



# ESCAPE

European Science Cluster of Astronomy &  
Particle physics ESFRI research Infrastructures

## WP2 DIOS contribution

to WP3 session at ESCAPE Progress meeting

Paul Millar (on behalf of WP2)



# WP2 overview

- DataLake concept involves **abstracting** data locality.

Data locality is controlled by an agent (**Rucio**), with another agent (**FTS**) orchestrating data movement.

- Domain-specific workflows need to **access** to their data.

Note that workflow code may be domain hosted, rather than in ESCAPE/WP3's gitlab.

- The DataLake concept has the potential to disrupt workflows.

With DataLake, data is handled differently from ad-hoc data copying.

- Therefore, we need to **verify** that workflows continue to work when data is placed in the DataLake.

- This requires:

- A testbed DataLake, within which data may be **stored**.
- The ability to **deploy** domain-specific workflows.
- The ability to **run** workflows.



# WP2 overview

- DataLake concept involves **abstracting** data locality.

Data locality is controlled by an agent (**Rucio**), with another agent (**FTS**) orchestrating data movement.

- Domain-specific workflows need to **access** to their data.

Note that workflow code may be domain hosted, rather than in ESCAPE/WP3's gitlab.

- The DataLake concept has the potential to disrupt workflows.

With DataLake, data is handled differently from ad-hoc data copying.

- Therefore, we need to **verify** that workflows continue to work when data is placed in the DataLake.

- This requires:

- A testbed DataLake, within which data may be **stored**.
- The ability to **deploy** domain-specific workflows.
- The ability to **run** workflows.



# Container usage in the LOFAR use-case



# LOFAR processing

- **First steps** of processing (flagging and averaging) happens on central processing
- Next steps: calibration, imaging and post-processing **elsewhere**
  - Have some resources for users to post-process (with software installed)
  - Partially relies on the same set of tools
- This will probably change in the future
  - More processing in remote facilities
- Currently LOFAR software is **one stack** with many dependencies
  - New releases tend to break build scripts
  - A lot of overhead to build even a portion of the code
  - Building takes lots of time (and hacking)
- Working towards
  - Modularity
  - Binary packages



# Containerisation to the rescue?

- Stack needs to be built for central processing anyhow, so why not directly build in **containers**?
  - But since hardware and software management have different responsible admins better to make the boundary as concrete as possible
- Docker is the de facto standard for containers. LOFAR stack isn't exactly a microprocess so it **doesn't really match** the idea of Docker.
- For the current survey pipelines, the setup is a CVMFS volume containing a **Singularity container** with the software
  - Singularity used because Docker is not considered acceptable on multi-tenanted systems
- Also worth mentioning: **KERN suite** which is a collection of radio astronomy software, packaged in Ubuntu and distributed as such, as well as in Docker and Singularity containers.



# Testbed vs Services



# WP2 services

Question: what **services** can WP2 DIOS offer for WP3 service/software catalogue?

There are two main services:

- The ability to store data within the DataLake
- The ability to test workflows against data stored in the DataLake.

Both of these are **potentially useful** for researchers.

However both are **problematic**, for different reasons



# DataLake as a service

DIOS has built up **testbed DataLake**, within which data is currently being populated with LOFAR data.

- The agents like Rucio and FTS are running at CERN,
- Support for running these agents beyond ESCAPE is **unclear**.

From WP3, catalogue registered services should be **sustainable**.

They should not “go away” after ESCAPE ends.

This is likely only possible if the **ESFRI communities** take over responsibility for running the agents (Rucio and FTS)

- Unclear at this stage, when and how this transition will take place.
- Should we delay registering services until this transition has taken place?



# Workflow testing as a service

DIOS is building up the ability to **run workflows**.

We are doing this in order to verify the DataLake concept.

The same service could be used “the other way around”

A researcher could use the service to verify their work-flow works with data stored in the DataLake.

The problematic part is that the computing resources are **volunteered by sites** (“scrounged”) for testbed validation testing.

- Intended specifically to support ESCAPE WP2 testing.
- Unclear how this could be extended to support external researchers.

Also unclear how this service would related to WP5’s Science Analysis Platform.



# Summary

**Containers** are useful for WP2 to validate workflows work with the DataLake.

DIOS has **some services** that may be presented to researchers:

- Storing data in the DataLake,
- Testing workflows against the DataLake concept.

Some **open questions** remain on whether either service, right now, provides something that may be presented within the WP3 catalogue.



Thanks for listening!

