University Of Florida CISE

HDF and NetCDF

Introductory Understanding Document

Ravishankar M S UFID: 6996-1313

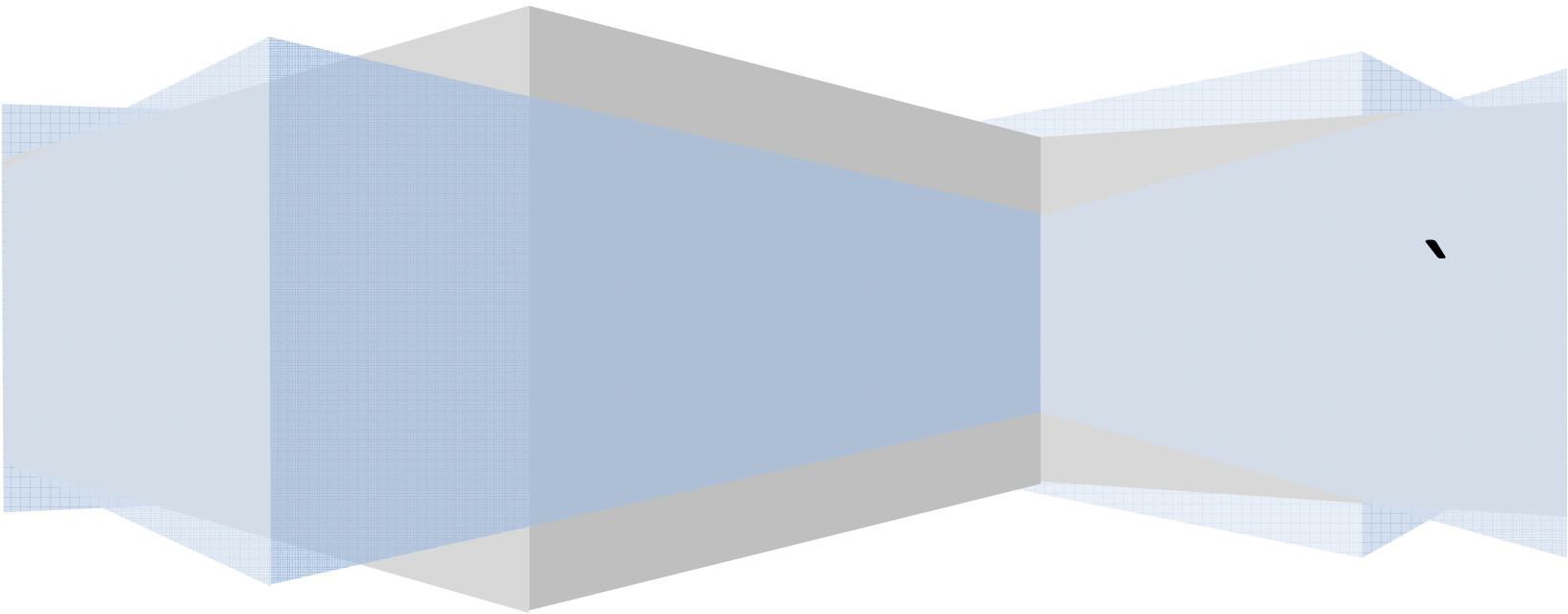


Table of Contents

H

1. Introduction

HDF (Hierarchical Data Format) and NetCDF (Network Common Data Format) Are Not Database Management Systems.

Relational database system is not suitable for complex large (CLOB/BLOB) data access.First, existing database systems that support the relational model do not support multidimensional (arrays) or hierarchical objects as a basic unit of data access. A quite different data model is needed for such data to facilitate its retrieval, modification, mathematical manipulation and visualization.

Related to this is a second problem with general-purpose database systems: their poor performance on large objects. Collections of satellite images, scientific model outputs and long-term global weather observations are beyond the capabilities of most database systems to organize and index for efficient retrieval.

