DISTRIBUTED DATA ANALYSIS FOR BETTER SCIENTIFIC COLLABORATIONS

International Series of Online Research Software Events (SORSE), March 3rd, 2021

Philipp S. Sommer

Helmholtz Zentrum Geesthacht
Institute of Coastal Research, Helmholtz Coastal Data Center

Helmholtz-Zentrum Geesthacht
Zentrum für Material- und Küstenforschung

AWI

JÜLICH
FORSCHUNGSZENTRUM

GEOMAR
Helmholtz-Zentrum für Ozeanforschung Kiel

GFZ
Helmholtz-Zentrum
POTSDAM

HELMHOLTZ
ZENTRUM FÜR UMWELTFORSCHUNG
UFZ

KIT
Karlsruher Institut für Technologie

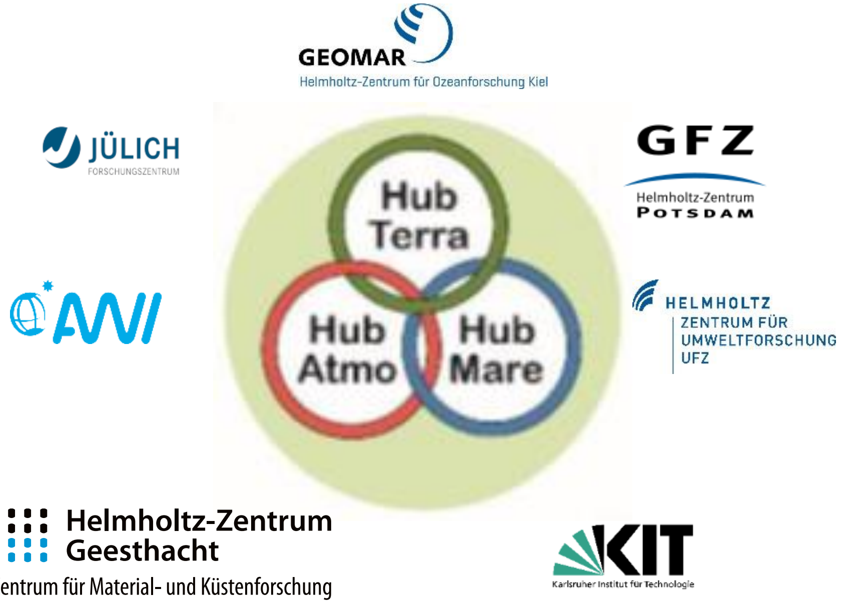
HCDC
Helmholtz Coastal Data Center

AK Datenanalyse

Distributed data analysis Working Group within Datahub

Contributors

- **HZG:** Philipp S. Sommer, Viktoria Wichert
- **GFZ:** Daniel Eggert (Digital Earth)
- **AWI:** Tilman Dinter, Brenner Silva, Angela Schäfer
- **Geomar**: Klaus Getzlaff, Andreas Lehmann
- **KIT:** Christian Werner
- **UFZ:** Lennart Schmidt



What is distributed Data analysis

Examples

Ship campaign

- Sonne (Geomar) and Ludwig Prandtl (HZG) measure real-time-data in a campaign.
- Sonne sends to internal area of Geomar, Ludwig Prandtl to HZG.
- How can people from HZG access and analyze the data at Geomar?

Model simulations

- Compare a COSMO-CLM-Simulation (HZG) with output of the Baltic Sea Model (Geomar)
- And with ship measurements
- How to share terra-bytes of data?
- How to get the latest version?

It’s about *analyzing* distributed data

The ideal world

- We all have one single big cloud
  - Run model simulations in the cloud
  - Store NRT data in the cloud
- Post processing and data analysis runs in the cloud
- Someone from HZG needs access to data from Geomar? *Just grant it.*

The real world

- We have many different clusters.
  - Every center (or even every scientist) has different requirements
  - We are behind VPNs
  - Each center has his own cluster for processing, storage, etc.
- Someone from HZG needs access to data from Geomar? *Ok, I upload it to Dropbox.*

Can we do it without the cloud?

What we need:

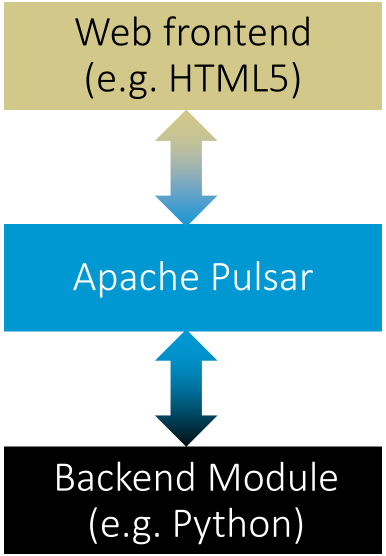
- Access to data in another research center
- Access to computing power in another research center

And:

- It must be safe
- It must be easy

We are not the first

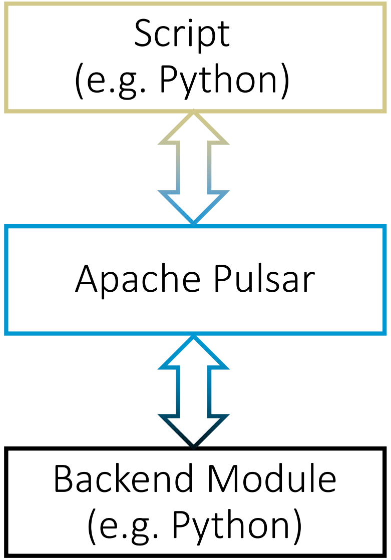
with this idea



Web frontend (e.g. HTML5)

Apache Pulsar

Backend Module (e.g. Python)

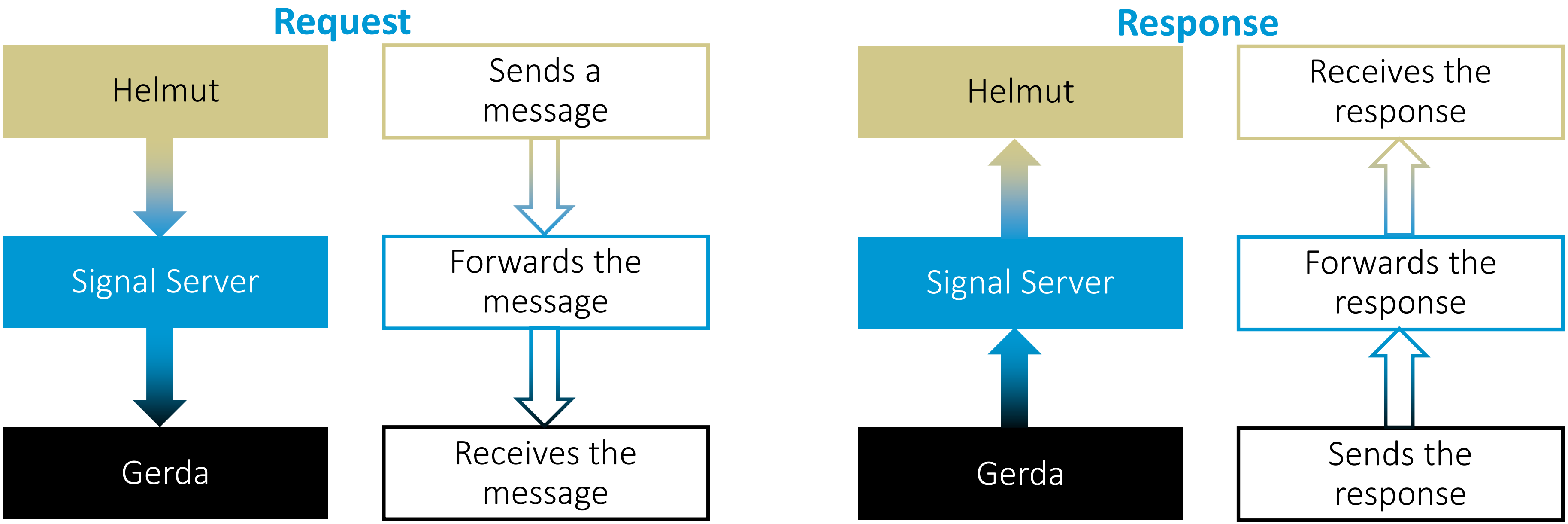


Script (e.g. Python)

Apache Pulsar

Backend Module (e.g. Python)

