Hierarchical Data Format for Water-related Big Geodata (HDF4Water)

Hao Li[1], Martin Werner[1]

[1]Chair of Big Geospatial Data Management, Department of Aerospace and Geodesy, Technical University of Munich

13 May 2022, required funding: 6 months

Abstract

Humans rely on clean water for their health, well-being, and various socio-economic activities. To ensure an accurate, up-to-date map of surface water bodies, the often heterogeneous big geodata (remote sensing, GIS, and climate data) must be jointly explored in an efficient and effective manner. In this context, a cross-platform and rock-solid data representation system is key to support advanced water-related research using cutting-edge data science technologies, like deep learning (DL) and high-performance computing (HPC). In this incubator project, we will develop a novel data representation system based on Hierarchical Data Format (HDF), which supports the integration of heterogeneous water-related big geodata and the training of state-of-the-art DL methods. The project will deliver high-quality technical guidelines together with an example water-related data repository based on HDF5 with the support of the BGD group in TUM, with which the NFDI4Earth will consistently benefit from this incubator project since the solution can serve as a blueprint for many other research fields facing the same big data challenge.

I. *Introduction*

Water plays a key role in human health, well-being, socio-economic activities, and Sustainable Development Goals (SDGs). In the past two years, the COVID-19 pandemic has demonstrated the substantial importance of hygiene rules, sanitation, and adequate access to clean water for reducing the spread of infectious diseases and preserving the public health of millions.

Recently, the development of deep learning (DL) techniques shows great potential to facilitate up-to-date and large-scale water-related research in Earth System Science (ESS), but also poses substantial challenges, especially when considering multi-sensor and multi-modal geospatial big data, which are all heterogeneous in their nature. Therefore, there is an unprecedented need for an efficient and flexible data representation and paradigm for both raster-based geodata (e.g., Multispectral satellite images), vector-based geodata (e.g., OpenStreetMap (OSM)), or even point-based geodata (e.g., LiDAR and Photogrammetry point clouds), while few satisfactory big data solutions exist so far.

In HDF4Water, we propose a concise and rock-solid data representation system based on the state-of-art Hierarchical Data Format 5 (HDF5), a multi-objects-based format originated from the National Center for Supercomputing Applications (NCSA). The key advantages of HDF5 mainly include: (i) support for heterogeneous data (e.g., any n-dimensional datasets); (ii) portable and easy-to-share, with no vendor or platform lock-in; (iii) cross-platform, from laptops to massively parallel systems; (iv) fast I/O allows for access time and storage space optimizations; (v) keep

metadata with data, streamlining big data lifecycles and pipelines. Moreover, HDF5 can provide seamless support for modern DL libraries (e.g, TensorFlow, PyTorch, etc.) in addition to a self-describing file system of raster, vector, and point-based geodata in a way such that water-related DL models can be directly and efficiently implemented on top of this data representation with limited (if at all) loss of lineage.

HDF4Water aims to bridge between the current relational model of GIS (e.g., spatial types as extensions of spatial database management systems) and the high-performance computing (HPC) domain (e.g., DL, distributed computing) in the water-related ESS research, where immediate memory random access is of importance. Therefore, it is something significantly different from the HDF5-based NetCDF representation, which is optimized for raster data only. As a final result, showcases of storing heterogeneous big geodata for various water-related applications in HDF5 and ZARR (an open-source modern implementation similar to HDF5) format shall be a paradigm for many other ESS research fields, or even outside the NFDI4Earth community, facing the same big data challenge.

II. *Incubator Project description*

Recently, the emergence of heterogeneous big geodata has led to significant performance boosting in the field of water-related research, which facilitates a larger-scale and faster-speed of monitoring surface water than ever before. Global Surface Water Layer (GSWL) demonstrated the promising capability of integrating remote sensing data (e.g., 30m Landsat data) with multi-sensor auxiliary data (e.g., digital elevation models (DEM), glacier data, urban settlement data, etc) in global-scale water mapping. In the meantime, intensive research efforts are dedicated to developing supervised machine learning (ML) methods, especially using deep learning, for accurate surface water mapping, which achieved superior performance than traditional indices-based method. Moreover, early attempt of harvesting open-source GIS data from OSM as training data for DL models can well address the lack of sufficient training labels during supervised learning. However, this big geodata, all of which are heterogeneous and in diverse formats, request distinct converting and processing steps in order to support a joint analysis, which could be time- and effort-consuming in any aspect.

