

InDetail



Talend Open Studio v3

An InDetail paper by Bloor Research
Author : Philip Howard
Publish date : January 2009

Talend Open Studio is worth serious consideration in more or less all circumstances.

Philip Howard

Executive summary

Talend is an open source provider of data integration products or, in traditional parlance, ETL (extract, transform and load), although the product also supports an ELT approach. The company's flagship product is Talend Open Studio but the product is also available as the core component within Talend On Demand, which is the company's software as a service (SaaS) offering; and the Talend Integration Suite, which is a bundled product and services offering that includes additional software capabilities for multi-developer environments and to provide high performance deployment options, including scheduling.

Key findings

In the opinion of Bloor Research the following represent the key facts of which prospective users should be aware:

- Talend Open Studio is a code generating product. That is to say, it generates either Java or Perl (plus SQL) that you can deploy as required (in parallel and/or across a grid, if appropriate). This means that Talend avoids the 'black box' approach of many other vendors that can result in the black box becoming a bottleneck.
- A major advantage of the code generating approach is that it is easy to encapsulate a job as a web service that can be invoked by third party applications, or to embed the generated Java into third party applications. This is more of an issue for conventional approaches because they require an engine (the black box) in order to run.
- Using a conventional ETL approach you would normally deploy your applications as close to the source as possible but ELT options allow you to deploy on the target. In the latest release, ELT options, based on SQL patterns (a combination of Java and SQL) that are specific to each database, have been significantly enhanced so that this approach can be used with any supported database environment.
- The recommended approach for development within Talend Open Studio starts with the Business Modeler before drilling down to developer functions. We particularly like this emphasis on the business analysts as a starting point for development.
- In most respects Talend Open Studio looks pretty much like other products: a graphical user interface, dragging and dropping onto a palette, a visual SQL builder, debugging facilities and so on. It actually uses the familiar Eclipse interface.
- There are data cleansing and data profiling capabilities built into the product.
- The latest release of the product supports significant parallelisation capabilities.
- Metadata management, based on a repository that supports reuse, is provided within Talend Open Studio though full multi-developer support with version control and check in and out is only available within Talend Integration Suite.
- As far as we know, Talend is the only open source vendor in this market to have offices outside its home market: in this case in China and the United States as well as France and Germany. This means that Talend can offer 24x7 support on all continents.

The bottom line

We have long been of the view that the advantages inherent in a code generating approach have been overlooked when compared to black box approaches—largely thanks to the marketing machines of the major ETL vendors—so we are glad to see it making a comeback. In particular, the ability to generate Java will be of particular appeal to companies wishing to embed relevant routines within their own applications.

Talend therefore has two main claims to fame: first, it is open source—if you like open source then we believe that Talend Open Studio represents an excellent option. Secondly, it is code generating and, again, if you are convinced by the merits of code generation then Talend is likely to appeal to you. The real issue will be if neither of these holds true, in which case the question is: how well does Talend Open Studio stack up when compared to the conventional suppliers within this market? In our opinion: surprisingly well, which suggests that the product is worth serious consideration in more or less all circumstances

Product availability

Talend Open Studio is currently in version 3.0. Typically, the company issues a 'milestone' release every month (similar to beta versions in the proprietary world, but available to all developers) and a major release three times a year. The product is made available under a GPL v2 license and the environment provided is based on Eclipse.

Note that Talend Open Studio is a code (Java, Perl and SQL) generating product and there is therefore no requirement for a run-time server to host Talend Open Studio, just a JVM for Java or an interpreter for Perl. For integration purposes it supports FTP, HTTP and web services (both as provider and consumer). Message oriented middleware is supported for near real-time processing while change data capture capabilities are also available with a complete publish/subscribe mechanism that supports multiple targets for updates with varying timescales.

More than 400 connectors are available out of the box in the current version, including application-specific connectors for SAP, salesforce.com, SugarCRM, Microsoft Dynamics and others. In addition, you can download community developed connectors that are published on the Talend Community web site. There is a direct connection from Talend Open Studio to this site and connectors downloaded from here can be plugged directly into your user environment.

