Inter-comparison of Big Data Technologies for Analysis of Earth Science Data

Kwo-Sen Kuo[1], Thomas L. Clune[2], Daniel Q. Duffy[3], Gyorgy Fekete[4], Rahul Ramachandran[5], John A. Rushing[6]
(alphabetical order after the first author)

Data intensive analytic workflows bridge between the largely unstructured mass of stored scientific data and the highly structured, tailored, reduced, and refined products used by scientists in their research. NASA is operating and planning many missions in space that continuously take Earth science observations. In addition to observations, NASA has a strong engagement in the modeling of geo-physical, -chemical, and -biological processes in order to understand and explain these observations. Both of these efforts generate increasingly large volumes of data.

To meet our Big-Data challenges, we are exploring the use of SciDB to optimize data intensive analytic workflows and provide a preliminary qualitative and quantitative assessment of SciDB as a means of enabling server-side climate data analysis. We carry out two separate yet related experiments, one at NASA Center for Climate Simulation (NCCS) with monthly mean fields of Modern-Era Retrospective Analysis for Research and Applications (MERRA) data and one at Information Technology and Systems Center (ITSC) of University of Alabama-Huntsville (UAH) with daily atmospheric fields derived from Special Sensor Microwave Imager (SSM/I) data.

The experimental cluster at NCCS is an 8-node system running Ubuntu 11.04 Linux operating system. Each node in the cluster has 4 dual-core AMD Opteron 280 processors running at 2.4 GHz and 8GB of RAM memory. Dedicated local disk space for array storage is 870GB per node for a total of approximately 7 TB. Both Hadoop File System (HDFS) and SciDB are installed on the cluster. MapReduce is used in association with HDFS. SciDB has built-in analytic capabilities, which are used to compare to MapReduce. The specific data set being used for the evaluation is 30+ years of MERRA monthly means. To develop performance metrics, we focus on a small set of canonical early-stage analytical operations that represent a common starting point in many analysis workflows in many domains: for example, average, minimum, maximum, and standard deviation operations over a given temporal and spatial extent.

[1] Kwo-Sen Kuo, Kwo-Sen.Kuo@nasa.gov, Caelum Research Corporation, Rockville, Maryland 20850, USA

[2] Thomas L. Clune, Thomas.L.Clune@nasa.gov, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

[3] Daniel Q. Duffy, Daniel.Q.Duffy@nasa.gov, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

[4] Gyorgy Fekete, Computer Science Corporation

[5] John A. Rushing, University of Alabama-Huntsville, Huntsville, Alabama 35899, USA

[6] Rahul Ramachandran, University of Alabama-Huntsville, Huntsville, Alabama 35899, USA

The UAH cluster consists of 70 nodes, each with two single-core processors. Each node has 120 GB of local disk space and 4GB of RAM. The cluster is running Rocks 5.2 (a variant of Linux). The goal of this experiment is to investigate the feasibility of real time queries, which may be used for exploration, visualization and analysis, across the entire data set. Our custom method loads the currently available data sets for all SSM/I platforms to the cluster and stripes them evenly across the compute nodes. An initial set of queries is implemented on the cluster using both the Message Passing Interface (MPI) and a custom daemon-based approach designed to eliminate latency. Queries used for our evaluation purpose include one-dimensional and two-dimensional histograms, threshold queries, and aggregate queries. These queries are performed using both our custom method and SciDB.

In this presentation, we report experiences gained from these experiments, show the results of performance comparisons between SciDB and {MapReduce-HDFS, custom-methodology}, and conclude with lessons learned from these experiments.