NEARDATA

Funded by the European Union

HORIZON EUROPE FRAMEWORK PROGRAMME

NEARDATA

(grant agreement No 101092644)

Extreme Near-Data Processing Platform

D6.2 Communication and standardization report

Due date of deliverable: 30-04-2024
Actual submission date: 30-04-2024

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

Summary of the document

Document Type	Report
Dissemination level	Public
State	v1.0
Number of pages	38
WP/Task related to this document	WP6 / T6.1, T6.2, T6.3
WP/Task responsible	SCO
Leader	Scontain UG (SCO)
Technical Manager	Vanesa Ruana (URV)
Quality Manager	Christof Fetzer (SCO)
Author(s)	Andre Miguel (SCO), Xavier Roca (URV), Vanesa Ruana (URV), Raúl Gracia (DELL), Sean Ahearne (DELL)
Partner(s) Contributing	SCO, URV, DELL
Document ID	NEARDATA_D6.2_Public.pdf
Abstract	Description of the first months of standardization activities with a summary of the main achievements of the project partners. Description of the dissemination and community involvement activities with lessons learned and progress reporting.
Keywords	communication, dissemination, engagement, near data processing, OMICs, genomics, transcriptomics, metabolomics.

History of changes

Version	Date	Author	Summary of changes
0.1	19-03-2024	Andre Miguel (SCO)	First draft.
0.2	13-04-2024	Andre Miguel (SCO), Xavier Roca (URV), Vanesa Ruana (URV), Raúl Gracia (DELL), Sean Ahearne (DELL), Max Kirchner (NCT), Aaron Call (BSC), André Martin (TUD), Paolo Ribeca (UKHSA), Maciej Malawski (SANO)	Contributions.
0.5	15-04-2024	Vanesa Ruana (URV), Christof Fetzer (SCO)	Internal review of the deliverable
1.0	30-04-2024	Andre Miguel (SCO)	Final version.

Contents

List of Abbreviations and Acronyms

AI	Artificial Intelligence
AMR	Anti-Microbial Resistance
API	Application Programming Interface
AWS	Amazon Web Services
BDVA	Big Data Value Association
BSC	Barcelona Supercomputing Center-Centro Nacional De Supercomputacion
CEC	Cloud-Edge Continuum workshop
CFP	Call For Papers
CPU	Central Processing Unit
CTF	Communication Task Force
D&E	Dissemination and Exploitation
DBMS	Database Management System
DELL	EMC Information Systems International Unlimited Company
DNA	Deoxyribonucleic Acid
DSN	Dependable Systems and Networks
EBVDF	European Big Data Value Forum
ECS	Elastic Container Service
EMBL	European Molecular Biology Laboratory
EU	European Union
GCP	Google Cloud Platform
HDFS	Hadoop Distributed File System
HPC	High-Performance Computing
ICNP	International Conference on Network Protocols
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
KIO	KIO Networks España S.A
KPI	Key-Performance Indicators
LLD	Last-Level Defense
MWC	Mobile World Congress
NCBI	National Center for Biotechnology Information

NCT	Deutsches Krebsforschungszentrum Heidelberg (German Cancer Research Center)
NFS	Network File System
NTU	Nanyang Technological University
OMICs	The set with genomics, transcriptomics and metabolomics
OR	Operation Room
OSDI	Operating Systems Design and Implementation
PoC	Proof of Concept
PR	Press Release
QMR	Quarterly Management Report
R&I	Research and Innovation
RNA	Ribonucleic Acid
SANO	Centre for Computational Medicine
SCO	Scontain UG
SEV	Secure Encrypted Virtualization
SGX	Software Guard eXtensions
SNP	Secure Nested Paging
SNSF	Intelligent Serverless and Cloud Applications Symposium
TEE	Trusted Execution Environment
TPC	Technical Program Committee
TUD	Technische Universität Dresden
UKHS	Department of Health of the United Kingdom
URV	Universitat Rovira i Virgili
WORKS	Workshop on Workflows in Support of Large-Scale Science
WoSC	Workshop on Serverless Computing

1 Executive summary

This deliverable aims to present the communication and dissemination strategies carried out during the first half of the project. Additionally, we present the standardization and interoperability activities on the software technologies that make up the project. Finally, we detail the exploitation plans on the progress and results of the project.

During the first months, the NEARDATA project consortium has focused on dissemination and communication activities to make the project known to different audiences (Scientific community, Industry and General Public). We present different tables with all the events and publications carried out, emphasizing those that have had a greater impact on the project. Next, we introduce the dissemination activities planned for the following months.

Finally, we present the standardization activities to understand the interoperability of the NEARDATA platform software technologies. Also, as the development of the project is as expected, each partner presents its exploitation plan with the progress made during the first half of the project along with the expected results for the end of the project.

2 Dissemination and Communication Activities

This section presents a report and updated communication strategies of the NEARDATA project. It provides a description of the activities performed until month 16 and details dissemination plans for month 16 - 36. This report oversees the implementation of the initial strategy outlined in "*D6.1 Communication plan*" (M6) to ensure that the project outcomes effectively reach diverse target audiences, facilitating their utilization and repurposing for the broader community's advantage.

Since the start of the NEARDATA project on January 1, 2023, the consortium has proactively raised awareness of the project and engaged with its target audience using the strategies and platforms for initial communication and dissemination described in "*D6.1 Communication plan*". The first part of the project has focused on establishing a distinctive identity for the project and leveraging various tools and channels to connect with potential stakeholders. This deliverable, "*D6.2 Communication and Standardization Report*", presents a detailed description of the activities carried out from M6-M16 and provides an update on the work planned to be carried out in the coming months.

2.1 Communication strategy

To implement the communication strategy, five main initiatives were established to facilitate engagement with different audiences: 1) project brand and identity 2) project website, 3) social networks, 4) promotional material and 5) press strategy. The consortium has exerted significant efforts to ensure that communication materials and tools are consistently refreshed with information spanning from highly technical content to content tailored for the general public.

2.1.1 Project brand and identity

Developing a distinct brand and establishing a project identity served as a crucial initial step in enhancing visibility and recognition for the project. To ensure the creation of a memorable and cohesive brand, the dissemination team of WP6 worked on a design to conceptualize and craft a unified graphic identity.

The brand guidelines, logos, and templates have been disseminated to all consortium members through the internal repository. All created materials feature the project name, website, EU acknowledgment, and a version of the project logo, all adhering to the project's signature colors and style. A comprehensive description of the graphic identity is accessible in the deliverable *D6.1 Communication plan*. Project partners have consistently integrated the project brand into their communication and dissemination endeavors.

2.1.2 Project website

The project's website can be found at `https://neardata.eu/`. Managed by the Universitat Rovira i Virgili (URV), it is updated to accurately reflect the consortium's endeavors. Serving as the primary public communication platform, the website offers comprehensive details about the project, including information about partners, branding, objectives, use cases, project outcomes (such as publications, deliverables, and demos), as well as news and events.

2.1.3 Social Networks

The project maintains a presence on two social media platforms: one on X (formerly known as Twitter) and another on LinkedIn. These platforms serve as avenues to heighten awareness regarding the project's initiatives and to drive traffic to the project website. Additionally, social media serves as a vital tool for broadening communication outreach to diverse audiences and fostering engagement with stakeholders in the political and industrial sectors.

Each social media platform is strategically utilized to reach distinct target demographics and convey tailored messages. LinkedIn serves as a platform to connect with industry professionals and researchers, facilitating networking opportunities. On the other hand, X (formerly Twitter) is leveraged to engage with a wider audience, emphasizing partner involvement in events and highlighting updates to the project website. Due to its character limit, X is particularly suited for concise and less technical announcements.

For both channels, we follow hashtags #compute #continuum #extremedata #cloud #edge #HorizonEurope. We engage with and follow projects from our cluster (HORIZON-CL4-2022-DATA-01-05) and the Xartec Salut, and the Big Data Value Association (BDVA), among others.

TWITTER. With its vast user base of 368 million individuals, this platform provides the project with a global reach spanning various demographics. Utilizing this channel enables the project to connect with diverse audiences, including the general public, policymakers, and industry stakeholders. The project's Twitter account has been operational since the project's inception and has achieved moderate success. Figure 1 shows the official NEARDATA account on the X platform (ex-Twitter). In M16, the account has amassed 45 followers.

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

The relatively modest follower count may stem from several factors not accounted for in *D6.1 Communication plan*. Firstly, since the project's initiation, this platform has undergone shifts in ownership, business strategies, and operational models, resulting in turbulent changes and conflicting agendas. Most recently, the platform underwent a rebranding from Twitter to X, potentially causing some organizations and individuals to reduce their usage or altogether abandon the newly branded X platform.

The project has relied on X analytics to evaluate the effectiveness of its communication strategy. Nonetheless, the availability of analytical tools is somewhat limited, and comparisons are constrained to three-month historical data. This limitation hampers our ability to conduct comprehensive analyses of success trends over an extended period.

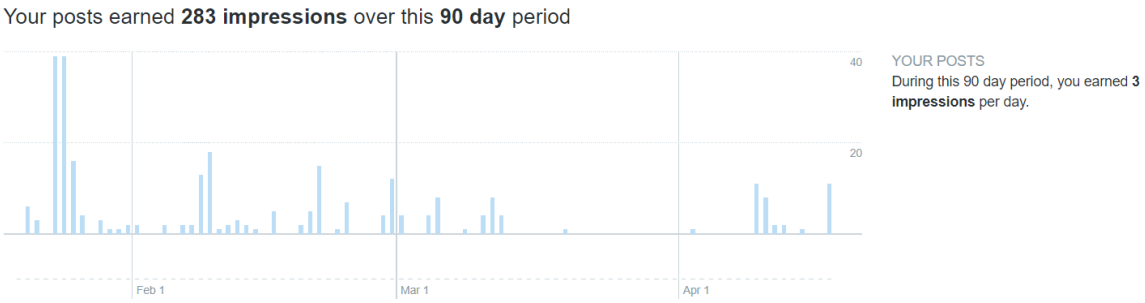

Figure 2: Twitter Analytics Providing Snapshot of the last 3 months impressions

Figure 2 shows the twitter analytics in relation to the impressions received on the official NEARDATA account in the last three months. The Twitter channel of the NEARDATA project garnered 283 impressions. The spikes in impression counts predominantly aligned with specific social media initiatives. For instance, the peaks observed in January coincide with tweets regarding NEARDATA consortium meeting, with several tweets strategically deployed to the meeting, which took place on 22-23 January 2024. Another notable surge occurred in February 2024, corresponding to the launch of the NEARDATA video. Figure 3 shows two examples of Twitter posts.

The dissemination team for WP6 will persist in tracking the analytics of both social media platforms and adjust as necessary to guarantee the dissemination of pertinent and current information to our designated audiences.

