A Lightweight, Microservice-Based Research Data Management Architecture for Large Scale Environmental Datasets

Alexander Götz, Johannes Munke, Hai Nguyen, Tobias Weber, Stephan Hachinger, Mohamad Hayek and Jens Weismüller

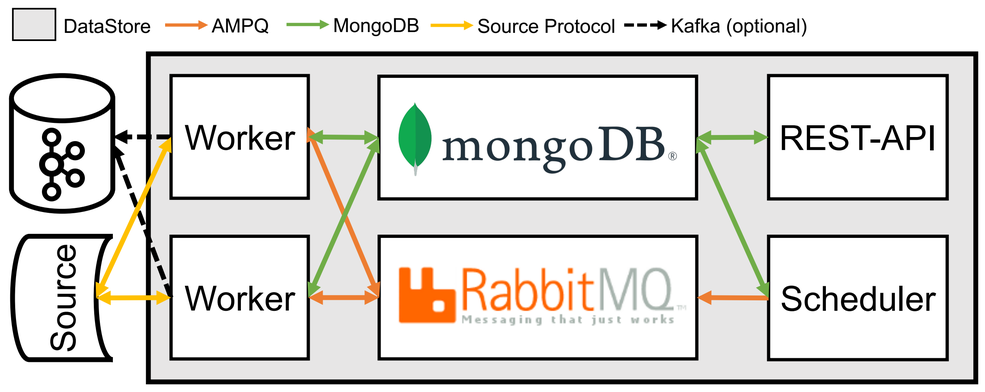
Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, Garching b. München

WHY TO DEVELOP A NEW RDM ARCHITECTURE?

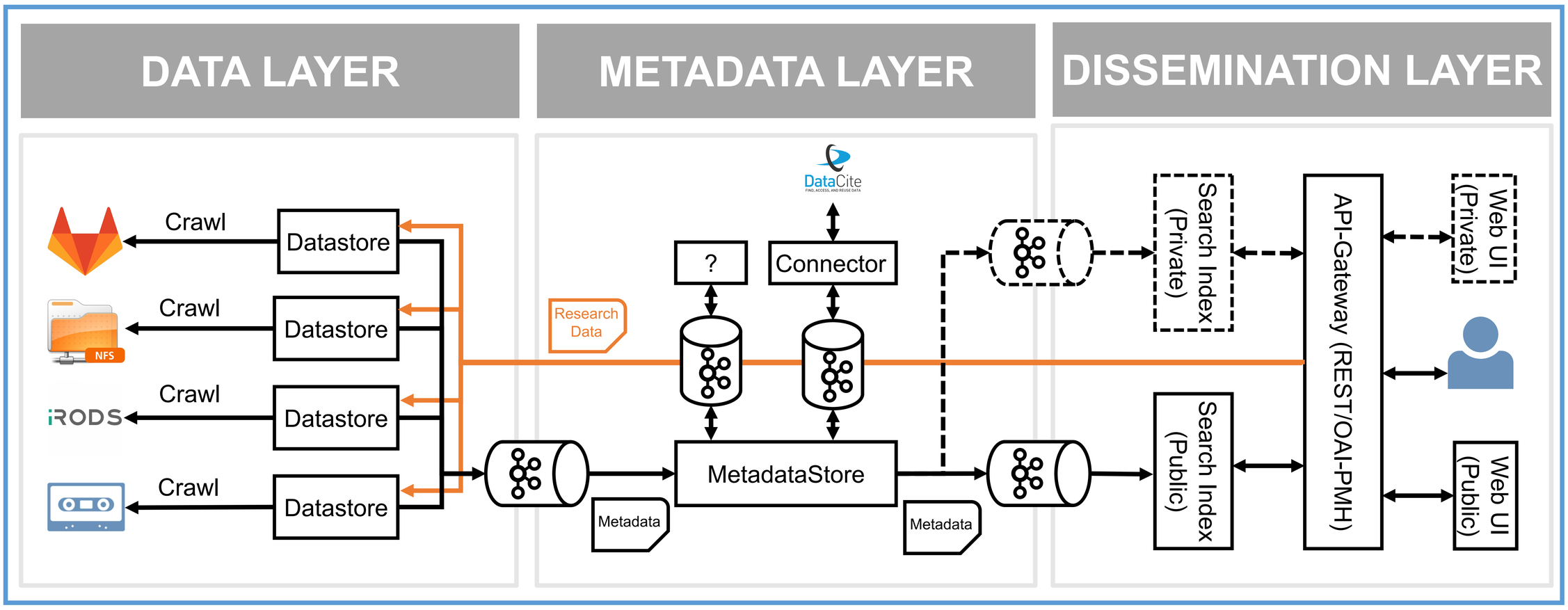
LTDS ("Let the Data Sing") is a lightweight, microservice-based **R**esearch **D**ata **M**anagement (RDM) architecture that makes previously isolated data stores ("data silos") into **FAIR** research data repositories. The core components of LTDS include a metadata store as well as dissemination services such as a landing page generator and an OAI-PMH server. As these core components were designed to be independent from one another, a central control system has been implemented, which handles data flows between components. LTDS is developed at **LRZ** (Leibniz Supercomputing Centre, Garching, Germany), with the aim of allowing researchers to make massive amounts of data (e.g. HPC simulation results) on different storage backends FAIR. Such data can often, owing to their size, not easily be transferred into conventional repositories. As a result, they remain "hidden", while only e.g. final results are published - a massive problem for reproducibility of simulation-based science.

