EUROPEAN BIG DATA VALUE FORUM

DataNexus:

Extreme Data Architectures Empowering AI Innovation Across the Compute Continuum

Moderated by: Nuria de Lama (IDC) and Janine Gehrig (BSC)

DATA NEXUS

EXA4MIND

EXTRACT
A distributed data-mining software platform for extreme data across the compute continuum

GRAPH MASSIVIZER

NEARDATA

ORGANISED BY

BDV BIG DATA VALUE ASSOCIATION

IN COLLABORATION WITH

DTU

SDU University of Southern Denmark

ebdvf.eu
#EBDVF25

Session overview

Intro

- Importance of extreme data
- What is DataNexus?
- Presentation of speakers

→

Use Case showcase: Problem, solution, impact

→

Panel and Q&A

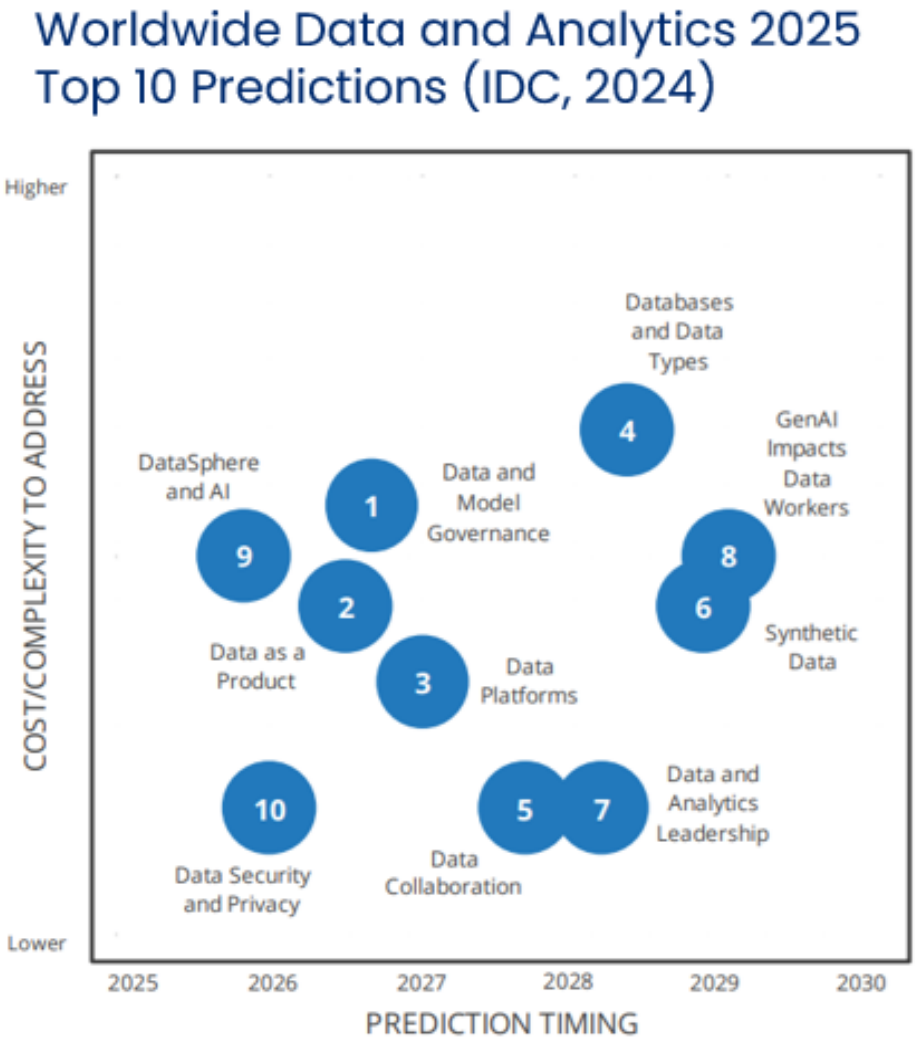
Session goal: Moving from Research to Impact

Showcase results: how 4 DataNexus projects are advancing beyond the state of the art

Promote update and use: highlight open platforms, tools and resources available for reuse

Foster collaboration: connect research outcomes to Europe’s broader data and AI ecosystem

Why is extreme data important?



- If we define "extreme data" as unstructured, streaming, or real-time data (the types that present the greatest challenges and opportunities for modern digital and AI-driven organizations), **then at least 80% of all data generated globally can be considered "extreme data".**
- The share is even higher in certain enterprise and AI contexts, where **unstructured and streaming data are the primary sources for analytics, automation, and AI model training**
  - **Unstructured Data Dominates**: By 2029, unstructured data will account for approximately 81.7% of all data generated globally. In 2024, this figure was 92%.
  - **Streaming and Real-Time Data**: In 2024, 67% of all data created globally was streaming data.
  - **AI-Related Data**: For AI workloads, organizations report that 44% of their AI-related data is unstructured, 28% is structured, and 17% is semi-structured.
  - **Raw/Unorganized Data**: Two-thirds of organizations indicate that the data used in workflows is in raw or unstructured forms, which can delay effective analysis and requires advanced data management.

Why is extreme data important?

Impact

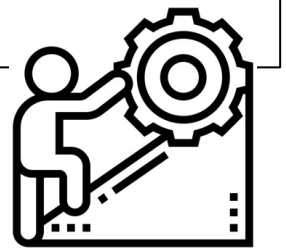
- AI and automation
- IoT and edge computing
- Business Insights and Monetization
- Regulatory and Security demands

Market Projections

- Volume of data created each year is forecast to increase at a CAGR of 24.9% from 2024 to 2029 (faster unstructured data)
- Data Integration and Intelligence SW: Worldwide revenue is projected to nearly double from $6.4B in 2024 to $12.2B in 2029 (11,8% for EMEA)
- AI Life-Cycle Software: CAGRs above 27% 2024-2029 (EMEA from $3B to $11B)

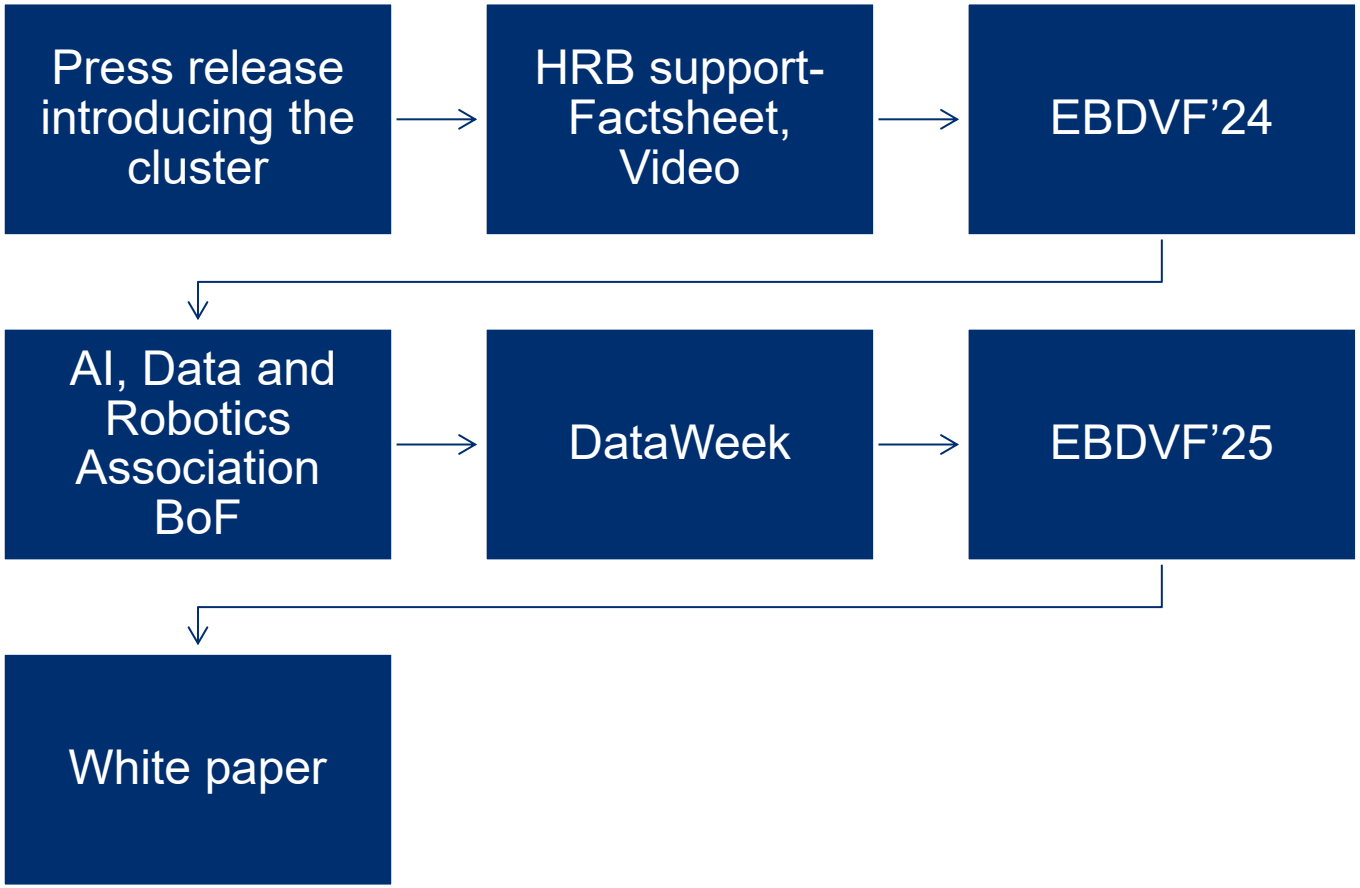
Main challenges

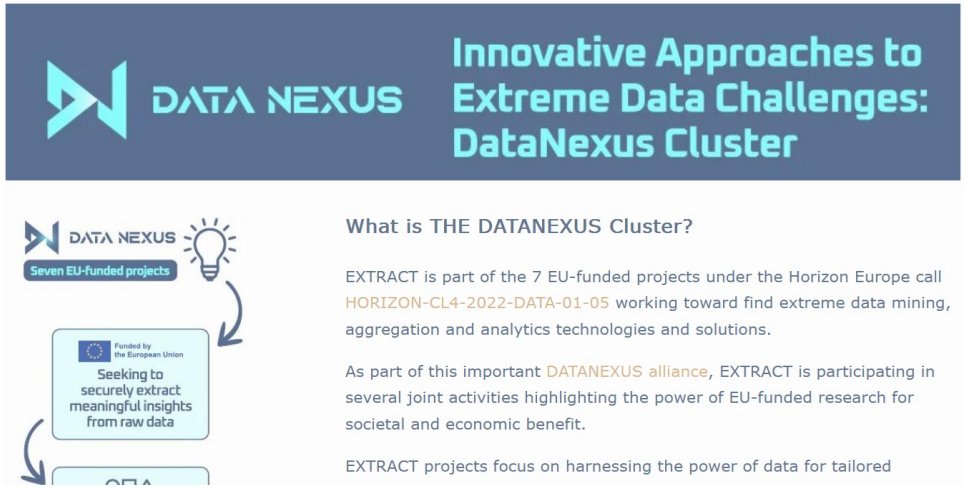
- Complexity and Integration
- Rising costs
- Regulatory and security concerns
- Talent and ecosystem gaps
- ROI and value extraction



The DataNexus Cluster

7 projects EFRA, EMERALDS, **EXA4MIND, EXTRACT, GRAPH-MASSIVIZER, NEARDATA &** SYCLOPS are funded under Horizon Europe call **HORIZON-CL4-2022-DATA-01-05**





Speakers

EXA4MIND

Stephan Hachinger,
Science and Co-Design Coordinator,
Leibniz Supercomputing Centre
(BADW-LRZ)

GRAPH-MASSIVIZER

Junaid Ahmed Khan,
PhD Candidate and Research Fellow,
University of Bologna

EXTRACT

Eduardo Quiñones,
Group Leader,
Barcelona Supercomputing Center

NEARDATA

Pedro García López,
Senior Researcher,
Universitat Rovira i Virgili

EUROPEAN BIG DATA VALUE FORUM

Use case showcase: DataNexus in Action

EXA4MIND

EXTRACT
A distributed data-mining software platform for extreme data across the compute continuum

GRAPH MASSIVIZER

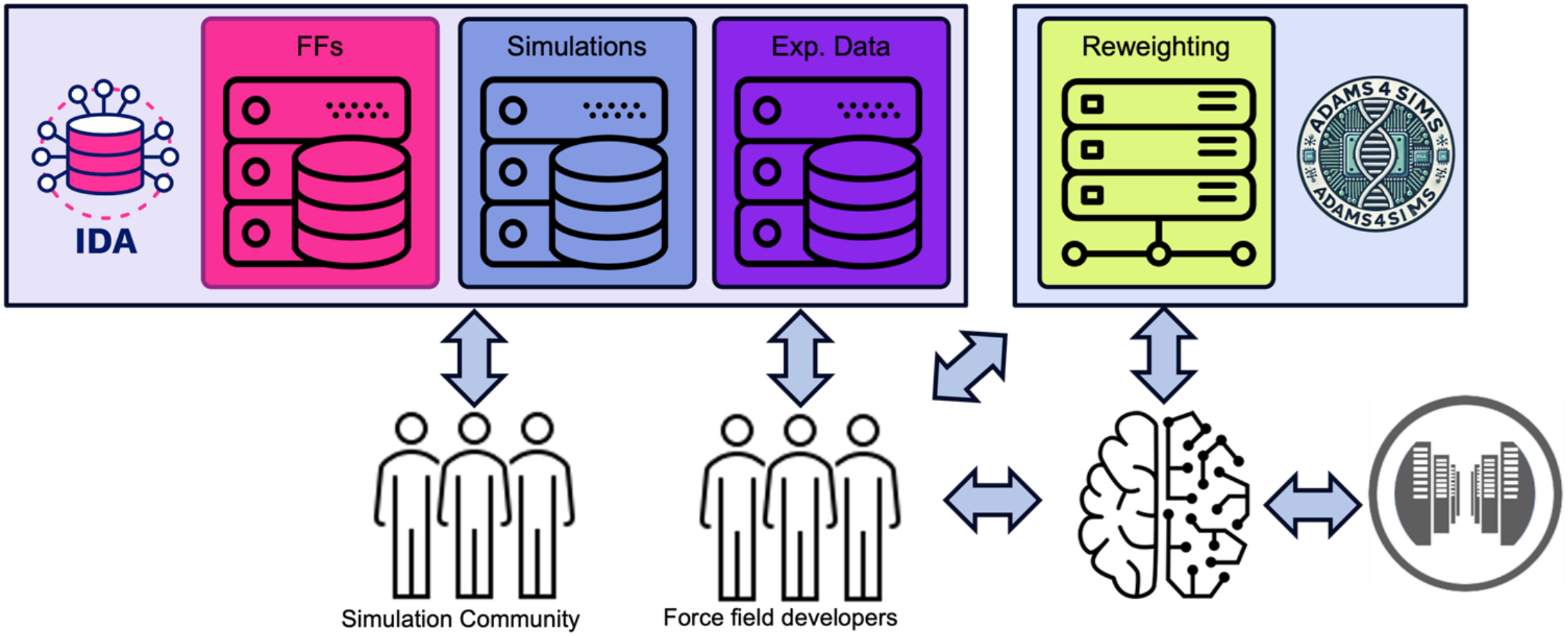
NEARDATA

Extreme Analytics for Mining Data spaces

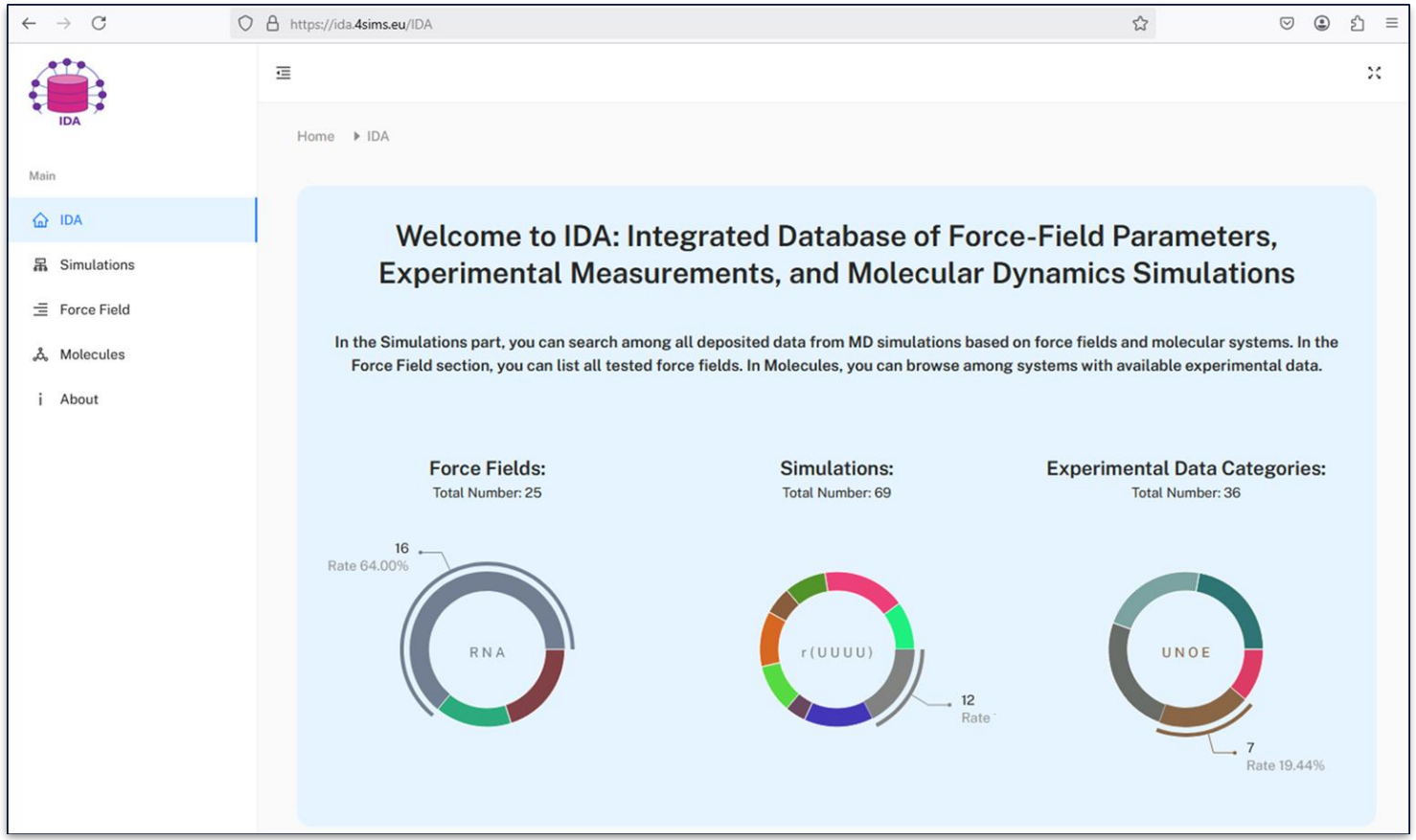
Stephan Hachinger
Science and Co-Design Coordinator
Leibniz Supercomputing Centre (BADW-LRZ)

Dr. Stephan Hachinger is Science and Co-Design Coordinator of EXA4MIND and leads the Research Data Management Team at Leibniz Supercomputing Centre (LRZ, Garching/D). He focuses on data-driven workflows for users to make the most of computing infrastructure, and on FAIR and Open Data. In EXA4MIND, he helps to enable Extreme Data analytics via supercomputing, optimised data backends and EU data ecosystems.