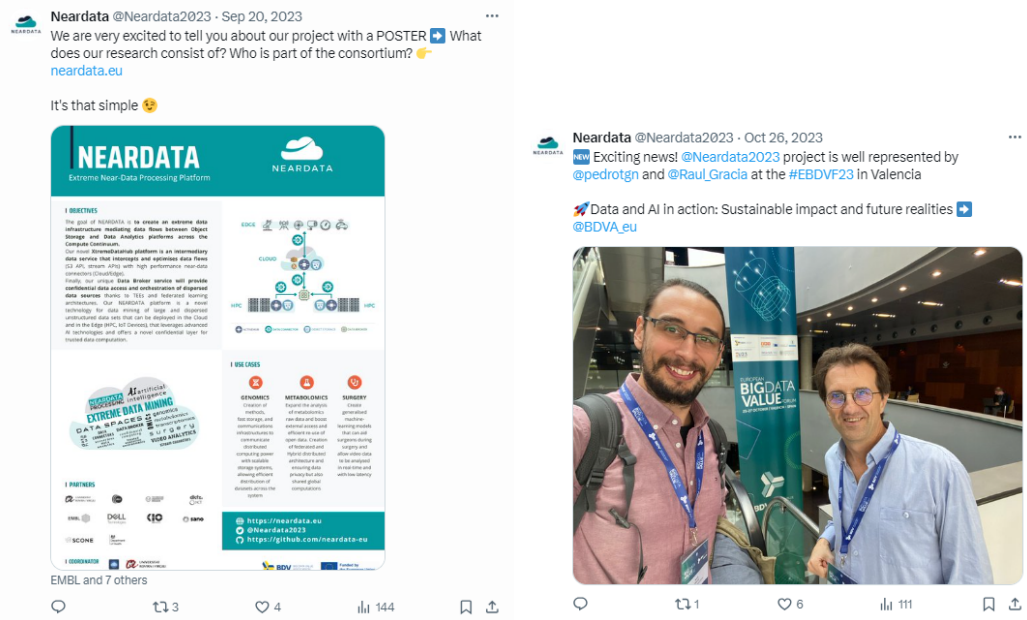

Figure 3: Examples of Twitter posts.

LINKEDIN. Linkedin is the other social media channel the project uses. Figure 4 presents the official NEARDATA account on the LinkedIn platform. Through this platform, we disseminate updates about our website, engage with the EU AI and research ecosystem, and connect with industry professionals. Additionally, we showcase the achievements and contributions of our consortium members. Our project regularly shares news and updates regarding our participation in various events, aiming to foster engagement with stakeholders from industry and academia who share similar interests. Figure 5 depicts two examples of posts on NEARDATA's LinkedIn account.

Figure 4: `https://www.linkedin.com/company/neardata-eu/`

Several project partners and other friendly organizations also use social media to share information about the NEARDATA project. Leveraging institutional networks helps the project leverage the reach of social networks. Figure 6 shows a repost of a NEARDATA LinkedIn post in the Xartec Salut institutional account.

2.1.4 Promotional material

We have developed communication and dissemination materials to assist partners in efficiently, consistently, and accurately (including proper acknowledgment and branding) creating and disseminating materials. These resources will enable them to effectively promote awareness of the project and its objectives in an engaging manner.

VIDEO. Promotional videos are a compelling and succinct method for showcasing the project. The first NEARDATA video[1] 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 7 depicts video snapshots from the first NEARDATA video.

[1] `https://www.youtube.com/watch?v=K37aMlcqPHw&t=3s`

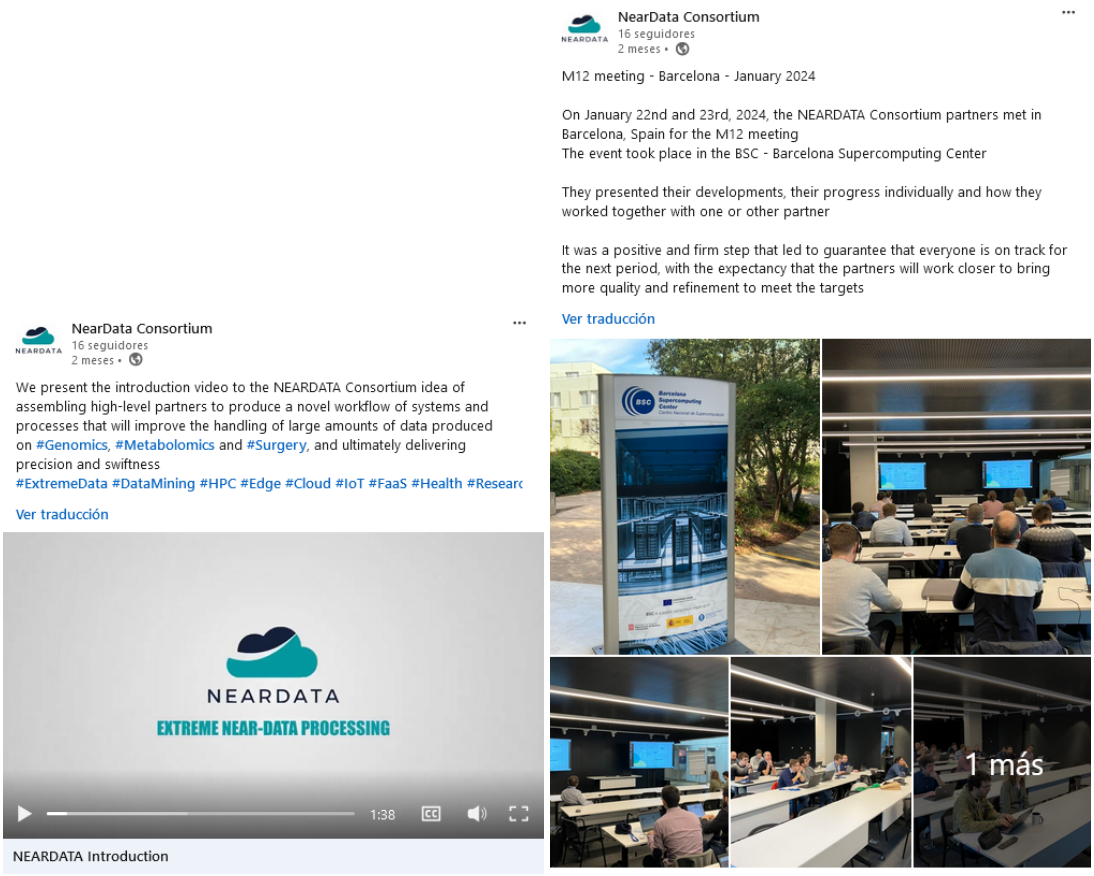

Figure 5: Examples of Linkedin posts.

Figure 6: Example of a repost of a NEARDATA post by Xartec Salut with 2.268 followers

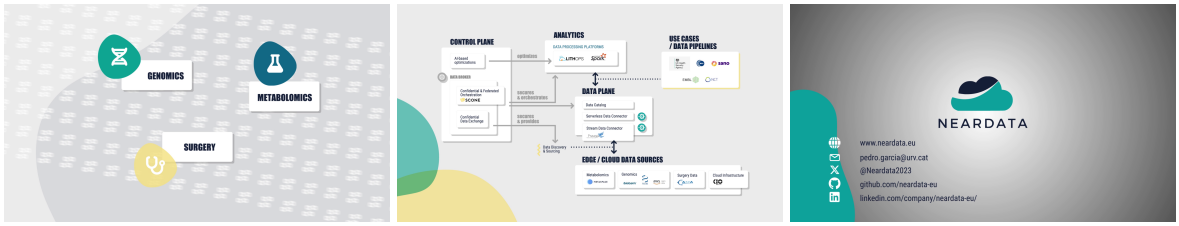

Figure 7: NEARDATA promotional video snapshots.

FLYER. A brochure has been developed that offers complete information about the project, its objectives and the implemented use cases, therefore produced to be very clear and ease the understanding of specific and general public alike. Figure 8 shows the NEARDATA brochure. They have been used in multiple events, including the Cloud-Edge Continuum Workshop, European Big Data Value Forum and Mobile World Congress, among others.

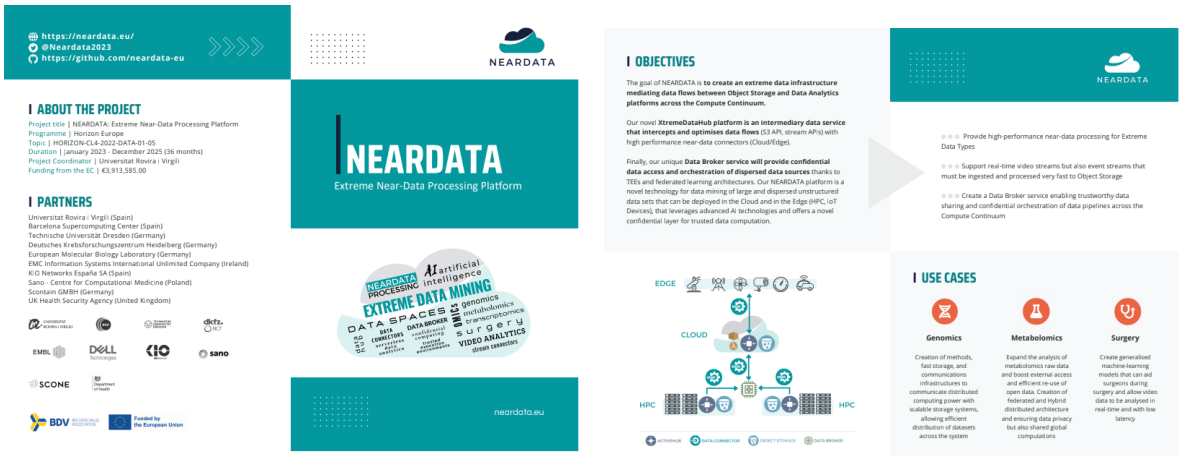

Figure 8: NEARDATA Brochure

POSTER. A comprehensive overview display was created for all partners to utilize. This display features project goals, partner information, software and use cases. Partners are encouraged to incorporate this display into their own communication efforts. Partner DELL printed the display for use in the Cloud-Edge Continuum Workshop' among others uses. Figure 9 depicts the NEARDATA poster.

OTHER MATERIAL. Presentation templates, quarterly management report templates (QMR), logos and the stylebook (containing font types) are available at the consortium's shared cloud drive. Figure 10 represents the NEARDATA slides template.

2.1.5 Press strategy

As part of the communication strategy for the NEARDATA project, a press strategy was considered to disseminate project updates and achievements to technical media outlets. This strategy involves the issuance of several press releases that serve as an effective means of conveying information about the project to technical media and their respective audiences. Furthermore, they can serve as valuable reference materials and foundational documents for project partners to utilize in their communications.

