Data-Intensive Scalable Computing Laboratory (DISCL)

Technical Report

Department of Computer Science

Texas Tech University

Fast Data Analysis with Integrated Statistical Metadata in Scientific Datasets

Jialin Liu and Yong Chen

jaln.liu@ttu.edu, yong.chen@ttu.edu

06/10/2012

Technical Report № TTU/DISCL-2012-06

http://discl.cs.ttu.edu

Fast Data Analysis with Integrated Statistical Metadata in Scientific Datasets

Jialin Liu and Yong Chen

Department of Computer Science, Texas Tech University, Lubbock, TX

Jaln.liu@ttu.edu, yong.chen@ttu.edu

Abstract

Scientific datasets, such as HDF5 and PNetCDF, have been used widely in many scientific applications especially for data intensive scientific discovery and innovations. These file formats and programming interfaces provide efficient access to large volume of data sets. Modern database techniques and parallel I/O have been recently started to integrate into the management of scientific datasets, in which I/O performance and query efficiency are both important criteria. In this research, we analyze how the subsetting partition can affect the access and analysis efficiency of datasets. Based on subsetting extraction, we present a new idea of adding statistical information into the datasets. The statistical information illustrates the data distribution features, and the parallel access code can utilize these metadata to perform fast query. The added metadata may increase the original data size, and we evaluate this trade-off issue though experiments. This research is the first study that utilizes statistical information with different ways of subsetting to perform fast query. It is currently evaluated with the PNetCDF library, but can also be implemented in other scientific data management libraries. The idea we present in this research can lead to a new dataset design and can have an impact on the scientific data management.

Keywords: high performance computing; data intensive computing; big data; statistical techniques; FASM