Product description

Introduction

Talend Open Studio is suitable for use in all sorts of data integration environments but it is particularly targeting operational requirements such as data migration, consolidations and real-time embedded capabilities in addition to the sort of facilities required to load data warehouses. This is not because Talend Open Studio is inferior in any way in the data warehousing space but that the company sees more opportunities within operational environments, especially given that it is a code generating product, which makes it much easier to embed the relevant code within a web service that can then be invoked by requesting applications.

On this topic we should comment on the fact that Talend Open Studio is code generating, as this technique has not been popular for some years. Briefly, Talend Open Studio provides a development environment that generates source code (Java or Perl plus SQL) that you can then compile and use on whatever systems are necessary (which may be deployed across a grid). This has a number of advantages. First, it means that you can run on any platform or combination of platforms, with a non-proprietary runtime environment.

Secondly, there are no limits to the degree of complexity that is supported. If you use a 'black box' style approach it is often the case that developers have to drop out of the environment to hand code part of the solution. However, since these products are not generally designed to support external coding then you immediately lose the audit trail, data lineage and impact analysis that is an essential part of data integration unless specific compensating features have been built into the product. In the case of code generating products, on the other hand, which are designed precisely for this purpose, you do not lose any of that capability.

Of course there is a downside. Because you are running generated code on the host system rather than an intermediate processor, there is an additional load on that processor. However, this should be mitigated by the fact that Talend applications can be deployed across a grid of servers and because of the parallelisation introduced with this release. Conversely, an intermediate processor can create a bottleneck. Also, Talend supports an ELT-based approach whereby all the transformation processes are hosted on the target system. This is particularly important in data warehousing environments because you take the load off the operational system where performance is likely to be most critical.

Source/target support

The product has native support for more than 30 databases (including not just the most well known products but also data warehouse appliance vendors such as Netezza, Vertica and Greenplum as well as the latest entrants into this market such as ParAccel, Kickfire and InfoBright) and ODBC/JDBC can be used otherwise. In this release, the actual process of connecting to databases is wizard-driven. Specific capabilities for supported databases are extensive. For example, for MySQL there are separate components for bulk loading, commits, reads and writes, rollback and so on. There is support for slowly changing dimensions (types 1, 2 and 3) and this applies regardless of whether you are using an ETL or ELT-based approach.

Note that there are also facilities to develop your own components (of which some 400 are provided out of the box), which can then be shared with the rest of the Talend community, as discussed above.

In addition to formal database support, Talend Open Studio also supports access to various file types, including unstructured data, zipped files, encrypted files and XML.

Architecture

Talend Open Studio has three major components: the Business Modeler, the Job Designer and the Metadata Manager. Before discussing each of these it is important to note the Business Modeler. As its name suggests, this is aimed at business users whereas the Job Designer is targeted at developers.

In principle, Talend recommends that you use these three components in the order stated. By starting at the business level and using a top-down approach you should help to eliminate any misunderstandings between the business user and the developers. However, this is not mandated and you can use the components in any sequence, as required.

Product description

Business Modeler

Figure 1 shows a screenshot of the Business Modeler.

