Contents

1 Project

I have chosen for project proposal 1.

Fighting with h5 files seems to me like a fun challenge. It is definitely not the easiest file format to work with. Luckily the great python ecosystem contains a library for these files but they are inherently complex. The live updates coming in from the government give a good reason to explore DAGs which interests me.

Two days ago the api also started returning .nc files. I have not yet looked as deeply at .nc properly, but it is another binary format that contains metadata and vector data. Expect another version of this document where I go into them more.

The goal of this project is to create a pipeline that takes in data from the pupsub api and combines it into a simple-to-use vector dataset. This dataset should then be easy to used to train machine learning models. Especially transformer models.

2 Quality

2.1 Ease of access

The files in the dataset are not easy to access. You first need to make an account on the knmi.nl website. After this you need to apply to get two different access tokens. If you get accepted you can then use token 1 to subscribe to a stream where kmni publishes new data. An item in this stream contains a link. Once you got this link you can make a request to the link with token 2. In the response of this request you get another link from which you can actually download the data.

I have mostly seen .h5 files. However, since a couple of days I have also seen .nc files. I have not yet had the time to fully investigate the structure of .nc files. Expect an updated version of this submission where this is described more.

2.2 H5 file

To become familiar with H5 I read these resources:

- https://docs.fileformat.com/misc/h5/
- https://github.com/christianversloot/machine-learning-articles/blob/main/how-to-use-h5py-and-keras-to-train-with-data-from-hdf5-files.md
- https://docs.hdfgroup.org/hdf5/develop/_f_m_t3.html#AttributeMessage
- https://docs.hdfgroup.org/hdf5/develop/_f_m_t3.html

I also tried out the online h5 file viewer at https://myhdf5.hdfgroup.org/ which was very helpful for my understanding.

After reading these resources I made the following description of H5 in my own words:

2.2.1 What is H5?

The H5 binary format is like a zip file. A file which contains other files. Files inside an H5 file have two kinds. These two kinds are groups and datasets. Each group and dataset has a name and is located in a certain path in the H5 file.

A group or dataset always contain a bit of json that describes the dataset or group. A dataset is a multidimensional array of a homogeneous type. So for instance a matrix of only integers. These datasets can be loaded as a numpy array.

A group can contain other groups and datasets. Datasets can not contain groups.

Groups and datasets can also contain attributes. Attributes can be used for the metadata of a file, but also as small datasets that are attached to a file. Attributes can contain many different types of data like stings, various types of numbers and more. According to the HDF5 spec the data in a single attribute should not be larger than 64kb.

