

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 position 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,

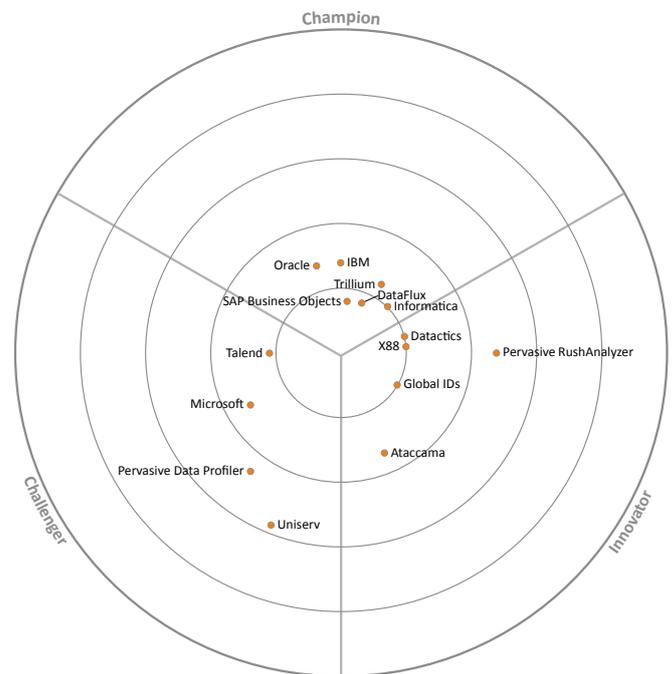


Figure 1: Data Profiling Landscape

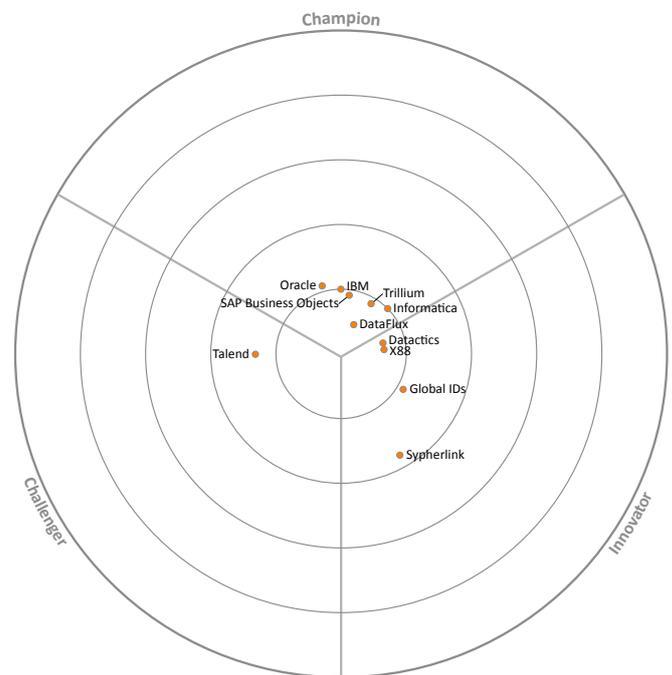


Figure 2: Data Profiling & Discovery Landscape

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.

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.

The vendors

There is not a great deal to distinguish pure-play profiling tools from one another. They are there principally to providing supporting capabilities for data cleansing and there is no appreciable difference in the statistics that the different vendors provide. Where there is a difference is that a couple of the vendors (Ataccama and Uniserv) provide data quality monitoring through separate tools rather than within the profiling tool itself, which is the approach typically adopted by other vendors. Ataccama, in particular, has taken this to extremes by offering its product as a free download while Microsoft's Data Quality Services (DQS) is also free, in the sense that it is bundled within SQL Server 2012.

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 Talend) to cost significantly more.

One of the biggest differences between products is with respect to integration. Some companies (Trillium, Oracle [because it resells Trillium] and x88 are examples) 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 (which 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, followed by 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.

For more details of the individual products, a discussion of the issues involved in this market, a copy of the questionnaire used in compiling this update, and for a breakdown of each product's scores against relevant criteria: see the [Bloor Research Market Report](#) on this subject.

*Philip Howard
Research Director, Data Management
Data Profiling & Discovery
June 2012*