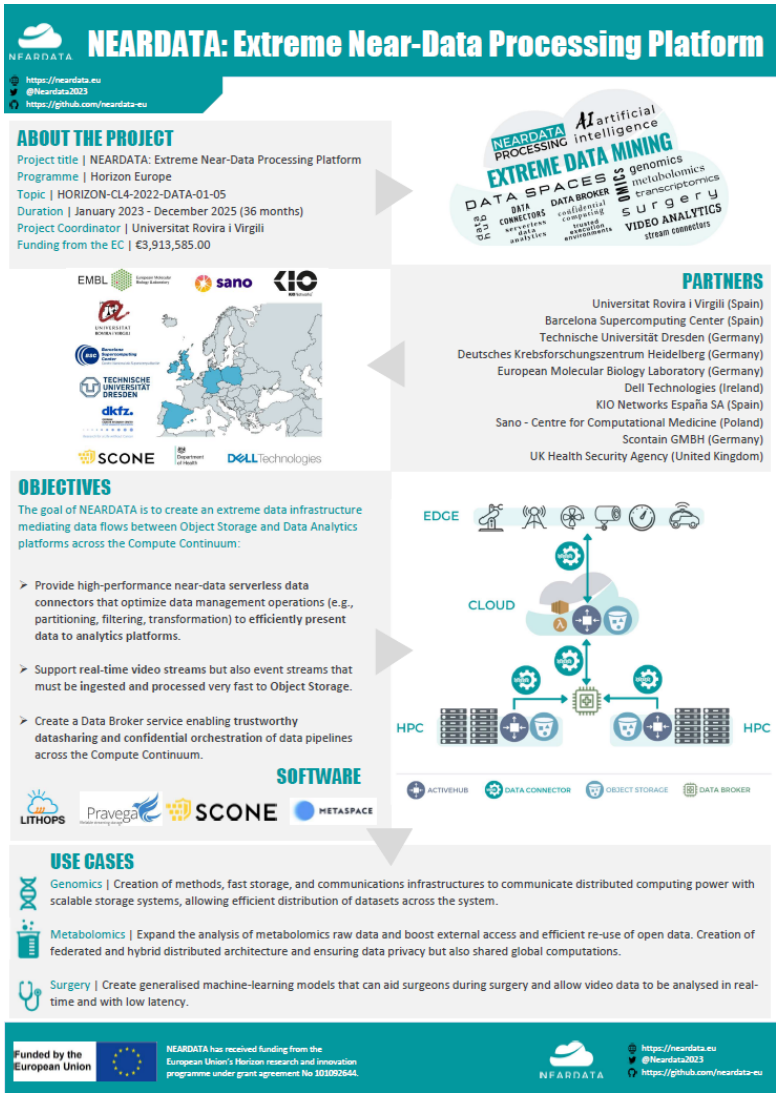

Figure 9: NEARDATA poster

As an example, three press releases published in different media are detailed below. Figure 11 depicts the press release published on the Xartec Salut website explains how the URV's projects, including NEARDATA, have a significant impact in the field of Health. The Xartec Salut network[2] is made up of 81 research groups that belong to 23 different institutions and it aims to be a catalyst for R+D+I in the field of HealthTech.

Figure 12 represents another example is the press release on the Sano website[3], explaining the work done by Sano on the project and how they successfully demonstrated the demos during the M6 NEARDATA meeting at the Dresden University of Technology, on the June 12-13, 2023.

Figure 13 shows another press release related to the project is published by DELL on the Pravega website[4] where they explain how Pravega has been included as part of the architecture of the NEARDATA projects.

2.2 Dissemination strategy

The project's dissemination strategy extends beyond merely communicating its goals and objectives, aiming to actively share and promote project outcomes with external stakeholders. The consortium

[2] https://xartecsalut.com/cloud-data-technologies-revolutionizing-healthcare/
[3] https://sano.science/m6-neardata-meeting-in-dresden/
[4] https://cncf.pravega.io/blog/2023/06/06/pravega-in-european-research-projects/

Figure 10: Slides template

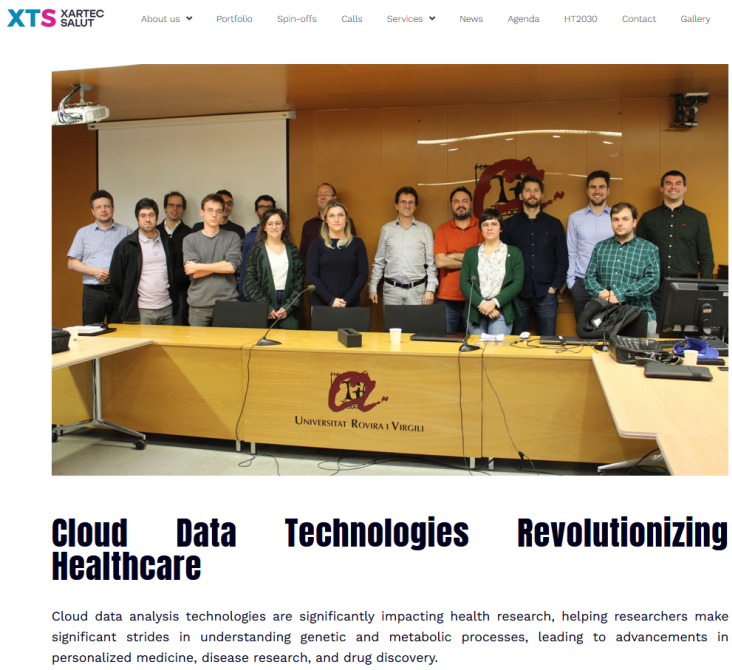

Figure 11: Press release on Xartec Salut website

has utilized various communication channels to engage with diverse stakeholders interested in leveraging NEARDATA technology. This includes research and innovation organizations throughout Europe, as well as industry players, technology providers, and service providers.

While the initial phase of the project primarily focused on executing and monitoring the communication strategy to raise awareness about the project's objectives, efforts were also made to advance the dissemination strategy. This strategy primarily revolves around disseminating project outcomes

M6 NEARDATA meeting in Dresden

It has been only 6 months since the launch of the NEARDATA project, but Sano already has demos of working software! Members of the Extreme-scale Data and Computing team: Maciej Malawski, Piotr Kica, Sabina Licholai, and Jan Przybyszewski, successfully demonstrated the demos during the M6 NEARDATA meeting at the Dresden University of Technology, 12-13 June 2023.

Figure 12: Press release on Sano website

Pravega About Pravega Getting Started

Pravega in European Research Projects

By Raúl Gracia on June 6, 2023 in News/Updates Use Cases

0 SHARES

Over the last 40 years, the European Union has built a powerful research framework through a variety of research programmes, such as FP1-9, Horizon 2020, and more recently, Horizon Europe, among others [1]. Research programmes are organized into calls that address timely and relevant societal, economic, and cultural challenges in the European landscape, including health care, aerospace research, agriculture, and automotive industry, to name a few. And, of course, addressing open challenges regarding IT and digital transformation has been one of the key areas of this European research strategy right from its inception. It is important to note that European research projects provide unique opportunities for academic institutions, companies, and startups to team up and work together towards a shared objective and a concrete set of outcomes that, in general, are open and accessible to the public. This clearly aligns with the spirit of open-source projects like Pravega.

Recently, Pravega has been included as part of the architecture of two new European research projects: NEARDATA (101092644) and CLOUDSKIN (101092646). In these projects, we collaborate with strong partners in the European ecosystem, including IBM, Imperial College London, Technical University of Dresden, European Molecular Biology Laboratory, and Barcelona Supercomputing

Figure 13: Press release on Pravega website

through participation in academic, industrial, and European AI ecosystem events, as well as through publications. Numerous meetings were convened between the project team and interested stakeholders, as well as with other EU-funded cluster projects, to explore the potential impact of project

results and foster future collaborations. Additionally, project findings were shared through keynote presentations and workshops.

Throughout Months 1 to 16, the consortium actively participated in 85 events and contributed to 17 publications. These endeavors laid the groundwork for the project, conveying its research direction and innovative aspects. During Months 17 to 24, these activities will be intensified as project results mature, with a focus on reaching out to communities that stand to benefit the most.

In the final phase of the project (Months 25 to 36), the WP6 team will proactively promote the adoption of project technologies to ensure long-term sustainability. This will involve conducting tutorials, participating in industry events, and forming strategic partnerships within the NEARDATA community and beyond, aiming for widespread dissemination and practical application of project outcomes.

2.2.1 Event Participation

Events play a pivotal role as a primary avenue for dissemination. Particularly, prestigious peer-reviewed conferences serve as platforms for the consortium to showcase the most recent advancements of the projects and engage various audiences in technology-focused dialogues from an early stage. A comprehensive list of dissemination events was compiled in the deliverable *D6.1 Communication plan*, and the WP6 team is actively seeking out additional opportunities that enable the consortium to connect with its intended audiences.

Between months 1 and 16, the consortium participated in 85 events aimed at academic, research, industrial and general public audiences. Below, we present highlights of the main contributions. Complete information about these events, including their details, can be accessed in the dissemination Appendix 6.1 Dissemination and Meeting Activities.

CLOUD-EDGE CONTINUUM WORKSHOP 2023 – CEC'23. This is the M12 open workshop of the project. Dell Technologies is making an important effort in disseminating research activities and promoting new partnerships that are beneficial for NEARDATA. Scientific conferences and workshops are an integral part of such activities. In addition to presenting our work in peer-reviewed venues, this year we set out to organize a workshop of our own that aligns with Dell's strategic research themes. The workshop, named 2023 Cloud Edge Continuum (CEC'23[5]) Workshop, was held on October 10th, 2023, in Reykjavik, Iceland and was co-located with 31st IEEE International Conference on Network Protocols (IEEE ICNP 2023[6]), which is a well-established scientific conference in the field of computer networks.

CEC'23 was supported by Dell Technologies and several European research projects that Dell participates in. These projects are funded by the EU commission and are conducting research in the areas of cloud, multi-cloud, and edge computing. Dissemination and publications in scientific events are a key activity and objective defined in these projects. To amplify the dissemination efforts of the team across multiple European research projects, we decided to organize a scientific workshop and bring together researchers and practitioners from academia and industry to discuss the latest research, trends, and challenges in ecosystems and environments based on Cloud-Edge Continuum paradigm.

The key themes of the workshop focused on challenges in a Cloud-Edge continuum revolving around AI-enabled resource allocation, security, energy and carbon footprint, system architectures, confidence, and optimization. The workshop program comprised of two keynote speeches, 12 presentations from peer-reviewed accepted papers, and 8 posters based on work and research from different EU research projects. Figure 14 represents pictures of the CEC Workshop (October 2023) organized by Dell Technologies.

