A Survey of Open Source Tools for Business Intelligence

Christian Thomsen and Torben Bach Pedersen

September 3, 2008

TR-23

A DB Technical Report

Title	A Survey of Open Source Tools for Business Intelligence
	Copyright © 2008 Christian Thomsen and Torben Bach Pedersen. All rights reserved.
Author(s)	Christian Thomsen and Torben Bach Pedersen
Publication History	Preprint of: Christian Thomsen and Torben Bach Pedersen: "A Survey of Open Source Tools for Business Intelligence" to appear in *International Journal of Data Warehousing and Mining*, 2009.

For additional information, see the DB TECH REPORTS homepage: ⟨`www.cs.aau.dk/DBTR`⟩.

The DB TECH REPORTS icon is made from two letters in an early version of the Rune alphabet, which was used by the Vikings, among others. Runes have angular shapes and lack horizontal lines because the primary storage medium was wood, although they may also be found on jewelry, tools, and weapons. Runes were perceived as having magic, hidden powers. The first letter in the logo is "Dagaz," the rune for day or daylight and the phonetic equivalent of "d." Its meanings include happiness, activity, and satisfaction. The second letter is "Berkano," which is associated with the birch tree. Its divinatory meanings include health, new beginnings, growth, plenty, and clearance. It is associated with Idun, goddess of Spring, and with fertility. It is the phonetic equivalent of "b."

A Survey of Open Source Tools for Business Intelligence

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 paper, we consider the capabilities of a number of open source tools for BI. In the paper, 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.

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 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 paper.

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 would also be of relevance in many BI settings, but data mining is a more advanced feature which should be considered in future work. 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 paper 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 paper 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 paper 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 relevant for all the categories of tools. Other more technical criteria only apply for 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 source code can or must be distributed. Although it is outside the scope of this paper to describe the different open source licenses, it is of interest to see which licenses are used for the particular products.

For a potential user, it is important if a certain tool can be used with her existing *platform*. It is thus of interest to consider with which hardware and software platforms the tools can be used.

For many professional users, it is important to know whether *commercial support, training, and consulting services* are available for a product, and the survey therefore considers these aspects. A related considered issue is the type and amount of *documentation* available. Many open source projects have strong user communities using *forums and/or mailing lists*. The survey therefore also considers if such active forums or mailing lists exist for the products.

Criteria for Extract-Transform-Load Tools

For an ETL tool, it is investigated if the tool is for *relational OLAP* (ROLAP) where relational database tables are loaded or if it is for *multidimensional OLAP* (MOLAP) where multidimensional cubes are loaded. The supported data sources and targets are obviously also of interest.

In many DW environments, it is of great practical interest to be able to load only the changes made to the source data since the previous load. The survey therefore considers the possibilities for doing such an *incremental load*. The survey also considers *how an ETL job is specified*, for example by means of a graphical user interface (GUI) or an Extensible Markup Language (XML) file. Also the possibilities for doing *transformations* and *data cleansing* are considered – both with respect to predefined transformations and user-defined transformations.

Due to the large data volumes in data warehousing, *parallel job execution* is also of great practical interest and it is investigated if the tools support parallelism.

Criteria for Database Management Systems

For DBMSs there are many interesting things to consider. In this paper, the scope is limited to investigate features and possibilities that are relevant to data warehousing. This includes investigating whether the DBMS can handle *large datasets* of many gigabytes. Related issues to consider for the DBMSs are their support for *materialized views, bitmap indices,* and *star joins* which all can improve performance for DW applications.

Possibilities for *replication* and *partitioning* are also of interest for many DW environments and the survey also considers if the tools support these features. Finally, it is considered which *programming languages* are supported for stored procedures/user-defined functions.

Criteria for On-Line Analytical Processing Servers

For an OLAP server, the survey investigates whether it is a *ROLAP* or a *MOLAP* server. It is also considered which *data sizes* it aims at handling and which underlying *data sources* it can be used with if any (for example a specific DBMS like MySQL). For performance reasons it can be very beneficial for an OLAP server to use pre-computed *aggregate tables*, and the possibilities for this are also investigated here.

It is also investigated how the user performs the *specification of cubes*, for example by means of a GUI or an XML file. Finally, it is considered what *application programming interface (API)* and *query language* the OLAP server offers.

Criteria for On-Line Analytical Processing Clients

For an OLAP client, the survey considers which *OLAP server(s)* the client can be used with and which *query language* it uses/generates. Further, the types of supported *reports* are investigated.

EXTRACT-TRANSFORM-LOAD TOOLS

This section presents the ETL tools that were found in the survey. In previous work (Thomsen & Pedersen, 2005), the category of ETL tools only had few possibilities and was considered to be the least mature and most difficult to use. At the time of this writing, many more tools are available and some of these are quite mature. All of the described tools but Pequel are implemented in Java and can thus be used on many different hardware and software platforms. In this and the following sections, the tools are presented in alphabetical order.

Apatar

Apatar (Apatar, 2008) is a data integration and ETL tool developed by the company also bearing the name Apatar. This survey considers Apatar version 1.1.9 from May 2008. Apatar seems to have a very fast release cycle as version 1.1.0 was released in October 2007, and version 1.1.10 in June 2008. Apatar is released under a dual-licensing scheme and is available under the GNU General Public License (GPL) or under a commercial license, if desired. The development company offers training, support, and consulting. The documentation exists in four PDF files (around 40 pages in total) as well as in some wikis. Further, some user forums exist.

Apatar is ROLAP-oriented and has direct support for a wide selection of relational DBMSs (as well as generic JDBC support). Further it supports file formats such a comma separated values (CSV) and Excel and ERP and CRM systems (Compiere ERP, SalesForce.com, and SugarCRM). The specification of a job is done in a GUI. However, it is not possible to do an incremental load. It is not possible to run jobs in parallel yet.

Apatar has built-in data quality tools for verification of US addresses, phone numbers and email addresses. It is possible for the user to define transformations in Java, although this is not as simple as in some of the other tools. To make her own transformation available in Apatar, the user must define two classes (both inheriting from provided base classes) and edit an XML file describing the available plug-ins.