EXA4MIND: Science Application Case (Molecular Dynamics)

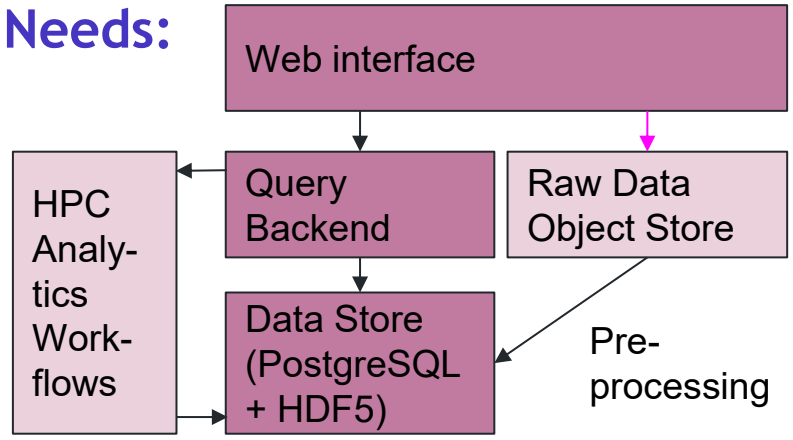


EXA4MIND: Science Application Case (Molecular Dynamics)

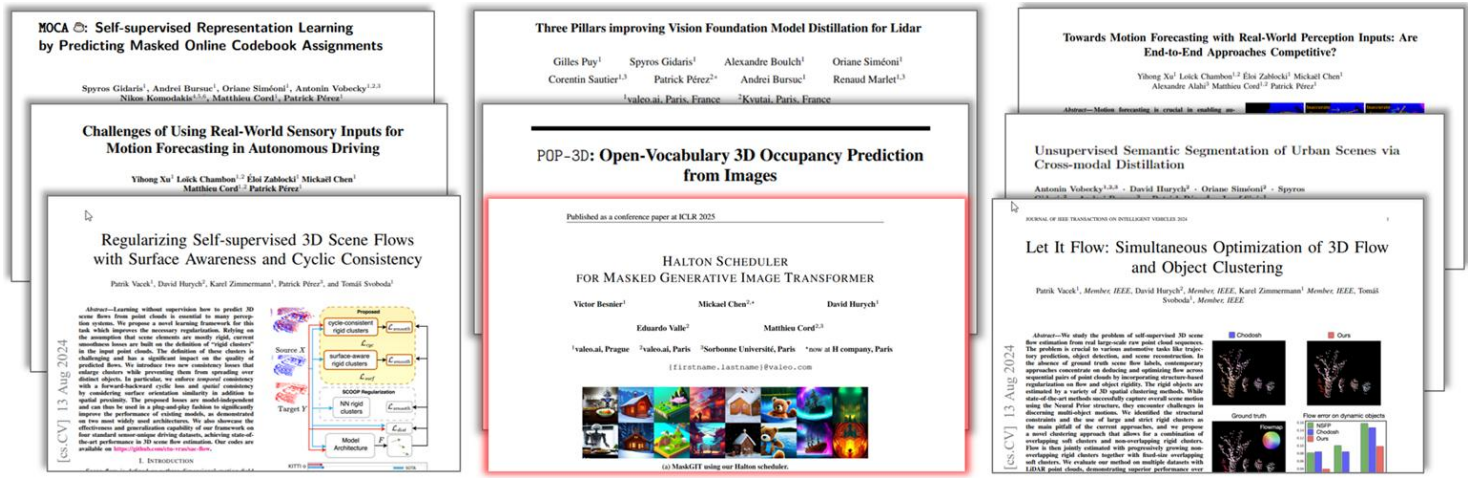


Automating analysis of Experimental vs. Simulated data on the way to optimisation of force fields.

Needs:



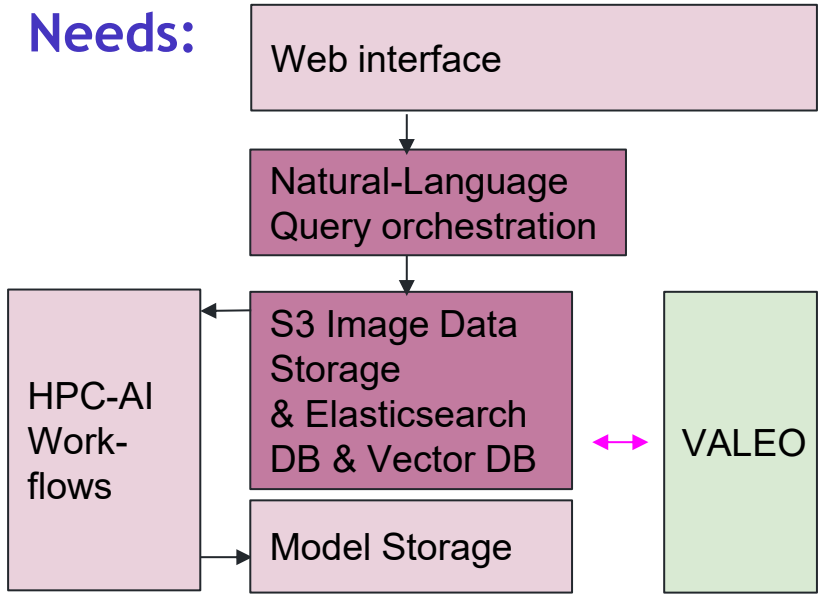
EXA4MIND: Autonomous Driving Application case



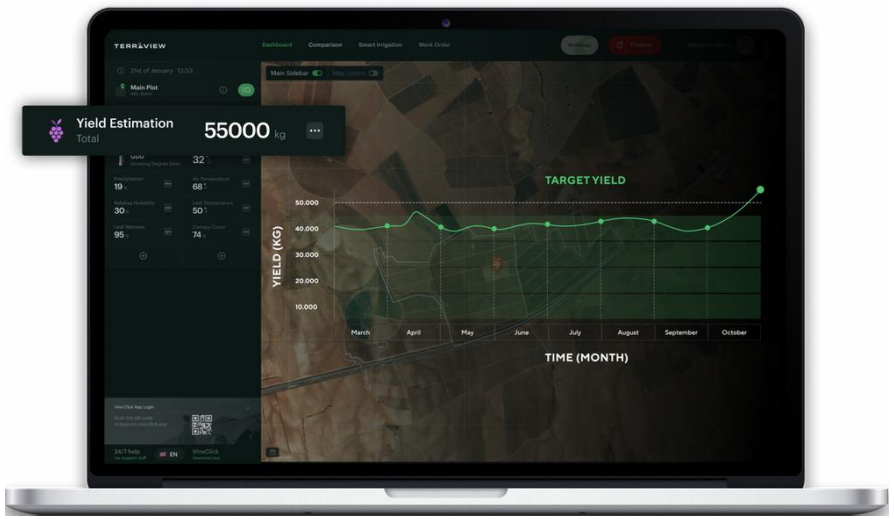
- AI model development enabled by staging & AI workflows on IT4Innovations supercomputing
- Interface for manual analysis of driving situations using combined queries on text and images

Developing AI models for Advanced Driving Assistance Systems: Novel image-processing algorithms + AI-aided failure analysis

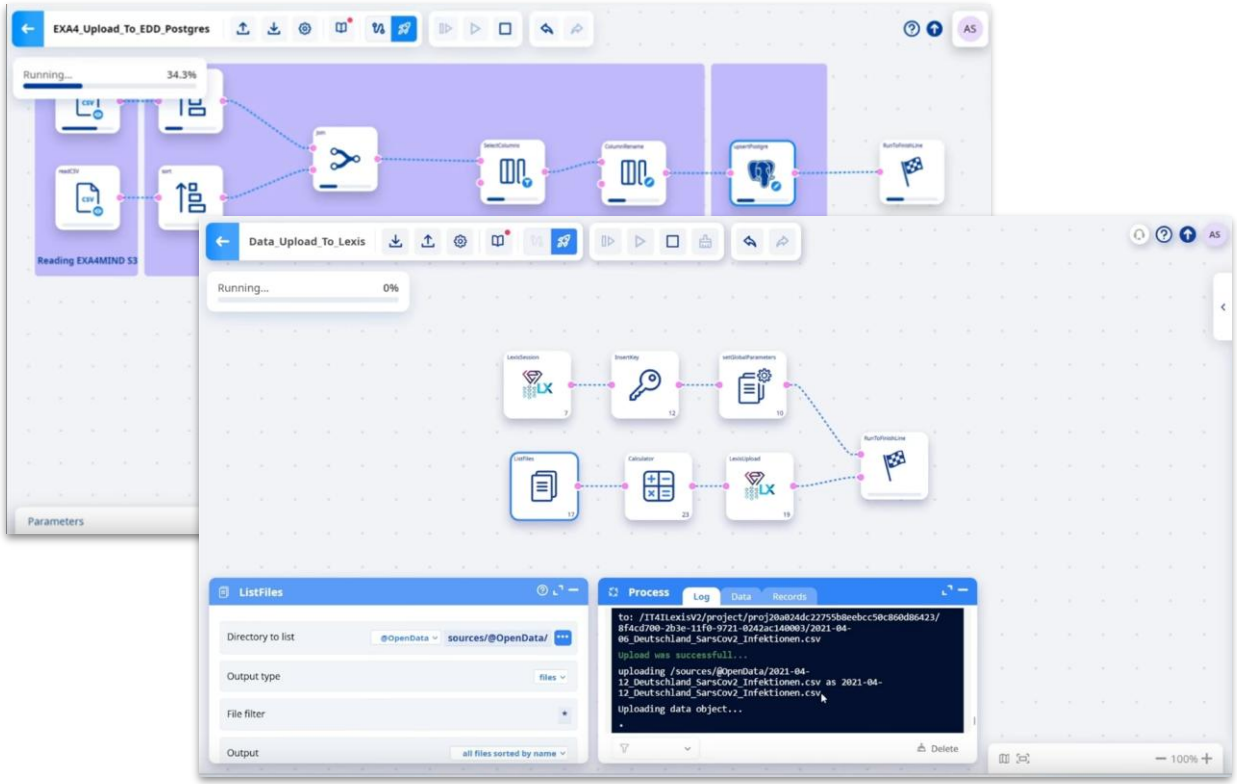
Needs:



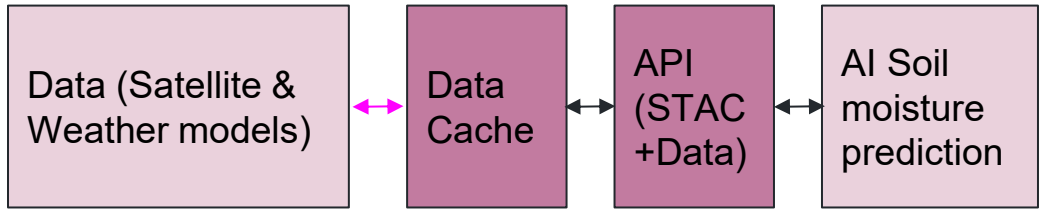
EXA4MIND: SME Application Cases: Smart Viticulture and Health Data Mining



- Satellite data cache & retrieval logic for soil moisture estimation



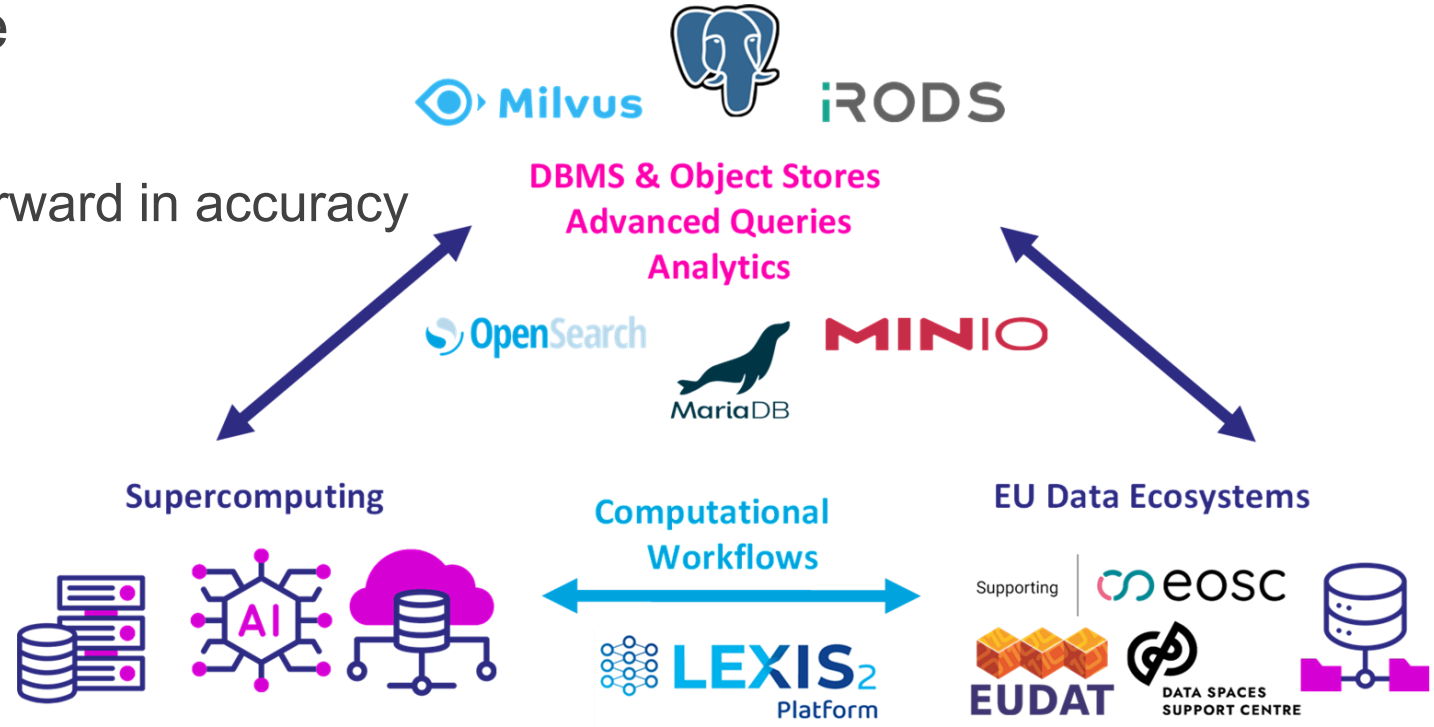
Needs:



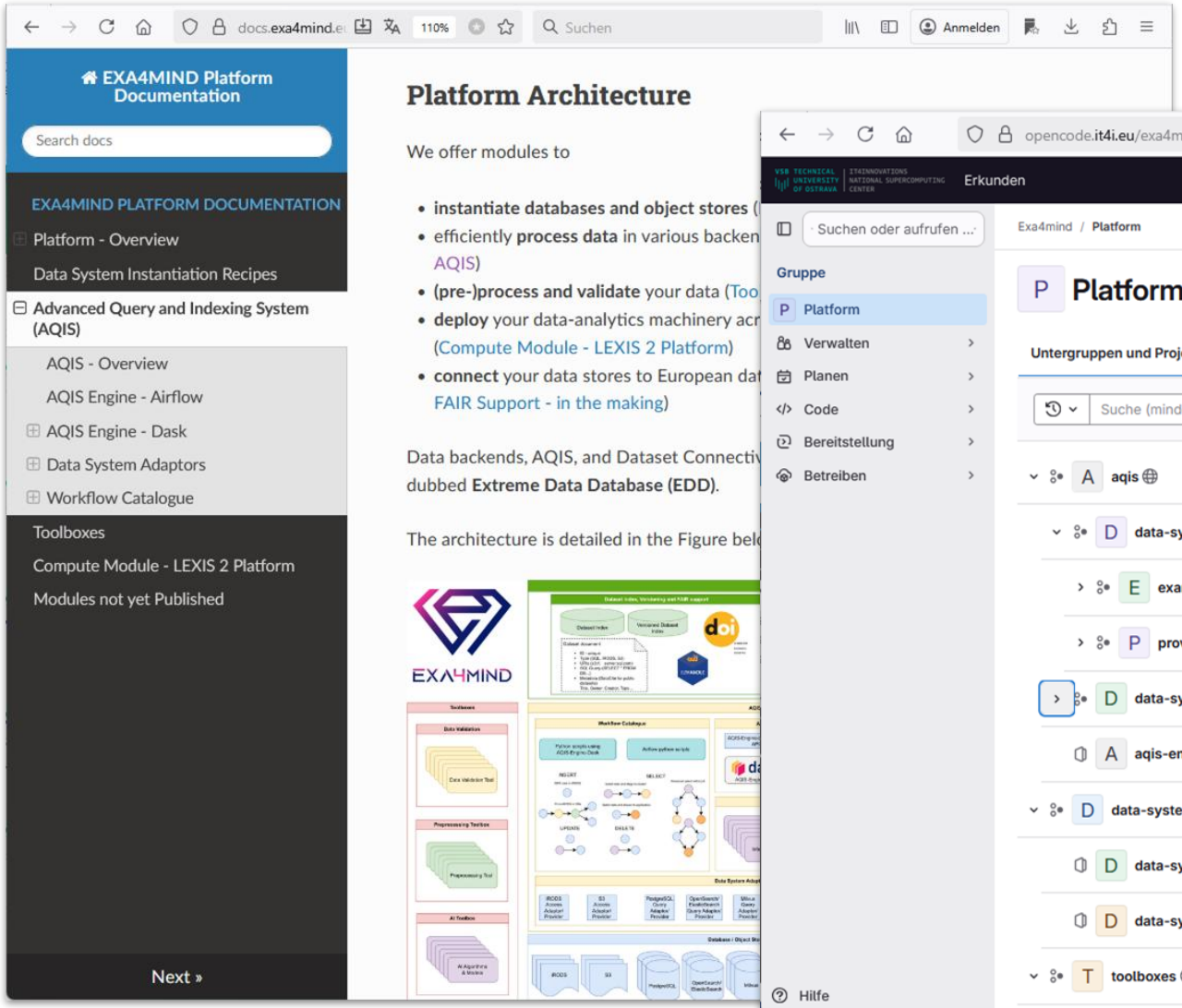
- Integration of ALTRNATIV visual ETL/Data-Mining Tool with
  - HPC/AI-System Offloading - LEXIS Distributed HPC/Cloud-Computing Platform
  - Optimum Data backends

