



The World's Leading  
Data Intelligence Platform

# FUELING THE AI FACTORY: CENTRAL ROLE OF THE METADATA CATALOG

[JTACQUAVIVA@DDN.COM](mailto:JTACQUAVIVA@DDN.COM) OSLO, MARCH 19, 2026

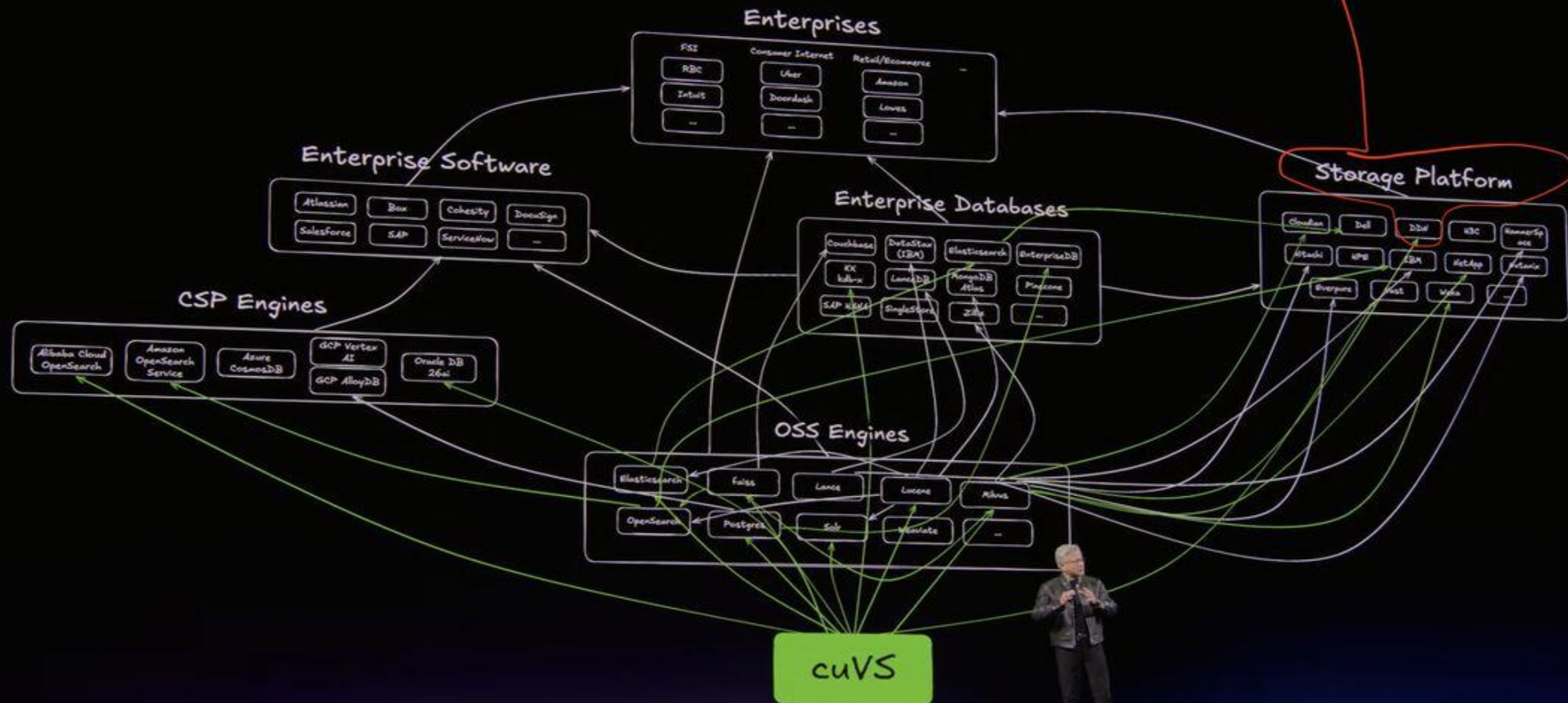
MINIMUM DATACENTER FOOTPRINT FOR MAXIMUM PERFORMANCE

# AI Factory: GPU + AI Model + Tons of Data = Science + Business



32 NVIDIA GB200 NVL72  
2304 B200 GPUS  
1 RACK OF DDN STORAGE  
16 DDN DATA APPLIANCES: AI400X3  
3 DDN METADATA APPLIANCES: AI400X3

