Contribution ID: **34**

Type: **Talk**

Enabling high performance DS capacity at the AWI's tier-3 HPC

Thursday 4 April 2024 16:25 (15 minutes)

The recent leaps in Data Science (DS) methodologies present a unique opportunity for many scientific topics to yield progress faster than ever, by leveraging increased capacity of analyzing data. However, major breakthroughs in most sciences are only achievable by combining these DS methodologies with large computing facilities, which have both the large computing capacity and the ability to fastly access and store the large data volumes. Many fields in science are lagging behind industry on the use of DS methods at large scale (e.g. in High Performance Computers). To a certain degree, the reason behind this is that the technical difficulties associated to using DS methods in distributed computing platforms pose a high entry barrier for many scientists who are already capable of training ML models or performing parallel analysis of their data in their local machines.

The HPC & Data Processing group at the Alfred Wegener Institute (AWI) is working to bridge the gap by offering a unified DS infrastructure for accessing and operating AWI's tier-3 High Performance Computer (HPC), Albedo. In this contribution to the Data Science Symposium we will present our approach to tackling the challenge of enabling high performance DS capabilities at our institute. Our strategy includes:

1. The building of modules, environments and containers that include the most-commonly used DS software (e.g. TensorFlow) and software for data processing/postprocessing (numpy, xarray, matplotlib, etc.)
2. The deployment of JupyterHub, with easy access to these environments, and kernels for the most common data processing/postprocessing and DS tools and programming languages, such as Python, R and Julia
3. Enabling Dask (and possibly Dask Gateway) to facilitate the execution of parallel data analysis workflows
4. Explore and train users in the utilization of file formats that allow for distributed high-thoughput I/O (for example, Zarr)
5. Train users in the utilization of the GPU resources and integrate GPUs in our interactive infrastructure

Primary authors: Dr GIERZ, Paul (Alfred Wegener Institute); Dr ANDRES-MARTINEZ, Miguel (Alfred Wegener Institute); Dr SILIGAM, Pavan (Alfred Wegener Institute); Dr BASAVA, Seshadri (Alfred Wegener Institute); Dr PINKERNELL, Stefan (Alfred Wegener Institute); Dr HARIG, Sven (Alfred Wegener Institute); Dr FRITZSCH, Bernadette (Alfred Wegener Institute)

Presenter: Dr GIERZ, Paul (Alfred Wegener Institute)

Session Classification: Session 2: Data Quality and HPC for Data Science

Track Classification: Data Quality and HPC for Data Science