POIR.04.02.00-00-D010/20

PCSS

Digital Transformation and the role of HPC, Data and Cloud

Norbert Meyer

Fundusze Europejskie Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska Europejski Fundusz Rozwoju Regionalnego

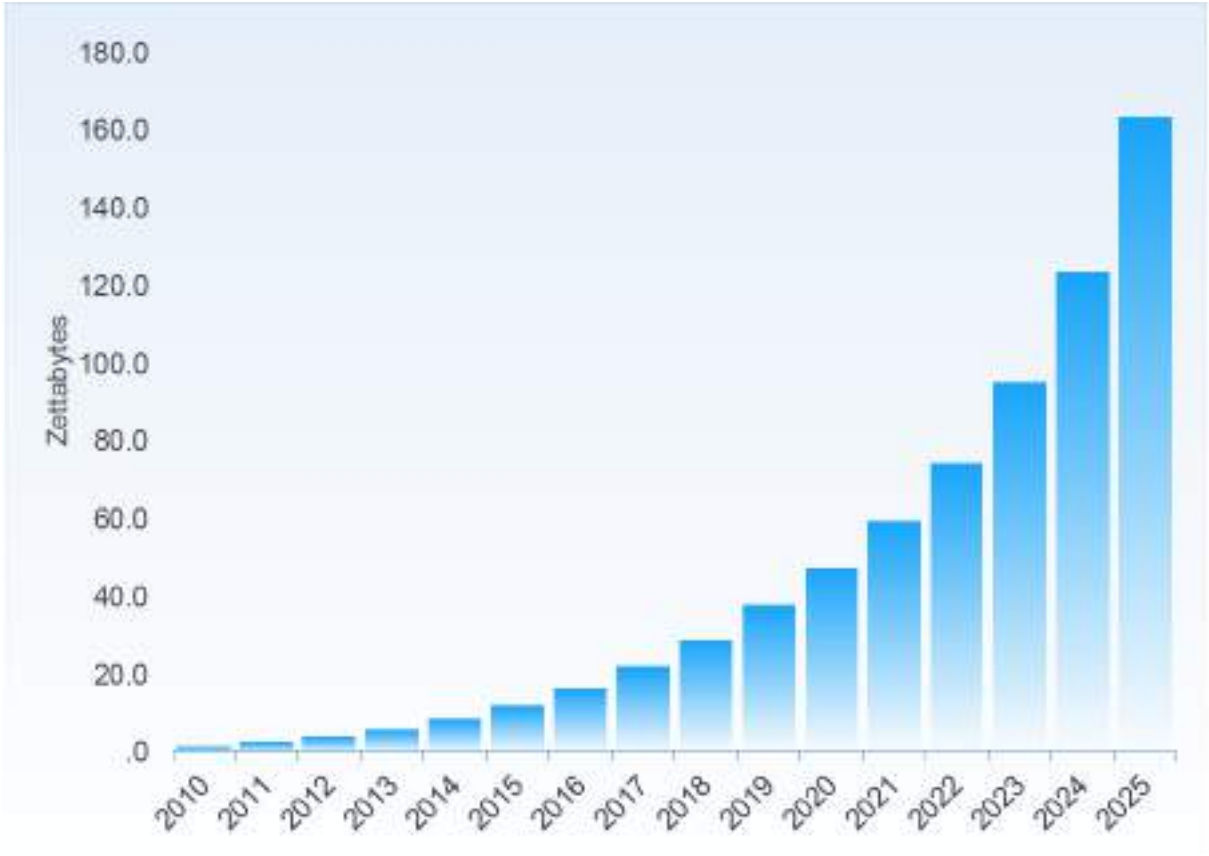
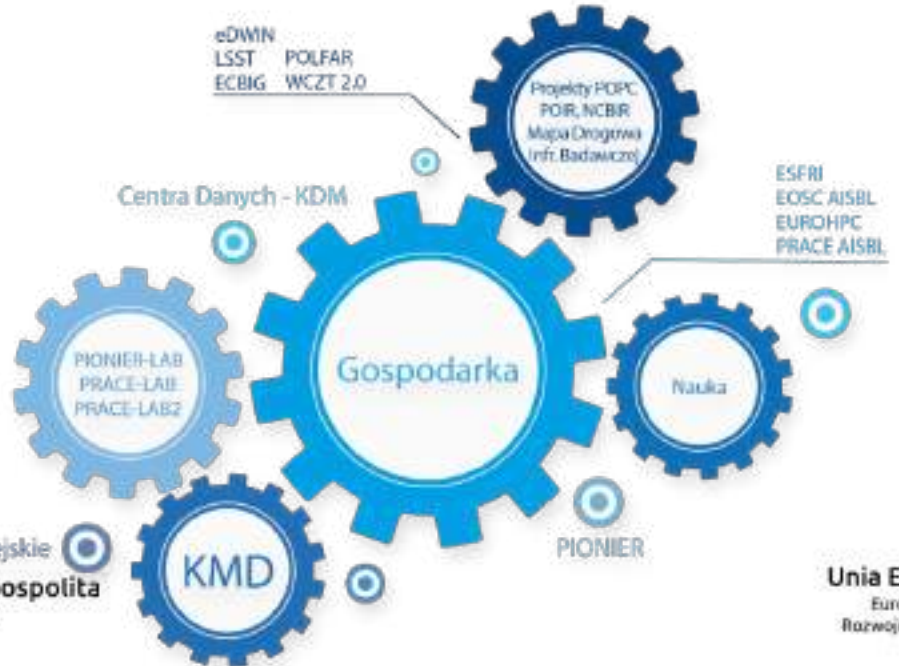