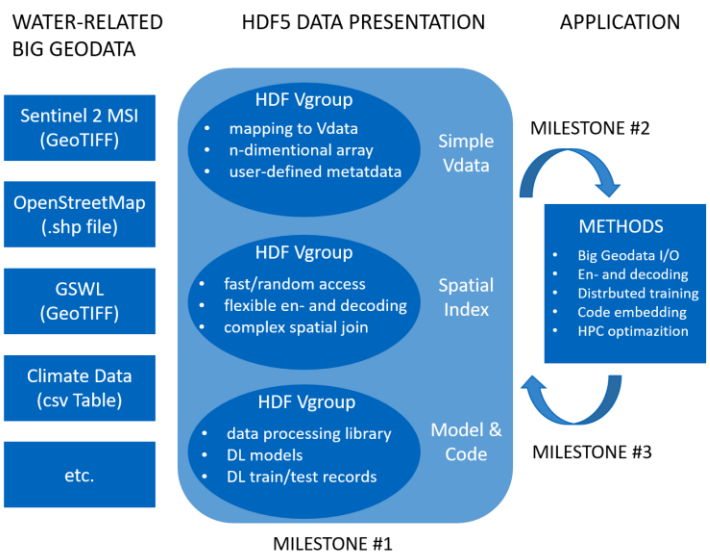

Figure 1: The data representation system based on HDF5

In the proposed Incubator project, we will develop a concise but robust data representation system (c.f. Figure 1) based on open-source technologies (HDF5 or ZARR) to facilitate DL applications with water-related big geodata. For this purpose, we draw on our preliminary work

based on the Copernicus' Sentinel-2 MSI, OSM vector data, GSWL occurrence map in automatic national surface water mapping, and combine them with diverse climate data within the NFDI4Earth repositories to support a broader water-related ESS research in the future. To ensure the transferability, we rely on simplicity as a design factor and will integrate data-related source codes (e.g., scripts, Jupyter notebooks, etc.) and documents (e.g., PDF, markdown, HTML) right into the HDF5 data format bringing a full-fledged, but conceptually simple, data science container format to life. In principle, we want the whole data science project to be captured in a single file.
The HDF4Water project will run as follows: (i) first, by specifying the processing pipeline of water-related big geodata using GIS tool to feed an ImageDataGenerator for DL models trained for automatic water mapping in Germany; (ii) next, we design and implement a concise mapping of heterogeneous data into HDF5 Vgroups (c.f. Figure 1) providing a suitable random access and spatial index, where the source code of data en- and decoding will be embedded together with the raw data file to support future data exploitation in a fully reproducible way; (iii) last, we implement and showcase a distributed training regime (benchmarked on SuperMUC Next Generation) to specify how the source code and documents can be incorporated directly with the heterogeneous geodata. In addition, we plan to provide an example Jupyter-notebook with which this data representation can be decoded into individual files and encoded into an HDF container to enable interfacing with traditional GIS software.

III. *Relevance for the NFDI4Earth*

DL is currently one of the techniques that are more and more adopted in ESS research at large. However, traditional data formats (either raster or vector) are not optimized for the access patterns of deep learning from big geodata, therefore, stakeholders, from almost every stage of the data lifecycle, started to prepare datasets that are very suboptimal in terms of them being spatial datasets (e.g., fragmented, loss of metadata, etc.) or to compromise with comparably slow training and inference performance by relying on libraries from our field that have not been optimized for the workloads typical for DL. In this context, our incubator project aims to provide a single-technology solution for DL with heterogeneous big geodata based on modern HDF5 technology. The choice for HDF5, aside from its excellent features, is also based on the observation that HDF5 is an essential part of, e.g. TensorFlow, also available on most if not all DL libraries.
Though we select water-related big geodata as a case study, the HDF4Water can be regarded as a blueprint for similar research fields, and the lessons learned in this project contribute to the long-term vision of NFDI4Earth to provide researchers with FAIR, coherent, and open access to all relevant ESS data, to innovative big data management and data science methods.

IV. *Deliverables*

A. Technical guidelines and documents (Water-related big geodata in Germany)
    1. Format description for spatial data mapping into groups
    2. Metadata implementation and specification
B. An example data repository based on HDF5 and ZARR
    1. Raw water-related geodata including imagery and GIS data
    2. Complete source code (e.g., processing, DL models, distributed training)
    3. Markdown documentations and tutorials