Big Data, organization and analysis

Hierarchical Data Formats

Spring 2023
Steffen Manfred Noe
Emílio Graciliano Ferreira Mercuri (post-doctoral researcher)
Estonian University of Life Sciences (EMÜ)

Outline

After this class you will be able to:

- Explain what the Hierarchical Data Format is.
- Describe the key benefits of the HDF5 format, particularly related to big data.
- Describe both the types of data that can be stored in HDF5 and how it can be stored/structured.

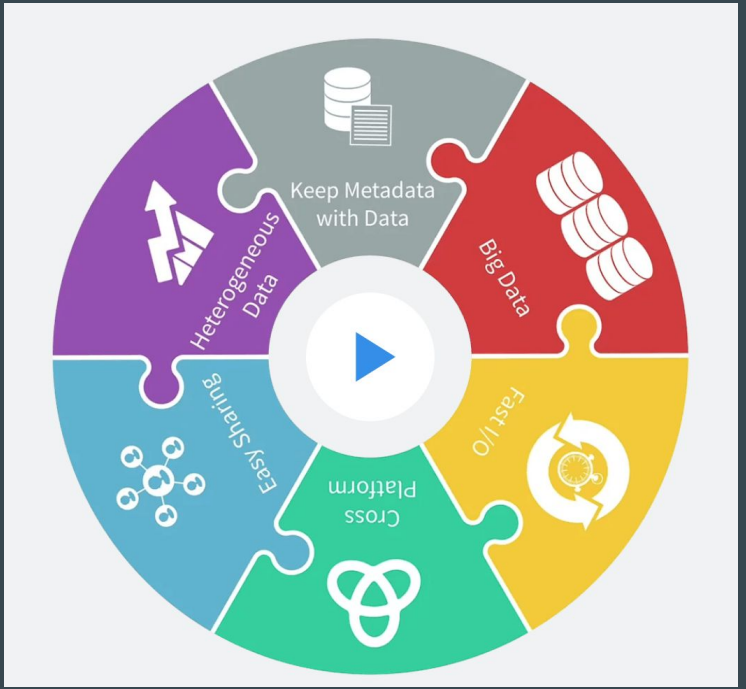
Hierarchical Data Format - HDF5

About Hierarchical Data Format

- HDF5 was created by **The HDF Group**
- HDF5 is an open source file format that supports large, complex, heterogeneous data.
- HDF5 uses a "file directory" like structure that allows you to organize data within the file in many different structured ways, as you might do with files on your computer.
- The HDF5 format also allows for embedding of metadata making it self-describing.

What is HDF5?

- Heterogeneous Data
- Easy Sharing
- Cross Platform
- Fast I/O
- Big Data
- Keep Metadata With Data



HDFGroup - What is HDF5®?
https://www.hdfgroup.org/solutions/hdf5/

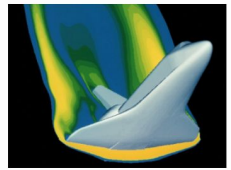
Users of HDF5

Examples of companies that use HDF5:

- NASA
- CERN
- National Institutes of Health (NIH)
- IBM
- Intel
- Google
- The MathWorks