Figure 2: Worldwide global DataSphere creation and replication, 2010-2025, Source: IDC, 2021

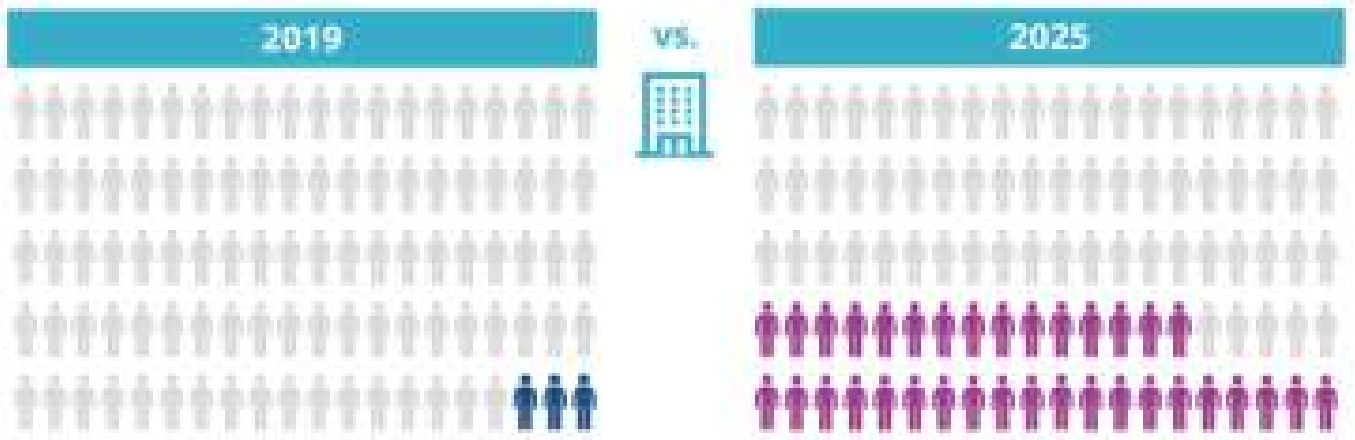
Trends 2025

170+ Zeta Bytes DATA

10+ Exascale HPC

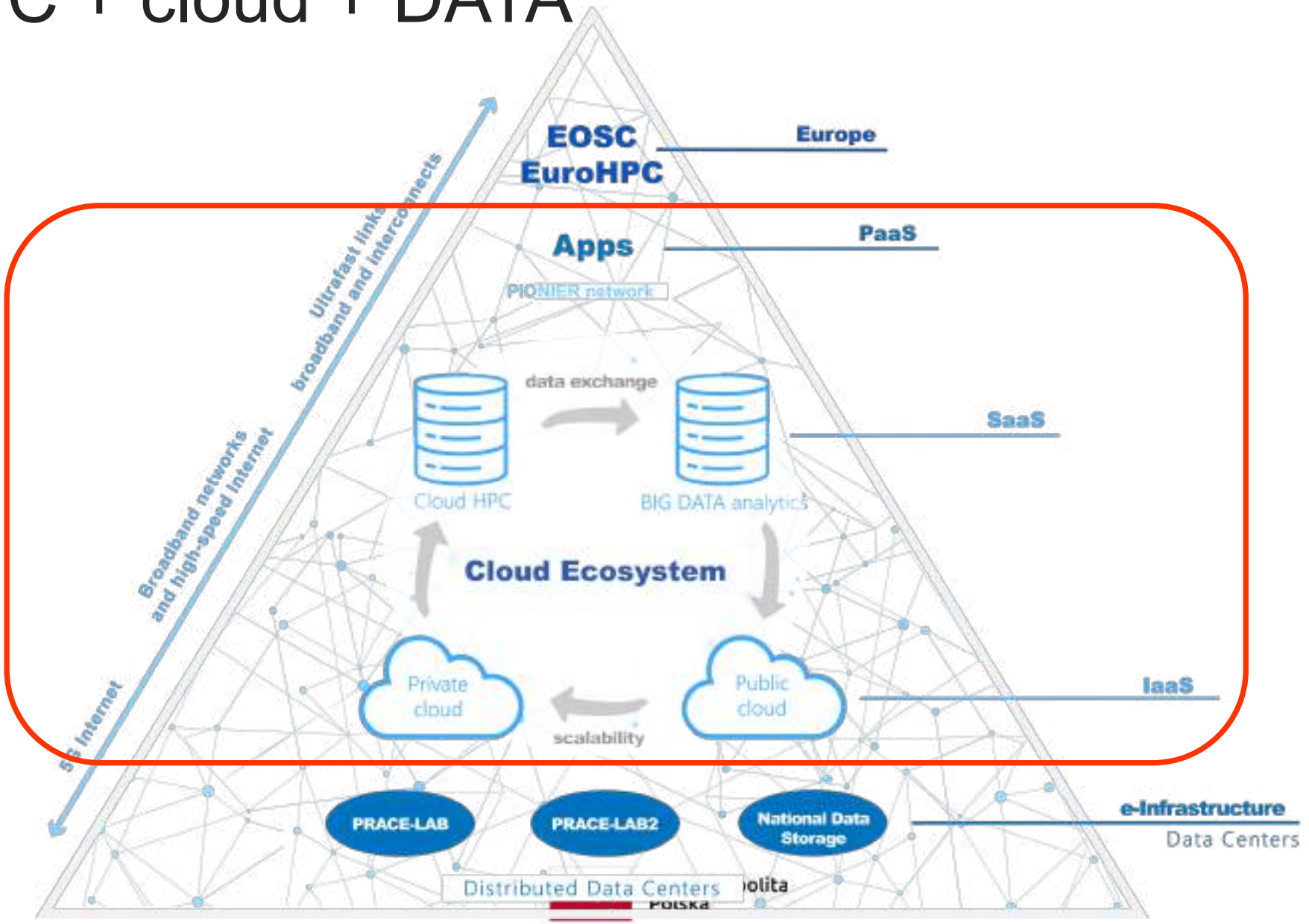


back to offices



Source: IDC, 2021

HPC + cloud + DATA



Data HPC Cloud

- **HPC, AI/ML: data processing**
  - **Big Data** for analysis, simulations, AI, deep-learning, …
