Introduction

Talend is an open source provider of data integration products. However, while many open source suppliers are built on little more than a man and his dog, Talend has a business model and organisation that more closely resembles that of a conventional software vendor, while still leveraging the open source model. Thus, for example, it is the only open source vendor in this market to have offices around the world (in China and the United States as well as France and Germany) so that it can offer 24 x 7 support to its corporate customers. Similarly, the company has an extensive partner network, both with other open source and proprietary vendors, as well as with a number of system integrators.

The company’s flagship product is Talend Open Studio. This is a code (Java and Perl plus SQL) generating tool that supports both ETL (extract, transform and load) and ELT (extract, load and transform) approaches to data integration. A notable feature of the product is Business Modeler which, as its name implies, makes it easier for business analysts to become involved in the development process and to enable collaboration. It is worth noting this in a broader context because Talend introduced this capability before either IBM or Informatica introduced anything similar. This is a good indicator of Talend’s intention to be treated seriously as an enterprise-class vendor, as are the product’s support for multi-developer environments (in the Talend Integration Suite, which bundles in value added services as well), metadata management, high-performance deployment options and the recently introduced support for change data capture to enable real-time operating environments.

The company also offers its capabilities through Talend on Demand, which is the company’s software as a service (SaaS) offering.

By April 2008 there had been over 250,000 “core product” downloads of Talend Open Studio and, while this is not necessarily indicative of usage (and still less of customers paying for support or other services) it is still a substantial number. Moreover, Talend has some significant accounts that it can name, such as Virgin Mobile, Alliance Healthcare, SNCF French Railways, US Cellular, Swiss Life and Shopping.com amongst others.

In any market, open source represents a potentially disruptive approach. However, relatively few vendors are effectively organised to take advantage of such disruption. Talend is an exception.
Introduction

At an enterprise level, the market for data integration is now less focused on traditional ETL (extract, transform and load) and more centred on data integration as a broader concept. A number of vendors have been preaching that message for some time but it is only recently that this idea had achieved broad acceptance within the marketplace.

This is not just a question of the major suppliers getting their marketing messages across but it can also be seen when you look at the spectrum of offerings available from different vendors. Today, the position is that there are a number of providers of data integration platforms on the one hand while, on the other, we are seeing the emergence of specialist products that focus on just a single aspect of data integration such as data migration, content ETL, web ETL, replication and synchronisation, data federation, and data services amongst others. In particular, there is a significant emerging sub-market for data migration and a Market Update on this subject is being published alongside this one. We expect that data services will emerge as a further sub-market in due course but this remains at an early stage.

We are also aware of new vendors planning to enter the market with innovative, potentially disruptive, approaches to data integration.

In this Update we are focusing on data integration platforms suitable for deployment on an enterprise-wide basis. We are not, therefore, considering any of these specialised solutions, nor broader integration solutions that include ETL capability (unless ETL is a specific focus), nor solutions focused on a specific target database or use case.

Key market issues

To genuinely offer an enterprise-ready data integration platform implies a long list of requirements. To be comprehensive the product (suite) needs to offer not merely ETL but data federation, replication and synchronisation as well as data services that can plug data retrieved by the platform into an SOA environment. Transformation services and similar facilities should also be able to be packaged as a web service. Further, the product needs to be able to support both batch and real-time operations and have scheduling, reporting, debugging, auditing, data lineage and similar facilities built in. Ideally, there should never be a requirement to leave the environment to code complex transformations as you then lose data lineage, impact analysis and where-used capabilities. There is a general trend towards introducing more business friendly capabilities into data integration offerings.

Products should also be able to handle a wide range of data sources and targets and will need to be able to handle not just structured data but also XML, semi-structured sources such as EDI and SWIFT messages, and unstructured data in the form of Word documents, Excel spreadsheets, PowerPoint presentations, and so on. It will also be helpful if the software has an understanding of the metadata that different content management vendors wrap around the content that they store so that the platform can be used to migrate from one content environment to another and also to support federated queries across those environments. Data connectors that have specialised knowledge of source environments, especially leading application products, will be an advantage.
We do not regard data quality provision (both data cleansing and profiling) as a necessary requirement for a data integration platform. However, if it is not provided by the vendor then a close integration with appropriate third party tools will be required; simply having a partnership is not enough.

