Market Report

Bloor

Data Profiling and Discovery

A Market Report by Bloor Research
Author : Philip Howard
Publish date : August 2012

Since we last reported on the data profiling and discovery markets a significant shift has taken place

Philip Howard

Free copies of this publication
have been sponsored by SAP

Introduction

This paper is a customised version of a market report into data profiling and discovery that was researched and published during the first half of 2012. The only difference between this paper and the original report is the inclusion of this introduction and an extended discussion of SAP Information Steward. In all other respects this paper is identical to our original report.

 © 2012 Bloor Research

Market trends

In 2009 we introduced the concept of data discovery as distinct from data profiling where we defined data discovery as "the discovery of relationships between data elements, regardless of where the data is stored". This distinction is important because data discovery has far wider application than just data quality. For example, data discovery is important when implementing MDM (master data management), it can be used to complement data modelling tools, it may be employed for business intelligence purposes, and has a significant role to play in supporting data migrations, data archival and data governance, amongst other areas of application. At that time there were data profiling tools that did a little of this, but not much, while there were data discovery tools that could discover relationships but did not do much in the way of statistical analysis and monitoring to support data quality initiatives.

That positioned has changed. Since we last reported on the data profiling and discovery markets a significant shift has taken place. It is apparent that many traditional data profiling vendors have been adding data discovery capabilities to their products while suppliers of data discovery tools have added statistical and profiling functions to their tools. While some vendors are clearly further down this path than others, you might therefore conclude that data profiling and discovery should be re-merged as a single market sector. However, that is not currently the case.

What appears to be happening is that while some vendors have opted to go down the route just described, others seem to be focusing just on profiling. Moreover, it is apparent that it is the vendors of less expensive products that are opting out. So what the market looks like today is a complete reversal of how it looked just a couple of years ago. Where we previously had a lot of profiling but not much discovery now most vendors offer some reasonable degree of discovery but there remain a few that focus specifically on profiling.

The second most important trend is towards the use of tools to discover personally identifiable information (PII), personal health information (PHI) and other data that needs to be subject to privacy and protection. This is typically done, in the case of credit card numbers for example, by defining the relevant pattern and then using a profiling tool to search for this. Relevant masking techniques can then be used to hide the data or the data can be flagged for remediation if it appears, for instance, in the middle of an address field. Note that this is a discovery technique that has nothing to do with relationships per se. However, any relationships that exist will need to be preserved during any masking process: you can't just mask willy-nilly.

Finally, the other most significant trend in the market (and not just this market) is towards support for big data. At present only around half of vendors have dipped their toes into this area and, almost invariably, the support offered is for Hadoop; and only Hadoop. Only one vendor supports MongoDB, no-one supports Cassandra and no-one supports any of the graph databases. No doubt this will change over time but support in this area can best be described as nascent.

Key market issues for data profiling

The key issues that distinguish products in this market are the extent to which the different tools extend beyond the core capability that you would expect from any product. This applies in a number of ways.

Firstly, the extent to which the product supports multiple, heterogeneous data sources. This is more widespread than it used to be. The ability to handle large numbers of sources is, if you like, a measure of scalability (as is the ability to perform appropriately with very large tables that contain rows numbered in the billions). There is also a question of the extent of heterogeneity supported: can you support flat files, XML (without a third party tool to flatten it), COBOL copybooks, spreadsheets, non-relational databases and so on?

More technical considerations are concerned with where the profiling takes place and against which sets of data. Ideally, you would like to profile in situ or by extracting the data, with discovery run against all of the data, or a sample, as required. There are also hybrid approaches where some profiling is done on the source systems but where you create cross-reference tables (say) that are held locally. Which is most suitable will depend on the number of sources, their complexity and the task you are trying to achieve. Flexibility will mean that the tool is more suitable for a wider range of tasks. If you are going to use data profiling as a part of broader data quality initiatives then you should be able to run data cleansing and matching routines without having to re-parse the information that you have already parsed for profiling purposes.

There are some basic data discovery capabilities that one would expect from a profiling tool. For example, you would expect to be able to identify redundant columns and primary/foreign keys (even when these are of different datatypes or field lengths). However, more complex discovery will typically be reserved for tools that are aimed at discovery as well as profiling and these issues are discussed in the next section.

Profiling is, in large part, a manual task. It is also tedious. Thus anything that can be done to reduce the amount of manual effort involved will be an advantage. This is particularly true if you have a large number of sources to analyse and/or if these are particularly complex. For example, if you are trying to determine candidates for primary/foreign key pairs then it would be nice if the software automatically tried all possible pairs for you and presented them to you in order of likelihood rather than just giving you a list of possibilities. Similar considerations apply to other requirements such as overlap analysis. In general, automation is particularly relevant when you do not know what you are looking for as opposed to looking for something that you already expect. For example, discovering exceptions to relationships (business rules) that have been predefined is one thing but looking for similar exceptions to rules that you do not actually know about is of an order of magnitude more complex and will therefore benefit from increased automation.

Data profiling is, or can be, an important collaborative tool. It is typically business analysts and domain experts who are best placed to validate business rules, for example, but on the other hand much of the information that is uncovered by data profiling is also of value directly to developers and to data management. It will be helpful therefore, if the product has functionality that will assist both of these constituencies. Support for a business glossary, an understanding of semantics, the discovery of attributes (constant, reference data and so forth) that may be of value to an analyst, workflow capability and the ability to visualise discovered relationships through entity-relationship diagrams (or something similar) will be useful. In addition, profiling may well be used to monitor data quality on an on-going basis. For example, you may decide to cut-over a data migration project only after data quality metrics have exceeded a particular threshold: in this case you will therefore also need dashboard capability and the ability to capture or use quality metrics.

 © 2012 Bloor Research

Key market issues for data profiling

On the statistics side, while there is commonality about statistics, such as the number of nulls, that doesn't mean that there are no issues with respect to these figures. For example, you would like to be able to distinguish between hidden sub-types. By way of illustration suppose that you have a table of financial instruments containing data on both bonds and equities, including a column for maturity date. A bond has a maturity date and thus must not be null but an equity doesn't, so it must be null. Simply reporting the number of nulls is not enough.

