An Introduction to HDF5

Aug 2023

Glenn Song & Gerd Heber
The HDF Group

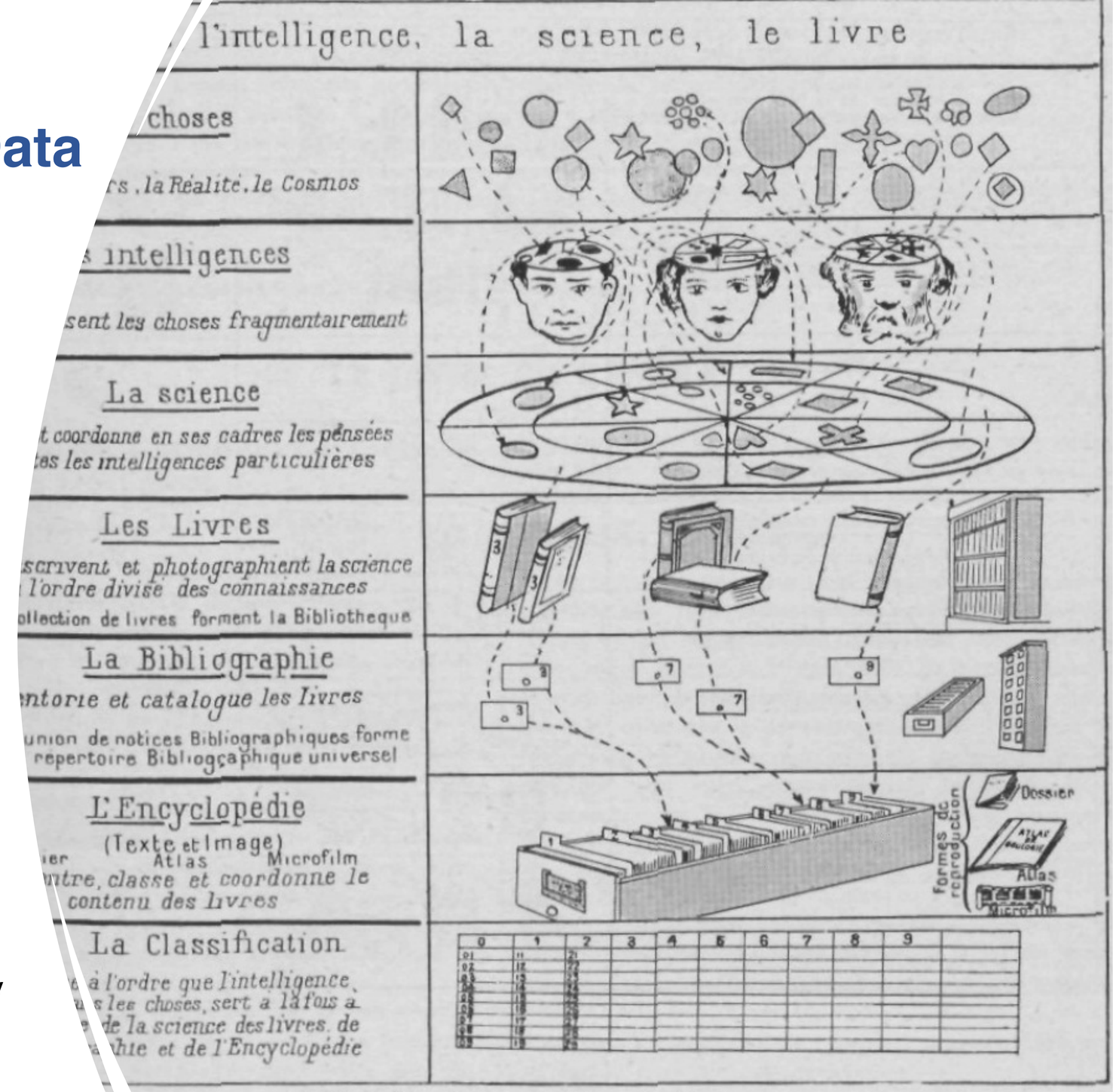
The problem of data

The Problem of Capturing Data

"… there are three realms of interest in data processing: the real world itself, ideas about it existing in the minds of men, and symbols on paper or some other storage medium. The latter realms are, in some sense, held to be models of the former.

