# DATACUBES AS A SERVICE PARADIGM[1]

*Peter Baumann, Angelo Pio Rossi*

Jacobs University, Bremen, Germany

## ABSTRACT

The data deluge we face does not only overwhelm us with sheer data volume, but also with an increasing variety of multi-dimensional datasets. As is more and more accepted, spatio-temporal data sets often can be represented through a common unifying paradigm, namely datacubes.
In this contribution, we present datacubes as a common modeling paradigm upon which flexible, scalable services can be offered based on the concept of a datacube query language. With the use case of an intercontinental datacube federation, EarthServer, we point out how the technicalities of the query language can be hidden, thereby allowing user-friendly visual data interaction.

***Index Terms***— datacube, Big Data, EarthServer, rasdaman, OGC WCS, OGC WCPS

## 1. INTRODUCTION

With the unprecedented increase of orbital sensor, in-situ measurement, and simulation data there is a rich, yet not leveraged potential for getting insights from dissecting datasets and rejoining them with other datasets, effectively establishing a "datacube" paradigm with the ultimate goal to allow users to "ask any question, any time" thereby enabling them to "build their own product on the go". Notably, the term datacube refers to multi-dimensional spatio-temporal datasets, not only to 3-D – for example, we find 1-D sensor timeseries, 2-D satellite imagery, 3-D x/y/t image timeseries and x/y/z geophysical voxel data, as well as 4-D x/y/z/t meteorological simulation output, to name but a few.

One of today’s most influential initiatives in Big Geo Data is EarthServer [5] which is paving the way for flexible, scalable datacube services based on innovative NewSQL technology (Fig. 1). Researchers from Europe, the US and recently Australia have teamed up to rigorously materialize the datacube paradigm.

EarthServer has established client and server technology for such spatio-temporal datacubes strictly based on open standards [4][2][3][5]. The underlying scalable array engine, rasdaman, enables direct interaction, including 3-D visualization, what-if scenarios, common EO data processing, and general analytics [1][7][9]. Conversely, EarthServer has significantly shaped and advanced the OGC Big Geo Data standards landscape based on the experience gained.

Phase 1 of EarthServer has advanced scalable array database technology into 100+ TB services [5]; in phase 2, a federation of Petabyte datacubes is being built in Europe and Australia to perform ad-hoc querying and merging.

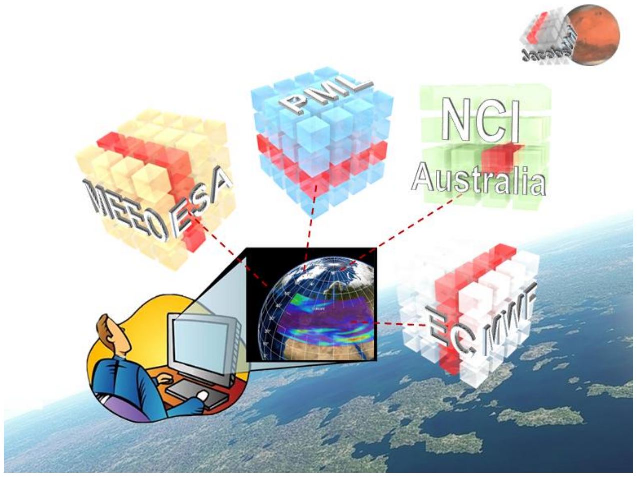

Fig. 1: Intercontinental datacube mix and match in the EarthServer initiative (source: EarthServer)

## 2. CONCEPTS

For modeling datacubes, EarthServer relies on the “Big Geo Data” standards of OGC (which, due to their success, are meantime also under adoption by ISO and INSPIRE). Key is the concept of *coverages* as representations of space/time varying phenomena, encompassing regular and irregular grids, point clouds, and general meshes (Fig. 2).

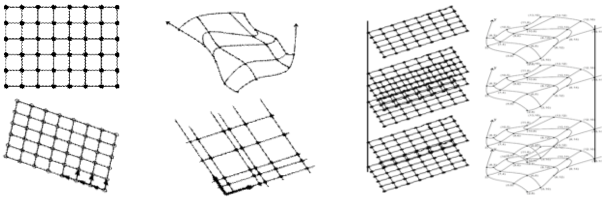

Fig. 2: Sample datacube grid types supported by rasdaman (source: OGC / Jacobs University)

[1] Research supported through EU H2020 project EarthServer-2

While such coverages can be accessed through many service interfaces, the *Web Coverage Service* (WCS) standard offers streamlined functionality. This modular suite of specifications starts with WCS Core allowing trimming and slicing of coverages, plus encoding the result in some data format for shipping. Various extensions add optionally implementable functionality facets, such as band ("channel", "variable") extraction, scaling, reprojection, and datacube maintenance. On the high end there is the WCS Processing extension which allows server-side filtering and processing based on the *Web Coverage Processing Service* (WCPS) geo raster query language.

In EarthServer, all data access is through WCS, WCPS, and WMS. Advanced visual clients enable point-and-click interfaces effectively hiding the query language syntax, except when experts want to make use of it.