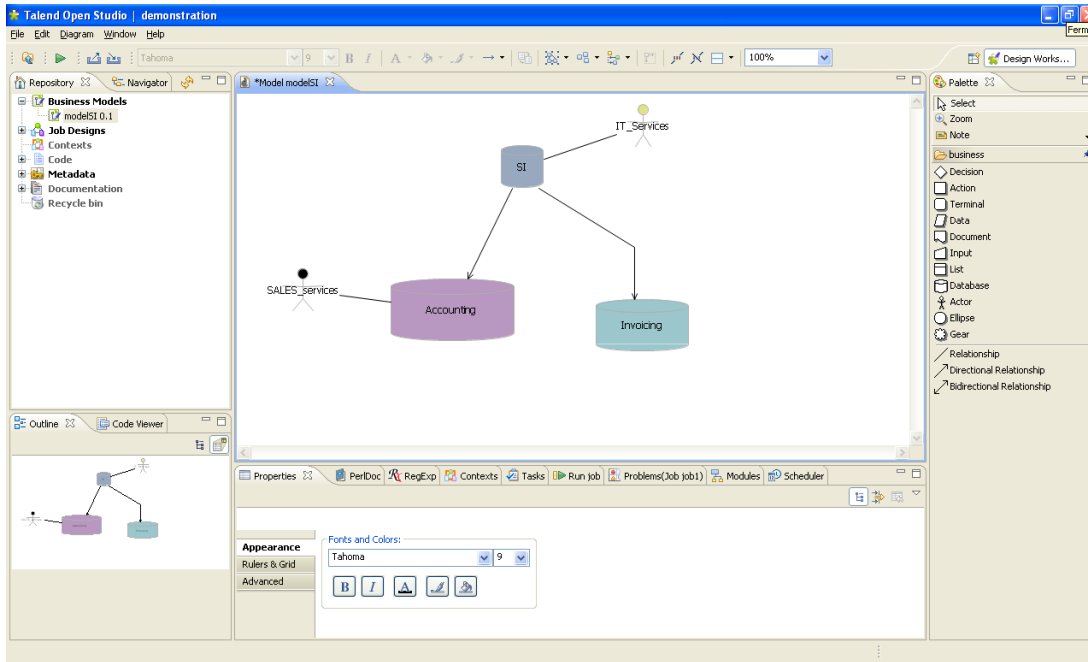


Figure 1: The Business Modeler

As can be seen here the Business Modeler has its own business oriented palette on the right hand side of the screen; the overall environment using a conventional drag-and-drop approach, but at a high level that is suitable for business analysts.

Job Designer

The Job Designer provides a fairly standard look and feel, both in terms of the industry as a whole and paralleling the Business Modeler, with a technical toolbox appearing on the right hand side of the palette, as illustrated in Figure 2.

In this screenshot note the various options within the technical toolbox: the ELT tab; the File tab, which is used to select non-database processing; the Log and error reporting tab, from which you can filter and from which you can generate emails; the Processing tab, which allows you to apply options such as normalisation or de-normalisation, and to call custom Perl or Java routines if needed; and so on. There is also a visual SQL Builder; debugging facilities that include a trace mode that allows you to step through your application,

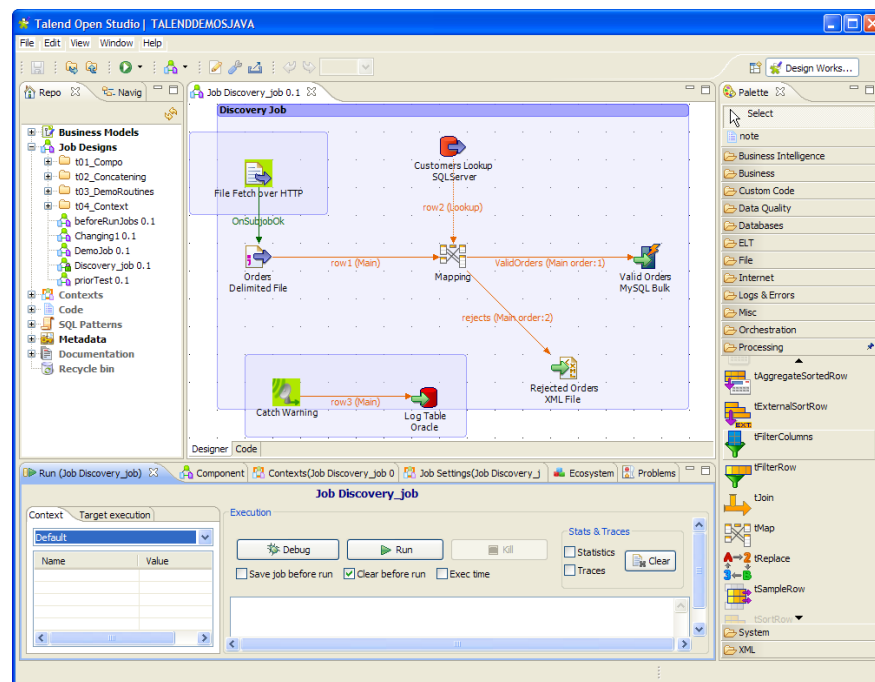


Figure 2: The Job Designer

breakpoints and the ability to inspect variables; and there are data cleansing facilities for matching data that include fuzzy matching (based on the Soundex, Metaphone and Levenstein algorithms), interval matching and others, plus data profiling capabilities. Also, name and address cleansing capabilities were recently introduced into the solution. Note that data quality processes can be embedded directly within a data movement process.

Product description

In addition to these, we want to highlight the mapping component, an illustration of which is shown in Figure 3.