…data are fragments of a theory of the real world, and data processing juggles representations of these fragments of theory."

George H. Mealy

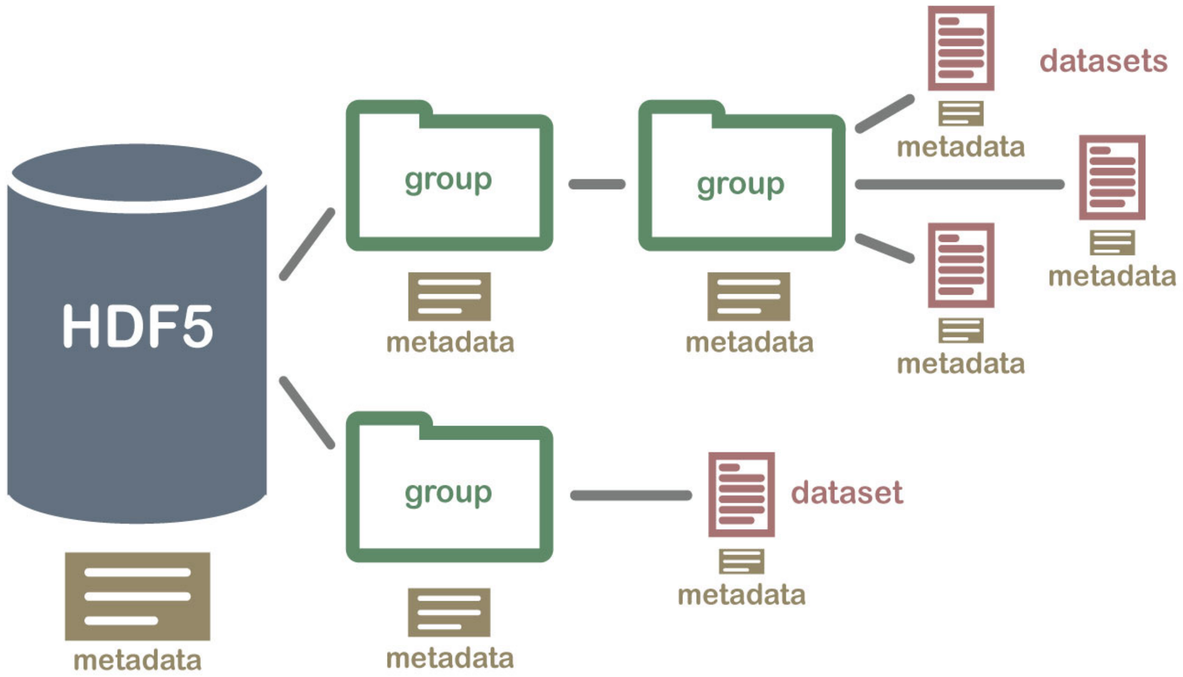


What is HDF5?

What is the HDF5 data model about?

The HDF Group

- It's like a game
  - There are components
  - There are rules
- Three* components
  - (Files)
  - Groups
  - Datasets
  - Attributes
- A few rules (simplified)
  - There must be a root group
  - Groups can be "nested"
  - Datasets must reside "in" 1+ groups
  - Attributes decorate groups and datasets
  - ...



In Practice...

The HDF Group

- Datasets and groups
  - Groups are like directories
  - Datasets are like files
- Attributes
  - Metadata object that describes the data
  - Can be attributed to the dataset or group
- An HDF5 dataset is a uniform multidimensional array of elements
  - Ranging anywhere from common data types (int, float, double, etc.) to more complex user-defined types

How is the HDF5 data model implemented?

