Contribution ID: **73** Type: **Oral**

Towards an Analysis-Ready, Cloud-Optimised service for FAIR fusion data

Wednesday, 17 July 2024 09:00 (30 minutes)

We present our work to improve data accessibility and performance for data-intensive tasks within the fusion research community. Our primary goal is to develop services that facilitate efficient access for data-intensive applications while ensuring compliance with FAIR principles [1], as well as adoption of interoperable tools, methods and standards.

The major outcome of our work is the successful creation and deployment of a data service for the MAST (Mega Ampere Spherical Tokamak) experiment [2], leading to substantial enhancements in data discoverability, accessibility, and overall data retrieval performance, particularly in scenarios involving large-scale data access. Our work follows the principles of Analysis-Ready, Cloud Optimised (ARCO) data [3] by using cloud optimised data formats for fusion data.

Our system consists of a query-able metadata catalogue, complemented with an object storage system for publicly serving data from the MAST experiment. We will show how our solution integrates with the Pandata stack [4] to enable data analysis and processing at scales that would have previously been intractable, paving the way for data-intensive workflows running routinely with minimal pre-processing on the part of the researcher. By using a cloud-optimised file format such as zarr [5] we can enable interactive data analysis and visualisation while avoiding large data transfers. Our solution integrates with common python data analysis libraries for large, complex scientific data such as xarray [6] for complex data structures and dask [7] for parallel computation and lazily working with larger that memory datasets.

The incorporation of these technologies is vital for advancing simulation, design, and enabling emerging technologies like machine learning and foundation models, all of which rely on efficient access to extensive repositories of high-quality data. Relying on the FAIR guiding principles for data stewardship not only enhances data findability, accessibility, and reusability, but also fosters international cooperation on the interoperability of data and tools, driving fusion research into new realms and ensuring its relevance in an era characterised by advanced technologies in data science.

[1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016) https://doi.org/10.1038/sdata.2016.18
[2] M Cox, The Mega Amp Spherical Tokamak, Fusion Engineering and Design, Volume 46, Issues 2–4, 1999, Pages 397-404, ISSN 0920-3796, https://doi.org/10.1016/S0920-3796(99)00031-9
[3] Stern, Charles, et al. "Pangeo forge: crowdsourcing analysis-ready, cloud optimized data production." Frontiers in Climate 3 (2022): 782909.
[4] Bednar, James A., and Martin Durant. "The Pandata Scalable Open-Source Analysis Stack." (2023).
[5] Alistair Miles (2024) 'zarr-developers/zarr-python: v2.17.1'. Zenodo. doi: 10.5281/zenodo.10790679
[6] Hoyer, S. & Hamman, J., (2017). xarray: N-D labeled Arrays and Datasets in Python. Journal of Open Research Software. 5(1), p.10. DOI: https://doi.org/10.5334/jors.148
[7] Rocklin, M. (2015). Dask: Parallel computation with blocked algorithms and task scheduling. In Proceedings of the 14th python in science conference

Speaker's Affiliation

UKAEA, Abingdon

Member State or IGO

United Kingdom

Primary authors: CUMMINGS, Nathan (UKAEA); Mr JACKSON, Samuel (UKAEA)

Presenter: Mr JACKSON, Samuel (UKAEA)

Session Classification: Data Storage and Retrieval, Distribution and Visulaization

Track Classification: Data Storage and Retrieval, Distribution and Visulaization