THE EARTHSERVER GLOBAL DATACUBE FEDERATION

Peter Baumann & the rasdaman team

Jacobs University | rasdaman GmbH
Bremen, Germany
baumann@rasdaman.com

IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium | 978-1-7281-6374-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IGARSS39084.2020.9324386

ABSTRACT

We present the EarthServer datacube federation which, based on Array DBMS technology, has established an intercontinental network of Earth data centers. Users get provided with a single, uniform information space acting like a local data holding, thereby establishing full location transparency. We present the technology and show a broad range of real-life distributed data fusion examples.

Index Terms— Datacube, federation, EarthServer, rasdaman, WCPS

1. OVERVIEW

Datacubes introduce function-rich services on spatio-temporally aligned, homogenized raster data assets typically coming as sensor, image (timeseries), simulation, and statistics data. Actionable datacubes support analysis (& visualization) through the paradigm of “any query, any time, on any size” [1] pioneered by the rasdaman team [2]. Meantime, rasdaman has matured into the de-facto standard for datacube engines being a full-stack Array DBMS implementation with Petabyte deployments served through standards-based datacube APIs [5]. Following this feasibility proof a series of epigon systems is emerging, with or without query language support. In 2019, the rasdaman query language has been adopted in the SQL datacube extension [4].

In the EarthServer federation, rasdaman provides the enabling platform for a worldwide networks of Earth data centers providing a rich variety of multi-dimensional raster data, including 2D DEMs, 3-D satellite image timeseries, and 4-D atmospheric data. Key distinguishing features of the platforms include complete location transparency (“any query, anywhere”), ad-hoc datacube analytics without programming, and seamless data/metadata integration via the OGC WCS datacube service suite including the OGC Web Coverage Processing Service (WCPS) datacube analytics language [3].

In this contribution, we present recent achievements in datacube federations based on rasdaman. We introduce the concepts underlying federated data fusion in rasdaman and show a series of real-life deployments in the EarthServer federation. Flexible extraction, visualization, and AI is available via standard clients ranging from OpenLayers over QGIS to R, python and tensorflow.

2. LOCATION-TRANSPARENT FEDERATION

Users experience a single, integrated information space: Intercontinental location-transparent, security-enabled federation and data fusion uniting research centers, agencies, large & small industry. The rasdaman engine resembles a fully-fledged DBMS architecture with a full-stack implementation in fast C++. Its architecture centers around the multi-parallel rasdaman worker processes which operate on arbitrarily tiled arrays (a tuning factor accessible to the administrator) stored in rasdaman or read from some legacy archive (thereby avoiding copies). When ingesting data they can be stored in a number of formats through an OGC WCS-T based ETL layer which homogenizes data and metadata, provides defaults, as well as the target tiling strategy. Further tuning parameters include compression, indexing, cache sizing, etc. The resulting OGC compliant coverages represent analysis-ready space-time EO objects.

In a rasdaman federation, worker processes can fork subqueries to other cloud nodes or other data centers for load sharing and data transport minimization. Users do not need to know the location of the datacubes in the federation; the rasdaman federation members continuously exchange about new, changed, or deleted datacubes so that they can at any time determine an optimal work distribution. Figure 1 shows a sample expression submitted; color codes show where each sub-expression will be evaluated.

3. DISTRIBUTED DATA FUSION IN PRACTICE

We present several data fusion cases demonstrating the added value of a location-transparent federation. In the EarthServer-2 project, heavy rainfall risk areas have been determined by combining data from National Computational Infrastructure Australia and ECMWF / UK. For visualization of the result, NASA WebWorldWind has been used (Figure 2). In another setup, the federated query, combining RGB from Sentinel-2 data on CODE-DE with a Digital Surface Model coming from commercial geo service provider cloudeo AG to generate a 4-band PNG image visualized with WebGL in the browser (Figure 5). The next example shows datacube aggregation: By combining data from Alfred Wegener Institute and Jacobs University an estimate of the sea ice temperature distribution is built from global sea ice and temperature data, effectively computing a histogram as a 1-D coverage output (Figure 3). Going one step further, a triple fusion is performed between Alfred Wegener Institute, Jacobs University, and CODE-DE. Sea ice, chlorophyll, and temperature get correlated (Figure 4). Finally, an NDVI forecast of Northwestern Taiwan can be obtained from data provided by both CODE-DE and GIS Taiwan (Figure 6).

