Particles, Universe, NuClei and Hadrons for NFDI

Facilitating FAIR data in Germany

C. Schneide, T. Schörner for the PUNCH4NFDI Consortium

Abstract

PUNCH4NFDI has as its central deliverable a **community-overarching** science data platform (SDP), in which **complex workflows** can be executed on **digital research products** (DRPs) in a transparent, automatised and **FAIR** (Findable, Accessible, Interoperable and Reusable) way. The SDP consists of several ingredients, not least the Compute4PUNCH and Storage4PUNCH federated infrastructures, that are more or less far advanced in their development and the interplay and interfacing of which is now being implemented.

Furthermore, the consortium is working on **coherent metadata** for PUNCH sciences, and providing **training in research data management** on different levels.

In this poster, we will give an overview of ongoing and planned activities by showcasing two example workflows to be executed on the SDP.

Contact

Are you interested in sharing expertise on FAIR data management? **Get in touch!**

www.punch4nfdi.de – info@punch4nfdi.de

Higgs Boson “Rediscovery” using CERN Open Data

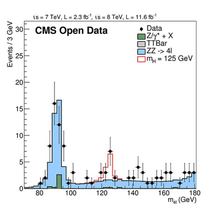
Main purpose: Demonstrate practical feasibility of a use case on the SDP going sgnificantly beyond what is already available outside PUNCH4NFDI (i.e. not just an import of thins already available elsewhere), unsing PUNCH4NFDI resources already now wherever possible.

Original CMS legacy research data (2 PB on CERN / eospublic via CERN Open Data portal)
- 2010 data (100%, legacy format 1)
- 2011/12 (70%, legacy format 2)

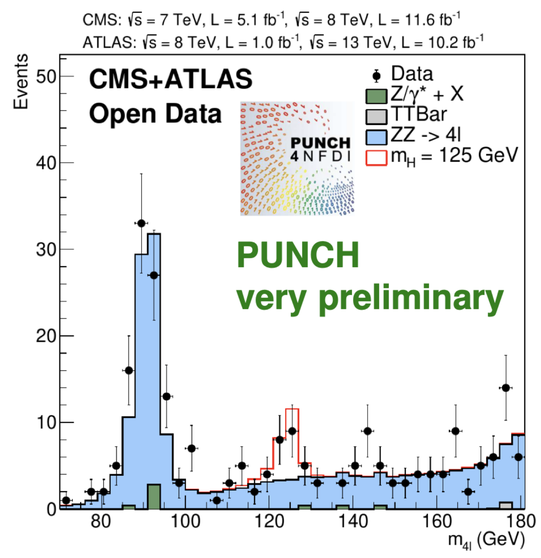
Original CMS legacy software (from public github via CERN Open Data portal)
(2 different versions, run on 2 different legacy VMs / containers)

Apply data transformation interface

Produce histograms many CPU months



Final perspective via the PUNCH4NFDI SDP:
- 76 data samples with unified data format via SDP
- Single script, < 1 CPU day, documentation + metadata

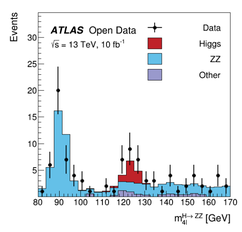


Non-public ATLAS legacy data

Simplified educational data sets 2012 / 2016 via CERN Open Data portal or ATLAS Open Data portal

