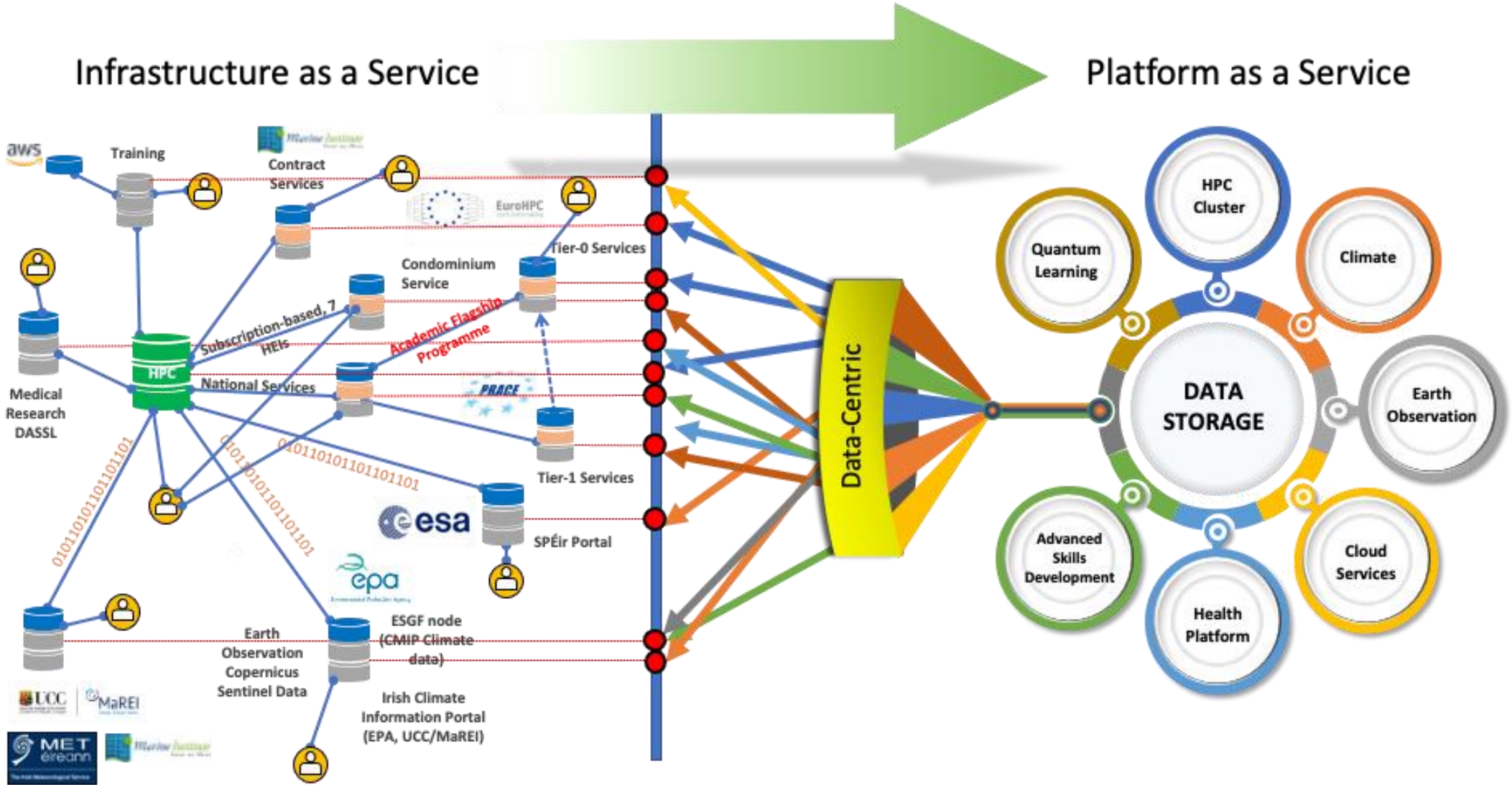
Application / workflow driven Infrastructure Design