4. LESSONS LEARNT

Users appreciate easy global access, but federation at scale needs several key features for eliminating the overhead of fiddling with heterogeneous archive structures and server technologies. Among the advantageous elements we found are:

- Location-transparency: for a common data & processing pool);
- strong standards support: benefits both federation members by reducing access overhead and increasing the number of clients available;
- codified governance rules for transparency among the federation partners and towards the users;
- as much as possible effortless administration;
- security capabilities, in particular: sub-cube access control.

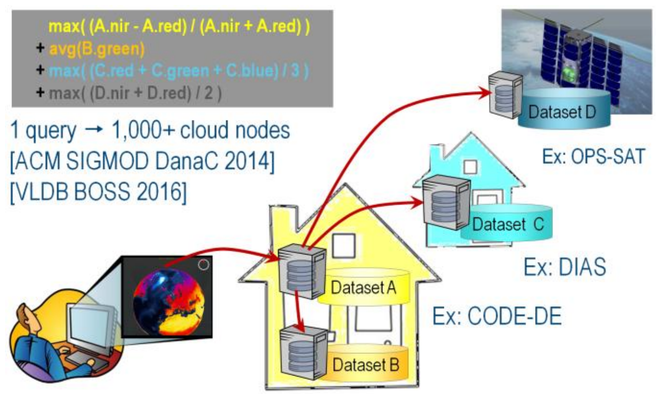

Figure 1 - rasdaman query splitting for cloud parallelization and distributed processing

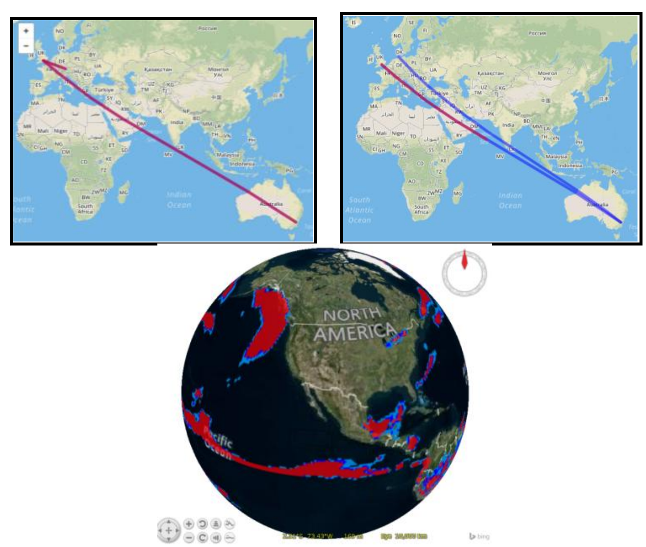

Figure 2 - WCPS fusion query path from Central Europe to ECMWF / UK, split query sent to Canberra / AUS (top left); query path to Canberra and subqery sent to ECMWF (top right); query result displayed in NASA WorldWind (bottom center)

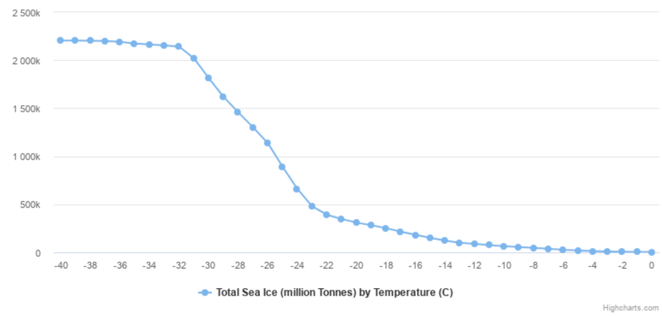

Figure 3 - WCPS global sea ice temperature classification histogram (left) and triple fusion query correlating sea ice, chlorophyll, and temperature (right)