MAKING DATA SILOS FAIR THE LTDS DATASTORE

- Connects (any) data source with the LTDS system
- Crawls data sources for data chunks, containing metadata
- Metadata can be defined by user or extracted automatically
- Data chunk indicated by existence of metadata file
- Core is an **E**xtract, **T**ransform and **L**oad (ETL) Process
- ETL process easy to extend due to modular structure
- Usually one DataStore service per data source
- Number of workers allows for horizontal scaling
- Application is available as Docker container



LTDS MICROSERVICE ARCHITECTURE



DEVELOPMENT STATUS AND EARLY ADOPTERS

LTDS is under active development

- LTDS is developed as a collection of independent microservices, which can communicate over a publish/subscribe model
- Services are equipped with REST-APIs for control and can publish and/or receive messages over dedicated Apache Kafka topics
- The Apache Kafka Message Broker allows the LTDS system to scale to millions of metadata sets that can be processed on a daily basis
- Data processing pipelines in the DataStore and MetadataStore consist of asynchronous task pipelines handled by the Celery framework.

ClimEx

- PI: Prof. R. Ludwig, LMU
- Web: www.climex-project.org
- Funded by the Bavarian Ministry of Environment & Consumer Protection
- One of the largest regional climate-simulation datasets (Bavaria/Quebec)

Requirements on LTDS:

- Indexing, annotation and publication of a 400TB dataset.

ViWA

Virtual Water Values

- PI: Prof. W. Mauser, LMU
- Web: viwa.geographie-muenchen.de
- Funded by the German Federal Ministry of Education & Research

Requirements on LTDS:

- >300 TB of hydrology simulation data
- Portal for browsing and downloading geo-selected data from the dataset

Alpine Environmental Data Analysis Centre (AlpEnDAC)

- PI: Prof. M. Bittner, DLR/UAU
- Web: www.alpendac.eu
- Funded by the Bavarian Ministry of Environment & Consumer Protection
- Platform with On-Demand Computing & Data Analytics facilities

Requirements on LTDS:

- LTDS system at LRZ as backend service
  - Minting of persistent identifiers (DOIs)
  - Dissemination to EUDAT and GeRDI

Alexander Götz: goetz@lrz.de
LTDS DEV-Team: rdm@lists.lrz.de

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften