Federating CODE-DE & DIAS Datacubes: Successes and Lessons Learnt

Peter Baumann, Dimitar Misev, Vlad Merticariu, Bang Pham Huu, Ursula Benz, Dimitris Bellos, Sven Jacobsen, Stefan Wiehle
Jacobs University, rasdaman GmbH, cloudeo AG, DLR

Datacube Federations: the rasdaman Approach

The rasdaman datacube engine can be configured into location-transparent federations offering a single, coherent information space without single point of failure. For users, this means they can approach any of the federation members and see all data made available by the federation members; they can access data on any node without further specifying location, and they can perform fusion of objects that may sit on different federation member nodes.

Data providers, when joining the federation, will configure their servers to specify / from directions for queries, thereby establishing the global part; also authentication needs to be agreed across nodes. Authorization, conversely, remains under exclusive control of the individual nodes, thereby retaining full authority. Technically, rasdaman transforms all incoming requests (such as OGC GetMap, Get-Coverage, ProcessCoverages, etc.) into array queries which then are split dynamically during query optimization. Subqueries are sent to the resp. Nodes holding the data ("ship code to data"); incoming partial results are integrated, and the final query result is presented to the caller.

Acknowledgement

BigDataCube is supported by German BMWI (Ministry of Economy and Energy). EOSC-hub is supported by the European Commission under H2020.

Lessons Learnt

- Homogenization of billions of scenes into few analysis-ready spatio-temporal datacubes is making access (in particular via OGC WCS/WCPS/WMS) substantially easier for casual users, experts, and machines.
- Datacubes can be supported in an efficient, scalable manner without extra burden to the user.
- Building datacubes moves burden from users to data providers – which makes sense (do it once, with larger resources, with better expertise); there should be strong tool support (such as datacube ETL) to minimize this burden.
- Formats like SAFE are good for archiving, but for a series of timeseries analytics. Hence, archives need to be reorganised (say, NetCDF cubelets) to become analysable.
- Federation raises novel security challenges on technology and governance level.
  - Authentication must be global (data fusion!) while authorisation (i.e., fine grain access control policies) must remain local; tools must support this.
  - Governance decisions are critical in a federated ecosystem; federation partners should exchange, agree, and streamline global aspects of governance.

Datacube History

Queryable datacubes were first proposed in 1992 [1]. Adoption, though long was limited to business analytics. Recently, science and technology also considers datacubes, in different terminologies: Multi-Dimensional Discrete Data (MDD), Array Databases, datacubes, etc.

The 2018 Research Data Alliance (RDA) datacube report investigates 19 tools. Results indicate distinct advantages of high-level query languages over scripting languages like python: flexibility, performance, scalability. The benchmark shows that rasdaman can be 300x faster [http://dx.doi.org/10.15497/RDA00024].