A Modern Data Platform to improve Workflows and Pipelines

DAMA, Tecnopolo di Bologna
19 September 2025

A Modern Data Platform to improve Workflows and Pipelines

Part 1 – MeteoHub: The Weather Data Platform

Giuseppe Trotta – Focus on Application and Functionality

Part 2 – The Vast Data Platform

Sven Breuner – Capabilities and Performance

INTRODUCTION & GOALS

CONTEXT

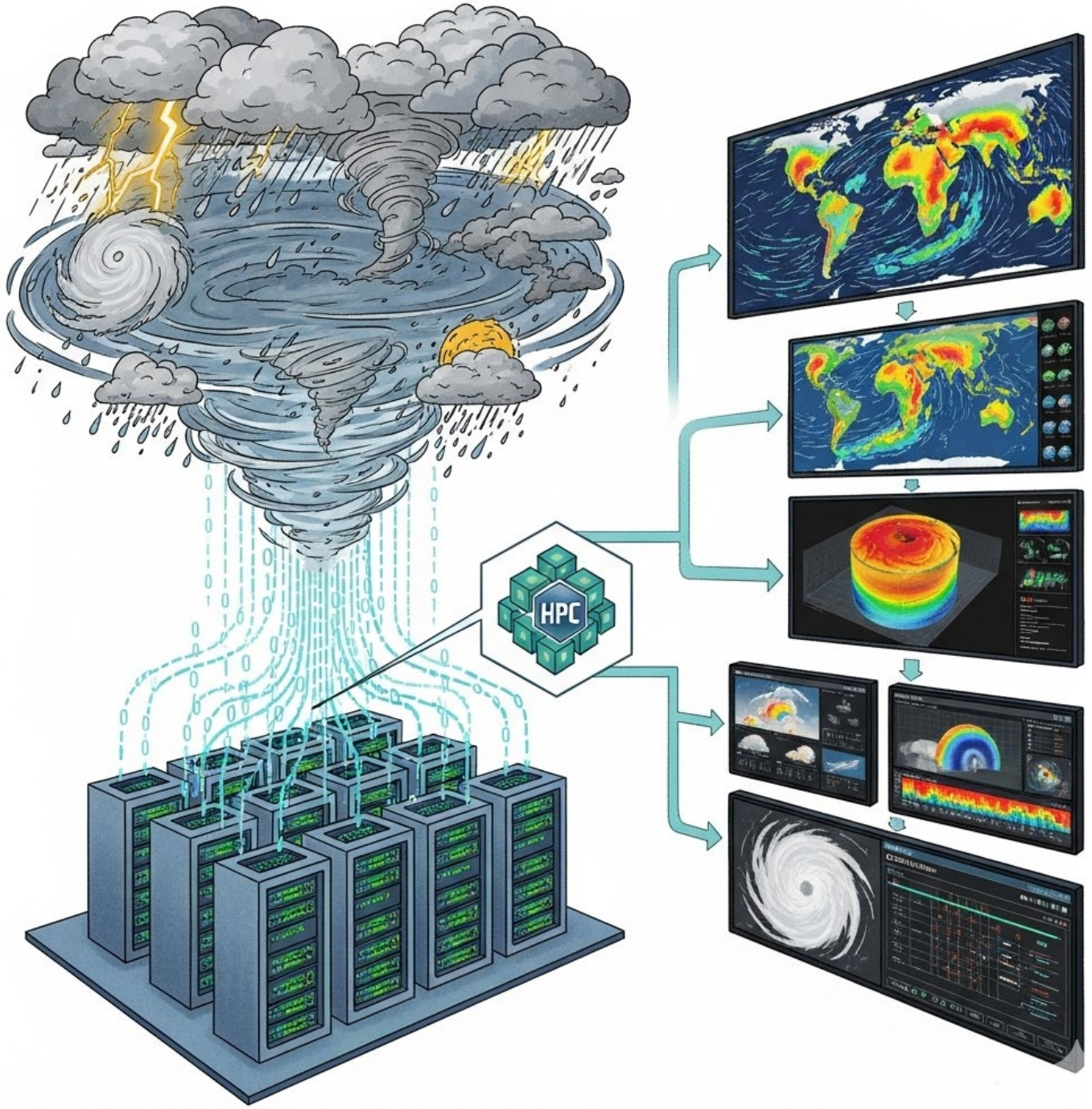
Meteorology, plus climate and
marine datasets