    - New challenges e.g. cybersecurity
  - While computing data kept in **storage tightly integrated with computing**

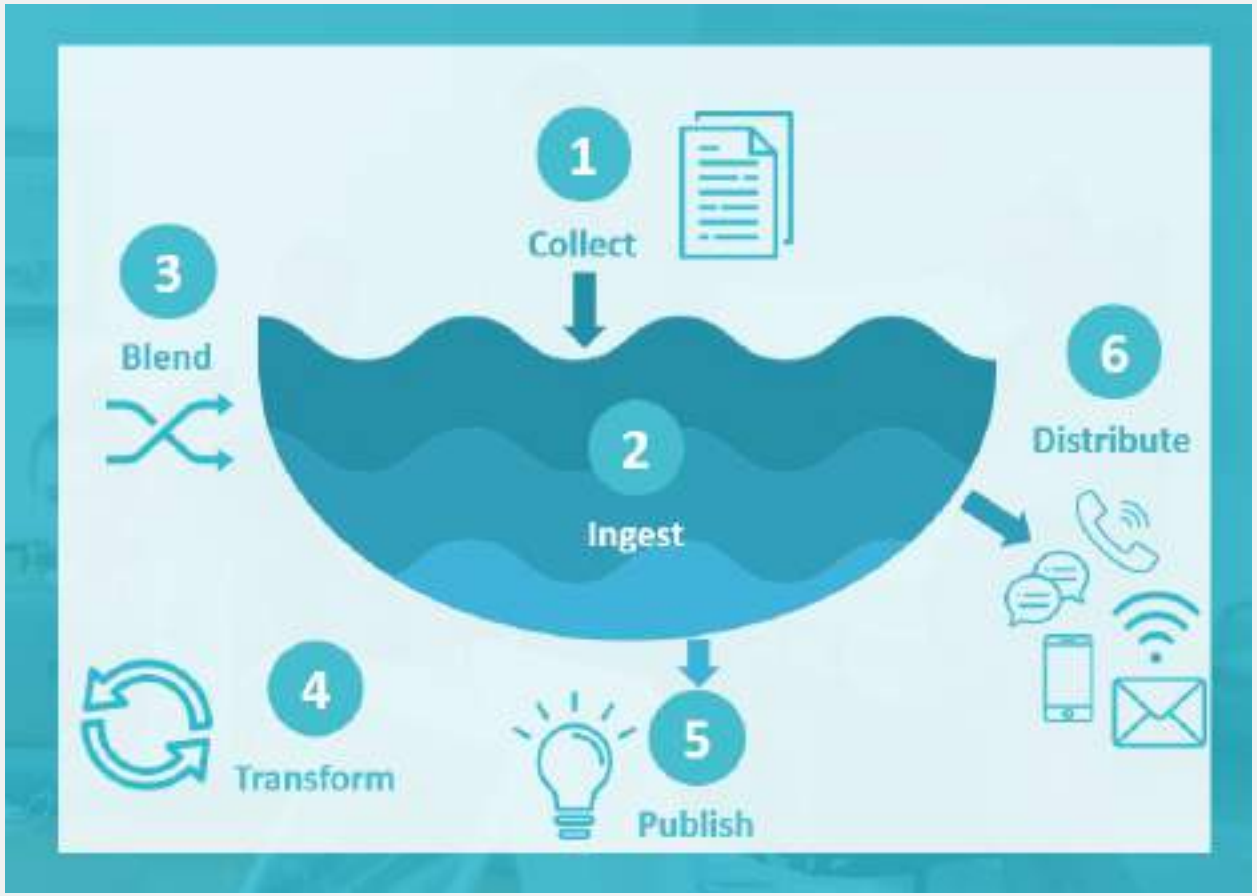
- **Before computing: data preparation**
  - Data has to be acquired, collected, cleaned, enriched, curated…
  - In order to be ready for exploration, exploitation, analysis and usage in computations

