Contribution ID: **36** Type: **Presentation**

Data analysis using Jupyter Notebooks for Open Science

Tuesday, 7 May 2019 11:10 (15 minutes)

The Jupyter Notebook and the Jupyter ecosystem provide a computational and data research and exploration environment, with great potential for open science, FAIR data and the European Open Science Cloud (EOSC).

Research facilities, such as those using Photons and Neutrons for imaging [1] have started to use these capabilities to support the data analysis for their users. Increasingly, the recorded data sets are so large that they cannot be 'taken home' by scientists after having visited the research facilities to record data. Remote data analysis is becoming more important: the data stays at the computing center of the facility, and analysis can be carried out, for example, through ssh and X forwarding. JupyterHub offers a technically attractive alternative for remote data analysis.

As part of the PaNOSC project [1], we propose to use the Jupyter notebook to allow remote exploration and analysis of data sets on the EOSC. We hope to demonstrate this for data from Photon and Neutron facilities, but expect the design to be useful for other types of data sets as well.

We assume that researchers or members of the wider public want to learn from a data set. Using the EOSC portal, they have found the data set, and now want to access it. They have the option to select a particular data analysis procedure from a list of available options that can be applied to the type of data set selected. The data analysis procedure is then provided as a Jupyter Notebook which the user can control and execute. In the simplest case, the pre-formulated data analysis procedure is all that the user is interested in and enables extraction of the meaning of the data immediately. However, there is the possibility to modify, extend and execute the notebook template interactively as usual for Jupyter Notebooks.

A particular use case are reproducible publications based on a data set: for such a data set and publication, we propose to make available data processing commands (ideally in form of a notebook or commands that can be driven from a notebook) that reproduce figures and other key statements in the paper. The combination of the data with the analysis within the proposed framework makes the data set and published research immediately re-usable.

One technical challenge in this proposed design is that the computational environment within which the notebooks execute needs to be preserved and made available on demand. Another technical challenge is that some of the data sets are so large (terabytes and upwards), that it will not be possible to move the data to the notebook server, but rather we need to move the notebook server to the data.

In this presentation, we describe some examples of current use of Jupyter Notebooks, and describe our vision for interactive open science data analysis.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823852

Type of abstract

Presentation

References

[1] Photon and Neutron Open Science Cloud (PaNOSC) https://panosc.eu

Primary author: FANGOHR, Hans (European XFEL)

Co-authors: CAMPBELL, Aidan (ESRF); GOETZ, Andy (ESRF); PELLEGRINI, Eric (ILL); HALL, Jamie (ILL); PERRIN, Jean-François (ILL); KIEFFER, Jerome (ESRF); SELKNAES, Jesper (ESS); WRONA, Krzysztof (EuXFEL); ROSCA, Robert (EuXFEL); BROCKHAUSER, Sandor (EuXFEL); KLUYVER, Thomas (EuXFEL); ROD, Thomas (ESS)

Presenters: FANGOHR, Hans (European XFEL); ROSCA, Robert (EuXFEL)

Session Classification: Jupyter Notebooks