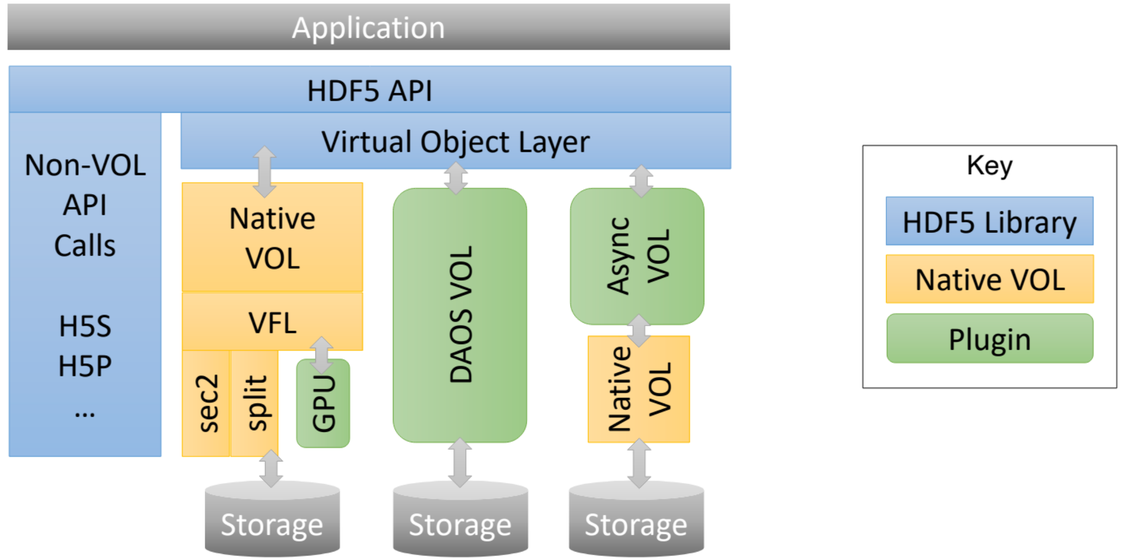
The HDF Group

- As Highly Scalable Data Service (HSDS)
- As a "marriage" of a library and file format (Jeff Kuehn)
- The format is self-describing
  - Data and metadata are together
  - Can be interpreted without "external references"
- The file format specification is public and freely available
- The library is FOSS (Free Open-Source Software) under BSD license
- Has bindings for many different languages
  - C++, Python, Fortran, R, etc.
- Supports many mainstream operating systems
  - Linux, Windows, MacOS, etc.
- Builds with CMake or Autotools

How can the implementation be extended?

The HDF Group

- (File, dataset) Storage layouts
  - Contiguous, chunked, virtual, ...
- Filters
  - Compression, encryption, etc.
  - More than one filter can be applied to a chunk
- Virtual File Drivers
  - Allow users to design and implement their own mapping between HDF5 format address space and storage (Subfiling, Core)
- VOL Connectors
  - Terminal and pass-through connectors (async, REST, DAOS)
  - Can map the physical storage of the HDF5 file and objects to storage depending on what the user needs



