November 6, 2007

Market Overview: Open Source ETL Tools

An Attractive Alternative To Custom Code

by Rob Karel and Michael Goulde
with Mike Gilpin and Norman Nicolson

EXECUTIVE SUMMARY

As open source, community-driven software continues its evolution from an interesting experiment to a viable technology alternative for enterprises, additional market segments develop open-source alternatives. The latest segment in the trend is open source extract, transform, and load (ETL) data integration. While the vendors and projects leading the way in this emerging market have their sights set on creating an alternative to commercially available packaged solutions, expect early adopters to be independent software vendors (ISVs) and systems integrators (SIs), as well as end users in both midmarket and enterprise-sized organizations that are looking for a more productive and efficient alternative to writing custom code.

BASIC ETL REQUIREMENTS ARE WELL SUITED TO THE OPEN SOURCE MODEL

The basics of extraction, transformation, and load are straightforward: Get some data, apply business transformation logic to the data, and deliver it to a target platform. The most familiar packaged ETL software on the market includes enterprise-class options from Ab Initio, Business Objects, IBM, Informatica, and SAS. These ETL tools are well suited for solving complex ETL requirements such as running on a wide variety of heterogeneous platforms, connecting with a multitude of sources and targets, and handling very high performance and scalability requirements. These tools are expensive with minimum list prices ranging anywhere between $45,000 to much more than $100,000.[1]

Less expensive packaged ETL options exist, but they often have limited support for heterogeneous environments. For example, Microsoft's SQL Server Integration Services (SSIS), an express version of which is bundled free with SQL Server, only runs on Windows platforms and only provides native database connectivity to Microsoft SQL Server. Similarly, Oracle's core Oracle Warehouse Builder (OWB) functionality is bundled at no extra charge with the purchase of an Oracle database. While OWB offers greater platform support than Microsoft, it focuses primarily on Oracle relational database management system (RDBMS) target environments. In addition, Oracle charges for customer relationship management (CRM) and enterprise resource planning (ERP) application connectivity (e.g., Oracle E-Business Suite, PeopleSoft, and SAP) and data quality services among other advanced options, bringing its costs more on par with those of some of its high-priced competitors.

These market dynamics make it a big challenge for integration architects to identify an ETL tool that can integrate data between the wide variety of data sources and targets within their IT environment at a reasonable cost. It's expensive to buy this connectivity, but building custom connectors for each source/target combination also requires significant resources and is not as conducive to connector reuse. In contrast, consider the open source model:

- **Open source implementations bring community power to ETL.** When a developer creates a particular connector, he contributes it to a community thereby benefiting others with a similar need. The community approach to building a collection of connectors is especially applicable to ETL because of the wide variety of environments that need to be connected, by spreading the cost of building so many connectors over the community.

- **Open source implementations promote development of standards.** Standards play a crucial role in integration data formats, metadata, and common transformations. External standards include industry file formats like XML, HIPAA, and SWIFT while reuse of data format, validation, and connectivity transformation objects is a major driver for internal efficiencies. The transparency of open source code and the large number of testers available in the community promote sound implementations of these standards. And when standards are still evolving, developing open source code in parallel helps ensure that those writing the standard specification are creating something that can actually be implemented in code.

- **Open source costs less.** The open source ETL products on the market today are all significantly less expensive than commercially licensed tools. In addition, application developers always have the option of downloading the source code, compiling it, and supporting it themselves if they think that will drive cost down further. Forrester does not recommend this approach for all companies, but it is a viable option for companies that have the resources, necessary skills, and a general policy favoring self-support — one that is not available with commercial products.

In the long run, developers who participate actively in ETL open source communities have the most to gain since their contributions will be tested, fixed, and enhanced by others in the community. The additional resources that a community can bring to bear on a product can be a valuable benefit. But with open source ETL offerings still very much in their infancy, the number of community-based contributions has been minimal to date.

END USERS WITH BASIC REQUIREMENTS, ISVs, AND SIs ATTRACTED TO OPEN SOURCE ETL

Many developers will find that one of the open source ETL tools will be more than adequate for their current needs. The four primary constituencies adopting open source ETL today are:

- **ISVs looking for embeddable data integration.** Open source enables ISVs to reduce the cost of product development. ISVs often combine freely available open source components, such as databases or logging functions, into complete platforms. These savings are passed on to customers in the form of a shift from product licensing revenue to lower margin support and maintenance subscriptions. Many vendors with business intelligence, data management, and enterprise application products incorporate data integration, migration, or transformation capabilities as an embedded component. Open source ETL provides an attractive alternative,

which also reduces the memory footprint of the resulting product compared to embedding a larger commercial offering.

- **System integrators looking for inexpensive integration tooling.** Many system integrators implement complex custom data management solutions for their customers. These solutions often require the development of point-to-point integration interfaces, and leveraging high-cost packaged ETL software is not always within a customer's budget. Using open source ETL software enables SIs to deliver these integration capabilities faster and with a higher level of quality than if they had to custom-build the capabilities.