CURRENT

Data workflow using HPC filesystem for GRIB2 datasets with access limitations

GOAL

Improve the platform for scalable, high-performance handling of large datasets

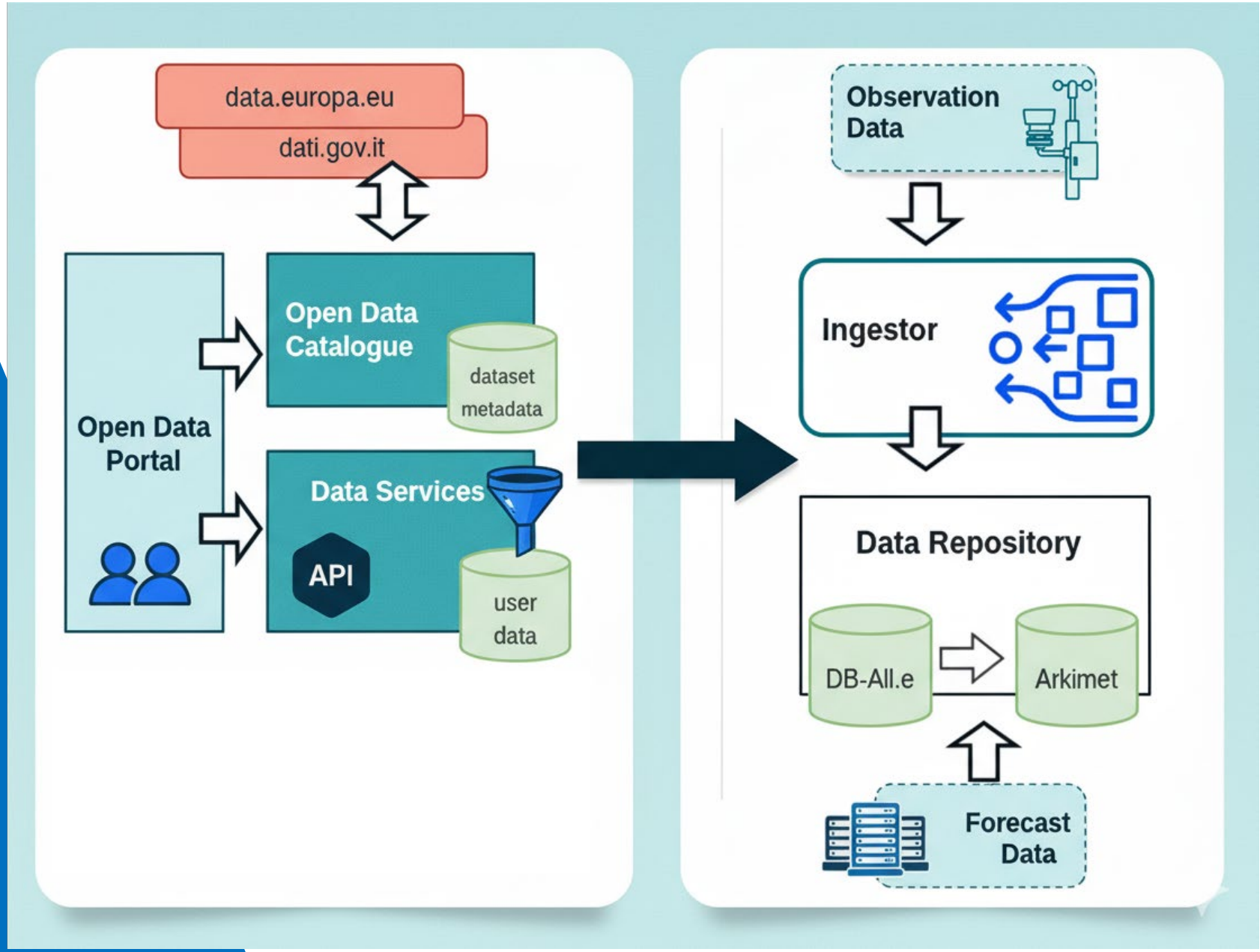


METEOHUB

Central hub for collection & distribution of observational and forecast data