Another major issue is that if you are checking rules about your data then most tools will simply tell you about any exceptions that have occurred. However, some tools cannot cope with multiple rule violations. What you would really like to know is what percentage of records have no violations, one violation, two violations, and so on. Going a step further, you would also like to monitor this over time and be able to compare these figures with a baseline to get comparative confidence levels for the data. This is essentially a part of data validation functionality that relatively few products build in but which should provide automated testing and validation, not only during the normal course of events but also to support product upgrades where you want to re-check the data and its rules for validity.

Key market issues for discovery

Some of the functionality required of discovery has already been discussed but other features you would like to see include business and transformation rule discovery, exception detection against discovered or predefined business and transformation rules, data validation, dependency analysis, overlap analysis, precedence analysis, the discovery of cross-source binding conditions, matching key evaluation, outlier analysis, clustering, sub-schema and sub-type profiling, recognition of join key values that match multiple times (which is an often overlooked reason for unexpected data multiplication) and so on. Needless to say, a number of these requirements are only relevant in multi-source environments. If you are only profiling a single source then many of these requirements will not be necessary.

The second and most commonly used approach to discovering relationships is data profiling. These tools are usually, though not always, marketed as part of a wider data quality offering. This has had unfortunate consequences in that the relevant vendors have tended to view data discovery simply as a function of data quality and have not leveraged its capabilities outside of that environment as much as they might have. This is surprising given how many data quality vendors are active in MDM (for example) where overlap and precedence analysis, as well as the discovery of matching keys, are of fundamental importance in determining the best source(s) for loading the data into the MDM hub, but which are not supported by most products.

In terms of collaboration, and to support data stewards in particular, facilities such as a business glossary and the ability to visualise discovered relationships will be important. This last is especially important where discovery is being used to support archival, migration or MDM initiatives because of the need to be able to visualise a business entity such as a customer or product in its entirety, in business terms.

 © 2012 Bloor Research

Evaluation

From a technical perspective we have evaluated products under five categories and otherwise with respect to the commercial strength of the vendor, the support offered, the ease of use and implementation, and the value offered by each product. The technical categories are as follows:

Architecture and performance, which includes details about the platform, whether data is profiled in situ or has to be extracted, the level at which the product works (data or metadata or both) and so on.

Integration, which includes integration with other environments such as data cleansing, master data management, data archival and so forth; as well as the sources supported (including big data sources) and the depth to which they are supported.

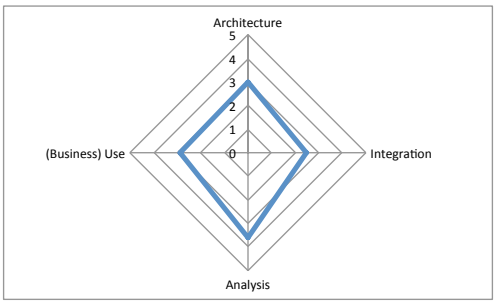
Analysis and rules, which includes statistics, actual profiling results, how extensive profiling can be, the degree of support for rules, and so on.

Relationships and discovery, which includes the identification of relationships and rules and exceptions to those relationships and rules. This heading also includes the establishment of business entities through discovery. This category has been given a higher weighting within the discovery landscape diagram as opposed to the profiling landscape diagram. There are also elements included in the former that are not included in the latter.

(Business) use, which includes functionality provided for different classes of users and the extent to which they are supported. There are some elements in this category that have not been included within the profiling landscape diagram as they are particularly pertinent to discovery.

Information has been gathered from discussions with vendors and users, and via a questionnaire. Details of the survey used (for technical aspects of each product) are shown in the Appendix. Users are explicitly exempt from copyright for this questionnaire and may replicate these questions for use in their own product evaluations.

The vendors



Data Profiling

Ataccama

Ataccama is a data quality and MDM vendor that originated in the Czech Republic, though its headquarters are now in the United States. Founded in 2007, the company is a part of the Adastra Group, which is Canada-based.

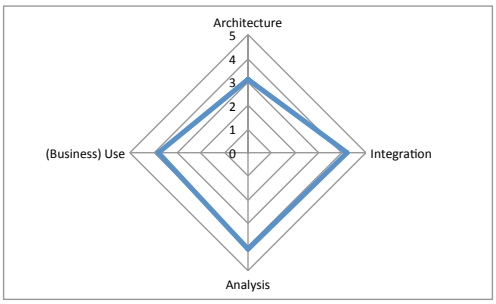
Ataccama's data profiling tool is DQ Analyzer and it is available as a free download from the company's website. This is compelling for users that want to demonstrate to their organisations that they really have a data quality problem but who don't believe it. Of course the purpose of making this available for free is to entice those users to subsequently license the company's DQ Center, as well as other products, in order to resolve the problems identified by DQ Analyzer.

As a product DQ Analyzer is clearly in the profiling and not the discovery camp. This is somewhat surprising given that Ataccama markets Master Data Center as an MDM offering: one would expect that features such as the ability to identify matching keys would be a useful ability in support of MDM. Also noteworthy is that DQ Analyzer only provides statistical analysis of your data. Unlike other profiling products it has no ability to monitor data quality on an on-going basis. This is provided through a separate product called DQ Dashboard, which is chargeable. One feature of the product suite that is worth special mention is DQ Issue Tracker, which supports data governance and is unique in the marketplace. To summarise: DQ Analyzer as a pure-play profiling tool is decent without being exceptional. On the other hand, given that it is free, it is sensational value for money.

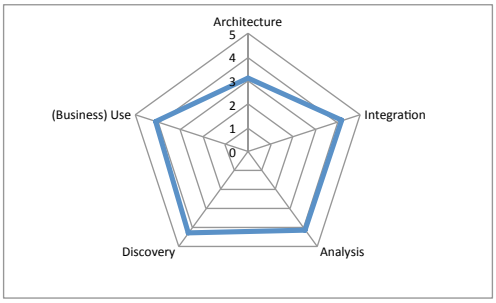
Finally, note that Ataccama software is resold by iWay (a division of Information Builders). iWay has embedded both DQ Center and Master Data Center into its enterprise service bus.