The first thing to note here is the colour coding of the mapping arrows. The arrow or arrows highlighted are the ones that we are currently working on, while the grayed out ones have been previously defined. Further, yellow arrows are used for standard mappings while orange is used for constraints and purple for relationships. Secondly, note the Expression Editor tab: this can be used to do things such as convert target data to upper case or to multiply data values by a constant.

We should also mention that Talend Open Studio includes context management capabilities. This is a facility that we have not seen in other products. What it does is to allow you to apply parameters depending upon the context you are working in. In particular, it means that you can set parameters such as "read first 100 records" when working in test mode as opposed to reading all records when in production. This obviates the need to make such changes manually and eliminates the possibility of forgetting to do so.

Once you are happy with your design you simply generate autonomous applications in Java or Perl and then deploy them to whatever servers are appropriate. The software will also generate technical documentation for you automatically (and dynamically), which is created in a zipped XML file.

Metadata Manager

Open Studio includes a centralised repository to enable reuse. In particular, it has the ability to support job properties. This is important because the alternative is to define these individually within each application. This can create problems when there is a change, in particular because you have to change every application that uses such properties, whereas this sort of reuse is automated if applied via the metadata manager. The actual process of defining job properties in the repository is wizard driven with automated facilities for recognising datatypes, field and row separators, footers and headers and so on.

In addition, in version 3.0 Talend supports so-called joblets. These are sub-processes that you want to reuse. For example, you might have the same process that you want to use across several different source databases. Using joblets you could define a joblet for the process itself and then create different jobs simply by adding in the relevant source data.

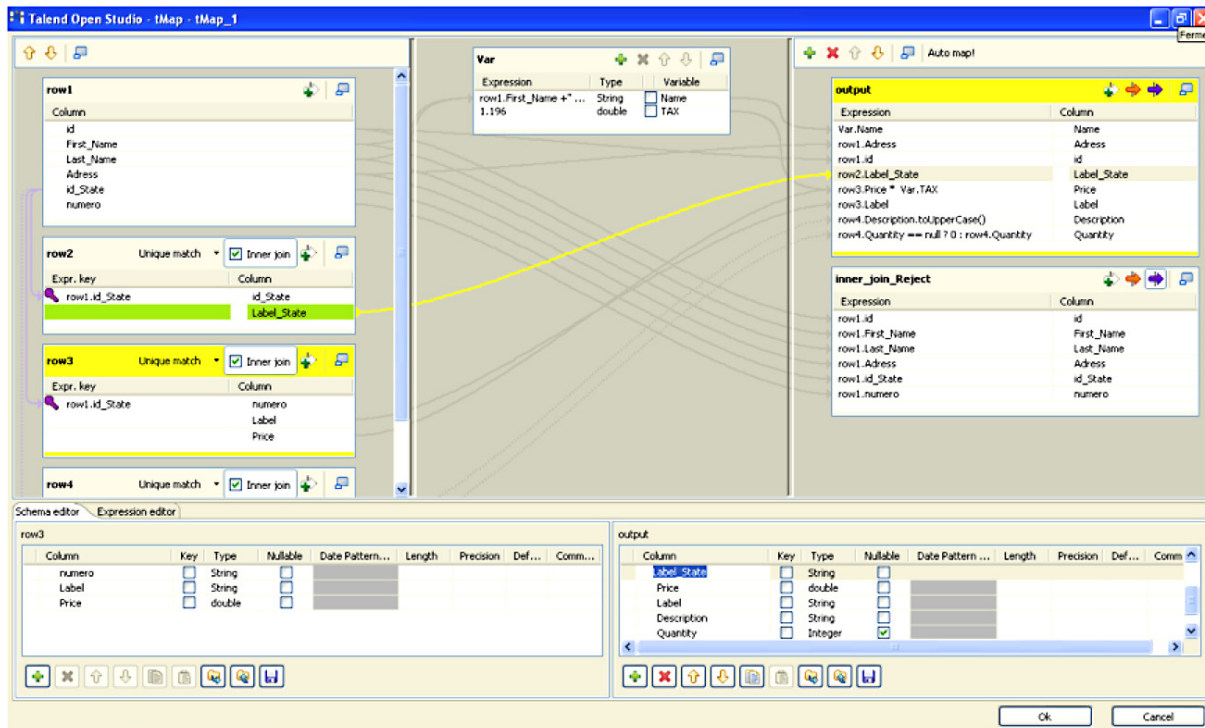


Figure 3: The mapping component of Talend Studio

Talend Integration Suite

As has been noted, the Talend Integration Suite combines Talend Open Studio with a support contract for large enterprise deployments. However, it also provides some extended capabilities, particularly with respect to metadata management, where Talend provides a full multi-user repository with version control, check in and out, support for roles and permissions and so forth.