Supports workflows for meteorology & climatology

Built on HPC and Cloud infrastructure

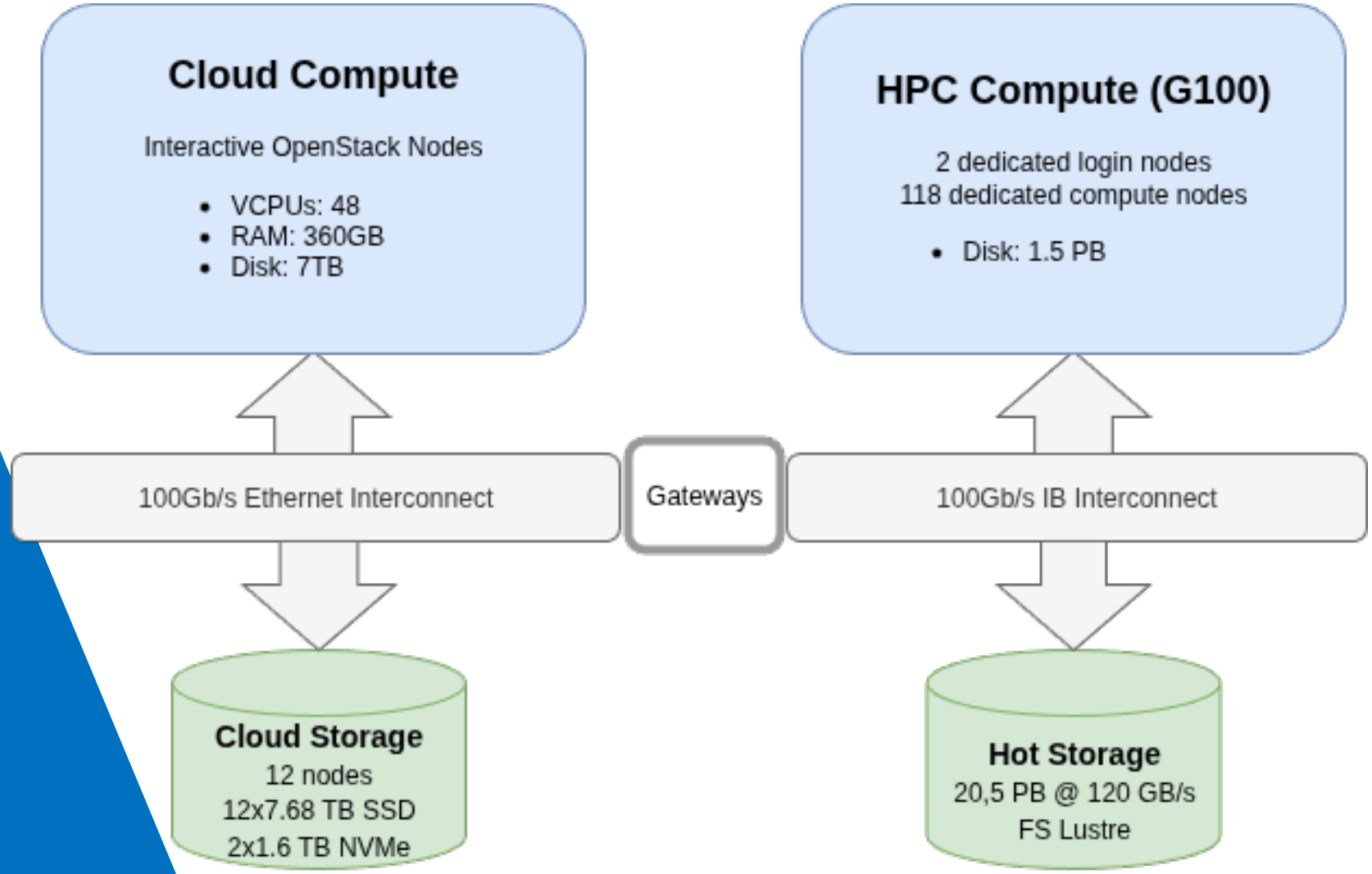


INFRASTRUCTURE DATA EXPOSURE

Forecast datasets stored on HPC parallel file system (Lustre)

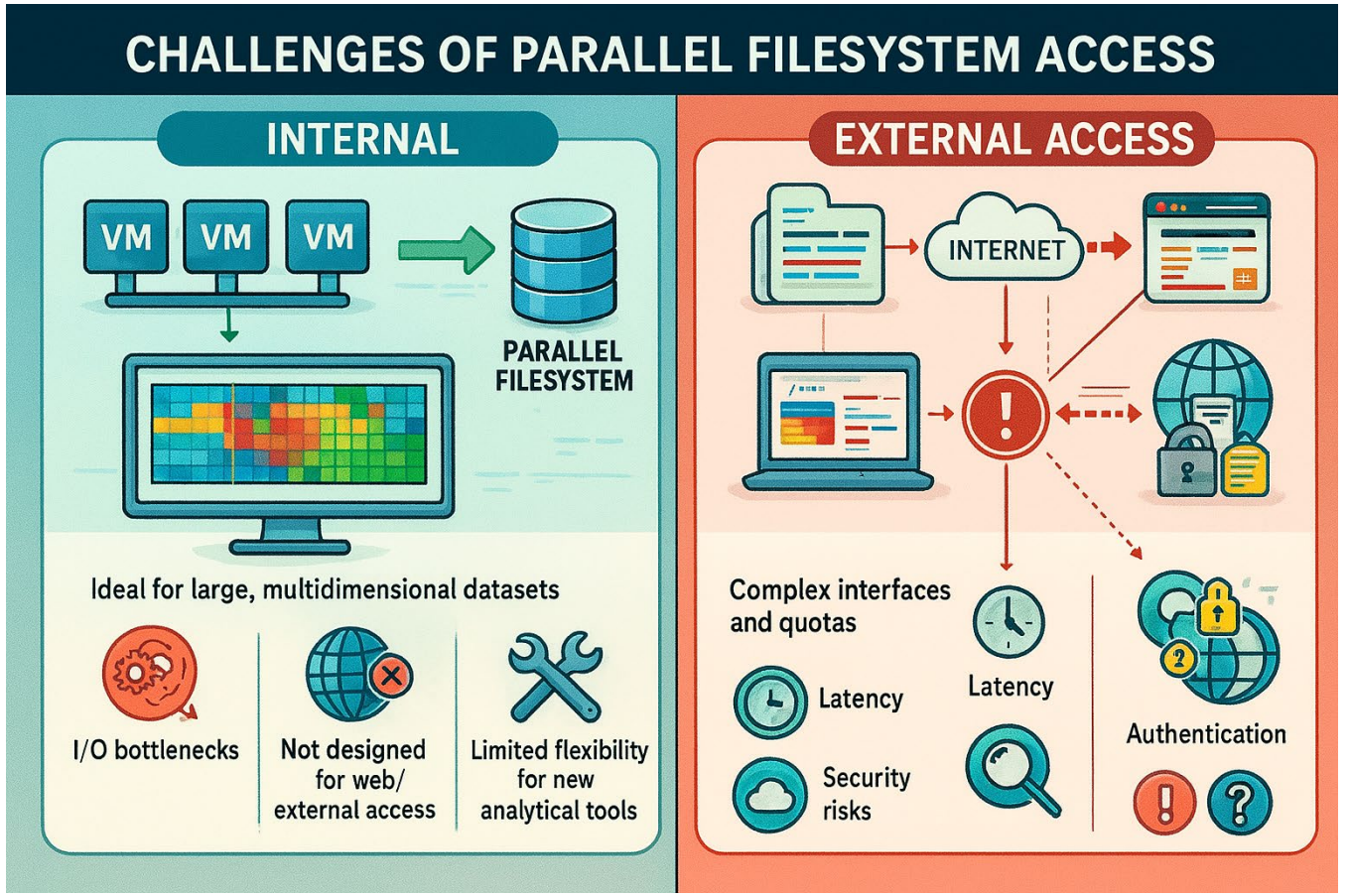
Accessible internally, but not optimized for external exposure

