can leverage this position to influence future standards in surgical video representation, metadata schemas, and federated-learning interoperability.

- **Industrial Innovation and Technology Transfer**
  Through collaboration with DELL, Scontain, and others industry partners, the consortium has established a technology pipeline that is well positioned for transfer into industrial products and clinical solutions. **StreamSense, for example, has clear potential to be integrated as a module within surgical video management systems or as part of digital OR platforms**, enabling efficient semantic search and retrieval in large surgical video archives.

Likewise, secure federated-learning workflows based on Trusted Execution Environments (TEEs) can evolve into **a service offering for hospitals and industry partners that need to train AI models without moving sensitive patient data**. This creates opportunities for consulting, licensing, integration into existing product portfolios, and long-term industrial partnerships. At the same time, these collaborations open avenues for co-innovation—for instance, the development of scalable surgical AI foundation models or real-time intraoperative assistance tools. Broad industrial uptake of these technologies would significantly accelerate the translation of project results into medical solutions with tangible benefits for patients.

- **Strengthening Clinical Translation and Workforce Empowerment**
  Through targeted dissemination activities—including public events, summer schools, dedicated teaching modules, and **close clinical collaborations—the project has created a strong foundation for sustainable clinical translation**. By training young researchers, medical students, and surgeons in AI-driven surgical assistance, the consortium is building a workforce that can not only use but also critically develop and assess surgical AI technologies.
  This establishes a long-term exploitation pathway: as clinicians increasingly work with data-driven tools, the demand for advanced AI systems in everyday surgical practice will naturally grow. In parallel, openly available datasets and software resources function as high-quality training material for emerging AI developers, **helping Europe cultivate a self-sustaining talent pool capable of driving the continued evolution of surgical data science, healthcare robotics, and intelligent operating room technologies**.

4.7.2 Challenges

At the same time, several structural and technical challenges must be overcome to fully exploit the project's results.

- Regulatory Integration and Clinical-Grade Robustness
  AI systems intended for surgical assistance face stringent requirements for **transparency, robustness, and validation**. The EU AI Act requires extensive risk management, human oversight mechanisms, and post-market monitoring. Achieving compliance requires large, diverse datasets and stable generalization across institutions, devices, and surgical workflows. Although the project has made progress in the development of federated learning and multi-center datasets, the **transition of these prototypes to medical device–ready AI systems will require additional clinical trials, rigorous performance benchmarking, harmonization between hospitals and regulatory certification**. This creates a significant time and resource burden that may slow translation into market-ready technologies.

- Data Fragmentation and Heterogeneity Across Healthcare Institutions
  Despite the advantages of federated learning, surgical data remain highly heterogeneous: **differences in surgical technique, video resolution, endoscope manufacturers, annotation quality, and institutional workflows introduce variability that can reduce model performance**. Furthermore, hospitals often differ in their data governance processes, IT infrastructure, and cybersecurity requirements, making it difficult to deploy uniform federated learning frameworks or streaming pipelines such as Pravega and GStreamer. Effective exploitation therefore requires **sustained collaboration, common data standards, and extensive engineering work to ensure interoperability**. Without harmonization, scaling up the project outcomes across Europe may be challenging.

- Integration into Clinical Workflow and Professional Acceptances
  For AI to meaningfully support surgeons, it must integrate seamlessly into the intraoperative workflow—offering reliable, interpretable, real-time insights without increasing cognitive load. Achieving this requires **substantial design effort, user-centered interface development, and**

comprehensive training for clinical staff. Moreover, professional acceptance is not guaranteed. Surgeons must trust that AI systems are robust, reliable, and beneficial. If AI assistance is perceived as slow, disruptive, or insufficiently validated, adoption will stagnate. Thus, exploitation depends on **coordinated efforts to create clinically meaningful use cases, co-develop tools with end users, and provide continuous education**. Without this, innovations risk remaining in research environments rather than translating into widespread clinical use.

4.8 Sano's Exploitation Plans

The main product of Sano's involvement in the NEARDATA project is the Transcriptomics Atlas pipeline, which includes implementation, cloud architecture design and deployment, as well as significant optimization (both application-specific and cloud-related). Moreover, this is extended by the Serverless Lithops solution for running the pipeline (pseudoalignment path with the Salmon tool). The exploitation of these products can be related to providing bioinformatics services, on-demand processing with the pipeline and others. Moreover, the partnership within the project and the results obtained, position Sano as an important center of competence combining expertise in both bioinformatics pipelines and cloud/HPC infrastructure developments. All this leads to multiple exploitation opportunities.

4.8.1 Opportunities

The project opens multiple paths for successful exploitation:

- Atlas as a Service: such an extension of the Transcriptomics Atlas would provide users with an easy way to process their RNA-sequences using the Transcriptomics Atlas pipeline. The pipeline developed within the project, together with a database of results, can be made publicly available as a service hosted by Sano, and providing the analysis services in a fee-based model. The main costs would be related to the usage of underlying computing resources. Based on our experience from the project, these resources can be obtained either from a commercial cloud provider, but also from a scientific HPC centre, such as Academic Computer Center Cyfronet-AGH. Such HPC centres, while focusing mainly on non-commercial scientific users, are also open to provide computing power for commercial users. This is well-aligned with the EuroHPC initiatives, such as AI Factories and other investments in sovereign computing platforms.