## Unstructured Data is the Context of AI 100's of Zettabytes Per Year of Unstructured Data – Growing Exponentially



# Where do we meet?

<b>Criteria</b>	<b>HPC Storage</b>	<b>AI Storage</b>	<b>CS3 Community</b>
Data volume	PB	PB	PB
Structure	Structured	Un-structured	Un-structured
API	Standard POSIX	Interoperable	Interoperable
Fault Tolerance	Best effort	Always on	Always on
Metadata	File system MD	Semantic MD	-/
<b>Platform</b>	<b>On-Prem</b>	<b>Multi-cloud</b>	<b>Multi-Cloud</b>
<b>Key Metric</b>	<b>Performance</b>	<b>Business outcome</b>	<b>Service</b>



# What do we see from Research: DaFab

Funded by EUPSA  
4 Academic Partners  
4 Industrial Partners

DaFab HORIZON-EUSPA-2022-SPACE

Grant Number: 101128693



# Earth Observation

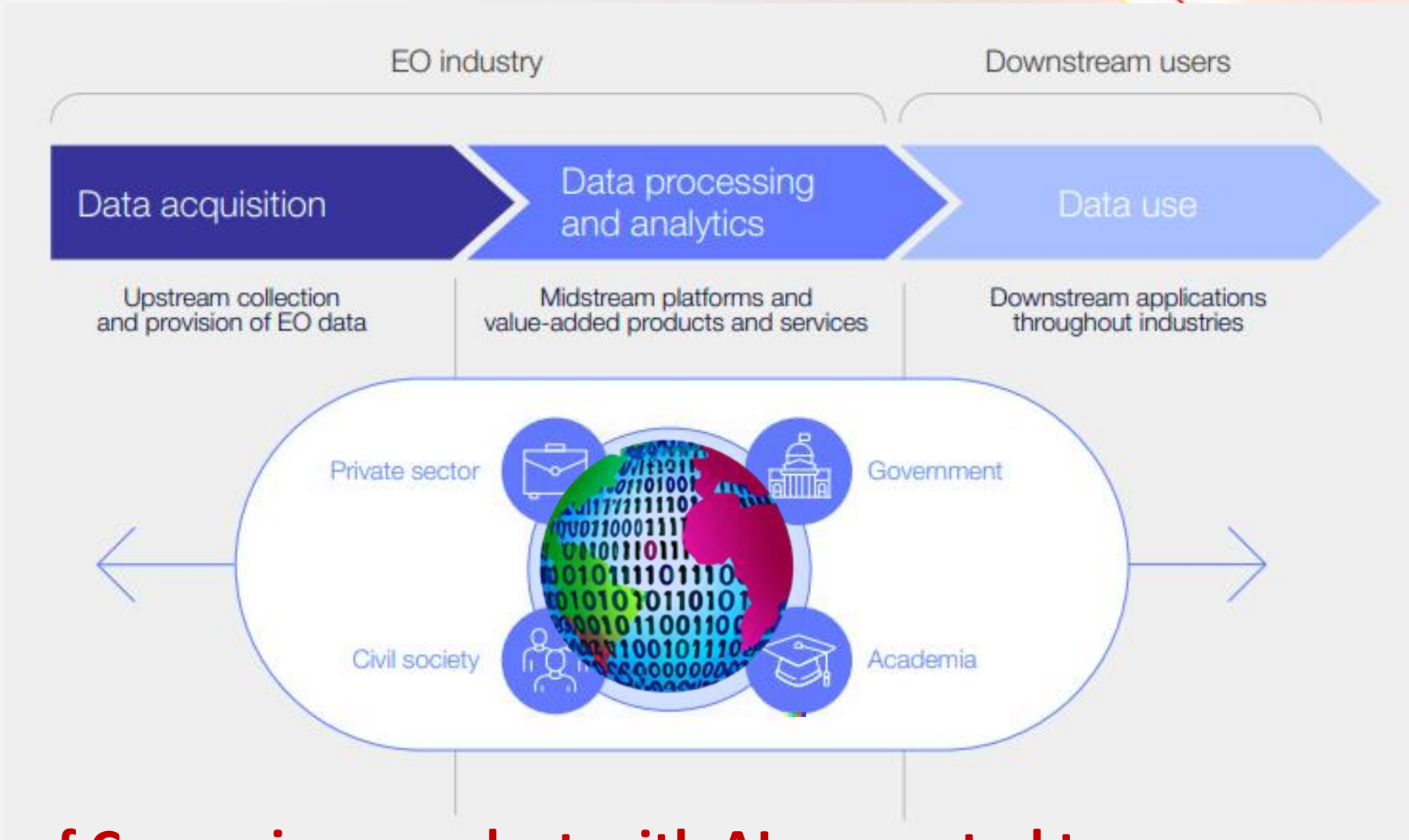
Data-intensive field with major economic and societal impact