Challenges: scalability, flexibility, external interoperability



Challenges of Parallel Filesystem Access

- I/O bottlenecks for large multidimensional datasets
- Not designed for web/external access(complex interfaces)
- Limited flexibility for new analytical tools



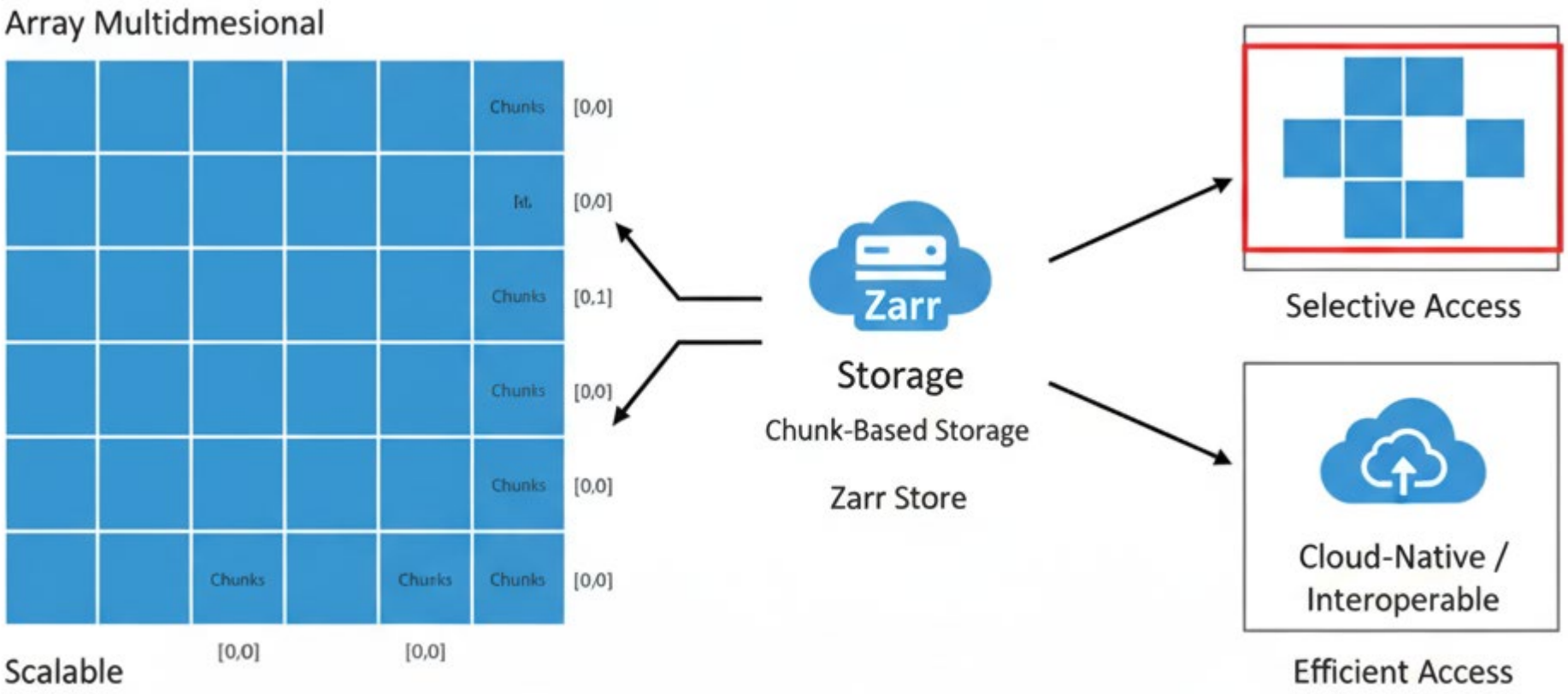
Zarr+S3-like Storage as an Enabler

- Object storage better suited for distribution & external access
- Scalable, multi-protocol, cloud-compatible
- Opens doors to new workflows & tools

Aspect	HPC Lustre	Zarr + S3
Data format	GRIB2 monolithic files	Chunked, compressed arrays
Scalability	Limited parallel I/O	Highly scalable
Access	Sequential/parallel jobs	Partial reads, parallel queries
Performance	I/O bottlenecks	Reduced I/O, faster analysis
Integration	Complex scripts	xarray, pandas, dask

