Bloor

Market Report

Market Report Paper by Bloor
Author **Philip Howard**
Publish date **January 2014**

‘Free’ Data Profiling Tools

Experian is one of the leading vendors in this market and it will come as no surprise that Experian Pandora is the highest rated of the products covered here.

Author **Philip Howard**

Introduction

There are a number of data profiling and discovery tools that are available for free download. Needless to say, a number of these are open source products but there are also, notably, proprietary vendors that make their products available as free to use offerings, albeit sometimes with restrictions on who can use them and how. In addition, of course, there are also suppliers that make their products available on a try-before-you-buy basis. In other words, you can use the product for free for a limited period of time.

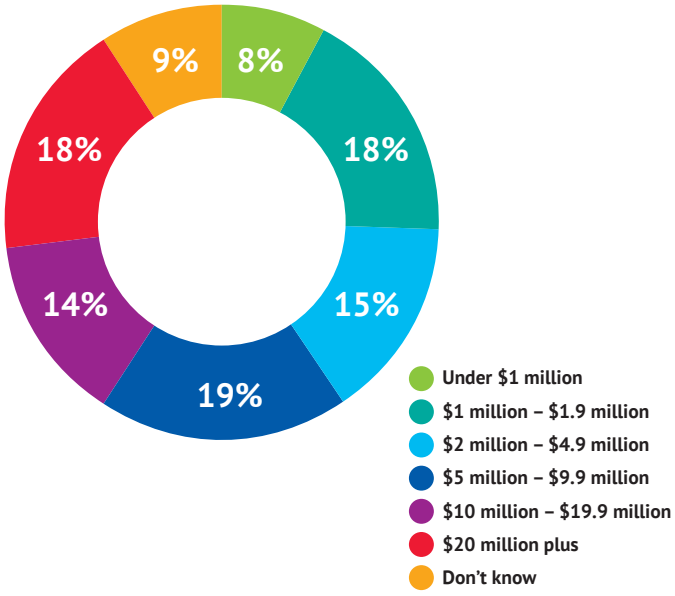
Such offers are commonplace across the software industry but they are especially apposite within the context of data profiling and discovery. It is worth considering why that is before we compare the various free-to-use products that are available.

Why profile?

Profiling and discovery software does three things:

1. It can analyse a database, or subset of a database, for errors.
2. It can monitor a dataset for errors on an on-going basis, typically presenting the results via a dashboard.
3. It can establish (discover) the relationships that exist between data elements, not just within a single database but also across and between heterogeneous data sources.

Figure 1:
What do you estimate that data-related issues cost your company annually?



The second of these capabilities is unlikely to be used by people just trying out the software, as opposed to those that are already committed to a data quality or governance initiative and, for this reason, we will confine our discussion to the first and third of the three things that profiling and discovery software can do. Moreover, we will discuss these capabilities from both a business and an IT point-of-view although we should advise readers of either persuasion to read both of the following sections and not just that which pertains to their role.

The business perspective

Despite the fact that data quality products have been in existence for the best part of 20 years it is still the case that a great many companies do not have any data quality or governance initiatives in place or, where they do exist, they are limited to a single (usually marketing) department.

According to "The data revolution – liberating lost budget" a report published in 2012 by Experian, "nearly 90% of companies admitted to wasting departmental budget as a result of duplicated mailings, lost contacts and missed sales opportunities, which is all down to inaccurate data. Departments such as Marketing, Sales, Operations and Customer Services report wasting 15% of their budget on average, while in IT and Data Management this rises to 18%, or about a sixth of the overall budget.

In customer-focused areas, the cost of poor data quality is particularly high. More than 80% of the companies quizzed operate customer loyalty programmes, and two thirds of these report that inaccurate data has had a negative impact on their programmes. In terms of lost custom and deteriorating reputation, this has a significant impact." Conversely, "companies investing in improving data accuracy believe that they generate an average of nearly £1 million in additional profit."

There are endless such examples of various bodies estimating the costs of poor quality data. The result of another, conducted by Forbes on behalf of SAP, is illustrated in **Figure 1**. Unfortunately, despite this wealth of evidence there remain significant numbers of executives who think that this is just an IT problem and not a business issue. As the examples quoted demonstrate, this is not the case.

A major potential benefit of free profiling is therefore to demonstrate to an unwilling management that there are issues with corporate data that are directly affecting the performance of your organisation. You should be able to install, get running and start profiling within less than a day so that, within a very short period of time and

with negligible cost (just some time), data irregularities can be demonstrated and, with appropriate software, you can assign values to data quality, not only so that you prioritise remediation but also in order to estimate the value of such a process. Your measurement of real data quality issues should allow you to justify the establishment of appropriately funded business-focused, business-sponsored data quality and data governance processes.