- Consulting services for bioinformatics and computational medicine: there is a possibility of offering consulting services to bioinformatics companies, especially in the context of computations for biomedical applications. Thanks to the expertise and solutions developed in the project, Sano is becoming recognized as a valuable expert in optimization of scientific computing, in particular in computational medicine and bioinformatics. The centre already collaborates with many industrial partners, and such commercial contracts are on of the pillars of financial sustainability of the centre.

4.8.2 Challenges

We also recognize several challenges to effective exploitation:

- Quality of the data for the Atlas extension – The Transcriptomics Atlas can still be extended with new tissues, which would improve its quality and applicability. However, gathering quality data for niche tissues may be challenging, and the currently available dataset may not be enough. This process is time consuming, as it requires a careful data quality assessment by the experts and manual verification of results.

- Atlas as a Service opportunity – to make it a production ready service requires a long-term support, development, and dedicated resources, including monitoring of the service. To fully assess the feasibility of this exploitation path, a more detailed cost-benefit analysis needs to be performed and a marketing research should be done.

4.9 UKHSA Exploitation Plans

The novel technologies explored by the UKHSA use case open an exciting range of possibilities to put the processing of data relevant to pathogen genomics on a firmer, more scalable basis, while at the same time reaping a number of straightforward advantages. We plan on exploring a more widespread adoption of such technologies within UKHSA; this will happen at the level of UKHSA core bioinformatics, which is the main internal provider of genomic services and beyond.

4.9.1 Opportunities

There are a number of opportunities offered by the technologies making up the NEARDATA platform that can be helpful to pathogen genomics at UKHSA and other public bodies and agencies focusing on public health:

- They might foster the transition to more modern and better supported technologies. They would also align well with the push for "cloud-first" systems set in motion by the UK government during the last years in order to make the UK administration more robust and more resilient to incidents and when rapid change is needed.
- They might provide easier scalability when loads are unusually high, for instance, at the beginning of a pandemic — as COVID-19 has shown, quickly ramping up service provision when everyone is trying to do the same and when the technical capabilities are not present in the first place is something very difficult to achieve.
- They would provide easier data tracking capabilities and real-time access across UKHSA federated systems to the many different professional profiles that need to collaborate and interact on the management of outbreaks at different institutions and across different levels of government.

Overall, our hope is that the adoption of these and other technical solutions for the handling of extreme data will result in a better and more cost-effective public health, ultimately providing benefits for the community as a whole.

4.9.2 Challenges

However, the regulatory landscape for health-related applications in the UK and elsewhere is extremely complex, with a large number of requirements that must be satisfied in order to certify any application for laboratory or clinical use.

In particular, clinical pipelines for the processing of pathogen genomic data typically require:

- Compliance with data protection regulations, such as the GDPR. This is of particular relevance to UKHSA, as clinical investigations ultimately depend on the collection of personal data. Rigorous precautions are internally followed to maintain confidentiality at each processing stage — for instance, samples names are pseudo-anonymized almost immediately upon reception and reads of human origin eliminated from sequencing files prior to export to public repositories, so as to eliminate any possibility of re-identification.
- Compliance with cyber-security directives, such as the Cyber Security and Resilience Bill currently being read in the UK Parliament. This has implications at the level of infrastructure design and operation, imposing several constraints on data handling and the choices of technology employed during data processing.
- Accreditation provided by UKAS (UK Accreditation Service) of the pipelines used to process sequencing data for clinical use. Additional certifications might be required, for instance if the clinical pipeline is considered a laboratory instrument. UKAS accreditation requires large-scale testing of the pipelines, and end-to-end and periodic physical inspections to the infrastructure used and the workplace to ensure that proper procedure is followed and adequate precautions are in place should any incident occur.

- ISO accreditation according to relevant standards such as ISO/IEC 17025 and 15189.

This generates some understandable internal reluctance to evaluate new technologies, in particular considering the significant investment required to actually incorporate them into tried and tested workflows (such as infrastructural investments, comprehensive testing for re-certification, and re-training of staff).

In this light, we appreciate how our collaboration with partners in the NEARDATA consortium made it possible for us to freely develop and test innovative solutions that are likely to be very helpful in the future. Hence, we will do our best to overcome regulatory issues as much as possible to get what we learned with NEARDATA widely deployed and used, across UKHSA and elsewhere.

5 Conclusions

This deliverable has brought a detailed yet comprehensive history of the activities on the domain of communication, dissemination and exploitation carried out by consortium partners during the period starting at month 17 up to month 35, when the project ended. It has been demonstrated that targeted communities (Scientific community, European Cloud Providers, Pharma Industry, Hospitals, Public Health Authorities, Hi-tech Health Equipment Manufacturers, General Public), ca. 15,000 people, were reached and got relevant information about NEARDATA throughout the consortium's lifetime.