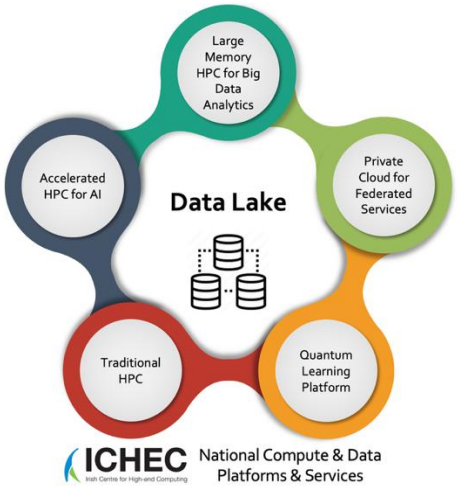
Venkatesh Kannan

Irish Centre for High-End Computing (ICHEC)

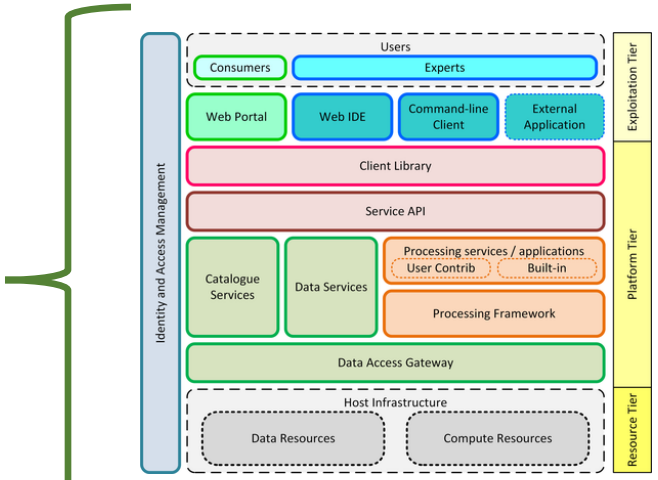
ICHEC Data & Compute Services



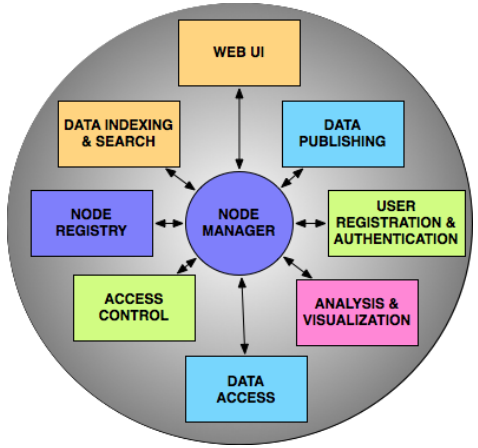
ICHEC Data & Compute Services



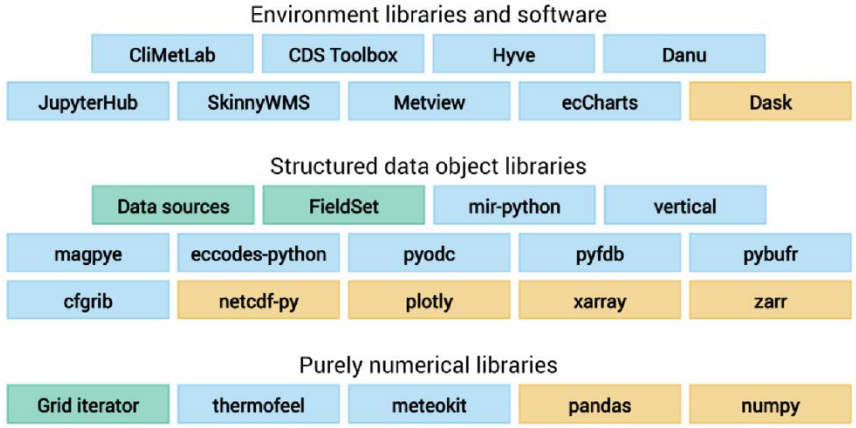
CASPIr



EO Data Cube



ESGF



ECMWF software roadmap

Operational Platforms & Services

ESGF Earth System Grid Federation

Copernicus Europe's eyes on Earth — sentinel open access hub >

Climate Ireland — Ireland's Climate Status Tool

CMIP The World Climate Research Programme's Coupled Model Intercomparison Project