 © 2012 Bloor Research

The vendors



Data Profiling

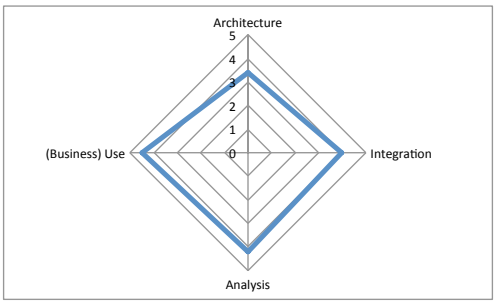


Data Profiling & Discovery

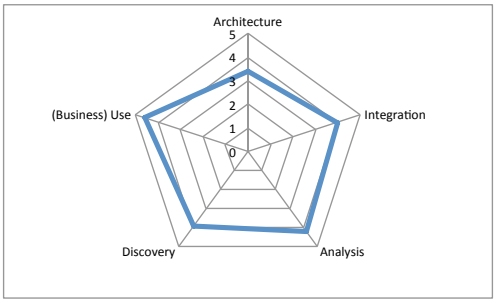
Datactics

We are particularly impressed by Datactics. Based in Northern Ireland (and with an office in Chicago) the company covers the gamut of data quality and master data management. It is a relatively small organisation compared to many of its competitors. However we rate it from a technical perspective as a leading contender for both profiling and profiling combined with discovery. We especially like what the company is doing with its data quality extensions and predictive data quality. While the former should be self-explanatory (and the details are confidential) the latter (also confidential) is effectively doing for data quality what in-database analytics does for data warehousing, with embedded capabilities provided for DB2, MySQL and EMC Greenplum at this time. Perhaps it is the fact that it is a relatively small company that allows it to be so innovative? The company also has an enviable reputation amongst its client base for the quality of its service.

The vendors



Data Profiling



Data Profiling & Discovery

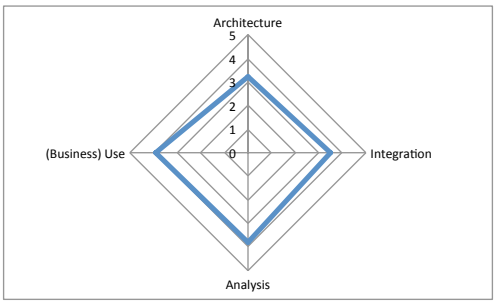
DataFlux

DataFlux is a division of SAS. It is a data quality and master data management provider and also provides data integration (ETL/ELT), data federation and event processing products. The company markets its products direct to end users but SAS also embeds DataFlux products within its own solutions.

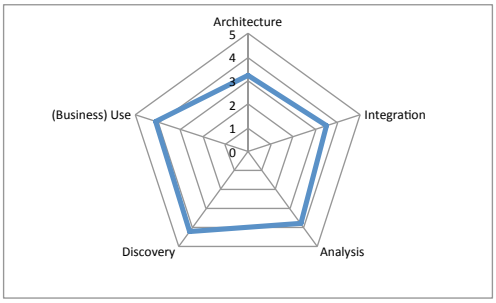
According to our research, the DataFlux Data Management Platform is the leading product in the market scoring the highest marks for its technical competency for both profiling and profiling combined with discovery. In particular, DataFlux out-scored all of its competitors in its support for users, both at a business and technical level, especially for profiling combined with discovery, where it was significantly in advance of its competition.

 © 2012 Bloor Research

The vendors



Data Profiling



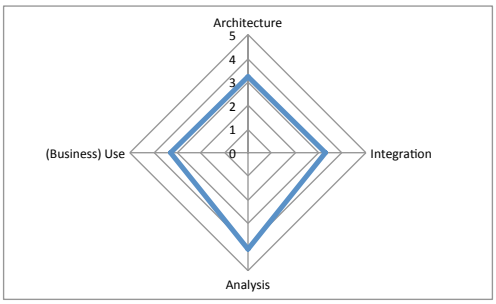
Data Profiling & Discovery

Global IDs

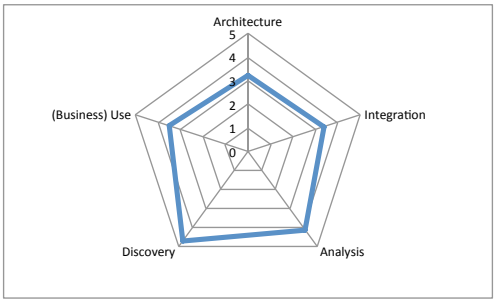
Global IDs is a small company based in New York that is quiet about its marketing and that does not have a presence outside North America. While it does not have a large number of customers, those that it does have are large and prestigious. This is because Global IDs focuses particularly on large and complex environments. It has a very broad product set that covers all data management functions (data quality, master data management, data governance and data integration), based on a single integrated platform. It is probably the most mature of such platforms currently in the market.

Because of the complex nature of the problems that Global IDs targets, its Data Profiling and Mapping Suite is probably more suitable for environments that require discovery as well as profiling, though its pure profiling capabilities are more than adequate. While the product scores highly for profiling and discovery our survey probably does Global IDs a disservice by emphasising more general capabilities rather than the features required to tackle more complex problems. Thus the scores for Global IDs potentially err on the low side.

The vendors



Data Profiling



Data Profiling & Discovery

IBM

Since our last report into this market IBM has acquired Exeros, which was market leading for discovery purposes at that time. It should therefore come as no surprise that IBM offers the best understanding of relationships of any product that we have examined. For pure profiling capabilities IBM InfoSphere Information Server is good without being outstanding but it is clearly one of the market leaders when discovery is required alongside profiling.

 © 2012 Bloor Research

The vendors

Informatica

Despite repeated promises to complete our questionnaire Informatica ultimately failed to do so. We trust that the company does not treat its potential customer's requests for information in a similar fashion. For this reason we are unable to provide a detailed technical breakdown and scoring for Data Explorer, which is the company's profiling and discovery product. However, the company is included in our overall landscape diagrams for both of these categories based on publicly available information, together with conversations and briefings with the vendor. Nevertheless, our general principle in such instances is to assume that a product does not have relevant features where there is doubt. It is likely, therefore, that Data Explorer is actually better suited to these markets than the product's overall score suggests.