Use case leaders concluded with important analysis on the work realized, how the outcome will benefit and influence the science, the society and the economy; and how the health-related *omics fields will absorb and use the results of the project. For example, from the societal and economic sides, UC leaders notice the potential of improvement in treatments, disease and death prevention, as well as costs reduction in all the targeted *omics areas. The quality of the data and of the systems that were made public also have the potential of being adopted as part of the present reference stack and to influence future standards.

With respect to the scientific impact, it is important to refer to the many presentations the partners offered in congresses like BDVA, Middleware, and MWC. It is remarkable to see the number of scientific articles produced (*42* papers published or pending approval) and how they are already being cited by other papers published in prestigious journals. With articles published in the top *A** or *A* venues – *3* and *13* respectively, – comprising *38%* in flagship and excellent venues; plus *6* and *2* papers published in venues with grades *B* or *C*, *57%* of the intellectual production has being made available via recognized venues. While *19* articles, i.e. *43%*, were published in yet to be classified journals, this is no demerit and quite the contrary. Important and well established journals, like "Nature Scientific Reports", and new initiatives, like "Cloud-Edge Continuum" and "Workshop on Serverless Systems, Applications and Methodologies" have the same strict filtering and will not accept poor quality papers.

Table 7: Communication Achievements M17 - M35

Type of communication	Category of audience	Achievements
Scientific Publications	Scientific community	25 publications
Conferences and Workshops	Scientific community and Industry	60 events
Community Building	Scientific community, Industry	2 events
Meetings outside the consortium	Scientific community, Industry	14 meetings
Events for society	General Public	6 events
Press releases and posts	General Public	9 publications

Together with community building efforts ("Horizon Results Booster" and "DataNexus Cluster") it is possible to see that, albeit the consortium officially ends now, the results will endure and influence other science projects in the near future.

6 Appendix

6.1 Dissemination and Meeting Activities (M17-M35)

Table 8: Dissemination and Meeting Activities

Event Type	Description	Date	Type of audience	Partner, audience
Congress	Keynote talk at a conference, Innovations in single cell omics' (ISCO)	01/05/2024	Scientific community	EMBL 150
Congress	6G-life General Assembly, oral presentation of robot-assisted surgery	06-07/05/202	Scientific community, Hi-tech Health Equipment Manufacturers	NCT 200
Congress	Talk at a conference, GRC Single-Cell Genomics	02/05/2024	Scientific community	EMBL 250
Presentation	Presentation in IBM Zurich "The many faces of locality in Big Data analytics"	17/06/2024	Scientific community	URV 30
Presentation	Presentation of the Project and collaborations - U. Neuchatel	25/06/2024	Scientific community	URV 30
Presentation	Presentation of the Project and collaborations - ETH	24/06/2024	Scientific community	URV
Workshop	the long night of science in Dresden. 3 demonstrations of AI-based robot-surgery, surgical training and intraoperative navigation system of liver in the field of translational surgical oncology have been presented	14/06/2024	Scientific community, General Public	NCT 200
Presentation	Overview of Research Topics at Sano - Keynote Presentation, Insigneo Showcase, University of SHeffield	14/06/2024	Scientific community	SANO 200
Congress	co-organizer of a conference / meetings with key opinion leaders and stakeholders in the field, VIB Spatial Omics	03/06/2024	Scientific community	EMBL 450
Congress	Berlin 6G-Conference, link-to	01-04/07/2024	Scientific community, General Public	NCT
Workshop	Presentation of our EndoMersion Demo (computer and Robotic assisted Surgery) to Saxony SMS-Minister Mrs. Petra Köpping in Experimental Operation Room	30/07/2024	Scientific community, Hospitals, Public Health Authorities	NCT 10

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Workshop	Reviews and organization of CEC Workshop at IEEE ICNP'24 with an open-session for EU Projects to disseminate results.	07/2024 to 09/2024	Scientific community	DELL 15
Meeting	Monthly meetings with the DataNexus cluster	07/2024 - 09/2024	Scientific community	URV 7
Presentation	Launchhub42 CeTI Opening, Presentation of our EndoMersion Demo (computer and Robotic assisted Surgery) to visitors, Dresden	08/08/2024	Scientific community, Hospitals, General Public	NCT 100
Press release	DataNexus cluster: press release and flyer	09/2024	Scientific community, General Public	URV 50
Congress	Presented latest findings and work in NEARDATA at EASD 2024 in Madrid	09/2024	Scientific community	BSC 200
Meeting	Presentation of the work with URV and NCT on embeddings indexing of video streams to Dell OCTO Research Council.	13/09/2024	Scientific community, European Cloud providers, General Public	DELL 20
Congress	NEARDATA presentation in European Big Data Value Forum (BDVAF 2024) accompanied by the other projects of the DATANEXUS cluster, link-to	02/10/2024	Scientific community, General Public	URV 470
Workshop	Presentation of Demo Remote surgery system Endomersion by Viszeralmedizin, Leipzig, link-to	03/10/2024	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 300
Congress	27th INTERNATIONAL CONFERENCE ON MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION (MICCAI 2024), Marrakesh – Presentation about FL Challenge	06-10/10/2024	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 2300
Congress	40. Congress of the German Society for Vascular Surgery, invited talk of "Computer-Assisted Assistance in the Operating Room of the Future", link-to	09-12/10/2024	Scientific community, Hospitals, Hi-tech Health Equipment	NCT 200