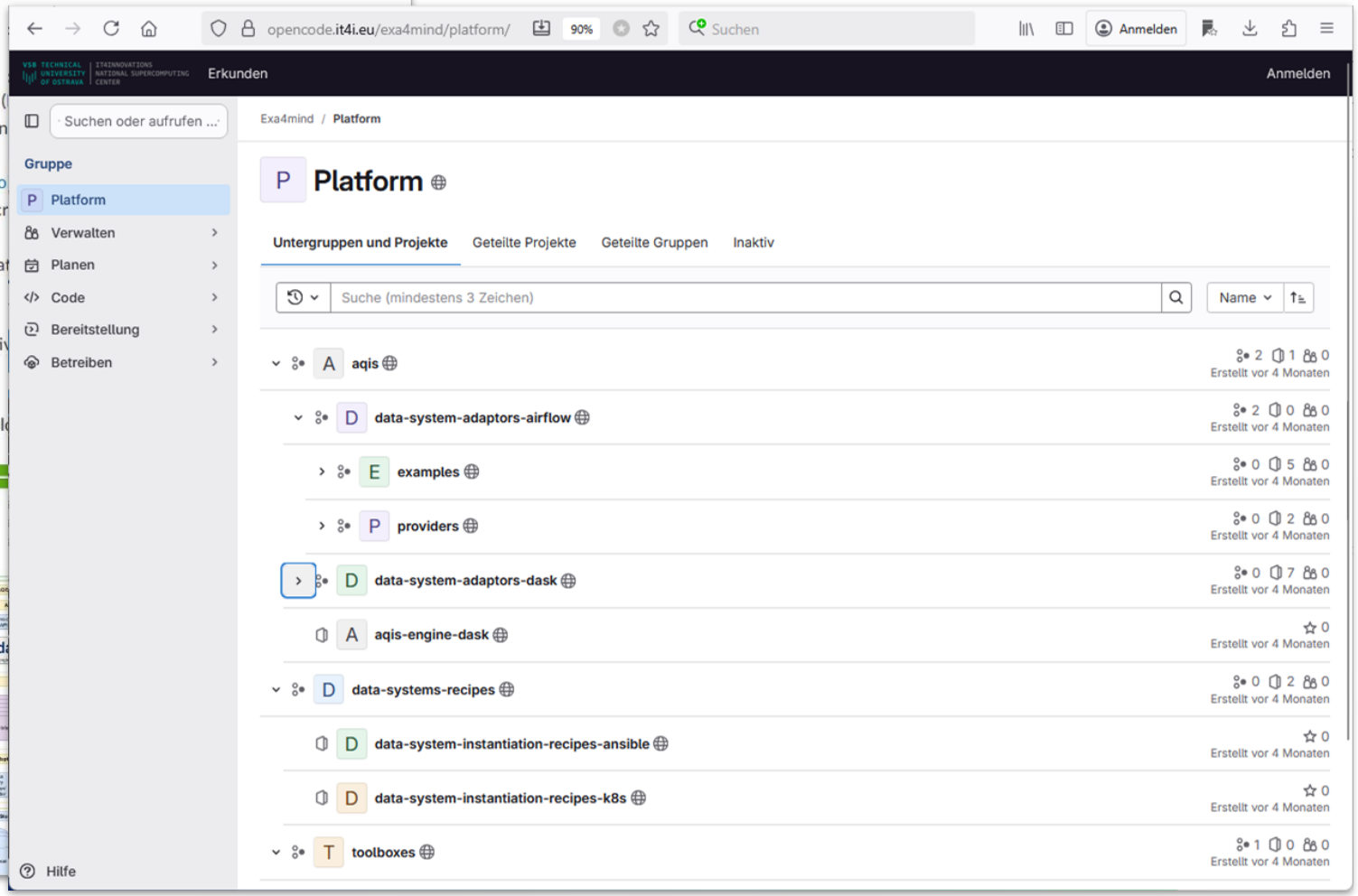
A glimpse on the impact

- EXA4MIND modules as enabler for Extreme Data Analytics: they bridge Supercomputing - Data backends - Data ecosystems
- Best practices for compute & Extreme Data backend usage
- Application cases impacts:
  - Molecular simulations: leaps forward in accuracy
  - Development of next-level driving assistance systems
  - Flexible leveraging of satellite data for smart viticulture
  - Health-related Extreme Data analytics via convenient GUI

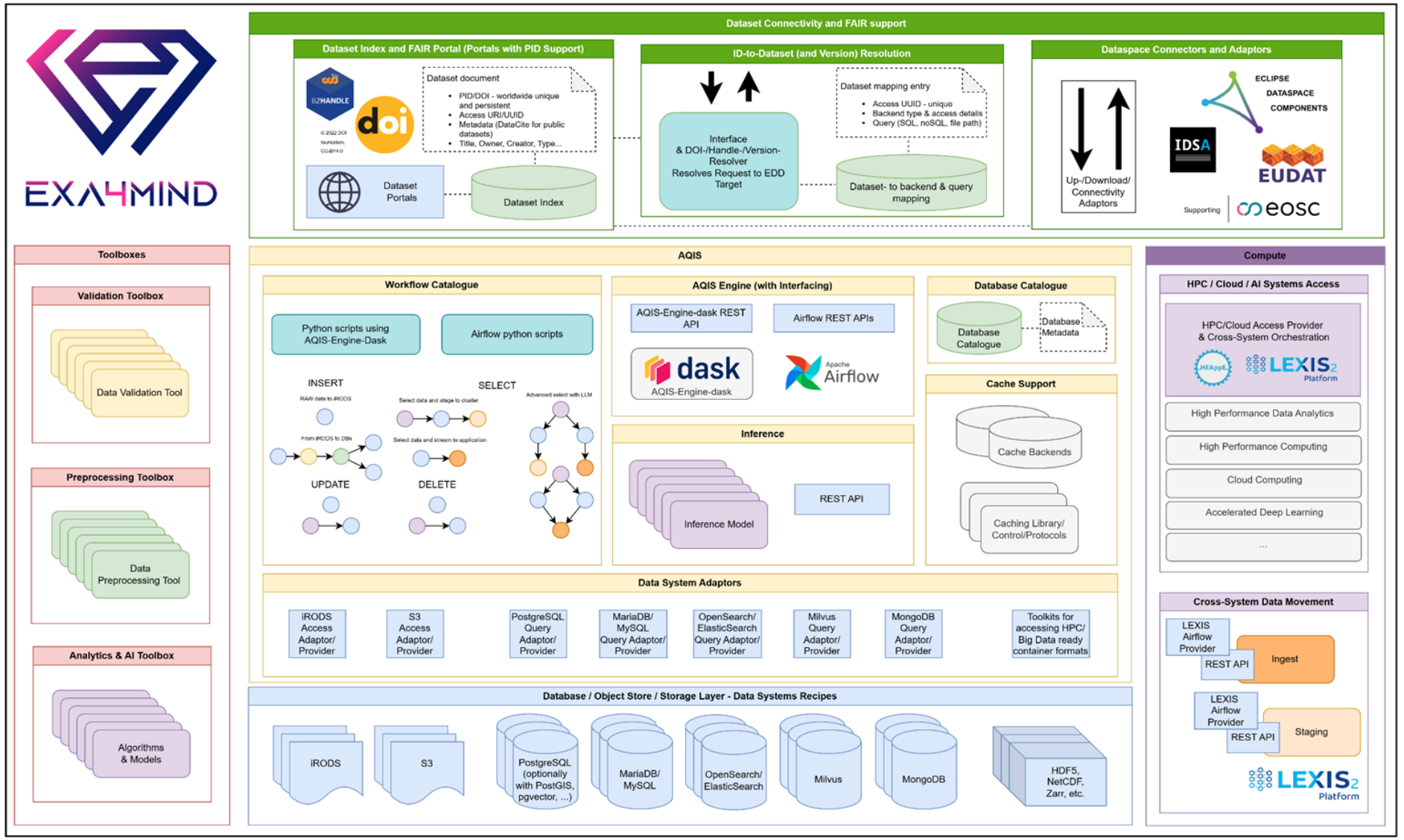


Open Source Modules & Documentation





Full Architectural Overview



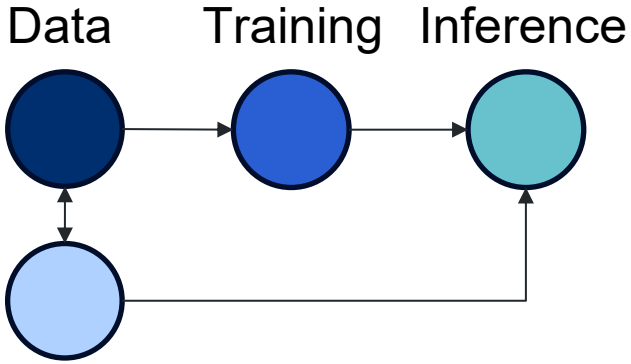
A distributed data-mining software platform for extreme data across the compute continuum

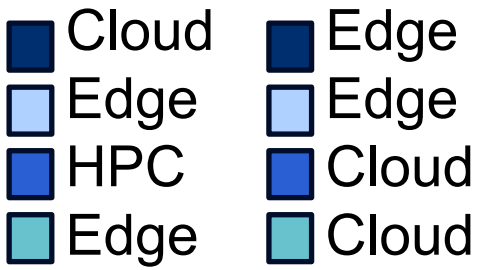
Eduardo Quiñones
Group Leader
Barcelona Supercomputing Center

Eduardo is coordinator of the EXTRACT project. He specializes in integrating cutting-edge technologies like AI, data mining, advanced computing, into real-world applications. He has a strong background in edge, cloud and HPC systems, and his work focuses on leveraging these technologies to enhance decision-making in urban environments, including in crisis management scenarios.

Workflow Description and Execution

- Describe workflows at different granularity levels independently of the infrastructure
- Guarantee efficient data and analytics processing across the continuum





Developments applied to 2 use-cases:

- **Personalized Evacuation Route (PER)**
- **Transient Astrophysics with the Square Kilometre Array (TASKA)**

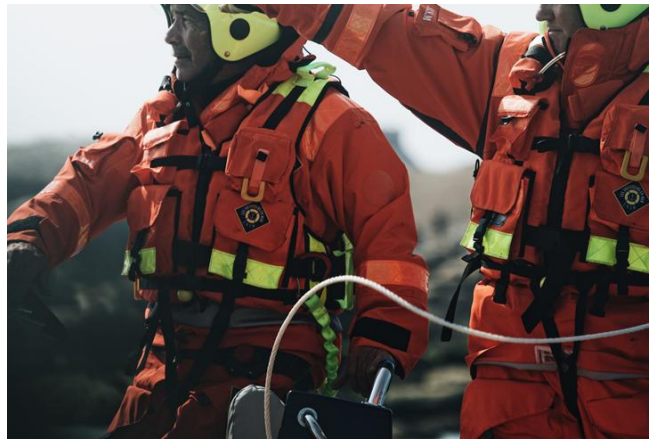
PER Use Case: Understanding the Crisis Response Gap

DISASTER

CONFUSION

RESPONSE GAP

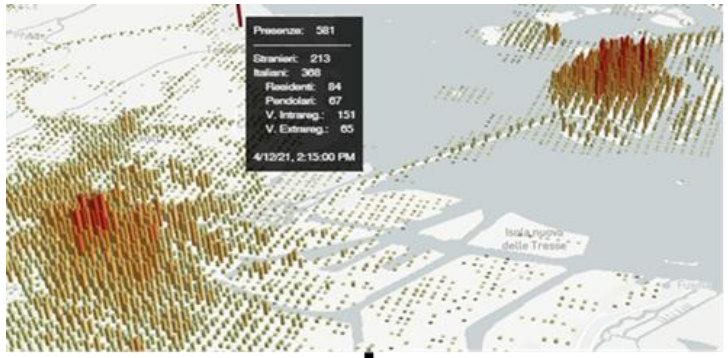
RESPONSE



Guide citizens in a complex urban mesh (Venice) through a safe route

Personalized Evacuation Route (PER) System

PER: Closing the crisis response gap

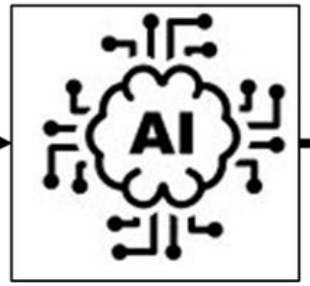


Real-time Data Gathering

REALITY

Specific Actions

Urban Digital Twin
Models the city gathered data

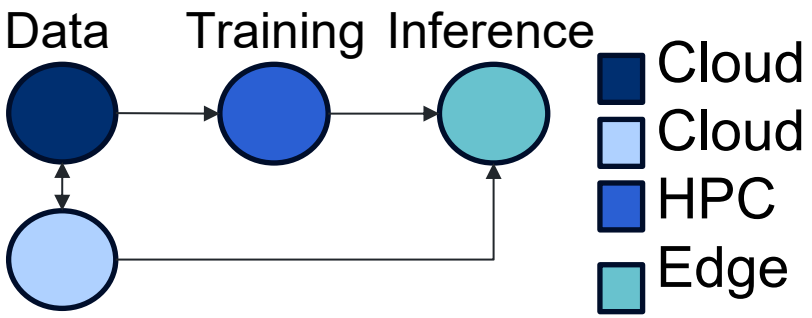


Artificial Intelligence

1. Learns to generate routes based on simulated emergency scenarios
2. Generates routes to citizens when emergency occurs

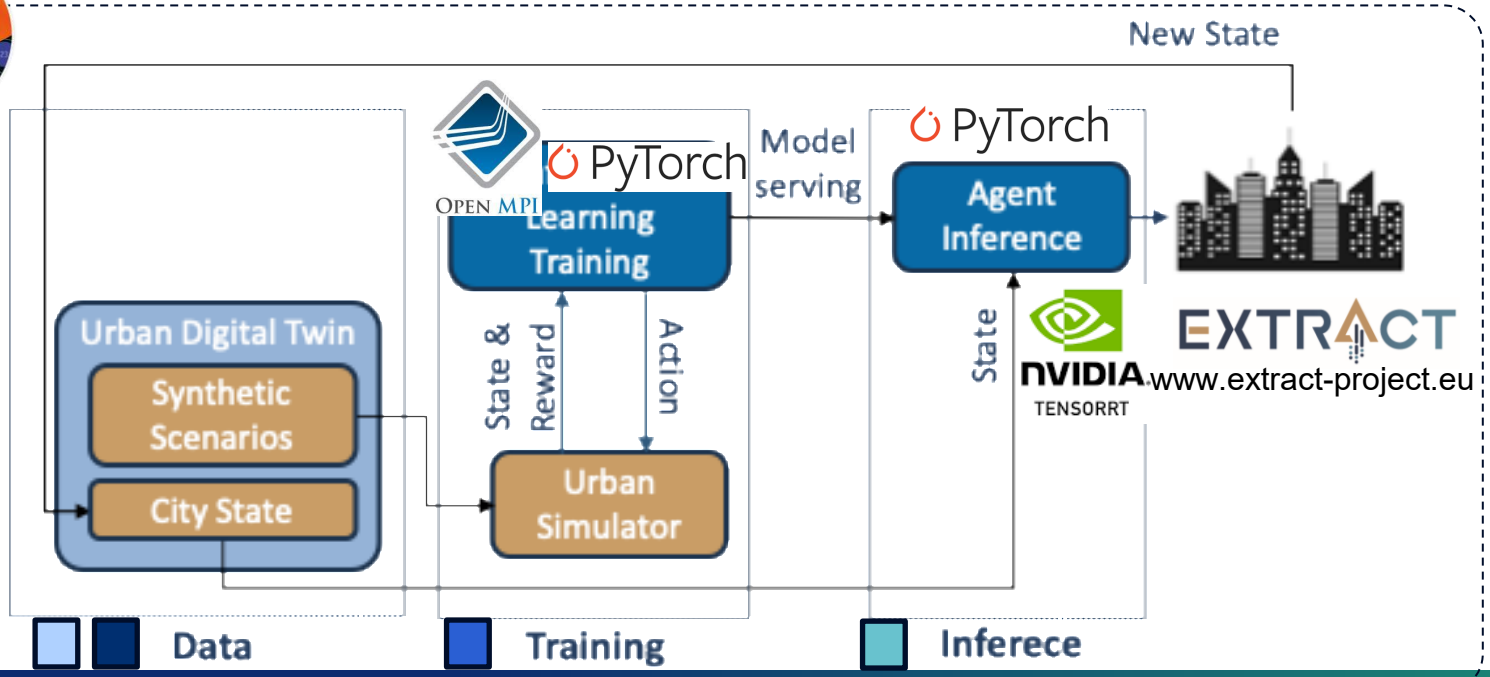
SIMULATED SCENARIOS

PER Workflow Description

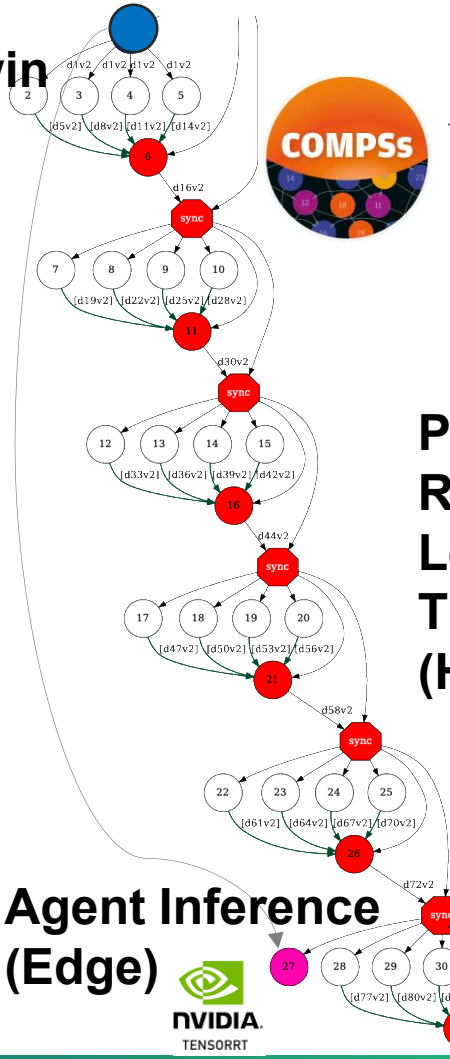


Cloud
Cloud
HPC
Edge

COMPSs



Urban Digital Twin (Cloud)