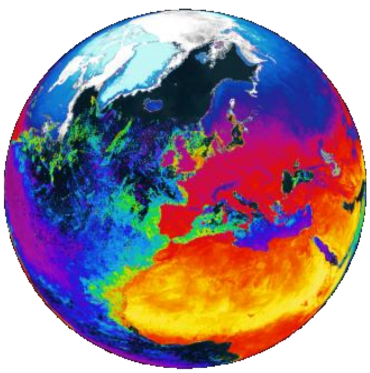

Figure 4 - WCPS global sea ice temperature classification histogram (left) and triple fusion query correlating sea ice, chlorophyll, and temperature (right)

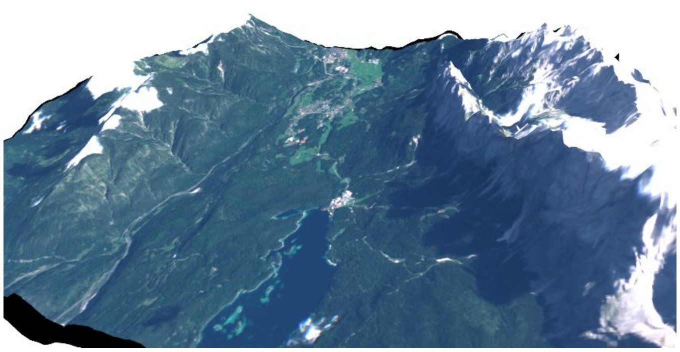

Figure 5 - WCPS query with result visualized through terrain draping in WebGL

5. NEXT STEPS

Aside from advancing rasdaman technically, an aggressive growth of the EarthServer federation is ongoing; a line-up of datacenters has expressed interest, and a charter for governance is being established.

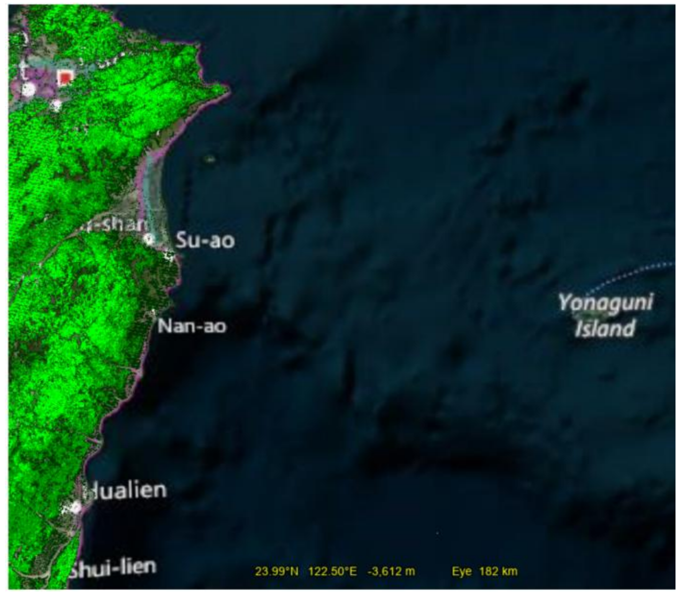

Figure 6 - WCPS query for NDVI forecast in Northwestern Taiwan

ACKNOWLEDGEMENT

Research supported by EU H2020 EarthServer-1/-2, LandSupport, CopHub.AC, PARSEC, German BigDataCube, BigPicture, ORBiDANSe, DeepRain; and several more.

REFERENCES

[1] P. Baumann, D. Misev, V. Merticariu, B. Pham Huu: Datacubes: Towards Space / Time Analysis-Ready Data.. In: J. Doellner, M. Jobst, P. Schmitz (eds.): Service Oriented Mapping - Changing Paradigm in Map Production and Geoinformation Management, Springer Lecture Notes in Geoinformation and Cartography, 2018

[2] P. Baumann: *Language Support for Raster Image Manipulation in Databases*. Intl. Workshop on Graphics Modeling, Visualization in Science & Technology, Darmstadt/Germany 1992, pp. 236 – 245

[3] P. Baumann: *The OGC Web Coverage Processing Service (WCPS) Standard*. Geoinformatica, 14(4)2010, pp 447-479

[4] ISO: *9075-15:2019 SQL/MDA (Multi-Dimensional Arrays)*. https://www.iso.org/standard/67382.html

[5] OGC: *OGC Spatio-Temporal Coverage / Datacube Standards*. http://myogc.org/go/coveragesDWG