- **After computing: data protection, long-term storage & access:**
  - Source **/ RAW data** – to be protected beyond the computing / data lifetime
  - Data Access – must be ensured by supporting (evolving) standards
  - **Re-use**: discoverability, refferability, accessibility – based on meta-data
  - **Migration support**: among infrastructures, for users: opt-in & opt-out

National Data Storage = Data Lake

Ensuring:

- **Capacity (PBs)**
- **Performance (10s GB/s)**
- **Scalability (distribution)**
- **Data access**:
  - over time (persistency)
  - across protocols (translation)
- **Data re-use**:
  - Discoverability
  - Search'ability
- **Extendibility:**
  - Functions & apps
  - **Interfaces vs compute**

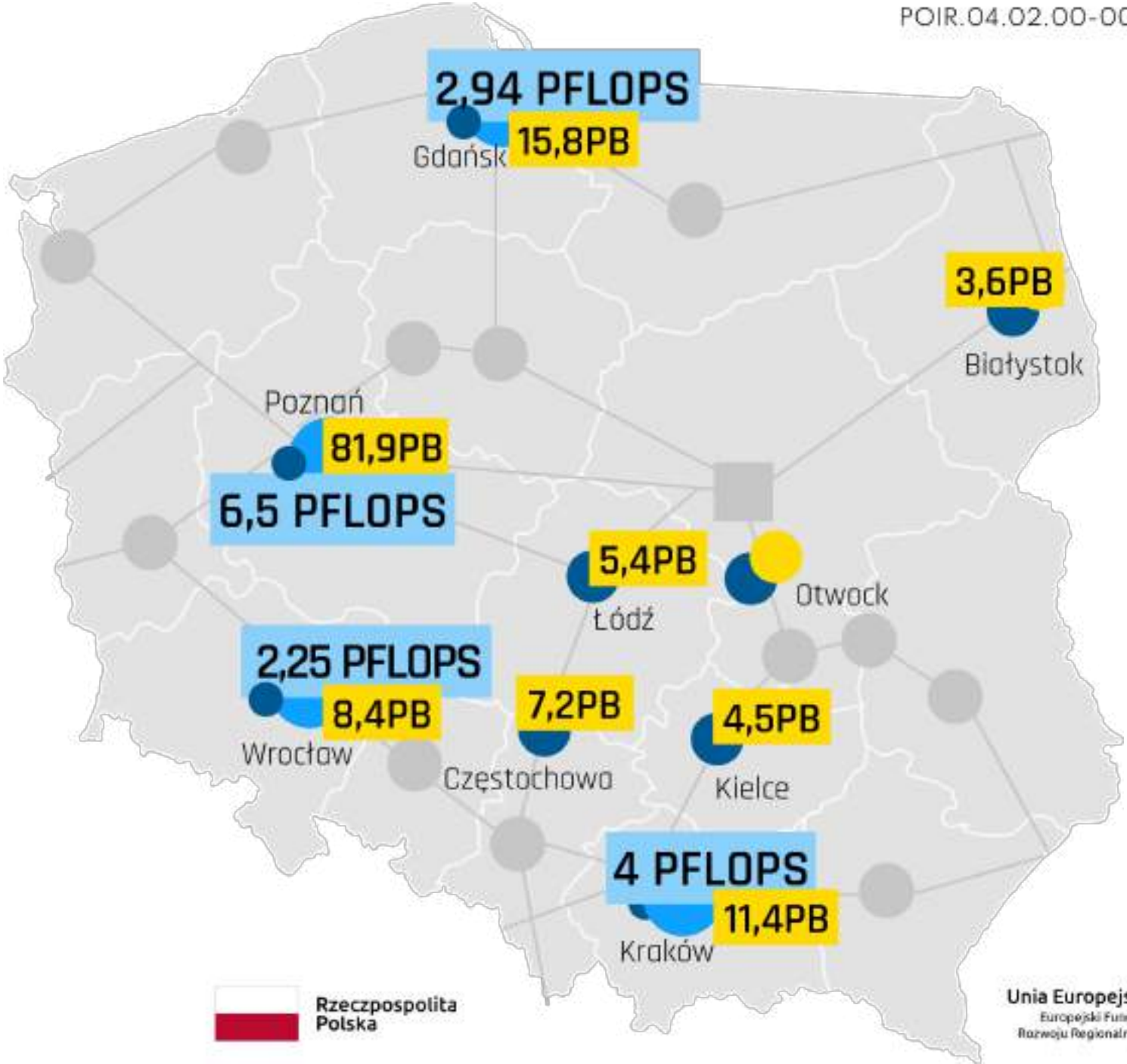


Source: https://www.ecloudvalley.com/what-is-datalake-and-datawarehouse/