- x3 in the next 5 years
- Saving 2 GT of carbon emission

Data are available; it is now a computational problem



# Value Chain: Data Broker from Acquisition to Exploitation



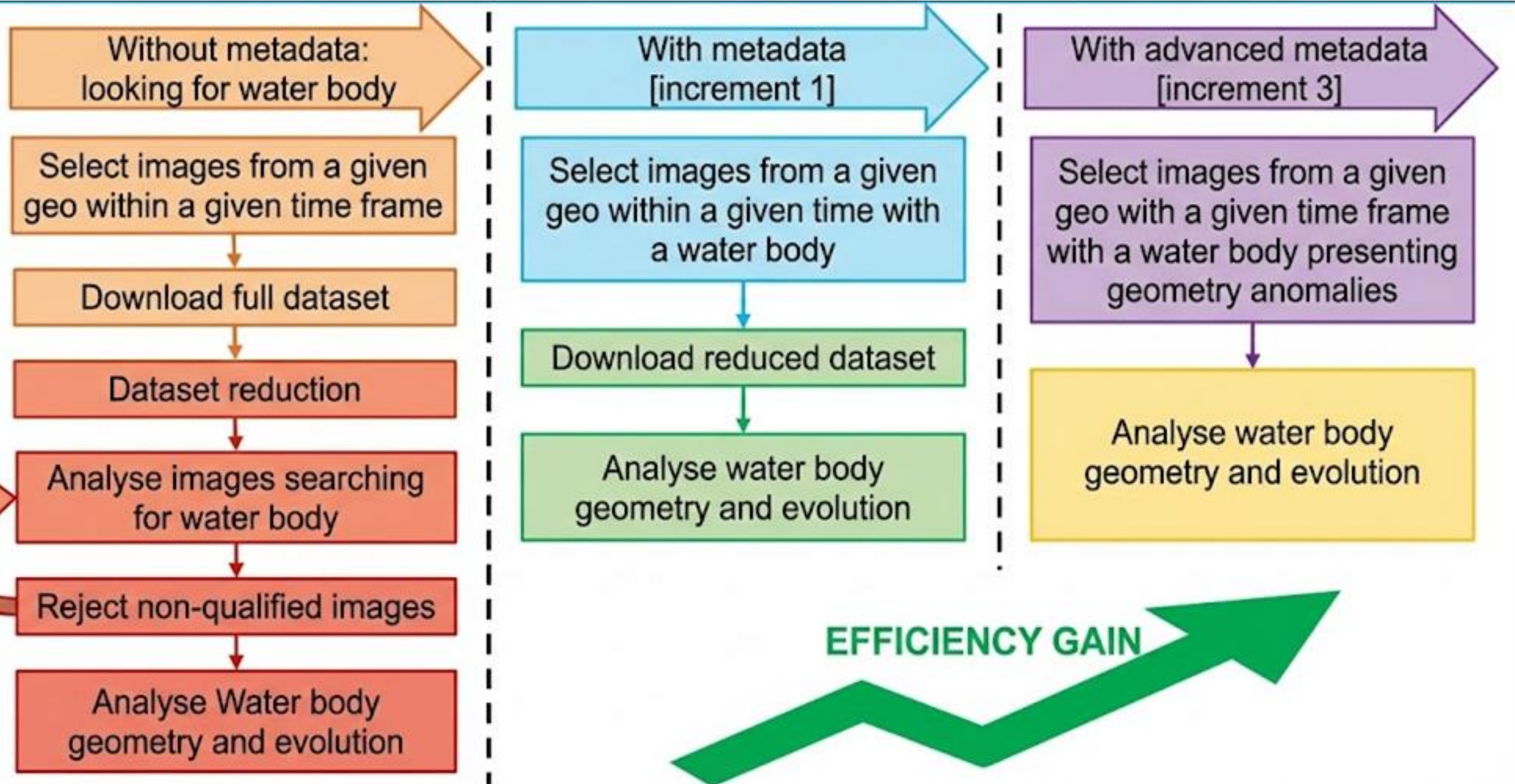
Copernicus 47 PB



**Annotation of Copernicus product with AI generated tags**

2 EU SMEs already using DaFAB data broker services

# Unified Metadata Catalog as an Enabler

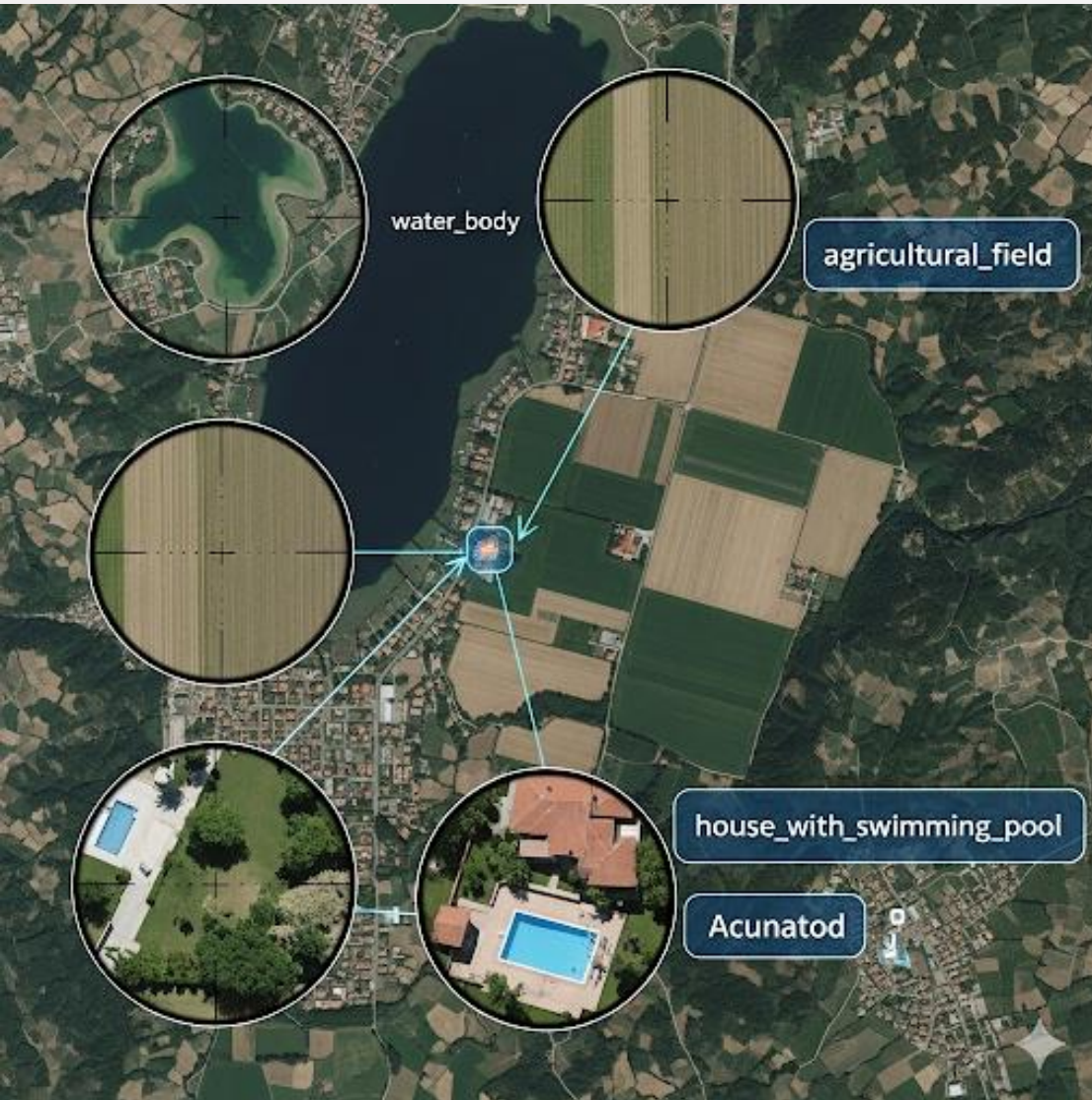


# Unified Metadata Catalog Key Properties

- **At a given scale, metadata becomes data**
  - Requires tool: Scientific Data Management
- **Ability to host a sufficiently large volume of information**
  - images annotation can be verbose (GeoJSON)
- **De-coupling of metadata catalog from data hub**
  - Ingest annotation from distributed processes across multiple datahubs
- **Mandatory to manage public/private data sources**
  - Queryable
- **Long-term support**
  - [sept. 2024] RUCIO selected by the SKA project