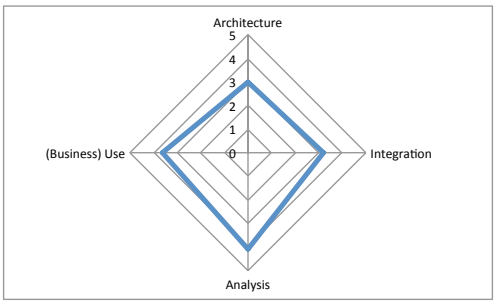
Microsoft

With SQL Server 2012 Microsoft has introduced Data Quality Services (DQS), which incorporates data profiling, as a part of the product. This new solution provides a significant enhancement to the sort of facilities that were in SSIS in previous versions of the database. The good news is that Data Quality Services come free with the database, but the bad news is that it is optimised specifically for Microsoft SQL Server sources (SQL Server, Excel and CSV files). Non-Microsoft sources can connect to DQS via SSIS workflows or Excel.

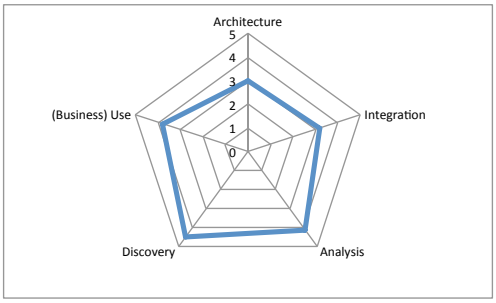
In so far as profiling is concerned this is very much an analysis and statistical tool: it has very limited discovery capabilities and is not really aimed at that market.

Microsoft did not complete our questionnaire so we are unable to provide a detailed technical breakdown and scoring for the product. For profiling of SQL Server-based data it represents excellent value but we cannot recommend its use for environments where there is a significant non-Microsoft component; further, it is not suitable for discovery purposes to support functions such as master data management, database consolidation or archival. For this reason Microsoft only appears on our profiling landscape diagram and not for discovery combined with profiling.

The vendors



Data Profiling



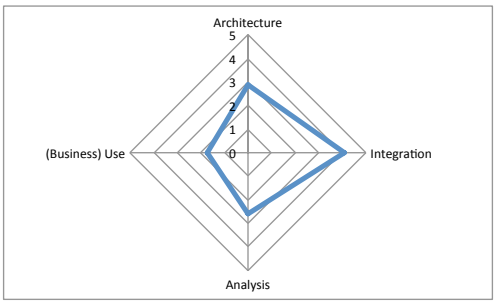
Data Profiling & Discovery

Oracle

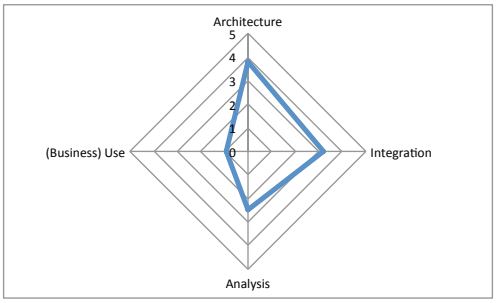
Oracle resells Trillium Software's solutions. As such there is not much to say that is not included in the section on Trillium below. The most notable point is that TS Discovery (and therefore Oracle) ranks as the leading product for functionality (analysis, discovery and business/use) in total. Our only consideration is that Oracle also offers other data cleansing (though not profiling and discovery) solutions that it has acquired in recent years, notably from Silver Creek Systems and Datanomic. Needless to say, one would not expect such good integration between TS Discovery and these other tools as exists within the Trillium product family. For this reason we have marked Oracle down on its integration capabilities compared with Trillium but otherwise the scores are the same. Of course, if a wholly Trillium-based solution is to be implemented then there would be no change in the marks.

 © 2012 Bloor Research

The vendors



Data Profiling: Data Profiler



Data Profiling: RushAnalyzer

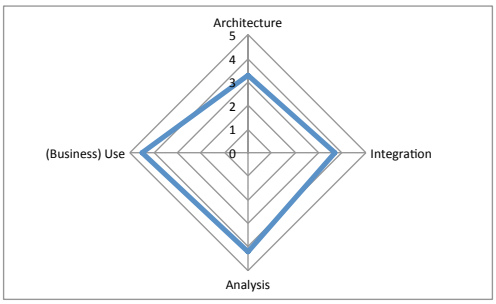
Pervasive

Pervasive, which is based in Texas and has offices worldwide, has two products in this category: Pervasive Data Profiler and Pervasive RushAnalyzer. The former is a traditional data profiling tool but the latter probably needs some description.

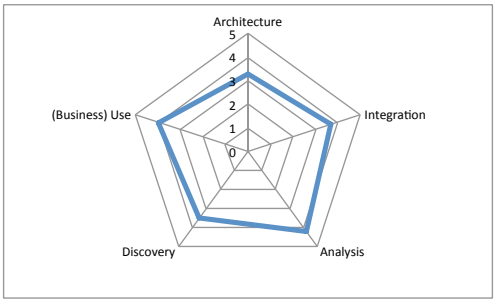
RushAnalyzer is an application of the Pervasive DataRush technology. This is a (very) high performance platform for parallelising a variety of tasks. In particular, the software has been built to parallelise across the cores within each processor so the product is not just very fast but also cost effective. One of the applications for DataRush is for high performance loading into a data warehouse for data mining and similar tasks. With increases in data volumes (big data), this is becoming an increasingly important requirement. In order to further support this environment, RushAnalyzer includes as a profiling capability that runs on DataRush and profiles the data for you during the load process. In effect, this means that by the time data is loaded into your warehouse you also have all the statistics about that data already calculated for you. While this is a specialist application of profiling and not for general-purpose use, it could be very valuable when the circumstances are right.

Data Profiler is firmly targeted at profiling and RushAnalyzer is targeted more generally at Big Data ETL and predictive analytics. Neither is intended to aid in discovery processes. In addition, it is worth bearing in mind that while Pervasive certainly has enterprise-class customers it tends to focus on its large partner community and SMBs. As a result, and at the price points you would expect from such a supplier, its products do not have all the bells and whistles that you might expect from a more expensive vendor. So, while the products do not score especially well, potential users should consider the value-for-money aspects of these solutions. The two exceptions to this are Data Profiler's integration (no more than you would expect from a vendor that specialises in connectivity) and that we would expect RushAnalyzer to outperform any other product in this report. Conversely, as a newly introduced product it is not surprising that it does not yet have all the features one might like. It is also worth noting that since RushAnalyzer is a more general tool for big data ETL and predictive analytics, it doesn't have all the data profiling functions found in a tool specifically dedicated to data profiling. Its score in this category is not really reflective of its main intended uses.