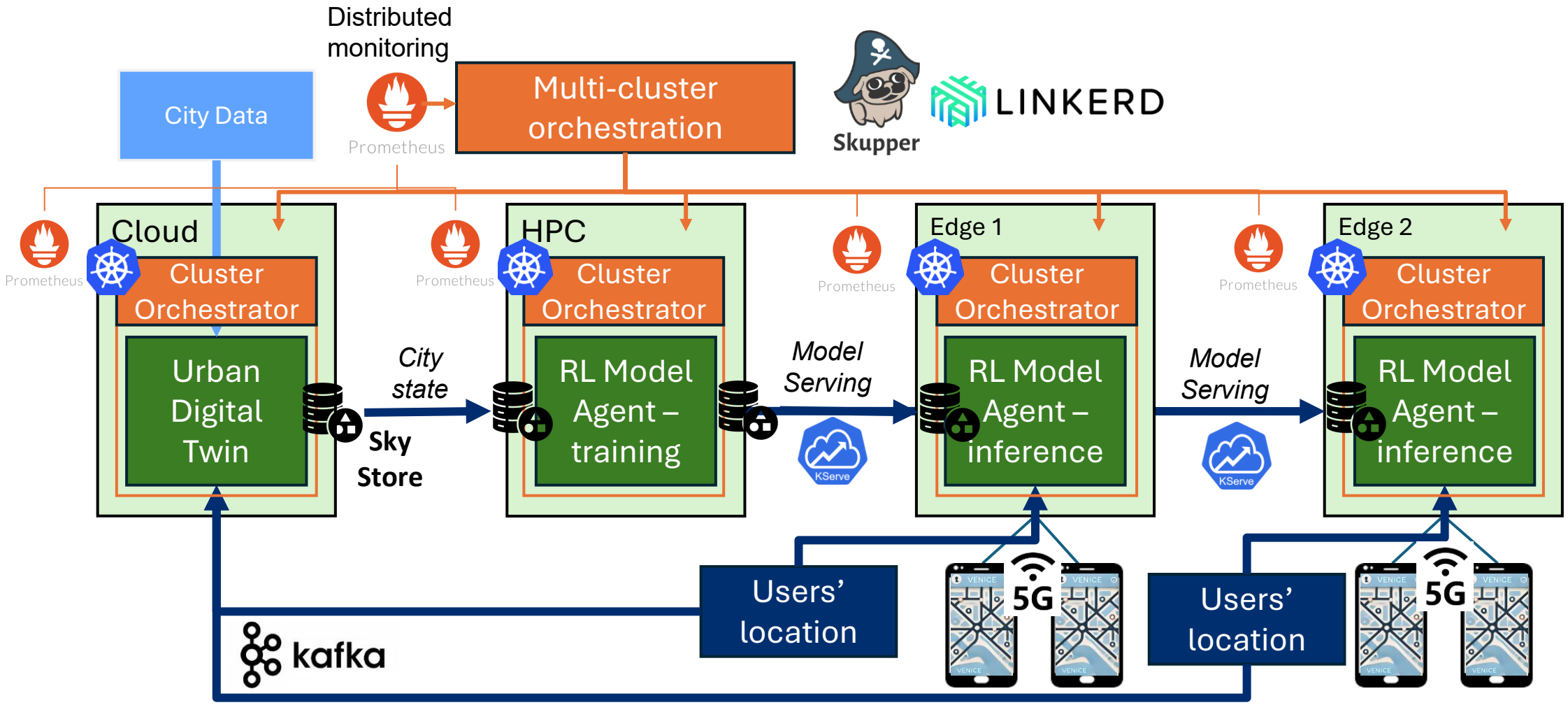
Parallel Reinforcement Learning Training (HPC PyTorch)

OPEN MPI

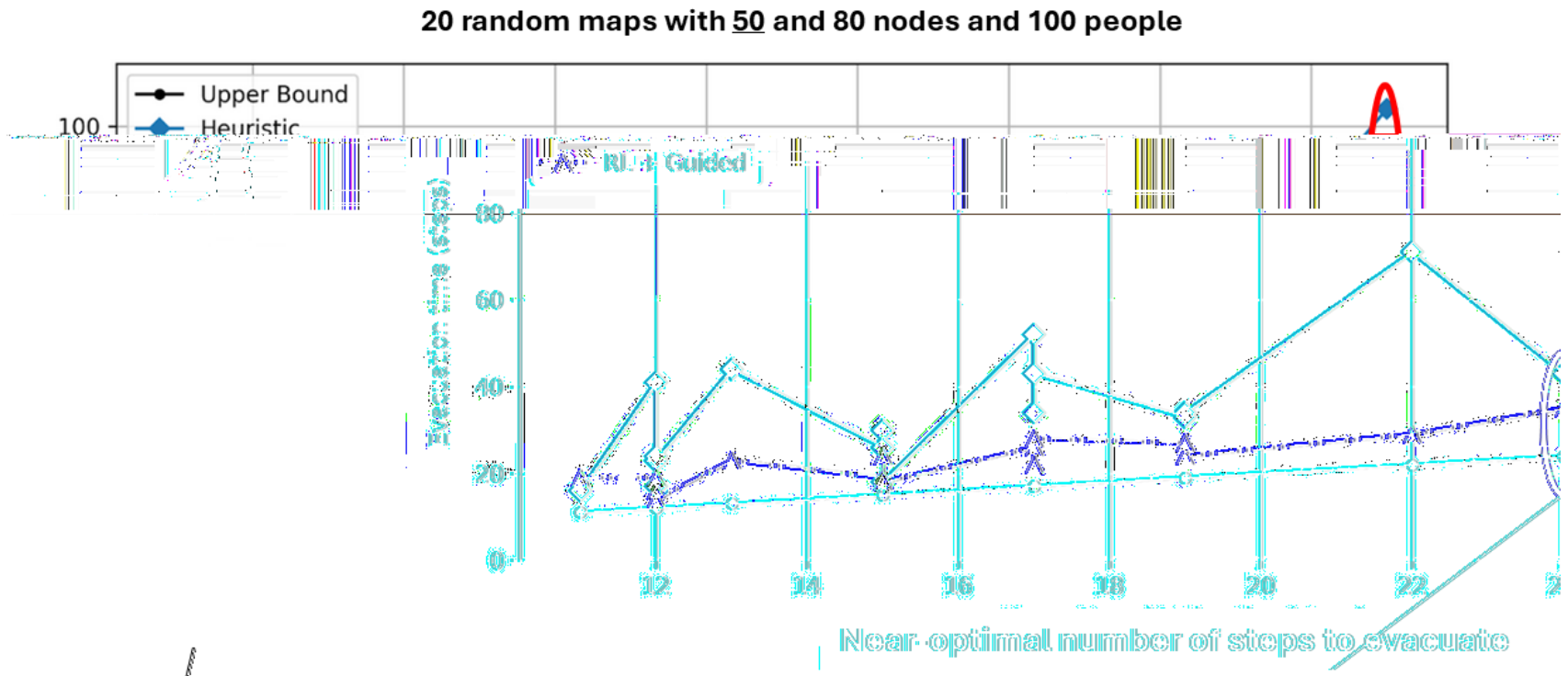
Agent Inference (Edge)

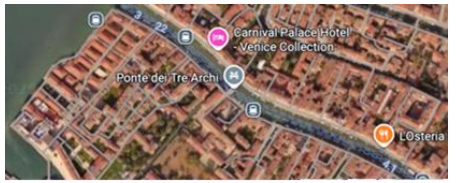
NVIDIA TENSORRT

PER Workflow Deployment and Orchestration



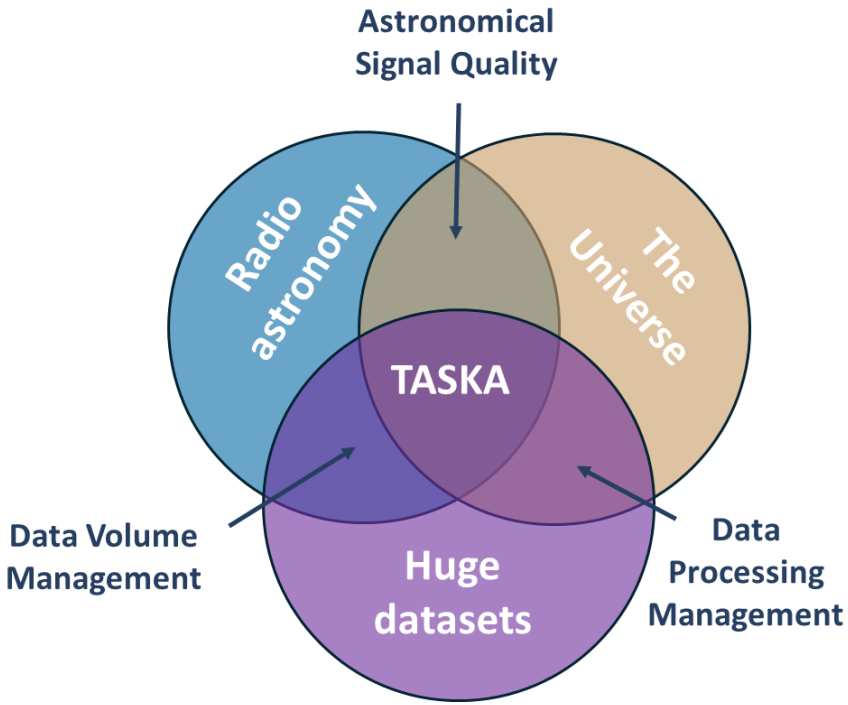
Impact-Closing the Response Gap



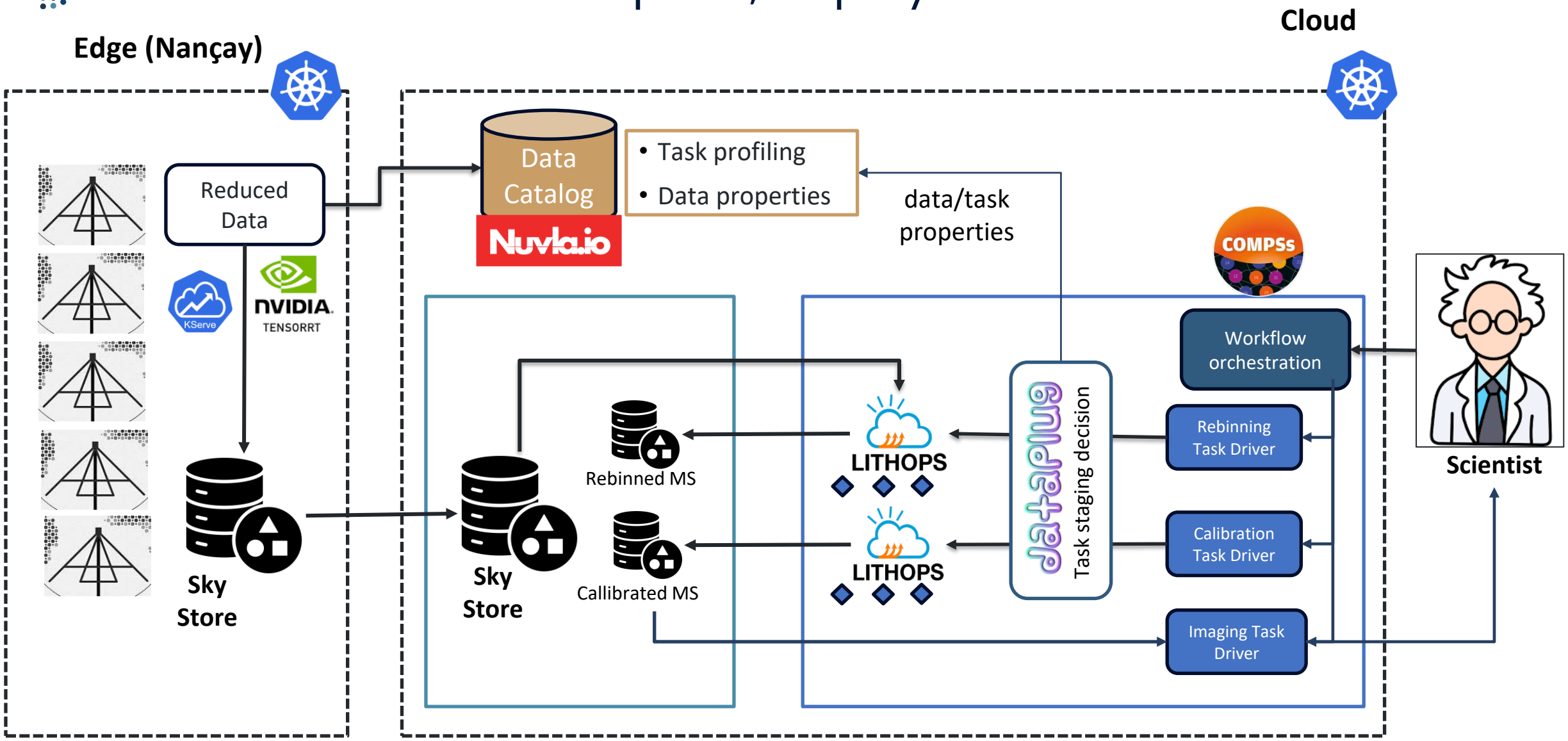


Use Case 2: Transient Astrophysics with the Square Kilometre Array Pathfinder (TASKA)

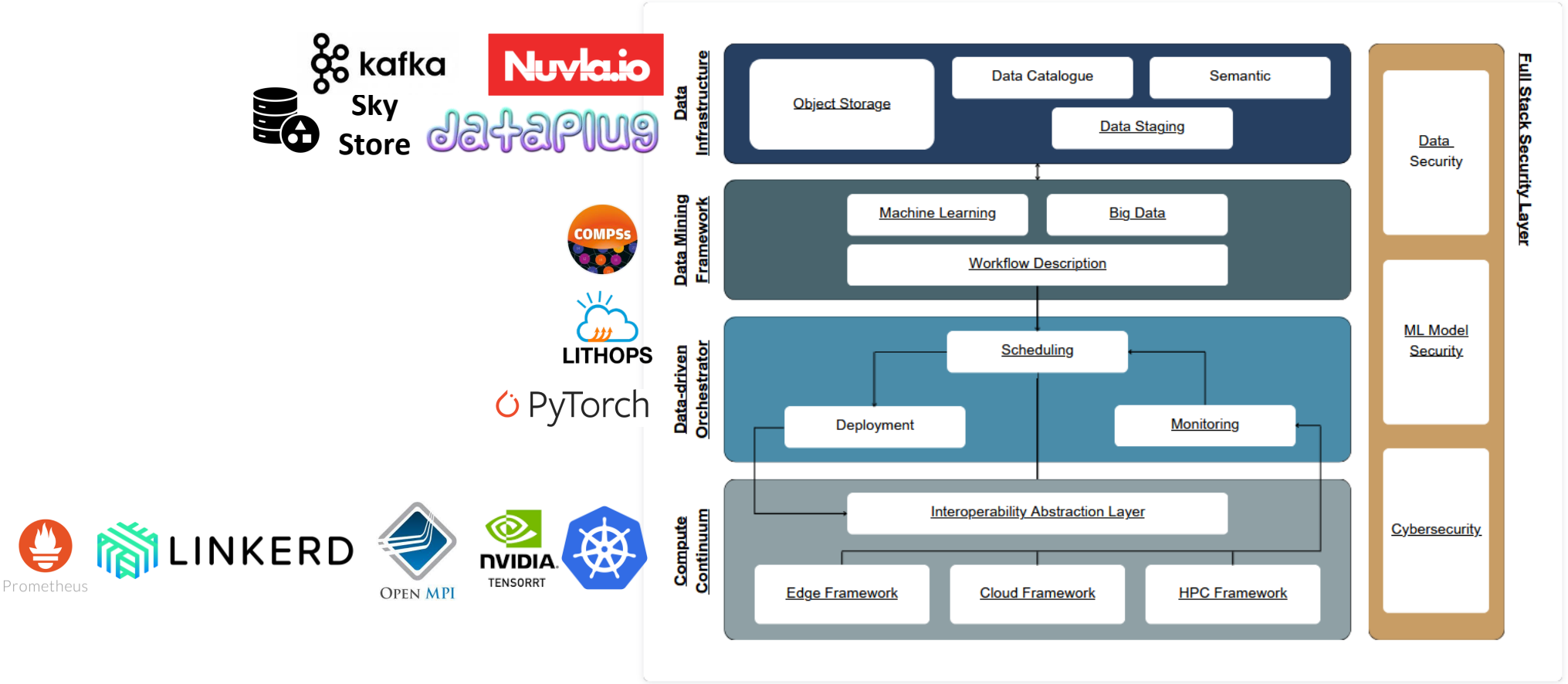
NenuFar radio telescopes in Nançay, France



