Geophysical Research Abstracts
Vol. 14, EGU2012-13942, 2012
EGU General Assembly 2012

Towards Direct Manipulation and Remixing of Massive Data: The EarthServer Approach

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

Complex analytics on "big data" is one of the core challenges of current Earth science, generating strong requirements for on-demand processing and fil
tering of massive data sets. Issues under discussion include flexibility, performance, scalability, and the heterogeneity of the information types invo
lved. In other domains, high-level query languages (such as those offered by database systems) have proven successful in the quest for flexible, scalable data access interfaces to massive amounts of data. However, due to the lack of support for many of the Earth science data structures, database systems are only used for registries and catalogs, but not for the bulk of spatio-temporal data.

One core information category in this field is given by coverage data. ISO 19123 defines coverages, simplifying, as a representation of a "space-time varying phenomenon". This model can express a large class of Earth science data structures, including rectified and non-rectified rasters, curvilinear grids, point clouds, TINs, general meshes, trajectories, surfaces, and solids. This abstract definition, which is too high-level to establish interoperability, is concretized by the OGC GML 3.2.1 Application Schema for Coverages Standard into an interoperable representation. The OGC Web Coverage Processing Service (WCPS) Standard defines a declarative query language on multi-dimensional raster-type coverages, such as 1D in-situ sensor timeseries, 2D EO imagery, 3D x/y/t image time series and x/y/z geophysical data, 4D x/y/z/t climate and ocean data. Hence, important ingredients for versatile coverage retrieval are given - however, this potential has not been fully unleashed by service architectures up to now.

The EU FP7-INFRA project EarthServer, launched in September 2011, aims at enabling standards-based on-demand analytics over the Web for Earth science data based on an integration of W3C XQuery for alphanumeric data and OGC-WCPS for raster data. Ultimately, EarthServer will support all OGC coverage types. The platform used by EarthServer is the rasdaman raster database system. To exploit heterogeneous multi-parallel platforms, automatic request distribution and orchestration is being established. Client toolkits are under development which will allow to quickly compose bespoke interactive clients, ranging from mobile devices over Web clients to high-end immersive virtual reality.

The EarthServer platform has been deployed in six large-scale data centres with the aim of setting up Lighthouse Applications addressing all Earth Sciences, including satellite and airborne earth observation as well as use cases from atmosphere, ocean, snow, and ice monitoring, and geology on Earth and Mars. These services, each of which will ultimately host at least 100 TB, will form a peer cloud with distributed query processing for arbitrarily mixing database and in-situ access.

With its ability to directly manipulate, analyze and remix massive data, the goal of EarthServer is to lift the data providers' semantic level from data stewardship to service stewardship.