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Congress	Teaming Club Conference, Overview of Research at Sano, link-to	9/10/2024	Scientific community	SANO 100
Meeting	Visit of Nobel prize winner Prof. Thomas Christian Südhof at NCT, presentation of computer- and robot-assisted surgery by Zhaoyu Chen	10/10/2024	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 20
Congress	IEEE Cluster Conference, Poster: Optimizing STAR Aligner for High Throughput Computing in the Cloud, link-to	24-27/10/2024	Scientific community, General Public	SANO
Congress	LSBC Conference Keynote Presentation, Healthcare of tomorrow, crafted today: AI and computational innovations in medicine, link-to	25/10/2024	Scientific community, Hospitals, Public Health Authorities	SANO 100
Workshop	Attendance and organization of the CEC Workshop at IEEE ICNP'24 with an open-session for EU Projects to disseminate results, link-to	28/10/2024	Scientific community, General Public	DELL 35
Congress	NEARDATA poster in CEC'24, link-to	28/10/2024	Scientific community, General Public	URV 15
Meeting	D-CREDO project meeting, Overview of Research at Sano,	6/11/2024	Scientific community	SANO 20
Forum	COSMO Research Forum AI in Surgery, Dresden	13/11/2024	Scientific community, Hospitals	NCT 10
Congress	Presentation at TEx14 Dell conference on "StreamSense: Policy-driven Semantic Video Search in Streaming Systems".	18/11/2024	Scientific community, General Public	DELL 25
Congress	Sano Science Day Conference, Poster: Optimizing Star Aligner for High Throughput Computing in the Cloud, link-to	27/11/2024	Scientific community	SANO 50
Congress	Sano Science Day Conference, Poster: Serverless Approach to Pseudoalignment with Salmon Tool, Kamil Burkiewicz, link-to	27/11/2024	Scientific community	SANO 50
Meeting	Internal meeting with ISG PM member to discuss future research challenges related to Dell Data Lakehouse product.	27/11/2024	General Public	DELL 1

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Workshop	Hamlyn Winter School on Surgical Imaging and Vision, London, presentation of "The context-aware assistant in the OR of the Future"	02-06/12/2024	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 36
Workshop	Sachsen Full-day School Program "Girls for Robots - CeTI", Dresden, Introducing our projects and the experimental operation room to young women to engage them to study STEM	03/12/2024	General Public	NCT 20
Press release	Monthly Newsletter, link-to	2025	General Public	URV 70
Press release	Lithops in the Innovation Radar, link-to	01/2025	General Public	URV 100
Congress	Presentation of NEARDATA at HiPEAC'25.	21/01/2025	Scientific community	DELL 25
Meeting	Internal progress update on EU Projects outcomes.	01/2025 – 03/2025	Scientific community	DELL 20
Meeting	Internal meeting with Dell PowerScale team to discuss progress on measurement paper.	07/02/2025	Scientific community	DELL 2
Meeting	IFIP Working Group 10.4 Dependable Computing and Fault Tolerance	14/02/2025	Scientific community	TUD
Workshop	Future-Ready event: On-Demand Solutions with AI, Data, and Robotics in Brussels, organised by Adra. Together with the DataNexus Cluster, NEARDATA was part of this workshop.	18/02/2025	Scientific community	URV 30
Meeting	Discussion with research director (Mike Robillard) on exploitable EU Project outcomes.	20/02/2025	Scientific community	DELL 1
Promotional material	MWC'25. Incorporated into BSC catalogue of initiatives.	03-06/03/25	General Public	BSC
Congress	BDVA Presentation on Big Data for AI.	05/03/2025	Scientific community	DELL 100
Congress	Presentation at IEEE Conference on Virtual Reality and 3D User Interfaces, Saint-Malo	10-12/03/2025	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 200
Congress	Presentation at Scientist for Future (Radiology meeting), Dresden	13/03/2025	Scientific community, Hospitals	NCT 50

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Congress	13th Privacy Enhancing Techniques Convention (PET-CON 2025.1) Workshop participation. Talking about confidential computing with federated learning.	14/03/2025	Scientific community, General Public	TUD
Meeting	BDVA Healthcare task force meeting.	20/03/2025	Scientific community, General Public	DELL 5
Workshop	Presentation at Robotic Institute Germany, Nürnberg	21/03/2025	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 100
Congress	Presentation at Retreat of TSO, Bastei Saxony	24-27/03/2025	Scientific community, Hospitals	NCT 25
Meeting	Explored collaboration with URV/Dell on how to apply this technology within their vector DB research	04/2025	Scientific community	UKHSA
Workshop	DKFZ Heidelberg, Event: Joint Workshop on Endoscopic Video Analysis, Title: Federated Learning for Surgical Foundation Models: A viable Alternative?	30/04/2025	Scientific community, Hospitals	NCT 20
Presentation	Presentation of NEARDATA to DELL OCTO Research Council, online	01/05/2025	Scientific community	DELL 10
Congress	LifeScience4EU Conference, Panel on KPIs for success: what metrics define progress in health innovation?, link-to	15/05/2025	Scientific community, Hospitals, Public Health Authorities	SANO 100
Congress	Dr. Sashko Ristov from the University of Innsbruck visited the URV and gave the talk on "Coding Modern Serverless Workflow Applications in Sky Computing: From Monoliths to AI-Assisted Devel"	15/05/2025	Scientific community	URV 15
Congress	CCGrid2025 conference - participation with poster and presentation, link-to	19-22/05/2025	Scientific community, General Public	SANO 160