Outcomes from the workshop

- Organization of a workshop focused on a key strategic area of Dell Technologies (Multi-cloud,

[5] https://cec23.github.io/

[6] https://icnp23.cs.ucr.edu/

Edge) and generating novel ideas and solutions in this domain.

- Building a technical program committee (TPC) for the workshop with 36 senior engineers and researchers from academia and industry.
- Setting up a public webpage for the workshop[7] and advertising the workshop call for papers (CFP) through various channels.
- The workshop accepted 10 original research papers (out of 13 submissions). 2 more papers were added to the program from a co-located workshop on a similar topic.
- The technical program committee carried out 30+ paper reviews to grade, select, and suggest improvements to the original submissions.
- 2 inspiring talks from keynote speakers on life and challenges at Edge. Prof. Cormac Sreenan (Head of Computer Science School, University College Cork) and Steve Todd (Fellow, Dell Technologies).
- 8 posters were displayed and presented at an interactive poster session during the workshop. The posters shared research, use-cases, challenges, and novel designs being developed in various EU projects.
- There was 25+ attendees in the CEC'23 workshop from various countries and backgrounds.
- The peer-reviewed accepted papers are being processed to appear in the IEEE Xplore proceedings. The paper titles and authors can be found in the Program section of the workshop web page[8].

Figure 14: Pictures of the CEC Workshop (October 2023) organized by Dell Technologies.

EUROPEAN BIG DATA VALUE FORUM – EBDVF 2023. The European Big Data Value Forum (EBDVF 2023) took place on 25-27 October 2023 in Valencia, Spain and counted with +400 speakers. It was a good opportunity for networking and let important players (Siemens, Amazon, IDC, to name a few[9]) in the big data market to learn about NEARDATA. One particular outcome from attending

[7] https://cec23.github.io
[8] https://cec23.github.io/index.html#program
[9] https://european-big-data-value-forum.eu/2023-edition/sponsors-and-partners/

this event was establishing new links with HPC and companies players around opportunities to the upcoming calls on Digital Twins for Destination Earth. Figure 15 depicts pictures of EBDFV 2023.

Figure 15: Pictures of EBDVF 2023 (October 2023).

MOBILE WORLD CONGRESS 2024 – MWC24. An important event where NEARDATA was presented is the Mobile World Congress 2024 (MWC24), one of the major international industrial events, that took place between 26th and 29th of February 2024 in Barcelona, Spain. Two partners, BSC and DELL, had a booth where the NEARDATA project was disseminated. According to the event organization: MWC convened over 88,500 attendees from 202 countries and territories, including policymakers and business leaders from the mobile ecosystem and beyond. Figure 15 shows pictures of Dell Technologies stand at MWC'24.

Figure 16: Pictures of Dell Technologies stand at MWC'24 (February 2024).

2.2.2 Other Activities

Besides these important activities described in previous sections, consortium partners attended several events and meetings, or wrote news, blog posts and white papers, where the NEARDATA project was presented. The full list of activities is presented in Appendix 6.1 Dissemination and Meeting Activities.

2.2.3 Publications

Peer-reviewed publications are an important way for the consortium to disseminate its findings and to ensure that NEARDATA technical results are shared openly and as widely as possible. This section presents the important conferences that had the participation of NEARDATA's partners. From M1-M16, 17 writings were published. Below, we will highlight some of them. We present the list with all the publications made in Appendix 6.2 Publications Released.

Congress "Middleware'23: Proceedings of the 24th International Middleware Conference". 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[10]. The following the papers were presented:

- **Glider: Serverless Ephemeral Stateful Near-Data Computation**[11]
  Abstract: Serverless data analytics generate a large amount of intermediate data during computation stages. However, serverless functions, which are short-lived and lack direct communication, face significant challenges in managing this data effectively. The traditional approach of using object storage to carry the data proves to be slow and costly, as it involves constant movement of data back and forth. Although specialized ephemeral storage solutions have been developed to address this issue, they fail to tackle the fundamental challenge of minimizing data movements. This work focuses on incorporating near-data computation into an ephemeral storage system to reduce the volume of transferred data in serverless analytics. We present Glider with the aim to enhance communication between serverless compute stages, allowing data to smoothly "glide" through the processing pipeline instead of bouncing between different services. Glider achieves this by leveraging stateful near-data execution of complex data-bound operations and an efficient I/O streaming interface. Under evaluation, it reduces data transfers by up to 99.7%, improves storage utilization by up to 99.8%, and enhances performance by up to 2.7×. In sum, Glider improves serverless data analytics by optimizing data movement, streamlining processing, and avoiding redundant transfers.

- **Scaling a Variant Calling Genomics Pipeline with FaaS**[12]
  Abstract: With the escalating complexity and volume of genomic data, the capacity of biology institutions' HPC faces limitations. While the Cloud presents a viable solution for short-term elasticity, its intricacies pose challenges for bioinformatics users. Alternatively, serverless computing allows for workload scalability with minimal developer burden. However, porting a scientific application to serverless is not a straightforward process. In this article, we present a Variant Calling genomics pipeline migrated from single-node HPC to a serverless architecture. We describe the inherent challenges of this approach and the engineering efforts required to achieve scalability. We contribute by open-sourcing the pipeline for future systems research and as a scalable user-friendly tool for the bioinformatics community.**??**

- **Practical Storage-Compute Elasticity for Stream Data Processing**[13]
  Abstract: Stream processing pipelines need to handle workload fluctuations (e.g., daily patterns, popularity spikes) by scaling up/down the resources contributed to running jobs. While there have been efforts proposing auto-scaling mechanisms for stream processing engines, prior work has overlooked the role of the storage system in ingesting and serving stream data. The absence of effective scaling for data streams is problematic given that the number of parallel partitions of a data stream limits both streaming data ingestion throughput and read parallelism for downstream streaming jobs. In this paper, we propose to augment the auto-scaling notion of stream processing engines with information about the source data stream. The key novelty of our approach lies in exploiting elastic data streams to ingest data, which is a unique feature of Pravega: a storage system for data streams part of the Dell's Streaming Data Platform. Pravega streams can dynamically change their parallelism based on the ingestion workload, and such information can in turn be exploited for auto-scaling the streaming job downstream. To this end, we have developed an Apache Flink connector for Pravega, as well as an auto-scaling orchestrator that feeds on data stream metrics. Our experiments show how a stream processing pipeline auto-scales by coordinating data stream and processing parallelism under workload fluctuations, with low operations cost.

[10]https://middleware-conf.github.io/2023/
[11]https://doi.org/10.1145/3590140.3629119
[12]https://dl.acm.org/doi/10.1145/3631295.3631403
[13]https://dl.acm.org/doi/10.1145/3626562.3626828

- **Pravega: A Tiered Storage System for Data Streams (Best Paper Award)**[14]
  Abstract: The growing popularity of the data stream abstraction entails new challenging requirements when it comes to data ingestion and storage. Many organizations expect to retain data streams for extended periods of time and to store such stream data in a cost-effective manner. It is also crucial to reconcile apparently opposite properties, like data durability and consistency, along with high performance. Furthermore, data streams should not only deal with a high degree of parallelism, but also adapt to fluctuating workloads with little or no admin intervention. To our knowledge, no storage system for data streams fully copes with all these requirements.

- **Trustworthy confidential virtual machines for the masses**[15]
  Abstract: Confidential computing alleviates the concerns of distrustful customers by removing the cloud provider from their trusted computing base and resolves their disincentive to migrate their workloads to the cloud. This is facilitated by new hardware extensions, like AMD's SEV Secure Nested Paging (SEV-SNP), which can run a whole virtual machine with confidentiality and integrity protection against a potentially malicious hypervisor owned by an untrusted cloud provider. However, the assurance of such protection to either the service providers deploying sensitive workloads or the end-users passing sensitive data to services requires sending proof to the interested parties. Service providers can retrieve such proof by performing remote attestation while end-users have typically no means to acquire this proof or validate its correctness and therefore have to rely on the trustworthiness of the service providers.

- **SinClave: Hardware-assisted Singletons for TEEs"**[16]
  Abstract: For trusted execution environments (TEEs), remote attestation permits establishing trust in software executed on a remote host. It requires that the measurement of a remote TEE is both complete and fresh: We need to measure all aspects that might determine the behavior of an application, and this measurement has to be reasonably fresh. Performing measurements only at the start of a TEE simplifies the attestation but enables "reuse" attacks of enclaves. We demonstrate how to perform such reuse attacks for different TEE frameworks. We also show how to address this issue by enforcing freshness – through the concept of a singleton enclave – and completeness of the measurements. Completeness of measurements is not trivial since the secrets provisioned to an enclave and the content of the filesystem can both affect the behavior of the software, i.e., can be used to mount reuse attacks. We present mechanisms to include measurements of these two components in the remote attestation. Our evaluation based on real-world applications shows that our approach incurs only negligible overhead ranging from 1.03% to 13.2%.

- **A Last-Level Defense for Application Integrity and Confidentiality**[17]
  Abstract: Our objective is to protect the integrity and confidentiality of applications operating in untrusted environments. Trusted Execution Environments (TEEs) are not a panacea. Hardware TEEs fail to protect applications against Sybil, Fork and Rollback Attacks and, consequently, fail to preserve the consistency and integrity of applications. We introduce a novel system, LLD, that enforces the integrity and consistency of applications in a transparent and scalable fashion. Our solution augments TEEs with instantiation control and rollback protection. Instantiation control, enforced with TEE-supported leases, mitigates Sybil/Fork Attacks without incurring the high costs of solving crypto-puzzles. Our rollback detection mechanism does not need excessive replication, nor does it sacrifice durability. We show that implementing these functionalities in the LLD runtime automatically protects applications and services such as a popular DBMS.

[14] https://dl.acm.org/doi/10.1145/3590140.3629113
[15] https://dl.acm.org/doi/10.1145/3590140.3629124
[16] https://dl.acm.org/doi/abs/10.1145/3590140.3629107
[17] https://arxiv.org/abs/2311.06154

Conference "Supercomputing 23 - SC23: International Conference for High Performance Computing, Networking, Storage, and Analysis". Traditional forum for bringing together the top minds in High Performance Computing, Networking, Storage, and Analysis to share ideas, contributions, and advancements in the fields that support HPC[18]. The following the papers were presented:

- **Transcriptomics Atlas Pipeline Cloud vs HPC**[19]
  Abstract: Transcriptomics studies the RNA present in a specific cell or tissue at a given time or condition. This dependence on time makes the problem computationally challenging, as the data generated by transcriptomics experiments is larger than the genomics studies on DNA sequences. The goal of the Transcriptomics Atlas project is to create a database of analyzed RNA sequences corresponding to given tissue and organ types based on the data from public repositories and make it available for researchers. We describe our transcriptomics atlas pipeline as an example of a new data- and compute-intensive scientific workflow. After analyzing the requirements of the tasks in the pipeline, we describe our proposed cloud architecture. We present the preliminary results of the experimental evaluation of the pipeline in the AWS cloud, and compare the performance results to the traditional execution on the HPC cluster.