Clover.ETL

Clover.ETL (*Clover.ETL*, 2008) is developed by OpenSys and Javlin and is offered under either a GNU Library General Public License (LGPL) or a commercial license. Support and consulting can be bought from the above-mentioned companies. Here, version 2.4.6 of Clover.ETL is considered (2.4 was released in Feb. 2008, 2.4.7 was released in June 2008). Unlike the previously described ETL tool, Clover.ETL does not have an open-source GUI. A closed-source GUI exists, but is only free of charge if not used commercially. So, for a solution fully based on open source, the user has to specify the ETL job in XML when using Clover.ETL. A free 181-page manual exists for the GUI, but for Clover.ETL itself, the user has to settle with the wiki-documentation and a User's Guide consisting of 104 slides from a presentation. Further, two forums exist.

Clover.ETL is a ROLAP tool and transfers structured data from different DBMSs and file formats. The user can create transformations in Java (together with XML descriptions of them) or in Clover.ETL's own *TL* language. Clover.ETL supports parallel execution and, for some DBMSs, also bulk-loading, while no support for incremental load was found.

ETL Integrator

ETL Integrator (*JBIWiki: ETLSE*, 2008) is an ETL tool developed by Sun Microsystems. It has a service engine that makes ETL operations available as web services and further it has an ETL editor which is integrated into the Netbeans integrated development environment (IDE) version 6.1. It is released under the Common Development and Distribution License (CDDL). No information about commercial support specific for ETL Integrator was found. But Sun will provide commercial support for its upcoming integration platform GlassFish ESB (Sun, 2008) which is built on the OpenESB project which ETL Integrator is part of. The documentation found consists of design documents, wikis, and video tutorials. Also, the User Guide is a tutorial. Further, forums for the OpenESB project exist, but only few postings are related to ETL Integrator.

ETL Integrator is a ROLAP tool outputting to relations. It supports different relational DBMSs as sources as well as different file formats and OpenESB components for connecting to ERP/CRM systems. The tool is integrated into the Netbeans IDE and ETL jobs can be specified graphically from there. ETL Integrator supports incremental load as well as parallel execution of parts. Also bulk loading is supported. In the descriptions of ETL Integrator it is said that its editor "has many predefined transformations as well as cleansing operators/functions" and further that it is possible to add user defined functions. It does offer name and address parsing and normalization, but these operators depend on SQL calls to the database except for flat files for which an internal engine is used.

KETL

KETL (Kinetic Networks, 2008) – not to be mistaken for Kettle described below – is developed by Kinetic Networks from which support can also purchased. The latest version of KETL is 2.1.24 from April 2008. The oldest generally available release in 2.1 series is 2.1.12 from April 2007. KETL is partly released under the GPL license and partly under the LGPL license. The homepage for KETL states that the documentation currently is being overhauled and that only the Installation Guide has already been updated. Older versions of the documentation are still available in the mean time. However, these (37 pages Administration Guide and 24 slides in a Training Presentation) fail to describe how ETL jobs are defined in the used XML language. More than 60 example XML files are available but they also lack documentation.

KETL is ROLAP oriented and can be used with JDBC sources (KETL has special support for three DBMSs) and flat files as well as XML files. The user must specify the ETL jobs in an XML file. Transformations are apparently possible, but due to the missing documentation it is not clear how to create them. Likewise, it is unclear if incremental loads are supported. KETL is capable of executing parts in parallel.

Kettle / Pentaho Data Integration

Kettle (Pentaho, 2008c) started as an independent open source ETL project, but was in 2006 acquired by Pentaho to be included in the Pentaho BI suite (Pentaho, 2008b). Thus Kettle is now also branded under the name Pentaho Data Integration. This survey considers version 3.0 of Kettle (3.0.0 was released in Nov. 2007, 3.0.4 in June 2008), but version 3.1 is expected to be released soon. Kettle is released under the LGPL. Kettle has a graphical designer for jobs and transformations. This designer has a manual of 274 pages and also 40 pages with frequently asked questions (FAQs) and answers. Further, very active forums exist (more than 22,000 posts in 5,000 threads in the last 2½ years). Pentaho offers support, training and consulting, but many Pentaho partners also offer such services.

Kettle is ROLAP-oriented, but another open-source project (Gimenez & Lopez, 2007) provides a plug-in that enables Kettle to output data to the Palo MOLAP server. Kettle supports around 35 different DBMSs (also generic JDBC and ODBC) as well as a variety of flat files. A 3rd party SAP connector is also available, but it is not yet ready for version 3.0 and it is commercial. Incremental load is possible in the sense that Kettle logs when a job was executed and it is possible to use this timestamp in the queries to only select new data. Furter, an "insert or update" step is available. ETL jobs are specified in a GUI. Kettle is shipped with more than 80 predefined transformations, and further the user can implement transformations in JavaScript. Kettle also supports debugging of these with breakpoints etc. It is possible to use "clustering" where a transformation step can be split into parts that are executed on distinct servers. Apart from that, parallel jobs are not supported in version 3.0, but planned for the up-coming 3.1 release. Junk dimensions and slowly changing dimensions of type 1 and 2 (Kimball & Ross, 2002) are supported by Kettle and for some DBMSs (experimental) bulk-loading can be applied.

Octopus

Octopus (Together Teamlösungen, 2007) is an ETL tool from Enhydra.org with the LGPL license. Here version 3.6-5 from Oct. 2007 is considered (3.6-1 was released in June 2006). A manual of 122 pages exist for Octopus and further a mailing list exists. The latter is, however, not very active and has had 2 posts in the first half of 2008 and 25 in all of 2007. Commercial support for Enhydra.org's products is available from Together Teamlösungen and other commercial vendors.

Octopus is also ROLAP-oriented and uses JDBC to connect to data sources and targets. ETL jobs are specified in XML files (a GUI for generating skeletons for these XML files as well as file dumps of the database content exists, though). Octopus can update values, but does not support incremental loads in more advanced ways. It is possible to implement transformations in Java and JavaScript. Further, Octopus has a few predefined transformations for setting a default value, ensuring a maximum length of a string, and for correction of foreign key values. Parallel job parts and bulk loading are apparently not supported.

Palo ETL Server

Palo ETL Server (Jedox, 2008a) is developed by Jedox AG from which commercial support and training options are also available. Palo ETL Server is released under the GPL. Version 1.0 was released in April 2008 and followed by 1.1 in July 2008. A manual with 56 pages is freely available. A forum with some activity also exists.