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Meeting	Academic visit of Dr. Alejandro Granados from King's College London to NCT Dresden and research discussion, Dresden, 25	21/05/2025	Scientific community	NCT 25
Presentation	Data Week session "DataNexus: Extreme data driven solutions and services - Updated perspectives and Challenges for Europe" in Athens, Greece - Together with the DataNexus Cluster, NEARDATA was part of this event by Dr. Pedro Garcia, link-to	27-28/05/2025	Scientific community	URV 30
Meeting	Dr. Ahmed M. A. Sayed from the Queen Mary University of London visited the URV and gave the talk on "Advancing Decentralized AI: Scalable, Adaptive, and Client-Centric Learning Systems"	30/05/2025	Scientific community	URV 15
Promotional material	Completed DELL internal NEARDATA White Paper FY26 (+200 recipients).	30/05/2025	Scientific community	DELL 200
Meeting	one-week visit of Prof. Xiangyu Chu from The Chinese University of Hong Kong and potential cooperation discussion, Dresden	02-06/06/2025	Scientific community	NCT 20
Workshop	Cloud Meets Innovation Hackathon co-hosted at the University of Innsbruck and URV, link-to	06-07/06/2025	Scientific community	URV 30
Congress	BISIH (Bank for International Settlements Innovation Hub) Technologies for the Future Financial System Workshop presenting TEEs in the Banking context, Basel	10-13/06/2025	General Public	TUD
Congress	Confidential Computing Summit 2025, Secure automatic updates for cloud-native apps with SCONE	16-18/06/2025	Scientific community, General Public	TUD
Workshop	Cloud Control Workshop 2025	16-17/06/2025	General Public	TUD

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Workshop	18th Cloud Control Workshop – Enabling Secure Rolling Updates for Confidential Kubernetes, Skåvsjöholm	17/06/2025	General Public	TUD
Workshop	18th Cloud Control Workshop – Is Confidential Computing Living Up to Its Potential?, Skåvsjöholm	18/06/2025	General Public	TUD
Workshop	Dresden Long Night of Science	20/06/2025	Scientific community, General Public	NCT 200
Congress	SciPy 2025 (Tacoma), link-to	07-13/07/2025	Scientific community	URV 100
Congress	2025 USENIX Annual Technical Conference in Boston, link-to	07-09/07/2025	General Public	URV 200
Congress	ICCS Conference, Accelerating Cloud-Based Transcriptomics: Performance Analysis and Optimization of the STAR Aligner Workflow, link-to	07/07/2025	Scientific community, General Public	SANO 200
Workshop	SurgicalDataScience summer school, Strasbourg	09/07/2025	Scientific community, Hospitals	NCT
Medium Post	Beyond 'Hello World': Powering Real-World Science and AI with PyRun, link-to	11/07/2025	Scientific community	URV 20
Medium Post	Effortless Serverless Python: Get Your Code Running in the Cloud in 3 Clicks with PyRun, link-to	11/07/2025	Scientific community	URV 17
Workshop	Organising InDeMed Workshop "Artificial Intelligence, Extended Reality, and Robotics in Healthcare: Emerging Pathways from Diagnosis to Therapy" supported by Indo-German Science & Technology Centre (IGSTC), more than 50 attendees from India and Germany, Dresden	16-18/07/2025	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 50
Tutorial	From Zero to Cloud in 5 Minutes: Your First Python Script on PyRun, link-to	18/07/2025	Scientific community, General Public	URV 13
Congress	Industrial keynote at IEEE ICDCS'25: "Streaming Storage in Practice: Engineering, Research, and the Road Ahead", link-to	19-24/07/2025	Scientific community, General Public	DELL 30

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Meeting	Meeting with Vrashank Jain (PM) regarding future storage challenges in the Dell Data Lakehouse product, online	28/07/2025	General Public	DELL 2
Meeting	Meeting with Teng Yu's team (SD) to disseminate serverless vector DB work, online	01/08/2025	Scientific community	DELL 4
Meeting	Meeting with Himmabindu Tummala (DE) to communicate our experimental analysis of S3 vectors (online, 3 attendees).	11/08/2025	Scientific community	DELL 3
Congress	EuroSciPy 2025 link-to	21/08/2025	Scientific community	URV
Congress	Euro-Par PhD Symposium, Heterogeneous computing, storage and network infrastructures for medical applications, link-to	26/08/2025	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	SANO 200
Congress	Huawei Global Software Summit - Confidential computing for financial services	01/09/2025	General Public	SCO
Workshop	Organising the Dr. Hans Riegel-Stiftung Workshop ""Fachseminar Dresden—Technologie schafft Verbindung"" at National Center for Tumor Diseases Dresden (NCT/UCC), Dresden	05/09/2025	Scientific community, Hospitals, Hi-tech Health Equipment Manufacturers	NCT 30
Congress	BIS Innovation Summit 2025 - Secure Rolling Updates for Cloud-Native Confidential Workloads, Basel	11/09/2025	General Public	SCO
Congress	2025 IEEE International Conference on eScience (eScience), Solutions for Distributed Memory Access Mechanism on HPC Clusters, link-to	15/09/2025	Scientific community	SANO 200
Release Notes	PyRun Roadmap, link-to	17/09/2025	General Public	URV 24.000
Promotional material	La Nit Europea de la Recerca, link-to	19/09/2025	General Public	URV 100
Workshop	Organization and Attendance to CEC'25 Workshop @ICNP'25, link-to	22/09/2025	Scientific community, General Public	DELL 30
Congress	PyConES 2025, link-to	18/10/2025	Scientific community, General Public	URV 750