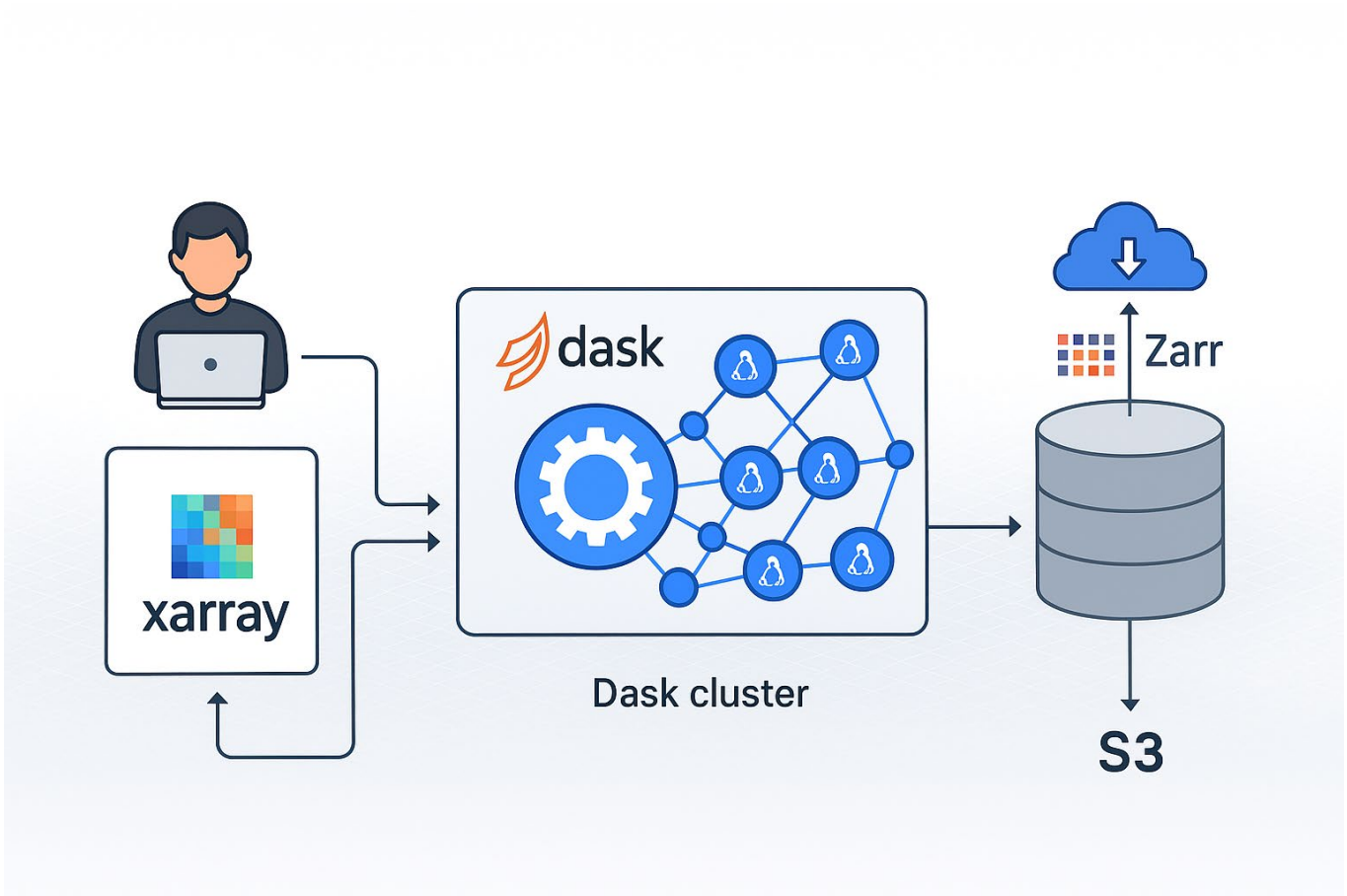
Why Zarr?

- **Scalable**: chunk-based storage for big arrays
- **Efficient**: partial access without reading full datasets
- **Flexible**: supports cloud-native workflows
- **Interoperable**: works with existing tools