SPÉIR Satellite Platform for Éire

marinerg-i

Climate Change Service climate.copernicus.eu

evdc esa

IRLDAT

Towards Data Spaces & Digital Twin Platforms

Dublin City Council Comhairle Cathrach Bhaile Átha Cliath — SMART DUBLIN

Smart City Digital Twin for Engagement

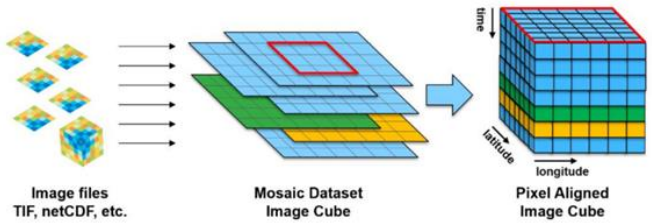
WILD ATLANTIC NATURE LIFE

Waters of LIFE

CCCENTRE Climate Biodiversity Water

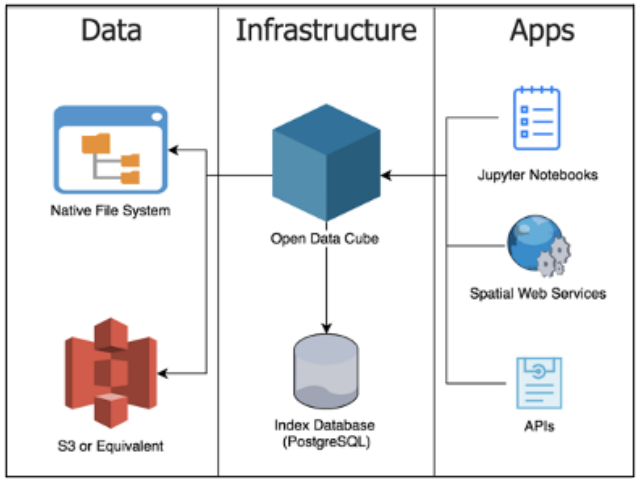
Special Area of Conservation (SAC)
Protect & enhance High-Status Waters
Results Based Payment Scheme (RBPS)

Earth Observation Data Cube

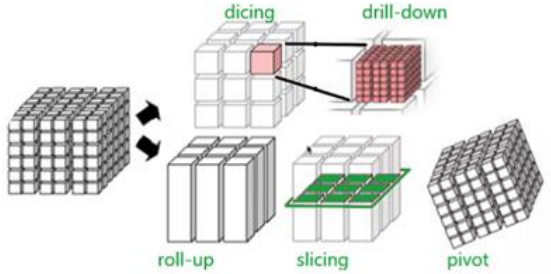


Prepare files for data cube

Run workflows continuously adding updates to data cube



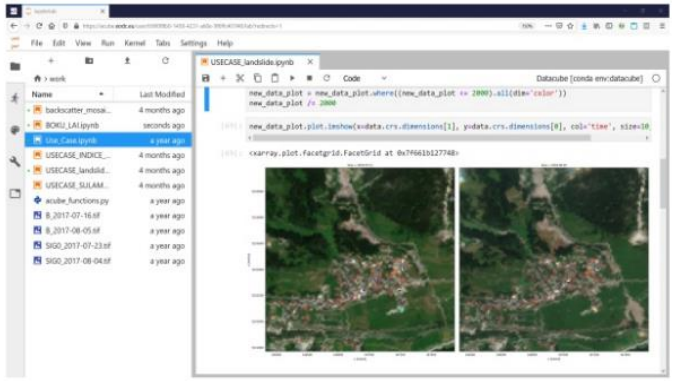
Data served for HPC-enabled analytics / ML



