Stable API's Are Critical to Evolving Data Storage Architectures

Dr. Ray E. Habermann,
Director of Earth Science, The HDF Group,
thabermann@hdfgroup.org

John Readey, Director Tools and Cloud Technology, The HDF Group,
jreadey@hdfgroup.org

Dr. Habermann worked at NOAA's National Geophysical Data Center for twenty-five years leading the development of data management and access systems for a wide-variety of NOAA observations. His experience spans CD-ROMS, scientific data formats and tools, relational and spatial databases, and web services. He is an active participant in development of ISO and Open Geospatial Consortium standards for geographic data. John Readey has worked at Intel from 1997 through 2006 were he developed the Intel Array Visualizer – an application and library for array data visualization. From 2006 through 2014 he worked at Amazon where he developed service-based systems for eCommerce and AWS. They both recently joined The HDF Group to leverage their commitments to long-term data access and understanding.

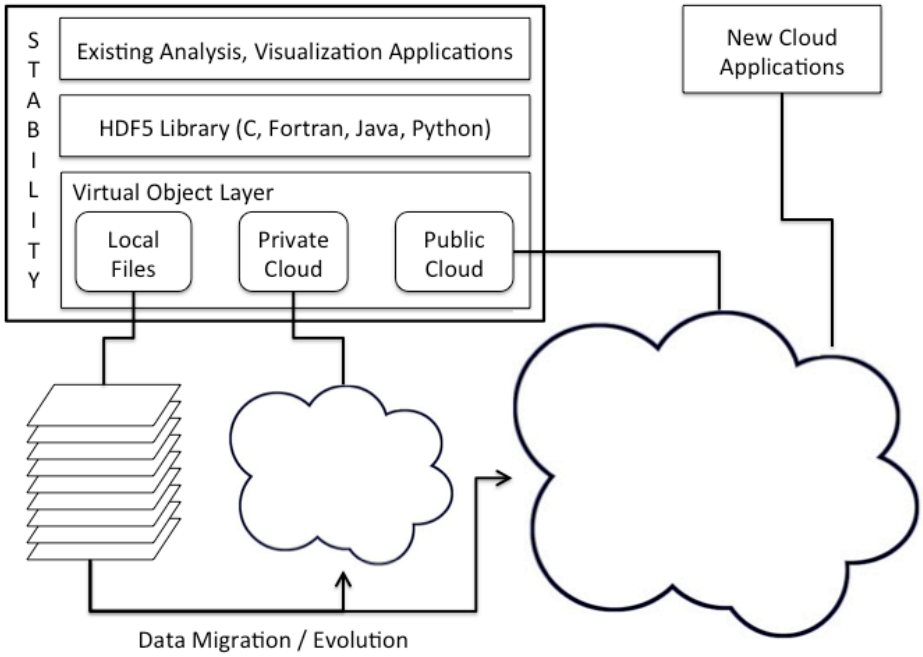
The HDF Group (www.hdfgroup.org) developed and maintained the Hierarchical Data Format that has been used for scientific and engineering data and model results for more than 25 years. HDF is a foundation for standard formats serving scientific and engineering communities in a broad range of disciplines (Habermann et al., 2014). The HDF5 format supports data of any size, shape or source. It also supports the metadata that are critical for understanding the data currently and in the future. HDF is the primary format for all NASA Earth Observations. As netCDF4, it is the primary format for many climate model outputs and NOAA coastal, oceanic and atmospheric data sets. As Bathymetric-Attributed Grid (BAG), it is the standard format for NOAA Hydrographic Surveys.

Technical evolution always involves organizational as well as technical elements. A migration path that preserves existing capabilities during a transition is critical for facilitating progress on difficult organizational change associated with such evolution. The user communities mentioned above, and many others in the U.S. Federal workforce, have well developed tool sets and workflows that are built around data in HDF5, the flagship HDF format. The creators and stewards of these high-value capabilities will resist evolution without a migration path. These people and systems form a significant obstacle to migrating the Federal environmental data infrastructure forward into the Big Data Paradigm.

Creating this migration path is critical to success in this challenging task, but migration path considerations are not included in the Visions and Priority Actions document

The HDF Group has addressed the need for a migration path using an architectural concept termed the Virtual Object Layer (VOL) between the application I/O library and the underlying storage system. It controls how and where HDF5 data objects are actually stored. Analysis and visualization applications using the HDF5 library are shown in the upper left corner of Figure 1. These applications use the HDF5 library, available in many programming languages, to read and write data objects in local HDF5 files. These applications, and their existing API's need to be reliable and stable through the transition to the new paradigm.

As data migrates into the cloud, across the bottom of Figure 1, the VOL provides this stability, allowing applications to read and write data and metadata in any underlying architecture. At the same time, new applications (e.g. Hadoop or other cloud tools) can access the cloud data using API's appropriate for them.



The HDF5 format supports rich metadata associated with any data object in the file. The VOL approach writes/reads that metadata in a way that is consistent with the underlying storage strategy. For example, for a data repository hosted on Amazon AWS it would be useful to store metadata (Attributes, Link names, Types) in DynamoDB (a NoSQL database) while storing dataset values in S3 (an object storage system).

In conjunction with traditional HDF5 applications this approach offers new opportunities that may be effective in expanding access to large data collections:

1. An API can provide access to interactive web based apps for PCs/Tablets/Smart Phones
2. NoSQL databases can be queried directly to effective search for specific attributes
3. Hadoop-based systems can import data directly from the object store
4. Systems such as Apache Solr can be leveraged to provide text based search

Habermann, Ted; Collette, Andrew; Vincena, Steve; Billings, Jay Jay; Gerring, Matt; Hinsen, Konrad; Benger, Werner; Maia, Filipe RNC; Byna, Suren; de Buyl, Pierre (2014): The Hierarchical Data Format (HDF): A Foundation for Sustainable Data and Software. fig**share**. http://dx.doi.org/10.6084/m9.figshare.1112485
Retrieved 00:42, Oct 15, 2014 (GMT)