PETER BAUMANN
RASDAMAN GMBH

EXECUTIVE DIRECTOR

WHAT IS THE ROLE OF RESEARCH DATA WITHIN YOUR COMPANY ACTIVITY?

A: As we claim to be technology leader in our segment – flexible, scalable datacube analytics engines – it is of utmost importance to continuously be aware of (and extend) the state of the art. Therefore, own research, as well as university collaborations, are at the heart of our strategy.

Documentation is key to survival. We use state of the art tools that we select on demand, and we combine these to obtain documentation that is as seamless as possible.

Our partners and customers provide us with their data, which they like to see managed by our datacube engine. We have cloud providers available where we can scale demo and prototype services quickly on demand.

The results can encompass new or enhanced functionality as well as performance and scalability accelerations, which we offer through new product versions.

WHAT ARE THE RECURRING PROBLEMS THAT YOU ENCOUNTER?

We mainly capacitate on Earth data, specifically: massive multi-dimensional raster data (including, but not limited to satellite imagery and weather forecasts). Our data sources are partially open, partially proprietary. Public data is superficially well structured, but we regularly find all kinds of flaws inside the data, sometimes trivial and sometimes highly involved, which the original providers were not aware of.

Mostly, there is sufficient documentation and metadata available to enable meaningful sharing and reuse, at least for basic use. However, a plethora of local conventions impede exploitation – actually, bringing such data into standards-based services and thereby unleashing their potential is an important part of our business model.

WHAT HAPPENS WITH STANDARDS?

We try to evangelize the standards in the field, and we feel that gradually the added value is getting seen. In fact, we extensively try to boost standards through active work in the relevant standardization bodies (OGC, ISO, INSPIRE) and often as editor of such standards.

WHAT ARE THE MOST DIFFICULT DATA MANAGEMENT REQUIREMENTS FOR YOUR COMPANY? WHAT ARE THE RECURRING BOTTLENECKS?

Regarding data management requirements, we are developing data management software, and here issues that arise include: how to best homogenize the plethora of diverging data ("variety") and ensure their quality ("veracity")? How can we make complex queries efficient in face of data center federations? As a recurring bottleneck: finding suitably skilled staff.

HOW DOES YOUR ORGANISATION DECIDE ABOUT THE CORRECTNESS AND AUTHENTICITY OF DATA FOR ITS RE-USE? FOR INSTANCE, DO YOU TAKE VALIDATION MEASURES OR USE PEER REVIEW?

As we don't offer data we don't do that. However, we offer means to our customers by helping them formulating quality assessment queries that the system can answer (and monitor).

CAN YOU ENVISAGE ANY BETTER WAYS TO IMPROVE THE QUALITY OF DATA? DO THESE REQUIRE COOPERATION FROM OTHER STAKEHOLDERS?

Absolutely so: capturing quality-relevant parameters along the generating pipeline in sufficient accuracy and completeness; spotting overlaps in standards so that redundancy in metadata is not harmful any longer; clearly express quality requirements so that they can be assessed automatically through standing queries (and not black-box ad-hoc code).

HOW DO YOU ESTABLISH TRUST IN THE DATA, ESPECIALLY IF THEY ARE OPEN?

This is the responsibility of the service provider (i.e., our customer). On another scenario (partially open) actually we support by allowing service providers to define fine-grain access policies which the system subsequently guarantees.

DOES YOUR COMPANY MAKE USE OF RELEVANT RDA RECOMMENDATIONS OR EU TECHNICAL SPECIFICATIONS?

It will make use of the outcome of the Array Database Assessment Working Group (ADA WG). Relevant EU specifications include INSPIRE Coverage Download Services. We contribute to establishing these, and we support, eg, the INSPIRE standards through our implementation. To make this feasible we work hard on homogenizing relevant standards in OGC, ISO, and INSPIRE so as to minimize implementation effort (and customer's perceived complexity of our software).

THERE IS A GROWING AVAILABILITY OF AND DEMAND FOR DATA. DOES THIS CREATE RISKS? VALUE? BOTH?

Surely both. And I will not be enough to just do "some" analytics (which is a trending hype), but to put the same effort into a solid basis of ground truth data and allowing users to query them as they like ("any query, any time").

WHAT KIND OF PREDICTIONS CAN BE MADE FOR YOUR INDUSTRIAL SECTOR IN TERMS OF ITS DEMAND FOR DATA AND THE EMERGENCE OF NEW MARKETS AND BUSINESS MODELS?

With the advent of tools like ours, management of Big Data becomes easier for the data and service providers. This allows services to pop up faster, and service providers can concentrate on their core business of value-adding services. One consequence is that domain experts (like data scientists) can operate such services, the burden of substantial own software development (and, hence, finding skilled computer scientists) is relieved to some extent. We observed such cases – for example, EOfarm is a Greek startup utilizing open-source rasdaman for their precision farming services. We believe that in the near future startups will bloom that concentrate on their core business, bet on open standards, and do not try to reinvent silo solutions.

DATA INTEGRATION FROM DIFFERENT SOURCES CAN BE QUITE EXPENSIVE. IS THIS BUSINESS MODEL SUSTAINABLE? IF INDUSTRIES ADOPTED GLOBAL INTEROPERABILITY STANDARDS, WHAT WOULD THE GAINS BE COMPARED TO THE LOSSES?

That is an extra value data centers can provide. Earlier users needed to do that by themselves, but with the trend towards "analysis ready data" this task migrates into data centers. This makes sense: data centers have both expertise and compute resources to do this. And there is a distinct quality of service where they might differentiate their offering against their competitors: as open data as such are available from many places, users will to go where the best service quality is offered.

CAN AN OPEN DATA FORUM BE CREATED WHERE DATA IS OFFERED TO FOSTER EXCHANGE/TRADING? HOW WOULD IT NEED TO BE STRUCTURED?

In my humble opinion not. Just adding another data provider wouldn't improve the situation. We have already too many data providers, often having copies of the same data. What is needed is data for transforming the innovation occurring in computer science into user services that make a difference. To exemplify this: In our main application domain, Earth data, technology is available today which allows data center federations; data quality assessment via user queries; strictly based on open standards; allowing users to keep with their existing paradigms and clients; etc.

This might be some existing facility, or a new one indeed – critical is to offer state of the art services offering "analysis ready data".

Q: HOW DO DATA CREATION, ORGANISATION AND MANAGEMENT PRACTICES NEED TO BE CHANGED TO MAKE DATA INTENSIVE PROJECTS MORE EFFICIENT AND THUS ENABLE RE-USE EFFECTIVELY? COULD YOU GIVE US SOME EXAMPLES OF SUCCESS STORIES REGARDING DATA MANAGEMENT?

Less in house development that just reinvents wheels in a typically degraded manner; rather, more collaboration with computer science experts, in particular: database innovators. A data scientist is not equivalent to a computer scientist. A data scientist is not trained to *design* complex systems, but to *use* them. The idea of DevOps unfortunately leads people (in particular data centers) to believe that they have expertise along the full software value generation chain. This is not the case, as can be proven easily. Again, collaborations across experts in the different domains involved leads to best results.