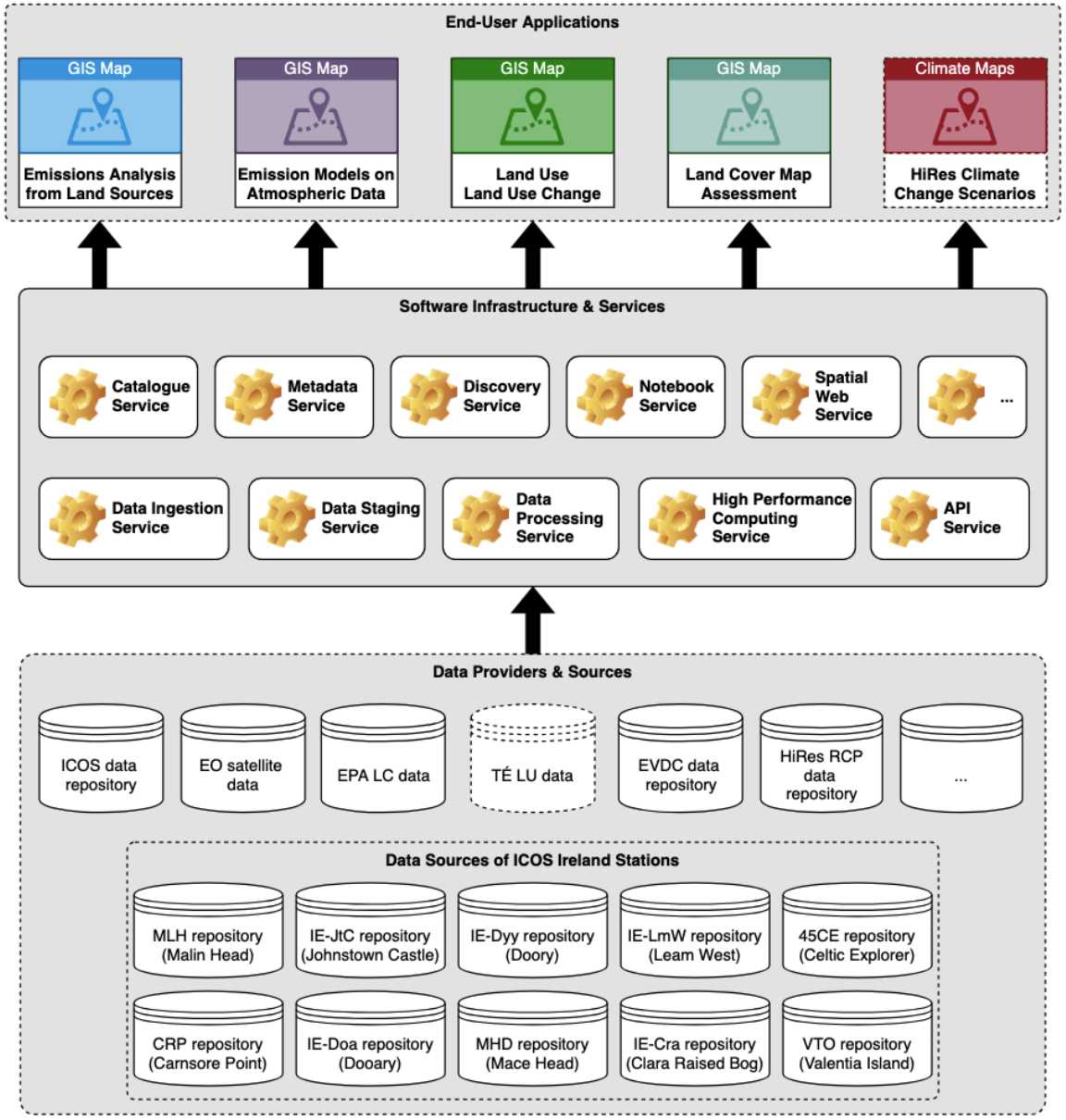
Map widgets in web pages, notebooks , ArcGIS, etc

