Technical requirements from the EOSC use cases

Diego Scardaci (EGI.eu), Giacinto Donvito (INFN)

EOSC-hub Technical Workshop, Amsterdam 25-27 June 2019

eosc-hub.eu

@EOSC_eu

Dissemination level: Public/Confidential *If confidential, please define:*
Disclosing Party: (those disclosing confidential information)
Recipient Party: (to whom this information is disclosed, default: project consortium)

EOSC-hub receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 777536.

Outline

- EOSC Use Cases
  - EOSC-Pilot scientific demonstrators
  - EOSC-hub Thematic Services
  - EOSC-hub Competence Centers

- Build EOSC for users
  - Defining EOSC technical architecture to satisfy users' needs
  - How EOSC can support user communities
  - Reference architectures, interfaces and interoperability guidelines for users
  - Useful documentation

EOSC Use cases

EOSC-Pilot Scientific Demonstrators (PanCancer, EPOS/VERCE, Photon & Neutron)

EOSC-hub Thematic Services

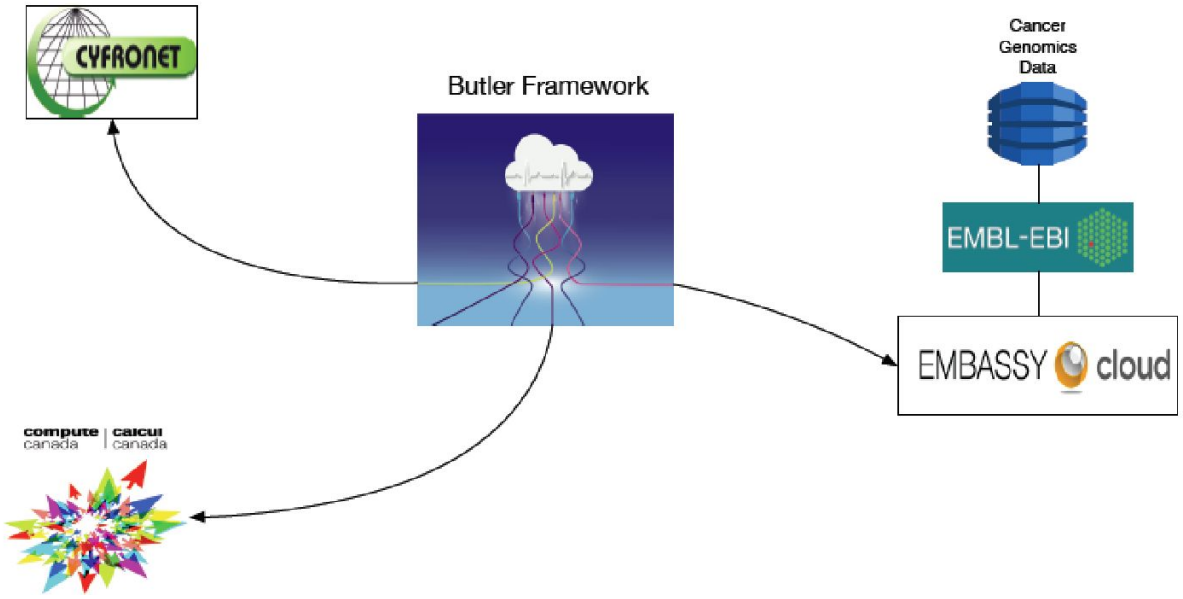
EOSC-hub Competence Centers (ELIXIR, EISCAT-3D)

PanCancer - Challenges

- Collect Next Generation Sequencing Data from several cohorts of cancer patients generated at multiple sequencing centres and across multiple cancer types.
- Reanalyze the data using a uniform and consistent data processing pipeline utilizing established best practices from the International Cancer Genomics Consortium.
- Analyze the integrated data set to identify patterns of germline and somatic mutation that act across cancer types in a PanCancer fashion.

PanCancer – Scientific Demonstrator Work

- Utilize Butler, a cloud based large scale scientific workflow framework developed in the context of ICGC's Pancancer Analysis of Whole Genomes project to perform a coordinated data analysis across multiple clouds.
  - Code https://github.com/llevar/butler
  - Paper https://doi.org/10.1101/185736
- Perform automated repeatable deployments and configuration of the entire processing infrastructure at three academic cloud (EMBL EBI, ComputeCanada, EGI/Cyfronet)
- Deliver a large dataset (>50 TB) to each cloud computing centre.

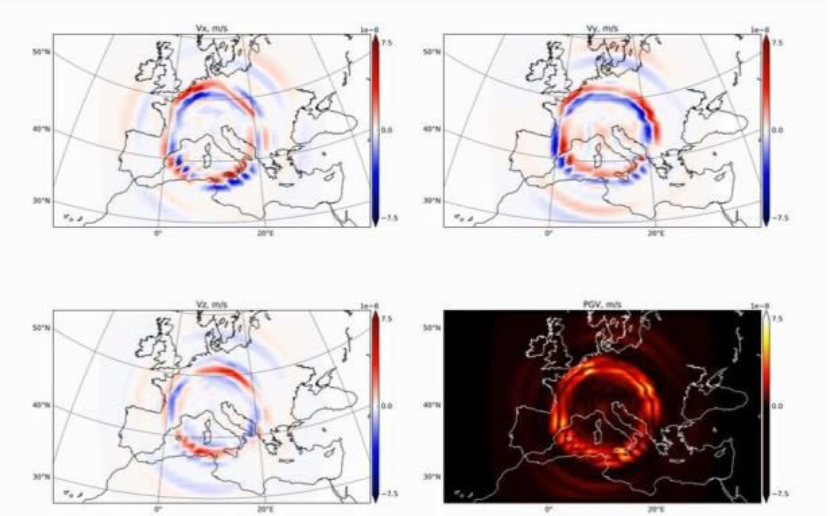


PanCancer - Technical Requirements

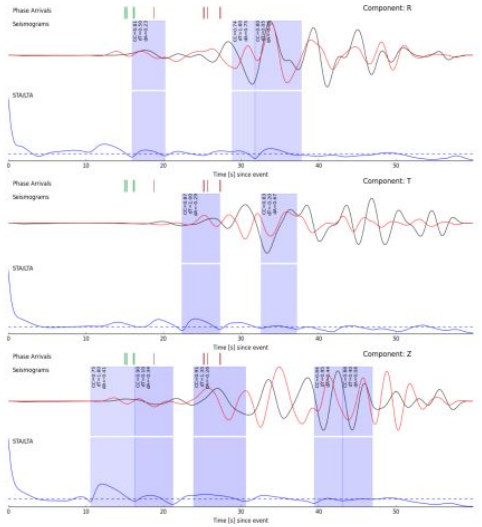
- Shortage of resources for operating at “cloud scale”.
  - Used 20% of data set that was utilized for PCAWG
  - < 0.5% of data set for 100k Genomes
- Repeatable provisioning of large clusters of VMs
  - ->10% of provisioning jobs experience failures
- Data movement and staging
  - 50 TB data set takes up to two weeks to move locations
  - Genomics data requires encryption and network security measures
- Shared access to network storage creates processing bottlenecks.
- Diverse data sets have diverse data handling requirements
- Automated detection and resolution of issues with infrastructure needed for effective operation at cloud scale

EPOS/VERCE Earthquake simulation Platform - Challenges

- **Earthquake Simulation**: Production of synthetic seismograms for public and custom Earth models and Earthquakes via the execution of HPC simulation codes (SPECFEM3D & Globe)



- **Raw data acquisition & Misfit**: The model is evaluated and further improved by comparing the synthetic data with real observations collected by institutional archives, adopting Data Intensive workflows.