- **Novel Approaches Toward Scalable Composable Workflows in Hyper-Heterogeneous Computing Environments**[20]
  Abstract: The annual Workshop on Workflows in Support of Large-Scale Science (WORKS) is a premier venue for the scientific workflow community to present the latest advances in research and development on the many facets of scientific workflows throughout their lifecycle. The Lightning Talks at WORKS focus on describing a novel tool, scientific workflow, or concept, which are work-in-progress and address emerging technologies and frameworks to foster discussion in the community. This paper summarizes the lightning talks at the 2023 edition of WORKS, covering five topics: leveraging large language models to build and execute workflows; developing a common workflow scheduler interface; scaling uncertainty workflow applications on exascale computing systems; evaluating a transcriptomics workflow for cloud vs. HPC systems; and best practices in migrating legacy workflows to workflow management systems.

2.2.4 Important meetings bringing NEARDATA to the industry and across the world

Meetings led by URV brought to Nanyang Technological University – NTU, from Singapore, and to Kookmin University, from Seoul, a presentation of EU projects, therefore letting NEARDATA be known across the world. July 2023 and September 2023 respectively.

Meeting with Dell's presales team to educate them in terms to disseminate NEARDATA by exploring the use-cases. November 2023.

2.2.5 Community building

Community building involves providing stakeholders with a more profound insight into project outcomes and engaging interested parties to become project stakeholders. The WP6 team, in collaboration with the entire consortium, has actively worked on identifying and expanding the NEARDATA community. Efforts were made to foster synergies and collaborations with broader European communities in AI, Data, industry, academia, and the Horizon Results Booster initiative, among others.

The NEARDATA project has engaged in two events organized by the Big Data Value Association (BDVA), namely the 'Get to Know' event [21] and European Big Data Value Forum (EBDVF 2023)[22] . Participation in these events provided valuable insights into ongoing projects and facilitated the identification of potential intersections. Furthermore, efforts have been made to collaborate with other

[18] https://sc23.supercomputing.org/attend/35-years-of-sc/
[19] https://sc23.supercomputing.org/proceedings/workshops/workshop_pages/ws_worksa104.html
[20] https://doi.org/10.1145/3624062.3626283
[21] https://bdva.eu/events/get-to-know-introductory-workshop-and-welcome-day-to-he-data-projects-wp2021-22/
[22] https://european-big-data-value-forum.eu/2023-edition/

projects within the HORIZON-CL4-2002 DATA-01-05 cluster to amplify NEARDATA's messages and explore common challenges that could be leveraged to expand each other's networks.

While the initial phase from Month 1 to Month 16 served as an introductory period to engage with communities and gain insights into their needs and preferences, the subsequent phase from Month 16 to Month 36 will focus on nurturing enduring relationships.

Horizon Results Booster. The European Commission's initiative, Horizon Results Booster (HRB)[23], is geared towards fostering a steady flow of innovation into the market and optimizing the influence of publicly funded research across the EU. Its objective is to assist projects in surpassing their Dissemination and Exploitation (D&E) obligations, guiding research endeavors towards tangible societal impact and solidifying the value of Research and Innovation (R&I) efforts in addressing societal challenges.

To achieve these goals, HRB offers free consultancy services to completed or ongoing research projects funded by the FP7, Horizon 2020 or Horizon Europe programmes. The NEARDATA project has used module A of this service with the Graph Massivizer[24] and EXTRACT[25] projects to create a portfolio dissemination and exploitation strategy in order to identify and create the portfolio of research and innovation project results. The projects group is now working on Module B to create a joint flyer.

Call HORIZON-CL4-2022-DATA-01-05 Sister projects. NEARDATA accepted the EXA4MIND project's proposal to enhance the collaborative efforts by creating a Communication Task Force (CTF) among sister projects, aimed at amplifying the visibility and impact of the projects and the respective networks. The support each other consists on social media engagement, website and newsletter features and regular collaboration calls, among others.

This initiative is made up of seven European projects from the same call: EMERALDS, NEARDATA, EFRA, EXTRACT, SYCLOPS, GRAPH-MASSIVISER and EXA4MIND.

Figure 17 presents a Twitter post of the first online meeting took place on February 29 2024.

Figure 17: First online meeting of the Sister projects

2.2.6 Science for society

It is incumbent upon the members to convey the project, its activities, and the achieved outcomes to the general public. The project partners are exerting genuine efforts to disseminate the project to the

[23] https://www.horizonresultsbooster.eu/
[24] https://graph-massivizer.eu/
[25] https://extract-project.eu/

society, some examples are:

Three demonstrations of AI-based robot-surgery, surgical training and intraoperative navigation system of liver in the field of translational surgical oncology have been presented by NCT in the **20th year of long night of science** in Dresden [26] with 200 attendees.

URV published the information and poster of the NEARDATA project in the European corner [27], as part of the **European Night of Research**. This event is celebrated every year on the last Friday of September in more than 300 cities in 30 countries across Europe. Its aim is to bring research, innovation and its protagonists, the scientists, to the public in a fun and flat way. Thus, all types of public, from schools, families and children to young people or adults of all ages, will be able to learn about and participate in the science of your territory through different activities such as workshops, talks, shows, experiments, astronomical observations and games.

URV has the **T-Systems Cloud Computing Chair**[28] whose main mission is to analyze the impact of Cloud technology on the business world and society in general, through research, training and dissemination. It seeks to deepen the knowledge about Cloud Computing and its implications in different aspects, train students and professionals in this area and bring knowledge about this technology to society in general. All this with the aim of contributing to the development and expansion of Cloud Computing in different areas of society. Along these lines, the T-Systems Cloud Computing Chair has given 21 presentations on four different cloud computing subjects, during which the NEARDATA project was presented to around 500 secondary school students in the Tarragona region. Figure 18 shows a presentation on Cloud Computing, Big Data and Artificial Intelligence that took place on September 28, 2023.

Figure 18: One of the 21 presentations of the T-Systems Cloud Computing Chair

2.3 Planned activities in the coming months

2.3.1 Cloud-Edge Continuum Workshop 2024

As we mentioned in Section 2.2.1, the organization of the Cloud-Edge Continuum Workshop 2023 was a success for the project in multiple ways. For this reason, we have applied to IEEE ICNP'24[29]

[26] https://physics-of-life.tu-dresden.de/events/2023/06/30/dresden-science-night-2023
[27] https://lanitdelarecerca.cat/neardata-extreme-near-data-processing-platform/
[28] https://www.urv.cat/ca/societat-empresa/catedres/cloud-computing/
[29] https://icnp24.cs.ucr.edu/

for repeating the workshop in this year's edition of the conference. Our proposal has been accepted and we are now starting to organize Cloud-Edge Continuum Workshop 2024, which will take place in Charleroi, Belgium on October 28-31, 2024.

2.3.2 International Workshop on Serverless Computing Experience

We are working on organizing the International Workshop on Serverless Computing Experience (WoSCx3) with IBM Watson within the "Intelligent Serverless and Cloud Applications Symposium" (SNSF) where some members of the consortium will present their research work carried out within the project. This event will take place on June 17, 2024 in Zurich, Switzerland and it will be a great opportunity to disseminate the research of our researchers to a wide audience.

3 Industry Standards, APIs, and Interoperability

3.1 Health Data Spaces

NEARDATA aims to establish three International Health Data Spaces in the fields of metabolomics, genomics and surgery to be adopted worldwide by the scientific community. The starting point of the NEARDATA platform is the reference architecture of the International Data Spaces[30]. In deliverable *D2.2 NEARDATA Architecture Specs and Early Prototypes* we have specified the requirements and new entities to develop our novel International Health Data Spaces.

We aim to interoperate with existing standards and APIs when possible. For example, in the metabolomics Data Space we provided data connectors to existing APIs in three popular data hubs (Metaspace, Metabolights, Metabolomics Workbench). In the second part of the project we will also study integrations with widely adopted standards in each field.

3.2 Industry standard APIs, data storage, containerization

At the platform level, NEARDATA strives to break silos and become broadly applicable across multiple environments by embracing *de-facto* industry standard APIs. For example, Lithops allows users to transparently execute serverless jobs across a broad variety of compute (*e.g.*, AWS, Azure, IBM Cloud, GCP, KNative) and storage (*e.g.*, AWS S3, Azure Blog Storage, IBM Object Store) back-ends. This is key, as it implies that NEARDATA will be exploited seamlessly in an technology landscape where multi-cloud is becoming increasingly important. In fact, in the project itself we have a real example of the potential of NEARDATA in this regard: METASPACE, the platform for metabolite annotation of imaging mass spectrometry data of EMBL, was executing data management and computing workloads on IBM Cloud infrastructure with Lithops. In the last months, they decided to switch to AWS for various reasons, including cost. Thanks to the multiple APIs that Lithops supports, this transition was smooth and the developer efforts for materializing it were significantly reduced.

Another example of the broad applicability of the NEARDATA platform relates to data storage. To wit, Pravega is the streaming storage engine of NEARDATA. One of the key features of Pravega is that it was the first system providing tiered storage for data streams. In terms of APIs and standards, its design allows tiering data to multiple storage services, including AWS S3, Dell ECS, Azure Blog Storage, GCP, HDFS, and NFS. This makes Pravega a perfect fit for NEARDATA, as the same streaming applications can interact with Pravega and transparently store data to different storage back-ends. In this sense, there is another standards/interoperability point that relates to the actual streaming APIs that applications use. Today, similarly to the case of AWS S3, the popularity of Apache Kafka has led the industry to assume the Kafka API as the *de-facto* standard. Companies like StreamNative, which features and maintains Apache Pulsar, are doing efforts to also support Kafka APIs in the Apache Pulsar-based platform [1]. In this sense, we are undertaking in Pravega a similar effort as other companies by providing a basic Kafka adapter for Pravega [2]. This adapter embodies an interesting avenue in which existing Kafka applications could start using NEARDATA streaming services seamlessly, without having to change their internal APIs. This strategy may be practical for boosting the applicability of the platform for legacy applications compared to trying to impose our own APIs in an industry with large and well-established players powering mature systems.

Finally, exploiting standard virtualization and containerization technologies, such as Docker, can boost the applicability of the NEARDATA platform as it becomes independent of the infrastructure. First, Lithops and Pravega can be both executed in containerized environments (*e.g.*, Kubernetes, AWS Lambda, etc.). Furthermore, Scone, the main security component of NEARDATA to provide confidential computing, also runs in containerized environments. Within a container environment like Docker, Scone can exploit Trusted Execution Environment technologies, such as Intel Software Guard eXtensions (SGX), to provide confidential computations. So far, Scone can be deployed in Azure Cloud as well as in other Kubernetes-based environments, whereas Lithops and Pravega can be deployed on almost any containerized cluster with no hardware restrictions.

