Contribution ID: **22** Type: **Poster**

Large-Scale Analysis of Heterogeneous Earth Observation Data

In recent years, the amount of data made available in the earth-observation domain has increased exponentially. In 2022, for example, data released from the observations of eight Sentinel satellites amounted to 6.64 petabytes [2]. Now, researchers all over the world are using these vast amounts of resources to further improve our understanding of the world. In their journey, the researchers often use code notebooks like Jupyter while conducting their analyses [1]. However, analyses have shown [3] that the code quality in these notebooks is generally mediocre. As analyses grow larger using more, higher-resolution datasets, it is now necessary to find new methodologies for handling the data.

In the past, it was feasible to simply download all the analysis-ready data (ARD), store it locally and run the required analysis. Now, as the amount of data to be processed has increased significantly, this approach of always downloading data to a local machine is not practical anymore. The main approach to address this issue is by pushing the computation to the cloud. Usually, these cloud platforms provide the user with a code notebook interface, an EO dataset catalog, and optionally a web interface for exploring datasets. However, as each pipeline has to contain some code that loads data, improving the pipeline by splitting up the execution pipeline in order to reduce latency and overall resource utilization can have a huge impact on users being able to perform their research.

To solve these issues, we want to create an algebra for Earth Observation queries. Based on this algebra, we will build an optimizer that distributes the workload between the data provider, platform provider, and user in order to minimize the query latency while maximizing efficiency. This is important for several reasons: First, it increases the productivity of an Earth Observation scientist. Second, it decreases the amount of resources necessary to complete the workload. And finally, it caters to the FAIR (findable, accessible, interoperable, reusable) data principles as it allows other scientists to validate datasets and perform analyses more efficiently.

References

[1] Caprarelli, G. et al. 2023. Notebooks Now! The Future of Reproducible Research. Earth and Space Science. 10, 12 (2023), e2023EA003458. DOI:https://doi.org/10.1029/2023EA003458.
[2] Castriotta, A.G. 2023. Copernicus Sentinel Data Access Annual Report Y2022. Technical Report #1. European Commission.
[3] Wang, J. et al. 2020. Better code, better sharing: on the need of analyzing jupyter notebooks. Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: New Ideas and Emerging Results (Seoul South Korea, Jun. 2020), 53–56.

Type of Poster

A challenge

Primary author: DUSELLA, Gereon (DIMA@TU Berlin)

Co-authors: Dr PANDEY, Varun (DIMA@TU Berlin); Prof. MARKL, Volker (DIMA@TU Berlin)

Presenter: DUSELLA, Gereon (DIMA@TU Berlin)

Session Classification: Poster

Track Classification: Poster