# Metadata Standard and Ontologies



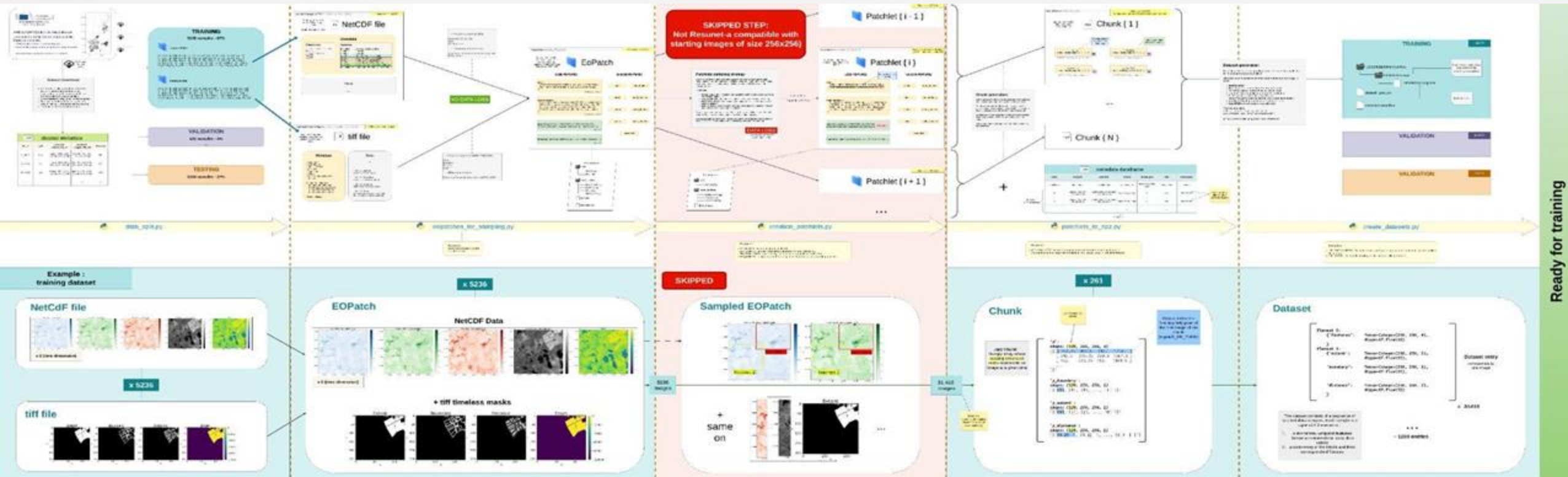
*metadata format follow standard*

- *GeoJSON*
- *STAC structure*

```
📦 catalog.json (Catalog)
├── 📁 sensors/ (Catalog)
│   ├── 📁 sentinel-2/ (Catalog)
│   │   ├── 📁 sentinel-2_l2a.json (Collection)
│   │   │   ├── 📁 2024/06/ (Catalog partition)
│   │   │   │   └── 📄 S2A_20240611...json (Item)
│   │   │   └── 📄 sentinel-2_l1c.json ...
│   └── 📁 landsat-9/...
└── 📁 derived/...
```

- Clients follow **self** → **child** → **item** to crawl downward.
- Every object also offers **root** to jump straight to the top.

# AI Pipeline: Queryable data and data gravity



On-going effort of multistage characterization *Workflow Roofline Model* to determine *the best data movement / computational efficiency tradeoff*