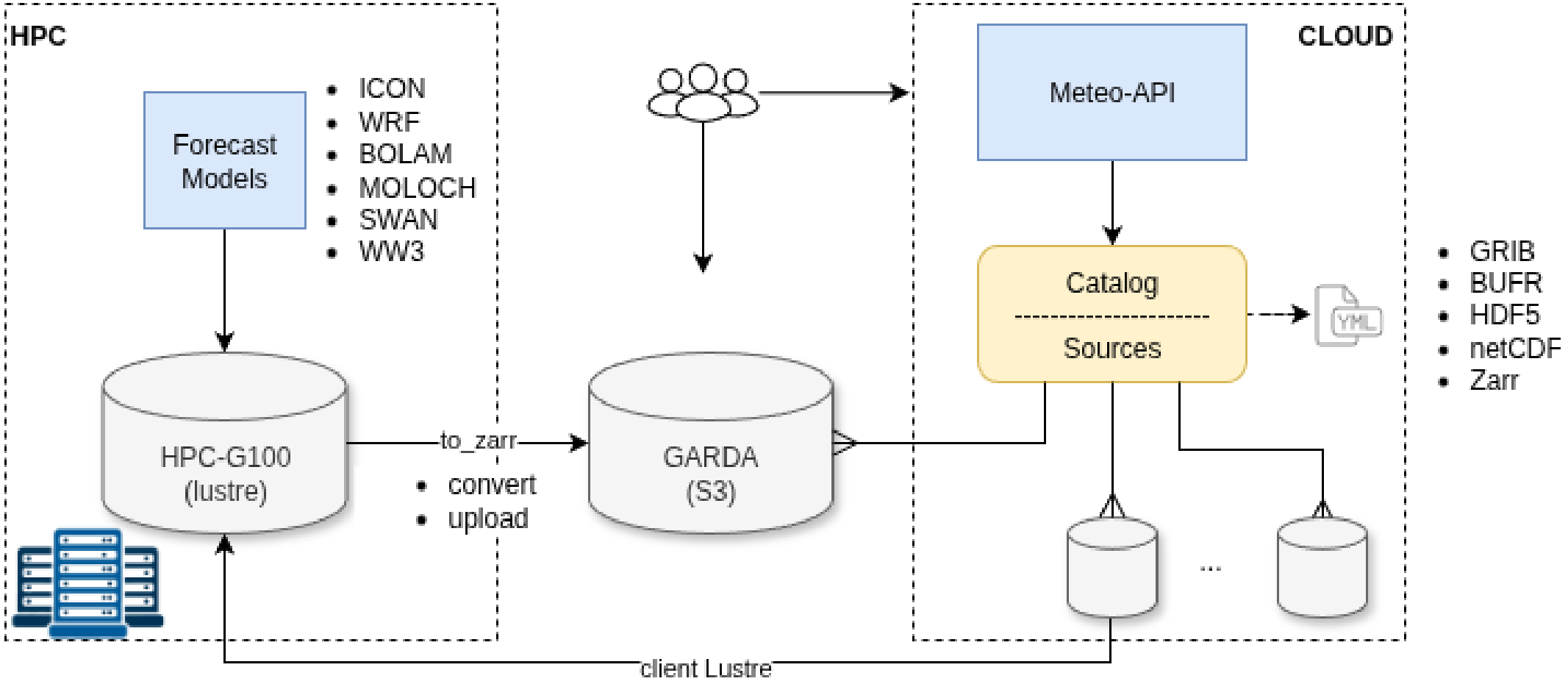
Tools for Data Access and Analysis

- **Xarray**: simplified access to multidimensional data (NetCDF, Zarr)
- **Dask**: parallel & distributed analysis for large datasets
- Supports scalable scientific pipelines



New Data Pipeline with Zarr and NetCDF

- **NetCDF**: standard for scientific data, widely used
- **Zarr**: chunked, compressed, cloud-native array format
- **GRIB2** → can be converted into Zarr without loss of data/quality



Summary & Transition

- **Current**: Meteohub with HPC file systems → limited for external access
- **Future**: **S3**-like backend + **Zarr** format = scalable, cloud-ready workflows
- Libraries (**Xarray**, **Dask**) unlock parallel analysis and new capabilities