Continued on next page

Table 8 (continued)

Event Type	Description	Date	Type of audience	Partner, audience
Congress	Smart Country Convention 2025 - Register-as-a-Service	30/10/2025	General Public	SCO
Workshop	Vodafone Tech Innovation Center Dresden, Cyber Horizon – Security, Sovereignty, Collaboration, Confidential computing with SCONE, Dresden	12-13/11/2025	Scientific community, General Public	TUD
Workshop	Supermicro, Confidential AI	17/11/2025	Scientific community, General Public	TUD
Medium Post	AI on Demand, link-to	28/11/2025	Scientific community, General Public	URV

6.2 Publications (M17-M35)

Table 9: Publications

Title	Authors	Publisher/Journal/ Magazine/Conference	Link
Triad - Trusted Timestamps in Untrusted Environments	Gabriel P. Fernandez, Andrey Brito, Christof Fetzer	2023 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)	Publication
Dataplug: Unlocking extreme data analytics with on-the-fly dynamic partitioning of unstructured data	Aitor Arjona, Pedro Garcia-Lopez, Daniel Barcelona-Pons	2024 IEEE 24th International Symposium on Cluster, Cloud and Internet Computing (CCGrid)	Publication
Serverless End Game: Disaggregation enabling Transparency	Pedro Garcia Lopez, Aleksander Slominski, Bernard Metzler, Michael Berhendt, Simon Shillaker	SESAME '24: Proceedings of the 2nd Workshop on Serverless Systems, Applications and Methodologies	Publication
CRISP: Confidentiality, Rollback, and Integrity Storage Protection for Confidential Stateful Computing	Ardhi Putra Pratama Hartono, Andrey Brito, Christof Fetzer	2024 IEEE 17th International Conference on Cloud Computing (CLOUD)	Publication
Optimizing STAR Aligner for High Throughput Computing in the Cloud	Piotr Kica, Sabina Lichołai, Michał Orzechowski, Maciej Malawski	Cluster2024 conference	Publication
Exploring Secure and Efficient Temporary Data Sharing between co-located Kubernetes Containers	Aitor Arjona, Bernard Metzler, Pascal Spörri	CEC Workshop (IEEE ICNP'24)	Publication

Continued on next page

Table 9 (continued)

Title	**Authors**	**Publisher/Journal/ Magazine/Conference**	**Link**
Towards Multi-tier Stream Data Tiering in the Cloud-Edge Continuum	O. Jundi, R. Gracia-Tinedo, S. Ahearne, P. Spörri, B. Metzler	CEC Workshop (IEEE ICNP'24)	Publication
Serverful Functions: Leveraging Servers in Complex Serverless Workflows (industry track)	Germán T. Eizaguirre, Daniel Barcelona-Pons, Aitor Arjona, Gil Vernik, Pedro García-López, Theodore Alexandrov	Middleware Industrial Track '24	Publication
StreamSense: Policy-driven Semantic Video Search in Streaming Systems	Finol, Gerard; Gabriel, Arnau; Garcia-Lopez, Pedro; Gracia-Tinedo, Raul; Liu, Luis; Docea, Reuben; Kirchner, Max; Bodenstedt, Sebastian	Middleware Industrial Track '24	Publication
Back to the Byte: Towards Byte-oriented Semantics for Streaming Storage	Gracia-Tinedo, Raul; Junqueira, Flavio; Kaitchuck, Tom	Middleware Industrial Track '24	Publication
A Comprehensive Study on the Impact of Vulnerable Dependencies on Open-Source Software	Shree Hari Bittugondanahalli Indra Kumar; Lília Rodrigues Sampaio; André Martin; Andrey Brito; Christof Fetzer	2024 IEEE 35th International Symposium on Software Reliability Engineering (ISSRE)	Publication
TICAL: Trusted and Integrity-protected Compilation of AppLications	Robert Krahn, Nikson Kanti Paul, Franz Gregor, Do Le Quoc, Andrey Monteiro Brito, André Martin, Christof Fetzer	2024 19th European Dependable Computing Conference (EDCC)	Publication
Burst Computing: Quick, Sudden, Massively Parallel Processing on Serverless Resources	Daniel Barcelona-Pons; Aitor Arjona; Pedro García-López; Enrique Molina-Giménez; Stepan Klymonchuk	USENIX Annual Technical Conference 2025	Publication
Federated EndoViT: Pretraining Vision Transformers via Federated Learning on Endoscopic Image Collections	Max Kirchner, Alexander C. Jenke, Sebastian Bodenstedt, Fiona R. Kolbinger, Oliver L. Saldanha, Jakob N. Kather, Martin Wagner, Stefanie Speidel	Computer Vision and Pattern Recognition	Publication
Serverless Approach to Running Resource-Intensive STAR Aligner	Piotr Kica, Michał Orzechowski, Maciej Malawski	CCGrid 2025	Publication
Dynamic Selection and Detection of Spreading Factors and Channels for End-Node Devices of LoRa Networks	Carles Aliagas, Roger Pueyo Centelles, Roc Meseguer, Pere Millán, and Carlos Molina	Electronics 2025	Publication
Quantifying Serverless Elasticity: The gumeter Benchmark Suite	Germán T. Eizaguirre, Enrique Molina, Gerard Finol, Carlos Molina, Pedro García-López	International Conference on Service-Oriented Computing (ICSOC 2025)	Publication

