Design your e-infrastructure!

https://indico.egi.eu/indico/event/4434/

Use case:

Break out group coordinator:
Bjorn / Enol
…

Amsterdam, 9. May, 2019.

Group members

- ASTRON
- NenuFAR
- PROCESS
- EISCAT_3D
- EGI

First break-out

Background and Users

Who will be the user? Can the users be characterised? How many are they?

	ASTRON	NenuFAR	PROCESS	EISCAT
Users	• Astronomers • Astrophysicists • Particle physicists	• As ASTRON	• Pipeline developers • Astronomers • Medical	• Space physics (near and far) • Plasma physics • Atmospheric physics
Characteristics	• Transition from local to cloud services (zero downloads) • Pipeline developers • Backend providers • Provide own workflows • Production managers • Support for L0 through L4 data is needed	• ±same as ASTRON • Experience using LOFAR → used to working remotely	• Prepare for Exa-scale • 2 sets of users • 1. Pipeline dev • 2. End-end users • User can build code (1) • Unable to scale up beyond local resources (1) • End-end users just use portals (high level of abstraction)	• 3 sets of users • 1. just download and use data locally • 2. Pipeline devs • 3. Operational services… job scheduling++ • Events trigger pipeline for storing data and not throwing it away
Numbers	• O(1,000s)	• O(100s)	• O(1,000s)	• 1. O(10,000) • 2. O(100)

What value will the envisaged system deliver for them (the whole setup)? What will the system exactly deliver to them?

	ASTRON	NenuFAR	PROCESS	EISCAT
What should EGI do	• Integrate Scientific Analysis Platform with **AAI** and Notebooks • HTC for batch jobs • Payment models? • Hosting platform? • Training to use EGI services • Training for service management • Training infrastructure / environment	• Data transfer and storage mechanisms. Archiving and online access for analysis • AAI and Group Management • Scientific Analysis Platform – Cloud Container Compute • Interactive analysis platform for pipeline development (Notebooks)	• High-throughput Data transfer • HPC access: schedule jobs from k8s clusters • AAI and SSO to access different systems (HTC, Cloud, HPC) • Cloud container compute (docker + k8s)	• Analysis platform • AAI and SSO to access different systems (HTC, Cloud, HPC) • See left
Value / Gaps			• A way to	

How should they use the system?

ASTRON	NenuFAR	PROCESS	EISCAT
• A user can log on to the SAP, he/she can search for available data, select the data, start up a Notebook / HTC batch processing platform, which can then see the data and he/she can analyse it easily, and publish the data/workflow via Research Object (group of of DOIs for different elements)	• As ASTRON + access to own data and public data. User can choose Notebook / container etc that contains preferred workflow environment, including data. • Need to deliver / fetch data container to workflow	• Pipeline developer logs on to the system and will have access to a db of containers/notebooks (configs) and resources to spin up a system for a end-end user • End-end user will log on to a webUI and select a dataset and pipeline launch the processing and retrieve the result from the webUI	• A user has to log on to a portal, search across different levels of data (events), then uses these data in an interactive analysis platform (Notebooks / HTC), publish the data in a Research Object.

What's the timeline for development, testing and large-scale operation?

(Consecutive releases can/should be considered.)

ASTRON	NenuFAR	PROCESS	EISCAT
Grab from PPTs			

Second break-out

Design and implementation plan

What should the first version include? - The most basic product prototype imaginable already bringing value to the users

(the so-called Minimal Viable Product - MVP)

ASTRON	NenuFAR	PROCESS	EISCAT
• Check-in (AAI) • Notebooks • Online storage for Notebooks (LOFAR) → Onedata/ DataHub?	• Check-in (AAI) • Data (online) storage and archiving (+4PB/year) • Data transfer service between infrastructures (I/O needs?)	• Check-in • Data transfer/staging (16TB) between HTC/HPC sites (want to move data from Poznan to Krakow) • Data storage • Cloud container compute	• Check-in • DIRAC • Data transfer • Notebooks (experimental) + Matlab kernel (NB to sort out license)

- MVPs should be easily scalable (at least think about scaling, i.e. don't make it difficult to scale).