The vendors



Data Profiling



Data Profiling & Discovery

SAP

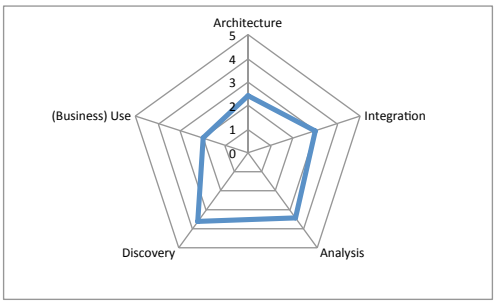
SAP Information Steward from SAP is one of the leading products in the marketplace. However, it is clear from our analysis that it is not quite as advanced as some of its competitors when it comes to adding discovery capability to its existing profiling. Thus SAP Information Steward scores the highest marks (4.5 out of 5) for business and technical use within a profiling context as well as having the highest overall score on our landscape diagram for pure profiling (4.2) but those scores drop to 4 and 3.9 respectively (still excellent but not best) when the same considerations are applied to discovery as well. SAP did inform us that a number of relevant new features are going to be introduced into the product in the future but, while we are pleased to hear this, we cannot give full credit for them at this time.

What the preceding does not make clear (because our report was focused exclusively on profiling and discovery) is that SAP Information Steward provides precisely what its name suggests. That is, a single place for data stewards and business analysts to collaborate, and to monitor and manage the organisation's information assets. This means that there is rather more to SAP Information Steward than many of the other products included in this report. In particular, SAP Information Steward supports the creation and maintenance of a business glossary, metadata management functionalities for data lineage and impact analysis, and, in addition to supporting the definition of business terms you would also use SAP Information Steward to define validation and cleansing rules as well as models. The single user interface for this entire environment is also where you assign data ownership, accountability and roles within the context of stewardship.

In summary, SAP Information Steward provides a single, holistic environment for all of the functions that are required by a data steward, including supporting his or her interaction with those members of the business that have responsibilities with respect to the management and governance of information.

 © 2012 Bloor Research

The vendors



Data Profiling & Discovery

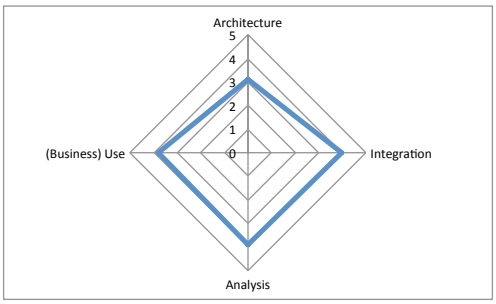
Sypherlink

Sypherlink is a US-based company with offices in the UK and India. It started life as a company focused on discovery and that remains core to its approach. However, it also has a specialised vertical practice catering to the Justice, Public Safety and National Security domains. To serve these environments, Sypherlink supports the NIEM (National Information Exchange Model) standard, which provides a common language, a universal vocabulary, and a framework by which state, local, tribal, and federal government agencies may share data in emergency as well as day-to-day situations.

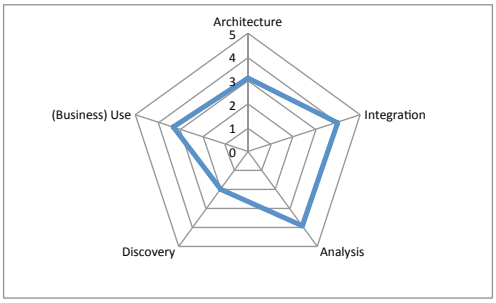
As we have noted, Sypherlink is primarily focused on discovery and mapping. It has some profiling and statistical capabilities but that is not its strength. As a result, Sypherlink is only included in our combined profiling and discovery landscape diagram. Unfortunately, this still does Sypherlink a disservice as there are clearly significant aspects of profiling included in this that Sypherlink does not do particularly well. As a result the figures shown underestimate the product's capabilities for pure discovery environments.

Sypherlink has a long-standing partnership with ASG (and that company's Rochade repository) and it also has the ability to generate ETL (extract, transform and load) transformations based on the relationships it discovers and can thus act as a precursor to the use of ETL tools.

The vendors



Data Profiling



Data Profiling & Discovery

Talend

Talend is the leading open source vendor of data management solutions that span data integration, data quality, master data management and, uniquely, it also provides business process management and an enterprise service bus. Thus it describes its range of products as a "holistic integration platform" that goes a long way beyond traditional data integration or even data management.

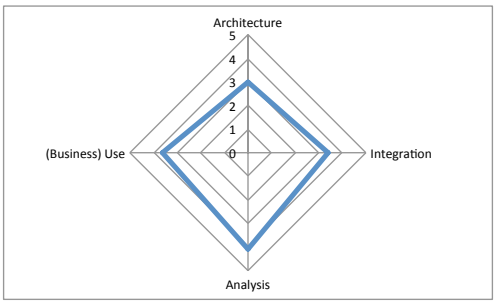
When we last looked at Talend's data quality (cleansing and profiling) solutions they were, to be frank, not very good. In a remarkably short space of time they have matured to the point of, if not being the best, certainly well able to compete. Moreover, we mean this on a pure feature/function basis without regard to costs, where Talend has a very significant advantage over its proprietary competitors. This rapid development of features is, in part, a result of the open source environment in which Talend operates, which significantly shortens development cycles.

In general, Talend is much stronger when it comes to profiling the data than the discovery of relationships. It is up there with the leading vendors in the former category but lags some way behind when it comes to discovery. Given the rate of development that we have already discussed we would expect this position to shift in Talend's favour quite quickly.