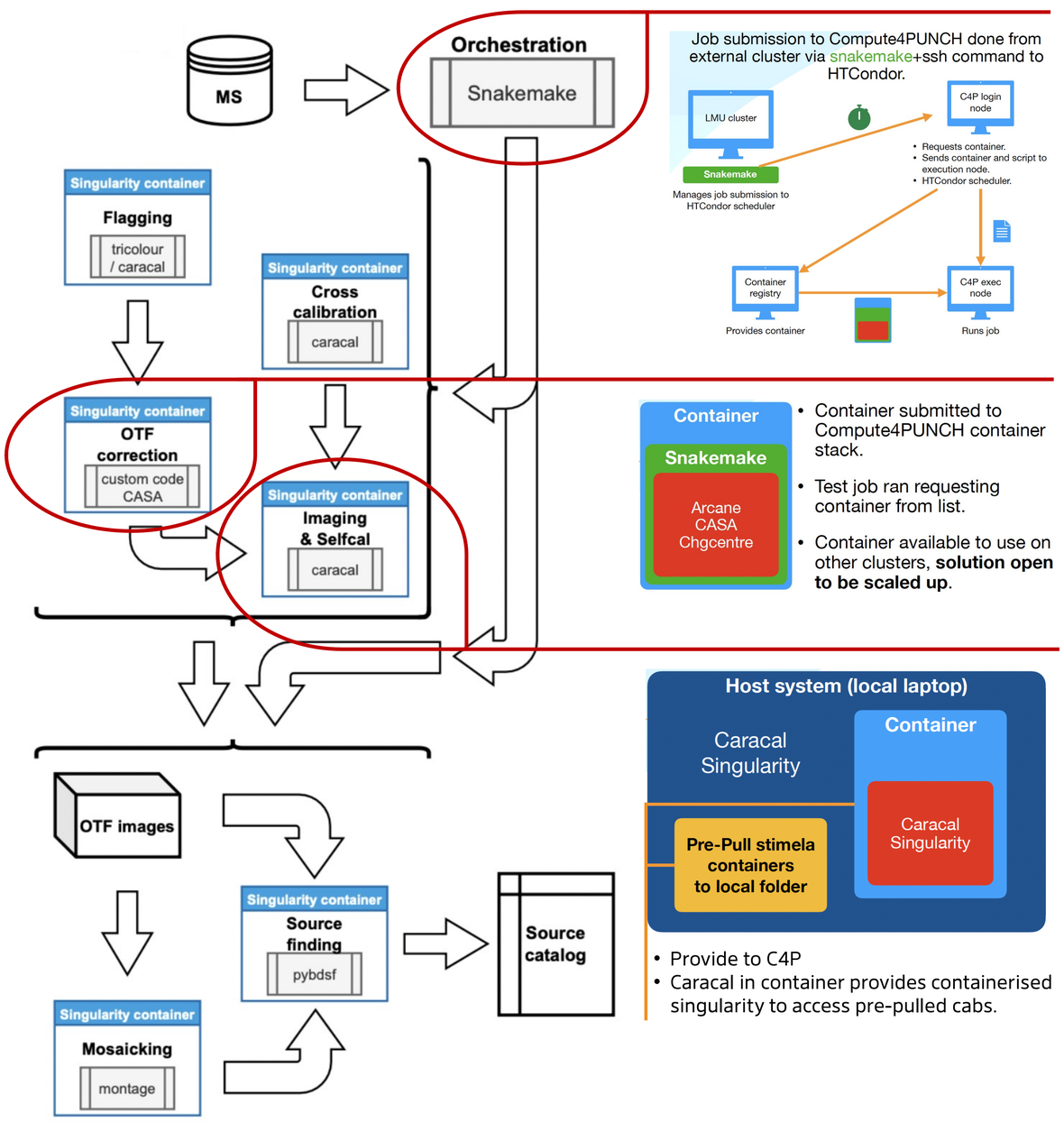
Download to S4P and transform

VM with dedicated software package or Jupyter notebook



Reduction of interferometric MeeerKAT Data – Achievements:

MeerKAT is a multi-instrument radio-astronomy measurement campaign involving massive data volumes, requiring on-the-fly data reduction and generation of radio images. The analysis chain involves numerous software packages. Automation and resources will be dealt with using the PUNCH4NFDI resources C4P and S4P.



Next steps:

- Scale up OTF step to large volumes on C4P, implement imaging step on C4P
- Use S4P to store data (streaming data from workflow)
- Set up orchestration – have entire workflow managed by workflow manager (use REANA instance to steer snakemake workflow)

Further use cases to be implemented soon:

- HiggsTools: provide an easy interface for researchers to combine and interpret various Higgs-related measurements to custom BSM models
- Belle II analysis: provide easy access to data and easy access to software
- Heavy quark diffusion: provide a digital research product including software, data, publications and corresponding metadata

Further activities in PUNCH4NFDI include ...

- Concepts for metadata and dynamical archives in light of irreversible data loss
- Machine learning and data processing on FPGAs
- Training
  - **Past courses**:
    - Software development training: Professional Git and Clean Code
  - **Future courses**:
    - The JupyterHub service: an introduction to the service provided by WWU Münster
    - BAT.jl: introduction to the Bayesian analysis toolkit in Julia
    - Pyhf: introduction to the pure-python implementation of the HistFactory
    - SciTraceWeb: a tool for tracing digital research products, ensuring reusability and reproducability of workflows
- In addition to these courses, we provide support for career planning via a One-to-one mentoring service and courses e.g. on Leadership or project management.

In Zusammenarbeit mit

nfdi Nationale Forschungsdaten Infrastruktur

Gefördert durch

DFG Deutsche Forschungsgemeinschaft

@punch4nfdi@nfdi.social

@punch4nfdi