Collaborative BI for dummies

Jérôme Darmont

Collaborative BI for dummies

Jérôme Darmont, ODBMS.org, August 2015

Business intelligence (BI) technologies such as data warehousing and on-line analysis processing (OLAP) used to necessitate heavy investment that was only achievable by big-sized organisms and companies. Nowadays, with cloud computing and on-demand payment of cheap storage and computing power, numerous cloud BI solutions as a service have popped up. Thus, initiating a BI project can be done at reduced costs. Yet, such tools are still aimed at BI specialists and remain out of the reach of small companies, non-governmental organizations and the average citizen.

Moreover, the cloud BI trend crosses with the growing demand for collaborative tools that could, e.g., enable users to easily share and reuse data and analyses. Mashing up private and public, open data is also a most probable requirement of cloud BI systems. Solutions do exist, mostly in the form of on-line spreadsheets such as Google Fusion Tables [1], but we could imagine further deploying a data warehousing process in software as a service mode.

To achieve this goal for the largest possible audience, a cloud BI service's interface must of course be very simple and user-friendly; but also intelligible, i.e., the terms used must belong to the user's vocabulary and thus presumably be extracted from data (IT jargon proscribed!); and finally actionable, i.e., navigating through, exploiting data and sharing analysis results must be straightforward. Thence, the data warehousing process should ideally be fully automated, and analytics assisted.

In this framework, classical data warehousing stages must still apply, albeit in a transparent way, starting with data integration. The NoETL (Extract, Transform, Load) approach [2] is a good start, but automatic data cleansing and integration is not enough. To actually crowdsource a data warehouse, semantics necessarily come into play. Yet, to the best of our knowledge, all related problems are not solved as of today. For instance, though extracting local ontologies from data is definitely feasible, automating ontology matching to achieve a global ontology remains tricky [3]. Moreover, so-called semantic data warehouses [4] require semantic data sources that are not always handy.

Once data has been cleaned, multidimensional modeling must take place. Although automatic data-driven warehouse schema generation has been quite well investigated [5], it mainly targets relational data, whereas source variety is likely in our scenario. XML data sources were considered [6], but since functional dependencies are not explicit as in relational data, the whole dataset must be traversed to infer schemas, which pose performance problems in an online context. Yet, emerging research on RDF analytics [7] might be a key to deriving multidimensional schemas from ontologies.

Finally, OLAP-like analysis must be available to users in a simple manner. Although there are efficient, open source (did I mention that, to be fully trusted, the whole system should be open?), web OLAP clients such as JPivot and its successors, such tools remain too complex for the non-specialist. Ideally, a sound, automatically selected visualization should be provided along with tabular data, and should allow dynamic navigation. Then, the collaborative aspect pops up again, and could be taken into account, e.g., through the recommendation of analysis paths [7].

In conclusion, truly simple, accessible and collaborative BI as a service is achievable. The remaining tasks are assembling the puzzle, but it might be more complex that it seems, since there is room for improvement in all above-mentioned individual components. Moreover, beside technical issues, privacy will most probably be a concern to would-be users. Imagine a scenario where several small business owners from some domain and/or area want to achieve insights from the data they own

(maybe complemented by open data), without disclosing individual data to competitors in the group. Then, a (presumably cryptographic) scheme should be devised to allow shared analysis while protecting source data.

[1] H. Gonzalez et al. Google fusion tables: data management, integration and collaboration in the cloud. In Proc. SoCC 2010, 175-180.

[2] M. Middelfart. The Inverted Data Warehouse based on TARGIT Xbone - How the biggest of data can be mined by "the little guy". In Proc. BIRTE 2013.

[3] P. Shvaiko& J. Euzenat. Ontology matching: State of the art and and future challenges. IEEE TKDE, 25(1):158-176, 2013.

[4] L. Bellatreche et al. Semantic Data Warehouse Design: From ETL to Deployment à la Carte. In Proc. DASFAA 2013, 64-83.

[5] C. Phipps & K.C. Davis. Automating data warehouse conceptual schema design and evaluation. In Proc. DMDW 2002, 23-32.

[6] M. Golfarelli et al. Data warehouse design from XML sources. In Proc. DOLAP 2001, 40-47.

[7] D. Colazzo et al. RDF Analytics: Lenses over Semantic Graphs. In Proc. WWW 2014, 467-478.

[8] J. Aligon et al. A collaborative filtering approach for recommending OLAP sessions. Decision Support Systems, 69:20-30, 2015.