The Talend Integration Suite also adds deployment capabilities driven through a common execution interface from which jobs can be started upon request, or you can use the built-in scheduler, which supports both time-based and event-based scheduling. Available execution servers are automatically mapped, with constant monitoring of their

resources, in order to intelligently load-balance the execution of jobs. In addition, execution monitoring is also provided, as illustrated in Figure 5. Further deployment features include grid optimisation, automatic fail-over and parallelisation across CPUs and cores, including synchronisation and wait points. In fact, a major new feature in this release is the product's extended support for parallelism. This allows you to create different threads for different processors and, optionally, to parameterise this process so that you can decide at run-time the extent of the parallelism that you wish to deploy. Remote execution of jobs on specified systems, for testing and running jobs upon request on specific systems, is also supported.

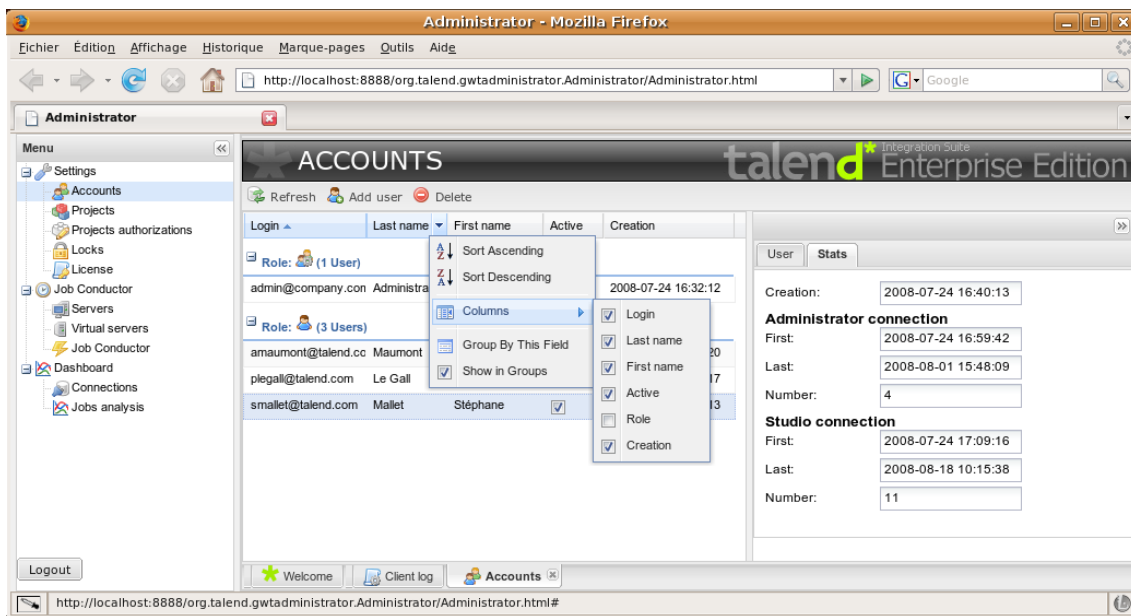


Figure 4: The Talend Integration Suite Administration Console.