EPOS/VERCE - Scientific Demonstrato Work (1/2)

- Enhancing the services of the VERCE portal and integrate the EGI FedCloud Infrastructure as the main data intensive computational service provider for Misfit Analysis Workflows
- These consists in three different phases
  - Assisted discovery preprocessing of the observed data and the correspondent synthetic results
  - Data pre staging from the FDSN network to an iRODS instance with metadata and provenance
  - Final comparison adopting different Misfit techniques
- AAI and delegation mechanisms are needed to submit executions and to connect to remote data stores (iRODS) from the Cloud

EPOS/VERCE - Scientific Demonstrato Work (2/2)

- Processing workflows enabling the Misfit analysis (data download, preprocessing and misfit) have been refactored to support their execution on EGI FedCloud resources
- The lineage services have been upgraded to a later version of the ProvFlow system improving the interactive exploration of lineage information and delivering PROV format for interoperable provenance analysis
- The portal has been extended to allow the retrieval of Per User Sub Proxy certificates from the eToken proxy certificate additionally to its community specific IdP Login via OpenID Connect through the EGI Check In service has been successfully validated

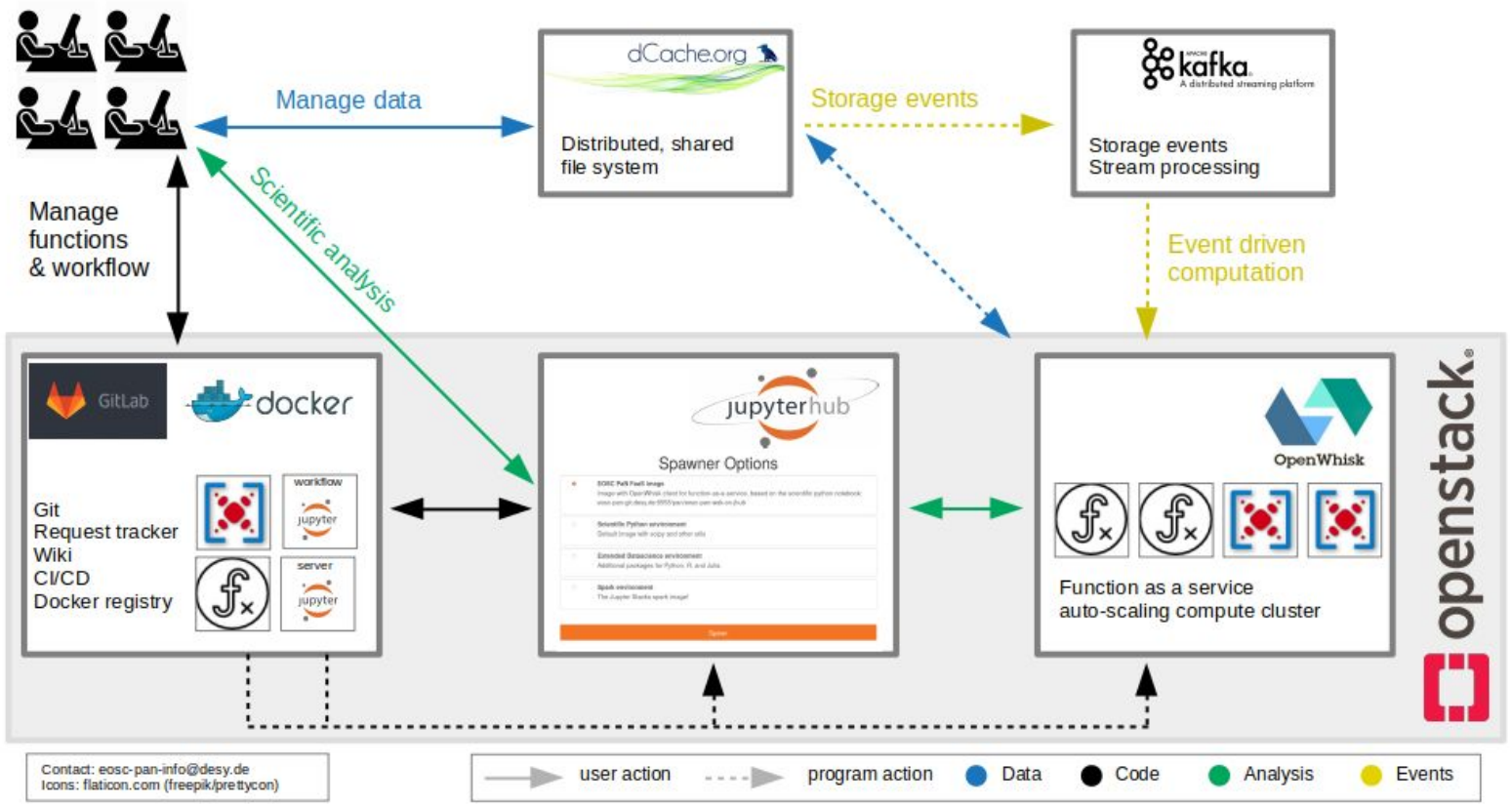
EPOS/VERCE - Technical Requirements

- No built in credential delegation e.g. proxy certificate delegation
- Differences in network implementations on FedCloud sites (required floating IP requests on some sites, not implemented by middleware at the time)
- EOSC could *feature* software and tools, promoting them to the communities with enough description about usage and transparency in its sustainability plans
- The reproducibility problem should start to be addressed structurally, scaling from ad hoc solutions to reusable and more general services Computational tools offered by EOSC should be aware of the existence of these service and use them

Photon & Neutron - Challenges

- Demonstrate the use of EOSC resources to support the technical requirements of CrystFEL24
  - a software suite created to address the processing needs of serial femtosecond crystallography (SFX).
- Raw data involved during the analysis ranges between 1-100TB resulting very difficult to be moved around.
- Efficient analysis of large datasets requires
  - bringing together the data with workflows
  - computing resources in an integrated platform.

Photon & Neutron - Scientific Demonstrator Work



Technical Capabilities:

- Compute - Cloud compute infrastructures - FaaS
- Storage - Integration of middleware for mass storage systems (e.g. dCache)
- Processing & Analysis - Jupyter Hub
- Networking - Local VPN to access VMs and containers in multi-cloud environment
- Integration with federated AAI