TASKA: Workflow Description, Deployment and Execution



EXTRACT Software Architecture



Extreme and sustainable graph processing for urgent societal challenges in Europe

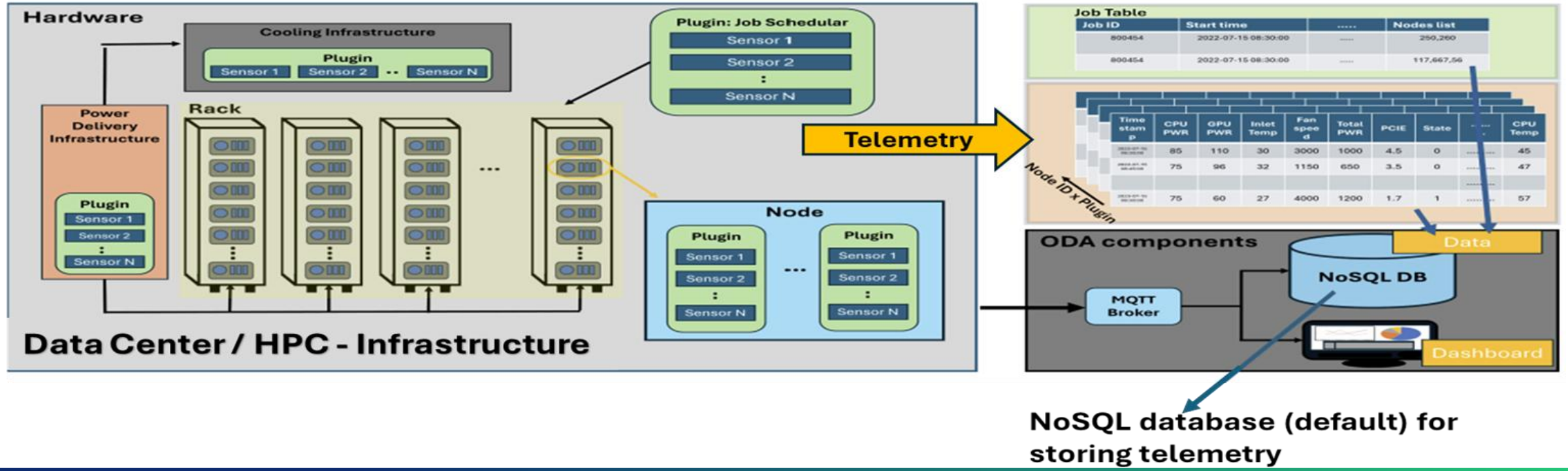
Junaid Ahmed Khan
PhD Candidate and Research Fellow
University of Bologna

Junaid is a final year PhD student is data science and computation at the university of Bologna. His main research focus has been on applying graph methodologies for operational data analytics in high-performance computing (HPC) system. He has been involved with the Graph-Massivizer project as one of the main developers for the data center use case.

Data Centre Digital Twin for Sustainable Computing

GRAPH MASSIVIZER

- HPC systems are the cornerstone of technological and scientific advancements
- HPC installations are massive (millions of sensors) => large-scale telemetry => difficult to manage.
- Operational data analytics (ODA) proposed, but it struggles to do more than fancy dashboarding.

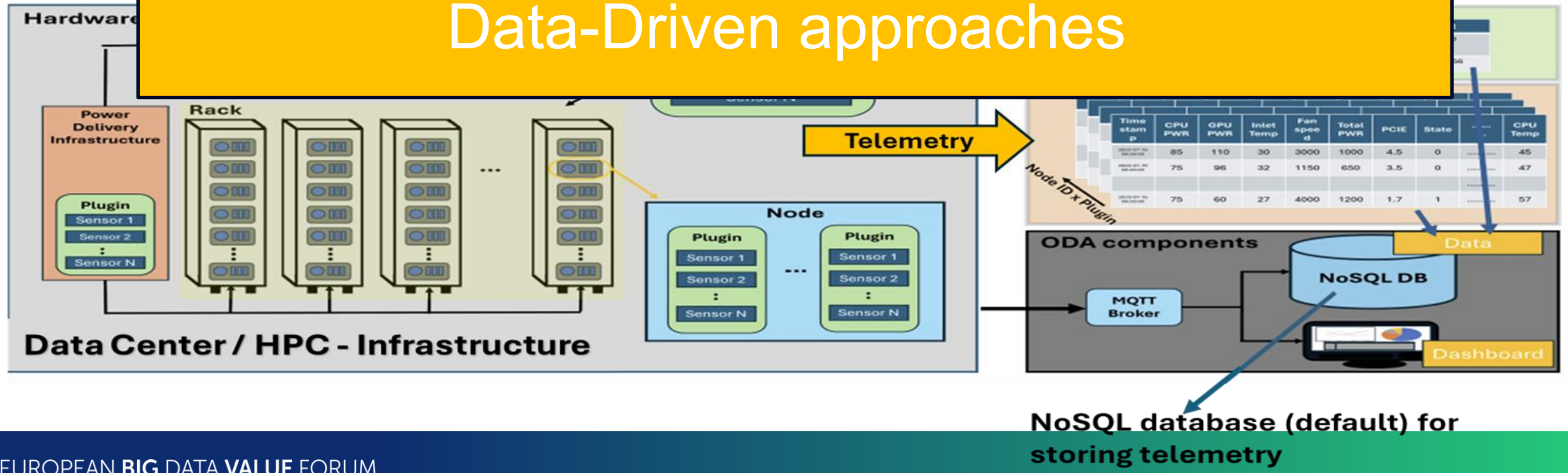


Data Centre Digital Twin for Sustainable Computing

GRAPH MASSIVIZER

- HPC systems are the cornerstone of technological and scientific advancements
- HPC installations are massive (millions of sensors) => large-scale telemetry => difficult to manage.
- Oper
  dash

AI for HPC sustainability - Leverage AI and Data-Driven approaches



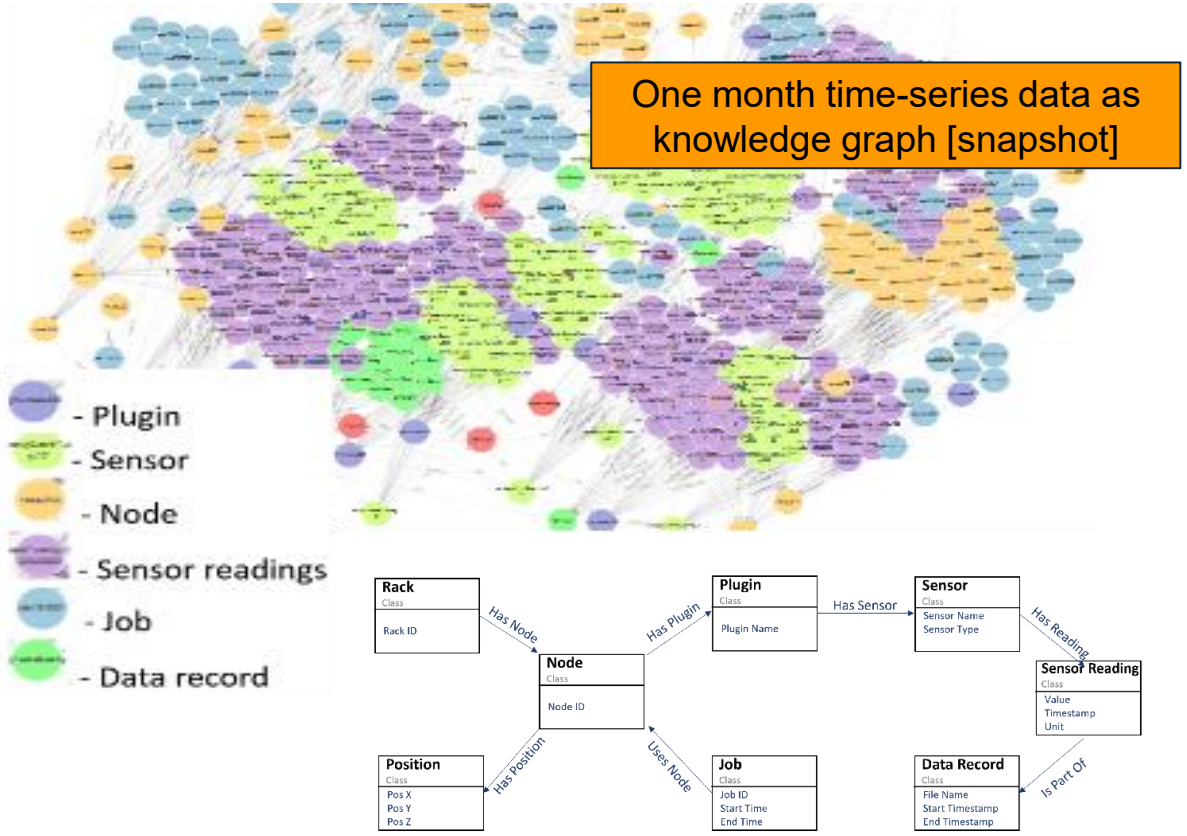
Data Centre Digital Twin for Sustainable Computing

- Focused on two operations for data center operational sustainability:
  - Prevention of system downtime:
    - Anomaly detection and prediction.
    - Improved system utilization => improved power efficiency.
  - Reduction in management complexity:
    - Neuro-symbolic data analysis assistants: Interaction with the data in natural language.
    - Reduced entry barrier to complex data analysis for post-mortem analysis and what-if evaluations.

Data Centre Digital Twin for Sustainable Computing

- ETL monitoring data ingestion pipeline:
  - Mapping operational data to an RDF Knowledge graph
  - RDF database: GraphDB
  - One month of telemetry, approx:
    - 11.6 billion nodes
    - 68.6 billion edges
- Dataset: ExaData [1]

Massive graph

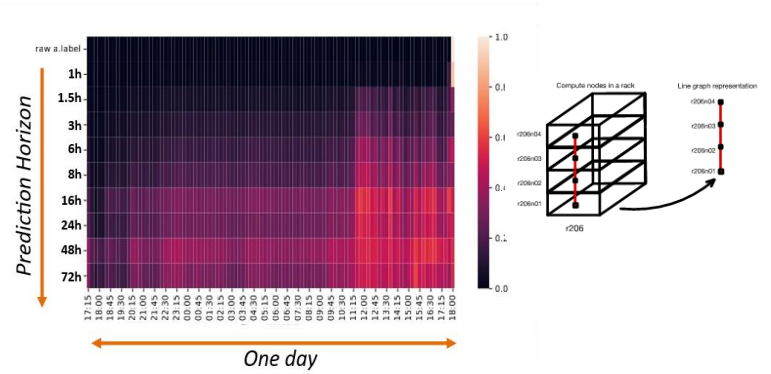


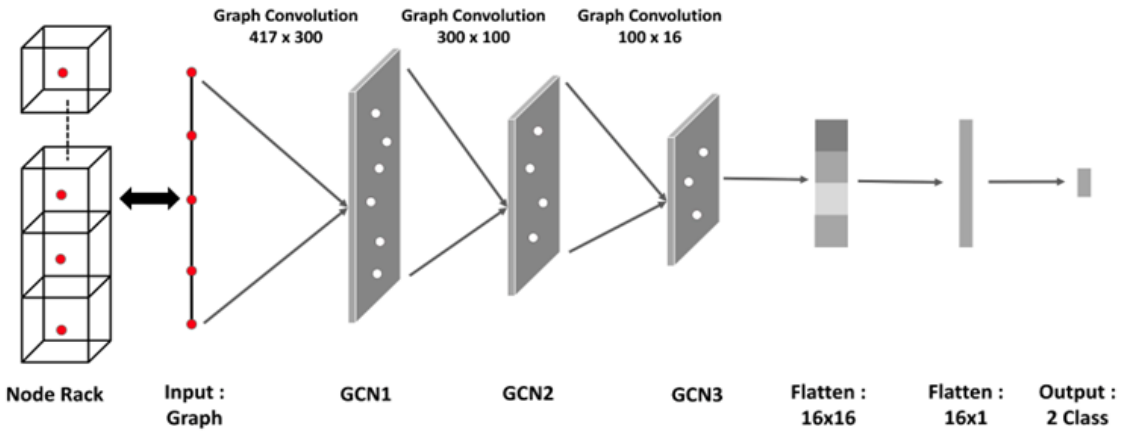
[1] Borghesi, A., Di Santi, C., Molan, M. *et al.* M100 ExaData: a data collection campaign on the CINECA's Marconi100 Tier-0 supercomputer. *Sci Data* 10, 288 (2023). https://doi.org/10.1038/s41597-023-02174-3

Data Centre Digital Twin for Sustainable Computing

GRAPH MASSIVIZER

- Predictive maintenance: Developed Graph Neural Network (GNN) based models called “GRAAFE” for anomaly prediction (much more impactful than anomaly detection).
- The GRAAFE models achieves an area under the curve (AUC) from 0.91 to 0.78, surpassing state-of-the-art (SoA), achieving AUC between 0.64 and 0.5.





GRAAFE Graph Neural Network (GNN) model structure [1]

Table 1. The rack-level GNN outperforms (achieves higher AUC score) all other methods across all future windows (FW).

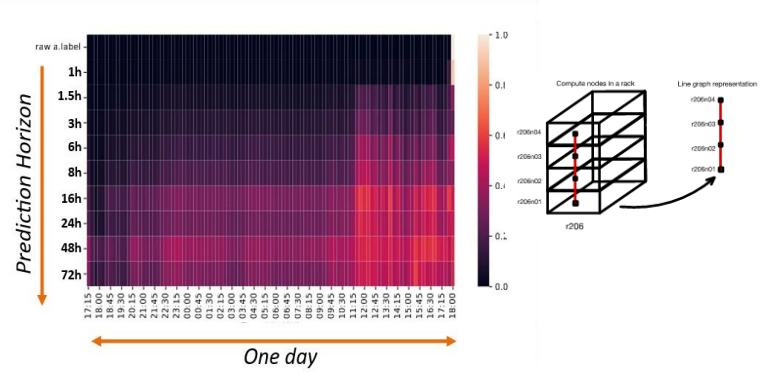
FW	rack GNN	room GNN	DNN	GB	RF	DT	MC
4	0,91	0,59	0,64	0,63	0,61	0,51	0,5
6	0,89	0.58	0,66	0,64	0,59	0,5	0,5
12	0,84	0.47	0,65	0,63	0,59	0,5	0,5
24	0,78	0.58	0,62	0,6	0,55	0,5	0,5
32	0,75	0.55	0,59	0,58	0,55	0,5	0,5
64	0,66	0.42	0,5	0,48	0,49	0,49	0,5
96	0,62	0.58	0,55	0,51	0,58	0,51	0,5
192	0,55	0.52	0,47	0,48	0,52	0,5	0,5
288	0,53	0.50	0,52	0,51	0,52	0,49	0,5

[1] Martin Molan, Mohsen Seyedkazemi Ardebili, Junaid Ahmed Khan, Francesco Beneventi, Daniele Cesarini, Andrea Borghesi, Andrea Bartolini, GRAAFE: GRaph Anomaly Anticipation Framework for Exascale HPC systems, Future Generation Computer Systems, Volume 160, 024, Pages 644-653, SSN 0167-739X,https://doi.org/10.1016/j.future.2024.06.032.

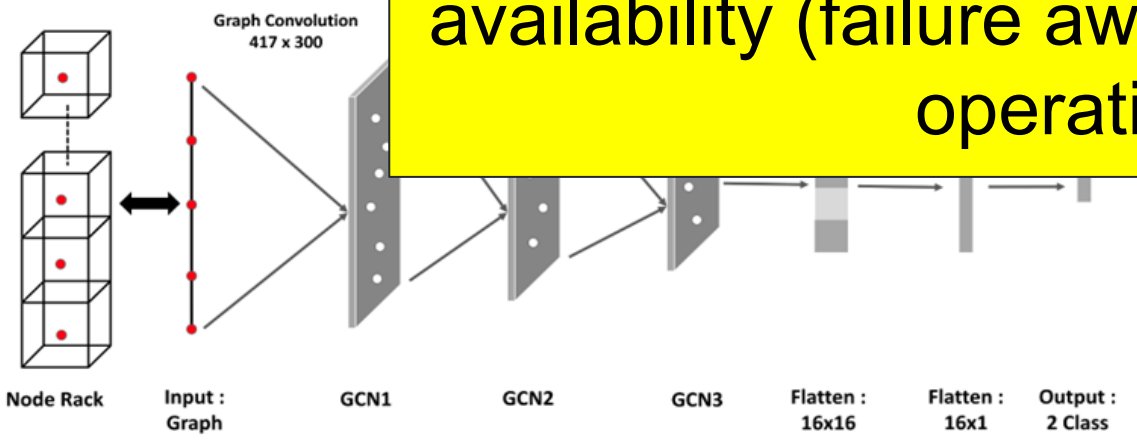
Data Centre Digital Twin for Sustainable Computing

GRAPH MASSIVIZER

- Predictive maintenance: Developed Graph Neural Network (GNN) based models called “GRAAFE” for anomaly prediction (much more impactful than anomaly detection).
- The GRAAFE models achieves an area under the curve (AUC) from 0.91 to 0.78, surpassing state of the art (SoA), achieving AUC between 0.64 and 0.5.