- **Enterprise departmental developers looking for a local solution.** Integration architects, developers, or project managers within large enterprises often consider using open source ETL technology to support small departmental initiatives. When a new data movement requirement involves at most one or two sources and a single target, these integration specialists recognize that custom code can be inefficient and sloppy. But attempting to acquire software licenses for a packaged ETL tool through a long procurement process can be even worse. Leveraging open source ETL can provide a more expedient solution yet retain good quality and performance.

- **Midmarket companies with less complex requirements and smaller budgets.** Smaller organizations are more likely to support a homogeneous IT environment with minimal data migration, integration, and transformation requirements, hence less need for high-cost data integration software. The integration projects that do arise typically revolve around early business intelligence (BI) initiatives, which is why JasperSoft and Pentaho — two of the leading open source business intelligence providers — now also offer open source ETL capabilities to complement their BI offerings.[2]

SEVERAL OPEN SOURCE ETL PROJECTS COMPETE FOR MIND SHARE

There are dozens of open source projects that perform one or more ETL functions. A smaller number of projects attempt to provide a more complete set of capabilities. Forrester spoke with representatives from clover.ETL, Kinetic Networks' KETL project, Pentaho's Kettle project, and Talend, which are among the leading contenders for open source ETL.[3] The technical characteristics of these projects are more similar than different. Their market strategies represent the bulk of their differentiation.

clover.ETL Dual License Well Suited For Embedded Use

The clover.ETL project is directed by OpenSys, a company based in the Czech Republic. clover.ETL is a Java-based framework that is dual licensed: LGPL or commercial. The commercially licensed version is the same code as the open source, but includes support and warranty.

clover.ETL offers a small footprint, making it easily embeddable by ISVs and systems integrators. The open source part of the clover.ETL project aims at creating a core library of functions including mappings and transformations. Its enterprise server edition and clover.GUI developer user interface are commercial offerings (although clover.GUI is offered for free for non-commercial usage).

KETL Developed To Meet A Service Provider Need

KETL is a project sponsored by Kinetic Networks, a professional services company. It began life as a tool to be used in customer engagements because commercial tools were too expensive. The core library is licensed under the LGPL license while the KETL server is licensed under a GPL license. Code is currently being developed by Kinetic employees, although outside contributions are expected in the future. Kinetic also developed additional modules, such as a data quality and profiling component, which are not placed under an open source license.

KETL was initially developed as a utility to replace custom PLSQL code used to move large volumes of data. It is a Java-based, XML-driven development environment with configuration capabilities that are most useful for skilled Java developers. A current limitation of KETL that may be relevant to some users is that they do not have a visual development GUI, everything is defined in XML, leveraging the Eclipse IDE's XML authoring capabilities.

Pentaho Offers More Than ETL

Pentaho positions itself as a business intelligence provider that offers ETL tools as a data integration capability. Its ETL capabilities are based on the Kettle project, which is currently licensed under the LGPL license. Pentaho sells subscriptions that include support services, management tools, and indemnification.

Pentaho's Kettle project has been able to incorporate a number of contributions from its community primarily focusing on connectivity, bulk loaders, and transformations. Examples of community-driven enhancements include the development of an Oracle bulk loader, a Web services lookup, and an SAP connector. Although integrated with Kettle, the SAP connector is not currently free: It is a commercially available plug-in offered by Proratio, the SI partner that created it.

Talend Focuses On Open Source ETL And Software OEMs

Talend is a French startup that positions itself as an open source data integration pure play, with its product Open Studio. It is available under the GPL V2 license, but Talend also has an OEM license agreement for vendors that wish to embed Open Studio's capabilities in their products. For example, JasperSoft has embedded Open Studio into its popular open source business intelligence software, creating an open source BI stack to compete with Pentaho's.[4] Talend is a commercial open source vendor that generates revenue from training, support, and consulting services.

Talend's Open Studio offers a user-friendly graphical modeling and development environment and offers both the traditional ETL as well as an extract, load, and transform (E-LT) approach for performance management. E-LT, also often referred to as pushdown optimization, is an architectural approach that allows users to bypass the cost of dedicated hardware to support an ETL engine or transformation server and instead allows users to leverage spare server capacity within the source or target environment to power the transformations.

WORK WITHIN THE LIMITS OF THESE TOOLS

Today's open source ETL tools are quite suitable when they are used within their limits. Future enhancements will undoubtedly extend these limits in directions that will be guided by feedback from their communities. Current limitations include:

- **Enterprise application connectivity.** A common enterprise ETL requirement is to extract data from apps like Oracle, PeopleSoft, and SAP. None of the open source ETL products include enterprise application connectivity free of charge. These connectors — if available — are usually a paid option. They are not compatible with open source principles but a common revenue source.