The IT perspective

One of the difficulties with justifying investment in data quality is precisely that business executives often deny the possibility of error, or underestimate its extent, or decry its potential benefits. They are also inclined to think that it is an IT problem. As the preceding discussion demonstrates this is not the case. However, there is some truth in it, as some IT functions will fail or run over-budget because of data quality issues.

The most well-known example of an IT function in which data quality is an issue is data warehousing. TDWI (The Data Warehousing Institute), back in 2002, estimated that poor data quality in data warehouses and data marts cost American businesses $600bn per annum. More recent estimates suggest that that figure is now around $700bn. Now, this may be an IT function but we would argue that with figures that big it is actually a business issue.

Another IT function that is really a business issue but is often relegated to IT and which depends on data profiling and discovery, is data migration. That is, where you are migrating from one version of an SAP, Oracle or similar application to another, when you are migrating from one database to another or when you are consolidating databases. Such projects can fail or significantly overrun their budgets if data quality is not up to scratch. It is important to understand why.

There are two types of data error. Data may be incorrect or invalid. For example, an email address may be wrong but it may still be in the correct format. Alternatively, that email address may not be in the right format at all (for instance, using "&" instead of "@"). If you load an email address that is valid but incorrect into your new system then that will not impact the running of that system – it may depreciate its value, that's what we have previously been discussing, but it will still run – but if you load an invalid value then the load may fail or, in the worst instance, it could actually crash the new system.

For this reason it is best practice to profile your data before you even start your migration (or archival: the same principles apply) project in order to determine the scale of the data quality problem you face. It is only after you have done that that you can reasonably estimate the resources and costs that the migration project is likely to require. Thus, profiling your data using a 'free' tool prior to a project of this type is another good use case.

One final consideration relates to the third – discovery – capability associated with profiling tools. In data migration (and archival) projects it is important to migrate 'business entities' – by which we mean, for example, a customer with his invoices, service history, delivery addresses and so on – as a whole. However, discovering what constitutes a business entity is a non-trivial exercise and it is something that is enabled by this discovery aspect of profiling tools. It should be noted that some vendors (see next section) have put much more effort into extending their discovery capabilities than others and this becomes especially critical in distributed environments where you are, perhaps, consolidating data from multiple different sources and understanding relationships across those sources becomes more complex.

The products

> We believe that limited time, try before you buy, offers are most suitable when you have already decided that you are going to go ahead with a data quality, migration or governance project and you want to decide which product to use as opposed to establishing that you have a data quality issue in the first place.

We should re-iterate that we are here concerned with free to use software as opposed to a limited period free trial offer. While they might appear to be comparable, we do not believe that is the case. The advantage of free software is that you can download the software when you want to and you can trial it when you want to. If you have a limited period offer then you have to be committed to use the software within that period; if something else comes up that you urgently need to attend to you can't put your data profiling software aside to be used when you feel like it, you have to use it now or you will lose it. You can of course extend a trial by contacting the sales team of the software vendor, but that may involve a lot more than a simple phonecall or email. With free to use software you can trial it whenever you need to. Moreover, if you are using the software to demonstrate to executives that you have a data quality problem that needs to be addressed then they will take more convincing than a single run through of the data. It is entirely possible that once the matter has been raised then it will be escalated to different levels within the organisation, which may require additional evidence. This whole process, like it or not, may take months. Thus we believe that limited time, try before you buy, offers are most suitable when you have already decided that you are going to go ahead with a data quality, migration or governance project and you want to decide which product to use as opposed to establishing that you have a data quality issue in the first place.

Having said all of this it would be disingenuous not to at least mention the vendors offering a free limited time trial version of their products and these are, most notably, CloverETL, Datiris and DataLynx. However, we will focus on the free to use market, which consists of:

Talend, which is the leading open source vendor in this market.

Ataccama, a proprietary vendor that makes its data profiling software free-to-use as an encouragement for those users to license its data quality software. Note that this (but not the free download) is also available from iWay (a division of Information Builders), which OEMs Ataccama's software.

Experian. A proprietary supplier of data profiling, discovery, quality and migration software. Its free-to-use software is a constrained version of the Pandora Profiling Edition, which only runs on Windows, the licence is for a single user, it is limited to 50 tables with no more than one million rows per table and the licence is, in fact, limited to an automatically renewable six months period. One further point is that the repository created with the free profiler can be carried forward and used by the full-function product. This is because they are actually the same product but controlled by licence key.

SQL Power, which is an open source provider of tools to support data warehousing. As we shall see, SQL Power Architect is a data modelling tool that has some data profiling features but the emphasis is very much on data modelling with profiling as a 'nice to have'.