Palo ETL Server is the only considered ETL tool that is MOLAP-oriented as it is made for loading data into the Palo MOLAP server also created by Jedox AG (described later). It loads data from JDBC sources and CSV files, LDAP servers, and MOLAP cubes/dimensions. Jobs are specified in XML (a GUI is planned for a future release), and transformations can be implemented in Java. The jobs can be parameterized such that incremental loads to some degree can be supported. Parallel jobs are not supported.

Pequel

Pequel (Gaffiero, 2007) is the only considered ETL tool that is not written in Java. It is implemented in Perl and runs on UNIX-like platforms and on Windows using Cygwin. Pequel's license is GPL. Version 3.0.94 of Pequel is considered here, but as the documentation is not complete, documentation for version 2.4.6 has also been used. The documentation for version 3.0 consists of a Programmer's Reference (30 pages) and a Pequel Type Catalog (54 pages) which are both quite technical and serve as documentation for the Pequel source code. For version 2.4, a User Guide (72 pages) also exists but with many sections that have been left empty. Forums exist, but the traffic is low (no posts from Sep. 2007 to June 2008). No commercial support offerings were found in this survey.

Pequel generates Perl and C code for the load job. It is mainly targeted at processing files to generate other files. However, support for relations using Perl's DBI module also exists. A job is specified either by means of a Perl API or by means of an XML file. Data conversion and rejection of records (based on regular expressions) are possible. Further, the user can use Perl's functions and operators. It is possible to distribute read data records to different Pequel processes and merge them again and in this way execute parts in parallel.

Scriptella

Scriptella (Kupolov, 2008) is an ETL and script execution tool. It is released under the Apache License. Here, version 1.0beta is considered. A manual of 23 pages exist as well as a forum. The developer offers commercial support and consulting.

Scriptella is intended for ROLAP-use (as well as output to files) and JDBC drivers for different DBMSs are included together with JDBC drivers for different flat files, XML files, and LDAP servers (but any other JDBC driver can also be used). ETL jobs are specified in an XML file where Java or any scripting language compatible with the JSR-223 standard can be used directly. In a Scriptella script, data rows can be fetched from multiple connections by queries. In the script it is specified what to do for each

row in the query result (e.g., perform an SQL statement using the values from the result of the outer query, or apply certain transformations). It is possible to nest queries and scripts written in different languages while still sharing variables. Scriptella's focus is on simplicity and it does not have out-of-the-box transformations available or incremental load support. On the other hand, the simplicity of using code for user-defined transformations and logic should be noted.

Talend Open Studio / JasperETL

Talend Open Studio (Talend, 2008) is developed by Talend that also offers support, training and consulting. The GPL-licensed Talend Open Studio is also included in the open-source BI package from JasperSoft (JasperSoft, 2008), there under the name JasperETL. The survey considers version 2.3 of Talend Open Studio (2.3.0 was released in Feb. 2008, 2.3.3 in May 2008), but since the data was collected, version 2.4.0 has been released. Compared to the other open source ETL tools, it has a large printable documentation with a User's Guide (161 pages) and a Reference Guide (550 pages). However, personal information, like name and email, must be given to access this documentation.

Talend Open Studio is primarily ROLAP-oriented, but output to the Palo MOLAP server is also supported. Talend claims to support more than 100 different source systems. These include different DBMSs, files, and web services, Subversion logs etc. Further CRM systems like SugarCRM, CentricCRM, SalesForce.com, and VtigerCRM are supported. The ETL jobs are specified in a GUI. Like Pequel, Talend Open Studio generates code for a stand-alone ETL application. The generated code is Java or Perl. It is possible for the user to specify transformations (also in Java or Perl). Further, Talend Open Studio comes with a set of predefined transformations, including six for data quality (matching, replacing, etc.) and from version 2.4 also name and address parsing (but only when Perl code is generated). The generated code can execute parts in parallel, but the parallelism support in Talend is still being extended. Talend supports slowly changing dimensions and bulk load, while incremental load is done by use of look-ups possibly followed by inserts or updates.

Summary

Compared to the previous survey (Thomsen & Pedersen, 2005), many more open source ETL tools are available today. Also, the quality of the existing tools seems to have increased a lot in the mean time. Four out of the ten tools include GUIs where ETL jobs are specified and for one of the remaining tools, a closed source (but free for non-commercial use) GUI exists.

Nine out of the ten described tools are implemented in Java and one in Perl and all the tools thus run on many different platforms. The tools are primarily targeted at ROLAP (Palo ETL Server being the exception) and in general support a variety of relational DBMSs as well as generic JDBC, common file formats like Excel and XML. Some of the tools can also extract data from ERP and CRM systems, but not all of these connectors are for free.

The most notable tools are Kettle and Talend which both have large user communities, comprehensive documentation, and many features and are included in BI suites.

DATABASE MANAGEMENT SYSTEMS

In previous work (Thomsen & Pedersen, 2005), the category of DBMSs was considered to be the most mature of the considered categories. Also, for the current survey many mature DBMSs have been found. More open source DBMSs than those described below exist, but they have a low visibility compared to the described ones and/or are mainly for use-cases that are less relevant for BI-usage, e.g., for embedded usage with smaller data sets.

Firebird

Firebird (Firebird Project, 2008) is based on the code base for the commercial DBMS InterBase version 6.0. The current version of Firebird is 2.1 (from April 2008). Firebird uses two licenses which are both similar to the Mozilla Public License (MPL). Firebird runs on Windows, Linux, FreeBSD, and MacOS X. Binary releases for Solaris and HP-UX are not yet available for version 2.1, but are available for the 2.0 series from Nov. 2006. Firebird is a commercially independent project, but a large part of the development is done by the company IBPhoenix which also offers support, training, and consulting. The documentation for Firebird is still not complete. On the homepage it is stated that the project is working on full user's and reference guides, but that the current documentation still consists of the manuals for InterBase 6.0 combined with the Firebird release notes that describe changes made to the Firebird code. All changes between InterBase 6.0 and Firebird 1.5 are documented, but updates for 2.0 and 2.1 are still in preparation. Apart from this documentation, different (rather short) guides and manuals exist in PDF and HTML format. Active mailing lists exist for the project (the support list has had around 95,000 posts since Nov. 2000).