Next step: Vast Data can provide these capabilities and extend them further

From Data to Insights

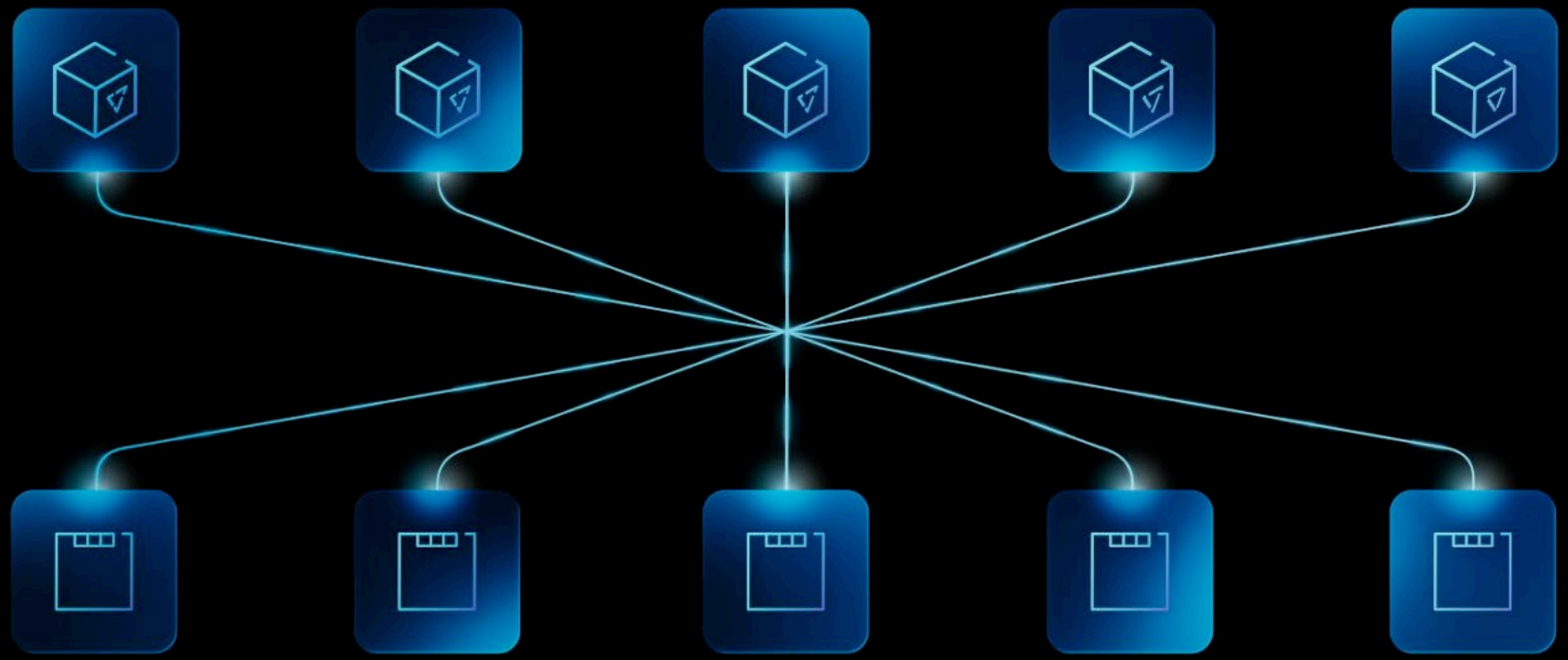
The VAST Data Platform

Sven Breuner
Field CTO International
sven@vastdata.com

A New Parallel Architecture for the AI Era

Solving For Challenges of Scale, Simplicity, Resilience and Efficiency

VAST Service Nodes
Stateless Containers



VAST Data Nodes
Exabyte-Scale Flash

DASE: VAST's Disaggregated, Shared-Everything Architecture

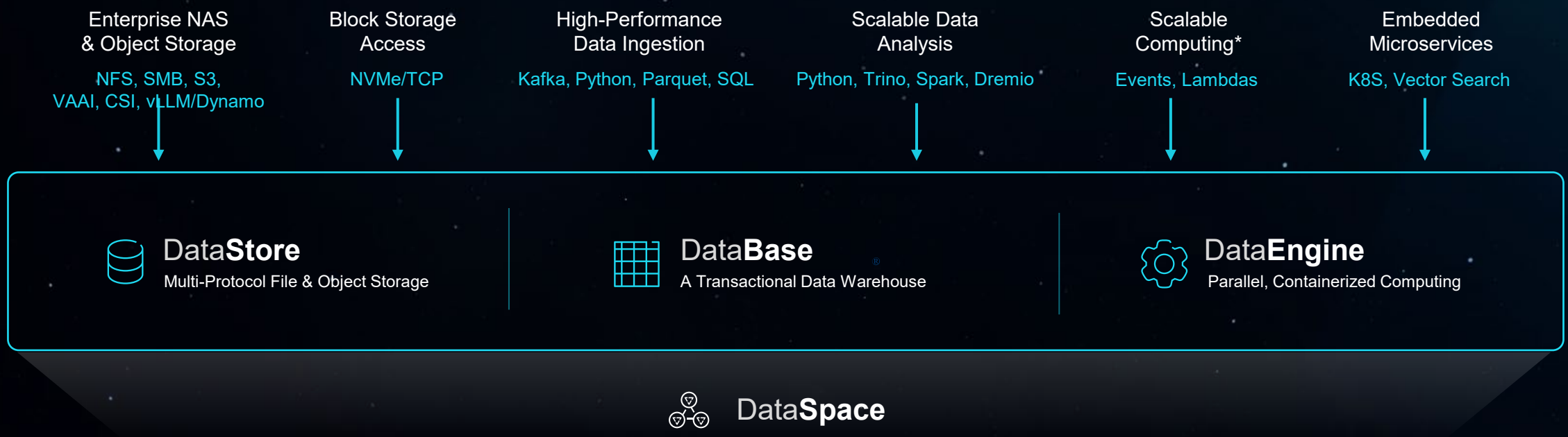
The VAST Data Platform

One Platform for All the Data


VAST
Data Platform
DataStore
DataEngine
DataBase
DataSpace

VAST Data Platform

Multi-Tenant, Zero-Trust, All-Flash



Hewlett Packard Enterprise · DELL Technologies · Supermicro · cisco · aws · Azure * · Google Cloud · EQUINIX · CoreWeave · core42 · Lambda · and more…

VAST SyncEngine: Your Data Bridge to AI

Universal Catalog

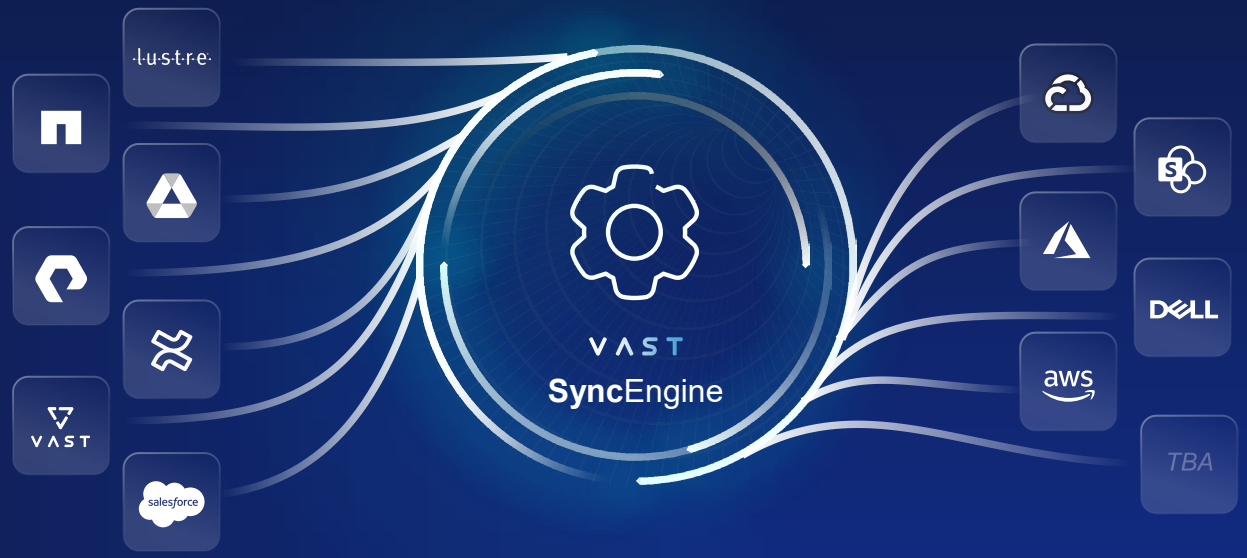
See and search across all enterprise data with the industries most scalable index

