Projection Indexes for HDF5 datasets

Rishi R Sinha, James Liard, Peter Cao, Quincey Koziol, Mike Folk
The HDF Group
{rsinha, jlaird, xcao, koziol, mfolk}@ncsa.uiuc.edu

Abstract

The increasing sizes of the data being stored in HDF files, necessciates methods for efficient retrieval of a subset of the data. An index datastructure provides a very effective method of increasing efficiency of subset data retrieval. Over the years other groups have come up with wrapper indexing APIs to serve this purpose. Hence a need was felt for a standardized, portable indexing API to be built into the HDF API so that the duplication of effort to create such APIs could be avoided.

This paper describes our attempts at creating a Single Dimensional Projection Index to improve the performance of Range queries over high cardinality attributes. Since most of scientific data is read/append only, a dataset once written is rarely changed. This allowed us to choose Projection Index as our indexing datastracture, as we did not need an easily updatable data structure. Using dataspace to represent the pointer into the dataset allows us to efficiently return subsets of data because of efficient dataspace retrieval in HDF. Our experiments show that the time to retrieve a small subset of the data is greatly reduced by the presence of a Projection Index. Along with the indexing API we have also developed extensions of HDF Viewer so that they can support creation and viewing of the index, and also for querying datasets directly through the viewer.