On-disk bitmap indexes are not supported, but Firebird can combine indexes and form bitmaps in memory. Firebird does not support materialized views, star joins or partitioning. Replication is not available in the Firebird distribution itself, but (both commercial and open source) 3rd party tools provide this. With respect to data sizes, it should be noted that tables are limited to 2 billion rows in Firebird, but that there is no limit on the byte-size of databases (the largest known database has more than 11TB data). It is reported that the current Firebird has problems with scalability on computers with multiple CPUs, but these problems will be solved in the coming Firebird version 3.0.

The user can implement stored procedures (SPs) in Firebird's procedural language PSQL. Further user-defined functions (UDFs) can be loaded from external shared object libraries and, thus, the user can implement in C, C++, Delphi, etc.

Ingres Database

Ingres Database (Ingres, 2008) developed by Ingres is available under a commercial license (this is the "Enterprise Edition" of Ingres Database) or under the GPL (the "Community Edition"). The current version is Ingres 2006 Release 2 but Release 3 is available in a beta version. Ingres supports its products for 15 years, but support and maintenance from Ingres is only available for the Enterprise Edition (however other independent companies also provide support and training). Ingres

Database runs on Windows and a variety of different UNIX-like platforms. Active community forums exist and Ingres also offers 25 free manuals with close to 8,000 pages in total.

Materialized views, bitmap indexes, and star joins are not supported. Multi-master replication and partitioning (based on range, value, or hash) including sub-partitioning are supported out-of-the-box in Ingres Database. With respect to scalability, Ingres claims that Ingres Database is capable of handling many terabytes of data easily.

SQL can be used for stored procedures and further user-defined functions can be implemented in C.

LucidDB

LucidDB (*LucidDB*, 2008) is developed by the software company LucidEra and the non-profit organization The Eigenbase Project. The LucidDB server is licensed under the GPL while the LucidDB client is licensed under the LGPL. The newest version of LucidDB is version 0.7.3 from March 2008. On LucidDB's homepage, it is stated that LucidDB is "purpose-built entirely for data warehousing and business intelligence". This is in contrast to the other considered tools apart from MonetDB (see below). LucidDB and MonetDB are also the only considered column-stores (Abadi, Madden, & Hachem, 2008). In a column-store, all data tables are split vertically at the physical layer such that each column is stored on-disk independently of other columns. This is different from traditional row-stores where data from different columns is stored together in rows.

LucidDB runs on 32 and 64 bit Linux and on 32 bit Windows (using Cygwin). LucidEra does explicitly not intend to sell support or commercial licenses for LucidDB. The documentation consists of relatively short wiki manuals and tutorials. Further, there is a mailing list but this has had less than 100 posts from May 2007 to May 2008.

LucidDB supports B-tree and bitmap indexes. LucidDB chooses itself which indexes to create and can also combine the two types. Star joins are also supported, while partitioning and replication is not supported. Support for materialized views is planned for a future release. It is reported that LucidDB has been tested with 10GB TPC-H data, but the performance results have not been found during this survey. User-defined functions can be created in Java. It is also possible to wrap external data sources like files or tables from another DBMS and use them as traditional tables from LucidDB.

While LucidDB offers many features relevant for data warehousing, it should be noted that it is still not a mature DBMS. For example, foreign keys, sub-queries, transaction handling, and support for altering table definitions are still missing.

MonetDB

MonetDB (CWI, 2008) is the second column-store considered in this survey. Like LucidDB, it is not a DBMS made for on-line transactional processing (OLTP) with highly concurrent workloads. Instead the focus is on efficient handling of query-intensive access patterns. MonetDB is developed by the research institute CWI and has a license similar to the MPL. It runs on Windows and different UNIX-like operating systems. In the implementation, care has been taken to use the hardware very

efficiently. In this survey the "Feb2008" release of MonetDB is considered (the "Jun2008" release became available at the end of June 2008). A manual with 260 pages is provided for MonetDB. Further, a manual of 113 pages exist for the SQL part of MonetDB. A commercial spin-off that offered support existed, but this has been acquired by another company and no commercial support for the current MonetDB has been found during this survey. An active mailing list also exists.

Like LucidDB, MonetDB itself picks which indexes to create. However, bitmap indexes seem to be unsupported. Partitioning, replication, and materialized views are also currently unsupported, but future additions are planned for these areas. The user can define stored procedures in SQL as well as external functions in MonetDB's proprietary MAL language and in C.

MySQL

MySQL (MySQL, 2008), developed by MySQL (now owned by Sun Microsystems), is available under a commercial license or under the GPL. It can be downloaded in two versions: The "Community Server" which is free and the "MySQL Enterprise" edition which is not free, but for which extra features and commercial services exist. The latest production release of the community server is 5.0.51 from April 2008 (the first production release from the 5.0 series was from October 2005), and version 5.1 is available as a release candidate. Sun offers a wide range of commercial support, consulting, and training. It is reported that MySQL-based data warehouses larger than 30 terabytes exist. MySQL runs on Windows and a large collection of UNIX-like systems. The manual has 2071 pages and further very active forums exist.

There is no support for star joins or materialized views. Also partitioning is not supported in the 5.0 series, but is available in the upcoming 5.1 series where range, list, hash, and key partitioning are supported. Statement-based master-slave asynchronous replication has been available in MySQL since version 3.23, but from release 5.1, row-based replication is also available. Synchronous replication is also possible using MySQL Cluster, but as stated in the manual "all live data storage is done in memory" then, but from version 5.1 non-indexed columns can be saved on-disk. On-disk bitmap indexes are not supported in MySQL, but MySQL can perform an index merge where a bitmap is built in-memory. The user can implement stored procedures in SQL and user defined functions in C/C++.

PostgreSQL

PostgreSQL (PostgreSQL Global Development Group, 2008) is released under the BSD license. The newest version is 8.3.3 from June 2008 (8.3.0 was released in Feb. 2008). PostgreSQL runs on Windows and on a large collection of UNIX-like operating systems. PostgreSQL's development is led by its community and not by a single company. Due to its non-restrictive BSD license, several (open source and commercial) derivatives exist. For example, Netezza (Netezza, 2008), EnterpriseDB (EnterpriseDB, 2008), and Greenplum (Greenplum, 2008) offer PostgreSQL-based products. Several companies also offer commercial support, training and consulting. The PostgreSQL manual consists of 1908 pages and very active mailing lists also exist as well as Internet Relay Chats (IRCs) in more

languages. It is reported that databases larger than 4 terabytes are used in production environments.