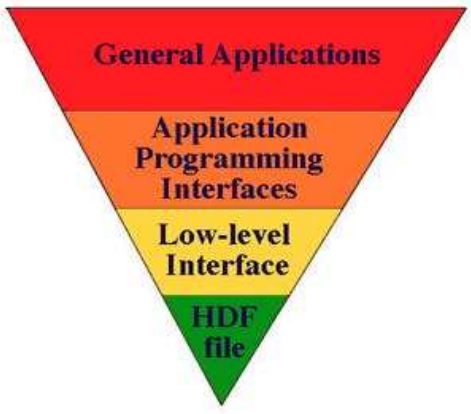
Finally, general-purpose database systems provide, at significant cost in terms of both resources and access performance, many facilities that are not needed in the analysis, management, and splay of array-oriented data. For example, elaborate update facilities, audit trails, report formatting, and mechanisms designed for transaction-processing are unnecessary for most scientific applications.

Thus, these two technologies have evolved to meet following Data/Storage challenges:

- Data is large
- Data is complex
- Data is heterogeneous
- Data is esoteric
- Access needs include parallel I/O
- Access needs include random access

2. HDF File Understanding

2.1 What is HDF?



At its lowest level, HDF (Hierarchical Data Format) is a physical file format for storing scientific data. At its highest level, HDF is a collection of utilities and applications for manipulating, viewing, and analyzing data in HDF files. Between these levels, HDF is a software library that provides high-level APIs and a low-level data interface.

HDF and HDF5 are two different products. HDF is a data format first developed in the 1980s and currently in Release 4.x (HDF Release 4.x). HDF5 is a new data format first released in Beta in 1998 and designed to better meet the ever-increasing demands of scientific computing and to take better

advantage of the ever-increasing capabilities of computing systems. HDF5 is currently in Release 1.x (HDF5 Release 1.x).

HDF5 is a completely new Hierarchical Data Format product consisting of a data format specification and a supporting library implementation.

2.2 HDF5 File Organization:

An HDF5 file is a container for storing a variety of scientific data and is composed of two primary types of objects: groups and datasets.

- **HDF5 group**: a grouping structure containing zero or more HDF5 objects, together with supporting metadata
- **HDF5 dataset**: a multidimensional array of data elements, together with supporting metadata

Any HDF5 group or dataset may have an associated attribute list. **An HDF5 attribute** is a user-defined HDF5 structure that provides extra information about an HDF5 object.

Working with groups and datasets is similar in many ways to working with directories and files in UNIX. As with UNIX directories and files, an HDF5 object in an HDF5 file is often referred to by its full path name (also called an absolute path name).

2.3 Why HDF5?

HDF5 is designed to address some of the limitations of the older HDF product and to address current and anticipated requirements of modern systems and applications. HDF5 is a unique technology suite that makes possible the management of extremely large and complex data collections.

The HDF5 technology suite includes:

- A versatile data model that can represent very complex data objects and a wide variety of metadata.
- A completely portable file format with no limit on the number or size of data objects in the collection.
- A software library that runs on a range of computational platforms, from laptops to massively parallel systems, and implements a high-level API with C, C++, Fortran 90, and Java interfaces.
- A rich set of integrated performance features that allow for access time and storage space optimizations.
- Tools and applications for managing, manipulating, viewing, and analyzing the data in the collection.

2.4 HDF5 Tools and Applications:

Following is a list of few HDF5 Tools and Software.

- HDF5 Utilities: Tools included with the HDF5 distribution
- HDFView: A visual tool for browsing and editing HDF4 and HDF5 files
- HDF Java Products: All of the HDF Java Products, including HDFView and HDF Java wrappers
- h5check: A tool to check the validity of an HDF5 file.
- H4-H5 Conversion Software: A library and tools for convert/h4toh5/ing to and from HDF4 and HDF5.
- HDF Web-browser Plugin (Windows): Extends a web browser to display HDF4 and HDF5 files
- 3rd Party Applications: User applications that read / write HDF5 files

2.5 HDF5 Utilities available for Users:

Following is a list of the HDF5 utilities that are available for users on most platforms supported with HDF5. These utilities are automatically built when building HDF5, and come with the pre-compiled binary distribution of HDF5.

- gif2h5 - Converts a GIF file into HDF5.
- h5import - Imports ASCII or binary data into HDF5.
- h5diff - Compares two HDF5 files and reports the differences. See also
- h5repack - Copies an HDF5 file to a new file with or without compression/chunking.
- h52gif - Converts an HDF5 file into GIF.
- h5cc, h5fc, h5c++ - Simplifies compiling an HDF5 application. [Also see FAQ]
- h5debug - Debugs an existing HDF5 file at a low level.