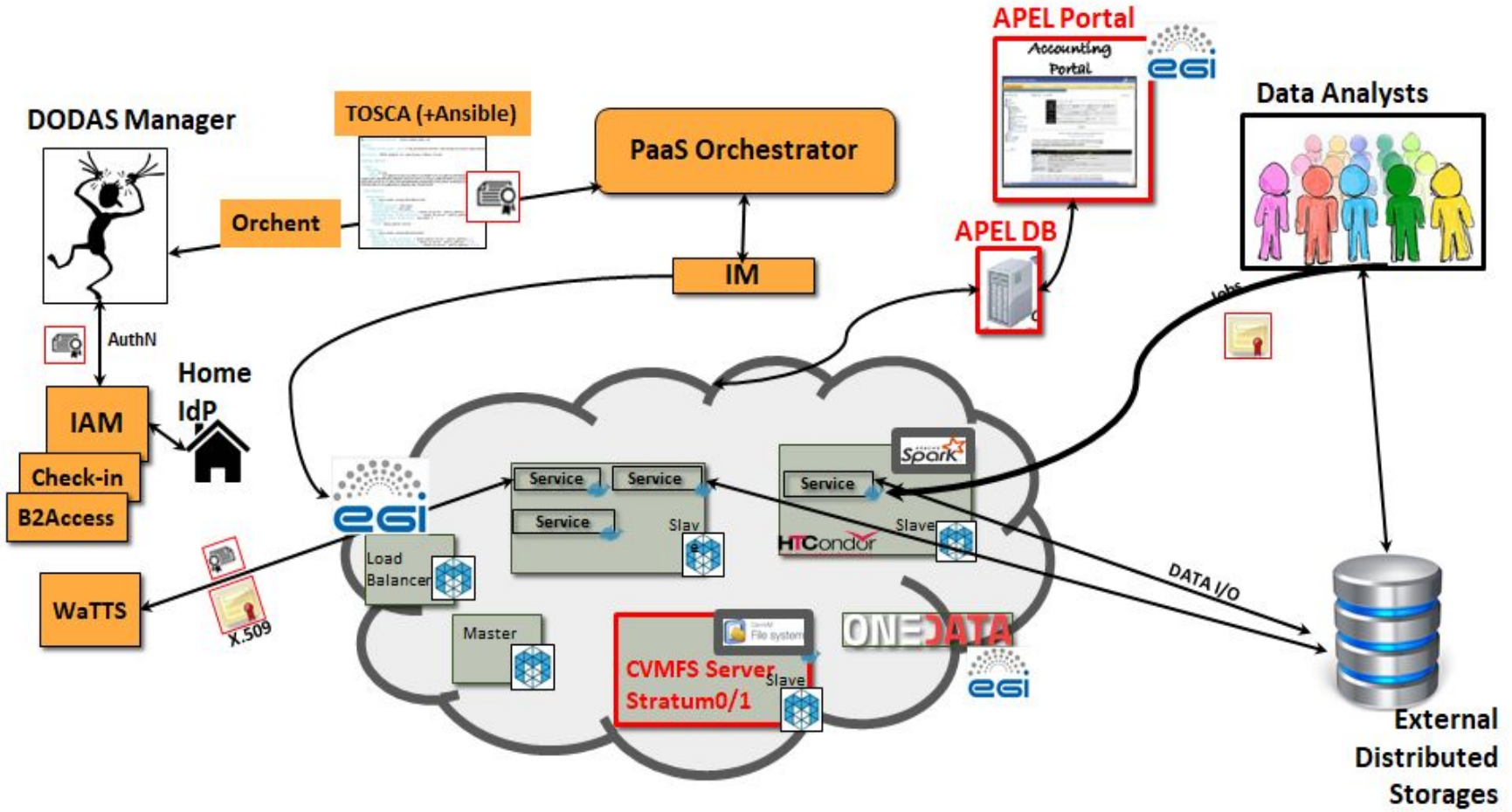
Photon & Neutron - Technical Requirements

- Reliable and effective inspection/surveillance service, across a global network, to validate intermediate analysis.
- Provide access to services based on license. It would be more scalable whether EOSC can provide such attributes in a central/federated way.
- Better integration of Jupyter notebooks with EOSC services would be beneficial.
- EOSC centralized Docker registry providing trusted solutions.

DODAS - Challenges

- **Dynamic On Demand Analysis Service (DODAS)** automates the process of provisioning, creating, managing and accessing a pool of heterogeneous computing and storage resources:
  - HTCondor based batch system as a Service
  - Big Data platform for Machine Learning as a Service
- DODAS has a highly modular architecture and its workflows are highly customizable.
- Adopted by the WLCG CMS and Alpha Magnetic Spectrometer experiments
  - Other interested communities

DODAS - Thematic Service integration activities



- The Services come-out from different experiences: EGI, EUDAT, INDIGO-DataCloud
- The end user is able to instantiate dynamic clusters exploiting heterogeneous computing resources and distributed storage

Integration: EGI DataHub, CVFMS, EGI Check-in, INDIGO IAM, Orchestrator Accounting, Monitoring, etc.

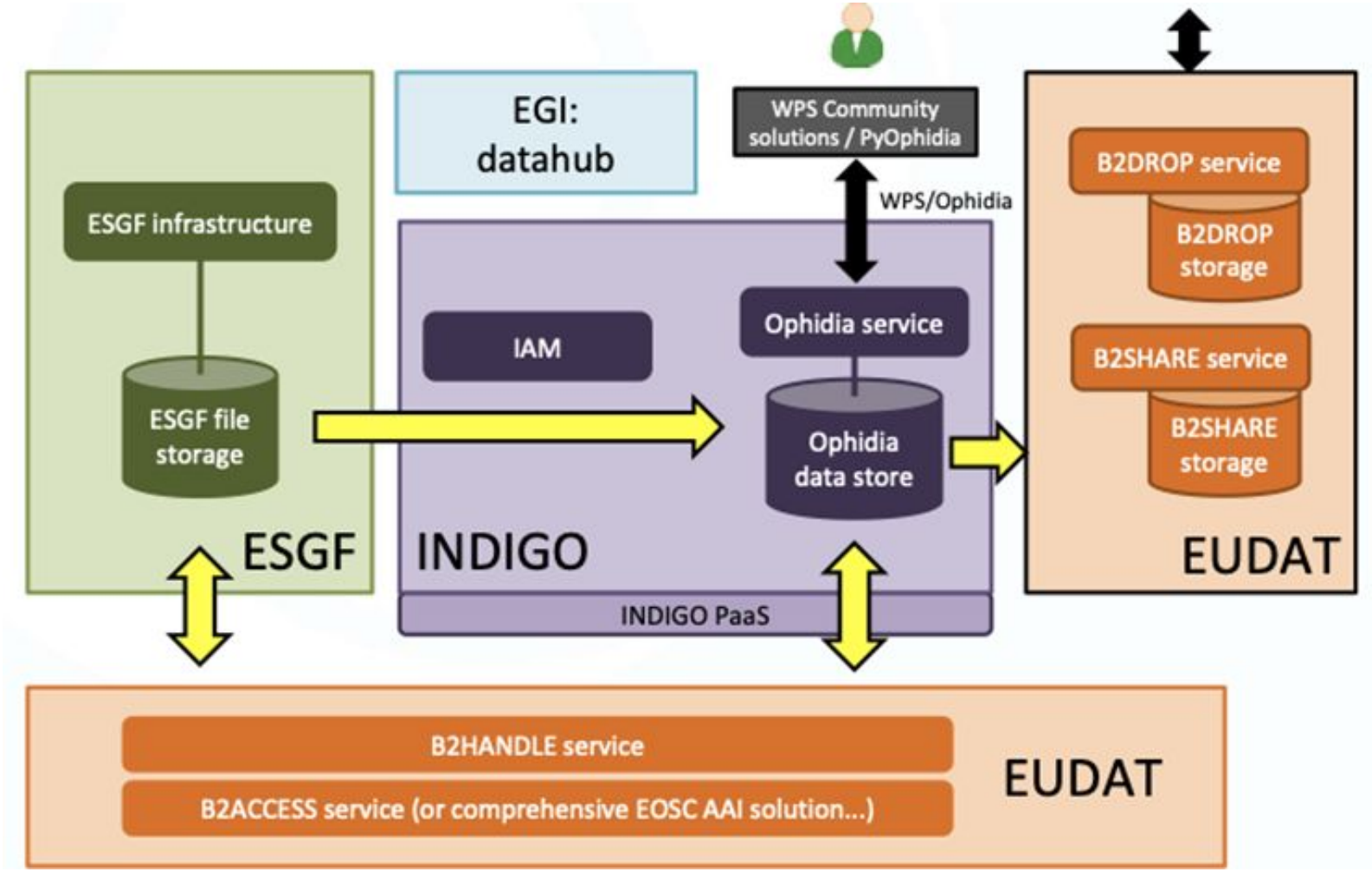
DODAS - Technical Requirements

- Solutions to implement transparent data access
  - reduce possible source of inefficiency coming from the latency during read operations of data hosted outside the cloud provider where CPU is
- Solution for data ingestion and temporary store of input/output data.
- Solution for user authentication and authorization
  - Delegation (services acting on behalf of users)
  - Federating IAM with other EOSC AAI
- Dynamic infrastructure accounting
  - Deploy accounting probes on the fly
  - Connect accounting probes to the central repository
- Dynamic extension of cluster through CLUES