• • •

Astronomy



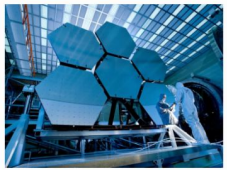
Computational Fluid Dynamics

Genomics

Medicine

Earth Sciences

Engineering



Physics

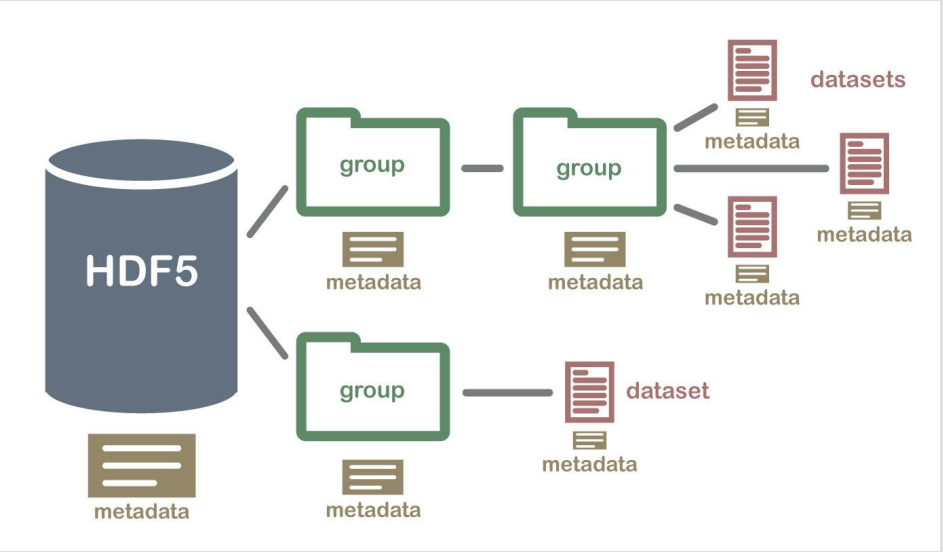
Finance

Hierarchical Structure

A file directory within a file

2 Important HDF5 Terms

- **Group**: A folder like element within an HDF5 file that might contain other groups OR datasets within it.
- **Dataset**: The actual data contained within the HDF5 file. Datasets are often (but don't have to be) stored within groups in the file.

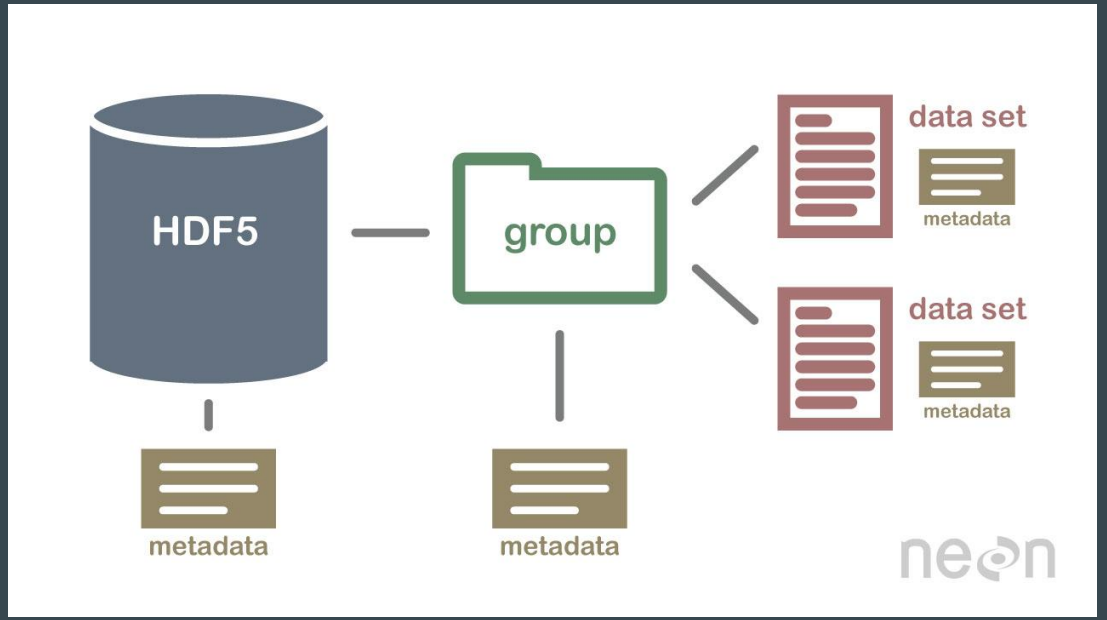


An example HDF5 file structure which contains **groups**, **datasets** and associated **metadata**.