POIR.04.02.00-00-D010/20

HPC/Cloud

pracelab



Fundusze Europejskie
Inteligentny Rozwój

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

National Data Storage

A complete environment for data-driven science:

Data preparation & preservation:
- Data acquisition, cleaning, enrichment
- Discovery & search
- Sharing & publishing
- Data re-use, discovery and exploitation
- Data protection / preservation

Data processing & analytics:
- High performance computing
- Data Analysis, Data Science…
- Machine Learning, AI

… in synergy with HPC computing projects in Poland

NDS project
infrastructure

Planned - overall:

250 PB tape (long-term storage/archive)

250 PB disk (on-line storage, data lake)

10 PB SSD/NVMe (access acceleration)

Target: **1 EB of storage capacity**

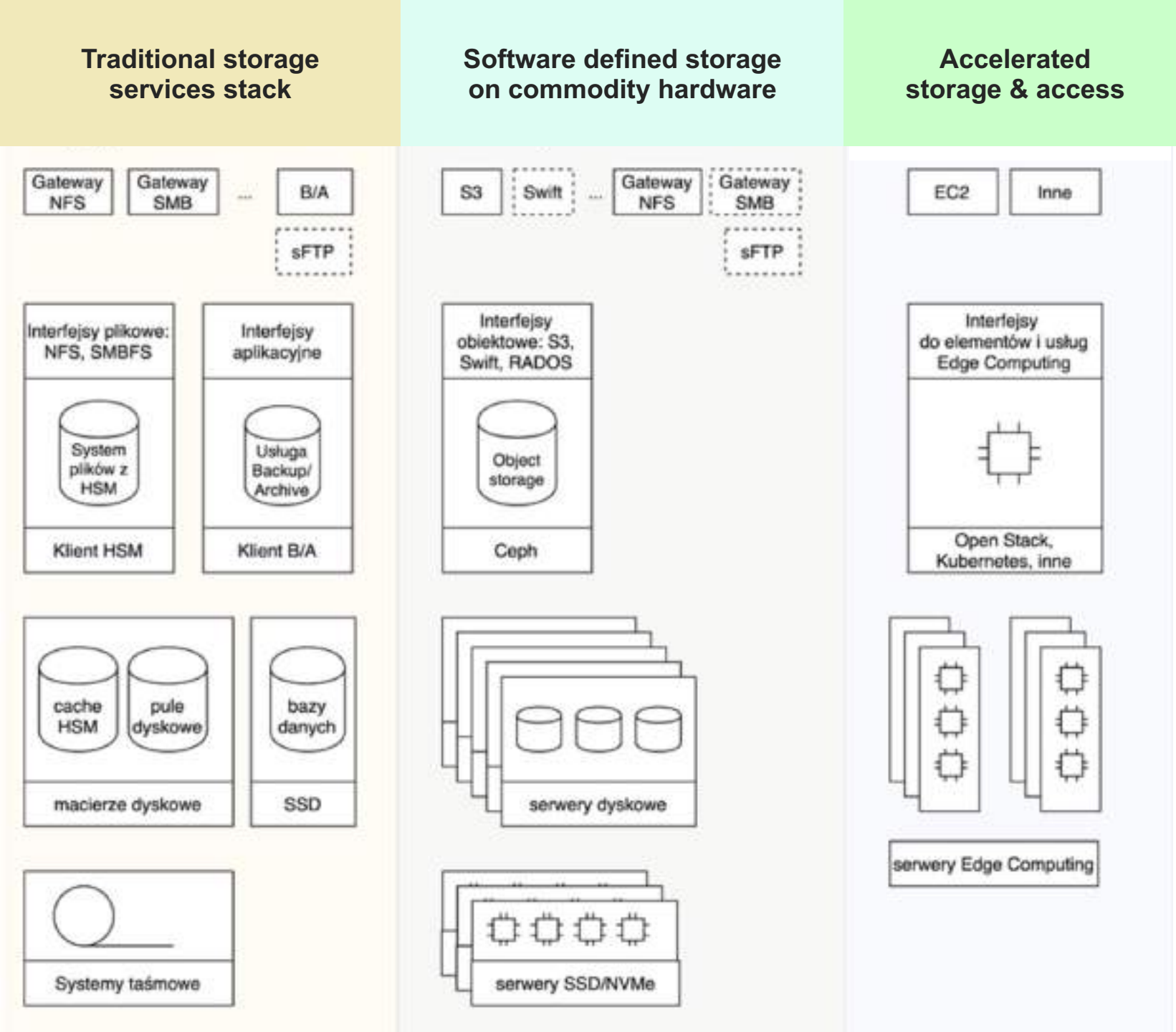
Partners: 5 HPC & 4 MAN sites:

HPC: PSNC, Cyfronet, TASK, WCSS, NCBJ

MAN: Białystok, Częstochowa, Łódź, Kielce

POIR.04.02.00-00-D010/20

NDS data services stacks



‘Traditional’:

- tape libraries w/ disk cache
- filesystem-like access

‘Software defined storage’ on commodity hardware

- Mostly Ceph
- Possibly MinIO (for AI workflows)

‘Accelerated’ data storage & access

- NVMe-based arrays & servers,
- FPGA-based accelerators

Unia Europejska
Europejski Fundusz Rozwoju Regionalnego

NDS – use cases (1)

Radio-astronomy (LOFAR):

- **Long-term storage in PBs:** PSNC holds several PBs of data as the Long Term Archive for LOFAR node
- **Computing:** HPC computing PoC on-going @PSNC, plans to extend computing beyond SURF and FZJ

Challenges:

- Processing large datasets (100s of GBs)
- Data access from archive to compute: 10s of GB/s

NDS features:

- dCache-based repository embedded in data lake
- Tieried storage: tape, object (Ceph), ‘regular’ disk
- Fast data transfer through high-capacity links (Eth, IB)

NDS - use cases (2)

Polish UV Satellite System (UVSAT)
on the Roadmap for RIs in Poland:

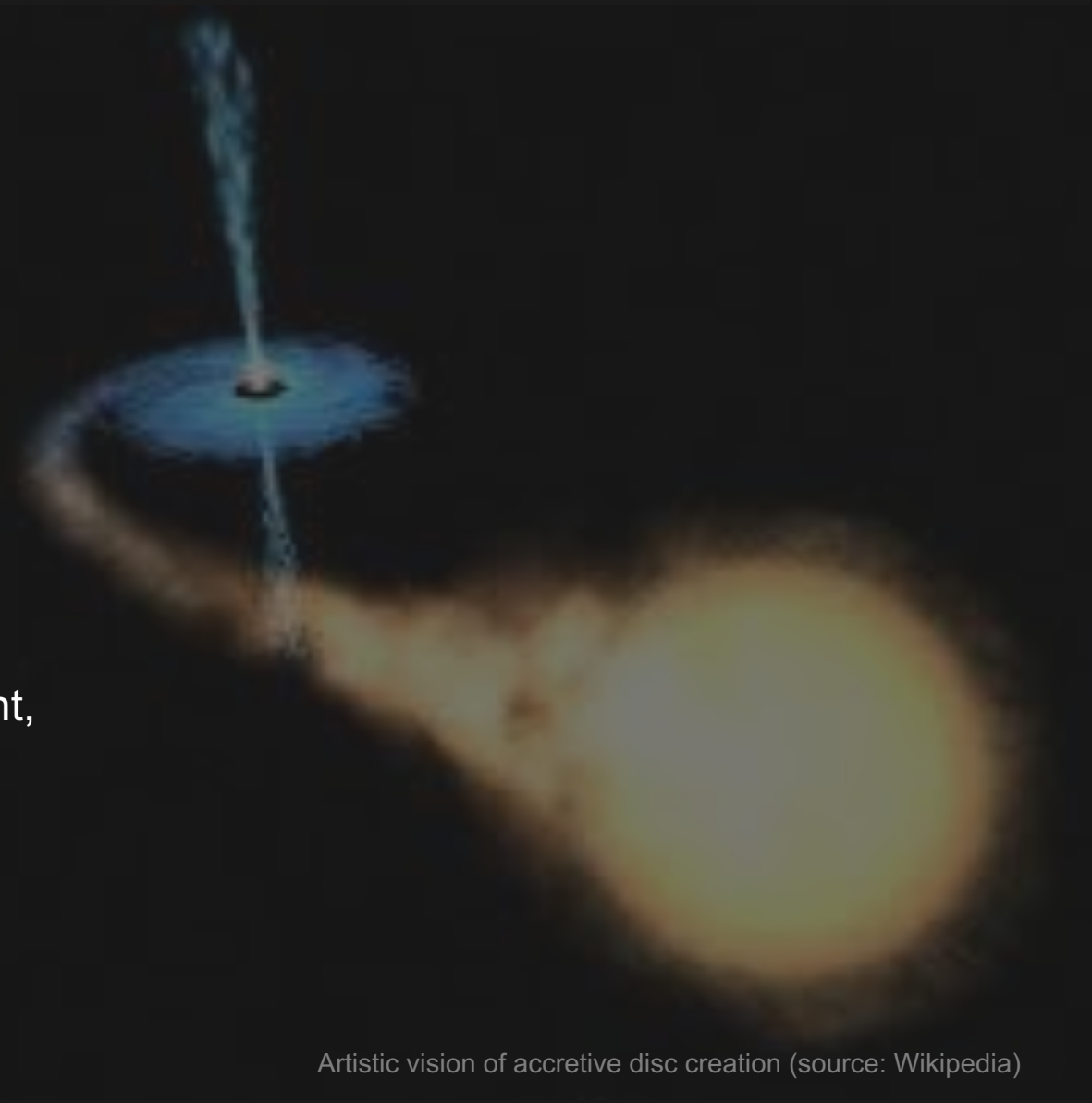
- Large datasets collection and storage (PBs):
- UVSat builds the data acquisition infrastructure with data buffers and data-centric processing workflow