ENES Climate Analytics Service (ECAS) - Challenges

- **ECAS Climate Analytics Service (ECAS)** enables scientific end-users to perform data analysis experiments on large volumes of multidimensional data:
  - PID-enabled, server-side, and parallel approach
  - Focus on data intensive analysis, provenance management, and server-side approaches
  - Ophidia Big Data Analytics framework
- 4 main use cases:
  - Users with no directly available computing or data analysis resources
  - Users from climate data communities
  - Climate data cross-model/large-scale ensemble analysis
  - Prototyping applications with datacube concept

ENES/ECAS - Thematic Service work



- Services provided by: EGI, EUDAT, INDIGO-DataCloud
- The user exploits the data previously stored in EUDAT Services, to process with an high level PaaS solution that provide easy and straightforward way of deployment of the analysis workflow

ENES/ECAS - Technical Requirements

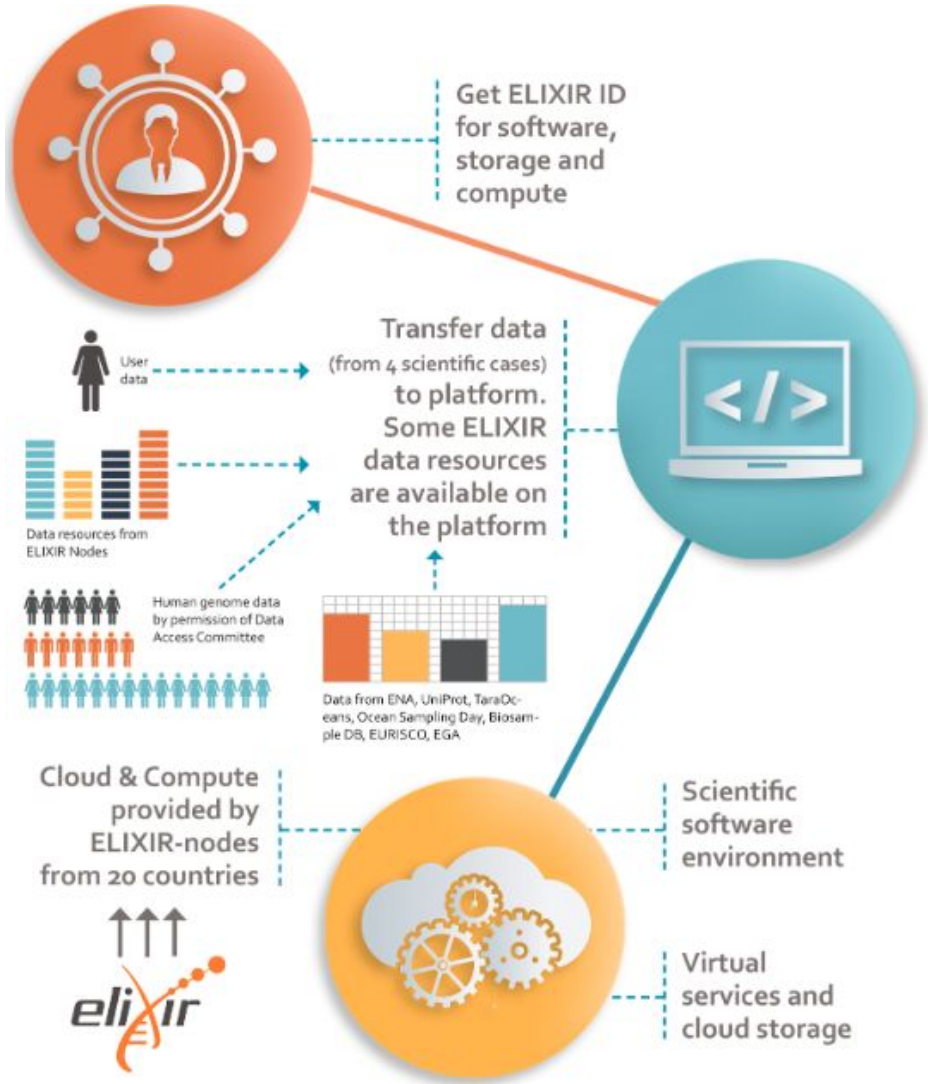
- Software containerization of the ECAS environment using Docker
- Providing good workflows for development and operations (Jupyter and Ophidia)
- Integration with the Earth System Grid Federation (ESGF)
- Integration with the EGI Data Hub (OneData)
- Data sharing (B2DROP)
- Integration with the EOSC-hub accounting and monitoring
- Integration with AAI (IAM)
  - Interoperability with other EOSC AAI solutions
- Persistent identifiers (PIDs) on output results and connecting them with input data (B2Handle)

