Array Databases: Agile Analytics (not just) for the Earth Sciences

Gridded data, such as images, image timeseries, and climate datacubes, today are managed separately from the metadata, and with different, restricted retrieval capabilities. While databases are good at metadata modelled in tables, XML hierarchies, or RDF graphs, they traditionally do not support multi-dimensional arrays.

This gap is being closed by Array Databases, pioneered by the scalable rasdaman ("raster data manager") array engine. Its declarative query language, rasql, extends SQL with array operators which are optimized and parallelized on server side. Installations can easily be mashed up securely, thereby enabling large-scale location-transparent query processing in federations. Domain experts value the integration with their commonly used tools leading to a quick learning curve.

Earth, Space, and Life sciences, but also Social sciences as well as business have massive amounts of data and complex analysis challenges that are answered by rasdaman. As of today, rasdaman is mature and in operational use on hundreds of Terabytes of timeseries datacubes, with transparent query distribution across more than 1,000 nodes. Additionally, its concepts have shaped international Big Data standards in the field, including the forthcoming array extension to ISO SQL, many of which are supported by both open-source and commercial systems meantime. In the geo field, rasdaman is reference implementation for the Open Geospatial Consortium (OGC) Big Data standard, WCS, now also under adoption by ISO. Further, rasdaman is in the final stage of OSGeo incubation.

In this contribution we present array queries a la rasdaman, describe the architecture and novel optimization and parallelization techniques introduced in 2015, and put this in context of the intercontinental EarthServer initiative which utilizes rasdaman for enabling agile analytics on Petascale datacubes.

Big Earth Data: the Film, the Experience, and some Thoughts

Scientists have to get out of the ivory tower and tell society, which ultimately finances them, about their work, their results, and implications, be they good or bad. This is commonly accepted ethics. But how would you "tell society" at large what you are doing? Scientific work typically is difficult to confer to lay people, and finding suitable simplifications and paraphrasings requires considerable effort. Estimating societal implications is dangerous as swimming with sharks, some of which are your own colleagues. Media tend to be not always interested - unless results are particularly spectacular, well, in a press sense. Again, sharks are luring. All this makes informing the public a tedious, time-consuming task which tends to receive not much appreciation in tenure negotiations where indexed publications are the first and foremost measure.

As part of the EU funded EarthServer initiative we tried it. Having promised a "video about the project" we found it boring to do another 10 minute repetition from the grant contract and started aiming at a full TV documentary explaining "Big Earth Data" to the interested citizens. It took more than one year to convince a TV producing company and TV stations that this is not another feature about the beauty of nature or catastrophies, but about human insight from computer-supported sifting through all those observations and simulations available. After they got the gist they were fully on board and supported financially with a substantial amount. The final 53 minutes "Big Earth Data"

movie was broadcast in February 2015 in German and French (English version available from). Several smaller spin-off features originated around it, such as an uptake of the theme (and material) in a popular German science TV series.

Of course, this is but one contribution and cannot be made a continuous activity. In the talk we want to present and discuss the "making of" from a scientist's perspective, highlighting the ups and downs in the process, in the hope that it contributes to our profession's quest for materializing its scientific outreach promise.

IN008: Beyond the power of one: collaborative efforts creating new standards and platforms to enable programmatic access to data for multiple use cases

EarthServer: an Intercontinental Collaboration on Petascale Datacubes

Previously Published:
https://ec.europa.eu/digital-agenda/

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. Obviously, the goal is to allow users to "ask any question, any time" thereby enabling them to "build their own product on the go".

One of the most influential initiatives in Big Geo Data is EarthServer which has demonstrated new directions for flexible, scalable EO services based on innovative NewSQL technology. Researchers from Europe, the US and recently Australia have teamed up to rigourously materialize the concept of the datacube. Such a datacube may have spatial and temporal dimensions (such as a satellite image time series) and may unite an unlimited number of scenes. 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. EarthServer has established client and server technology for such spatio-temporal datacubes. The underlying scalable array engine, rasdaman, enables direct interaction, including 3-D visualization, what-if scenarios, common EO data processing, and general analytics. Services exclusively rely on the open OGC "Big Geo Data" standards suite, the Web Coverage Service (WCS) including the Web Coverage Processing Service (WCPS). 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; in phase 2, Petabyte datacubes will be built in Europe and Australia to perform ad-hoc querying and merging. Standing between EarthServer phase 1 (from 2011 through 2014) and phase 2 (from 2015 through 2018) we present the main results and outline the impact on the international standards landscape; effectively, the Big Geo Data standards established through initiative of EarthServer include OGC, ISO, and INSPIRE.

Big Earth Data Analytics: How an Idea Has Made Its Way

It was at a time when "search" was only used in conjunction with metadata, because the data themselves - such as imagery and climate datacubes - were only meant for download. Databases wanted to understand tables only, all else was "unstructured" and not really considered.
At that time, an idea was phrased that databases might support massive multi-dimensional arrays as a new kind of attribute, with more semantics than just the usual BLOBs (Binary Large Objects). At a scientific database conference in 1998 the Principal Architect was told "it won't work, and if it will, nobody will need it". Still, the system named rasdaman (for "raster data manager") was built, and it did work.
Meantime, rasdaman has pioneered a new research field, Array Databases, which has been taken up by Stanford, MIT, and further high-profile resarch labs.
Successors like SciQL, SciDB, and Teradata have prototypes, too. The rasdaman concepts from the basis for the forthcoming ISO SQL extension with n-D arrays.
Being a university's basic research project in the beginning, meantime it is commercially supported by a spinoff company going by the same name, and among the customers having acquired licenses is the European Space Agency, German Weather Service, and more; also in Asia and the US requests are frequent, although an open-source edition is available in addition, estimated to a value of 10 million US$ by OpenHub.

In the talk we present the concepts, architecture, and history of Array Databases and show their impact and potential for a game change in agile analytics on Big Earth Data.