Challenges:

- Gathering and protecting large datasets (100s of TBs)
- Integrating seamlessly: initial processing, format unification, meta-data normalisation & enrichment, enabling automation, visualisation & analysis

NDS features:

- Repository software to be embedded in the datalake
- Tight integration with computing infrastructure to enable HPC computing and data analytics
- Automation of initial data processing steps



Artistic vision of accretive disc creation (source: Wikipedia)

NDS - summary

NDS - complete environment for data-driven science:

- Implements a vision of ‘scientific data lake’
- Tightly integrated with computing resources
  - *HPC, cloud, Edge Computing*

Services

Składowanie danych

Obiektowe

Plikowe

Hierarchiczne

Niezawodne

Repozytoria danych

EOSC - otwarty dostęp

Kompatybilny z FAIR

Przetwarzanie danych niestrukturalnych

Bezpieczeństwo danych

Szyfrowanie danych

Repliki geograficzne

Niezawodne składowanie z redundancją

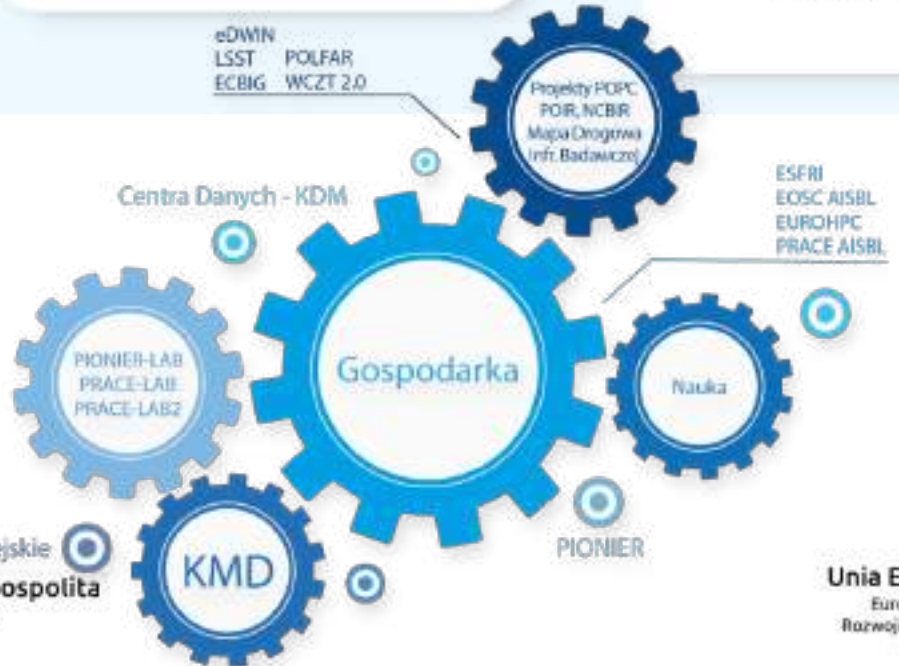
Analityka BIG DATA

Sztuczna inteligencja

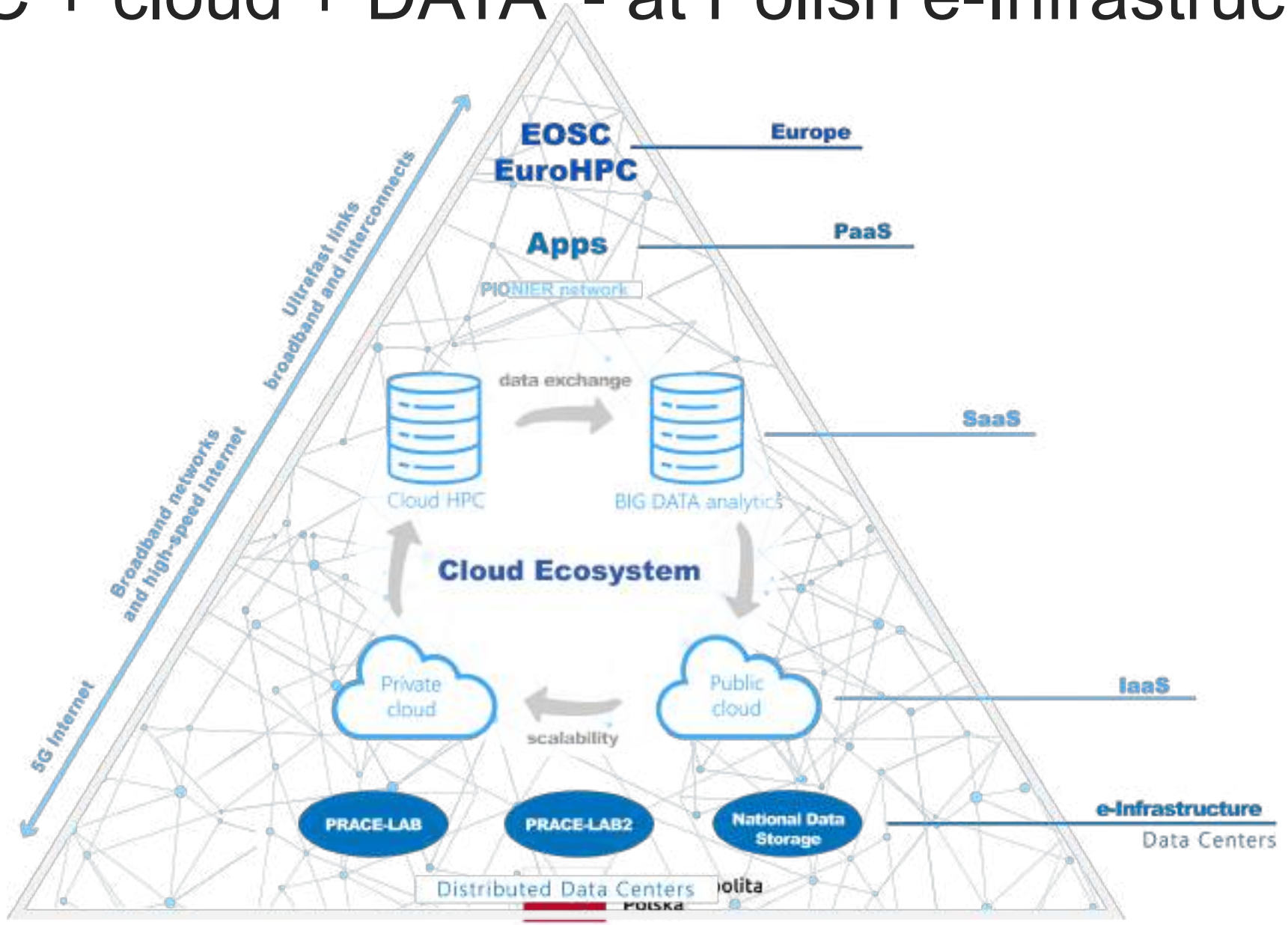
Model Data Lake

Środowiska HPC i chmura

Przetwarzanie brzegowe



HPC + cloud + DATA - at Polish e-Infrastructure



Commercial part

- **40% of infrastructure,**
- enterprises, SMEs, R&D, central and local government administration,
- Industry 4.0, automotive, security, power engineering, medicine, agriculture and bioinformatics, etc.,
- CFD and MES simulations, Big Data processing and analysis (including elements of AI), optimization of business and production processes based on sensory data (IoT) and support for designing and testing new and / or improved products and services.

Scientific part

- ***60% of infrastructure,***
- R&D at universities, institutes of the Polish Academy of Sciences and National Research Institutes,
- physics, computational biology and chemistry, bioengineering, nuclear physics, astrophysics, mathematics, climate change, humanities, etc..
- New methods of model optimization for selected AI / ML tools for different hardware architectures.

NDS supports EOSC

NDS is important piece of Polish inkind contribution to EOSC

NDS - National

Common services and NDS infrastructure
Access rule defined by FAIR

NDS Repositories

Sustainability policy at NDS

EOSC – European level

EOSC Core
EOSC Exchange, EOSC FAIR

EOSC Data
ESFRI, EuroHPC

EOSC Association

Poznańskie Centrum
Superkomputerowo-Sieciowe

Norbert Meyer

meyer@man.poznan.pl