3 NetCDF File Understanding

3.1 What is NetCDF?

NetCDF (**Network Common Data Form**) is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. The data format is "self-describing". This means that there is a header which describes the layout of the rest of the file, in particular the data arrays, as well as arbitrary file metadata in the form of name/value attributes.

NetCDF data is:

- Self-Describing: A netCDF file includes information about the data it contains.
- Portable: A netCDF file can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- Direct-access: A small subset of a large dataset may be accessed efficiently, without first reading through all the preceding data.

- Appendable: Data may be appended to a properly structured netCDF file without copying the dataset or redefining its structure.
- Sharable: One writer and multiple readers may simultaneously access the same netCDF file.
- Archival: Access to all earlier forms of netCDF data will be supported by current and future versions of the software.

3.2 Why NetCDF?

NetCDF access has been implemented in about half of Unidata's software, so far, and it is planned that such commonality will extend across all Unidata applications in order to:

- Facilitate the use of common datasets by distinct applications.
- Permit datasets to be transported between or shared by dissimilar computers transparently, i.e., without translation.
- Reduce the programming effort usually spent interpreting formats.
- Reduce errors arising from misinterpreting data and ancillary data.
- Facilitate using output from one application as input to another.
- Establish an interface standard which simplifies the inclusion of new software into the Unidata system.

3.3 NetCDF Applications:

It is commonly used in climatology and meteorology applications (e.g., weather forecasting, climate change) and GIS applications. It is an input/output format for many GIS applications, and for general scientific data exchange.

A wide range of application software has been written which makes use of netCDF files. These range from command line utilities to graphical visualization packages.

- A commonly used set of Unix command line utilities for netCDF files is the NetCDF Operators (NCO) suite, which provide a range of commands for manipulation and analysis of netCDF files including basic record concatenating, slicing and averaging.
- NcBrowse is a generic netCDF file viewer that includes Java graphics, animations and 3D visualizations for a wide range of netCDF file conventions.
- The NCAR Command Language is used to analyze and visualize data in netCDF files (among other formats).
- Ferret is an interactive computer visualization and analysis environment designed to meet the needs of oceanographers and meteorologists analyzing large and complex gridded data sets.

3.4 The NetCDF Data Model

A netCDF dataset contains **dimensions, variables,** and **attributes**, which all have both a **name** and an **ID number** by which they are identified. These components can be used together to capture the meaning of data and relations among data fields in an array-oriented dataset. The netCDF library allows

simultaneous access to multiple netCDF datasets which are identified by dataset ID numbers, in addition to ordinary file names.

Groups, like directories in a Unix file system, are hierarchically organized, to arbitrary depth. They can be used to organize large numbers of variables. Each group acts as an entire netCDF dataset in the classic model. That is, each group may have attributes, dimensions, and variables, as well as other groups. The default root is the root group, which allows the classic netCDF data model to fit neatly into the new model.

Dimensions are scoped such that they can be seen in all descendant groups. That is, dimensions can be shared between variables in different groups, if they are defined in a parent group.

4 NetCDF/HDF5 Comparison

One of the goals of netCDF is to support efficient access to small subsets of large datasets. To support this goal, netCDF uses direct access rather than sequential access. This can be much more efficient when the order in which data is read is different from the order in which it was written, or when it must be read in different orders for different applications. The use of HDF5 as a data format adds significant overhead in metadata operations, less so in data access operations.

HDF supports n-dimensional datasets and each element in the dataset may itself be a complex object. Relational databases offer excellent support for queries based on field matching, but are not well-suited for sequentially processing all records in the database or for sub setting the data based on coordinate-style lookup.

NetCDF does not support compression directly but allows users to use HDF5 interface for data compression.

5 Appendix

Document referred:

- http://hdf.ncsa.uiuc.edu/HDF5/index.html
- http://www.unidata.ucar.edu/software/netcdf
- http://en.wikipedia.org/wiki/NetCDF
- http://en.wikipedia.org/wiki/Hierarchical_Data_Format