DataCleaner is another open source tool, which, as its name suggests, covers the full gamut of data quality capabilities. Like Talend and SQL Power it offers Professional and Enterprise Editions in addition to its Community Edition. However, the company does not provide commercial support for its product, which is instead provided by Human Inference, the Dutch MDM specialist. One notable downside of this product is that in the Community Edition information is stored in a file-based repository whereas the other two editions use a database-based repository. This will impair performance in the Community Edition and make migration to one of the paid-for options more complex.

Open Source Data Quality and Profiling. This is downloadable from SourceForge. In other words, there is no commercial version or support for this product. That makes it difficult to get information about the product – indeed, experience has proved that the founders and developers of small open source projects do not

have the time or inclination to respond to analyst's requests for information – so we are having to rely on publicly available information here. In fact, we had initially thought that this was the same product as DataCleaner but since the product versions numbers are different we have had to assume that this is not the case.

AMB Data Profiling. This is another open source project downloadable from Source Forge. It is an open source project started in 2010 but its web site has not had an update in nine months, which suggests that nothing much is happening with it.

In the Bullseye diagram that follows, we have relied on Bloor Research's recently published Market Update on Data Profiling and Discovery for the positioning of Talend, Ataccama and Experian. However, it is worth briefly outlining the strengths and weaknesses of these products:

Ataccama: there is no feature of this product that is not also available in one or other, or both, of the other two products.

Talend: has extensive support for NoSQL, will run on a Hadoop platform, has support for non-normalised and specialised data-types (such as nested tables and small integers respectively – both in Oracle databases) and support for COBOL copybooks. In terms of functionality the product is not as rich as Experian Pandora and, where it does have equivalent features these may be in its data quality rather than its data profiling product.

Experian: lacks the source support offered by Talend and relies wholly on JDBC. Offers a number of features not included in either of the other two products, such as overlap and precedence analysis, discovery of matching keys, the ability to discover personally identifiable information (PII) and similar data, support for the definition and use of business terms and the ability to construct reference data based on profiling results. We would also expect Pandora to out-perform either of the other two, based on its architecture.

However, the remaining products were not included in that paper and nor has Bloor Research previously published any material relating to any of these products, so it is appropriate to also briefly discuss these. The first thing to say is that all of these lack cross-database discovery capabilities and they also lack support for joins, patterns (some of them, DataCleaner is an exception), or drilldown / navigation and there is no support for multi-column key analysis and relationship discovery / analysis / validation. DataCleaner does have some nice visualisations. Both this and the Open Source Data Quality and Profiling software have support for some NoSQL sources but it is unclear whether this includes data profiling as well as cleansing. In the case of the latter product it is SQL-based, which suggests not. It would also imply that profiling of non-relational sources such as CSV or Excel files (which are supported for cleansing purposes) will be difficult.

Figure 2: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator segment if their innovation rating is over 2.5 and Challenger if it is less than 2.5. The exact position in each segment is calculated based on their combined innovation and overall score.

Conclusion

> **You should be able to install, get running and start profiling within less than a day and most of the products offer training videos and documents, so it is possible to try them out quickly to see if they suit your needs.**

As mentioned previously, you should be able to install, get running and start profiling within less than a day and most of the products offer training videos and documents, so it is possible to try them out quickly to see if they suit your needs. More specifically, a great deal will depend on how likely you think your company is to adopt a data quality programme. If that is a real possibility then being able to carry forward your results into production systems could be a significant factor, as could commercial support.

These considerations will tend to mean that one of the three major vendors (Talend, Experian and Ataccama) should be preferred. Conversely, if the chances are remote then you can only go on the features of the products as they stand. Of course, Pandora will be ruled out if you want to exceed the constraints that Experian places on its free-to-use version or if you want to work with NoSQL databases but otherwise Experian is one of the leading vendors in this market and it will come as no surprise that Pandora is the highest rated of the products covered here.

Nor should it be a shock that Talend and Ataccama (in that order) are more richly featured than any of the other products featured in this paper. More generally, Talend may be more familiar and reassuring to a technical audience whereas Experian's ease of use is appreciated by less technical data analysts. If we were going to try a product other than from the three major vendors, it would be DataCleaner. SQL Power Architect is an interesting product but not for data profiling per se.

FURTHER INFORMATION
Further information about this subject is available from
www.bloorresearch.com/technology/data-profiling-discovery/

About the author

PHILIP HOWARD
Research Director / Information 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 foreign travel.

Bloor overview

Bloor Research is one of Europe's leading IT research, analysis and consultancy organisations, and in 2014 celebrated its 25th anniversary. 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 25 years 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.

Copyright and disclaimer

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

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloor.eu**
email: **info@Bloor.eu**