ELIXIR - Challenges

Single-sign on to ELIXIR services across all platforms

User need to be trained of their possibilities

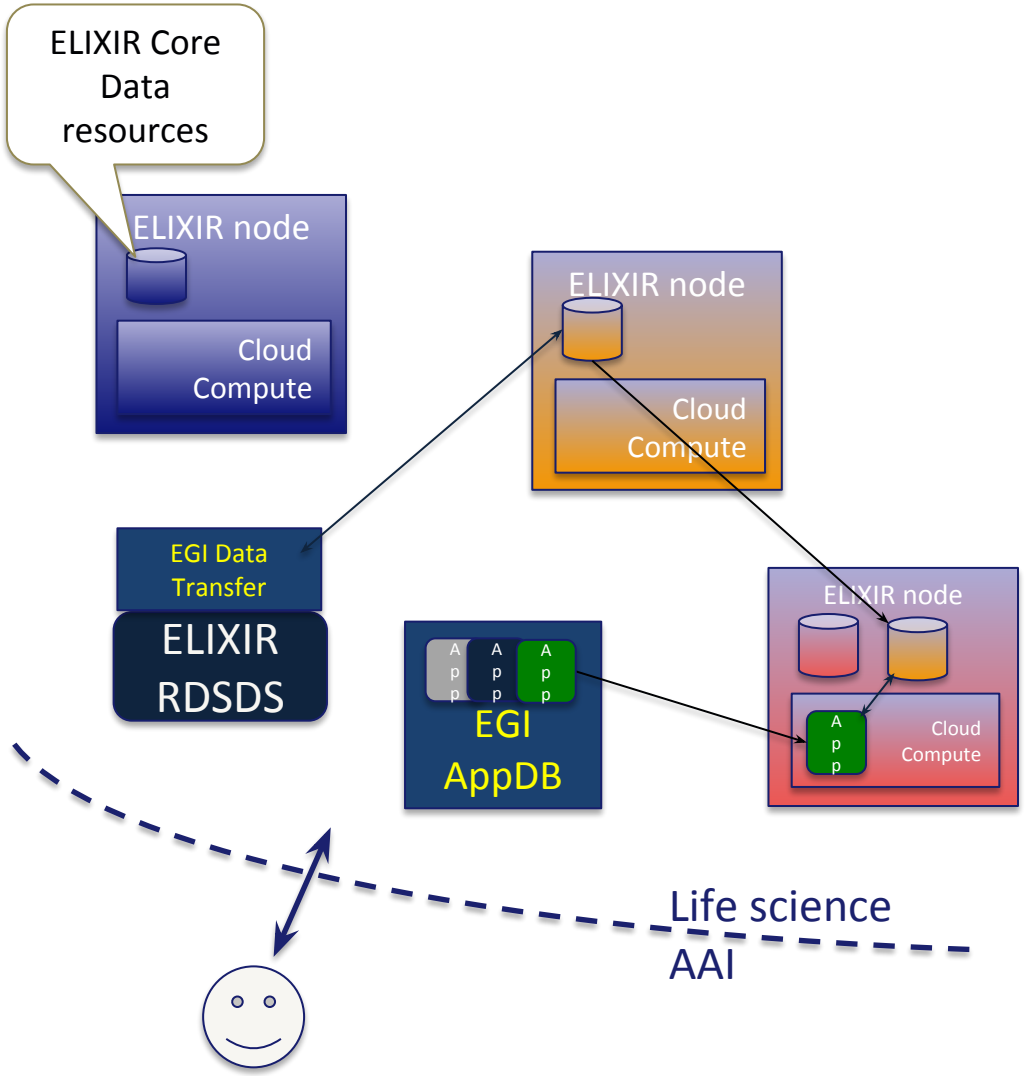
Common data formats, compute interfaces, service processes increase interoperability



Scientific software-level access to compute on data: Internationally compatible access to data in collaboration between Tools, Comp and Data

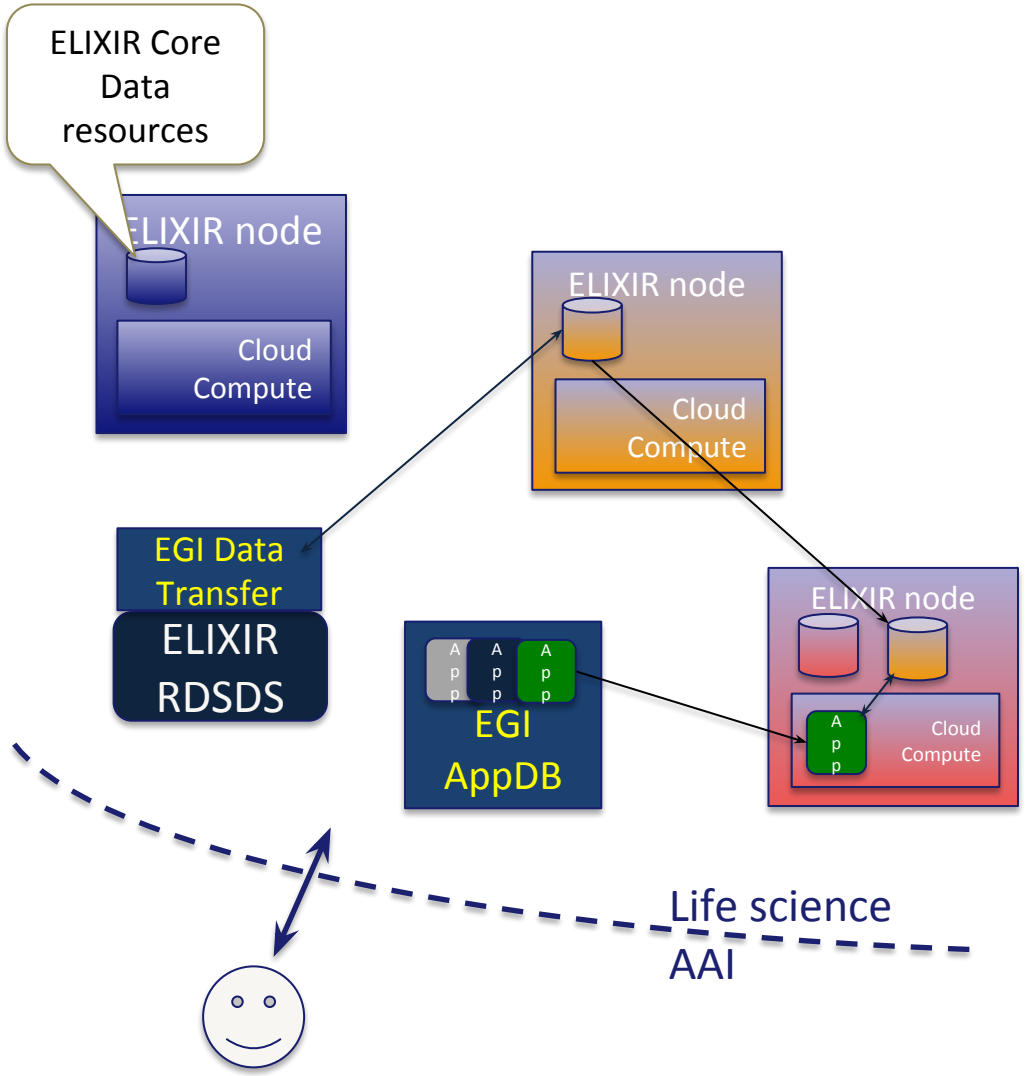
ELIXIR needs to leverage EOSC and e-infrastructure experts and services

ELIXIR – Competence Center Work



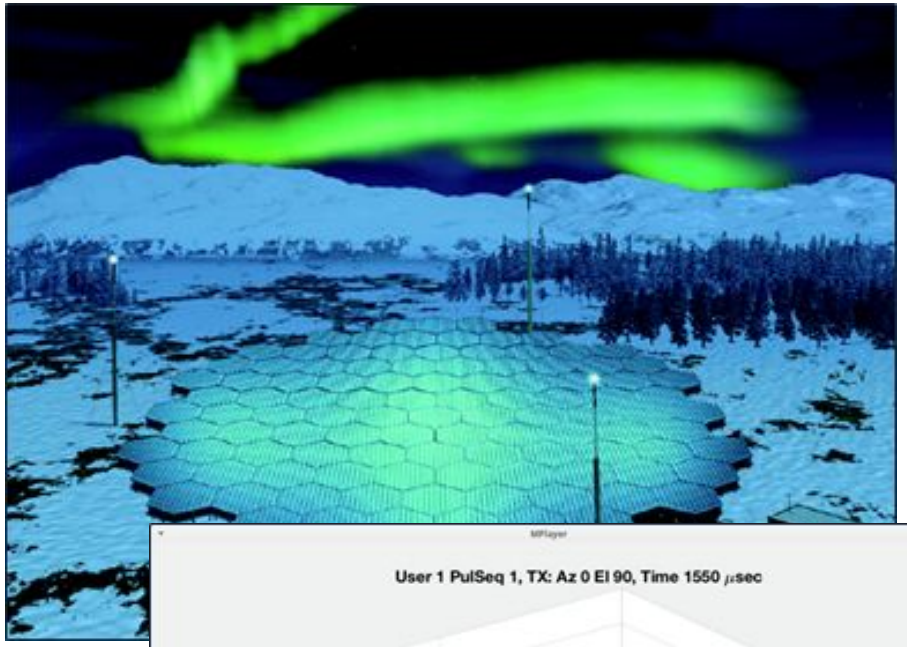
1. Technical integration
   - EOSC-hub AAI -- Life Science AAI
   - EGI Cloud (cloud federation)
   - EGI AppDB (virtual application store)
   - EGI Data Transfer Service (file replication)
2. Business model
   - Capacity allocation to users and to user projects projects
3. Data access policies
   - Sensitive data
4. Training
   - Train the resource centres