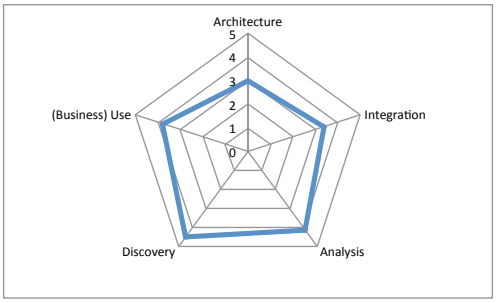
One final point that is worth noting is that Talend is more advanced in the range of NoSQL and Big Data connectivity that it can offer than any other vendor. It is the only supplier, for example, that has integrated with MongoDB and it has a close relationship with all the major Hadoop distributions. Given that most NoSQL developments are based on open source projects this lead is perhaps not surprising.

 © 2012 Bloor Research

The vendors



Data Profiling

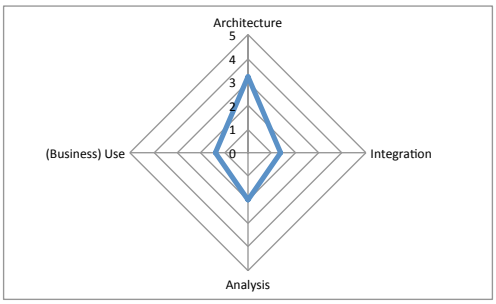


Data Profiling & Discovery

Trillium

Trillium Software is a division of Harte-Hanks, a multi-billion dollar worldwide organisation. The company focuses solely on data quality and it is a leading vendor in this marketplace. While Trillium tops no single category in our survey TS Discovery ranks as the leading product for functionality (analysis, discovery and business/use) in total. The one weakness that Trillium appears to have compared to some of its competitors is that it only offers a limited number of direct database connectors and it relies heavily on ODBC. Also missing from Trillium's armoury is a lack of direct connectivity to NoSQL sources. It does offer facilities for working with such data but the company relies on third parties and partners to connect to it. Had we left these considerations aside it is entirely possible that Trillium would have emerged as the leading vendor in this report: so if you don't have non-standard environments to support then you may well wish to consider Trillium.

The vendors



Data Profiling

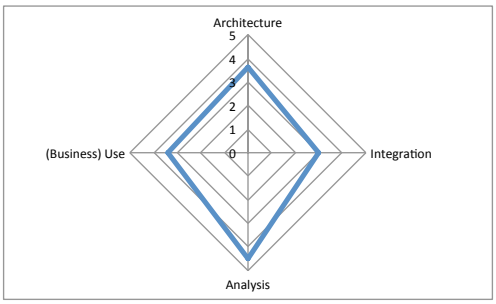
Uniserv

Uniserv is a German company, founded in 1969, that now has offices and resellers across Europe and in North America. Its original focus was on address management and from there it has extended into data quality more generally and, most recently, into data integration. Data Quality Explorer, which is the company's product used for profiling, does exactly what it says on the tin: it is used to explore data in order to support data quality processes. It is not intended as a discovery product and nor does it providing on-going monitoring of the data, which is provided by a separate product called Data Quality Monitor.

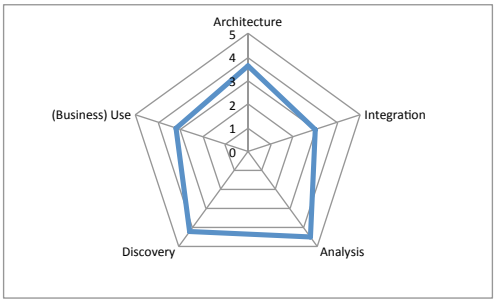
Both of these factors (lack of discovery and lack of monitoring) militate against a good score in this report. Moreover, Data quality Explorer is a brand new product, launched only in November 2011 and you would not expect it have all the features and functions of a more mature product. For all of these reasons it is therefore not surprising that Uniserv does not score well: the timing has not suited it.

 © 2012 Bloor Research

The vendors



Data Profiling



Data Profiling & Discovery

X88

X88 is a relatively young company, based in the UK, but the team has a significant pedigree in this market; the company's founder, Tony Rodriguez, created data profiling pioneer Avellino Discovery, a tool that was subsequently acquired and evolved to become TS Discovery (see above). The X88 product, Pandora, has installations in Europe and North America, and is resold, sometimes 'white labelled', by partners in many parts of the world.

In the context of a product that provides both profiling and discovery, Pandora has the highest scores for both its architecture and its profiling capabilities and, with its proven performance and scalability, it is a leading contender overall. Pandora relies on JDBC interfaces for most sources, and native adapters for XLS and delimited files, however it does not provide support for non-relational databases (including NoSQL sources). Although Pandora has been designed for non-technical users, the company does not yet offer any user-specific interfaces (you can tailor them to your own purposes easily enough, but it is not out of the box). The upcoming release of Pandora allows customisable web-based interfaces through a REST API. Pandora has its own role-based security but does not support external authentication such as Active Directory or LDAP.

Two major features that the product offers are the ability to design and generate data integration (ETL) specifications, and the ability to associate a monetary value with data quality rules.

Other vendors

There are two other vendors that we would like to draw particular attention to. These are:

Silwood Technology

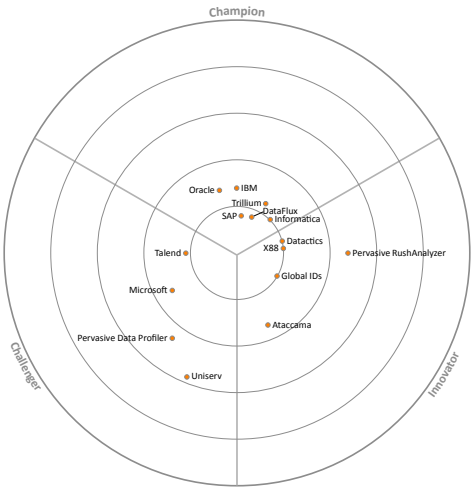
Silwood Technology, a UK-based company, has a product called Saphir which is used to discover data relationships within application environments (SAP, Oracle, Siebel, JD Edwards and PeopleSoft) rather than within databases. This is useful because many relationships are defined in the application layer as opposed to the database. These may be discoverable by implication but it is much easier and simpler if they can be discovered using Saphir which, as far as we know, is a capability unique to Saphir. Note that Saphir is re-badged and sold by IBM, but only for Oracle environments.

