MARKET UPDATE

Data profiling and discovery

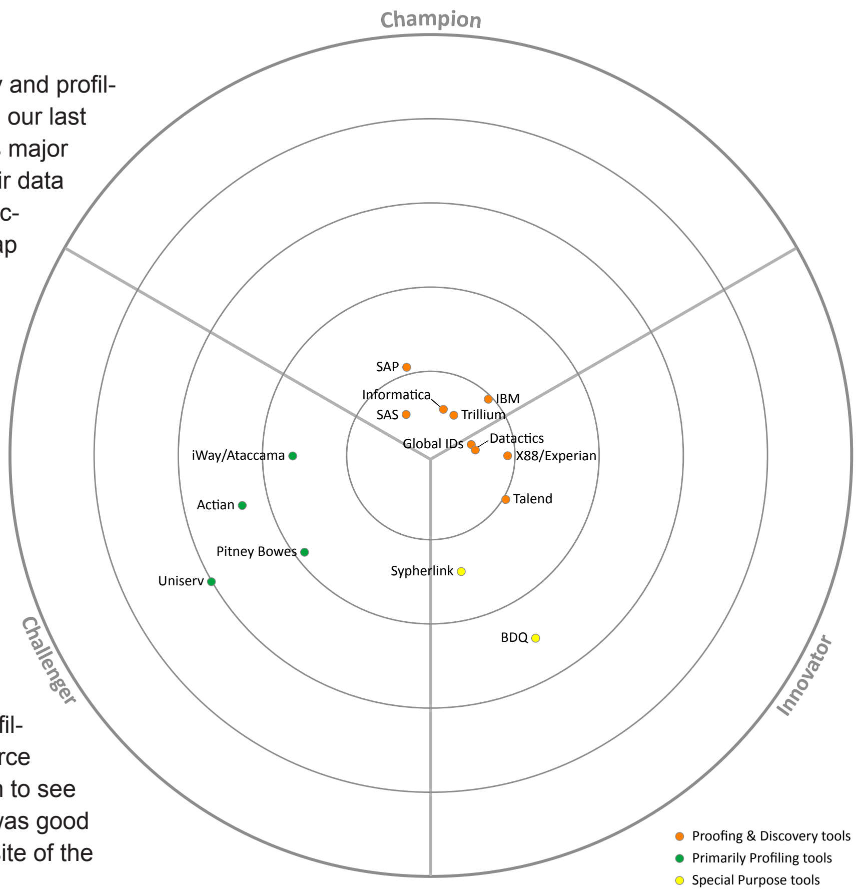

Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator segment if their innovation rating is 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.

Market trends

There are two major trends within the data discovery and profiling market at present. The first is that, as reported in our last Market Update, there continues to be a shift towards major vendors adding more discovery capabilities onto their data profiling capabilities. Similarly, there remains a distinction between those suppliers that have made this leap and those that appear content to simply offer data profiling. Certainly, there are use cases for this but the trend is definitely towards a broader set of capabilities. Note that there are pure-play discovery products that could be used to complement single function profiling tools.

The second major trend is with respect to NoSQL databases and, especially, Hadoop. There is a chasm here between vendors who have implemented significant support for big data platforms and those that have not. In some cases suppliers have dipped a toe into the water but have little real capability. What we want to see is not just the ability to profile NoSQL data but also to be able to host profiling databases (where the data is extracted from source systems) on Hadoop or HDFS. Not all vendors seem to see the point of this—one claimed that its performance was good enough not to need to—which is precisely the opposite of the point: it's about cost savings.

One other major difference between products, which is perhaps more of a philosophical position than a trend, is between those companies that have extensive native connectors and adapters for different sources and those that rely on JDBC/ODBC. It is noteworthy that many of the laggards in supporting NoSQL are also those companies that rely on these standard interface products. In our view, the use of native adapters will provide superior performance.

An interesting area that is almost orthogonal to this market is that we are starting to see the growth of profiling tools for other purposes than those traditionally associated with data quality, data migration and so forth. For example, Business Data Quality (BDQ) has introduced a profiling tool that is more focused on allowing you to measure and report on the performance and compliance of your data versus user-defined tests. Another example is Grid-Tools, which has a data profiling capability (not available for stand-alone use) to support synthetic test data generation. We rather suspect that we may see more such: for example, tools specifically designed to support data masking, or built into such products.

One trend that we had hoped to see emerge has not done so. In our last Market Update we reported that one vendor (Datactics) was introducing the option of in-database profiling capabilities. That is, profiling in the database kernel in the same way that in-database analytics has to become standard for processing queries in your data warehouse. This should provide much improved performance and, at least in some instances, mean that in-situ profiling (which we have never liked because of the performance implications) becomes a viable proposition.