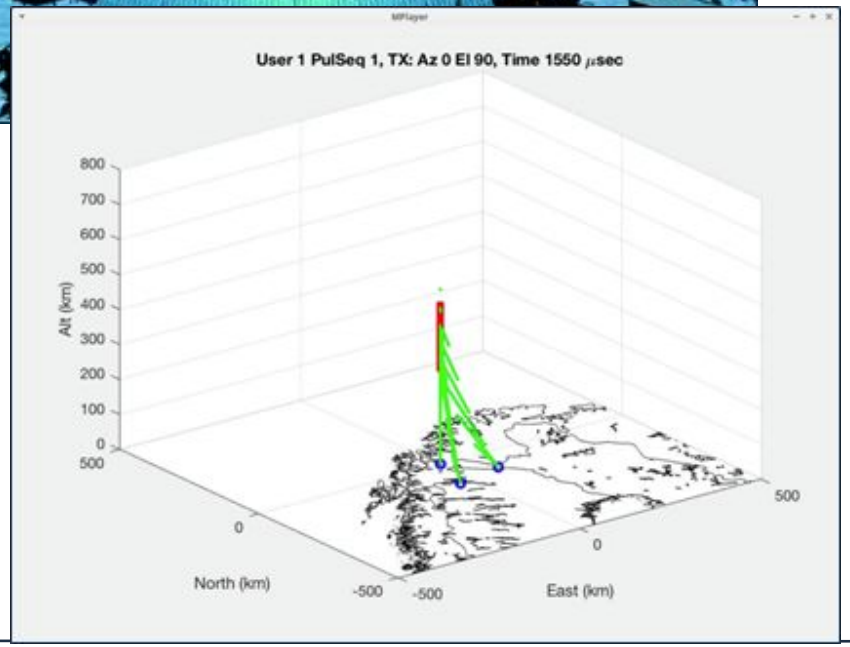
ELIXIR – Technical Requirements



1. EOSC AAI <-> Life Science AAI interoperability
2. Connect ELIXIR cloud compute or data storage location with EOSC
3. Data and application replication backbone for community centres
   - EOSC centrally provided data and applications distribution services ?
4. Kubersents as a Service
5. Handling sensitive data

EISCAT-3D - Challenges



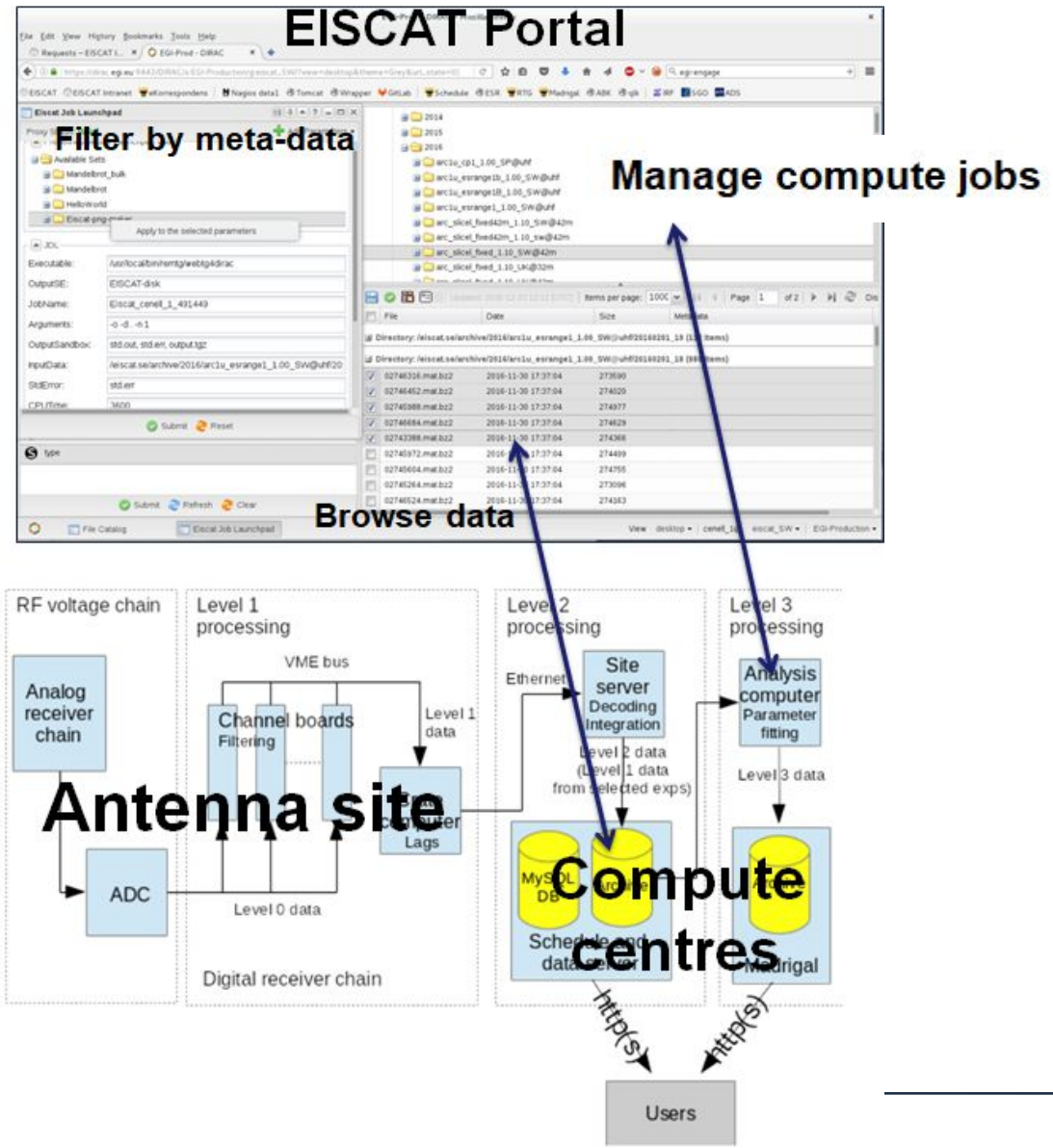


- Project of the European Incoherent Scatter Scientific Association (EISCAT)
- Studying interactions in the auroral ionosphere and magnetospheric cusp regions
- 109 Arrays of 91 antennas (+2 receiver stations)
- Up to 100 simultaneous beams
- Maximum data rate after beamforming > 50 Gb/s
- First data expected in 2021

