Confluence Overview of a Virtual Dataset

The purpose of this file is to hold a working copy of the overview of the Virtual Dataset feature of HDF5. This overview will appear on an external Confluence page to introduce users to the feature.



Copyright 2014 by The HDF Group.

**All rights reserved.**

For more information about The HDF Group, see [www.hdfgroup.org](http://www.hdfgroup.org).

Contents

1. HDF5 Virtual Dataset 4

2. Revision History 5

# HDF5 Virtual Dataset

With a growing amount of data in HDF5, a new need has emerged to access data stored across HDF5 files using standard HDF5 objects such as groups and datasets without rewriting and rearranging the data.

While the ability to build hierarchical structures across existing HDF5 files has been available in HDF5 for quite some time through the mounting and external link features, the ability to present data stored in several HDF5 datasets and files as a single HDF5 dataset and to access the data via HDF5 APIs without rewriting and rearranging the data does not exist.

To address this, wathe (VDS)

The feature is a logical next step in the development of HDF5 that will enable HDF5 users to access and work with data stored in a collection of HDF5 files using well-known tools and existing HDF5 applications such as h5py, MATLAB, and IDL without changing the way the data is collected and stored.

In what kinds of situations might a virtual dataset be useful?

* Synchrotron centers such as DLS and DESY will be generating and storing terabytes of experimental data per day in HDF5 files. Because of the nature of the experiments and hardware constraints, the data representing, for example, an X-ray image will be stored across different HDF5 datasets in multiple HDF5 files. The whole image could be be accessed by an application without any specific knowledge of where data for each part of the image is stored.
* Climatologists who study and analyze climate variations (temporal changes at a given location) will be able to use the VDS feature to describe and access a “data rode” - data of interest stored in a series of HDF5 files organized by time stamps without rewriting the data into new HDF5 file. The ”data rode” will be accessible as a regular HDF5 dataset via their applications without any special knowledge “coded” into the applications.

For more information about the proposed HDF5 Virtual Dataset feature, see "RFC: HDF5 Virtual Dataset."

# Revision History

Use the table below to hold revision history information.

|  |  |
| --- | --- |
| December 10, 2014 | Editing version 1. |
|  |  |
|  |  |
|  |  |
|  |  |