

## DOLAP 2007 Summary

On November 7<sup>th</sup> 2007, I had the privilege of attending DOLAP 2007 in Lisbon, Portugal. Since my project “Optimizing Business Intelligence with the OODA Framework” is embedded in the OLAP field, it only seemed relevant to attend the conference that exposes the top of the academia in this field.

The conference was divided into five sessions, namely: “Data Warehouse Design”, “Physical Data Organization”, “Data Warehouse Processing”, “Spatio-Temporal Data Warehouses and Data Mining” and finally a panel discussion about “Research Challenges for DW and OLAP Seen from Industry and Academia” in which I was personally invited to participate.

In general, it seems that academia is very preoccupied with the challenges that arise when building a data warehouse in terms of physical storage, design and ETL, since this was the attention of all sessions with the exception of “Spatio-Temporal Data Warehouses and Data Mining”. However, even though my project has a strong focus on the usability of OLAP applications, I found the sessions very interesting and inspiring.

A concept in the design of data warehousing that I found fascinating was the presentation “Automating the Multidimensional Design from Ontologies” by Oscar Romero and Alberto Abello. The idea is to create a semi-automated method that points out multidimensional concepts from an ontology that then represents a business domain. Using this more abstract approach means that more data sources of varying types can be incorporated with less effort, and such heterogeneous data sources might be very relevant in a future that can bring us the semantic web and perhaps other external sources.

With regards to the processing of the ETL flow, which I have definitely done my share of on the physical level, I was impressed by the work of Vasiliki Tziouvara, Panos Vassiliadis and Alkis Simitis titled “Deciding the Physical Implementation of ETL Workflows”. This project proposes that we implement a declarative language for ETL in order to obtain detail independence and leave it to the system to implement the actual workflow, including the application of sorters. Again, from personal experiences, I have learned the hard way how important sorting and indexing can be during the implementation of a physical level ETL flow. This led me to suggest a “sort upon load” in the sense of applying a primary index that enforces a given sort before loading a table rather than sorting afterwards; this intuitively means that the sorting cost of each record will start small and end up at the level as the average sorting cost per record in a full table sort. I felt that my suggestion was taken in by the authors.

Another ingenious approach that can assist the generic optimization of an ETL flow was the categorization using a butterfly depiction of the ETL diagram. This approach allows more intelligent optimization decisions to be made, depending on the workload in the different phases of Extract, Transform and Load. A final note from the authors was that a commonly agreed benchmark was needed for ETL, since none of the commonly agreed tests so far made sense for this discipline. Such a benchmark would allow validation of performance improvements, and thus would logically stimulate the machine optimization of ETL flows.

Even though the presentation “Optimal Chunking of Large Multidimensional Arrays for Data Warehousing” by Ekow Otoo, Doron Rotem and Sridhar Seshadri was very different from my field, I still found it quite interesting and entertaining. This presentation revisited the work of Sarawagi & Stonebraker from 1994 and stated that their formula for calculating chunk overlap was wrong! The impact of this finding is actually quite significant, if we have a system that relies on their formula in an I/O access cost model; the presentation demonstrated potential errors of more than 100% inaccuracy! The presentation naturally ended up suggesting revised mathematical modeling for the chunking problem, and the overall presentation was done in a very vibrant and inspiring way.

In the final panel discussion, where I had the fortune of participating myself, I felt that one participant in particular saw some of the same usability challenges in OLAP that I see. Stefano Rizzi talked about "OLAP Preferences: A Research Agenda" and how we need to become better at capturing the data and visualization preferences of the business users, he even mentioned experiences from another project where users were asked what they wanted, and in this project commented with a smile that "Users lie". These are, in my opinion, indeed very good words to a very real usability problem, namely the fact that very few users know what they want until they see it. Rizzi's talk was, however, primarily focused on the impact of preferences as a parameter in an OLAP implantation that caters multiple users through an e-service model which seemed to be rather "push oriented"; nevertheless, preferences seemed to be a very viable approach to minimize information flooding and improve decision making speed in this scenario.

Finally, I had a chance to convey the message that I think the research community should focus more on broad usability of OLAP solutions, by proposing a heuristic approach to identify new desired technologies that can improve the speed and quality in any given decision making process through the OODA concept. I felt that people were very open-minded, and that the idea was taken seriously. In the following hours after the conference, I had the fortune of discussing these ideas even further with some of the leading scientists in the OLAP field.

This was my first DOLAP conference; I am definitely going to participate in the next one, and hopefully, I will have a chance to contribute with an article at that time...