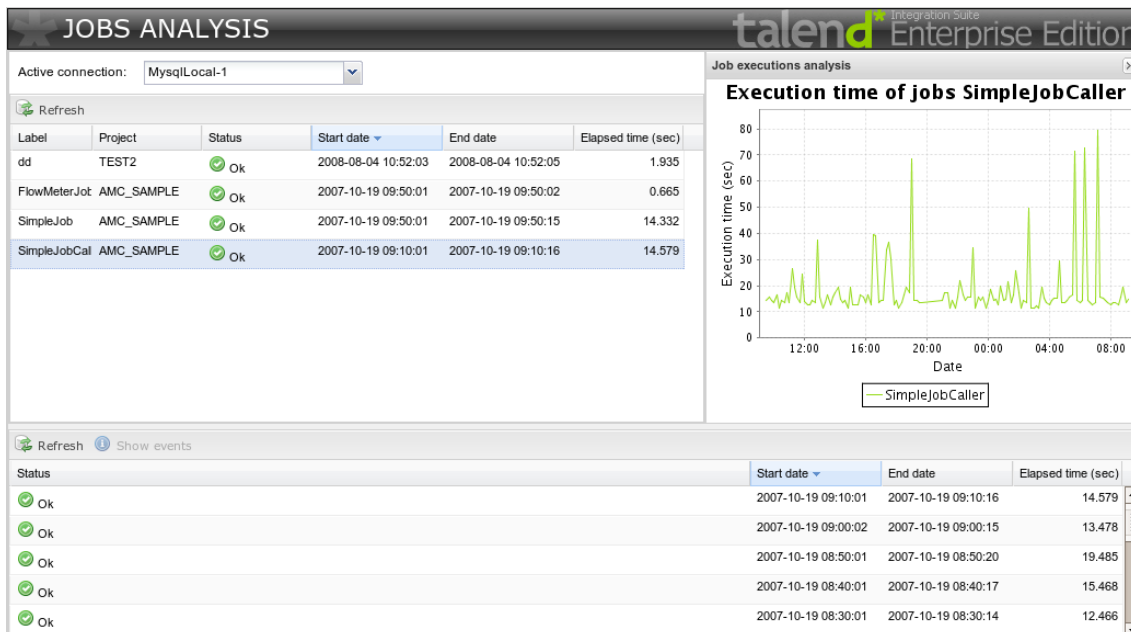


Figure 5: Execution monitoring

Vendor information

Talend was established, in Paris, in 2002 and it first came to market in August 2005 with the beta version of Talend Open Studio. The company is privately owned and backed by venture capital. It is an open source vendor and the product is available simply as a free download from the Talend web site. There is therefore the question of where the company will make its revenues? In practice there are three answers to this: the first is that Talend Open Studio is marketed through indirect channels to partners that want to embed Talend Open Studio's capabilities within their own product sets. These partners may or may not be open source vendors themselves, as is JasperSoft, for example. In any case, Talend provides chargeable support to these partners.

Secondly, Talend charges for training, support and technical expertise (consulting) via a direct sales model. To this end the company has opened offices in the United States, Germany and China as well as its native France, and is opening one in the United Kingdom in early 2009.

Talend web address: www.talend.com

Thirdly, Talend has two additional offerings: Talend On Demand, which is a software as a service (SaaS) offering whereby Talend charges for its hosting facilities; and the Talend Integration Suite, which is a combined services and product offering for large enterprises and is sold as an annual subscription. This includes some features (for example, support for team-based development) that are not available in the standard Talend Open Studio product. The same product structure applies to the company's data quality offerings: Talend Open Profiler is provided under the GPL license whereas Talend Data Quality is sold as an annual subscription.

Finally, it is worth noting that Talend has partnerships with other open source vendors, such as MySQL, Ingres and SugarCRM, as well with a number of systems integrators; and that Talend is a co-founding member of two important consortia: the Open Solutions Alliance and OW2.

As is indicated by the fact that Talend has offices around the world, it is able to offer 24x7 support.

Talend community web address: www.talendforge.org

Summary

In our view, Talend is the most enterprise-oriented of the open source data integration vendors. You can see this from the features in its Integration Suite and from its expansion into overseas offices. The advantage, of course, is that you can try out Talend Open Studio for free, perhaps for a departmental project, and then you can upgrade to the Talend Integration Suite for further projects and enterprise-wide deployment if you like what you have found. This potentially represents a major challenge for established vendors who normally charge six figures for their software, since Talend can easily and inexpensively penetrate their user base. How successful Talend will be with this strategy remains to be seen but the features and capabilities of Talend Open Studio and the Talend Integration Suite certainly make this a real possibility.

Further Information

Further information about this subject is available from <http://www.BloorResearch.com/update/996>

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.



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.

Copyright & disclaimer

This document is copyright © 2009 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.



2nd Floor,
145-157 St John Street
LONDON,
EC1V 4PY, United Kingdom

Tel: +44 (0)207 043 9750
Fax: +44 (0)207 043 9748
Web: www.BloorResearch.com
email: info@BloorResearch.com