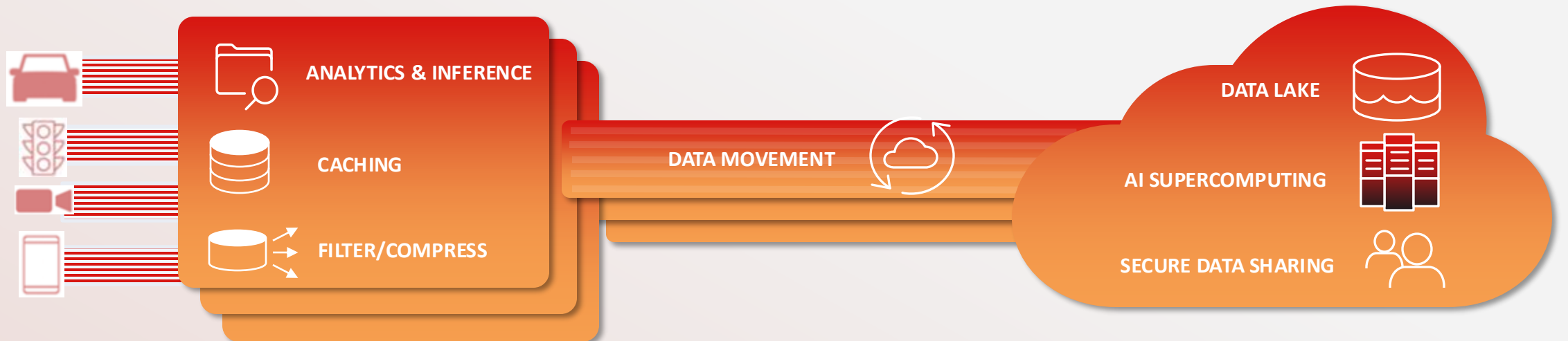
# Where is heading the Industry? DDN Infinia

Reduce Egress Costs by 10X

Free Up Infrastructure

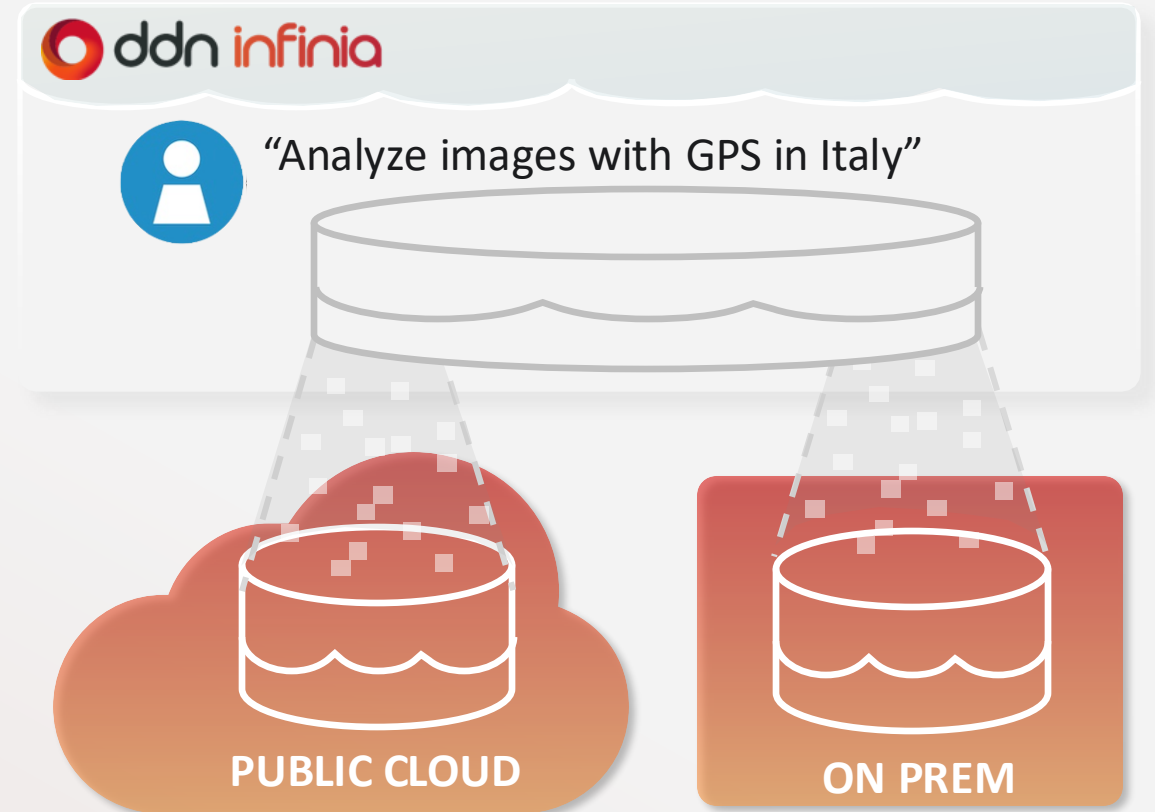
Simplify Distributed Data

- Distributed metadata and exposes it as a native Data Service
- Monitor Data at remote locations, only moving the data that are really needed to be process locally
- Subscription mechanism to remote metadata catalog