Prediction of anomalies => Improved system availability (failure aware scheduling) => increased operational efficiency



GRAAFE Graph Neural Network (GNN) model structure [1]

) all other methods

						DT	MC
						0,51	0,5
						0,5	0,5
12	0,84	0.47	0,65	0,63	0,59	0,5	0,5
24	0,78	0.58	0,62	0,6	0,55	0,5	0,5
32	0,75	0.55	0,59	0,58	0,55	0,5	0,5
64	0,66	0.42	0,5	0,48	0,49	0,49	0,5
96	0,62	0.58	0,55	0,51	0,58	0,51	0,5
192	0,55	0.52	0,47	0,48	0,52	0,5	0,5
288	0,53	0.50	0,52	0,51	0,52	0,49	0,5

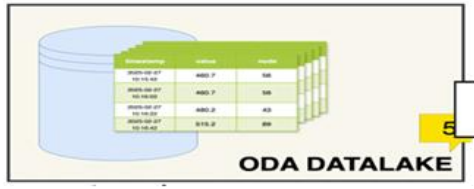
[1] Martin Molan, Mohsen Seyedkazemi Ardebili, Junaid Ahmed Khan, Francesco Beneventi, Daniele Cesarini, Andrea Borghesi, Andrea Bartolini, GRAAFE: GRaph Anomaly Anticipation Framework for Exascale HPC systems, Future Generation Computer Systems, Volume 160, 024, Pages 644-653, SSN 0167-739X,https://doi.org/10.1016/j.future.2024.06.032.

Data Centre Digital Twin for Sustainable Computing

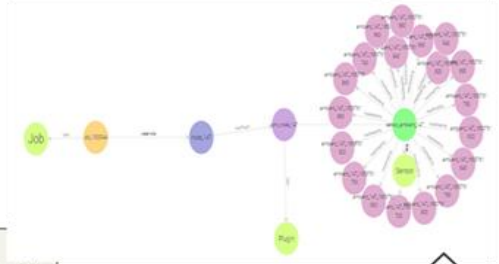
GRAPH MASSIVIZER

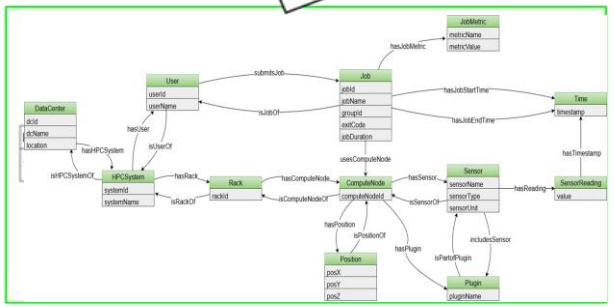
Objective: reduce management complexity

Graph to structure the unstructured ODA/telemetry data



EXAMON to KG (RDF)





ODA Ontology [2]

SoA ODA query :

"find all Jobs which caused a compute node overheating in the last XX hours"

```
def get_nodes_list(jobId,time_period):
  data = sq.SELECT('*') \
      .FROM('job_info_marconi100').WHERE(job_id=str(jobId)).TSTART(time_period
      [0]).TSTOP(time_period[1]).execute()
  df = pd.read_json(data)                    # create df of the query result
  # get the allocated nodes list
  dict_of_nodes = df['cpus_alloc_layout'][0]
  try: nodes = list(dict_of_nodes.keys())
  except: pass
  df = pd.read_json(data)                    # create df of the query result
  df['cpus_alloc_layout'][0]
  nodes = list(dict_of_nodes.keys())
  return nodes
sq.jc.JOB_TABLES.add('job_info_marconi100')        # Setup for Marconi100
data = sq.SELECT('*').FROM('job_info_marconi100').TSTART({start_time}).TSTOP({
      end_time}).execute()
df = pd.read_json(data)
job_ids = df['job_id'].to_numpy()
node_used_in_job_list = []
for job_id in job_ids:
  try: nodes_list = get_nodes_list(job_id,time_period)
    if {node_to_check in nodes_list):
      print(job_id,nodes_list)
      node_used_in_job_list.append(job_id)
  except: pass
def get_data(node_to_get,metric,start_time,end_time):
  data = sq.SELECT('*').FROM(metric).WHERE(node=node_to_get).TSTART(str(
      start_time)).TSTOP(str(end_time)).execute()
  value = data.df_table['value']
  return value
def get_job_time(jobId):
    data=sq.SELECT('*').FROM('job_info_marconi100').WHERE(job_id=str(jobId),
      node=node_to_check).TSTART({start_time}).TSTOP({end_time}).execute()
  df = pd.read_json(data)
  start_time = correct_TS_format(str(df['start_time'][0]))
  end_time = correct_TS_format(str(df['end_time'][0]))
  return start_time,end_time
each_job_df = []
for job in node_used_in_job_list:
  start_time,end_time = get_job_time(job)
  try: df = get_data(node_to_check,metric,start_time,end_time)
      each_job_df.append((max(df),min(df),(df.sum()/len(df))))
  except: print("error")
```

GM enabled

Query execution

```
query = f"""SELECT ?nodeId (AVG(?temperature) as ?avgTemperature) (MIN(?
      temperature) as ?minTemperature) (MAX(?temperature) as ?maxTemperature)
WHERE {{?job rdf:type cineca_m100:Job ;
            cineca_m100:startTime ?jobStart ;
            cineca_m100:endTime ?jobEnd ;
            cineca_m100:usesNode ?node .
        ?node cineca_m100:hasPlugin/cineca_m100:hasSensor ?sensor ;
            cineca_m100:nodeId ?nodeId .
        ?sensor cineca_m100:sensorName "temperature" ;
            cineca_m100:hasReading ?reading .
        ?reading cineca_m100:value ?temperature ;
            cineca_m100:timestamp ?timestamp ;
            cineca_m100:unit ?unit .
        FILTER(?jobStart <= "{end_time}"^^xsd:dateTime && ?jobEnd >= "{
      start_time}"^^xsd:dateTime)
}}GROUP BY ?nodeId"""
```

SPARQL (Graph) query

NoSQL/SQlite query - Standard query language in SoA

[1] J. Khan et al., ExaQuery: Proving Data Structure to Unstructured Telemetry Data in Large-Scale HPC, ACM GraphSys24

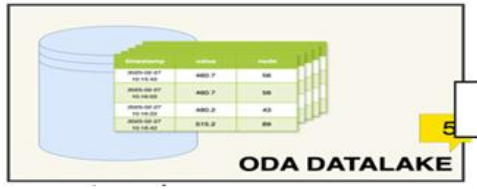
[2] J. Khan et al., A Unified Ontology for Scalable Knowledge Graph-Driven Operational Data Analytics in High-Performance Computing Systems, ARXIV

Data Centre Digital Twin for Sustainable Computing

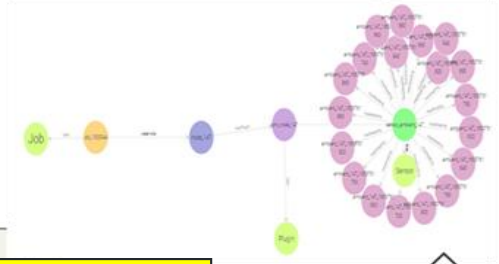
GRAPH MASSIVIZER

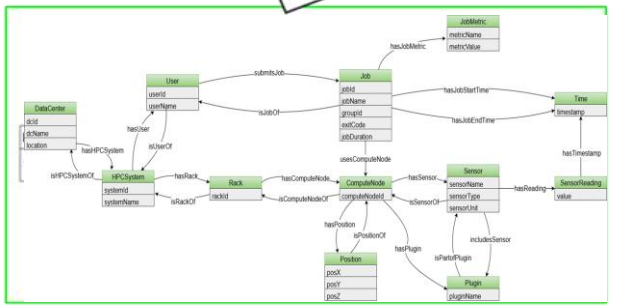
Objective: reduce management complexity

Graph to structure the unstructured ODA/telemetry data



EXAMON to KG (RDF)

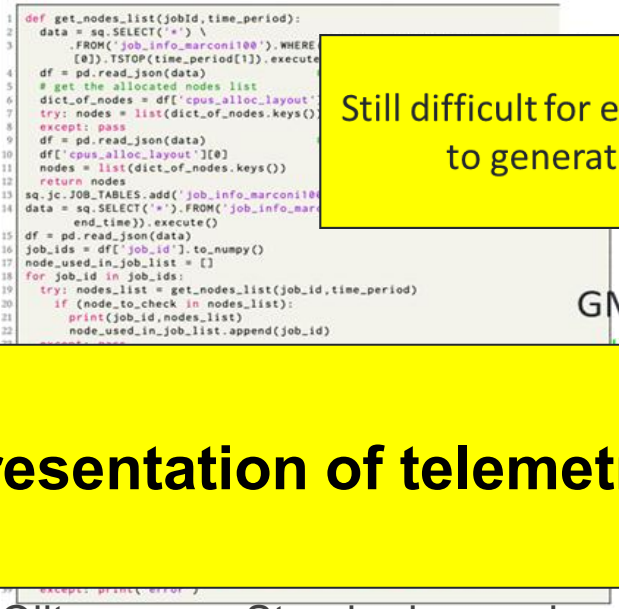




ODA Ontology [2]

SoA ODA query :

"find all Jobs which caused a compute node overheating in the



Still difficult for end user to generate

Query execution

Faster than NoSQL query

GM enabled

Graph representation of telemetry -> lowers knowledge access costs

NoSQL/SQlite query - Standard query language in SoA

[1] J. Khan et al., ExaQuery: Proving Data Structure to Unstructured Telemetry Data in Large-Scale HPC, ACM GraphSys24

[2] J. Khan et al., A Unified Ontology for Scalable Knowledge Graph-Driven Operational Data Analytics in High-Performance Computing Systems, ARXIV

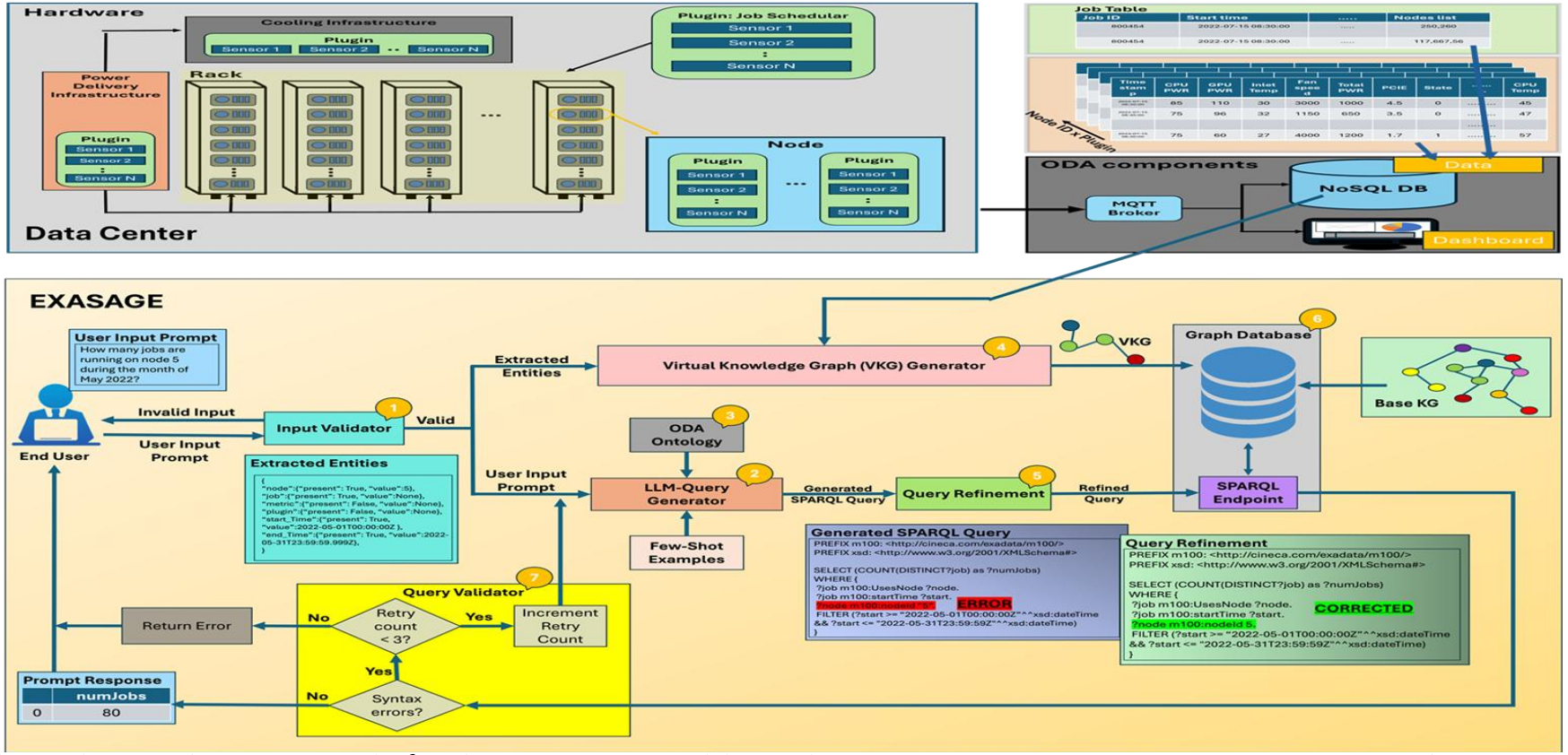
Data Centre Digital Twin for Sustainable Computing

GRAPH MASSIVIZER

Objective: Provide a conversational interface to Examon/ODA collected data.

Interact with the ODA in natural language

Tool: EXASAGE: The first operational data analysis assistant [1]



[1] Junaid Ahmed Khan, Martin Molan, Andrea Bartolini. *EXASAGE: The first data center operational data analysis assistant.* Future Generation Computer Systems, Volume 176, 2026, Article 108185. ISSN 0167-739X. https://doi.org/10.1016/j.future.2025.108185

[2] J. Khan et al., From Data Center IoT Telemetry to Data Analytics Chatbots -- Virtual Knowledge Graph is All You Need (https://arxiv.org/abs/2506.22267)

Data Centre Digital Twin for Sustainable Computing

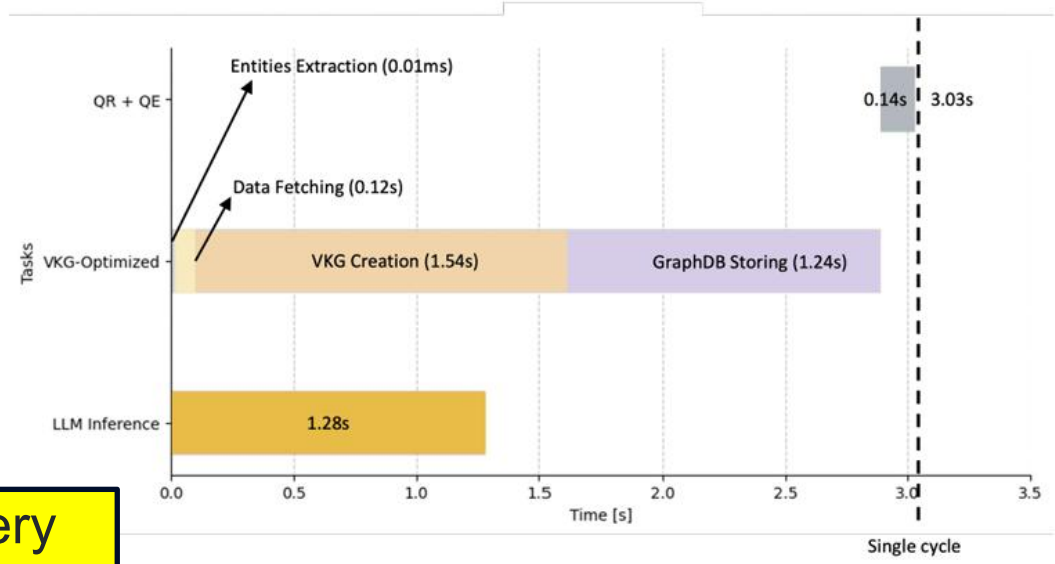
GRAPH MASSIVIZER

Objective: Provide a conversational interface to Examon/ODA collected data.
Tool: EXASAGE: The first operational data analysis assistant [1]

End-to-End latency = 3.03 s

	IoT Datalake	End-to-End Latency [s]	Accuracy [%]	Storage Size [GiB]
LLM-to-NoSQL/SQLite	NoSQL		25	-
LLM-to-SPARQL	Knowledge Graph	1.42	93.6	2979.84
LLM-to-SPARQL via virtualization of knowledge graphs	NoSQL	9.04	93.6	0.17
LLM-to-SPARQL via virtualization of knowledge graphs	Parquet	6.92	93.6	0.17

LLM (llama 3 8B) query generation accuracy [2]



Single input cycle - Time diagram [2]

Significant performance in automation of graph query (SPARQL) generation, showcasing the benefits of the Graph-Massivizer approach

[1] Junaid Ahmed Khan, Martin Molan, Andrea Bartolini. *EXASAGE: The first data center operational data analysis assistant.* Future Generation Computer Systems, Volume 176, 2026, Article 108185. ISSN 0167-739X. https://doi.org/10.1016/j.future.2025.108185
[2] J. Khan et al., From Data Center IoT Telemetry to Data Analytics Chatbots -- Virtual Knowledge Graph is All You Need (https://arxiv.org/abs/2506.22267)

Data Centre Digital Twin for Sustainable Computing- KPI Achievement

KPI-4.1.1	Details
Description	100% spatial interaction captured by the massive DC-MG
Goal	Achieve full (100%) spatial interaction coverage within the DC-MG.
Achievement	The system achieved a complete representation of spatial information across the entire DC-MG. By incorporating the concepts of Rack, Node, and Position, along with their respective relationships within the proposed UC4 ontology, a full 100% spatial interaction coverage was successfully achieved.

KPI-4.2.1	Details
Description	over 4,000 active users.
Goal	Exceed 4,000 active users.
Achievement	The target of 4,000 active users was surpassed, with Leonardo HPC system recording **6,925** active users in 2024–an increase of 1,617 compared to 2023.

KPI-4.2.3	Details
Description	20% better utilisation, modelling two partitions with 90% critical node identification.
Goal	Achieve 20% better utilisation while accurately identifying 90% of critical nodes.
Achievement	Higher utilisation was enabled through predictive compute node availability using GRAAFE, a GNN-based anomaly prediction framework. The model operates on all Marconi100 nodes every 120 s and achieves an **AUC of 0.91–0.78**, indicating high accuracy in predicting node availability. This is achieved with only **30% additional CPU** and **<5% extra RAM** compared to standard monitoring.