Just like ~~WhatsApp~~Signal



Request

Helmut → Sends a message

Signal Server → Forwards the message

Gerda → Receives the message

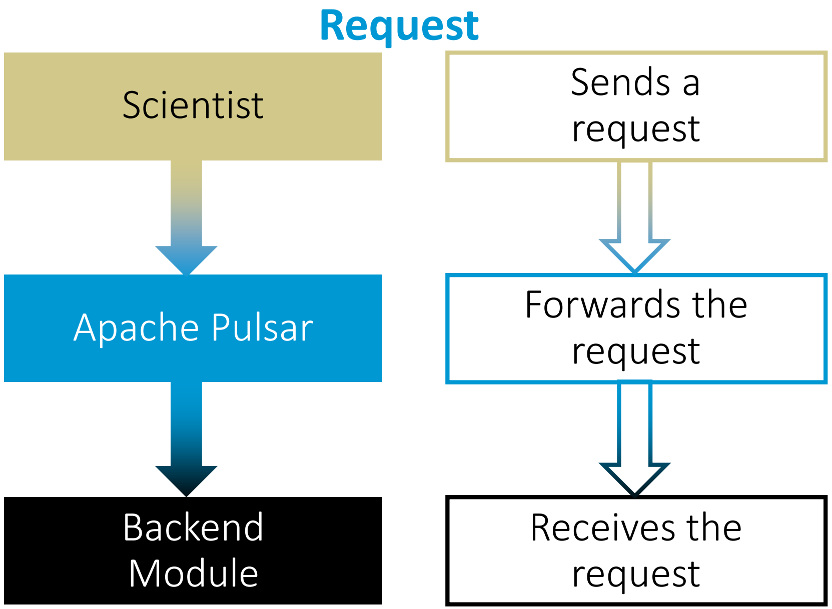
Response

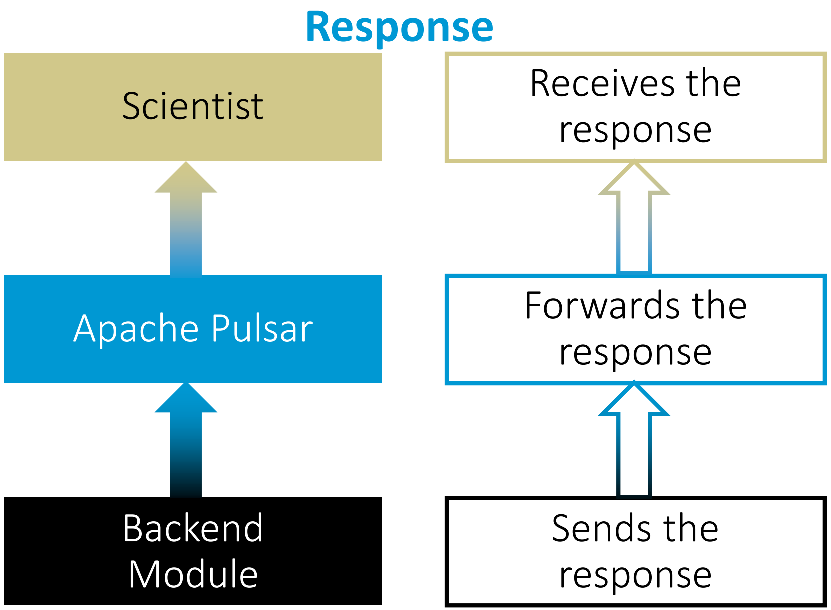
Helmut ← Receives the response

Signal Server ← Forwards the response

Gerda ← Sends the response

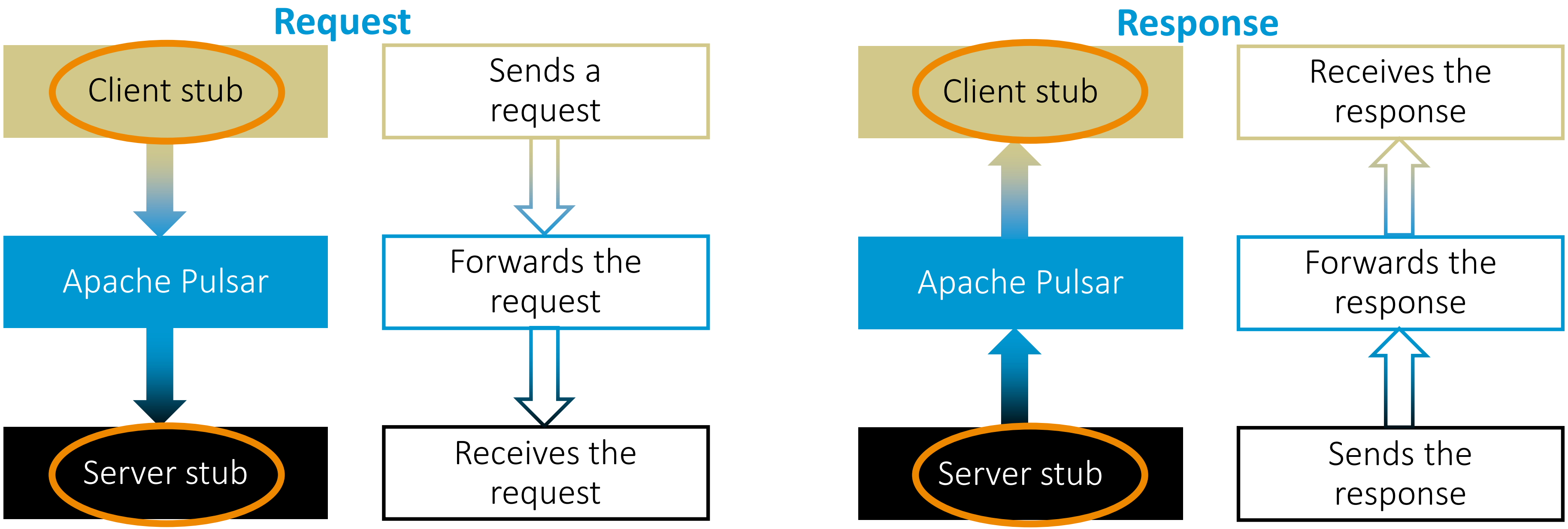
Just like ~~WhatsApp~~Signal





Just like ~~WhatsApp~~Signal

A Remote Procedure Call (RPC)



Pros and Cons

Advantages

- Scientist can simply send a request and retrieve the response on any other machine
- Backend Module can run everywhere, not necessarily on a dedicated web server (e.g. on the cluster)

Disadvantages

- Scientists are not familiar with web requests (nor are the backend module developers)
- Request needs serialization (transformation to JSON)
- Potential vulnerability for internal computing resources

- Scientists do have better stuff to do

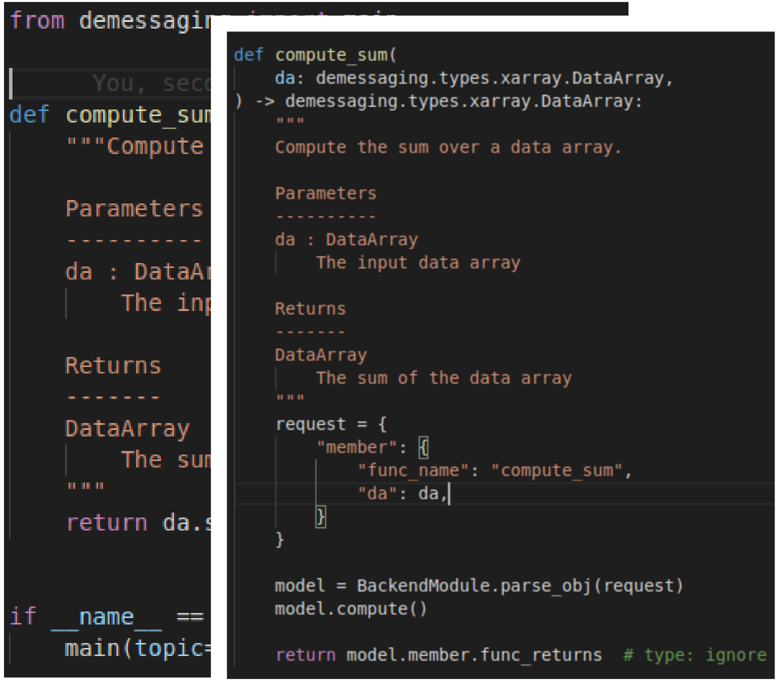
Be nice

and do not add more work

Helmholtz-Zentrum Geesthacht
Zentrum für Material- und Küstenforschung

Use the scientists methods

- abstract standard python functions and classes into web requests
- everything’s basic python, (almost) no need for special stuff
- Client stub is automatically generated
