Geophysical Research Abstracts
Vol. 15, EGU2013-13417, 2013
EGU General Assembly 2013

Towards Big Earth Data Analytics: The EarthServer Approach

Peter Baumann
Jacobs University Bremen gGmbH, EECS, Bremen, Germany (p.baumann@jacobs-university.de)

Big Data in the Earth sciences, the Tera- to Exabyte archives, mostly are made up from coverage data whereby the term "coverage", according to ISO and OGC, is defined as the digital representation of some space-time varying phenomenon. Common examples include 1-D sensor timeseries, 2-D remote sensing imagery, 3D x/y/t image timeseries and x/y/z geology data, and 4-D x/y/z/t atmosphere and ocean data. Analytics on such data requires on-demand processing of sometimes significant complexity, such as getting the Fourier transform of satellite images. As network bandwidth limits prohibit transfer of such Big Data it is indispensable to devise protocols allowing clients to task flexible and fast processing on the server.

The EarthServer initiative, funded by EU FP7 eInfrastructures, unites 11 partners from computer and earth sciences to establish Big Earth Data Analytics. One key ingredient is flexibility for users to ask what they want, not impeded and complicated by system internals. The EarthServer answer to this is to use high-level query languages; these have proven tremendously successful on tabular and XML data, and we extend them with a central geo data structure, multi-dimensional arrays.

A second key ingredient is scalability. Without any doubt, scalability ultimately can only be achieved through parallelization. In the past, parallelizing code has been done at compile time and usually with manual intervention. The EarthServer approach is to perform a samentic-based dynamic distribution of queries fragments based on networks optimization and further criteria.

The EarthServer platform is comprised by rasdaman, an Array DBMS enabling efficient storage and retrieval of any-size, any-type multi-dimensional raster data. In the project, rasdaman is being extended with several functionality and scalability features, including: support for irregular grids and general meshes; in-situ retrieval (evaluation of database queries on existing archive structures, avoiding data import and, hence, duplication); the aforementioned distributed query processing. Additionally, Web clients for multi-dimensional data visualization are being established. Client/server interfaces are strictly based on OGC and W3C standards, in particular the Web Coverage Processing Service (WCPS) which defines a high-level raster query language.

We present the EarthServer project with its vision and approaches, relate it to the current state of standardization, and demonstrate it by way of large-scale data centers and their services using rasdaman.