Continued on next page

Table 9 (continued)

Title	**Authors**	**Publisher/Journal/ Magazine/Conference**	**Link**
Accelerating Cloud-Based Transcriptomics: Performance Analysis and Optimization of the STAR Aligner Workflow	Piotr Kica, Sabina Lichołai, Michał Orzechowski, Maciej Malawski	Distributed, Parallel, and Cluster Computing	Publication
Appendix300: A multi-institutional laparoscopic appendectomy video dataset for computational modeling tasks	Fiona R. Kolbinger, Max Kirchner, Kevin Pfeiffer, Sebastian Bodenstedt, Alexander C. Jenke, Julia Barthel, Matthias Carstens, Karolin Dehlke, Sophia Dietz, Sotirios Emmanouilidis, Guido Fitze, Martin Freitag, Fabian Holderried, Thorsten Jacobi, Weam Kanjo, Linda Leitermann, Sören Torge Mees, Steffen Pistorius, Conrad Prudlo, Astrid Seiberth, Jurek Schultz, Karolin Thiel, Daniel Ziehn, Stefanie Speidel, Jürgen Weitz, Jakob Nikolas Kather, Marius Distler, Oliver Lester Saldanha		Publication
Exhaustive Variant Interaction Analysis Using Multifactor Dimensionality Reduction	Gómez-Sánchez, G.; Alonso, L.; Pérez, M.Á.; Morán, I.; Torrents, D.; Berral, J.L.	Appl. Sci. 2024	Publication
Solutions for Distributed Memory Access Mechanism on HPC Clusters	J. Meizner and M. Malawski	2025 IEEE International Conference on eScience (eScience)	Publication
Building Stateless Serverless Vector DBs via Block-based Data Partitioning	Daniel Barcelona-Pons, Raúl Gracia-Tinedo, Albert Cañadilla-Domingo, Xavier Roca-Canals, and Pedro García-López	ACM SIGMOD/PODS International Conference on Management of Data	Publication
Serverless Data Analytics (Finally) Bridging the Gap: Introducing the Ortzi Dataframe	Germán T. Eizaguirre, Marc Hostau, Marc Sánchez-Artigas	IEEE 17th International Conference on Cloud Computing (CLOUD)	Publication

Continued on next page

Table 9 (continued)

Title	**Authors**	**Publisher/Journal/ Magazine/Conference**	**Link**
Genetic Profiling and Early Detection of Type Diabetes Subtypes through Sex-Straified GWAS and Explainable AI	Lorena Alonso-Parrilla, Miguel Ángel Pérez-Elena, Mohammed Yousef Salem Ali, Maedeh Mashhadikhan, Nicolás Gaitán, Leila Satari, Rodrigo Martín, Anthony Piron, Xavier Farré, Natalia Blay, Lydia Ruiz, Aikaterini Lymperidou, Cecilia Salvoro, Rafael de Cid, Josep Lluís Berral, Juan R González, Ignasi Morán, Miriam Cnop, David Torrents		Publication
Lithops-HPC: Extending the Serverless Paradigm to High-Performance Computing for Accessible Resource Management	Andrés Benavides, Aaron Call, Pietro Morichetti, Daniel Barcelona-Pons, Ramon Nou	FGCS Conference	Pending

References

[1] F. R. Kolbinger, M. Kirchner, K. Pfeiffer, S. Bodenstedt, A. C. Jenke, J. Barthel, M. R. Carstens, K. Dehlke, S. Dietz, S. Emmanouilidis, G. Fitze, L. Leitermann, S. T. Mees, S. Pistorius, C. Prudlo, A. Seiberth, J. Schultz, K. Thiel, D. Ziehn, S. Speidel, J. Weitz, J. N. Kather, M. Distler, and O. L. Saldanha, "Appendix300: A multi-institutional laparoscopic appendectomy video dataset for computational modeling tasks." ISSN: 3067-2007 Pages: 2025.09.05.25335174.