Materialized views and star joins are not supported in PostgreSQL. On-disk bitmaps are not supported yet, but they are planned for inclusion in a future version. PostgreSQL does already support creation of an in-memory bitmap when combining other, existing indexes. Partitioning is to some degree supported by means of PostgreSQL's table inheritance features. The user has to create each partition manually and to create logic for redirecting inserts to the correct partition. The solution does not work well with parameterized queries and enforcement of integrity constraints. In the current PostgreSQL distribution, replication is not supported out-of-the-box as this deliberately has been left to let 3rd party tools offer competing solutions. As this standpoint is now considered to hinder acceptance of PostgreSQL, it has been decided to include simple asynchronous replication in the future standard distributions. But for the current PostgreSQL and for more advanced use-cases in the future, existing 3rd party tools (for example, Slony-I (Slony Development Group, 2008) and Pgpool-II (Ishii et al., 2007)) already offer replication.

PostgreSQL offers several languages for stored procedures: PL/pgSQL, PL/Tcl, and PL/Python. Further, it is possible to add new language support (for example, Java is supported this way). The user can also create external functions in C libraries.

Summary

Many open source DBMSs are available and overall they have reached a high maturity, also with respect to commercial support etc. Some features offered by leading commercial DBMSs are, however, still missing from the open source DBMSs. For example, none of the considered DBMSs support materialized views and most of them do not support on-disk bitmaps, star joins, and partitioning. Nevertheless, the open source DBMSs are usable for many production DWs and are being used for large BI projects in industry. Ingres Database, MySQL, and PostgreSQL are notable for their documentation, large user communities and rich feature sets.

<div align="center">

ON-LINE ANALYTICAL PROCESSING SERVERS

</div>

Not many open source on-line analytical processing (OLAP) servers are available. For this survey, only two open source OLAP servers with running code were found. This might be due to the success of the first of them, Mondrian. This server is included in the leading open source BI packages and uses de-facto standards and is a very popular choice for ROLAP usage.

Mondrian / Pentaho Analysis Services

Mondrian (Pentaho, 2008a) started as an independent open source project developing an OLAP server in 2002. In late 2005, Mondrian joined forces with Pentaho and is now being developed as part of Pentaho's BI package (Mondrian can be downloaded and used without the rest of the Pentaho software). Mondrian is a relational OLAP (ROLAP) server. The most recent version of

Mondrian is 3.0.3 from May 2008 (the 3.0 series was released in March 2008). It is released under the Common Public License (CPL). As Mondrian is implemented in Java, it runs on many platforms and uses JDBC such that it can be used with most DBMSs. The documentation consists of HTML pages with relatively large contents – close to 200 pages of text in printing. Further, active forums exist. Commercial support, consulting, and training are available from Pentaho and its partners.

The Mondrian project is involved in the standardization work for the olap4j specification which is intended to become a common API for OLAP servers (a kind of JDBC for multidimensional data). The primary API to Mondrian is olap4j. Queries to Mondrian are expressed in the de-facto standard in industry, MultiDimensional eXpressions (MDX) (Spofford, Harinath, Webb, Huang, & Civardi, 2008). Cubes are specified by means of an XML file. GUIs for creating these XML files do exist, but they are still described as incubator projects and have version numbers starting with 0.

With respect to scalability, it is stated in the FAQ for Mondrian, that large data sets can be handled if the underlying RDBMS can handle them, since all aggregation is delegated to the DBMS. Mondrian does have support for use of pre-computed aggregate tables. Users are reporting that Mondrian performance is good even when handling hundreds of gigabytes data with hundreds of millions rows in industrial settings.

Palo

Palo (Jedox, 2008b) is a multidimensional OLAP (MOLAP) server developed by Jedox AG. It is available under a commercial license or under the GPL. Windows and Linux are the primarily supported platforms. Commercial support and consulting is available from Jedox. A manual of around 376 pages exist, but costs €29.50. Active forums for Palo also exist. For this survey, version 2.0 was considered, but version 2.5 was released in early July 2008.

Palo loads its data set completely into memory and thus the memory on the host computer limits the supported data sets. Proprietary programming interfaces to communicate with Palo exist for Java, .NET, PHP, and C. There is also a free, but closed-source add-in for Microsoft Excel – the manual for Palo even states that "Palo was developed for Microsoft Excel". Using the plug-in, Palo-specific constructs like PALO.DATAC(…) can be used in formulas. The Excel add-in can also be used to specify cubes.

Summary

Only two open source OLAP servers with running code were found. They are targeting different segments as Mondrian is a ROLAP server which can handle large data volumes while Palo is a memory-based (and thus also memory-limited) MOLAP server. The Mondrian OLAP server seems very popular as it is not only included in the developer company's (i.e., Pentaho's) BI suite, but also in other open source BI-suites. Both of the OLAP servers are used in industry for BI projects.

ON-LINE ANALYTICAL PROCESSING CLIENTS

Compared to the previous survey (Thomsen & Pedersen, 2005) where only two open source OLAP clients were found, many more products are available at the time of this writing. From the previous survey, JPivot is still actively developed and parts of it are also used in other products.

FreeAnalysis

FreeAnalysis (BPM Conseil, 2008) is developed by BPM Conseil from which commercial support, training, and consulting also can be bought. From the downloadable code, the license is not clear. Previously the MPL and a license derived from the MPL have been used. A (currently empty) project created for FreeAnalysis on Google Code also states that the license is MPL. Version 1.14 from June 2008 is considered here. FreeAnalysis is implemented in Java and can be used as a stand-alone application or as a web application. A manual of 13 pages is available in French and another two-page document with an example connection specification is also available. It is, however, said that a URL to a manual will be given to those who subscribe to FreeAnalysis's mailing list.

FreeAnalysis works with Mondrian and servers that use XML for Analysis (XMLA) (Microsoft & Hyperion Solutions, 2002) and FreeAnalysis generates MDX queries. Reports in FreeAnalysis consist of pivot tables and graphs (the JFreeChart package (Object Refinery Limited, 2008) is included in the distribution). FreeAnalysis does also support generation of cube definitions for the Mondrian OLAP server (described above).