[30] https://internationaldataspaces.org/

All these efforts at the NEARDATA platform level, ranging from the adoption of standard APIs to the integration of virtualization technologies, provide a great substrate to increase the exploitability and applicability of the platform across use-cases and infrastructures.

3.3 Future standards

NEARDATA is open and receptive to nascent standards, constantly striving to incorporate novel technologies in the field of computer science and, especially, ways of describing and representing rich or complex data.

To that end, NEARDATA is keeping a liaison with the ongoing standardisation work of ISO/IEC 23092 [3], also known as MPEG-G, which aims to describe and encode genomic information in a logically coherent, efficient and (re)usable way. As previously explained in other deliverables, genomic information is a perfect target for NEARDATA, as it is inherently extreme data – for instance, public repositories storing data produced as a result of the sequencing of biological organisms are increasing their size exponentially, having now reached hundreds of petabytes [4]. This poses a constant challenge to efficient data analysis, storage and distribution, and as a matter of fact the most relevant public databases, such as the NCBI Short Read Archive, have recently been moved to the cloud. NEARDATA has multiple use cases considering such data sources, explicitly providing data connectors to access them in the cloud and bring computation closer to them.

In addition, genomic information is intrinsically heterogeneous, bringing together not only sequencing data but also metadata and annotations relevant to many different fields of biology and bioinformatics. This wealth of information is usually represented in a variety of formats and databases, making its access, use and sharing problematic. As a result, MPEG-G is striving to standardise a number of use cases and provide solutions to several scenarios in the field of genomics, such as:

1. Data format and compression
2. Data streaming
3. Compressed file concatenation
4. Incremental update of sequencing data and metadata
5. Selective access to compressed data, e.g. fast queries by genomic range
6. Metadata association
7. Enforcement of privacy rules
8. Selective encryption of data and metadata
9. Annotation and linkage of genomic segments.

NEARDATA is following the ongoing work of MPEG-G with great interest, and plans to evaluate possible mutually advantageous exchanges of information and implications on its own work at some later point in the future of the two projects.

4 Explotation

Partners have thought on initial exploitation ideas arising from our current progress and what we are planned to achieve.

DELL has worked with NCT on the computer-assisted surgery use case and that has had a positive impact on Dell products. The technology being developed in this branch has translated into new demos and improvements in *Pravega* and its ecosystem (i.e. the streaming storage engine of Streaming Data Platform (a new Dell product launched in 2020). The Streaming Data Platform is being delivered in *NativeEdge*, which is Dell's new Edge platform for winning customers in that market. Moreover, data management and multi-cloud is a strategic research topic for Dell. *Lithops* has been internally disseminated as a great example of a multi-cloud serverless framework for solving data management challenges. This is raising interest in the company and it could be an influential insight for future Dell products in the data management space (as the recent *Dell Data Lakehouse* product).

One key beneficial aspect for Dell from this consortium is the exposure to cutting-edge health-related use cases (e.g., computer assisted surgery, genomics, metabolomics). As part of the dissemination work DELL has been doing in the project, internal talks have been given to architects, presales, and sales teams about how DELL is exploiting the technologies in NEARDATA to solve real problems in the health sector. These talks are raising significant interest in the company and helping to set a new standard of what Dell can do to solve complex technology challenges for customers. This is progressively crystallizing into contacts from Dell presales/sales with partners for exploring collaboration opportunities (e.g., Dell may be sponsoring the *MCAAI conference endoscopic vision challenge* for NCT).

The work DELL is doing in NEARDATA, such as the computer-assisted surgery PoC, is expanding the catalog of demos that presales and sales can use to target new customers. Similarly, DELL is also helping the company's presales and sales teams to strengthen their links with current customers, like URV and BSC, to explore new business opportunities.

NCT, in line with their works with DELL concerning *Pravega* and *GStreamer* has explored the possibilities of leveraging these systems to establish robust pipelines for data collection and processing during surgeries, aiming to significantly improve the data management processes and enhance the efficiency of storing and processing surgical videos in *real-time* and in long term prospects.

Throughout the activities and after the conclusion, NEARDATA's results could be exploited *1.)* with storage and data access during and after surgeries, as long as Pravega and GStreamer to be set up in the Operation Room – OR which is not the case currently; 2.) and to scaling computational workloads (inference of ML models as well as other processes) across the available compute continuum infrastructure, e.g., 10 ORs have simultaneously a large demand on resources and there needs to be a system to leverage the available computing resources for the tasks. Additionally, the *secure federated learning* after maturing could be exploited in case of eventual engaging in real federated learning setups involving multiple participants from various institutions.

Another significant way to exploit NEARDATA for NCT is more particularly on advancing science: this could open up new avenues for collaborative research and development. In general, the aim of NCT (the aspect of Translational Surgical Oncology) is to bring such research topics closer to application in the OR. Therefore, both current efforts and the expected outcome of the project, particularly w.r.t. Pravega (provides a flexible infrastructure for the surgical unit of tomorrow) and SCONE (securing the use of cloud applications and federated learning) should make NCT's daily research work easier.

SCO has initially benefited from NEARDATA because of the effort of porting Lithops to *SCONE*; it demanded an improvement of the product for better support of asynchronous/parallel processes executions of a program's instructions. The intensive use of futures-like library in *Lithops* presented an extreme scenario of parallel instructions execution within the scope of confidential computing,

which prompted the need for the framework's adaptation. SCO's suite of systems to support confidential computing entirely (policy driven program execution, processing, memory, network, storage) has been leveraged by the addition of *Keycloak*, an open source identity and access manager that is used to cover the aspect of user-focused security, for the cases where enforcing privacy of data has to be pushed onto deeper granularity, where some users might be allowed to perform certain activities, but forbidden to do others.

The versatility of SCONE to support existing programs and port them to the scope of confidential computing has been employed in a new and privacy sensitive industry (e.g. processing genomics data). This brings confidence to SCO on challenging themselves to prospect new markets with clients anxious for solutions of the kind developed for NEARDATA. Keycloak is an important addition to SCO's portfolio; the research done in NEARDATA on how to employ it in a public health related industry to support a workflow to grant or deny access to resources will be exploited to demonstrate that the SCONE suite can match a variety of challenges and offer solutions beyond reassuring confidentiality in programs execution. Furthermore, throughout the next stages, when more systems integrations will be made, SCO will evaluate eventual new needs for the product improvement.

URV's exploitation is focused on the creation of a new startup (datoma.cloud) that is creating an ambitious Cloud Data analytics platform for Metabolomics. Datoma leverages many of the technologies developed in NEARDATA like data connectors to different datahubs, and cloud optimized execution of OMICs pipelines. Furthermore, companies like Astra Zeneca showed interest in confidential execution of metabolomics pipelines.

During these months, DATOMA is being exposed and demonstrated to different metabolomics research groups around the world. In the second half of 2024 the project will be disseminated in a major conference in Metabolomics, and we will arrange demostrations with the major stakeholders (Pharma industry, Instrumentation companies, hospitals, research centers).

UKHSA believes genomic analysis, which is inherently an extreme-data field, to be essential for public health, as recently demonstrated by the COVID-19 pandemic, the Mpox epidemic, and the current ongoing global surveillance of avian flu. In particular, without genomic analysis one would be unable to track the emergence of novel strains and variants, which can be more transmissible or virulent and carry specific genes or mutations that make them resistant to treatment, such as anti-microbial resistance (AMR) genes in bacteria. Without genomic analysis it would also be much more difficult to track, and find the sources of, cryptic outbreaks that originate from industrial or animal sources —- genomic profiling and matching is more often than not essential to reconstruct the trajectory of an outbreak.

However, public health systems are complex and federated entities by their own nature, as they involve a number of actors working at diverse institutions — GPs, hospitals, the NHS, bioinformaticians, epidemiologists and modellers, and public decision makers. This is why the next step in the cooperation between UKHSA and NEARDATA, which is to prototype a unified approach to intervention and incident management based on real-time event-based methods such as Pravega, will benefit all the actors involved in the public response to health threats, keeping everyone synchronised even when a large amount of data is flowing into UKHSA from a large number of partners that are not geographically co-located. In addition, reimplementation of analysis workflows and operating procedures based on secure frameworks such as SCONE would drastically simplify the need for the complex data anonymisation that is currently needed when handling sequencing data and patient metadata due to their sensitivity. Ultimately, the proposed exploration of novel technical solutions for the handling of extreme data will have obvious implications in terms of better and more cost-effective public health, eventually resulting in benefits for the community as a whole.

BSC sees that the integration of serverless platforms and their integration in *Mare Nostrum* has facilitated the ingestion of the input data in the use case of Genome-Wide discovery. The vast amount of data necessary to analyse the genetics of Type 2 Diabetes represent one of the current bottlenecks

for some type of analysis. Here, within the NEARDATA project, we have found which are the best tools to approach this problem and we have applied and integrated them in Mare Nostrum 4, thus allowing the efficient split, process and merge back of the information faster than before. In addition, we have studied and found the best methodology to accelerate and parallelise the analysis of this data. Particularly, we have chosen and prepared the bioinformatic tools behind the Genome-Wide discovery analysis to use COMPSs and, therefore, to exploit the inherent parallelism of Mare Nostrum during the analysis.

There are diverse factors which limit the study of the genetics behind Type 2 Diabetes and the generation of genetic prediction models to early detect and predict the development of the disease. Among these factors, there are diverse computational challenges that we are approaching under the scope of the NEARDATA project. Particularly, although the current solutions adopted during this period of time have been only applied in small test datasets, the results obtained during the tests are the promise of breaking the current barriers and bottlenecks that surround the Genome-Wide discovery analysis. For this reason, in the short term we expect a complete integration of all the solutions suggested and, in the long term, we expect to broaden to a genome-wide level both the genomic analysis of Type 2 Diabetes and the generation of genomic predictors. This will provide the community with new computational frameworks and protocols to improve current analysis, will enhance the creation of new Type 2 Diabetes prediction models to improve current detection and prediction protocols, and will broaden the understanding of the pathophysiology of the disease.

TUD's exploitation plans revolve around *1) Technology developments*, seeking to improve performance and lower overhead when applications run in Intel SGX enclaves, and, if applicable, provide support also for other Trusted Execution Environment technologies such as ARM TrustZone or other CPU vendors such as RISC-V; *2) Business strategy* subdivided in *Academic and Research* focusing in becoming world leader for confidential compute; *Sustainability through secured funding and new R&D projects* where the artefacts of SCONE and its tools are reused in other projects such as 6G-life, CPEC, CETI; *Disseminate NEARDATA results in academia* via Confidential compute lectures, part of the DSE (Distributed Systems Engineering) programme, and PhD and master theses aligned to exploit the project results. *Scientific dissemination* by making contributions to top level academic conferences and journals such as EuroSys, DSN, OSDI etc.; *Tutorials & Documentation* production of open-source development of demonstrators and tutorials around SCONE to gain community traction; *Technology transfer* integrating into the SCONE product for commercial exploitation and consulting; and *Increased interoperability of the developments* to foster additional innovations for future proposals.

