RDA PROPOSAL FOR
Array Database Assessment Working Group (ADA:WG)
Peter Baumann, Jacobs University

Co-Chairs

Initial chairs will be: Peter Baumann (Jacobs University, Germany) and Kwo-Sen Kuo (Bayesics, US). Replacement of chairs, or extending the chairs list, is possible anytime through democratic voting.

Background

Multi-dimensional data arrays play a core role in many, if not most science and engineering domains where they typically represent spatio-temporal sensor, image, simulation output, or statistics data. The research field of Array Databases has emerged in the attempt to augment the traditional set-driven paradigm with modeling and query support for large, n-D arrays. Such systems attempt to combine the best functionality and performance of different worlds: the long-standing experience of array handling in the sciences, the flexibility of database query languages, and the parallelization and scalability methods developed in HPC, HPD and Cloud, as well as using new hardware. Various implementations are known [1][4][2][5][6][10][14], deployments of Array Databases today are in the hundreds of Terabytes [13], and single queries reportedly have been parallelized over more than 1,000 nodes [3]. Therefore, Array Databases have to be considered as a serious option for Big Data management in science, engineering and beyond.

---

1 Several colleagues contributed important facets to this proposal; aside from the TAB reviewers, Peter Wittenburg, Ben Evans, and several more.
1. WG Charter

The Array Database WG will inspect the emerging technology of Array Databases to provide support for technologists and decision makers considering Big Data services in academic and industrial environments (such as in large-scale data centers) by establishing best-practice guidelines on how to optimally serve multi-dimensional gridded Big Data through Array Databases. This will be accomplished through a neutral, thorough hands-on evaluation assessing available Array Database systems and comparable technology ...

- based on relevant standards, such as the NIST Big Data Reference Architecture [2], ISO “Array SQL” [8], and OGC Web Coverage Processing Service (WCPS) [9] for the geo domain;
- comparing technical criteria like functionality, thereby eliciting the state of the art;
- establishing and running a combination of domain-driven and domain-neutral benchmarks that will be run on each platform;
- as well as real-life, publicly accessible deployments at scale.

The result, consisting of the AD-WG report together with the open-source benchmarking software and the services established, will establish a hitherto non-existing overview on the state of the art and best use of Array Databases in science, engineering, and beyond.

2. Value Proposition

Open-source and proprietary Array Databases are now becoming more readily available to the research community, and they promise to provide a new way of managing and analyzing regularly and irregularly gridded data in Earth, Space, and Life sciences, and beyond. For example, in the Earth sciences we find 1-D sensor time-series, 2-D satellite imagery, 3-D x/y/t image time-series and x/y/z geophysical voxel cubes, as well as 4-D x/y/z/t climate simulations. Similar data are used within industry, such as oil and gas exploration, insurance, pharmaceutical, automotive, shipping, aerospace industries, and may provide increased uptake of such technology.

The analysis of existing technology will focus on one question: to what extent can data scientists and engineers benefit from Array Database technology? To this end, the Working Group will address at least the following questions (greater detail to be determined during the first phase, i.e., until Milestone 1):

- **What systems are out there?** (currently considered are: rasdaman [1][4], SciQL [10], SciDB [5], PostGIS Raster [6], Ophidia[14], possibly Extascid [2]; further, a comparison with alternative technology is foreseen, such as SciHadoop [11]), HDF5 on filesystems [12]) and Python scientific data formats.
- **What are their features?** Currently recognized aspects include: query language power; scientific tool integration; adaptive data partitioning and distribution support, including alignment for data fusion; parallelization in clouds and data center federations; support for modern hardware; etc. – list to be completed by Milestone 1.
- **What is their performance, measured through objective, open tests?** To this end, a combination of domain-specific and domain-neutral (“kernel”) benchmarks will be developed, agreed in the WG, and applied to the systems.
- **What are relevant tuning parameters, and what are the recommended settings for these systems?**
- **How can these systems be used in large-scale deployments?** This will be answered by setting up publicly accessible installations\(^2\) (utilizing pre-existing installations where possible, to minimize

---

\(^2\) This will depend on the WG member resources. It is expected that each entity participating will pick and run at least one system. Therefore, further members will be chosen so as to achieve a broad range of such installations. (Note that this applies only to the large-scale deployment – all other investigations will be done on several systems in any case.)
effort) and evaluating them from the perspective of IT, service operator, and user. A list of known installations will be compiled.

- Generally, what are the strengths and weaknesses of Array Databases for Big Data in science & engineering?

In the final report, the findings will be consolidated into a state-of-the-art report, accompanied by the benchmarks, as well as guidance to service operators on using an Array Database system. Any code developed will made available as open-source along with documentation and samples. It is hoped that this will be an important decision aid for science and industry alike.

3. Engagement with existing work in the area

Engaging and collaborating with other related existing works are important to our mission. This will include interaction with consortiums, alliances, and standards bodies. The knowledge gained is expected to enhance our development by aligning with high quality best practices. These include at a minimum:

- OGC [9]
- EU INSPIRE

The list will be revisited and likely amended until Milestone 1 (see below).