JPalo Client and JPalo Web Client

JPalo Client (Tensegrity Software, 2008) is a stand-alone application while JPalo Web Client is an Ajax-based web-application. Both products are developed by Tensegrity Software and available under a commercial license or the GPL. The most recent version for both is 2.0 from June 2008. JPalo is here used to refer generically to JPalo Client and JPalo Web Client. As suggested by the name, JPalo works with the Palo server. It offers modeling and administration for the Palo server as well as tabular exploration of the data (in JPalo Client the data can also be represented in charts). But JPalo can also be used with XMLA-enabled sources. In any case, the user can explore the data in the GUI without typing queries manually. JPalo is implemented in Java and JPalo Web Client does not require any installation on the end-users computer. During this survey, no manual for JPalo was found. Forums for JPalo exist and have some activity.

JMagallanes Olap & Reports

JMagallanes Olap & Reports is an open source component of Grupo Calipso's JMagallanes suite (Grupo Calypso, 2006) where other components are closed source. The latest version is 1.0 from May 2006. The development company sells support and installation services. The documentation consists of videos and some rather short HTML documents. Forums do exist, but have had little activity lately. JMagallanes Olap & Reports is distributed under the BSD license.

The open source part of JMagallanes reads data from JDBC sources as well as XML and Microsoft Excel. Apart from static reports based on JasperReports, the user can explore the data in pivot tables and with charts based on JFreeCharts. Data can be exported to PDF, XML, Excel, CVS and HTML. It is possible to schedule reports and have them sent by email.

JPivot

JPivot (Tonbeller, 2008), developed by Tonbeller, is among the first open source OLAP clients and parts of its code is also used in some of the more recent clients. JPivot is also included in BI suites like those from Pentaho and JasperSoft. JPivot is licensed under the CPL. The newest version of JPivot is version 1.8 from March 2008. It is a web-application implemented in Java and JavaServer Pages (JSP) and thus the end user uses a normal web browser to explore the data. Some HTML documentation exists for the JSP tags and for the Java source code. Apart from this, active forums exist for JPivot (and for the BI suites that include it).

JPivot was originally developed to be used with Mondrian, but it can now be used with other XMLA-enabled servers as well (the JPivot project is also involved in the olap4j specification). JPivot generates MDX queries. The user explores data by means of pivot tables and graphs (the latter are based on JFreeChart) and can choose to enter MDX queries manually. Support for exporting data to Excel format and PDF is present.

JRubik

JRubik (*Introduction to JRubik*, 2005) is an OLAP client which is based on JPivot components. JRubik is, however, a stand-alone Java application while JPivot is a web-application. It is licensed under the CPL. The documentation is in HTML format and is relatively short. A forum with some activity exists. Version 0.9.4 was released in December 2006. That version only worked with Mondrian. In May 2008, a new version of JRubik using the new olap4j definition was released (with version number 0.0.0). In both of these version, the tool generates MDX queries.

The user explores data by means of pivot tables, charts (again from JFreeChart), or in a map component. Tabular data can be exported to PDF, XML, HTML, and Excel. Chart data can be exported to XML and HTML.

OpenI

OpenI (OpenI.Org, 2008) was developed by the company Loyalty Matrix from which commercial support was also available. The company has now been acquired by another company and its technology is being integrated into a closed source application. It is therefore not known if the OpenI project will continue. The existing source code is available under a license similar to the MPL. The documentation is in HTML form and rather short. Forums exist and have been active. Lately the activity has been limited, though.

OpenI is implemented in Java and thus runs on different platforms. OpenI connects to XMLA sources and generates MDX queries. The user explores data in tables and charts which are based on components from JPivot and JFreeChart, respectively.

REX

REX (*SourceForge.net: REX*, 2007) is an – untraditional – abbreviation of "wa*r*ehouse *ex*plorer". The latest release is 0.7 from November 2007. It is released under the LGPL. The documentation consists of a tutorial. Some forums do exist, but they are not very active with less than 60 messages in three years. No commercial support offerings for REX were found during this survey.

REX is implemented in Java and runs on many different platforms. It works with XMLA sources and generates MDX queries. Data is browsed in pivot tables and charts using JPivot and JFreeChart components, respectively.

Summary

All the considered OLAP clients are implemented in Java. Six of the eight tools run directly on the client, while the two remaining run on a webserver. JPivot is widely spread since it is included in different BI suites. JPivot components are also used in four of the other OLAP clients. While many more clients are available compared to the previous survey (Thomsen & Pedersen, 2005), it is still the case that the OLAP client category leaves some room for improvements. The available documentation is limited and in some cases also the available support possibilities.

<div align="center">

INTEGRATED BUSINESS INTELLIGENCE SUITES

</div>

The focus of this survey is not on pre-packed, integrated BI suites, but instead on the individual tools that can be used at the different layers in a full BI solution. This provides flexibility for a completely customized solution. Integrated open source BI packages do, however, also exist. In this section integrated packages are briefly described.

JasperSoft Business Intelligence Suite

JasperSoft Business Intelligence Suite (JasperSoft, 2008) from JasperSoft ships with MySQL and the web server Tomcat (The Apache Software Foundation, 2008) such that it can be used out-of-the-box. Further, it includes JasperServer for ad-hoc queries, reports, charts, crosstabs and dashboards. It is also possible to schedule, share, and interact with reports using JasperServer. JasperSoft Business Intelligence Suite also includes JasperAnalysis for OLAP and JasperETL. These are based on Mondrian/JPivot and Talend, respectively. Finally, the JasperReports reporting tool is included in the suite.

Pentaho Open BI Suite

Pentaho Open BI Suite (Pentaho, 2008b) from Pentaho does not have a DBMS in the package. Pre-configured setups that use Firebird or MySQL do exist for easy testing, though. The suite includes Pentaho Data Integration, also known as Kettle. It also includes Pentaho Analysis with Mondrian and JPivot as well as Pentaho Dashboards. A reporting tool (based on JFreeReports which are now developed as part of the Pentaho suite) and the data mining tool Weka (Pentaho, 2008d) are also included.

SpagoBI

SpagoBI (Engineering Ingegneria Informatica, 2008) from OW2 Consortium is an integration platform. As it is an integration platform, and not a product platform, it is not made to use a certain tool set. Instead different "engines" (closed or open source) can be used, even at the same time. Thus, SpagoBI has drivers that integrate other tools into the platform such that, for example, Talend can be used as ETL tool and Mondrian as OLAP server in a SpagoBI project. SpagoBI's behavioural model regulate visibility of data and documents. Also, administration tools for scheduling, configuration, etc. are included. A tool for designing and maintaining analytical documents and module for metadata management are also included in SpagoBI.