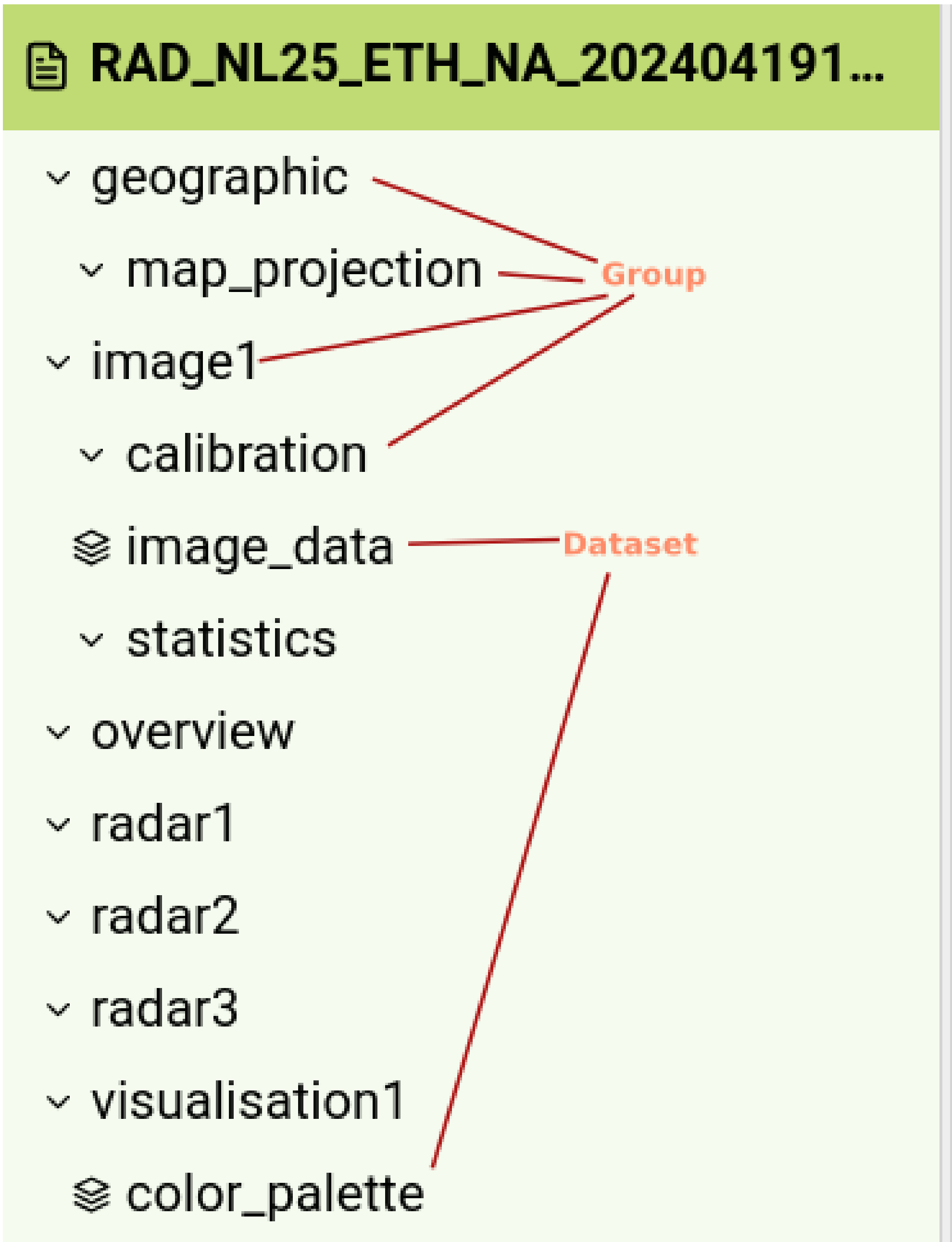

Figure 1: H5 file system

2.2.2 H5 file format quality assessment

The fact that the data is in H5 format is good and bad for the quality of the data. An H5 file should be good for quality because the format can store a massive amount of vectors in such a way that it is efficient to load. However, in practice it is not good for quality because a data scientist has to actually learn about HDF5 files before they can use them. From my experience you need a pretty good understanding of h5 files before you can actually use them. The average data scientist did not learn about H5 files in their education. Thus, the average data scientist will find it difficult and annoying to load this type of file and load the data into a model. There is a trade-off here between speed of loading and convenience.

What also does not help the quality of the data is that the data comes from the government in multiple h5 files.

2.2.3 Nc file format

The NC (NetCDF) file format is a binary data format used to store multidimensional arrays. This format is often used to store climate data along with the metadata that describes how the vector data was collected.

2.2.4 Data quality assessment

The quality of the data itself is high. It comes from scientific measurement tools. The data itself is vectors of climate data with are basically just lists of lists of numbers along with some meta data describing the list of lists of numbers. Description contain units of data like for instance pressure units or temperature in c. There do not seem to be null values. This makes sense because it is mostly just the raw measurements from the weather station sensors distributed over the Netherlands. This makes all parts of the data seems to be usable and in theory very suitable for machine learning models. Machine learning models do well with vectors.

However, accessing the data is hard as I have hopefully described. You first have to connect to the api which is quite some steps and then extract the data from these file formats that I have never seen before which is also not trivial.

3 Improving the Quality

The data is annotated with the source it came from. Making a time series per location sounds like a good idea.

I will improve the quality of the data by making it much easier to load. I plan to do this by making a pipeline that joins the .nc data and .h5 data into an of the shelf modern vector database that is optimized to create machine learning models.

Besides a single large vector database the pipeline shall also transform every file that comes in into a parquet file.