4. Work Plan

- **Milestones**

<table>
<thead>
<tr>
<th>milestone</th>
<th>month</th>
<th>outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>3</td>
<td>Evaluation criteria established, candidate systems chosen, external engagements defined</td>
</tr>
<tr>
<td>M2</td>
<td>6</td>
<td>Testbeds established</td>
</tr>
<tr>
<td>M3</td>
<td>9</td>
<td>Evaluations done</td>
</tr>
<tr>
<td>M4</td>
<td>12</td>
<td>Final report: results and recommendations</td>
</tr>
</tbody>
</table>

- **Mode and Frequency of Operation**

  It is anticipated that all discussions will be conducted via emails plus bi-weekly teleconferencing through a Web-based conferencing tool. The date/time of the bi-weekly telecon will be determined by Doodle Poll.

- **Membership**

  Participation in the WG is open to all interested parties. There are no membership fees.

- **Coordination/Interaction**

  The WG will function in close coordination with other Big Data related standards and best practices from industry, academia and government at the international level.

- **Standing Rules**

  All information exchanged within the WG will be freely available (using CC-BY license).
  All information exchanged within the WG will contain non-IPR materials.
  WG members should assume that all materials exchanged will be made public.
  Documents will be publicly accessible from the WG portal.
5. Adoption Plan
Following completion of the report (and likely at intermediate milestones, too) the AD-WG will perform outreach activities to promote insights into stakeholder communities. This includes standardization meetings (such as OGC, ISO, INSPIRE) as well as conferences (such as database and geospatial conferences). The outreach portfolio will be elaborated in the WG and will evolve over its lifetime.

6. Initial Membership
Currently confirmed members (note the balanced participation of academia, supercomputing centres, and industry SMEs):

<table>
<thead>
<tr>
<th>First Name</th>
<th>Last Name</th>
<th>Organization</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>Baumann</td>
<td>Jacobs University</td>
<td>Germany</td>
</tr>
<tr>
<td>Dimitar</td>
<td>Misev</td>
<td>rasdaman GmbH</td>
<td>Germany</td>
</tr>
<tr>
<td>Morris</td>
<td>Riedel</td>
<td>Juelich Supercomputing Centre</td>
<td>Germany</td>
</tr>
<tr>
<td>Oliver</td>
<td>Clements</td>
<td>Plymouth Marine Laboratory</td>
<td>UK</td>
</tr>
<tr>
<td>Mike</td>
<td>Grant</td>
<td>Plymouth Marine Laboratory</td>
<td>UK</td>
</tr>
<tr>
<td>Stephan</td>
<td>Siemen</td>
<td>European Centre for Medium-Range Weather Forecasts</td>
<td>UK</td>
</tr>
<tr>
<td>Julia</td>
<td>Wagemann</td>
<td>European Centre for Medium-Range Weather Forecasts</td>
<td>UK</td>
</tr>
<tr>
<td>Simone</td>
<td>Mantovani</td>
<td>MEEO s.r.l.</td>
<td>Italy</td>
</tr>
<tr>
<td>George</td>
<td>Kakaletris</td>
<td>CITE s.a.</td>
<td>Greece</td>
</tr>
<tr>
<td>Panagiota</td>
<td>Koltsida</td>
<td>ATHENA Research Centre</td>
<td>Greece</td>
</tr>
<tr>
<td>Patrick</td>
<td>Hogan</td>
<td>NASA Ames</td>
<td>US</td>
</tr>
<tr>
<td>Ben</td>
<td>Evans</td>
<td>National Computational Infrastructure</td>
<td>Australia</td>
</tr>
<tr>
<td>Joseph</td>
<td>Antony</td>
<td>National Computational Infrastructure</td>
<td>Australia</td>
</tr>
<tr>
<td>Sandro</td>
<td>Fiore</td>
<td>Euro Mediterranean Center on Climate Change (CMCC)</td>
<td>Italy</td>
</tr>
<tr>
<td>Kwo-Sen</td>
<td>Kuo</td>
<td>Bayesics. Inc.</td>
<td>US</td>
</tr>
</tbody>
</table>

7. References
[7] NIST Big Data Reference Architecture, M0266

---

will invite further members once WG has initial agreement from RDA
**Annex: Responses to public comment period**

The following comments below have been posted. Those excerpts containing a “do message” have been extracted, and are commented on.

<table>
<thead>
<tr>
<th>Suggestion</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>deepseadawn:</td>
<td>• The WG, by its definition, will be looking at data that can be represented as arrays (i.e., gridded rasters). Therefore, general unstructured data (such as text) is out of scope. That said, many people in the past have considered array data (when coming as images or data) have been (wrongly) considered as “unstructured”. In this sense, such data are under consideration, too.</td>
</tr>
<tr>
<td>Peter Wittenburg:</td>
<td>• For sure there will be experience provided on ease/difficulty of installation and operation. Price lists etc. cannot be provided (interested parties should contact vendors for conditions). • See above – non-database concepts are included, such as SciHadoop. Generally, the more activists we find the more systems and facets can be addressed. Examples from different communities are planned, but again: we need stakeholders from domains in order to investigate them. Of course, the more specific you suggest items the better we can try planning them in. For example, we will be most happy if you can provide more domain stakeholders to us in addition to those on board! • Who is going to test: this will be the WG members. Detailed work distribution to be done as a first workpackage of the group (see M1). Of course, it critically depends on the level of commitment of partners.</td>
</tr>
<tr>
<td>• I wonder if this WG will be considering unstructured (big) data as well?</td>
<td>• •</td>
</tr>
</tbody>
</table>