<div align="center">

CONCLUSION

</div>

Compared to the findings in the last survey of open source BI tools from 2005 (Thomsen & Pedersen, 2005), there has been a strong development. Many more tools are available and their maturity has improved. In the current survey, ten ETL tools, six DBMSs, two OLAP servers, and seven OLAP clients were considered. As in the previous survey, the DBMS category is the most mature. The DBMSs have many advanced features, and good commercial support and documentation are available. The ETL tool category has improved a lot compared to the previous survey. ETL tools with advanced features and GUIs now exist. The OLAP servers are still dominated by the ROLAP server Mondrian which has been developed further and now offers better and faster functionality. For the OLAP clients, there are also many tools available, but this category seems to lack a little behind and leave room for new "killer applications".

While this survey considers many pre-defined general criteria for the different categories, there are other factors to consider for the specific cases. Issues like stability and performance are also important for BI projects. As the tools were not configured and run in this survey, it was not possible to investigate these issues.

Future work includes building two full BI solutions for the same purpose and data sets. One of them will be built using open source tools while the other will be built with commercially licensed tools. For such two solutions, it is interesting to investigate the possible differences in development time, ease-of-use, features and problems.

ACKNOWLEDGMENTS

REFERENCES

Abadi, D. J., Madden, S. R, & Hachem, N. (2008). Column-stores vs. row-stores: How different are they really? In Wang, J. T.-L. (Ed.), *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2008* (pp. 967-980). New York: ACM.

BPM Conseil (2008). *http://freeanalysis.org.* Retrieved August 7, 2008 from http://freeanalysis.org/.

Clover.ETL – open source data integration tool (2008). Retrieved August 7, 2008 from http://cloveretl.org/.

CWI (2008). *Query Processing at Light Speed.* Retrieved August 7, 2008 from http://monetdb.cwi.nl/.

Engineering Ingegneria Informatica (2008). *Spago Solutions – SpagoBI*. Retrieved August 7, 2008 from http://spagobi.org/.

EnterpriseDB (2008). *Postgres Plus – Open Source Database | EnterpriseDB.* Retrieved August 7, 2008 from http://www.enterprisedb.com/.

Firebird Project (2008). *Firebird – The RDBMS that's going where you're going*. Retrieved August 7, 2008 from http://firebirdsql.org/.

Gaffiero, M. (2007). *SourceForge.net: Pequel ETL Data Transformation Engine.* Retrieved August 7, 2008 from http://sourceforge.net/projects/pequel/.

Gimenez, J. & Lopez, J. (2007). *SourceForge.net: PaloKettlePlugin*. Retrieved August 7, 2008 from http://sourceforge.net/palokettleplug/.

Greenplum (2008). *Greenplum database redefines data warehousing, the bi database, and the data warehouse appliance*. Retrieved August 7, 2008 from http://greenplum.com/.

Grupo Calypso (2006). *JMagallanes*. Retrieved August 7, 2008 from http://jmagallanes.sourceforge.net/en/.

Ingres (2008). *Enterprise Open Source Database – Ingres.* Retrieved August 7, 2008 from http://ingres.com/.

Introduction to JRubik (2005). Retrieved August 7, 2008 from http://rubik.sourceforge.net/jrubik/intro.html.

Ishii, T., et al. (2007). *pgpool-II README*. Retrieved August 7, 2008 from http://pgpool.projects.postgresql.org/.

JasperSoft (2008). *JasperSoft – Open Source Business Intelligence*. Retrieved August 7, 2008 from http://jaspersoft.com/.

Jedox (2008a). *Introduction to ETL | Jedox – Enterprise Spreadsheets für Excel.* Retrieved August 7, 2008 from http://www.jedox.com/en/enterprise-spreadsheet-server/etl-server/introduction.html.

Jedox (2008b). *Palo Open-Source OLAP for Excel – Multidimensional Database for Budgeting, Forecasting, Planning, Reporting, MOLAP, Analysis Software - Business Performance Management (Excel-friendly OLAP) | Jedox - Enterprise Spreadsheets für Excel.* Retrieved August 7, 2008 from http://www.jedox.com/en/enterprise-spreadsheet-server/excel-olap-server/palo-server.html.

JBIWiki: ETLSE (2008). Retrieved August 7, 2008 from http://wiki.open-esb.java.net/Wiki.jsp?page=ETLSE.

Kimball, R. & Ross, M. (2002). *The Data Warehouse Toolkit* (2nd ed.) . New York: John Wiley & Sons.

Kinetic Networks (2008). *KETL.org – Designed to support the community that uses the KETL™ open source ETL product*. Retrieved August 7, 2008 from http://ketl.org/.

Kupolov, F. (2008). *Welcome to Scriptella ETL Project*. Retrieved August 7, 2008 from http://scriptella.javaforge.com/.

LucidDB Home Page (2008). Retrieved August 7, 2008 from http://luciddb.org/.

Microsoft, & Hyperion Solutions (2002). *XML for Analysis Specification*. Retrieved from http://www.xmlforanalysis.com/xmla1.1.doc.

MySQL (2008). *MySQL :: The world's most popular open source database*. Retrieved August 7, 2008 from http://mysql.com/.

Netezza (2008). *Data warehouse appliances, data warehousing, Netezza, Netezza.com*. Retrieved August 7, 2008 from http://netezza.com/.

Object Refinery Limited (2008). *JFreeChart*. Retrieved August 7, 2008 from http://www.jfree.org/jfreechart/.

Open Source Initiative (2006). *Open Source Licenses*. Retrieved from http://opensource.org/licenses.

OpenI.Org (2008). *openi.org - Open Source Web Application for OLAP Reporting*. Retrieved August 7, 2008 from http://openi.sourceforge.net/.

Pentaho (2008a). *Pentaho Analysis Services: Mondrian Project*. Retrieved August 7, 2008 from http://mondrian.pentaho.org/.

Pentaho (2008b). *Pentaho Commercial Open Source Business Intelligence: Home*. Retrieved August 7, 2008 from http://pentaho.org/.

Pentaho (2008c). *Pentaho Commercial Open Source Business Intelligence: Kettle Project*. Retrieved August 7, 2008 from http://kettle.pentaho.org/.

Pentaho (2008d). *Pentaho Commercial Open Source Business Intelligence: Weka Project*. Retrieved August 7, 2008 from http://weka.pentaho.org/.

