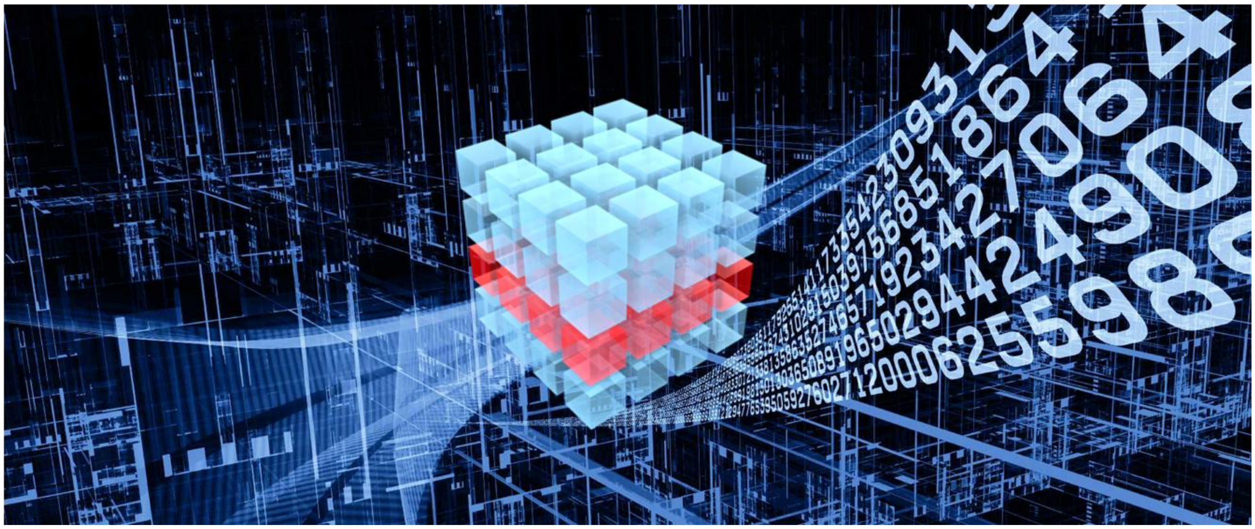

Array Databases: Concepts, Standards, Implementations

Peter Baumann[1], Dimitar Misev[1], Vlad Merticariu[1], Bang Pham Huu[1], Brennan Bell[1], Kwo-Sen Kuo[2]

[1] Jacobs University
Large-Scale Scientific Information Systems Research Group
Bremen, Germany
{p.baumann,d.misev,v.merticariu,b.phamhuu,b.bell}@jacobs-university.de

[2] Bayesics, LLC / NASA
USA
kwo-sen.kuo@nasa.gov

Corresponding author: Peter Baumann, p.baumann@jacobs-university.de

This report is proposed for an official RDA Recommendation
produced by the RDA Array Database Assessment Working Group.

Executive Summary

Multi-dimensional arrays (also known as raster data or gridded data) play a core role in many, if not all science and engineering domains where they typically represent spatio-temporal sensor, image, simulation output, or statistics "datacubes". However, as classic database technology does not support arrays adequately, such data today are maintained mostly in silo solutions, with architectures that tend to erode and have difficulties keeping up with the increasing requirements on service quality.

Array Database systems attempt to close this gap by providing declarative query support for flexible ad-hoc analytics on large n-D arrays, similar to what SQL offers on set-oriented data, XQuery on hierarchical data, and SPARQL or CIPHER on graph data. Today, Petascale Array Database installations exist, employing massive parallelism and distributed processing. Hence, questions arise about technology and standards available, usability, and overall maturity.

To elicit the state of the art in Array Databases, Research Data Alliance (RDA) has established the Array Database Assessment Working Group (ADA:WG) as a spin-off from the Big Data Interest Group. Between September 2016 and March 2018, the ADA:WG has established an introduction to Array Database technology, a comparison of Array Database systems and related technology, a list of pertinent standards with tutorials, and comparative benchmarks to essentially answer the question: *how can data scientists and engineers benefit from Array Database technology?*

Investigation shows that there is a lively ecosystem of technology with increasing uptake, and proven array analytics standards are in place. Tools, though, vary greatly in functionality and performance as investigation shows. While systems like rasdaman are Petascale proven and parallelize across 1,000+ cloud nodes, others (like EXTASCID) still have to find their way into large-scale practice. In comparison to other array services (MapReduce type systems, command line tools, libraries, etc.) Array Databases can excel in aspects like service friendliness to both users and administrators, standards adherence, and often performance. As it turns out, Array Databases can offer significant advantages in terms of flexibility, functionality, extensibility, as well as performance and scalability – in total, their approach of offering "datacubes" analysis-ready heralds a new level of service quality. Consequently, they have to be considered as a serious option for "Big DataCube" servicees in science, engineering and beyond.

The outcome of this investigation, a unique compilation and in-depth analysis of the state of the art in Array Databases, is supposed to provide beneficial insight for both technologists and decision makers considering "Big Array Data" services in both academic and industrial environments.