We expect the following impact together with the related business KPIs: *1)* having 2 new adopters of SCONE where SCONE is used in different application scenarios, e.g., training or inference; *2)* improved performance of SCONE in general making it applicable to a wide range of application scenarios. Furthermore, NEARDATA will enable TUD and Scontain – SCO, a startup which has been specifically created in order to transfer the results in concrete industry products, with the following approach for exploitation: *1)* Exploiting results through customisations, adoption and integration of SCONE into existing applications through subcontracting; *2)* Commercial licences for companies that use SCONE in a commercial setting.

Technology transfer will be achieved through SCO. One of the potential application domains where the approach will be leveraged is the German health system with the electronic patient record system as well as the eRezept.

SANO has gained a better understanding of the costs over using cloud for transcriptomics commutations, whilst also improved their vision on identifying potential optimizations.

In the long term, SANO expects new research topics in applications of cloud technologies for biomedical research and the possibility to analyze larger datasets to gain new knowledge and develop new treatments.

KIO NETWORKS is a company that operates Edge TIER IV Data Centers in Spain. As part of the

NEADATA project, KIO has provided IaaS (Infrastructure as a Service) resources and the necessary capabilities to create testbeds that each partner has required for their research work. In this sense, confidential computing capabilities have been made available to the SCONE/SCONTAIN project, which is being crucial for developing functionalities that NEARDATA project aspires to. KIO network will benefit from the technologies developed in NEARDATA to create new innovative services for our clients. Specifically, we are considering SCONE as a confidential stack for our container offering. This project also opens interesting possibilities to offer our infrastructure for compute intensive workloads in extreme data settings.

5 Conclusions

This deliverable presents the description of the dissemination and communication strategies, the standardization and interoperability activities of the software technologies presented in NEARDATA and the initial exploitation plans of the project by each of the partners forming the consortium.

During the first months, the consortium has made significant efforts to establish a recognizable project brand and to facilitate engagement with different audiences (Scientific community, Industry and General Public) through established communication channels, such as the project website, social media, dissemination activities and publications.

As the project is progressing according to plan, communication activities have increased favorably. We can distinguish up to 85 dissemination events and 17 publications made to date. In this deliverable we have highlighted the events and publications with the highest impact along with the list of all activities and publications.

Community building is essential to understanding the NEARDATA project. The consortium has actively worked to establish synergies and collaborations between broader European communities. The NEARDATA project has engaged in two European events organized by BDVA called "Get to Know" and EBDVF 2023. Additionally, we have made efforts to collaborate with other projects in the HORIZON-CL4-2002 DATA-01-05 cluster to expand the NEARDATA project.

During the first half of the project, efforts were focused on building a recognizable project brand, encourage involvement with different audiences and the development of the NEARDATA platform software. A summary of the communication achievements reached by the consortium is illustrated in Table 1. We expect in the second half of the project an increase in communication and dissemination activities in order to better achieve the KPIs.

Table 1: Communication Achievements

Type of communication	**Category of audience**	**Achievements**
Scientific Publications	Scientific community	17 publications
Conferences and Workshops	Scientific community and Industry	44 events
Community Building	Scientific community, Industry	4 events
Meetings outside the consortium	Scientific community, Industry	23 meetings
Events for society	General Public	29 events
Press releases	General Public	6 publications

Finally, standardization activities on the NEARDATA platform have been presented to expose the interoperability of the software technologies. Then, based on current developments, the partners established an initial exploitation analysis on the results and future developments to maximize the impact of the NEARDATA project on their own benefits.

6 Appendix

6.1 Dissemination and Meeting Activities (M1-M16)

Table 2: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Workshop. CZI Workshop	Metabolism Across Scales	14/2/23	Scientific community, General Public	EMBL. Attendance: 200
Workshop. Regional Mass Spectrometry Imaging	Keynote	27/3/23	Scientific community, General Public	EMBL. Attendance: 200
Conference. EASL Liver Cancer Summit	Conference	14/4/23	Scientific community, General Public	EMBL. Attendance: 500
Meeting. Alterna	—	14/4/23	Industry	URV. Attendance: Pedro Garcia Lopez. Jose Miguel Garcia
Meeting. ORO	Overview of Pravega and NEARDATA	20/4/23	Scientific community, Industry	DELL. Attendance: 30+
Meeting. Xartec Salut	—	2/5/23	Industry	URV. Attendance: 10
Meeting. Technology of Data Spaces	Kick Off meeting with coordinators and representatives of the HE projects of the: HORIZON-CL4-2021-DATA-01-0X topics - the DSSC Thematic Groups – BDVA	4/5/23	Scientific community	URV. Attendance: 15
Workshop. CeTI General Assembly	Demonstration the AI-based surgery scene recognition	8/5/23	Scientific community, General Public	NCT. Attendance: 100
Workshop. 6G-life General Assembly	Demonstration the concept of AI-based robotic surgery	11/5/23	Scientific community, General Public	NCT. Attendance: 200
Meeting. Xartec Salut	Presentation of projects and collaborate in dissemination	12/5/23	Scientific community	URV. Attendance: 3

Table 3: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Blog post. Xartec Salut website	Cloud Data Technologies Revolutionizing Healthcare	12/5/23	Scientific community, General Public	URV
Conference. German Society of Mass Spectrometry (DGMS)	Plenary	14/5/23	Scientific community, General Public	EMBL. Attendance: 500
Meeting. Presentation to UKHD	Possibilities of Streaming Frameworks in the Surgical Domain	16/5/23	Scientific community, General Public	NCT. Attendance: 3
Conference. SDSI Conference	Key role of data in AI	17/5/23	Scientific community, Industry, General Public	SANO. Attendance: 100
Presentation. Huawei Research Summit	Confidential Computing	31/5/23	Scientific community, General Public	TUD. Attendance: 500
Conference. Single-Cell Proteomics Conference (SCP2023)	Spatial single-cell metabolomics reveals metabolic cell states	1/6/23	Scientific community, General Public	EMBL. Attendance: 100
Conference. ASMS	ThermoFisher User Meeting	4/6/23	Scientific community, General Public	EMBL. Attendance: 1000
Workshop. Surgical and Interventional Engineering	Cognitive sensor-guided robotically assisted surgery	5/6/23	Scientific community, General Public	NCT. Attendance: 20
Presentation. Nvidia Horizon Research Paper Presentations	Confidential Computing	5/6/23	Scientific community, General Public	TUD. Attendance: 50
Workshop. Jornadas de Concurrencia y Sistemas Distribuidos – JCSD23	NEARDATA internal workshop	19/6/23	Scientific community, General Public	URV
Blog post. Information and poster of the project in the European corner	NEARDATA: Extreme Near-Data Processing Platform	19/6/23	Scientific community, General Public	URV

Table 4: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Meeting. Dell	Discussion of data reduction opportunities in NEARDATA	21/06/23	Scientific community, Industry	DELL. Attendance: 2 Philip Shilane (DELL)
Workshop. Dresden Science Night 2023	Retreat on Endoscopic Vision with SYMIC	21/6/23	Scientific community General Public	NCT. Attendance: 50
Workshop. The 20th year of long night of science in Dresden	Demonstrations of AI-based robot-surgery, surgical training and intraoperative navigation	30/6/23	Scientific community General Public	NCT. Attendance: 200
Workshop. Data Management Tech Forum	NEARDATA	10/7/23	Scientific community, Industry	DELL. Attendance: 15-20
Workshop. NCT Scientific Advisory Board Meeting	NEARDATA project was presented to the board	10/7/23	Scientific community, General Public	NCT. Attendance: 100
Conference. European Cloud, Edge & IoT initiative	Dissemination and community involvement	11/7/23	Industry, General Public	KIO
Meeting. NTU Singapore	Presentation of EU projects	12/7/23	Scientific community	URV. Attendance: 15
Meeting. 6G communication networking	Discussion on the implementation of medical testbeds for use cases of low latency, high resilience,privacy	27/7/23	Scientific community, General Public	NCT. Attendance: 15 from NCT, TU Munich and TU Dresden
Meeting. Horizon Results Booster	Kick Off meeting with coordinators and representatives of the HE projects of the: HORIZON-CL4-2021-DATA-01-0X topics - the DSSC Thematic Groups – BDVA	1/8/23	Scientific community	URV. Attendance: Alba Balla (Graph-Massivizer project) and Feredico Drago (HRB Consultant)
Meeting. IBM Research	Presentation in the scope of CloudStars	15/8/23	Scientific community, Industry, General Public	SANO

Table 5: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Workshop. 20 Years University Cancer Center NCT/UCC	Artificial intelligence in improving the outcome of surgical oncology	21/8/23	Scientific community, General Public	NCT. Attendance: 100
Workshop. Summer School on Non-targeted Metabolomics Data Mining for Biomedical Research	Workshop	21/8/23	Scientific community, General Public	EMBL. Attendance: 200
Workshop. Dell ISG Technical conference	NEARDATA	6/9/23	Scientific community, Industry	DELL
Meeting. BDVA platform	Strategy in BDVA	7/9/23	Scientific community	URV. Attendance: 2
Workshop. SECAI/CeTI International Summer School	Workshop given by S. Bodenstedt	11/9/23	Scientific community, General Public	NCT. Attendance: 100
Workshop. EMBO Workshop on Lipid Droplets	Workshop	17/9/23	Scientific community, General Public	EMBL
Meeting. University of Warsaw	HPC-Whisk	20/9/23	Scientific community, Industry	SANO
Presentation. European Researchers' Night	Cloud Computing, Big Data, Artificial Intelligence	29/9/23	Scientific community, General Public	URV. Attendance: 20
Meeting. Kookmin University	Presentation of EU projects	29/9/23	Scientific community	URV. Attendance: Kyungyong Lee (Kookmin University in South Korea)
Presentation. Nit de la Recerca a Berga	NEARDATA project to the general public	29/9/23	Scientific community, General Public	BSC. Attendance: 100
Article. Report of the 2022-2023 academic year	Memòria del curs 2022-23	29/9/23	Scientific community, General Public	URV

