The Hierarchical Data Format for EOS (HDF-EOS)

Richard Ullman
NASA Goddard Space Flight Center
Greenbelt, MD USA

19th APAN Meetings, Bangkok, Thailand
eScience Workshop, January 27, 2005
HDF is a file format and a software library for science data storage, management, exchange, and archiving
- Highly adaptable, generalized object based data model
- Completely portable file format - read/write on any computing platform
- Supports large datasets, as simple or complex in structure as required
- Designed for high efficiency.
- Runs on virtually any scientific research computing system
- Enables a greater degree of collaboration than any other science data format or library available today.

HDF is written and maintained by the National Center for Supercomputing Applications (NCSA)
- Rigorous design and testing maintain the library and format for continually evolving scientific computing environments.
- A highly stable and talented staff are dedicated to assuring that scientific users of HDF receive world-class support.
• Unlimited size, extensibility, and portability
  - HDF5 does not limit the size of files or the size or number of objects in a file.
  - The HDF5 format and library are both extensible and designed to evolve gracefully with the articulation of new demands.
  - HDF5 functionality and data is portable across virtually all computing platforms used in scientific research and is distributed with C, C++, Java, and Fortran90 programming interfaces.

• General data model
  - HDF5 has a very simple but versatile data model. The HDF5 data model is compatible most competing formats.
  - Through its grouping and linking mechanisms, the HDF5 data model enables complex data relationships and dependencies.
  - HDF5 accommodates the inclusion of many common types of metadata and arbitrary types and quantities of user-defined metadata.
• Flexible, efficient I/O
  – HDF5, through its virtual file layer (VFL), offers extremely flexible storage and data transfer capabilities by means of special-purpose file configurations and powerful I/O mechanisms, including standard I/O, parallel I/O, and network I/O.
  – An application writer can add additional drivers to implement customized data storage or transport.
  – The parallel I/O driver for HDF5 makes it possible to write data in parallel directly to HDF, resulting in improved access times on parallel systems.

• Flexible data storage
  – HDF5 employs various data compression, data extensibility, and chunking strategies to enhance data access, management, and storage efficiency.
  – HDF5 provides for external storage of raw data, often saving disk space and allowing raw data to be shared among HDF5 files and/or applications.
• Unlimited variety of datatypes
  – HDF5 either offers or enables the creation of a virtually unlimited variety of datatypes and imposes no limit on the complexity of a user-defined datatype.
  – Any datatype can be stored in an HDF5 file and shared among other objects in the file, providing a powerful and efficient mechanism for describing data.
  – Datatype storage includes all relevant information, such as endianness, size, and architecture (e.g., IEEE, STD, MIPS).

• Data transformation and complex subsetting
  – HDF5 enables datatype and spatial transformation during I/O operations.
  – HDF5 data I/O functions can operate on selected subsets of the data.
A diagram showing the structure of an HDF file. The root group is labeled as "/". Below it are subgroups and objects:

- "/foo"
- 3-D array
- Raster image
- Palette
- 2-D array
- Table

The diagram also includes a note at the bottom:

January 27, 2005
19th APAN Meetings, Bangkok, Thailand
HDF Storage Options

chunked
Better subsetting access time; extendable

compressed
Improves storage efficiency, transmission speed

extendable
Arrays can be extended in any direction

Split file
Metadata in one file, raw data in another.
Dimensionality: 5 x 3

Datatype: int8, int4, int16, 2x3x2 array of float32

Record

January 27, 2005
19th APAN Meetings, Bangkok, Thailand
Schematic of an NPOESS HDF5 File

File

Secondary Data Attributes

Secondary Data

Root

Data Product

Root Attributes

Data Product Attributes

Data

Aggregated Granules

Granule 1

Granule N

Secondary Data

Data Product

Array Datatype

Array Datatype

Array Datatype

Data Ref 1

Data Ref 2

Data Ref 3

January 27, 2005
19th APAN Meetings, Bangkok, Thailand
Attributes and/or Dimension Scales applied to any field

Global Attribute
<SwathName>::<AttrName>

Group Attribute
<DataFields>::<AttrName>

Local Attribute
<FieldName>::<AttrName>

Data Fields
- Data Field.1
- Data Field.n

Profile Fields
- Profile Field.1
- Profile Field.n

Geolocation Fields
- Longitude
- Latitude
- Time
- CoLatitude

Shaded Objects' representation is fixed by library.
Landsat 7 (ETM+)
Terra (ASTER, CERES, MISR, MODIS, MOPITT)
Meteor-3M (SAGE III)
Aqua (AIRS, AMSR-E, AMSU-A, CERES, MODIS)
Aura (HIRDLS, MLS, OMI, TES)
OrbView 2 (SeaWIFS)
TRMM (CERES, PR, TMI, VIRS)
Quickscat (SeaWinds)
EO-1 (ALI, Hyperion)
ICESat (GLAS)
CALIPSO
NPP (ATMS, CrIS, OMPS, VIIRS)
• Folk, Mike, et al, National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign (UIUC) “HDF5 Nomination for the R&D 100 Award 2002”, February, 2002.