PostgreSQL Global Development Group (2008). *PostgreSQL: The world's most advanced open source database.* Retrieved from http://postgresql.org/.

Slony Development Group (2008). *Slony-I*. Retrieved August 7, 2008 from http://www.slony.info/.

SourceForge.net: REX – waRehouse Explorer (2007). Retrieved August 7, 2008 from
http://sourceforge.net/projects/whex/.

Spofford, G., Harinath, S., Webb, C., Huang, D. H., & Civardi, F. (2006). *MDX Solutions* (2nd ed.). New
York: John Wiley & Sons.

Sun (2008). *GlassFish ESB*. Retrieved August 7, 2008 from http://glassfishesb.org/.

Talend (2008). *Talend – first provider of open source data integration software*. Retrieved August 7,
2008 from http://talend.com/.

Tensegrity Software (2008). *JPalo – Palo Java World – Business Intelligence Reporting*. Retrieved
August 7, 2008 from http://www.jpalo.com/en/products/start_products.html.

The Apache Software Foundation (2008). *Apache Tomcat – Apache Tomcat*.Retrieved August 7, 2008
from http://tomcat.apache.org/.

Thomsen, C. & Pedersen, T. B. (2005). A Survey of Open Source Tools for Business Intelligence. In
Tjoa, A. M. & Trujillo, J. (Eds.), *Proceedings of the 7th International Conference on Data
Warehousing and Knowledge Discovery* (pp. 74-84). Berlin Heidelberg: Springer.

Together Teamlösungen (2007). *JDBC Data Transformations*. Retrieved August 7, 2008 from
http://www.enhydra.org/tech/octopus/.

Tonbeller (2008). *JPivot – Home*. Retrieved August 7, 2008 from http://jpivot.sourceforge.net/.

Yuhanna, N. (2006). *The Forrester Wave™: Open Source Databases, Q2, 2006.* Cambridge, MA, USA:
Forrester.

APPENDIX: TABULAR SUMMARY

Criteria / ETL tools	Apatar	Clover.ETL	ETL Inte-grator	KETL	Kettle	Octopus	Palo ETL Server	Pequel	Scriptella	Talend
License	GPL or com.	LGPLG or com.	CDDL	GPL	LGPL	LGPL	GPL	GPL	Apache	GPL
Platform	Java	Java	Java	Java	Java	Java	Java	Perl	Java	Java
Support	Yes	Yes	Not found	Yes	Yes	Yes	Yes	?	Yes	Yes
Documentation and forums	~40 pages PDF	181 pages for GUI, wiki and slides	Design docs, videos, tutorials.	Deprecated	274 + 40 pages	122 pages	56 pages	72 pages User's guide (deprecated)	23 pages	161 + 550 pages
Category	ROLAP	ROLAP	ROLAP	ROLAP	ROLAP	ROLAP	MOLAP	Files and ROLAP	ROLAP	ROLAP
Sources	DBMSs, files, ERP, CRM	DBMSs, files	DBMSs, files, ERP, CRM	DBMSs, files	DBMSs, files	JDBC	JDBC, CSV, Palo, LDAP	Files and DBMSs	DBMSs, files	DBMSs, files, CRM, Palo
Targets	Same	Same	DBMSs	Same	DBMS, files	JDBC	Palo	Same	Same	Same
Inc. load	No	No	Yes	?	Yes via timestamp	Can update	Not directly	No	No	Can update
Job spec.	GUI	XML (GUI avail.)	GUI	XML	GUI	XML	XML	XML	XML	GUI
Own transformations	Java	Java and TL	Java	?	JavaScript	Java, JavaScript		Perl	Scripting lang.s	Java, Perl
Parallel jobs	No	Yes	Yes	Yes	From next version	No	No	Yes	No	Yes

DBMSs / Criteria	Firebird	Ingres	LucidDB	MonetDB	MySQL	PostgreSQL
License	MPL-like	GPL or commercial	GPL + LGPL	MPL-like	GPL or commercial	BSD
Platform	Windows, Linux, MacOS X, FreeBSD, Solaris, HP-UX	UNIX-like, Windows	Linux, Windows (Cygwin)	UNIX-like, Windows	UNIX-like, Windows	UNIX-like, Windows
Support	Yes	Yes	No	No	Yes	Yes
Documentation	Deprecated	~8000 pages	Wikis and tutorials	260 + 113 pages	2071 pages	1908 pages
Mat. views	No	No	No (but planned)	No	No	No
Bitmap indexes	In-memory when combining indexes	No	Yes	No	In-memory when combining indexes	In-memory when combining indexes, on-disk planned
Star joins	No	No	Yes	No	No	No
Partitioning	No	Yes	No	No	From next version	Partly supported
Replication	3rd party tools	Yes	No	No	Yes	3rd party tools
Stored procedures and user-defined functions	PSQL, C, Delphi	SQL, C	Java	SQL, MAL, C	SQL, C	PL/pgSQL,PL/Tcl, PL/Python (and extensible to other languages), C

OLAP servers Criteria	Mondrian	Palo
License	CPL	GPL or commercial
Platform	Java	Primarily Linux and Windows
Support	Yes	Yes
Documentation	HTML (~200 pages of text)	~376 pages
Category	ROLAP (uses underlying DBMS)	MOLAP (memory-based)
DBMS	Any JDBC-compliant	n/a
Specification of cube	XML	Via Excel add-in or via API
Aggregate tables	Yes	No
API + query language	Olap4j, MDX	Proprietary

OLAP clients / Criteria	FreeAnalysis	JPalo Client and Web Client	JMagallanes Olap & Reports	JPivot	JRubik	OpenI	REX
License	MPL	GPL or commercial	BSD	CPL	CPL	MPL-like	LGPL
Platform	Web	Java and Web (Java), respectively	Java	Web (Java)	Java	Java	Java
Support	Yes	Yes	Yes	Yes	No	No	No
Documentation	?	Not found	Videos and short HTML documents	Short HTML documents	Short HTML documents	Short HTML documents	Tutorial as HTML document
OLAP servers / API	XMLA	Palo and XMLA	No OLAP server support. Connects to DBMSs, XML, and Excel	XMLA, olap4j	Mondrian and experimental olap4j support	XMLA	XMLA
Reports	Tables and charts	Tables and charts (only tables in JPalo Web Client)	Tables and charts.	Tables and charts	Tables, charts, and maps	Tables and charts	Tables and charts