Market position

Because of the marked difference in capabilities between those companies that just provide profiling and not discovery and between those that do and do not support NoSQL environments, we have included over the page two market maps, the first illustrating the different sectors the various vendors are involved in and the second the strength of their support for big data (NoSQL) environments. In addition, there is a Bullseye chart that illustrates different supplier's strengths. We have omitted Oracle from all of these because a) the company did not respond to our requests for information and b) although the company offers its own profiling capabilities it also continues to resell Trillium Software products. It is therefore unclear as to the company's approach to this market. We can say that there are no references to discovery in Oracle's publicly available literature—at least that we can find—so it is reasonable to assume

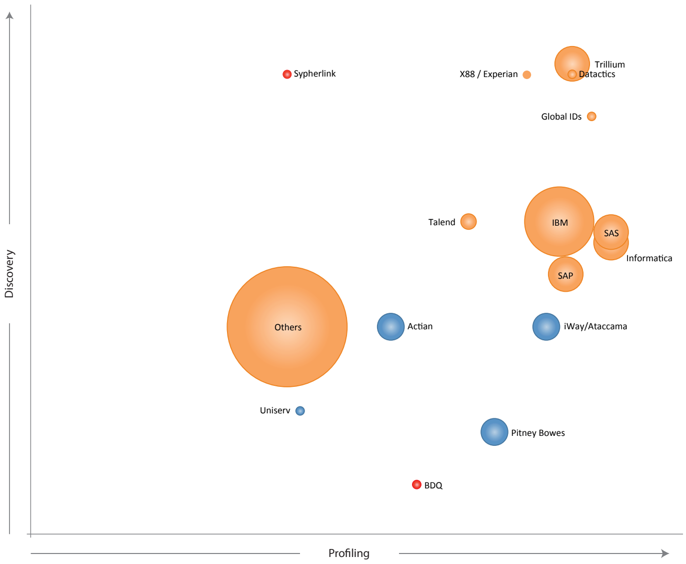

Figure 2: The size of the circle in the above diagram is an indication of the size of the company in terms of the marketplace and gives an indication relative to the others in the space. The stronger the vendor is in terms of profiling means they are more to the right of the diagram; the stronger vendors in discovery are towards the top. We have colour-coded the circles as: Orange - Profiling and Discovery tools, Blue - primarily Profiling tools and Red - Special Interest tools.

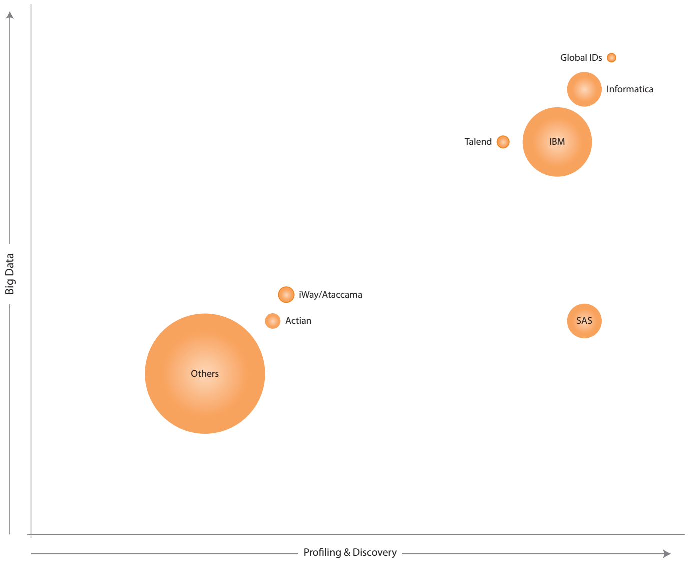

Figure 3: The size of the circle in the above diagram is an indication of the size of the company in terms of the marketplace and gives an indication relative to the others in the space. The stronger the vendor is in terms of profiling and discovery means they are more to the right of the diagram; the stronger vendors in big data are towards the top.

that its native capabilities are limited to conventional profiling. The same thing also appears to be true for Microsoft in that there is no mention of anything other than conventional profiling in its DQS (Data Quality Services) documentation. In neither case is there any reference to support for NoSQL sources.

We should clarify a few points, particularly that both Uniserv's and Business Data Quality's products are newly released, that Actian (Pervasive—acquired by Actian since our last report) actually has two products in this space and that iWay and Experian Data Quality embed Ataccama and X88's products respectively.

Summary

While there are trends towards including more discovery capability and towards supporting NoSQL data sources the vendors are significantly fragmented at present. Only a relative handful of suppliers have extensive support for the latter—particularly if you are interested in anything beyond Hadoop—while there remain a number of vendors that do not offer much in the way of discovery and the ability to understand relationships in and across data sources. Indeed, even across the leading companies in this market, the top 5 vendors for profiling only average 9% higher scores than the next 5; conversely, the leading 5 providers of discovery capabilities average 27% better ratings than the next 5 suppliers. In effect, profiling is on the way to becoming commoditised but that is very far from true when it comes to discovery.

Philip Howard
Research Director, Data Management
Data Profiling & Discovery
December 2013