HDF5 is a Self Describing Format

Self Describing Format

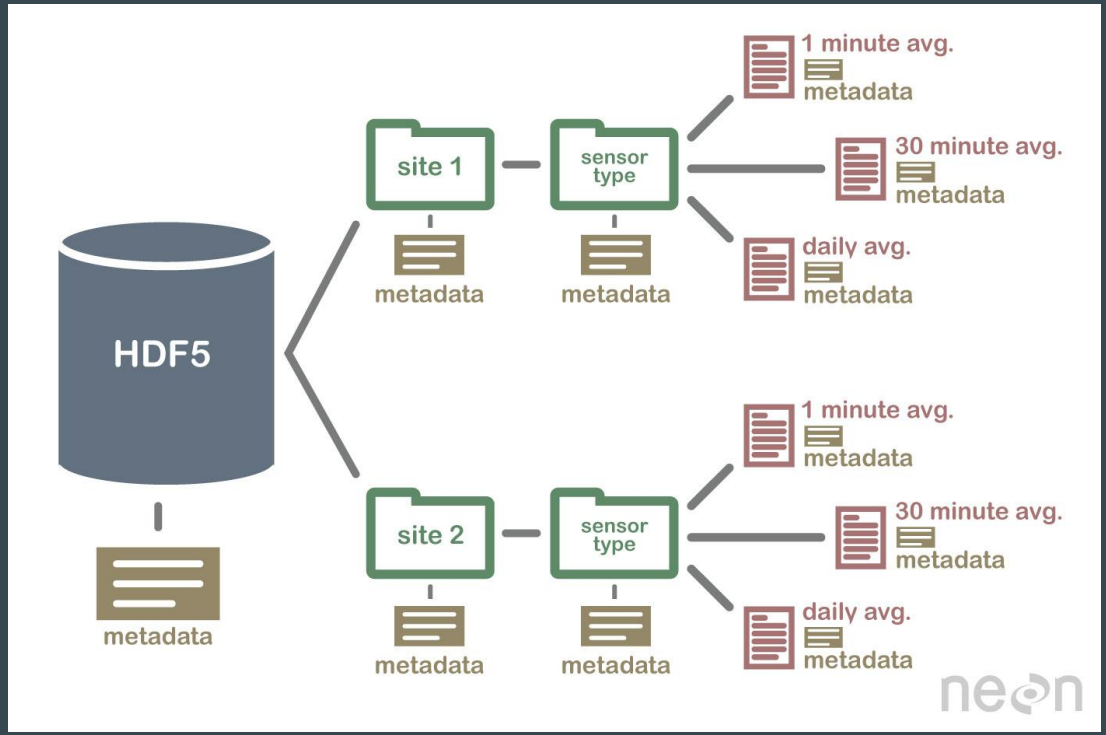
All elements (the file itself, groups and datasets) can have associated metadata that describes the information contained within the element.



National Ecological Observatory Network, or NEON
https://www.neonscience.org/resources/learning-hub/tutorials/about-hdf5

Example of HDF5 file containing datasets

An HDF5 file containing datasets, might be structured like this:



National Ecological Observatory Network, or NEON
https://www.neonscience.org/resources/learning-hub/tutorials/about-hdf5

Compressed & Efficient subsetting

The HDF5 format is a compressed format.

A powerful attribute of HDF5 is data slicing, by which a particular subsets of a dataset can be extracted for processing.

This means that the entire dataset doesn't have to be read into memory (RAM); very helpful in allowing us to more efficiently work with very large (gigabytes or more) datasets!

Heterogeneous Data Storage

HDF5 files can store many different types of data within in the same file.

For example, one group may contain a set of datasets to contain integer (numeric) and text (string) data. Or, one dataset can contain heterogeneous data types (e.g., both text and numeric data in one dataset).

Open Format

The HDF5 format is open and free to use.

The supporting libraries (and a free viewer), can be downloaded from the **HDF Group website.** As such, HDF5 is widely supported in a host of programs, including open source programming languages like R and Python, and commercial programming tools like Matlab and IDL. Spatial data that are stored in HDF5 format can be used in GIS and imaging programs including QGIS, ArcGIS, and ENVI.

Summary Points - Benefits of HDF5

- Self-Describing
- Supports Heterogeneous Data
- Supports Large, Complex Data
- Supports Data Slicing
- Open Format

Let's code!

Google Colab

Next **Thursday** and Next **Monday** there will be no Lecture.

In the end of today's class I will show an assignment (homework).

Next Monday I will be here to answer questions about the assignment.