Powerful Migration

Move data effortlessly between data sources and the VAST AI OS

Robust ETL

discover, transform and prepare data for analytics and AI workloads



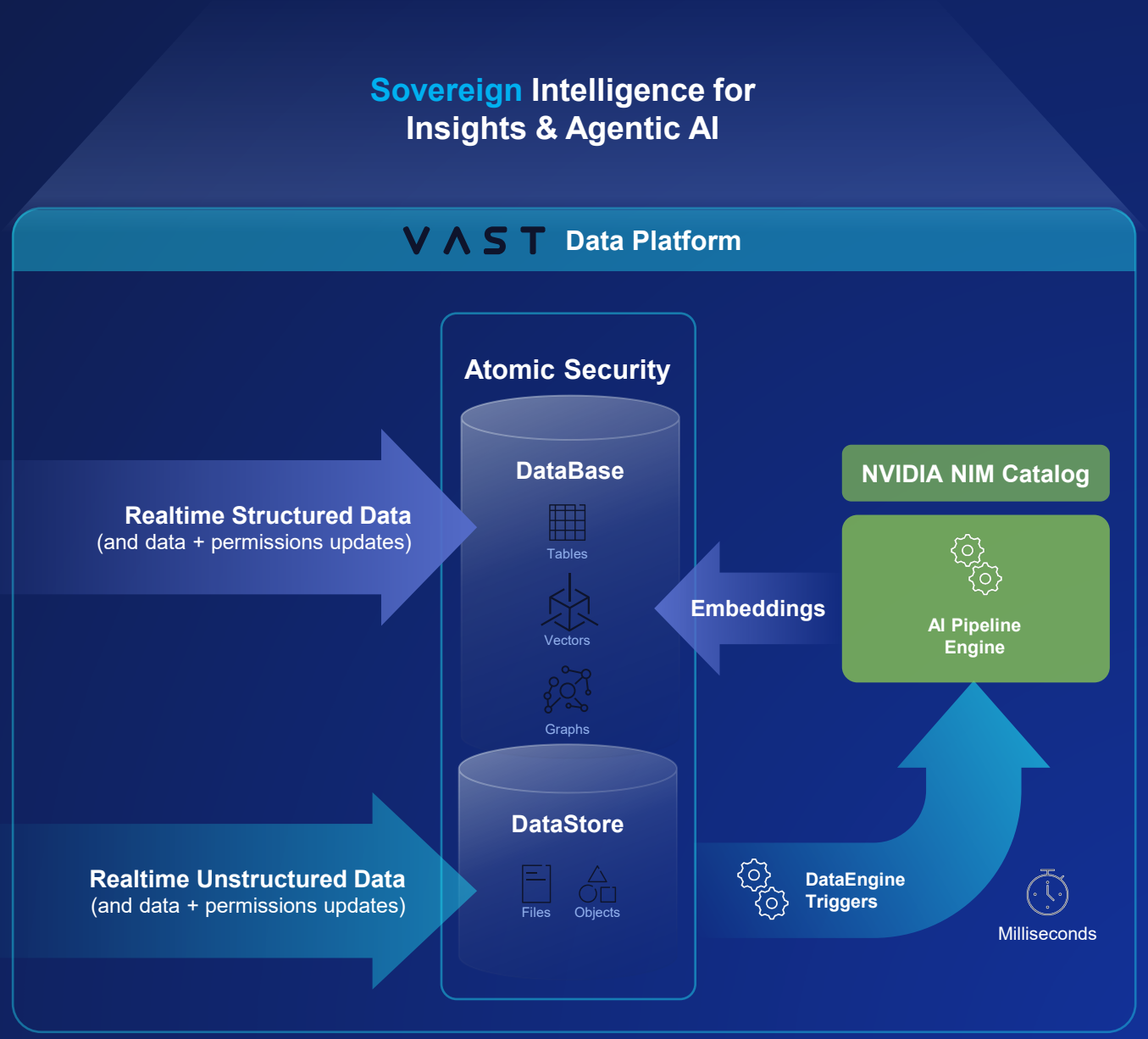
See, search and mobilize your entire data estate


NEW
VAST
InsightEngine
WITH NVIDIA

VAST InsightEngine

Secure & Real-Time Enterprise RAG

- Instantly get semantic understanding of any enterprise data (docs, images, videos)
- “No-code” retrieval augmented generation with enterprise security (RBAC) end-to-end:
  - Data ingestion handling in real-time
  - Embedding & indexing of unstructured data
  - Search and retrieval APIs for AI/agentic apps
- **Powered by NVIDIA AI Enterprise**



CINECA

Grazie