# Accelerate AI Outcomes with Global Data Visibility

- Query billions of files instantly — **without waiting for data**
- **Train/inference faster** with precise, policy-driven data access across cloud, edge, and core
- **Reduce Egress fees** and infrastructure waste by activating only the data that matters



# Data Lakehouse: Queryable metadata

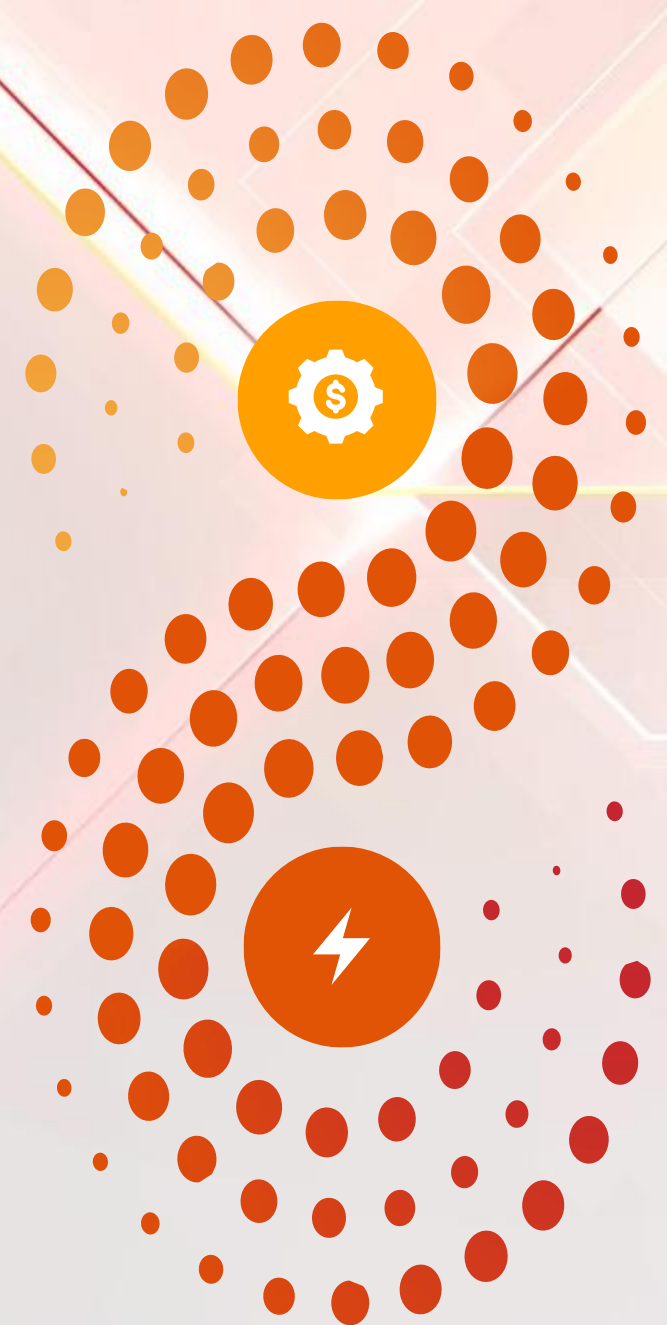
- Faster Time to Insight sub-ms response times on massive datasets
- Unified Access Across Data Types: SQL, Parquet, Object, File
- Seamless Scalability Linear scaling billions of rows
- Simplified Data Management & Superior Cost Efficiency Through License and Admin Reduction

```
Infinia@Infinia
-- Find top-performing models trained in the last 30 days
SELECT model_name, accuracy, training_time
FROM ai_models
WHERE training_date >= CURRENT_DATE - INTERVAL '30 days'
ORDER BY accuracy DESC
LIMIT 5;
```

# AI is reshaping the HPC storage industry

- **Metadata as first-class citizen**
- **Closer to end-user semantic**
- **Interoperability and programmability**
- **Service at scale**
- **From storage to Data platform**

*Are we converging?*





# AI-Powered Industries Run Faster with DDN