Session: C13 Poster: T04B

From global observations to using a pure object database to store and manage massive remote sensing data sets

Glen Grant[†]; David Gallaher
[†] National Snow and Ice Data Center (NSIDC), USA
Leading author: glen.grant@nsidc.org

Change over time, especially with respect to climate change, is becoming an increasingly important issue to society. The goal of most climate research is to detect these changes over time, diagnose the underlying causes, and make projections into the future. Those of us in the Earth sciences have been aware for some time that we are being overwhelmed with data. We developed The Data Rod model is an object-oriented approach that provides the ability to query grid-cell changes and their relationships to neighboring grid-cells through time. This resolves the long-standing problem of managing time-series data and opens new possibilities in temporal analysis of the data. Grid-cells are the atomic data structures that can be queried for the time of a specific sensor response. Grid-cells can be used to encode the sensor details and the image metadata. The pure object database structure allows core scientific queries to run at least an order of magnitude faster than existing relational database systems (if the relational database can store the data at all). There are no tables in an object database and no need for table joins. Millions of grid-cells can be queried simultaneously over any geographic area. A Data Rod is a collection of data that can be visualized as a record of infinite length that contains all the information known at a pixel through time. Our initial databases are extensible to any field of study with spatially referenced data. Several databases were created that contained greater than 2 billion pixel objects. Our prototype databases have been created for climate change analysis, however it is extensible to any field with spatially referenced, imagery, gridded files, point files or vectors. Several databases containing billions objects have been built covering portions or all of Greenland, some for the entire temporal extent of the available data (~28 years). Each database was limited to three terabytes of data for administrative purposes. The new databases were built from 250m Modis data, 5km AVHRR, and 25km SSM/I data. Overall the database can handle massive databases at the continental scale but not the global scale, yet there is no limit to the number of regions that could be contained in separate databases. Performance so far is impressive, with typical queries producing results in seconds and data rendering in less than 20 seconds. The database design has evolved to produce an optimized configuration that maximizes data ingestion and retrieval speeds. Grid-cells can be selected in mass across geographic areas, but are not restricted to fixed time segmentations. The grid cells values (x,y) are stored in fixed spatial intervals, and time is stored sequentially but not necessarily as a fixed interval. Time is stored as discrete events; the data is organized in sets that include multiple time stamps, sensor values, and resolutions. The database can independently identify complex space-time relationships across millions of grid cells.