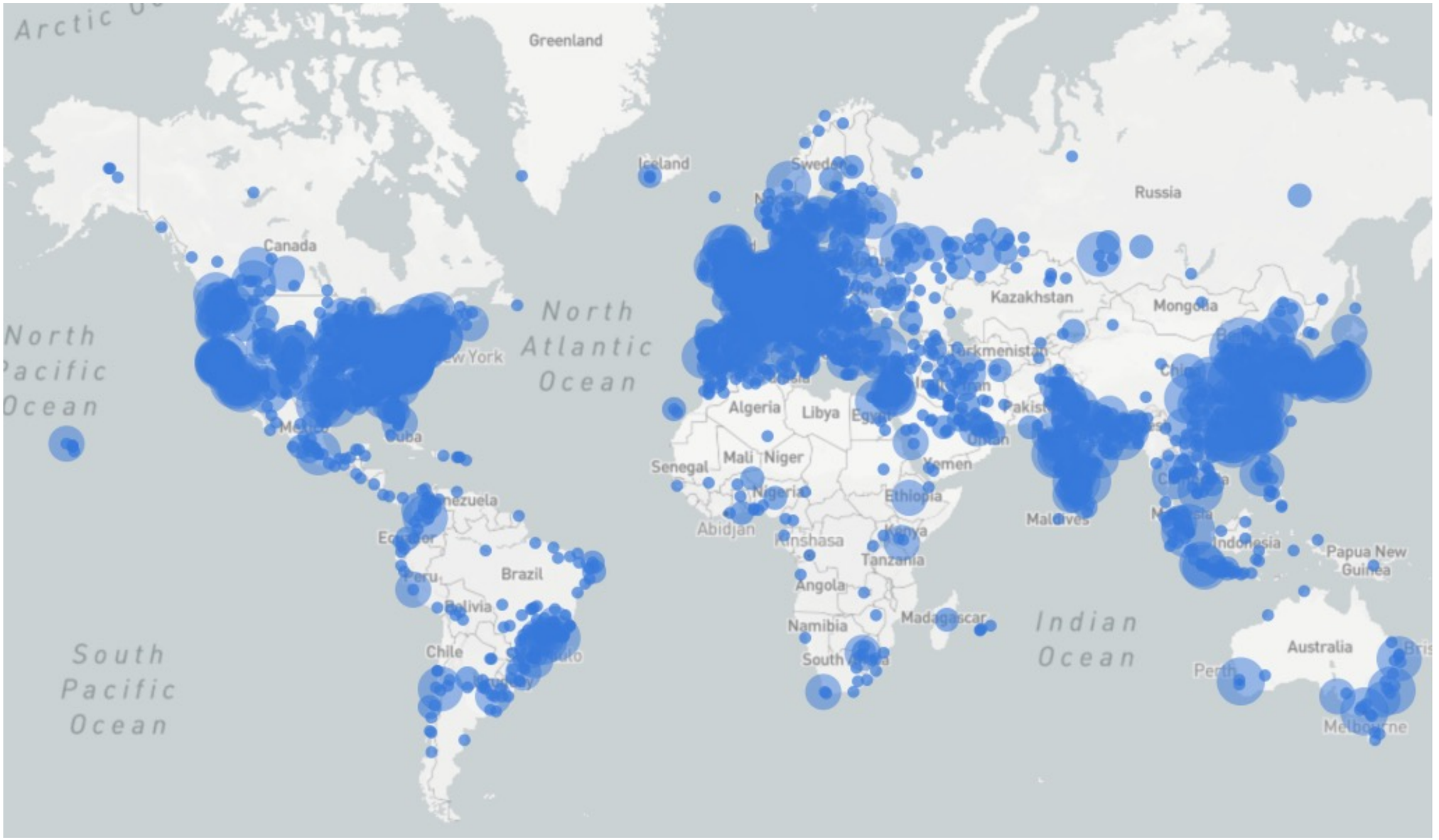
When do you want to use HDF5?

When do you want HDF5?

The HDF Group

- Lots of large, complex datasets
- Heterogeneous data (image, audio, video, time series, …)
- Need fast access and efficiency in dealing with data
- When you want parallel I/O

Who uses HDF5?



