Geophysical Research Abstracts
Vol. 14, EGU2012-13008, 2012
EGU General Assembly 2012

Large-Scale, Parallel, Multi-Sensor Data Fusion in the Cloud

B. Wilson, G. Manipon, and H. Hua
JET PROPULSION LABORATORY, Atmospheric Remote Sensing, PASADENA, United States (Brian.Wilson@jpl.nasa.gov)

NASA's Earth Observing System (EOS) is an ambitious facility for studying global climate change. The mandate now is to combine measurements from the instruments on the "A-Train" platforms (AIRS, AMSR-E, MODIS, MISR, MLS, and CloudSat) and other Earth probes to enable large-scale studies of climate change over periods of years to decades. However, moving from predominantly single-instrument studies to a multi-sensor, measurement-based model for long-duration analysis of important climate variables presents serious challenges for large-scale data mining and data fusion. For example, one might want to compare temperature and water vapor retrievals from one instrument (AIRS) to another instrument (MODIS), and to a model (ECMWF), stratify the comparisons using a classification of the "cloud scenes" from CloudSat, and repeat the entire analysis over years of AIRS data. To perform such an analysis, one must discover & access multiple datasets from remote sites, find the space/time "matchups" between instruments swaths and model grids, understand the quality flags and uncertainties for retrieved physical variables, assemble merged datasets, and compute fused products for further scientific and statistical analysis. To efficiently assemble such decade-scale datasets in a timely manner, we are utilizing Elastic Computing in the Cloud and parallel map/reduce-based algorithms.

"SciReduce" is a Hadoop-like parallel analysis system, programmed in parallel python, that is designed from the ground up for Earth science. SciReduce executes inside VMWare images and scales to any number of nodes in the Cloud. Unlike Hadoop, in which simple tuples (keys & values) are passed between the map and reduce functions, SciReduce operates on bundles of named numeric arrays, which can be passed in memory or serialized to disk in netCDF4 or HDF5. Thus, SciReduce uses the native datatypes (geolocated grids, swaths, and points) that geo-scientists are familiar with. We are deploying within SciReduce a versatile set of python operators for data lookup, access, subsetting, co-registration, mining, fusion, and statistical analysis. All operators take in sets of geo-arrays and generate more arrays.

Large, multi-year satellite and model datasets are automatically "sharded" by time and space across a cluster of nodes so that years of data (millions of granules) can be compared or fused in a massively parallel way. Input variables (arrays) are pulled on-demand into the Cloud using OPeNDAP or webification URLs, thereby minimizing the size of the stored input and intermediate datasets. A typical map function might assemble and quality control AIRS Level-2 water vapor profiles for a year of data in parallel, then a reduce function would average the profiles in bins (again, in parallel), and a final reduce would aggregate the climatology and write it to output files.

We are using SciReduce to automate the production of multiple versions of a multi-year water vapor climatology (AIRS & MODIS), stratified by Cloudsat cloud classification, and compare it to models (ECMWF & MERRA reanalysis). We will present the architecture of SciReduce, describe the achieved "clock time" speedups in fusing huge datasets on our own nodes and in the Amazon Cloud, and discuss the Cloud cost tradeoffs for storage, compute, and data transfer.