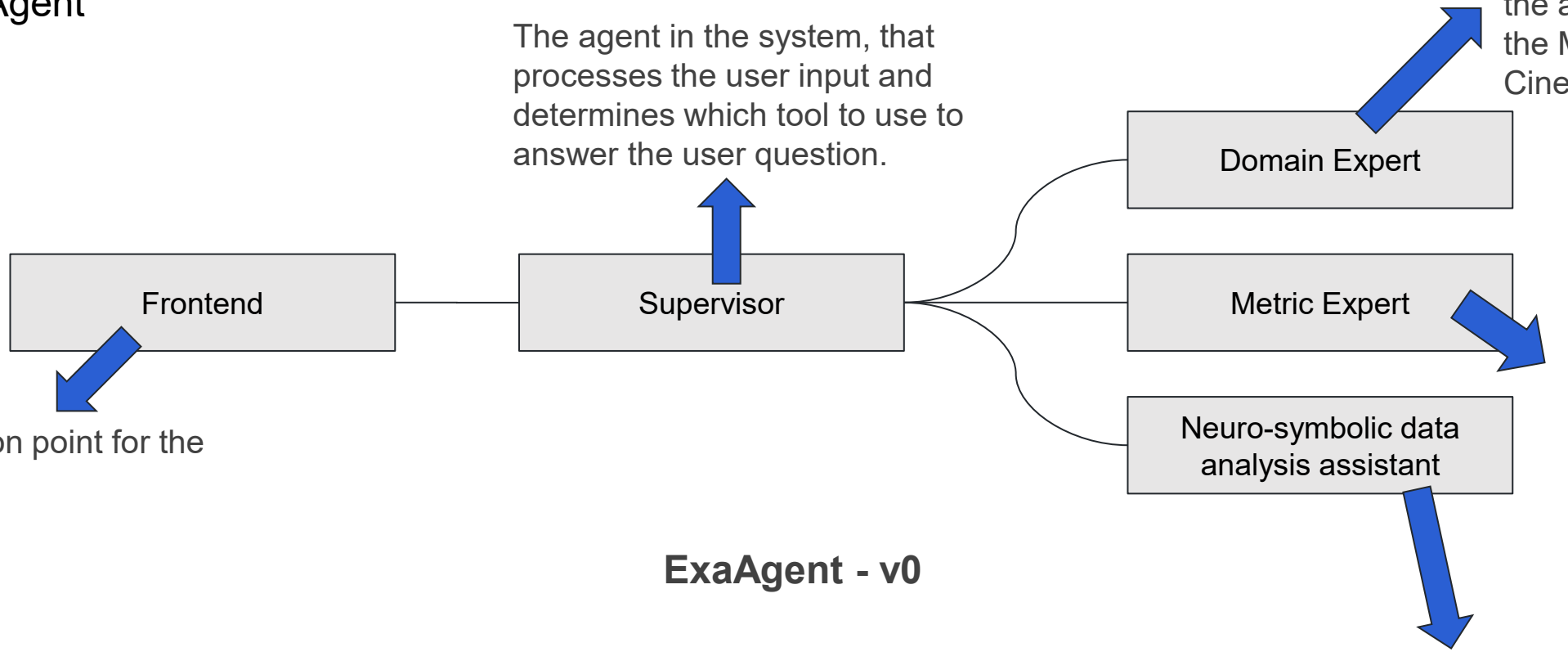
[1] Junaid Ahmed Khan, Martin Molan, Andrea Bartolini. *EXASAGE: The first data center operational data analysis assistant.* Future Generation Computer Systems, Volume 176, 2026, Article 108185. ISSN 0167-739X. https://doi.org/10.1016/j.future.2025.108185

[2] J. Khan et al., From Data Center IoT Telemetry to Data Analytics Chatbots -- Virtual Knowledge Graph is All You Need (https://arxiv.org/abs/2506.22267)

Data Centre Digital Twin for Sustainable Computing

Objective: Provide an agentic framework for the data center

Tool: ExaAgent



The agent in the system, that processes the user input and determines which tool to use to answer the user question.

Provides answers to questions related to data center and HPC domain. For example: “What is the architecture of the M100 system at Cineca?”, etc

The interaction point for the user

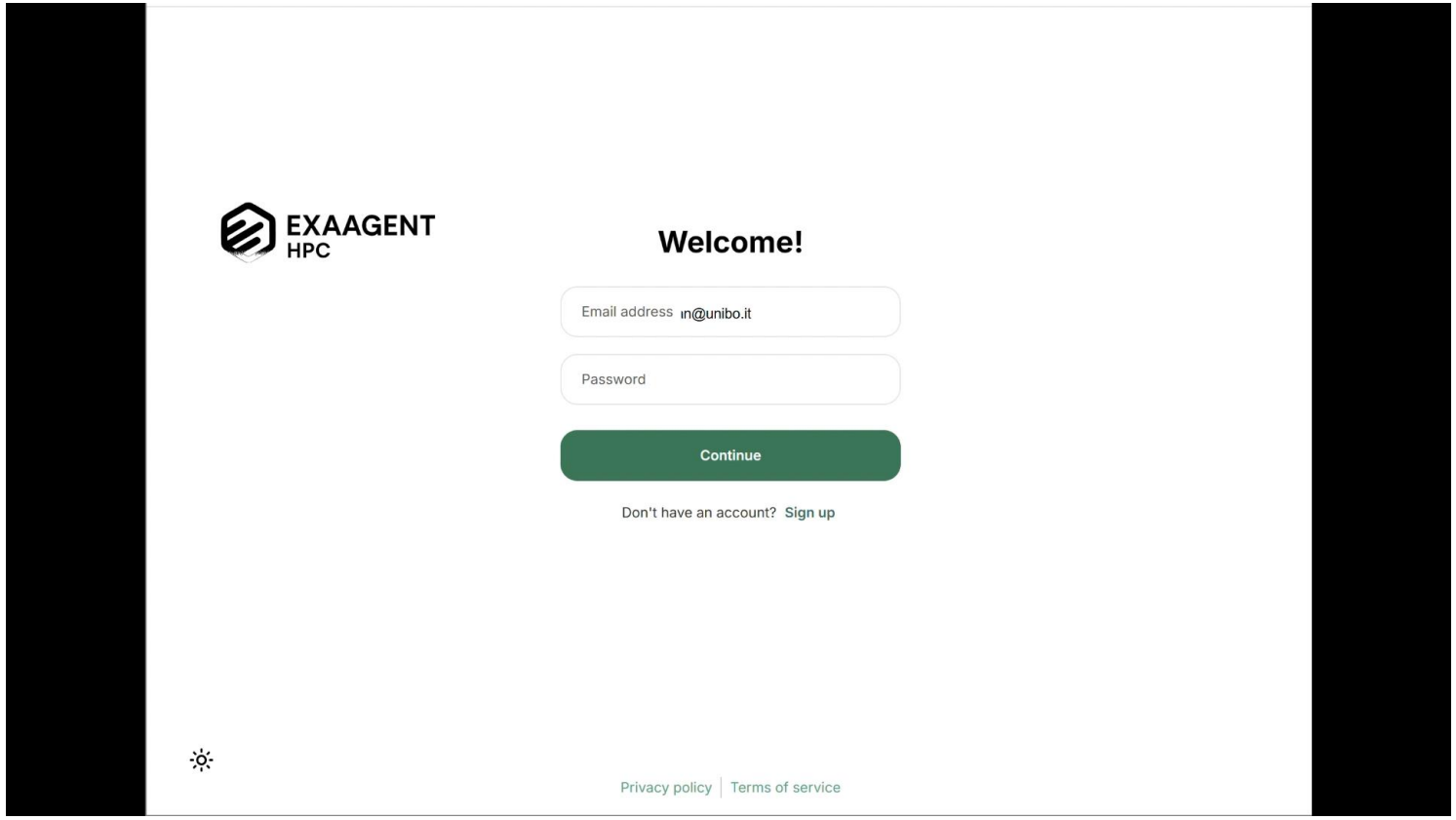
Provides answers to metrics related queries of the user For example: “What sensors are listed in the IPMI plugin?”, etc

ExaAgent - v0

EXASAGE: The first data analysis assistant

Data Centre Digital Twin for Sustainable Computing

Demo:
ExaAgent
https://gitlab.com/ecs-lab/exaagent/-/blob/main/ExaAgent__demo_.gif



NEARDATA

neardata.eu

Extreme Near-Data Processing Platform

Pedro García
Senior Researcher
University Rovira i Virgili

Pedro García Lopez is full professor of the Computer Engineering and Mathematics Department at the University Rovira i Virgili (Spain). He leads he “Cloud and Distributed Systems Lab” research group and has coordinated three large European research projects in the last years. His research topics are distributed systems, cloud computing, data analytics, software architectures and middleware.

Use cases I: problem solved, who benefits
METABOLOMICS

NEARDATA cloud platform for spatial metabolomics

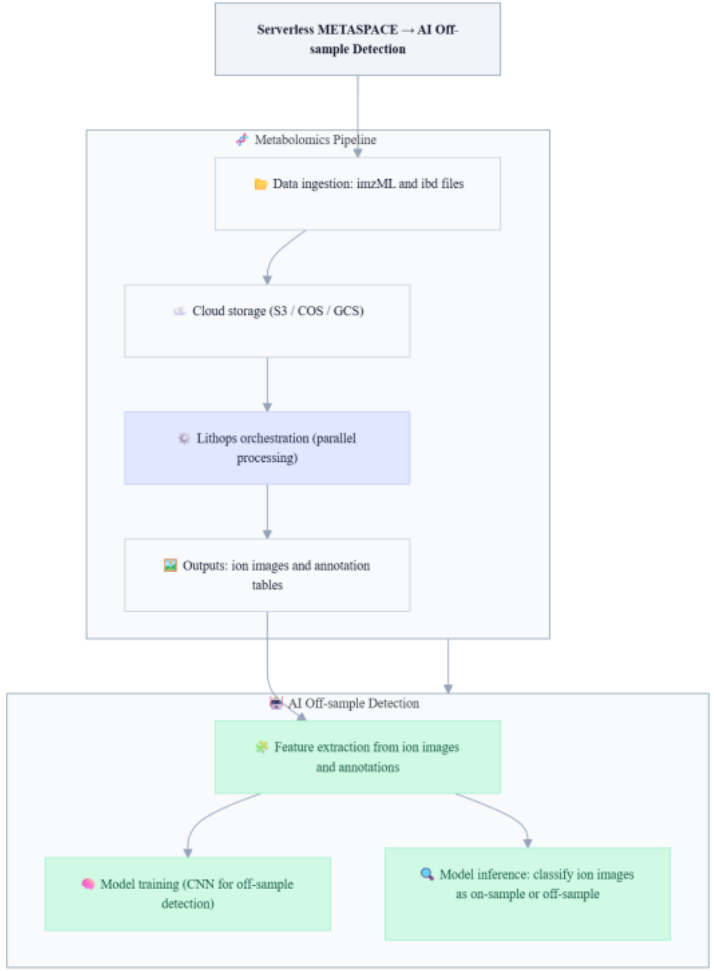
Problem Solved:
This use case solves the challenge of unifying, scaling, and securely processing fragmented metabolomics data from multiple sources to enable faster and more reliable molecular analysis.

Who Benefits:
Scientists and healthcare researchers benefit from this solution, as it allows them to access and analyze large, diverse metabolomics datasets more easily and securely, leading to better insights into health and disease.

AI Innovations Achieved:

- **Smarter metabolite detection**: Replaced rules with machine learning for better accuracy.
- **Scalable AI workflows**: Serverless pipelines handle large datasets efficiently.
- **Fast data prep**: Automated processing enables quick, reliable analysis.

Serverless data processing connected to AI pipeline



Use cases II: problem solved, who benefits
GENOMICS

Optimizing large genomic pipelines

Problem Solved:
The challenges of handling and analyzing massive genomic datasets by creating a faster, more flexible, and easier-to-use system for studying genetic variants linked to complex diseases.

Who Benefits:
Scientists and medical researchers benefit from this solution, as it helps them more quickly uncover genetic factors that contribute to human health and disease.

Extreme data analytics
Serverless Data Processing in HPC SuperComputers

AI Innovations Achieved:

- **Combinatorial Machine Learning:** Detects complex interactions among genetic variants influencing disease risk.
- **Automated Optimization:** Uses AI to speed up model tuning and improve prediction accuracy.
- **Intelligent Autoscaling:** Adjusts computing resources in real time for faster, more efficient analysis.

LITHOPS - HPC

Use cases II: problem solved, who benefits

SURGERY

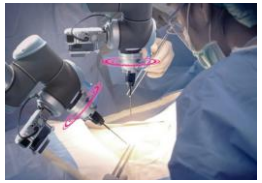
Real time surgery video analytics

Problem Solved:

Federated learning enables hospitals to collaboratively train powerful AI models for surgical assistance without sharing sensitive patient data, overcoming strict privacy regulations like GDPR.

Who Benefits:

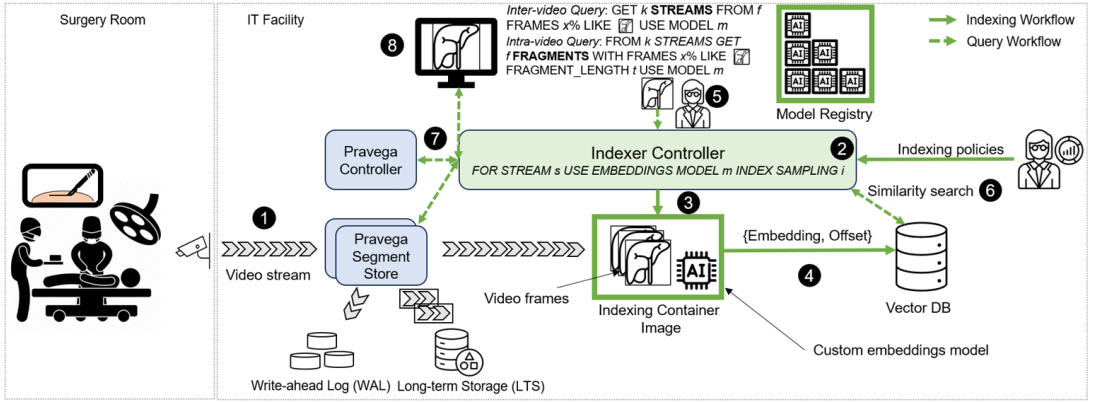
This approach benefits both surgeons, who gain access to more accurate and adaptable decision-support tools, and patients, who receive safer and more personalized surgical care.



Confidential federated learning
Semantic search
Serverless VectorDB

AI Innovations Achieved:

- **Smart video search**: AI indexes surgical videos for fast, semantic search and training.
- **Privacy-preserving training**: Federated learning trains models across hospitals without sharing patient data.
- **Real-time surgical support**: AI models assist surgeons live with phase, tool, and anatomy detection.



Impact, noteworthy results

KPIs Summary of the KPIs achieved by each of the use cases

KPI-1 - Significant performance improvements (data throughput, data transfer reduction) in Extract-Transform-Load (ETL) phases validated with near-data connectors over extreme data volumes (genomics, metabolomics).

KPI-2 - Significant data speed improvements (throughput, latency) in real-time video analytics validated using stream data connectors.

KPI-3 - Demonstrated resource auto-scaling for batch and stream data processing validated thanks to data-driven orchestration of massive workflows.

KPI-4 - High levels of data security and confidential computing validated using TEEs and federated learning in adversarial security experiments.

KPI-5 - Demonstrated simplicity and productivity of the software platform validated with real user communities in International Health Data Spaces.

KPI	Use Case	Summary of results
KPI-1	Epistasis(GWD)	Lithops-HPC connector improves data ingestion in GWD pipeline by **24x**.
KPI-1	Epistasis(GWD)	Hyperparameter selection improves performance by **5x**.
KPI-5	Epistasis(GWD/MDR)	Cyclomatic-Complexity reveals **1.5x** fewer execution paths.
KPI-5	Epistasis(GWD/MDR)	Yaqin's metrics reveals **1.6x** fewer branches, loops and nesting depth.
KPI-3	Epistasis(GWD/MDR)	The auto-scaler improves execution time of MDR use-case by **1.5x**.
KPI-2	Surgery	StreamSense achieve low-latency ranging between **63ms** and **360ms**
KPI-2	Surgery	Semantic video search latency: reduced latency up to **51%** lower.
KPI-1	Surgery	Data transfer savings in AI loading. Integrating with PyTorch data transfers are reduced **between 83.79% and 99.83%**.
KPI-4	Surgery	Enhanced encryption of files with TEE, access control meachisnms. The security of the system was rigorously validated through adversarial testing and TEE attestation, confirming that both the confidentiality and integrity of model updates and training data were consistently enforced.
KPI-1	Transcriptomics	Early stopping technique have increased alignment throughput by **19.5%**.
KPI-1	Transcriptomics	Use of a newer release of human genome index has resulted in execution times improvements of up to **12x** and smaller STAR index file (from 85GB to 30GB).
KPI-1	Transcriptomics	The usage of spot instances reduced, on average, **50%** of the execution cost.
KPI-1	Genomics	The Nexus FASTQGzip streamlet provides a good trade-off on compression and processing time, **3.8x** better compression ratio and **12.1** of processing time for $\lambda = 38$.
KPI-1	Genomics	FaaStream is on average **65.14%** cheaper than Flink.
KPI-1	Metabolomics	Depending on the size, we get a speed-up on processing time ranging from **1.13x to 1.22x** faster.
KPI-4	Metabolomics	Achieved full confidential computing support (data at rest, in transit and in use) on cloud storage service (MinIO) and FaaS Lithops Singularity, aided by SCONE mechanisms and using the TEE; achieved partial confidential computing support (data at rest and in transit) due to limitations of the ported system (Metaspace is not fully compliant with Lithops Singularity + SCONE).