The HDF Group

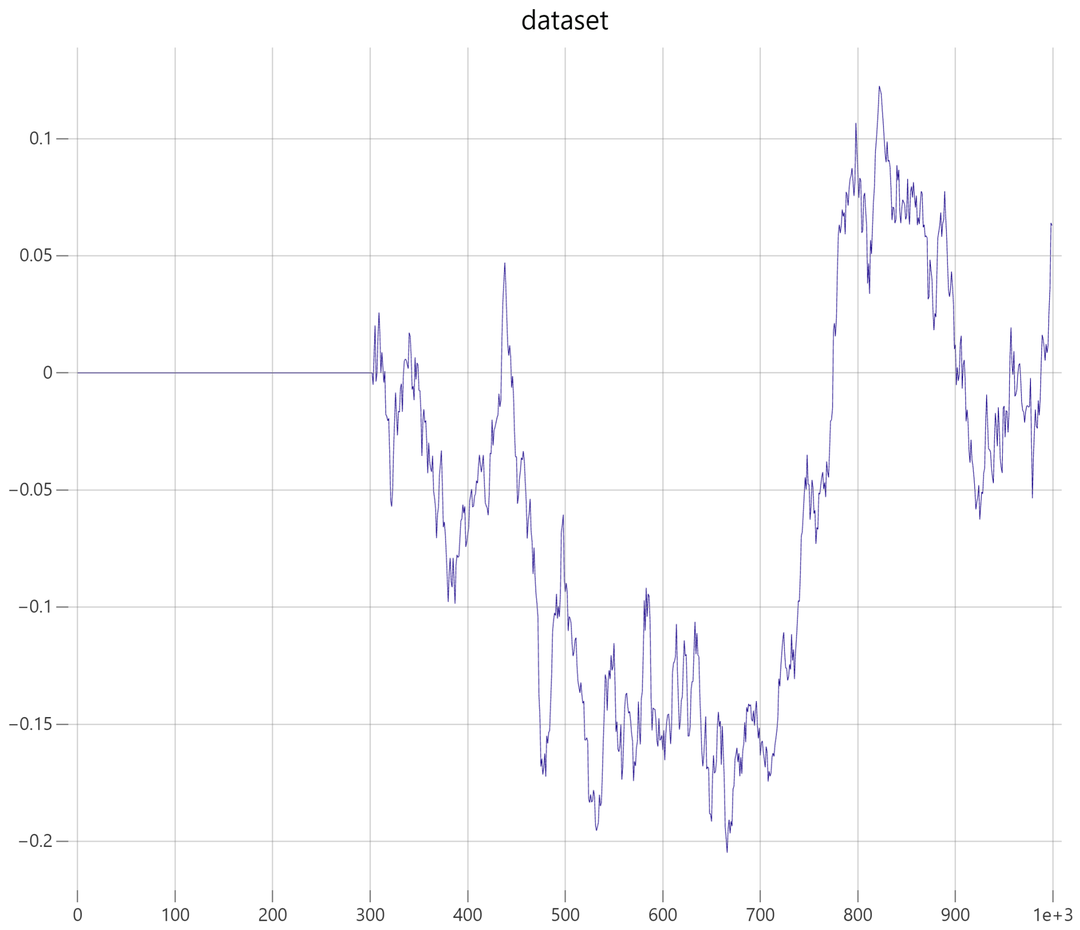
HDF Downloads from unique IP addresses from our website

A common thread for the oncoming tutorials

The Ornstein-Uhlenbeck Process

The HDF Group

- A stochastic process first used as a model for the velocity of a massive Brownian particle under friction
- Mean-reverting, depends on the random variables and inputs used
- We are using this process to generate lots of data to quickly populate HDF5 files and demonstrate the different things HDF5 can do



ChatGPT: A Helpful Tool

ChatGPT as a Starter Tool for HDF5

The HDF Group

- If you are interested in HDF5, but don’t know how to start, try ChatGPT!
- A lot of this tutorial’s code was generated using ChatGPT and then slightly altered to fit together

ChatGPT as a Starter Tool for HDF5

The HDF Group

- Some of the prompts we used:
  - Can you write a C++ program that creates 100 sample paths of an Ornstein-Uhlenbeck process and write that into an HDF5 file using C++?
  - Can you write some code to use command line arguments in a C++ program?
  - Can you create some double attributes in HDF5 using C++?

Visualizing the data

Visualizing the Data

The HDF Group

- First, we can look at the output of the code using a web visualizer tool
  - https://myhdf5.hdfgroup.org/
- We can also use built-in command line tools like h5dump and h5repack
- Or HDFView, which is a popular visual tool for viewing and editing HDF5 files
  - https://www.hdfgroup.org/downloads/hdfview/

Let’s look through some tutorial C++ code

Where to find materials and resources?

The HDF Group

- For ease of access, all the following tutorials will be listed in the following repo: https://github.com/HDFGroup/hdf5-tutorial/tree/main

- Specifically, we will be using this code: https://github.com/HDFGroup/hdf5-tutorial/blob/main/hdf-tutorial.cpp

The HDF Group

Questions?

THANK YOU!

Questions & Comments?

Proprietary and Confidential. © 2016, The HDF Group.