jupyter

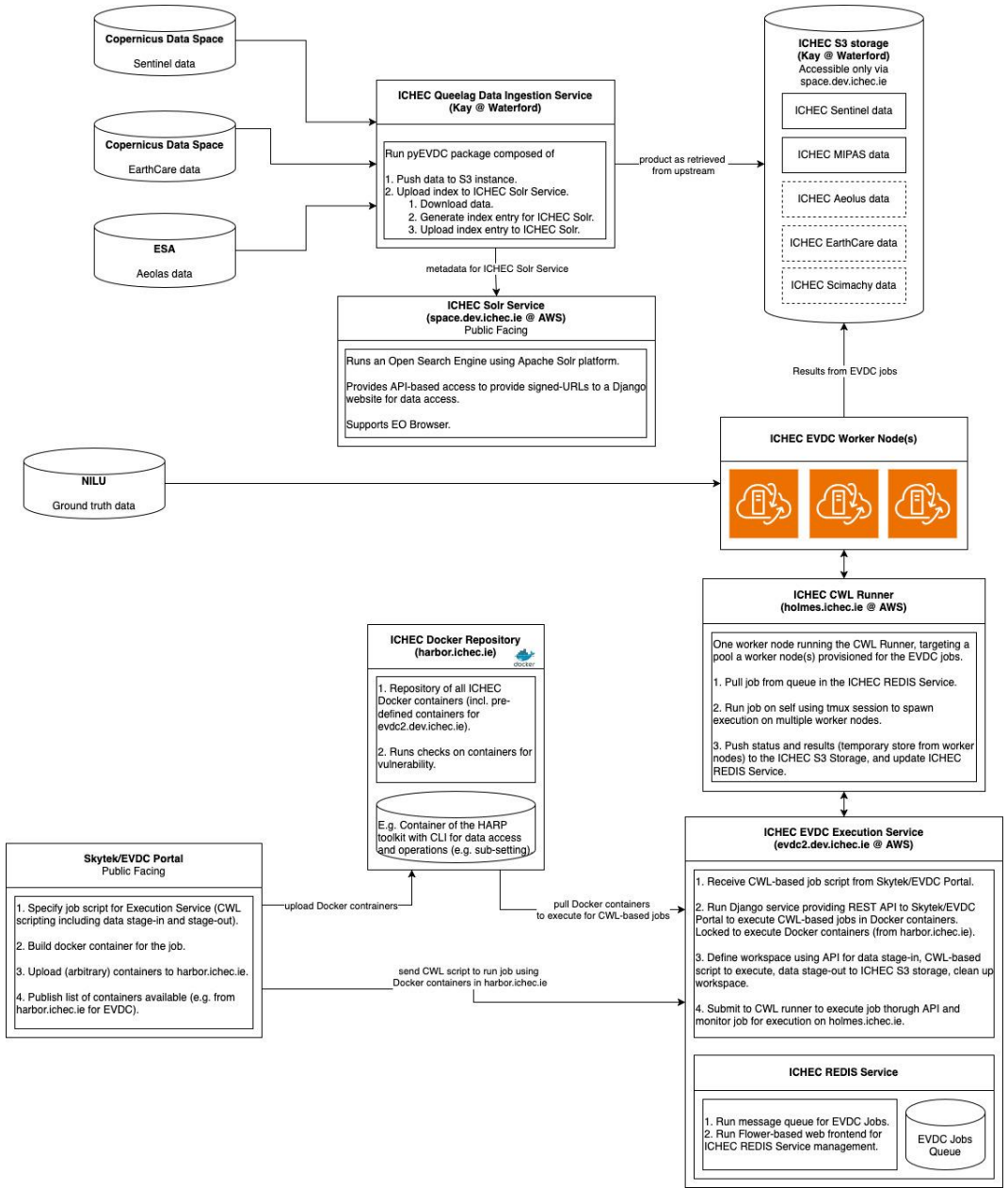
OPEN DATA CUBE



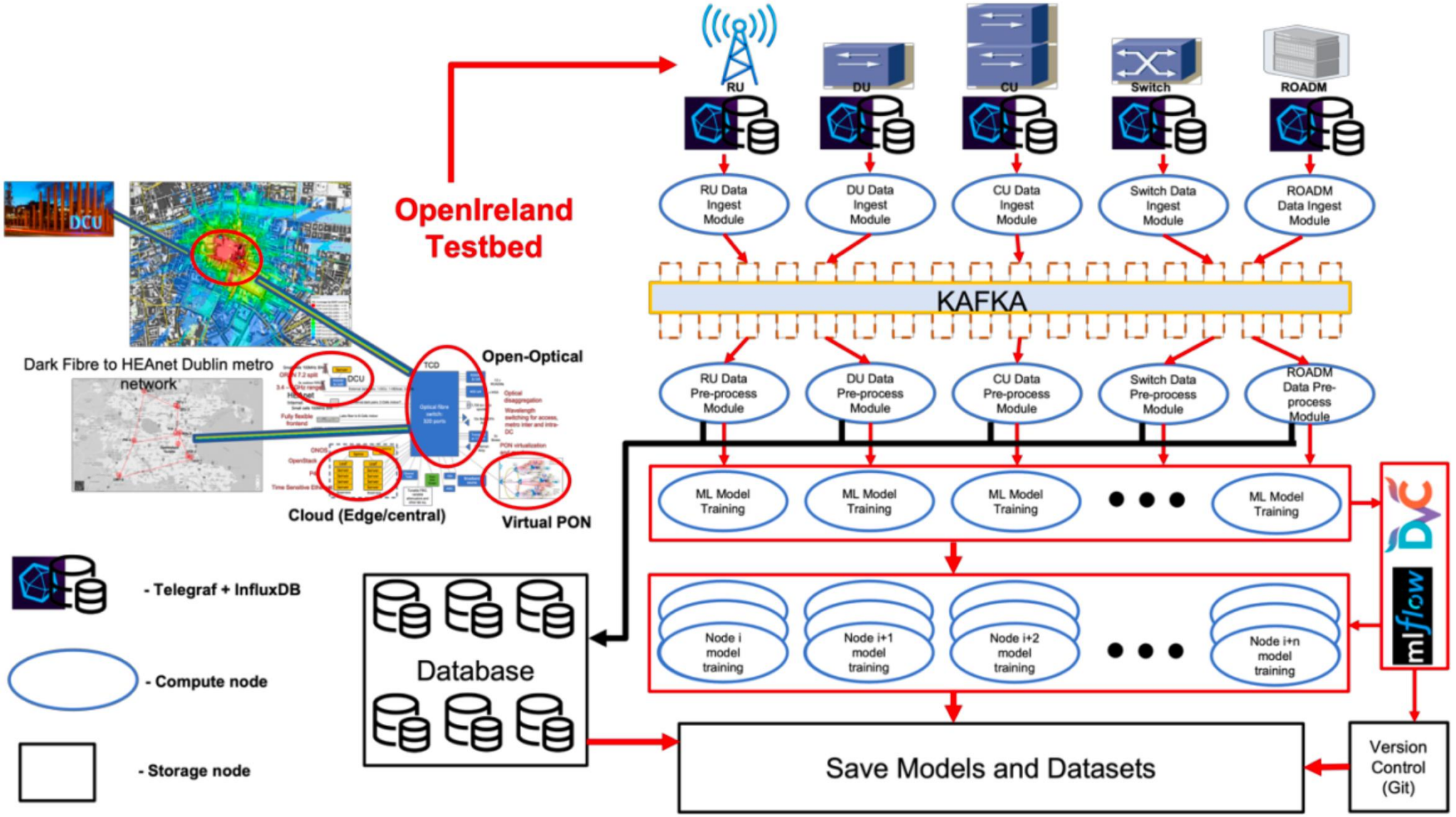
GeoCube


End-User Applications
GIS Map
Emissions Analysis from Land Sources
GIS Map
Emission Models on Atmospheric Data
GIS Map
Land Use Land Use Change
GIS Map
Land Cover Map Assessment
Climate Maps
HiRes Climate Change Scenarios
Software Infrastructure & Services
Catalogue Service
Metadata Service
Discovery Service
Notebook Service
Spatial Web Service
...
Data Ingestion Service
Data Staging Service
Data Processing Service
High Performance Computing Service
API Service
Data Providers & Sources
ICOS data repository
EO satellite data
EPA LC data
TÉ LU data
EVDC data repository
HiRes RCP data repository
...
Data Sources of ICOS Ireland Stations
MLH repository (Malin Head)
IE-JtC repository (Johnstown Castle)
IE-Dyy repository (Doory)
IE-LmW repository (Leam West)
45CE repository (Celtic Explorer)
CRP repository (Carnsore Point)
IE-Doa repository (Dooary)
MHD repository (Mace Head)
IE-Cra repository (Clara Raised Bog)
VTO repository (Valentia Island)