Xbridge Systems

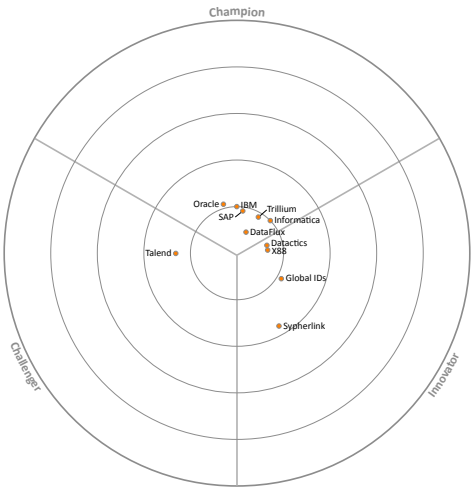
Xbridge, which is US-based, provides a product called DataSniff, which is a tool for searching mainframe environments for compliance purposes. Specifically, the company targets PCI-compliant information by algorithmic searching through the mainframe environment, against both structured and unstructured data. However, while the company is focused on PCI-based compliance, the software could equally well be used to discover PII (personally identifiable information) or PHI (personal health information) data, which would, in this context, make it competitive with the profiling and discovery tools discussed in this report.

 © 2012 Bloor Research

Summary and conclusions



Data Profiling Landscape Diagram



Data Profiling & Discovery Landscape Diagram

There is not a great deal to distinguish pure-play profiling tools from one another. They are there principally to provide supporting capabilities for data cleansing and there is no appreciable difference in the statistics that the different vendors provide. Vendors offering both profiling and discovery are a different kettle of fish in the sense that there are real differences in the sort of capabilities that they provide. However, it is also worth noting that these vendors, without exception, outscore all the pure-play profiling products, even when considered just for profiling. On the other hand, they all tend (the exception being the open source vendor Talend) to cost significantly more.

One of the biggest differences between products is with respect to integration. Some companies tend to rely on third party providers (notably Data Direct) for much of their connectivity. This is a downside compared to companies that develop native connectors and, in particular, it means that they are dependent on their partners to introduce integrated capabilities with NoSQL and Big Data sources in particular. This means that some other companies (such as Informatica and especially Talend) have forged ahead in this respect.

In so far as the discovery of relationships is concerned, there are of course some differences between products but these are largely down to which has been doing it longest (the answer is IBM through its acquisition of Exeros), though the others are running hard to catch up and are not far behind. More generally, there are a number of very strong products in this category, which are closely grouped, headed by DataFlux and then Datactics and X88. DataFlux is especially worth commending for its support for business users and overall ease of use, while X88 is particularly strong in its analytic capabilities.

Nevertheless, perhaps the most interesting company in this market right now is Datactics, which has introduced two innovative new concepts to the market. Firstly, it is now deploying in-database (for DB2, EMC Greenplum and MySQL) profiling and discovery. This works in much the same way as in-database analytics and is likely to prove just as popular for the performance benefits it brings. Secondly, it is now deploying what it calls predictive data quality. Given that a number of companies in this space (SAP Business Objects, DataFlux [part of SAS] and IBM) are all in the predictive analytics and data warehousing spaces, it seems likely these suppliers will be closely watching the success of Datactics to see if they should emulate them: we confidently expect this to be the case.

Further Information

Further information about this subject is available from
http://www.BloorResearch.com/update/2141

Appendix

The following are survey questions that companies evaluating profiling and/or discovery products may reuse:

Vendor/product details	Company name	
	Product name or names	
	Version number(s) of product(s)	
	What operating systems does the software run on?	
	What is the size of the installed base for this product?	
	In what geographies is the product marketed?	
	What are the license and maintenance/support costs for your software?	
Architecture/ technical details	Is the tool data-based, metadata-based or both?	
	Is discovery based on sampled data or always the whole data source, or are both available as options?	
	Is data & or metadata profiled in-situ or is it extracted or both? If both please state what is extracted and what is not?	
	In either case, what sort of overheads are imposed (in situ for run-time processes, extraction: while downloading)?	
	If data/metadata is extracted what sort of database or file system is it stored in?	
	For what data sources do you have native drivers?	
	Do you support ODBC/JDBC?	
	Can you understand and profile non-normalised data (such as non-first normal form NF2) e.g. nested tables within Oracle?	
	Does the software recognise native constraints such as 'small integer' datatype not supported by this database?	
	What level of support is provided for various languages and character sets? i.e. double byte, Unicode?	
	Can you recognise non-Ascii and/or unprintable characters?	
	Can profiling functions be automated?	
	Can profiling functions be scheduled by date/time etc?	
	What support is there for large data volumes? What limits are there to the number of rows or columns that can be profiled?	
Performance	Do you take full advantage of hardware - 64-bit, clusters, multiple cores? Please state which.	
	What support for parallelism do you provide?	
	What other performance features would you like to mention?	
Integration	If you also offer cleanse/match software can this work directly off the profiled data or do you have to re-process it?	
	Can you report profile results based on 'DQ Dimensions' rather than files/tables? e.g. completeness rules, integrity rules, ...	
	Can you choose items to profile based on metadata or business metadata?	
	Can you show data views behind profile statistics?	
	Can you construct reference data based on profiling results?	
	Can profiling queries be invoked as a real-time service?	
	Can profiling queries be included in a quality/cleansing workflow?	
Sources (integration)	What multi-valued sources (eg UniVerse, Unidata) do you support?	
	Which hierarchical databases do you support?	
	Can you profile against COBOL copybooks and can you extract metadata therefrom?	
	Can you profile against XML-based data?	
	Can you profile against flat files?	
	Can you profile spreadsheets?	
	Any other source types supported.	