Which components/services already exist in this architecture?

ASTRON	NenuFAR	PROCESS	EISCAT
• Check-in (RCAuth included) • Notebooks • DataHub (Onedata)	• Check-in (needs to talk to eduTEAMS) • DataHub (Onedata) • FTS	• Check-in • FTS • Cloud container compute (k8s) • DataHub (Onedata)	• Check-in • DIRAC (data catalogue) • FTS • Notebooks (Matlab license dependent)
Gaps in components / services?			
• Access to e.g. LOFAR LTA via Check-in (being worked on in EOSC LOFAR CC)	• Capacity allocation (+4PB/yr)	• HPC token translation in Check-in	• Checking out data from DIRAC to make visible in Notebooks and vice versa (→ Onedata)

- DIRAC can manage batch jobs to HTC
- Submit Notebook as a job via DIRAC to HTC

Which components/services are under development (and by who)?

ASTRON	NenuFAR	PROCESS	EISCAT
• Finding data from different communities (ESCAPE). • SAP interface needs to be built (ESCAPE WP5) – python/django framework, OIDC • SAP integration with Check-in, Notebooks (EGI/ESCAPE)	• Data management (NenuFAR) • Reproduce existing workflow configurations in EGI infrastructure (EGI/NenuFAR)	• Configure Check-in on pipeline development platform (PROCESS/EGI) • Some Pipelines under dev (PROCESS) • webUI under dev (PROCESS)	• Tool to populate DIRAC with metadata (EISCAT) • Use cloud services to run containerised applications (EOSC-hub). • Adapt VM environment specific to EISCAT applications (EOSC-hub)
Beyond the MVP			
• Connect to HTC platforms (EGI/ESCAPE) • Publish Research Objects (data + pipeline + publications, assign everything with a DOI)	• Notebooks • Publishing Research Objects	• Scalability • More pipelines	• Publish Research Objects • Notebooks with Matlab

Are there gaps in the EGI service catalogues that should be filled to implement the use case? Which service provider could fill the gap?

ASTRON	NenuFAR	PROCESS	EISCAT

- PRACE for HPC
  - Seamless integration
- Main work is about combining/integrating existing services – there were no major gaps identified…
- How to deal with system heterogeneity?
- Information portal about service providers.
  Up-to-date documentation about service providers and available resources.
  Machine readable specs.
  Users need to know where to submit jobs
  - Available hardware
  - Available soft / middleware
  - HTC technology support
- Pilot framework to help users define requirements (sandbox / fedcloud)
  - Improve documentation
- CI/CD (Continous Integration/Delivery) training

Proposed proof of concept

Baseline infrastructure:

- 2 HTC sites; 2 cloud sites; 1 HPC site
  - Enable federated AAI with Check-in
- Enable Data Transfer Service / DataHub to work with all the sites

Use-case specific extension:

- Configure DIRAC and Jupyter Notebooks as user interfaces
- Connect PROCESS environment
- Configure DIRAC and Notebooks to access the 5 sites:
  - User to discover dataset with DIRAC
  - Check-out test dataset from DIRAC and import them to Notebooks
  - Run exploratory analysis in Notebooks
  - Go back to DIRAC, discover full dataset
  - Run analysis with full dataset in DIRAC or Notebooks

Tasks

- Compute sites who can support us (ideas)
  - EOSC-hub LOFAR CC members
  - PROCESS members: Krakow and Poznan
  - EISCAT_3D VO sites

  Gergely

- Enable federated AAI on HPC (where exactly?)

- Integrate ‘data import/export’ feature into Notebooks (from DIRAC)
  - Enol, Michal, Andrei

- Integrate batch computing back-end with Notebooks
  - Enol (new EGI Task force!)

- Run Notebook code in DIRAC?

+ complete based on Bjorn’s slide #11 ?

NenuFAR

- Storage + Archive is the priority
  - France Grilles is evaluating it