NEARDATA

Impact, noteworthy results

SOFTWARE OUTCOMES

Summary of NEARDATA Software Components and Tools

Component / Tool	Category	Brief Description
Lithops	Production-Ready	Serverless function execution framework deployed close to the data.
Scone	Production-Ready	Confidential computing framework enabling secure, policy-driven execution of containerized workloads in Trusted Execution Environments (TEEs) across multi-cloud infrastructures.
Pravega	Production-Ready	Scalable and elastic stream storage system.
Metaspace	Production-Ready	Metadata catalog designed for large, unstructured datasets.
PyRun	Production-Ready	Serverless Python studio that enables users to write, run, and scale data science or AI workloads effortlessly on their own cloud accounts without managing infrastructure.
DataPlug	Production-Project	A framework that enables efficient, read-only, cloud-aware partitioning of unstructured scientific data in object storage through dynamic, parallel, and metadata-driven data slicing for elastic workloads.
DataCockpit	Production-Project	An interactive widget that allows scientists to browse, partition, and benchmark datasets from Amazon S3 or Metaspace directly within Jupyter notebooks for elastic, parallel data processing.

Component / Tool	Category	Brief Description
Serverless Vector DB	Research-Project	Prototype of a serverless vector database enabling similarity search over large embeddings with elastic scaling.
Lithops-HPC	Research-Project	Lithops framework to run on HPC environments.
FaaStream	Research-Project	Experimental framework combining serverless execution with stream-based storage for low-latency, elastic pipelines.
Nexus	Research-Project	Data mesh prototype introducing streamlets and swarmlets to support metadata-driven orchestration of heterogeneous backends.
Glider	Research-Project	Serverless system that enables stateful near-data computation on ephemeral storage, drastically reducing data movement and improving the performance and efficiency of serverless data analytics pipelines.
FaaSTs	Research-Project	Experimental framework to predict resource usage on FaaS environments leveraging AI and time-series.
Burst Computing	Research-Project	Experimental framework extending the serverless model with group-based invocations and efficient inter-worker communication, enabling synchronized, massively parallel workloads with reduced startup latency and data movement.

Impact, noteworthy results

PUBLICATIONS: more than 30 publications

USENIX ATC

- "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'25**

ACM/IFIP MIDDLEWARE

- "Pravega: A Tiered Storage System for Data Streams", Raúl Gracia-Tinedo, Flavio Junqueira, Tom Kaitchuck, Sachin Joshi. **ACM/IFIP Middleware'23 (Best Paper Award)**
- "Practical Storage-Compute Elasticity for Stream Data Processing", Raúl Gracia-Tinedo, Flavio Junqueira, Brian Zhou, Yimin Xiong, Luis Liu. **ACM/IFIP Middleware'23 Industry Track**
- "StreamSense: Policy-driven Semantic Video Search in Streaming Systems", G. Finol, A. Gabriel, P. García-López, R. Gracia-Tinedo, L. Liu, R. Docea, M. Kirchner, S. Bodenstedt. **ACM/IFIP Middleware'24 Industry Track**
- ""Back to the Byte": Towards Byte-oriented Semantics for Streaming Storage", R. Gracia-Tinedo, F. Junqueira, T. Kaitchuck. **ACM/IFIP Middleware'24 Industry Track**

ACM SIGMOD

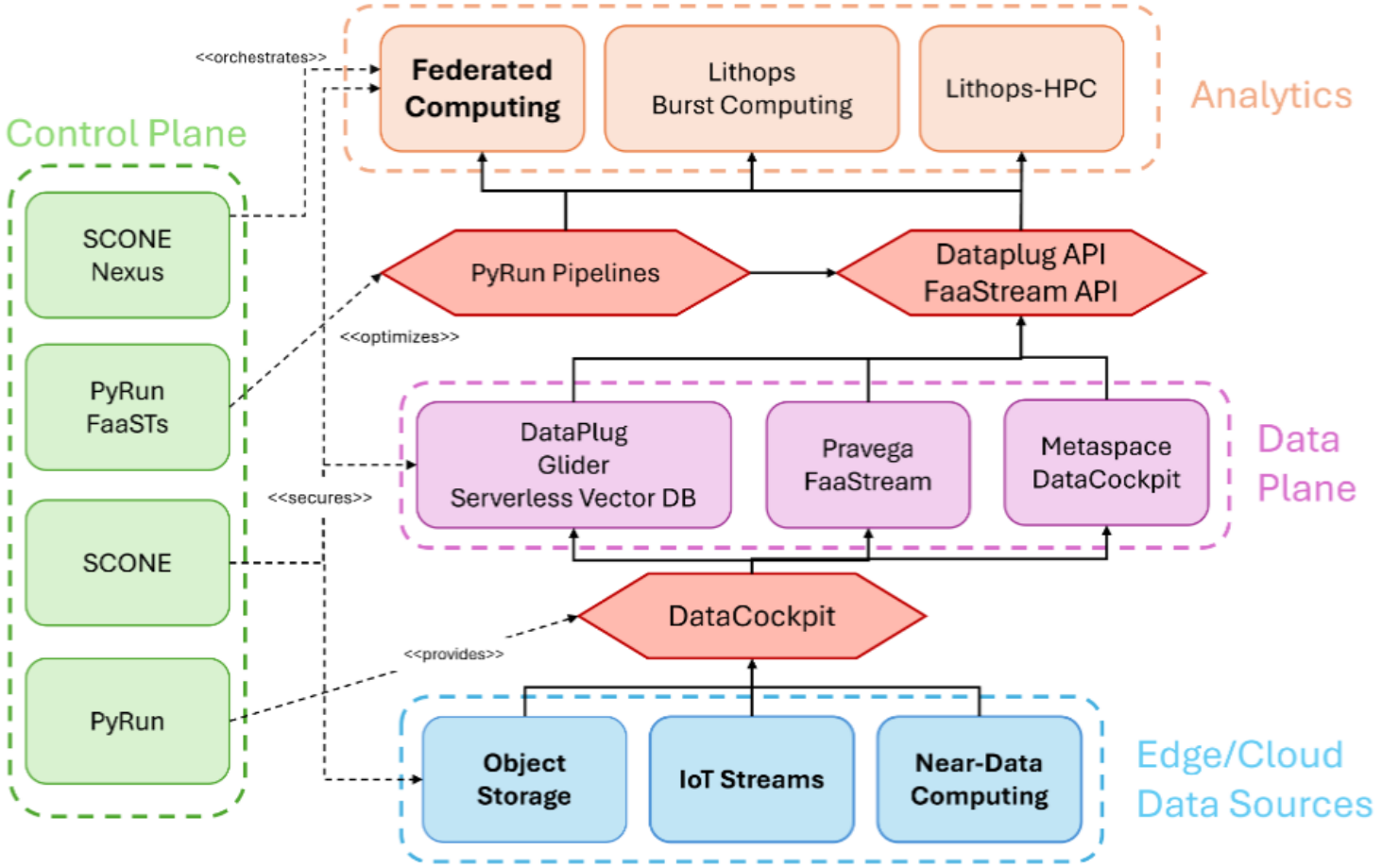
- "Building Stateless Serverless Vector DBs via Block-based Data Partitioning". D. Barcelona-Pons, R. Gracia-Tinedo, X. Roca-Canals, A. Cañadilla-Domingo, P. García-López. **ACM SIGMOD'26,** [Experiments & Analysis] (Accepted)

NATURE

- "Spatial metabolomics: from a niche field towards a driver of innovation", Theodore Alexandrov, **Nature Metabolism**, 2023
- "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**, 2023
- "METASPACE-ML: Metabolite annotation for imaging mass spectrometry using machine learning", Bishoy Wadie, Lachlan Stuart, Christopher M. Rath, Theodore Alexandrov, **Nature Communications**, 2024

https://neardata.eu/publication

Innovation, technology developed



LITHOPS

METASPACE

Pravega
Reliable streaming storage

SCONE

PyRun

Innovation, what technology was developed

PyRun

pyrun.cloud

Core Concepts

Effortless Cloud Computing
Run scalable Python workloads (Data, AI, Distributed) on your AWS account

Integrated
Unified, Integrated & Automated Platform
VS Code-like UI, Auto Runtime Management

Unified Platform
Managed Lithops (FaaS), Dask, Ray and Cubed

Platform Features

Integrated Web IDE
VS Code-like experience

Automated Runtime Management | Conda, Pip, Dockerfile | Auto-detect & rebuild

Real-Time Monitoring
CPU, Mem, Disk, Net, Gantt

Templates & Pipelines | Quick-start projects & real-world examples

Simplified Configuration
Lithops backend setup via UI

User community

Metabolomics | EMBL (Germany) METASPACE, URV(Spain)
Genomics | Hutton Institute (UK), UK Health Security Agency (UK), BSC (Spain), SANO (Poland)
Geospatial | Alterna (Spain), Kookmin (Korea)
Climate science | Pangeo (US)
Astronomics | Observatoire de Paris (France)
Finance sector | IBM (Israel)
Data/event streaming | Dell (Ireland, Spain)
Video analytics | NCT (Germany)

Target Users & Use Cases

Users | Data Scientists, ML Engineers, Python Devs
AI/ML | TensorFlow, PyTorch, Dask-ML, Ollama
Data Processing | ETL (Lithops), Large Scale Analysis (Lithops & Dask)
Scientific Computing | Climate (CMIP6), Geospatial (NDVI), Simulations (Mandelbrot, Vorticity), Metabolomics (METASPACE)

Data Cockpit

Interactive widget for data selection & partitioning
Sources | User S3, Public Registries (AWS Open Data), Metaspace
Partitioning (via Dataplug) | Intelligent slicing for parallel processing | Auto/Manual batch size, Benchmarking.
Supported Formats | COG, COPC, LIDAR, CSV, FASTA, FASTAQ.GZ, imzML, VCF, Zarr

Key Integrations & Frameworks

Core Execution
Lithops (FaaS/Serverless)
Dask (Distributed Clusters)

AI/ML Libraries
TensorFlow, Keras
Scikit-learn, Dask-ML, Polars
Ollama (for LLMs)

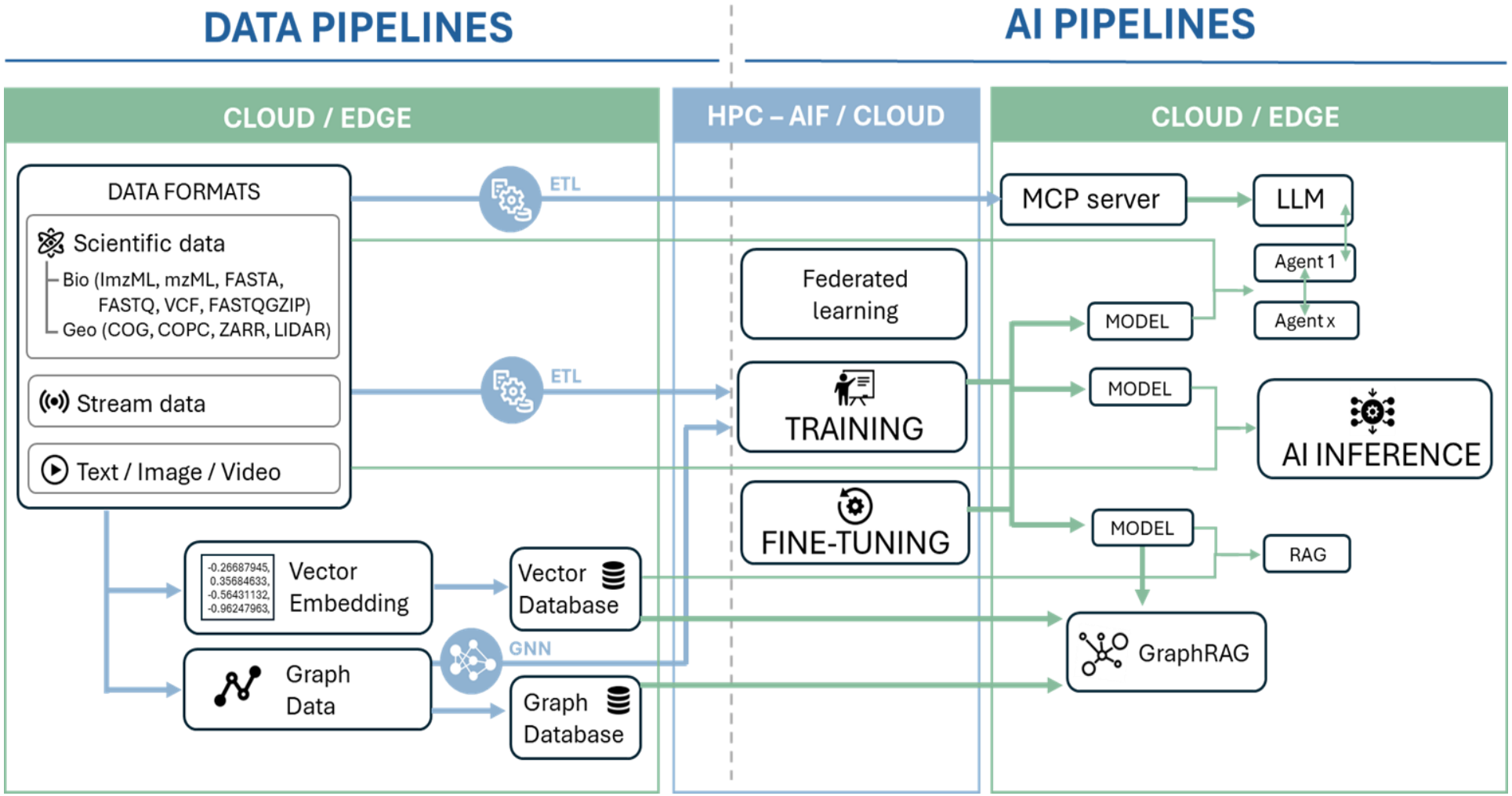
Data Handling
Data Handling:
Xarray, Pandas, NumPy
Rasterio, Matplotlib

Backends

Cloud backends: aws (Amazon EC2, AWS Fargate), IBM Cloud® Code Engine

On premise backends: OpenNebula

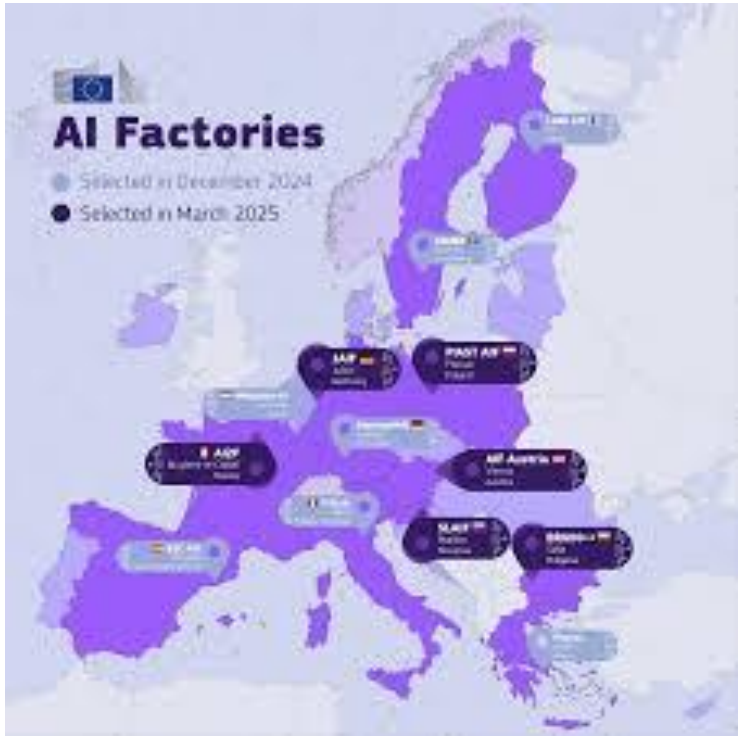
AI Factories: It's time to rethink the Cloud-HPC divide



EUROPEAN BIG DATA VALUE FORUM

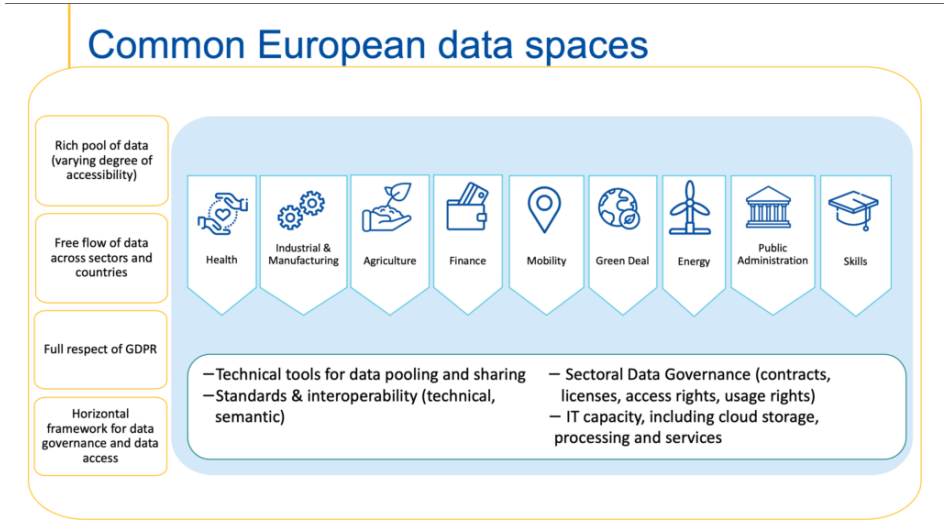
Panel: Complementarities, Overlaps, Shared Impact

How do these initiatives fit within the larger ecosystem?



eosc

AI on Demand



BDV BIG DATA VALUE ASSOCIATION

EUROPEAN BIG DATA VALUE FORUM

Q&A

EXA4MIND

EXTRACT
A distributed data-mining software platform for extreme data across the compute continuum

GRAPH MASSIVIZER

NEARDATA

Learn more

EXA4MIND

EXTRACT

NEARDATA

GRAPH MASSIVIZER

EUROPEAN BIG DATA VALUE FORUM

THANKS!

ORGANISED BY

BDV BIG DATA VALUE ASSOCIATION

IN COLLABORATION WITH

DTU

SDU University of Southern Denmark

ebdvf.eu
#EBDVF25