The Journal of Open Source Software

DASF: A data analytics software framework for distributed environments

Daniel Eggert[1], **Mike Sips**[1], **Philipp S. Sommer**[2], **and Doris Dransch**[1]

1 Helmholtz Centre Potsdam - GFZ German Research Centre for Geosciences **2** Helmholtz-Zentrum Hereon

DOI: 10.21105/joss.04052

Software

- Review
- Repository
- Archive

Editor: Martin Fleischmann

Reviewers:

- @uellue
- @sk1p
- @cjwu

Submitted: 17 December 2021
Published: 14 February 2022

Summary

The success of scientific projects increasingly depends on using data analysis tools and data in distributed IT infrastructures. Scientists need to use appropriate data analysis tools and data, extract patterns from data using appropriate computational resources, and interpret the extracted patterns. Data analysis tools and data reside on different machines because the volume of the data often demands specific resources for their storage and processing, and data analysis tools usually require specific computational resources and run-time environments. The data analytics software framework `DASF`, which we develop in Digital Earth (Digital-Earth-Project (2021b)), provides a framework for scientists to conduct data analysis in distributed environments.

Statement of need

The data analytics software framework `DASF` supports scientists to conduct data analysis in distributed IT infrastructures by sharing data analysis tools and data. For this purpose, DASF defines a remote procedure call (RPC, White (1976)) messaging protocol that uses a central message broker instance. Scientists can augment their tools and data with this protocol to share them with others. `DASF` supports many programming languages and platforms since the implementation of the protocol uses WebSockets. It provides two ready-to-use language bindings for the messaging protocol, one for Python and one for the Typescript programming language. In order to share a python method or class, users add an annotation in front of it. In addition, users need to specify the connection parameters of the message broker. The central message broker approach allows the method and the client calling the method to actively establish a connection, which enables using methods deployed behind firewalls. `DASF` uses Apache Pulsar (Apache-Pulsar (2021)) as its underlying message broker.

The Typescript bindings are primarily used in conjunction with web frontend components, which are also included in the `DASF-Web` library. They are designed to attach directly to the data returned by the exposed RPC methods. This supports the development of highly exploratory data analysis tools. `DASF` also provides a progress reporting API that enables users to monitor long-running remote procedure calls.

One application using the framework is the Digital Earth Flood Event Explorer (Digital-Earth-Project (2021a)). The Digital Earth Flood Event Explorer integrates several exploratory data analysis tools and remote procedures deployed at various Helmholtz centers across Germany.

The Journal of Open Source Software

Structure

The Data Analytics Software Framework (DASF) facilitates using data analysis tools in distributed IT infrastructures. The framework consists of three major modules:

`DASF-Web` (Eggert (2021b)) collects all web components for the data analytics software framework DASF. It provides ready-to-use interactive data visualization components like time series charts, radar plots, stacked-parameter-relation (spr), and map components to support the visual analysis of spatio-temporal data. Moreover, `DASF-Web` includes the web bindings for the DASF RPC messaging protocol. It is implemented in Typescript and uses Vuejs/Vuetify, Openlayers and D3 as a technical basis.

`DASF-Messaging-Python` (Eggert & Sommer (2021)) is a RPC (remote procedure call) wrapper library for the python programming language. As part of the data analytics software framework DASF, it implements the DASF RPC messaging protocol.

`DASF-Progress-API` (Eggert (2021a)) provides a lightweight tree-based structure to be sent via the DASF RPC messaging protocol. Its generic design supports deterministic as well as non-deterministic progress reports. While `DASF-Messaging-Python` provides the necessary implementation to distribute the progress reports from the reporting backend modules, `DASF-Web` includes ready-to-use components to visualize the reported progress.

Acknowledgements

We acknowledge funding from the Initiative and Networking Fund of the Helmholtz Association through the project `Digital Earth`.

References

Apache-Pulsar. (2021). Apache pulsar: A cloud-native, distributed messaging and streaming platform originally created at yahoo! And now a top-level apache software foundation project. In *Website*. Apache Software Foundation. https://pulsar.apache.org/

Digital-Earth-Project. (2021a). Digital earth flood event explorer. In *Gitlab repository group*. Gitlab. https://git.geomar.de/digital-earth/flood-event-explorer

Digital-Earth-Project. (2021b). Digital earth: Towards SMART monitoring and integrated data exploration of the earth system - implementing the data science paradigm. In *Website*. Helmholtz Gemeinschaft. https://www.digitalearth-hgf.de/

Eggert, D. (2021a). DASF-progress-API: A progress reporting structure for the data analytics software framework. In *Gitlab repository*. Gitlab. https://git.geomar.de/digital-earth/dasf/dasf-progress-api

Eggert, D. (2021b). DASF-web: Web components for the data analytics software framework. In *Gitlab repository*. Gitlab. https://git.geomar.de/digital-earth/dasf/dasf-web

Eggert, D., & Sommer, S. P. (2021). DASF-messaging-python: A python RPC wrapper for the data analytics software framework. In *Gitlab repository*. Gitlab. https://git.geomar.de/digital-earth/dasf/dasf-messaging-python

White, J. E. (1976). RFC 707. A high-level framework for network-based resource sharing. *Proceedings of the 1976 National Computer Conference*.