Aside from data extraction and processing, another important facet is integration of data and metadata handling. Traditionally (and technologically induced), data and metadata are stored, offered, and manipulated separately and through completely different methods. In EarthServer, an integration of WCPS with XQuery is being pursued (to be proposed as WCPS 2.0) which enables queries involving both XML metadata and gridded data simultaneously.

Actually, EarthServer is not only using these standards, but also shaping them: the Scientific Coordinator is editor of the WCS/WCPS suite. Further, in ISO the SQL standard is being enhanced with n-D arrays in a domain-neutral way. This has prompted reviewers at the end of phase 1 to state that "with no doubt" EarthServer "has been shaping the Big Earth Data landscape ".

An example may illustrate the use of WCPS: "*From MODIS scenes M1, M2, M3 over Bavaria, the difference between red & near-infrared bands, encoded as TIFF, but only those where near-infrared exceeds threshold 127 somewhere*." The corresponding query reads as follows:

```
for $c in doc("http://acme.com/wcs")//coverage
where
        some( $c.nir > 127 )
    and $c/metadata/@region = "Bavaria"
return
    encode( $c.red - $c.nir, "image/tiff" )
```

Such results can conveniently be rendered through WebGL in a standard Web browser (Fig. 3).

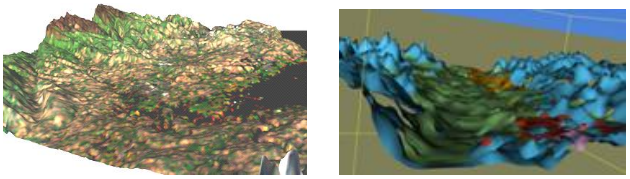

Fig. 3: 3-D rendering of geological query results (data: BGS, database: rasdaman)

## 3. SERVICES

EarthServer is committed to establishing operational services by augmenting existing data holdings with datacube functionality. Specifically, the following services are already existing or under development, resp. (Fig. 4): ocean color analysis and further marine offerings by Plymouth Marine Laboratory (PML); weather and climate data by the European Centre for Medium-Range Weather Forecasts (ECMWF); Sentinel (and further satellite image) data by the European Space Agency, managed by MEEO s.r.l.; reprocessed Landsat data (and further imagery) by the National Computational Infrastructure (NCI) of Australia. Finally, Jacobs University is extending its planetary science data service, PlanetServer, from Mars to further solar system bodies, such as Moon and Vesta [11].

While in phase 1 of EarthServer the 100 TB barrier has been exceeded, in its phase 2 offerings will cross the Petabyte frontier. Altogether, these services comprise the nucleus for a worldwide data federation; further data centers have already expressed interest in joining to bring in their data holdings. Ultimately, the EarthServer network will allow users to send queries to any node while transparently combining data sitting anywhere in the federation.

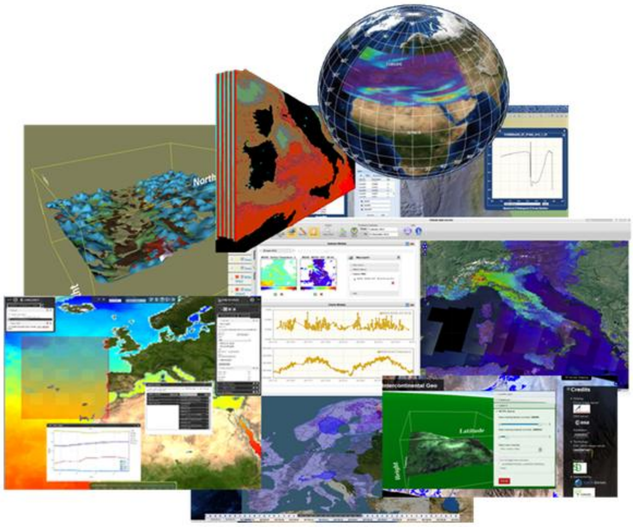

Fig. 4: EarthServer (and further rasdaman-based) datacube portals (source: Jacobs University)

## 4. TECHNOLOGY

The common engine underlying EarthServer is the rasdaman Array Database [1]. It extends SQL with support for massive multi-dimensional arrays, together with declarative array operators which are heavily optimized and parallelized [7] on server side (Fig. 5). A separate layer adds geo semantics, such as knowledge about regular and irregular grids and coordinates, by implementing the OGC Web service interfaces. For WCS and WCPS, rasdaman acts as official OGC reference implementation.

On storage arrays get partitioned (“tiled”) into sub-arrays which can be stored in a database or directly in files. Additionally, rasdaman can access pre-existing archives by only registering files, without copying them.

