Chapter 3.9
A Survey of Open Source Tools for Business Intelligence

Christian Thomsen
Aalborg University, Denmark

Torben Bach Pedersen
Aalborg University, Denmark

ABSTRACT

The industrial use of open source Business Intelligence (BI) tools is becoming more common, but is still not as widespread as for other types of software. It is therefore of interest to explore which possibilities are available for open source BI and compare the tools. In this survey article, we consider the capabilities of a number of open source tools for BI. In the article, we consider a number of Extract-Transform-Load (ETL) tools, database management systems (DBMSs), On-Line Analytical Processing (OLAP) servers, and OLAP clients. We find that, unlike the situation a few years ago, there now exist mature and powerful tools in all these categories. However, the functionality still falls somewhat short of that found in commercial tools. [Article copies are available for purchase from InfoSci-on-Demand.com]

INTRODUCTION

The use of Business Intelligence (BI) tools is popular in industry. However, the use of open source tools for BI is still quite limited compared to other types of software. The dominating tools are closed source and commercial. Only for database management systems (DBMSs), there seems to be a sizeable market where open source products are used in industry, including business-critical systems such as online travel booking, management of subscriber inventories for telecommunications, etc. (Yuhanna, 2006). Thus, the situation is quite different from, for example, the web server market where open source tools as Linux and Apache are very popular.

To understand the limited use of open source BI tools better, it is of interest to consider which tools are available and what they are capable of. This is the purpose of this article.

In the survey, we will consider products for making a complete solution with an Extract-Transform-Load (ETL) tool that loads data into a database managed by a DBMS. On top of the DBMS, an On-Line Analytical Processing (OLAP) server providing for fast aggregate queries will be running. The user will be communicating with the OLAP server by means of an OLAP client. We limit ourselves to these kinds of tools and do not consider, for example, data mining tools or Enterprise Application Integration (EAI) tools. Use of data mining tools is also of relevance in many BI settings, but open-source data mining tools have been considered in a recent survey (Chen, Ye, Williams, & Xu, 2007). EAI tools may have some similarities with ETL tools, but are more often used in online transactional processing (OLTP) systems. We focus on the individual components such that a "customized" solution is built, not the integrated BI suites – but the integrated BI suites are briefly described later.

The article is an updated version of a previous survey done in late 2004 (Thomsen & Pedersen, 2005). In comparison with the status in 2004, there are now mature and powerful open source tools available in all four categories (in 2004, only the DBMS category had sufficiently mature tools), so it is now for the first time possible to make a complete BI solution using only open source tools. More detailed findings are reported in the sections for each tool category.

The rest of the article is structured as follows. First, the method for conducting the survey is described. Second, the ETL category is described. Third, the DBMS category is described. Fourth, the OLAP server category is treated. Fifth, the OLAP client category is surveyed. Finally, the article describes the available integrated BI suites, before concluding remarks are offered.

CONDUCT OF THE SURVEY

To collect data about the tools, the Internet was searched for open source tools in each category. Some projects were left out of the survey if they only stated goals for *future* development and did not provide any working code at the moment. Also, projects for which no activity had taken place for years were left out. The presented data was found by inspecting the products' official homepages as well as their documentation (if any), mailing lists and forums. Finally, the source code was also inspected in some cases to clarify questions. For time reasons, the tools were, however, not evaluated by configuring and running each of them. The findings were collected from mid May to mid June 2008 with smaller updates in July 2008.

The data about the tools was collected carefully but nevertheless it is possible that certain information about products was misunderstood or not found and thus not considered correctly in this survey. The authors therefore disclaim any liability arising from omissions or errors and do not give any guarantees about completeness or accuracy. It should be emphasized that the authors are not involved in developing any of the described tools or any of their competitors and that the authors do not have any interests in recommending certain tools instead of other tools.

In the following, the criteria used for the evaluations of the considered products are described. Some general criteria are of relevance to all the categories of tools. Other more technical criteria are only of relevance to a specific product category.

Criteria for All Tool Categories

There exist many different open source *licenses* (Open Source Initiative, 2006), for example the GNU Public License and the Mozilla Public License. The different licenses vary widely with respect to what they allow and how modified