We do not have any particular view about the benefits or otherwise of a "black-box" solution as opposed to code generating approaches; there are arguments on both sides, though it will clearly be an advantage, in some instances, to be able to generate code (such as Java) that can be embedded within third party applications. We are aware that certain government agencies do not like the idea of black box.

On the performance side, parallelism is de rigueur in larger environments and there is a strong case for supporting ELT as well as ETL (plus other combinations), though some data warehousing vendors (for example, Vertica) do not recommend the use of ELT in their environments.

Unfortunately, there is no product that does all of the things just outlined; and, let’s face it, we have only skimmed the surface in terms of what one would like to see in a data integration platform.

**Vendor landscape**

The vendor landscape for data integration and ETL has changed dramatically over the last few years. On the one hand there has been significant consolidation with Sunopsis, Business Objects, Solonde, DataMirror and Hummingbird all having been acquired (by Oracle, SAP, Sybase, IBM and Open Text respectively) while on the other we have seen a growth in specialised vendors, as previously noted. There has also been a significant increase in the number of open source suppliers.

Of the leading players, IBM and Informatica remain the dominant vendors. The former is making a broader play, integrating its data integration capabilities with other functionality such as Master Data Management, while the latter is continuing to focus on its core competency in data integration. Amongst other things, it has particularly focused on improving its performance over recent releases and it is the only one of the major vendors to have introduced specialised general-purpose capabilities to support data migration. While both companies have an increased emphasis on providing capabilities at the business analyst level, IBM is currently ahead of Informatica in this respect.

Also of interest are the moves by other suppliers, notably Business Objects (now SAP) and SAS, to join these leading two. Both of these companies, of course, are saddled with the subjective impression that they are merely in the BI/ETL business (as is/was the case with Cognos) but both are keen to break out of this mould and have the products to do so. Business Objects, we believe, is better placed to do this, at least in the short term, in part thanks to its recent acquisition by SAP. Business Objects also has advanced capabilities aimed at business analysts.

Ab Initio is also a significant vendor at the high-end of the market but, as the company has a policy of not talking to analysts, we can only guesstimate that company’s position based on conversations with users and partners.

Two other large vendors making a major move into this market are Oracle and Sybase. The former has offered Oracle Warehouse Builder for some time. However, this is very much focused on Oracle as a target, which is why we have not included it in this report (and the same applies to Microsoft SQL Server Integration Services). Nevertheless, if that is your only target then Oracle Warehouse Builder must be the default option. However, Oracle owns so many applications, which are hosted on all sorts of platforms that it needs a more general-purpose data integration platform, which it has introduced in the form of Oracle Data Integrator. This is based on its acquisition of Sunopsis plus tight integration with Trillium’s data quality products. This combination is particularly powerful and we see Oracle Data Integrator as a serious competitor to the likes of Informatica and IBM and as a potential market leader along with these two.

Sybase has come from a slightly different position. It is the leading provider of heterogeneous replication and synchronisation and to that it added the acquisition of Avaki to provide federated query capability and then it bought the Solonde ETL tool (which is now also bundled with the company’s Sybase IQ product), all of which have been brought together in the company’s data integration suite, which also includes the PowerDesigner data modelling tool. However, it does not offer data quality capability and it does not have close partnerships with any data quality vendors. Nor have the various elements in this suite been as tightly integrated as one might wish. On the other hand, Sybase clearly shows potential and it has strengths where others do not.
Lower down the market two products that had seemed to be defunct now appear to have been resurrected by their respective owners. These are Genio (previously Hummingbird, now Open Text) and the other is Pitney Bowes Software’s Data Flow (previously Sagent and Group 1). Both of these have recently had significant new releases, which suggest that they could become serious players once again. Genio, which has historically had a strong user base in France, has been extended to support content ETL; while Data Flow, which always had a particularly strong business focus, has also been resurrected, and has been tightly integrated with Pitney Bowes’ CDQ (customer data quality) software. Pitney Bowes has also been extending its reach outside the US, which has been its traditional heartland. However, it remains to be seen how much headway these companies can make.

