Geophysical Research Abstracts
Vol. 18, EGU2016-16985, 2016
EGU General Assembly 2016

"Science SQL" as a Building Block for Flexible, Standards-based Data Infrastructures

Peter Baumann
Jacobs University, CASE (Center for Advanced Systems Engineering), Bremen, Germany (p.baumann@jacobs-university.de)

We have learnt to live with the pain of separating data and metadata into non-interoperable silos. For metadata, we enjoy the flexibility of databases, be they relational, graph, or some other NoSQL. Contrasting this, users still "drown in files" as an unstructured, low-level archiving paradigm. It is time to bridge this chasm which once was technologically induced, but today can be overcome.

One building block towards a common re-integrated information space is to support massive multi-dimensional spatio-temporal arrays. These "datacubes" appear as sensor, image, simulation, and statistics data in all science and engineering domains, and beyond. For example, 2-D satellilte imagery, 2-D x/y/t image timeseries and x/y/z geophysical voxel data, and 4-D x/y/z/t climate data contribute to today's data deluge in the Earth sciences. Virtual observatories in the Space sciences routinely generate Petabytes of such data. Life sciences deal with microarray data, confocal microscopy, human brain data, which all fall into the same category.

The ISO SQL/MDA (Multi-Dimensional Arrays) candidate standard is extending SQL with modelling and query support for n-D arrays ("datacubes") in a flexible, domain-neutral way. This heralds a new generation of services with new quality parameters, such as flexibility, ease of access, embedding into well-known user tools, and scalability mechanisms that remain completely transparent to users. Technology like the EU rasdaman ("raster data manager") Array Database system can support all of the above examples simultaneously, with one technology. This is practically proven: As of today, rasdaman is in operational use on hundreds of Terabytes of satellite image timeseries datacubes, with transparent query distribution across more than 1,000 nodes.

Therefore, Array Databases offering SQL/MDA constitute a natural common building block for next-generation data infrastructures. Being initiator and editor of the standard we present principles, implementation facets, and application examples as a basis for further discussion. Further, we highlight recent implementation progress in parallelization, data distribution, and query optimization showing their effects on real-life use cases.