- Requests are abstracted and standardized (JSONschema)



```
from demessagi

    You, seco
def compute_sum
    """Compute

    Parameters
    ----------
    da : DataAr
        The inp

    Returns
    -------
    DataArray
        The sum
    """
    return da.s


if __name__ ==
    main(topic=
```

```
def compute_sum(
    da: demessaging.types.xarray.DataArray,
) -> demessaging.types.xarray.DataArray:
    """
    Compute the sum over a data array.

    Parameters
    ----------
    da : DataArray
        The input data array

    Returns
    -------
    DataArray
        The sum of the data array
    """
    request = {
        "member": {
            "func_name": "compute_sum",
            "da": da,
        }
    }

    model = BackendModule.parse_obj(request)
    model.compute()

    return model.member.func_returns  # type: ignore
```

HCDC Helmholtz Coastal Data Center

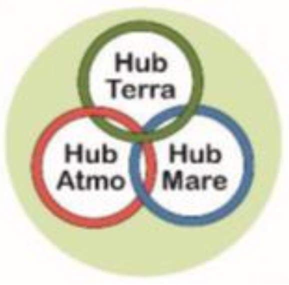
Live Demo

de-messaging-python

Summary

- Remote Procedure Call
- High-level API to easily create server and client stubs
- Very close to scientists common workflows

Thanks you!



Outlook

- More effort into security
  - User management for backends
  - End-to-End encryption
- How to handle large amounts of data
- We are looking for use cases and project that may use our framework!