[2] M. Kirchner, H. Hoffmann, A. C. Jenke, O. L. Saldanha, K. Pfeiffer, W. Kanjo, J. Alekseenko, C. d. Boer, S. R. Kolamuri, L. Mazza, N. Padoy, S. Bano, A. Reinke, L. Maier-Hein, D. Stoyanov, J. N. Kather, F. R. Kolbinger, S. Bodenstedt, and S. Speidel, "Federated learning for surgical vision in appendicitis classification: Results of the FedSurg EndoVis 2024 challenge."

[3] A. C. Jenke, S. Bodenstedt, F. R. Kolbinger, M. Distler, J. Weitz, and S. Speidel, "One model to use them all: Training a segmentation model with complementary datasets."

[4] M. Kirchner, A. C. Jenke, S. Bodenstedt, F. R. Kolbinger, O. L. Saldanha, J. N. Kather, M. Wagner, and S. Speidel, "Federated EndoViT: Pretraining vision transformers via federated learning on endoscopic image collections."

[5] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. d. Gusmão, and N. D. Lane, "Flower: A friendly federated learning research framework."

[6] A. Arjona, A. Gabriel-Atienza, S. Lanuza-Orna, X. Roca-Canals, A. Bourramouss, T. K. Chafin, L. Marcello, P. Ribeca, and P. García-López, "Scaling a variant calling genomics pipeline with faas," in Proceedings of the 9th International Workshop on Serverless Computing, WoSC '23, (New York, NY, USA), p. 59–64, Association for Computing Machinery, 2023.

[7] X. Didelot and P. Ribeca, "KPop: accurate and scalable comparative analysis of microbial genomes by sequence embeddings," Genome Biology, vol. 26, p. 170, June 2025.

[8] A. Arjona, P. García-López, and D. Barcelona-Pons, "Dataplug: Unlocking extreme data analytics with on-the-fly dynamic partitioning of unstructured data," in IEEE CCGrid'24, pp. 567–576, 2024.

[9] R. Gracia-Tinedo, F. Junqueira, and T. Kaitchuck, ""Back to the byte": Towards byte-oriented semantics for streaming storage," in ACM/IFIP Middleware'24 (Industrial Track), pp. 43–49, 2024.

[10] "Iso/iec 23092." https://en.wikipedia.org/wiki/MPEG-G, 2024.

[11] V. Amaral, B. Norberto, M. Goulão, M. Aldinucci, S. Benkner, A. Bracciali, P. Carreira, E. Celms, L. Correia, C. Grelck, H. Karatza, C. Kessler, P. Kilpatrick, H. Martiniano, I. Mavridis, S. Pllana, A. Respício, J. Simão, L. Veiga, and A. Visa, "Programming languages for data-intensive hpc applications: A systematic mapping study," Parallel Computing, vol. 91, p. 102584, 2020.

[12] O. OnDemand, "Open ondemand," 2025.

[13] E. Comission, "Ai factories," 2025.

[14] N. S. Foundation, "National artificial intelligence research resource," 2025.

[15] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, "Pegasus, a workflow management system for science automation," Future Generation Computer Systems, vol. 46, pp. 17–35, 2015.

[16] E. Tejedor, Y. Becerra, G. Alomar, A. Queralt, R. M. Badia, J. Torres, T. Cortes, and J. Labarta, "Pycompss: Parallel computational workflows in python," The International Journal of High Performance Computing Applications, vol. 31, no. 1, pp. 66–82, 2017.

[17] Y. Babuji, A. Woodard, Z. Li, D. S. Katz, B. Clifford, R. Kumar, L. Lacinski, R. Chard, J. M. Wozniak, I. Foster, M. Wilde, and K. Chard, "Parsl: Pervasive parallel programming in python," in Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing, HPDC '19, (New York, NY, USA), p. 25–36, Association for Computing Machinery, 2019.

[18] M. Rocklin, "Dask: Parallel computation with blocked algorithms and task scheduling," in SciPy, 2015.

[19] E. Shaikh, I. Mohiuddin, Y. Alufaisan, and I. Nahvi, "Apache spark: A big data processing engine," in 2019 2nd IEEE Middle East and North Africa COMMunications Conference (MENACOMM), pp. 1–6, 11 2019.

[20] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache flink : Stream and batch processing in a single engine," IEEE Data(base) Engineering Bulletin, vol. 36, pp. 28–33, 2015.

[21] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, and I. Stoica, "Ray: A distributed framework for emerging ai applications," 2018.

[22] L.-H. Team, "Lithops-hpc: Enabling serverless into hpc environments," 2025.

[23] J. Sampé, M. Sánchez-Artigas, G. Vernik, I. Yehekzel, and P. García-López, "Outsourcing data processing jobs with lithops," IEEE Transactions on Cloud Computing, vol. 11, no. 1, pp. 1026–1037, 2023.

[24] M. A. Yaqin, R. Sarno, and S. Rochimah, "Measuring scalable business process model complexity based on basic control structure," International Journal of Intelligent Engineering & System, vol. 13, no. 6, pp. 52–65, 2020.

[25] M. Shepperd, "A critique of cyclomatic complexity as a software metric," Software Engineering Journal, vol. 3, no. 2, pp. 30–36, 1988.