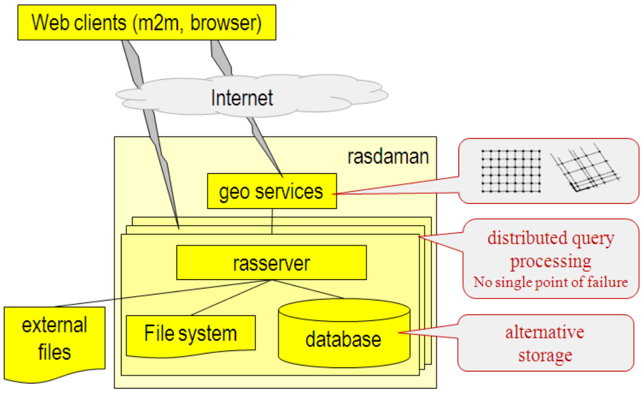

Fig. 5: rasdaman overall architecture (source: rasdaman)

On client side, NASA WorldWind has been chosen as a primary user interface due to its capabilities to allow lay people and experts alike a smooth, fast, and easily understandable visualization of datacube extracts.

A demonstration service for 1-D through 5-D coverages is available for studying the WCS / WCPS universe [14].

## 5. CONCLUSION

Datacubes are a convenient model for presenting users with a simple, consolidated view on the massive amount of data files gathered – “a cube tells more than a million images”. Such a datacube may have spatial and temporal dimensions (such as a satellite image time series) and may unite an unlimited number of individual images. Independently from whatever efficient data structuring a server network may perform internally, users will always see just a few datacubes they can slice and dice.

In the EarthServer initiative, the datacube paradigm is used as the central paradigm for fast, versatile access to massive 3-D satellite image timeseries and 4-D meteorological data.

The underlying rasdaman technology has proven that Array Databases constitute a suitable platform for flexible and scalable datacube extraction, analysis, and fusion. WCS and WCPS prove to be convenient tools for rapidly attaching clients and do versatile ad-hoc analytics. However, aside from using the OGC “Big Geo Data” standards for its client/server interfaces, EarthServer in turn massively shapes datacube standards in OGC, ISO, INSPIRE, and beyond. In the abovementioned review, experts attested rasdaman to "significantly transform the way that scientists in different areas of Earth Science will be able to access and use data in a way that hitherto was not possible".

Current work involves implementation of the forthcoming version 1.1 of the OGC coverage model, supporting data centers in establishing rasdaman-based services, and enhancing further the data and processing parallelism capabilities of rasdaman.

## 6. REFERENCES

[1] P. Baumann, A. Dehmel, P. Furtado, R. Ritsch, N. Widmann; “Spatio-Temporal Retrieval with RasDaMan”, Very Large Data Bases (VLDB), Edinburgh, Scotland, UK, September 7-10, 1999

[2] P. Baumann, “OGC WCS Interface Standard – Core”, version 2.0.1, OGC standard 09-110r4, 2010

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

[4] P. Baumann, “OGC GML 3.2.1 Application Schema - Coverages”, version 1.0, OGC standard 09-146r2, 2012

[5] P. Baumann, P. Mazzetti, J. Ungar, R. Barbera, D. Barboni, A. Beccati, L. Bigagli, E. Boldrini, R. Bruno, A. Calanducci, P. Campalani, O. Clement, A. Dumitru, M. Grant, P. Herzig, G. Kakaletris, J. Laxton, P. Koltsida, K. Lipskoch, A.R. Mahdiraji, S. Mantovani, V. Merticariu, A. Messina, D. Misev, S. Natali, S. Nativi, J. Oosthoek, J. Passmore, M. Pappalardo, A.P. Rossi, F. Rundo, M. Sen, V. Sorbera, D. Sullivan, M. Torrisi, L. Trovato, M.G. Veratelli, S. Wagner, “Big Data Analytics for Earth Sciences: the EarthServer Approach”, International Journal of Digital Earth, 0(0)2015, pp. 1 – 27, 2015

[6] P. Baumann and E. Hirschorn, “OGC Coverage Implementation Schema”, version 1.1, OGC candidate standard 09-146r3, 2015.

[7] A. Dumitru, V. Merticariu, P. Baumann, “Exploring Cloud Opportunities from an Array Database Perspective”, Proc ACM SIGMOD Workshop on Data Analytics in the Cloud (DanaC'2014), Snowbird, USA, June 22 - 27, 2014

[8] EarthServer, “EarthServer: Big Datacubes at Your Fingertips”, www.earthserver.eu, visited 20.11.2015

[9] D. Misev, P. Baumann, “Enhancing Science Support in SQL”, Workshop Data and Computational Science Technologies for Earth Science Research (co-located with IEEE Big Data), Santa Clara, US, October 29, 2015

[10] OGC, “Web Coverage Service”, www.opengeospatial.org/standards/wcs, visited 10.1.2016

[11] PlanetServer, “PlanetServer”, www.planetserver.eu, visited 10.1.2016

[12] J. Melton, P. Baumann, D. Misev, “ISO 9075 SQL Part 15: Multi-Dimensional Arrays”, ISO candidate standard, 2015

[13] rasdaman, “rasdaman”, www.rasdaman.org, visited 10.1.2016

[14] WCS demo, standards.rasdaman.com, visited 10.1.2016