EISCAT-3D - Data Policy

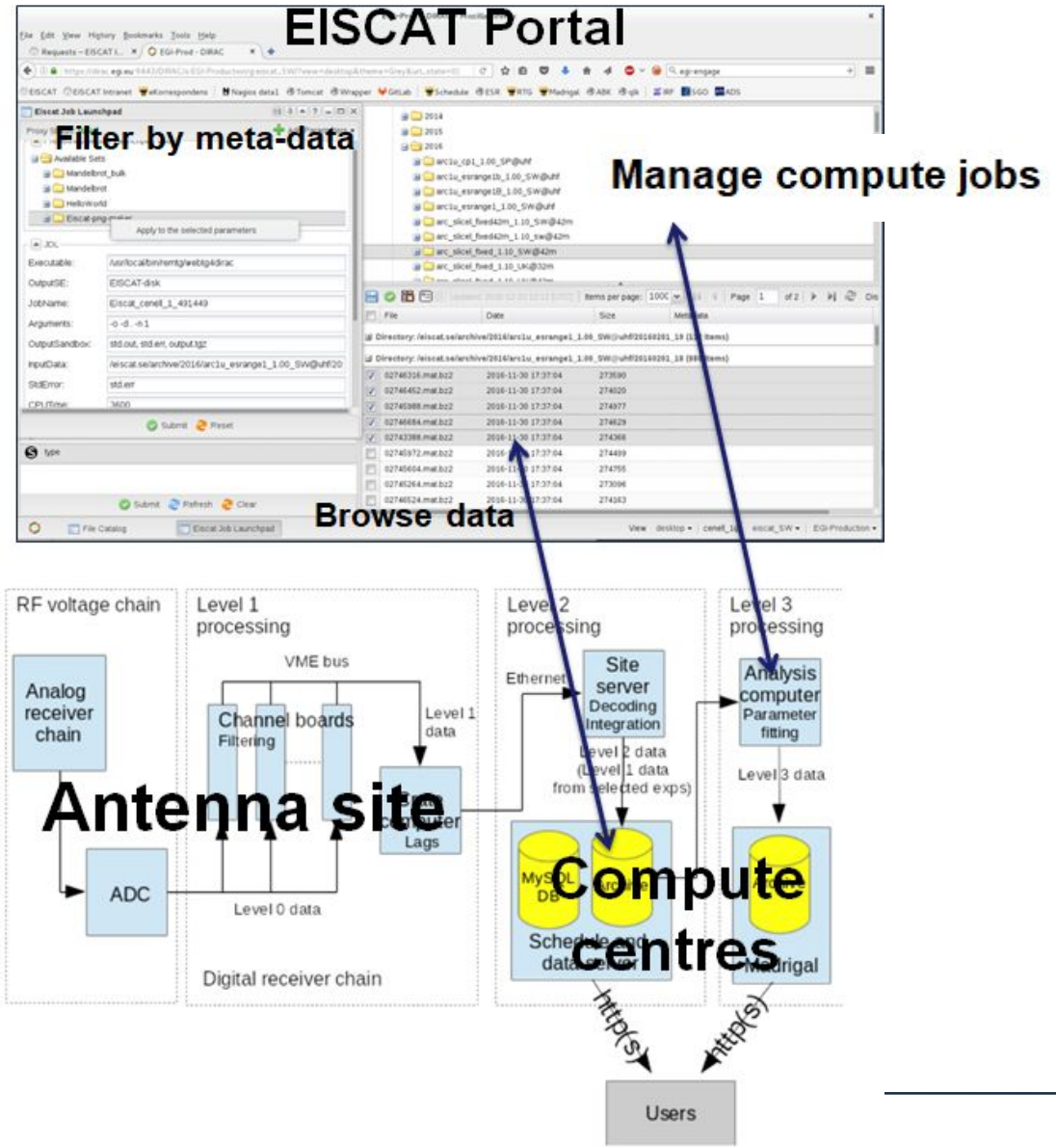
- Low-level data from each experiment embargoed for defined periods (typically 1 year for EISCAT member carrying out experiment, 3 years within EISCAT membership)
- Analysis of data either close to Data Centres or “spare” on-site computing.
- Data used in research should be given Persistent Identifiers (PIDs) according to a common standard
- A 4 months period is selected as this is the estimated time required to perform a “real-time” analysis on low-level data.
- A portion of the level 1 data will also be archived permanently, on the order of 1% of the level 1 data rate
- Strong recommendation to follow ENVRI-FAIR principles - FAIR = Findable, Accessible, Interoperable, Reusable

EISCAT-3D - Competence Center Work



- Service integration
  - Data transfer
  - Cloud computing
  - Container computing
  - AAI
  - User portal (DIRAC)
- User engagement
  - Early adoption
  - Technical training
  - Continuous service improvement

EISCAT-3D - Technical Requirements



- Temporary file storage close to computing
- Run VMs and/or containers for portal and data processing
- Registration of digital objects: data collections
- Federated AAI

EOSC-hub

Build EOSC for users

Defining EOSC technical architecture to satisfy users' needs

User technical requirements - some considerations

Requirements analysis allow to identify a set of common needs. Some (*non exhaustive!*) examples

- **Compute**
  - Federation models
    - Interfaces between federations
  - Dynamic Clusters
  - Containers:
    - Docker and Kubernetes
  - Automation for effective operation
    - Deploy on demand
    - Automatic re-submission in case of failure
  - Effective inspection/surveillance service, across a global network, to validate intermediate analysis.

User technical requirements - some considerations

- **Workflow**
  - Engines
  - Jupyter Notebook integrated with EOSC services
- **Fast Network links**
  - E.g. for data movements
- **Federated AAI**
  - Interoperability between AAI solutions
  - Delegation
    - in particular service to service and in different AAI realms
- **Federation Tools:**
  - (Dynamic) accounting and monitoring infrastructure system

User technical requirements - some considerations

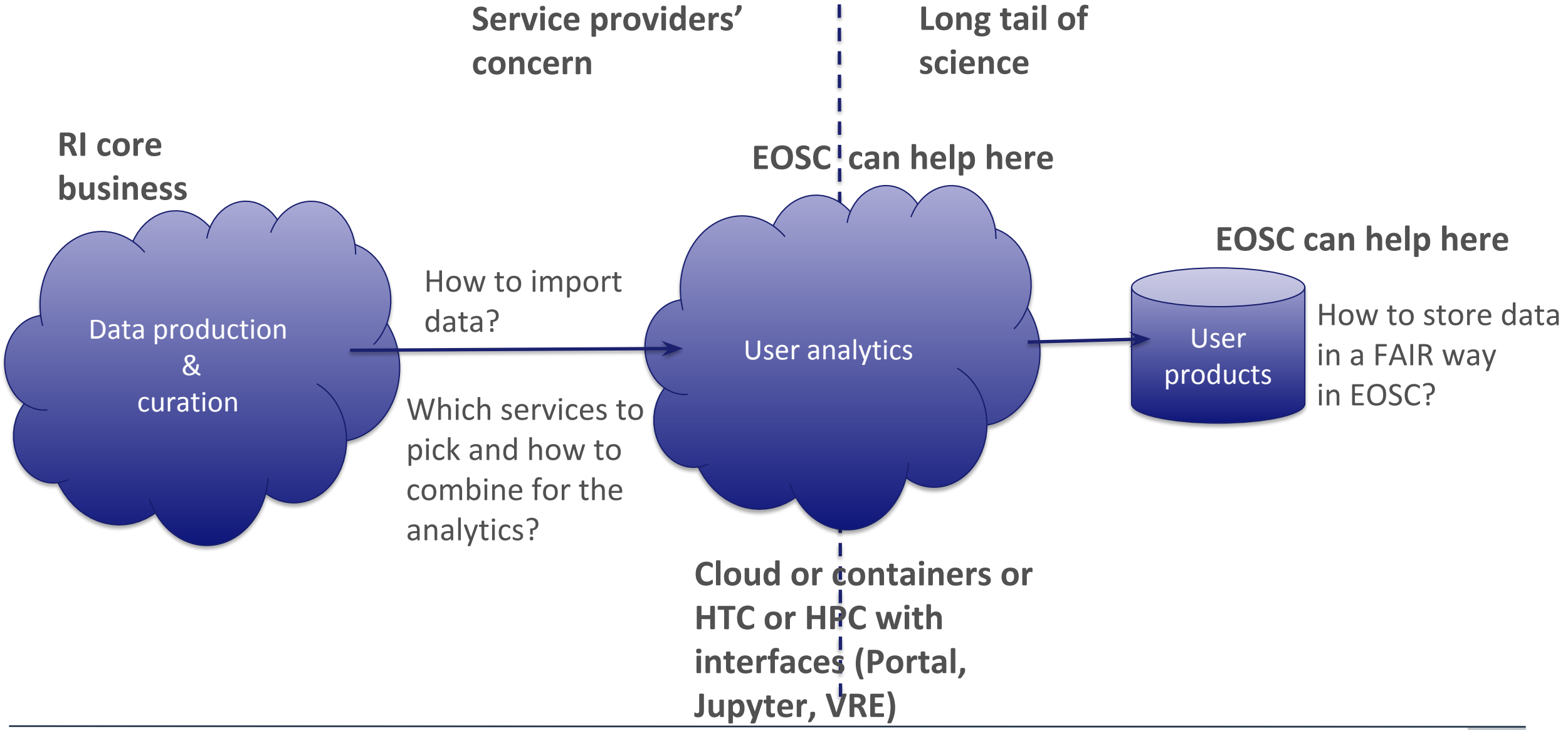
- **Data**
  - Transparent data access
  - Movement and staging/ Data ingestion
  - Metadata management
  - Data sharing
  - Integration with external data sources
  - PID
  - Handling sensitive data