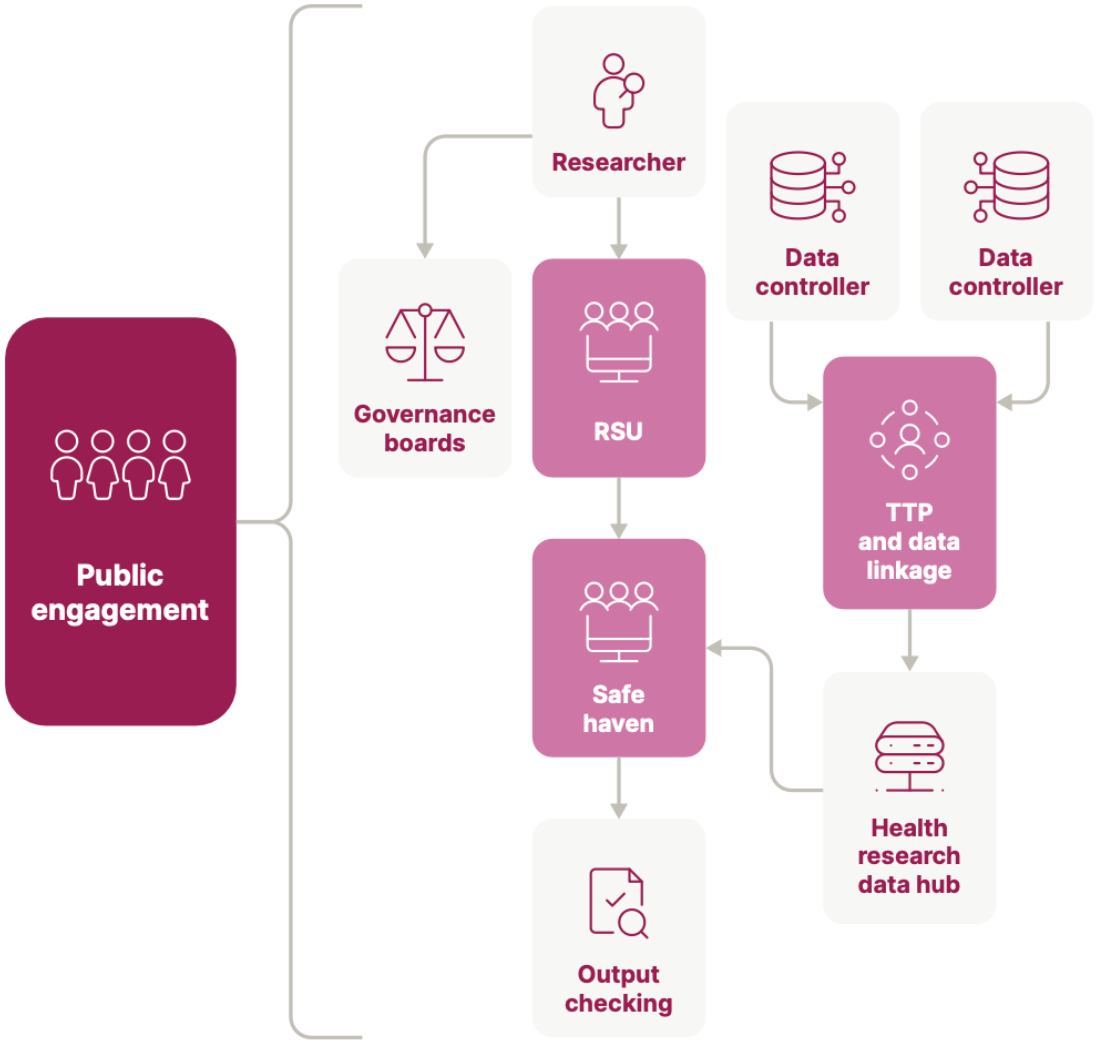
ESA Validation Data Centre



Telecommunication Testbed



Trusted Research / Secure Data Environments



- Technical infrastructure for data management and processing
  - Secure data environments (SDE/TRE/VDR)
  - Health research data hub
  - ISO 27001 certification
  - Co-location with other research infrastructure vs. secure cabinets/cages
  - Others?

- Institutional governance (of ICHEC) to implement and operate technical infrastructure for health R&I data platforms, services
  - (E.g.) RSU/TTP/DLU/HDAB ⇔ technical infrastructure provider
  - Handle DPIA, data input/management, data sharing agreements, repatriation of genomic data
  - Regulations/governance frameworks specific to themes (e.g., clinical research, cancer studies, genomics, etc.)
  - Others?

Mapping data, compute, usage characteristics

Data characteristics

- Average size (order of magnitude)
- Growth rate (frequency and size of accumulation)
- Ingestion rate (speed of new data ingested into system; recurrent, bulk?; throughput-critical?)
- Processing speed (based on use-cases on data; e.g., streaming, batch?)
- Data types and formats (e.g., satellite data, maps, images, tables, structured/unstructured, etc.)
- Read/write intensity (e.g., read-heavy, write-heavy, balanced?)
- Persistence (long-term preservation, warm/cold storage, purge after session, etc.)
- Data security (encryption, sensitivity, access control, authentication/authorization, etc.)
- Data flow patterns (e.g., ingest-curate-store, ingest-curate-analyse-storeresults-purgeinputs, ingest-curate-transform-store-analytics, etc.)

Compute characteristics

- Workload type (batch, real-time, interactive, hybrid)
- Intensity known (parallelism, memory requirement, acceleration, compute vs. IO)
- Input and output data characteristics (average per user job)
- Processing throughput (requests to handle per unit time, etc.)