- **Non-RDBMS connectivity.** Not all of your most critical information lives in relational databases that can be easily accessed through the open database connectivity (ODBC) or Java database connectivity (JDBC) connectors made available through most open source ETL solutions. Connectivity to mainframe and legacy apps, message queues (TIBCO, JMS, WebSphere MQ), and support for industry file format standards (HIPAA, SWIFT, ACCORD) are offered by more complex commercial data integration environments.

- **Large data volumes and small batch windows.** If you manage the flow of hundreds of gigabytes or more of data with service-level agreements (SLAs) that require the data to be processed within a very short time frame, you will need to consider a proven enterprise-class ETL platform. Not surprisingly, most of the open source ETL vendors indicate that they can support extremely high data volumes — and they likely can in certain scenarios, but there has not been enough adoption, testing, and experience in using these tools to verify this more broadly. If you are considering an open source product for an initiative requiring high performance and scalability, be extremely thorough in your testing for high availability and failover support.

- **Multirole collaboration.** A significant benefit of enterprise-class ETL software is the ability for a large number of architects, modelers, developers, and data stewards to collaborate, share metadata, and reuse mapping and transformation objects within and across complex projects. If your data integration staff consists of only one or two developers, with no plans to increase in the near future, this collaboration functionality may be unnecessary. But if you are managing a larger, more distributed team playing a variety of roles, multirole collaboration support can be extremely valuable.

- **Complex transformation requirements.** Most open source ETL tools require scripting and custom code to define and configure transformation rules. The more expensive packaged ETL products offer much more user-friendly rules wizards and robust transformation libraries, which represents a significant value proposition to those who need to manage a large volume of complex rules.

RECOMMENDATIONS

CONSIDER OPEN SOURCE ETL TO REPLACE CUSTOM CODE

Open source ETL does not yet provide the robust suite of heterogeneous data management capabilities needed to be considered for a cross-enterprise standard for data integration. The capabilities missing today include advanced connectivity, real-time data integration techniques like enterprise information integration (EII) and change data capture (CDC), enterprise-class collaboration, and integrated data quality management and profiling. With that said, many enterprises large and small, as well as ISVs and SIs, are not looking for a large, expensive data integration suite. If your goal is to find a cheap, efficient, and reliable alternative to custom code for your data integration needs, you should seriously consider open source ETL technologies.

WHAT IT MEANS

OPEN SOURCE ETL IS NOT YET COMMUNITY-DRIVEN, BUT STAY TUNED

The most popular open source ETL vendors mentioned here are not truly community-driven open source projects yet: A majority of new functionality is developed and released from the companies that sponsor or manage the software. Expect increased investment by these vendors to build out and encourage development from the wider community — especially for the development of connectivity modules to the near-infinite number of legacy and evolving source systems.

SUPPLEMENTAL RESOURCES

Companies Interviewed For This Document

Kinetic Networks

OpenSys

Pentaho

Talend

ENDNOTES

[1] Forrester evaluated leading enterprise extract, transform, and load vendors across 68 criteria and found that IBM and Informatica maintain leadership positions in enterprise ETL thanks to their ability to scale and perform batch and operational data integration (DI) in complex environments, as well as to maintain a consistent focus on providing robust data management capabilities. Business Objects and Oracle have also emerged as Leaders with significant usability and scalability improvements, but they are still primarily used in data warehousing environments and have not been widely adopted for operational DI needs. SAS is a Strong Performer, but it remains most attractive as an integrated piece of a SAS BI platform. Ab Initio offers a highly scalable and configurable data-processing platform, but its secretive corporate culture limits prospective customers' visibility into its strategy. iWay Software, Microsoft, Pervasive, and Sunopsis (acquired by Oracle) round out the Strong Performers best suited for a more targeted subset of DI professionals. As a new player in enterprise ETL, Sybase has some of the raw materials needed to develop a competitive solution, but it must integrate the tools it has acquired into a DI suite with a clear differentiation strategy to gain traction in this crowded market. See the May 2, 2007, "The Forrester Wave™: Enterprise ETL, Q2 2007" report.

[2] Pentaho supports the Kettle open source ETL project, and JasperSoft OEMs Talend's open source ETL software.

[3] For more details on each of the open source ETL projects interviewed, see www.cloveretl.org, www.ketl.org, http://kettle.pentaho.org, and www.talend.com. Other open source ETL projects of interest include Apatar (www.apatar.com), Octopus (www.enhydra.org/tech/octopus/index.html), and Scriptella (http://scriptella.javaforge.com). Smaller projects also include Babeldoc, DataSift, ETL Integrator, JETSTREAM, Joost, mec-eagle, netflux, OpenDigger, Xephyrus Flume, and Xineo.

[4] In a recent report, Forrester introduced the BI Stack — comprising all the architectural components required to build a comprehensive business intelligence strategy. See the October 23, 2007, "It's Time To Reinvent Your BI Strategy" report.