Table 6: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Conference. Int. Single-Cell Mass Spectrometry conference	Conference	7/10/23	Scientific community, General Public	EMBL. Attendance: 200
Workshop. Internal Dell conference	TEx12 presentation on Lithops	18/10/23	Scientific community, Industry	DELL. Attendance: 30+
Forum. BDVA Forum	Attendance and networking	25/10/23	Scientific community, Industry	DELL
Workshop. Zuse Schools Autumn Event	Keynote and Demonstrations	26/10/23	Scientific community, General Public	NCT. Attendance: 100
Congress. Thinknet 6G Summit 2023	Presentation of NCT	26/10/23	Scientific community, General Public	NCT. Attendance: 100
Workshop. Internal Dell conference	TEx12 presentation on Lithops	27/10/23	Scientific community, Industry	DELL. Attendance: 5
Attendance. European Big-data Value Forum	Data and AI inaction: Sustainable impact and future realities	27/10/23	Scientific community, Industry, General Public	URV
Workshop. DIU DresdenConference	Presentation of the use of AI and platforms	6/11/23	Scientific community, General Public	NCT. Attendance: 100
Publishing. Smart City Expo	Included NEARDATA into the BSC project's porfolio	7/11/23	Scientific community, General Public	BSC
Blog Post. Pravega	Pravega in European Research Projects	7/11/23	Scientific community, General Public	DELL
Workshop. IEEE ICNP	Cloud-Edge Continuum	7/11/23	Scientific community, General Public	DELL
Workshop. CeTI all-day programme Girls for Robots	Hands-on Experiments of computer-and-robot-assisted surgery	7/11/23	Scientific community, General Public	NCT. Attendance: 20

Table 7: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Workshop. Pravega Community meeting	Workshop	08/11/23	Scientific community, Industry	DELL. Attendance: 30+
Congress. HealthTech 2030	Congress	10/11/23	Scientific community, Industry	DELL
Workshop. WORKS23 at Supercomputing 23	Transcriptomics Atlas Pipeline: Cloud vs HPC	12/11/23	Scientific community, Industry, General Public	SANO. Attendance: 100
Presentation. Advanced Digital Technologies	I2cat, Gencat and Mobile World Capital	25/11/23	Scientific community, General Public	URV. Attendance: 50
Meeting. Dell	Engagements with presales to disseminate NEARDATA and the use-cases	–/11/23	Scientific community, Industry	DELL. Attendance: 4 different presales contacted
Congress. Middleware'23	Congress	14/12/23	Scientific community, Industry	DELL
Meeting. MICCAI FedSurg	MICCAI FedSurg 2024 Challenge Meeting	16/1/24	General Public	NCT. Attendance: 6
Workshop. Scontain	Present the progress of NEARDATA Consortium	19/1/24	Industry	SCO. Attendance: 8
Conference. DANEMO symposium	Presented METASPACE	25/1/24	Scientific community	EMBL. Attendance: 200
Workshop. LiverSeminars	Presented METASPACE	—–/1/24	Scientific community	EMBL. Attendance: 200
Publishing. X.com Posting	Manuscript link: Large-Scale Evaluation of Spatial Metabolomics Protocols and Technologies	5/2/24	Scientific community, General Public	EMBL

Table 8: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Meeting. Project Officer	Coordinator presented the project to the new PO	7/2/24	Scientific community	URV
Meeting. Internal Dell conference	ZEISS group of Health solution and Central Marketing visit at NCT	7/2/24	Industry	NCT. Attendance: 15
Meeting. Interview recordings for NEARDATA	—	12/2/24	Industry	DELL
Conference. SPIE Medical Imaging 2024	Talk at the Conference	18/2/24	Industry	NCT
Meeting. MICCAI FedSurg	MICCAI FedSurg 2024 Challenge Meeting	20/2/24	General Public	NCT. Attendance: 6
Conference. German Cancer Congress	Talk at the Conference	23/2/24	Scientific community, General Public	NCT. Attendance: 50
Conference. PITTCON	Presented METASPACE	24/2/24	Scientific community	EMBL. Attendance: 50
Congress. Mobile World Congress 2024	Presentation of NEARDATA	26/2/24	Scientific community, Industry	BSC
Congress. Mobile World Congress 2024	Attendance	26/2/24	Scientific community, Industry	DELL
Conference. Italian Society for Immunology and Clinical Allergology (SIICA)	Presented METASPACE	27/2/24	Scientific community	EMBL. Attendance: 200
Meeting. HORIZON-CL4-2022-DATA-01-05	Call HORIZON-CL4-2022-DATA-01-05 Sister projects meeting - EMERALDS, NEARDATA, EFRA, EXTRACT, SYCLOPS and EXA4MIND	29/2/24	Scientific community	URV
Conference. VIB Conference “Applied Bioinformatics in Life Sciences”	Presented METASPACE	7/3/24	Scientific community	EMBL. Attendance: 200

Table 9: Dissemination and Meeting Activities

Event Type	**Link**	**Date**	**Type of audience**	**Comments**
Workshop. Saxony SPIN2030 science festival	Talk at Workshop	8/3/24	Industry	NCT. Attendance: 200
Workshop. University College Cork	Participation in poster session as part of EU Glaciation Project dissemination event	12/3/24	Scientific community, Industry	DELL
Conference. 16th ACC Cyfronet AGH HPC users' conference	Transcriptomics Atlas Pipeline: HTC	13/3/24	Scientific community, Industry	NCT. Attendance: 50
Meeting. Educational	Student visit at NCT	14/3/24	General Public	NCT. Attendance: 200
Conference. Keystone Symposium on Immunometabolism	Presented METASPACE	18/3/24	Scientific community	EMBL. Attendance: 200
Meeting. MICCAI FedSurg	MICCAI FedSurg 2024 Challenge Meeting	19/3/24	General Public	NCT. Attendance: 6
Conference. Internal Dell conference	TEx13 presentation entitled "Edge Video Analytics with Pravega: Supporting a Computer-Assisted Surgery Use Case"	26/3/24	Scientific community, Industry	DELL. Attendance: 20
Conference. Internal Dell conference	TEx13 presentation entitled "Edge Video Analytics with Pravega: Supporting a Computer-Assisted Surgery Use Case"	26/3/24	Scientific community, Industry	DELL. Attendance: 20

6.2 Publications Released (M1-M16)

Table 10: Publications Released

Title	Authors	Publisher/Journal/ Magazine/Conference	Link
Glider: Serverless Ephemeral Stateful Near-Data Computation	Daniel Barcelona-Pons, Pedro García-López, Bernard Metzler	Middleware '23: Proceedings of the 24th International Middleware Conference	Publication
Exploiting Inherent Elasticity of Serverless in Algorithms with Unbalanced and Irregular Workloads	Gerard Finol, Pedro Garcia Lopez, Marc Sanchez Artigas	Journal of Parallel and Distributed Computing	Publication
MLLess: Achieving Cost Efficiency in Serverless Machine Learning Training	P. Gimeno Sarroca, M. Sánchez-Artigas	Journal of Parallel and Distributed Computing	Publication
A Seer Knows Best: Auto-tuned Object Storage Shuffling for Serverless Analytics	G. Eizaguirre, M. Sánchez-Artigas	Journal of Parallel and Distributed Computing	Publication
An Exhaustive Variant Interaction Analysis using Multifactor Dimensionality Reduction	Gonzalo Gómez-Sánchez, Ignasi Morán, Lorena Alonso, Miguel Ángel Pérez, David Torrents, Josep Ll. Berral	Nature Scientific Reports	Publication
Challenges and Opportunities for RISC-V Architectures towards Genomics-based Workloads	Gonzalo Gómez-Sánchez, Aaron Call, Xavier Teruel, Lorena Alonso, Ignasi Moran, Miguel Ángel Pérez, David Torrents, Josep Ll. Berral	ISC High Performance 2023	Publication
On Data Processing through the Lenses of S3 Object Lambda	P. Gimeno Sarroca, M. Sánchez-Artigas	IEEE International Conference on Computer Communications	Publication
Practical Storage-Compute Elasticity for Stream Data Processing	Raúl Gracia-Tinedo, Brian Zhou, Yimin Xiong, Luis Liu	Middleware '23: Proceedings of the 24th International Middleware Conference	Publication
Pravega: A Tiered Storage System for Data Streams	Raúl Gracia-Tinedo, Flavio Junqueira, Tom Kaitchuck, Sachin Joshi	Middleware '23: Proceedings of the 24th International Middleware Conference	Publication
SinClave: Hardware-assisted Singletons for TEEs	Franz Gregor, Robert Krahn, Do Le Quoc, Christof Fetzer	Middleware '23: Proceedings of the 24th International Middleware Conference	Publication
The Nanoservices Framework: Co-locating Microservices in the Cloud-Edge Continuum	Eric Caron, Raúl Gracia-Tinedo	IEEE ICNP'23	Publication

Table 11: Publications Released

Title	Authors	Publisher/Journal/ Magazine/Conference	Link
METASPACE-ML: Metabolite annotation for imaging mass spectrometry using machine learning	Bishoy Wadie, Lachlan Stuart, Christopher M. Rath, Theodore Alexandrov	bioRxiv	Publication
One model to use them all: Training a segmentation model with complementary datasets	Alexander C. Jenke, Sebastian Bodenstedt, Fiona R. Kolbinger, Marius Distler, Jürgen Weitz, Stefanie Speidel	IPCAI24	Publication
A Last-Level Defense for Application Integrity and Confidentiality	Gabriel P. Fernandez, Andrey Brito, Ardhi Putra Pratama Hartono, Muhammad Usama Sardar, Christof Fetzer	Middleware '23: Proceedings of the 24th International Middleware Conference	Publication
Novel Approaches Toward Scalable Composable Workflows in Hyper-Heterogeneous Computing Environments	Kica,P., Lichołai S., Malawski M	Supercomputing23 – SC23	Publication
Trustworthy confidential virtual machines for the masses	Anna Galanou, Khushboo Bindlish, Luca Preibsch, Yvonne-Anne Pignolet, Christof Fetzer, Rüdiger Kapitza	Middleware '23: Proceedings of the 24th International Middleware Conference	Publication
Scaling a Variant Calling Genomics Pipeline with FaaS	Aitor Arjona, Arnau Gabriel-Atienza, Sara Lanuza-Orna, Xavier Roca-Canals, Ayman Bourramouss, Tyler K. Chafin, Lucio Marcello, Paolo Ribeca, Pedro García-López	WoSC '23: Proceedings of the 9th International Workshop on Serverless Computing	Publication

References

[1] StreamNative, "Kafka on streamnative: Bringing enterprise-grade kafka support to streamnative pulsar clusters." https://streamnative.io/blog/kafka-on-streamnative-bringing-enterprise-grade-kafka-support-to-streamnative-pulsar-clusters, 2023.

[2] "Cncf pravega - kafka adapter." https://github.com/pravega/kafka-adapter, 2024.

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

[4] G. Dall'Alba, P. L. Casa, F. P. d. Abreu, D. L. Notari, and S. de Avila e Silva, "A survey of biological data in a big data perspective," Big Data, vol. 10, no. 4, pp. 279–297, 2022. PMID: 35394342.