© 2012 Bloor Research

Appendix

Big Data (integration)	What NoSQL sources do you support?	
	If you support Hadoop, which distributions and/or file types (e.g. HDFS, GPFS, Hive etc.)?	
	Is it the same product that profiles Hadoop as, say, relational data? Or a different version?	
	Can the software run on a Hadoop cluster? Does this also apply when you are profiling non-Hadoop data?	
	If you do support Hadoop can you process your discovery logic externally (source data out of Hadoop), in situ (push down discovery logic to run natively on Hadoop), or do you offer both options?	
Analysis	Does the product work directly with data elements?	
	Does the product work at row level?	
	Does the product work at the column level?	
	Does the product work cross-table within a single database?	
	Does the product work cross-table across databases?	
	If so, can these be heterogeneous databases?	
	Does the product work across different types of data source e.g. relational to XML? Please state which combinations?	
	Are there data filtering capabilities for profiling e.g. limit profiled data to certain rows/ attribute values?	
	Can you transform and re-profile data on the fly? e.g. keep only the postcode from an address and profile them.	
	Can profiling results be grouped by specific values, ranges etc.?	
	Can profiling results be stored for further analysis?	
	Does the tool support drill-down to specific rows/columns?	
	Is there support for time-series/trending of profiling results?	
Rules	Do you allow customer-specified validation rules?	
	If yes are these script/program-based or graphical?	
	Can rules be created that apply conditional logic to more than one field?	
	Can rules be created using mathematical formula?	
	Can rules be reused?	
	Can rules be accessed by external applications?	
	Can rules be applied to different sub-sets/populations of a data source? e.g. some rules could be only for active customers or current address or a particular category of contract.	
	Is an overall score produced?	
	Can rules be weighted vs. other rules?	
	Can you browse & export records which fail validation?	
	Can rules incorporate data from external sources such as D&B?	
Relationships/discovery	Can you identify duplicated or redundant columns?	
	If redundant columns are there for a reason (e.g. performance) can you turn off alerting?	
	Can you recognise duplicated or near-duplicated columns across different sources?	
	If columns in different data sources have the same name can you recognise that they are actually different (if they are) and disassociate them?	
	Do you support orphan analysis?	
	Can you recognise potential primary-foreign key relationships with a single source? Across multiple sources? Even if the sources are heterogeneous?	
	Can you rank potential primary-foreign key relationships by percentage likelihood?	
	Can you detect relationships that are not part of 3NF?	
	Can you recognise synonyms?	
	Can you recognise homonyms? If so, do you support Soundex, Metaphones, others?	
	Do you support distance measures in identifying potential synonyms/homonyms? Do you use any other techniques?	

Appendix

	Do you support join analysis? If so, can you detect cardinality? Values that do not match? Duplicates? Number of rows?	
	Does join analysis include self-joins? Many-to-many joins? Can you exclude coincidental joins (e.g. "m" stands for both male and married)?	
	Is there support for non-exact joins, fuzzy joins, phonetic joins?	
	Do you support overlap analysis? Precedence analysis?	
	Can you discover matching keys?	
	Can you discover business rules and violations thereof?	
	Can you discover transformation rules and their violations?	
	Can you automatically discover 'statistically unusual' data?	
	Can you discover relationships not defined in a data source (e.g. COBOL copybooks or other applications)? If so, please state which languages and/or applications are supported in this way.	
Business Entities	Can you discover 'business entities' (e.g. a customer with his orders, payments, service history and everything else pertaining to that customer)?	
	Do you offer facilities for (graphically) constructing business entities from discovered relationships?	
	Can you discover all of the applications that reference particular data elements, and how often?	
	Can you discover the users (business departments) that reference particular data elements, and how often?	
	Can you represent data elements (and business entities) in business terms (e.g. via a Business Glossary)?	
	Can you identify where data quality issues apply differently to different applications (e.g. what may be invalid data for one application is still valid for another)?	
Users	Can you register and assign issues to particular stewards?	
	Can multiple users work and collaborate in the same environment, sharing rules, results and metadata?	
	Can multiple users have different security and access rights?	
	What security protocols, such as LDAP, do you support?	
	What graphical techniques are provided for visualising profiles?	
	It is assumed that you offer dashboard capabilities for monitoring on-going data quality. Are there any special capabilities of this that you wish to mention?	
	Do you offer specific interfaces for business analysts? For data stewards?	
	If so, are these 100% thin-client based? And, if that is the case, is there any limit on the functionality provided—if so, please elaborate.	
	Similarly, can dashboards be viewed by a web interface?	
	What report formats are supported?	
	Is there support for user-defined reporting of DQ scores—aggregations at multiple levels, weightings for business-relevance reporting etc.?	
	Can alerts/notifications be configured on exceptions?	
	Do you provide capabilities for identifying/locating Personally Identifiable Information (PII) and/or other sensitive data? If so is customisation allowed?	
	Can you identify embedded values? e.g. offensive words in an address or credit card numbers in a comment.	
	Do you have the ability to discover the location of critical data elements?	
	Is there support for user-defined documentation, annotations?	
	Do you offer any specific vertical industry capabilities that you wish to detail? If so, please do so here	
Other	We do not believe that there is any significant difference between the error statistics capabilities (i.e. % of nulls, invalid datatypes etc.) of different products in this market. If you think differently please explain why. Note that we are not talking about presentation capabilities or ease of use, just functionality.	
	Please describe any further capabilities of your product that you believe to be important:	

© 2012 Bloor Research

Bloor Research overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations. We explain how to bring greater Agility to corporate IT systems through the effective governance, management and leverage of Information. We have built a reputation for 'telling the right story' with independent, intelligent, well-articulated communications content and publications on all aspects of the ICT industry. We believe the objective of telling the right story is to:

- Describe the technology in context to its business value and the other systems and processes it interacts with.
- Understand how new and innovative technologies fit in with existing ICT investments.
- Look at the whole market and explain all the solutions available and how they can be more effectively evaluated.
- Filter "noise" and make it easier to find the additional information or news that supports both investment and implementation.
- Ensure all our content is available through the most appropriate channel.

Founded in 1989, we have spent over two decades distributing research and analysis to IT user and vendor organisations throughout the world via online subscriptions, tailored research services, events and consultancy projects. We are committed to turning our knowledge into business value for you.

About the author

Philip Howard
Research Director - Data Management

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 his own company 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 focused on Data Management.

Data management refers to the management, movement, governance and storage of data and involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration (including ETL, data migration and data federation), data quality, master data management, metadata management and log and event management. Philip also tracks spreadsheet management and complex event processing.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to IT-Director.com and IT-Analysis.com and was previously 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 written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), dining out and walking Benji the dog.

Copyright & disclaimer

This document is copyright © 2012 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