Usage characteristics

- Access mechanisms (CLI/GUI, batch/XaaS)
- Usage characteristics (EPA/Agencies/researchers, number of concurrent users, peak concurrent requests, burst traffic patterns)
- QoS and SLA (uptime, response time, certification, security compliance, data sensitivity, data redundancy/replication, etc.)
- Scaling estimations (horizontal vs. vertical, auto-scaling, etc.)

Infrastructure Design Considerations

- Hardware
  - Compute | Data storage | Interconnects | Modularity
- Data Centre
  - Multi-site | Multi-tenancy | Tier-4
- Caging / Sandboxing
  - Software control | Data access control | Network restrictions | Execution isolation | Reproducibility
- Resource management / Scheduling
  - Workflow engines (DAG, job dependencies) | Heterogenous jobs | Hierarchical scheduling
- Software platform engineering
  - Containeraisation | Heterogenous jobs | Security | Modularity | Reusability | Micro-services
- User access & management
  - Batch, interactive, automated workflows | Portals, notebooks, CLI | Quotas, on-demand, shared
- Monitoring
  - Hardware resource utilisation | Application workflows | Software platform components
- Governance
  - Data | Institutional

IRL-DataSpaces

Framework & Roles	Architecture & Design
Security & Privacy	Lifecyle management
Compliance & Regulation	Stewardship
Ownership & Accountability	Metadata Management
Access & Usage Policies	Quality Management

Data Governance

Security & Privacy	Standards & Interoperability
Regulations & Compliance	Ecosystem & Services
Lifecycle & Provenance	Gravity & Performance
Residency Requirements	Open/Closed/Sensitivity
Risk Management	Producers, Users & Use-cases

Sectoral Specifics ⟷ **Technical Infrastructure**

Resource Registration	Data Services
Catalogue & Discovery	VRE / Data Rooms
Metadata Services	HW & SW Infrastructure
PID Services	Certification & Cybersecurity
Data Ontology	Federated AA(A)I

IRL-DataSpaces