User technical requirements - some considerations

- **Featured EOSC software and tools**
  - Large capacity
  - Large databases (e.g. ESGF, Copernicus Data, etc.)
  - Data and application replication backbone for community centres
  - Analysis reproducibility tool
  - Central docker registry
  - License service

Shared resources in the EOSC Federating Core

How EOSC can support user communities



Define the EOSC Technical Architecture to satisfy users' needs

From user requirements:

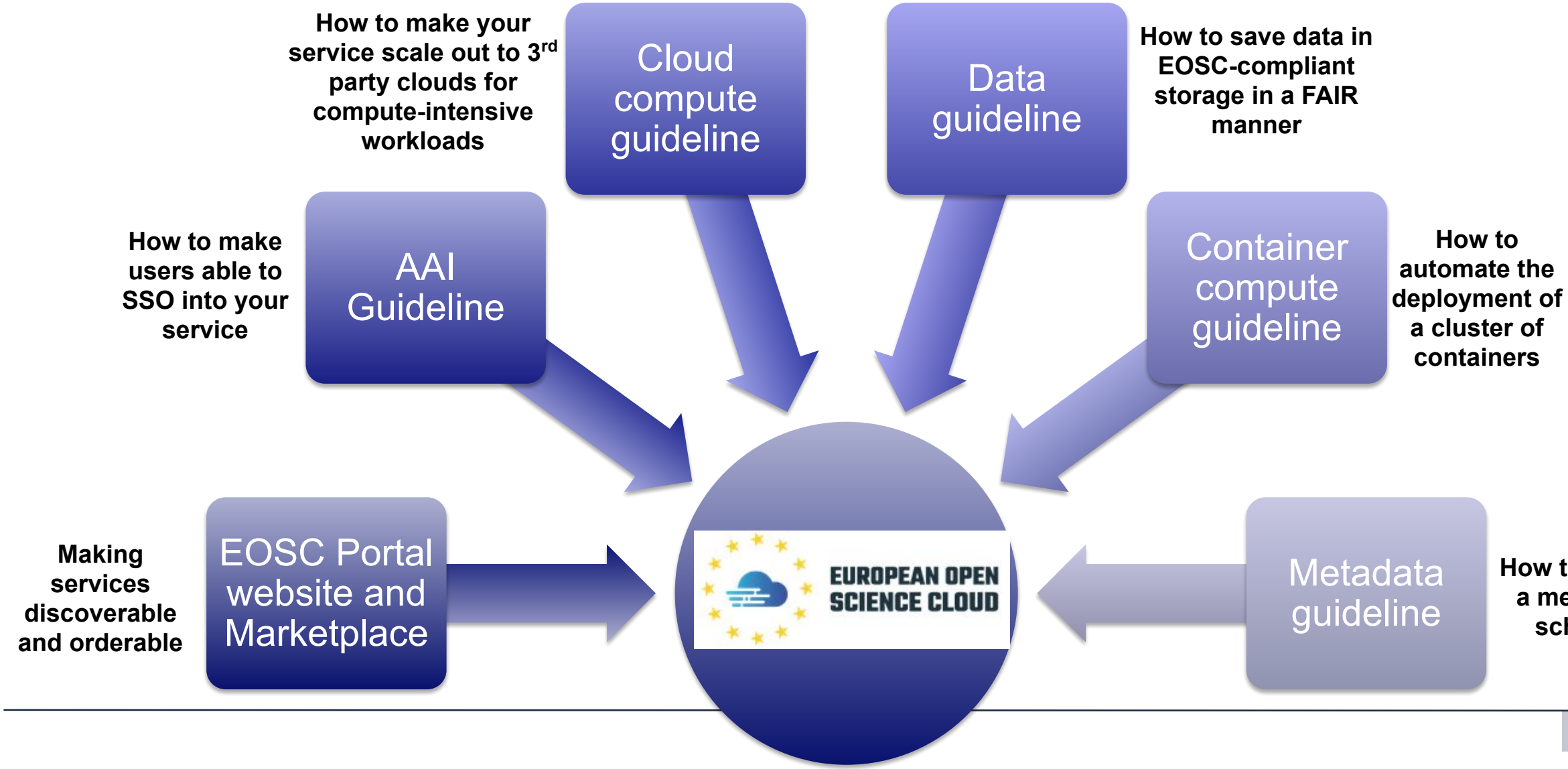
- Access enabling and federation tools
  - Define clear interfaces to:
    - Use/exploit these tools
    - Interconnect various instances of these tools (e.g. AAI, accounting, monitoring, …)
- Most requested **macro-features** for each technical area
  - Priority to define reference architecture and **interoperability guidelines**
- Identify **common combinations of composable services**
  - Which are the most common combinations ?
    - Plan and foster integration
  - Define and (then) leverage on interoperability guidelines

Define the EOSC Technical Architecture to satisfy users' needs

From user requirements:

- Shared resources
  - Large capacity is a must
  - Public large data repositories
  - Other commodity services
    - Data and application replication backbone for community centres
    - Analysis reproducibility tool
    - Central docker registry
    - License service
  - Each technical area can suggest components for the **Shared Resources of the EOSC Federating Core**

Reference architectures, interfaces and interoperability guidelines



Useful documents and links

Each technical area should use the following documents as references for technical requirements:

- **EOSC-Pilot D5.6:** Evaluation report of service pilots
- **EOSC-hub D7.2:** First Report on Thematic Services
- **EOSC-hub D8.1:** Report on progress, achievements and plans of the Competence Centres (is under finalisation)
- **EOSC-hub Community Requirements Database:** https://wiki.eosc-hub.eu/display/EOSC/Community+requirements+DB

Thank you for your attention!

Questions?

Contact

eosc-hub.eu @EOSC_eu