More generally, the mid-market is dominated by Microsoft (if you happen to have SQL Server as your only target) and Pervasive for more general, heterogeneous environments. Actually, in terms of functionality, we wouldn’t put Pervasive that far away from the leading suppliers, and what you might lose on performance you will certainly gain on ROI. Also active at this level of the market are ETL Solutions and ETI. Both are code generating solutions but both are limited by the geographic coverage that they can offer. iWay is also active in this market though it is more interested in selling connectors and adapters than data integration per se. There are also a number of products from companies such as Baycastle and Sesame Solutions that are more likely to be suitable for project-based solutions than as platforms. Embarcadero’s DT/Studio product is still available as a solution but it is no longer a focus for the company.

There are a significant number of open source solutions. The most mature of these and the ones most likely to be suitable for deployment as enterprise platforms are from Talend and Clover.ETL. The former has always been aimed at the higher end of the market while Clover has introduced an enterprise edition called Clover.Server. The Kettle product embedded within Pentaho is similarly mature but tends to be focused on the BI market alone rather than as a general purpose product, while most other open source products are most likely to be selected on a project by project basis. These include IKAN, Enhydra Octopus, ActiveWarehouse ETL (based on Ruby), Apatar and Kinetic Networks’ (which is a data quality supplier) KETL, and Snaplogic (which is focused on mashups).

Finally, there are a myriad of companies that offer solutions that encompass ETL, but where that is not their specific focus. These include broader integration companies such as TIBCO, Sun (SeeBeyond), Software AG (WebMethods), and Jitterbit; as well as specialised providers like GoldenGate (replication, synchronisation and data migration), Celona, Rever, Datalynx, Legeon (all data migration), XAware (data services), QL2 (web ETL), EntropySoft (content ETL), Composite Software (federation) and Denodo and Kapow (both of which support integration to enable mashups).

Summary and conclusions

While there are more than 40 companies from which one can acquire ETL or other data integration capabilities that might be appropriate for a particular project, there are less than half that number that might be suitable for deployment as a data integration platform that will support reuse for multiple projects across the enterprise. Of course, this depends on the size of the organisation. For the very largest companies there are half this number again that might be suitable for such deployments.

However, as noted previously, there is no product on the market that currently supports all of the data integration functionality that one might like. This means compromise and it may mean adopting a platform for most purposes but augmenting that, at least in the short term, with a specialist provider for particular projects. Thus the most important points in selecting a provider of a data integration platform will not only be functionality, price, performance, scalability, the stability of the vendor and all of those other things that one normally considers, but also the supplier’s roadmap and how and when it intends to fill in those holes (whatever they may be) in its offering.

Philip Howard
Data Integration Platforms
May 2008

For additional information relating to this subject visit http://www.bloor-research.com/update/949
Bloor Research overview

Bloor Research has spent the last decade developing what is recognised as Europe’s leading independent IT research organisation. With its core research activities underpinning a range of services, from research and consulting to events and publishing, Bloor Research is committed to turning knowledge into client value across all of its products and engagements. Our objectives are:

- Save clients’ time by providing comparison and analysis that is clear and succinct.
- Update clients’ expertise, enabling them to have a clear understanding of IT issues and facts and validate existing technology strategies.
- Bring an independent perspective, minimising the inherent risks of product selection and decision-making.
- Communicate our visionary perspective of the future of IT.

Founded in 1989, Bloor Research is one of the world’s leading IT research, analysis and consultancy organisations—distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services and consultancy projects.

About the author

Philip Howard
Research Director - Data

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up what is now P3ST (Wordsmiths) Ltd in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director. His practice area encompasses anything to do with data and content and he has five further analysts working with him in this area. While maintaining an overview of the whole space Philip himself specialises in databases, data management, data integration, data quality, data federation, master data management, data governance and data warehousing. He also has an interest in event stream/complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to www.IT-Director.com and www.IT-Analysis.com and was previously the editor of both “Application Development News” and “Operating System News” on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and published a number of reports published by companies such as CMI and The Financial Times.

Away from work, Philip’s primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master) and walking the dog.
This document is copyright © 2008 Bloor Research. No part of this publication may be reproduced by any method whatsoever without the prior consent of Bloor Research.

Due to the nature of this material, numerous hardware and software products have been mentioned by name. In the majority, if not all, of the cases, these product names are claimed as trademarks by the companies that manufacture the products. It is not Bloor Research’s intent to claim these names or trademarks as our own. Likewise, company logos, graphics or screen shots have been reproduced with the consent of the owner and are subject to that owner’s copyright.

Whilst every care has been taken in the preparation of this document to ensure that the information is correct, the publishers cannot accept responsibility for any errors or omissions.